MASTER

APPLIED ECONOMETRICS AND FORECASTING

MASTER'S FINAL WORK

DISSERTATION

# CAUSAL INFERENCE IN HIGH DIMENSIONAL PANEL MODELS: AN APPLICATION OF DOUBLE MACHINE LEARNING ESTIMATOR TO LABOR MARKET EFFECTS OF IMMIGRATION

GABRIEL DIAS MEDEIROS PEREIRA

SUPERVISION:

ISABEL PROENÇA

SEPTEMBER - 2024

# Abstract

This study focuses on estimation and inference in panel data models with unobserved individual-specific heterogeneity in a high-dimensional context. The framework accommodates scenarios where the number of regressors is comparable to the sample size. Crucially, we model the individual-specific heterogeneity as fixed effects, allowing it to correlate with observed time-varying variables in an unspecified manner and to be non-zero for all individuals.

Within this setup, we propose methods that provide uniformly valid inference for coefficients on a predetermined vector of endogenous variables in panel data instrumental variables (IV) models with fixed effects and numerous instruments. Central to the development of these methods is the application of machine learning algorithms within a semiparametric regression framework, enabling estimation in a grouped data structure where inter-group independence is assumed, and intragroup dependence is unrestricted. Simulation results support the theoretical framework, and we demonstrate the application of these methods in estimating the impact of immigration by non-European Union (EU) citizens on the employment of EU natives.



Keywords: panel data, machine learning, instrumental variables, high dimensional-sparse regression, clustered standard errors

JEL Codes: C23; C45; C55; C36; C52; C12.

# Glossary

**2SLS** Two stage least squares.

**CV** Cross-validation.

**DML** Double Machine Learning.

**EU** European Union.

**IV** Instrumental Variables.

**ML** Machine Learning.

**OLS** Ordinary Least Squares.

**RMSE** Root Mean Square Error.

Table of Contents

**Abstract** **1**

**Glossary** **2**

**Table of Contents** **3**

**List of Figures** **5**

**List of Tables** **6**

**1 Introduction** **7**

**2 Literature Review** **9**

2.1 Causal Inference and Machine Learning . . . . . 9

2.2 Labor market effects of immigration . . . . . 11

**3 Methodology** **13**

3.1 Frisch-Waugh-Lovell theorem . . . . . 13

3.2 Neyman Orthogonality and Moment Conditions . . . . . 14

3.3 Sample Splitting . . . . . 14

3.4 Hyperparameter Tuning . . . . . 15

3.5 Double Machine Learning . . . . . 17

3.6 LASSO . . . . . 19

3.7 Random Forests . . . . . 20

3.8 Boosting . . . . . 21

**4 Simulation Study** **22**

4.1 Simulation Design . . . . . 22

4.2 Simulation Results . . . . . 23

**5 Empirical Analysis** **27**

5.1 Data . . . . . 27

5.2 Two Stage Least Squares Regression Model . . . . . 28

5.3 Double Machine Learning Regression Model . . . . . 31

5.4 Labor market effects of immigration in EU countries . . . . . 32

**6 Conclusion** **37**

**Appendices** **43**

**Appendix A Appendices** **43**
A.1 Descriptive statistics . . . . . 43
A.2 Hyperparameters configuration . . . . . 44
A.3 Simulation study model comparison . . . . . 49
A.4 Empirical study model comparison . . . . . 51
A.5 Implementation . . . . . 51

## List of Figures


1 The cross-validation method. Based on (Pedregosa et al. 2011). . . . 15
2 The causal graph representation for IV DML, based on (Shao et al. 2024). . . . . 18
3 Simulation metrics for the different estimators. Author's own elaboration. . . . . 26
4 Panel A: Without interaction with institutions. Author's own elaboration. . . . . 34
5 Panel B: With interaction with institutions. Author's own elaboration. 35
6 Simulation bias for different panel specifications. Author's own elaboration. . . . . 49
7 Simulation distributions for different panel specifications (N = 15, 50, 100, 200). Author's own elaboration. . . . . 50
8 Notes: Author's own elaboration. RMSE results for the three different ML model specifications . . . . . 51

## List of Tables


I Hyperparameter Tuning . . . . . 17
II Panel IV, $p = n \times (T - 2)$ . . . . . 25
III Effect estimates of the non-EU immigrants on employment of EU natives . . . . . 33
IV Descriptive Statistics by Country . . . . . 43
V Descriptive Statistics for Barriers to Entrepreneurship, Labor Standards, and Replacement Rate . . . . . 43
VI Selected Hyperparameters for Models (Empirical Analysis) . . . . . 44
VII Selected Hyperparameters for Simulation Study (N = 15) . . . . . 45
VIII Selected Hyperparameters for Simulation Study (N = 50) . . . . . 46
IX Selected Hyperparameters for Simulation Study (N = 100) . . . . . 47
X Selected Hyperparameters for Simulation Study (N = 200) . . . . . 48

# 1 Introduction

The intersection of econometrics and machine learning has been progressing quickly, particularly in addressing causal inference questions within high-dimensional datasets. While empirical economic research traditionally relies on linear models and manual variable selection to mitigate biases from omitted variables, the introduction of machine learning methods provides a complementary approach by automating and improving these processes. This thesis explores the application of Double Machine Learning (DML) estimators in high-dimensional panel models, specifically focusing on labor market effects of immigration.

One of the primary objectives in empirical economics is to determine the causal effect of specific variables on outcomes of interest. Traditional regression models often struggle with high-dimensional data, where the number of potential covariates can be comparable or even exceed the number of observations. This challenge is particularly pronounced in observational studies where the inclusion of numerous controls is necessary to avoid biased estimates. Machine learning methods, designed to handle large datasets and complex interactions, present a viable solution by automating the selection of relevant variables and regularizing estimates to prevent overfitting. However, standard machine learning techniques prioritize predictive accuracy over causal inference, potentially introducing biases when applied directly to economic models.

Recent econometric literature has begun to bridge this gap by adapting machine learning methods for causal inference. Notable contributions include the works of Chernozhukov et al. (2018), Athey et al. (2018), which integrate machine learning algorithms with econometric models to enhance the reliability and interpretability of causal estimates. These approaches, particularly Double Machine Learning, leverage the strengths of machine learning for variable selection and regularization while maintaining the econometric focus on causal inference. Despite these advancements, the empirical economics literature has yet to fully embrace these methods.

According to Wooldridge (2010), panel data is widely used in economics because it not only allows for the analysis of dynamic relationships over time but also helps in controlling for unobserved heterogeneity across individuals. This advantages make it a powerful tool for studying complex economic phenomena that evolve across both time and entities. High-dimensional panel data, which includes a large number of time-varying covariates, pose additional challenges. Traditional methods, such as the linear fixed effects model, may struggle with these datasets due to the potential overfitting and multicollinearity issues that arise when the number of covariates is

comparable to the number of observations. The introduction of approximate sparsity and regularization techniques, as highlighted by Belloni et al. (2014*a*), provides a framework for addressing these issues by focusing on a smaller subset of relevant variables.

In high-dimensional settings, causal machine learning methods such as DML offer systematic approaches to model selection and inference. These methods allow researchers to handle a large number of covariates, including nonlinear transformations and interactions, ensuring that important confounders are not omitted. By combining machine learning techniques with traditional econometric models, DML provides a comprehensive tool for causal inference that is data-driven and allows valid statistical inference (Chernozhukov et al. 2018).

To address the question of how immigration impacts labor market outcomes in high-dimensional settings, this thesis applies the DML framework to examine these effects in European Union countries, revisiting the study made by Angrist & Kugler (2003). This application is significant because immigration studies often involve high-dimensional datasets, with numerous demographic, economic, and social variables influencing both the extent of immigration and labor market outcomes. Traditional econometric methods may fall short in capturing these complex relationships, leading to biased or incomplete estimates. By employing DML, this research aims to provide more accurate and reliable estimates of the causal impact of immigration on labor market variables such as employment, wages, and job displacement.

This thesis brings two main contributions. First, a Monte Carlo simulation study is conducted for high-dimensional panel data with fixed effects and endogeneity. This study follows the simulation framework of Belloni et al. (2016) and Chernozhukov et al. (2015) but extends it by incorporating various machine learning models—Random Forests, Boosting, and LASSO—within the DML framework. This extension allows for a comparative analysis of these models' performance in handling high-dimensionality and endogeneity, offering insights into their applicability and effectiveness in econometric research.

Second, the thesis revisits the Angrist & Kugler (2003) study on the labor market effects of immigration, expanding the model specification by adding more variables and estimating a DML model using the machine learning methods mentioned above. This revision aims to capture a more comprehensive set of covariates and interactions, leading to improvements for the accuracy and robustness of the causal estimates. By doing so, the thesis not only validates the use of DML in a real-world empirical context but also contributes to the empirical literature by providing

alternative estimates and methodologies for assessing the impact of immigration on labor market outcomes.

By integrating advanced machine learning techniques with traditional econometric models, this thesis aims to contribute to the growing field of causal inference in high-dimensional settings, offering new insights into the labor market effects of immigration and demonstrating the value of DML in empirical economic research. This research not only addresses the challenges of high-dimensional data but also provides a novel application of DML to a pertinent policy issue, highlighting the potential of machine learning methods to enhance econometric analysis and policy decision-making.

The thesis is structured as follows: Chapter 2 reviews the relevant literature on high-dimensional econometric models and machine learning methods for causal inference. Chapter 3 details the methodological framework of Double Machine Learning and its application to panel data. Chapter 4 presents the findings of the Monte Carlo simulation analysis. Chapter 5 discusses the empirical analysis, including data description, model specifications, results, and their implications for policy and future research. Finally, Chapter 6 concludes with a summary of the key contributions and potential avenues for further study.

# 2 Literature Review

## 2.1 Causal Inference and Machine Learning

Recent advancements in machine learning have significantly influenced econometric methods, particularly in addressing high-dimensionality problems in causal inference. As Belloni et al. (2014*a*) mention, nowadays, high-dimensional problems have been arising through a combination of two phenomena: 1) hundreds or even thousands of individual characteristics are being collected by surveys and from official bureaucratic institutions all over the world and 2) researchers rarely know the exact functional form with which the variables should be specified in the model. Researchers are then faced with a large set of potential variables formed by different ways of interacting and transforming the underlying variables.

One of the foundational works in semi/nonparametric regression was conducted by Frisch & Waugh (1933). They demonstrated through the Frisch-Waugh-Lovell theorem that a partially linear regression model can be decomposed, enabling the use of ordinary least squares estimates on the residuals. This theorem illustrates that after removing the effect of certain control variables, the relationship

between the remaining variables can be accurately estimated using standard OLS techniques. The theorem paved the way to more sophisticated methods to address the problem of high-dimensionality by allowing semi/nonparametric regressions. However, as outlined by Athey et al. (2018), the need of regularization methods (or dimensionality reduction) without proper care can introduce bias to causal inference estimates. For instance, Neyman (1979) introduced the orthogonality conditions in low-dimensional settings to manage roughly estimated parametric nuisance parameters, showing that the orthogonality condition could eliminate the bias introduced from regularization. In the seminal work of Robinson (1988), new methods for obtaining root n-consistent and asymptotically normal estimates for low-dimensional components within traditional semiparametric frameworks were demonstrated. However, as mentioned by Baiardi & Naghi (2021), kernel regressions often break down in the presence of a large number of covariate candidates. One of the first works to use orthogonality conditions with shrinkage methods to address the problem of regularization bias with the uniform post-selection inference in high-dimensionality was Belloni et al. (2011), using the LASSO model, as introduced by Tibshirani (1996). In Belloni et al. (2014*a*), it was proposed an augmented variable selection method to avoid this effect, starting the discussion of approximately sparse regression models in high-dimensional data, where it is shown that valid post-selection inference is generally available when estimation is based on orthogonal estimating equations. Further on, Belloni et al. (2017) provided a framework where any high-quality, machine learning methods (e.g., boosted trees, deep neural networks, random forest, and their aggregated and hybrid versions) can be used to learn the nonparametric/high-dimensional components.

In their 2022 work, Angrist & Frandsen (2022) identified three key areas where machine learning can enhance econometric research, particularly in labor economics. Firstly, they suggest using ML for data-driven selection of ordinary least squares control variables. The post-double selection lasso estimator introduced by Belloni et al. (2014*b*) effectively addresses this issue by selecting the most relevant control variables for OLS estimation. Secondly, ML can be employed for the choice of instruments in IV estimation. This approach is motivated by the bias inherent in two-stage least squares estimates (2SLS) in models with many instruments. Lastly, ML can aid in selecting control variables in IV models, especially when there are numerous potential control variables but only a few instruments available. These applications of ML help improve the precision and reliability of causal inferences in econometric analysis.

In their work, Chernozhukov et al. (2018) dealt with two common pitfalls

of introducing ML models naively for causal inference tasks: overfitting and regularization biases. Since ML algorithms have intrinsically regularization methods embedded and could be prone to overfitting, the authors introduced sample-splitting and also developed orthogonal moment functions to deal with both problems. Accordingly to Clarke & Polselli (2024), this bias occurs because machine learning algorithms typically minimize metrics like mean squared error, rather than directly addressing bias reduction, leading to either regularization or overftting. In Mackey et al. (2018), they describe DML as a two-stage process. In the first stage, nuisance parameters are estimated using various statistical ML techniques on an initial data sample. In the second stage, the low-dimensional parameters of interest are estimated using the generalized method of moments (GMM). A key requirement is that the moments in the second stage satisfy a Neyman orthogonality condition, providing robustness to errors in estimating the nuisance parameters.

The literature of high-dimensional panel models is vast. Belloni et al. (2016) allowed the assumption of clustered data structure, where data across groups are independent and dependence within groups is unrestricted for static panel models. Kock & Tang (2019) made contributions for high-dimensional dynamic panel models. Some interesting applications of DML for high-dimensional panel models can also be found at Klosin & Vilgalys (2023), where they applied machine learning algorithms to model high-dimensional relationships in a climate study. Moreover, they showed in a simulation study that, in applying the DML framework, the resulting modeling approach has low bias even in nonlinear settings. Furthermore, Semenova et al. (2022) also provided inference methods for high-dimensional dynamic panel settings with DML. Their procedure was composed by orthogonalization, where they partial out the controls and unit effects from the outcome and the base treatment and take the cross-fitted residuals. In this step, they advise that any machine learning method can be used (given that it learns the residuals well enough). The second step uses a novel generic cross-fitting method designed for weakly dependent time series and panel data. Additionaly, Clarke & Polselli (2024) consider causal estimation for static panels with DML in order to approximate high-dimensional and non-linear nuisance functions of the confounders, enabling to infer the effects of policy interventions from panel data.

## 2.2 Labor market effects of immigration

Angrist & Kugler (2003) examine the impact of immigration on native employment in Europe, focusing on how labor market institutions affect this relationship. Their key findings include that labor market rigidities and high

business entry costs tend to exacerbate the negative impact of immigration on native employment. They also mention that while some institutions can play a protective role, reduced labor market flexibility generally fails to shield natives from job losses due to immigration and may even worsen these effects.

Recent research suggests that immigration can have positive effects on native employment. Moreno-Galbis & Tritah (2016) found that the employment rate of natives increases in occupations and sectors that receive more immigrants. This effect is particularly pronounced for new immigrants and those from non-EU15 countries. The authors attribute this phenomenon to immigrants' lack of host-country-specific assets, which weakens their bargaining position with employers and consequently improves employment prospects for natives. Interestingly, the employment creation effect is more robust in countries where there are larger disparities in unemployment benefit take-up rates between immigrants and natives. To establish these findings, Moreno-Galbis & Tritah (2016) employed an instrumental variable approach based on historical settlement patterns across host countries and occupations by origin country, providing a robust methodological foundation for their conclusions.

Furthermore, D'Amuri & Peri (2014) found that immigration led to occupational upgrading for native workers, pushing them towards more complex (abstract and communication-intensive) jobs as immigrants filled manual-routine occupations. This job upgrade was associated with a 0.7% increase in native wages for a doubling of the immigrant share. The authors argue that this reallocation protected native wages from immigrant competition and allowed natives to benefit from the creation of jobs complementary to immigrants' manual tasks. The complexity of jobs offered to new native hires increased relative to the complexity of lost jobs (D'Amuri & Peri 2014). Also, according to the same authors, the reallocation process was stronger in countries with more flexible labor laws and this effect was particularly prominent for less-educated workers in flexible labor markets.

Moreover, in Ortega & Peri (2009) an instrument for migration flows was constructed that is exogenous to the economic conditions in the destination country. It was found that there was no evidence of crowding-out of native workers and that could even increase employment in receiving countries. Furthermore, the authors mention that investment responded rapidly and vigorously to immigration, where the capital adjusts to maintain the capital-labor ratio. They also highlight that the quick adjustment of capital is key in determining the short-run effects of immigration on wages.

# 3 Methodology

## 3.1 Frisch-Waugh-Lovell theorem

The Frisch & Waugh (1933) and Lovell (1963) theorem is a fundamental result in econometrics that provides another way to understand and compute the Ordinary Least Squares estimators in a linear regression model, opening possibiities for nonparametric regression as well.

Consider the linear regression model:

$$y = X_1\beta_1 + X_2\beta_2 + u$$

where:

- $y$ is the $n \times 1$ vector of the dependent variable,
- $X_1$ is the $n \times k_1$ matrix of regressors of interest,
- $X_2$ is the $n \times k_2$ matrix of additional regressors,
- $\beta_1$ and $\beta_2$ are the coefficient vectors,
- $u$ is the $n \times 1$ vector of error terms.

The Frisch-Waugh-Lovell theorem states that the OLS estimate of $\beta_1$ can be obtained by:

1. Regressing $y$ on $X_2$ and saving the residuals $M_2y$,
2. Regressing $X_1$ on $X_2$ and saving the residuals $M_2X_1$,
3. Regressing the residuals from step 1 on the residuals from step 2.

Here, $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$ is the projection matrix onto the orthogonal complement of the column space of $X_2$.

The resulting estimator for $\beta_1$ is:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2y$$

The Frisch-Waugh-Lovell theorem is closely related to the DML approach. Both involve the idea of partialing out the effects of confounders to obtain a more accurate estimate of the parameter of interest. In the context of DML, machine learning algorithms are used to estimate nuisance parameters, which are then used to orthogonalize the main regression problem.

## 3.2 Neyman Orthogonality and Moment Conditions

As already mentioned, the orthogonalization principle is fundamental to generate unbiased estimates under the use of regularization methods. The orthogonalized or double/debiased ML estimator $\tilde{\theta}_0$ solves

$$\frac{1}{n}\sum_{i\in I}\psi(W;\theta_0,\eta_0)=0, \tag{1}$$

where $\hat{\eta}_0$ is the estimator of the nuisance parameter $\eta_0$, $\psi$ is an orthogonalized or debiased "score" function and where $W=(Y,D,X,Z)$. The score function satisfies the property that the Gateaux derivative operator with respect to $\eta$ vanishes when evaluated at the true parameter values:

$$\partial_\eta \mathbb{E}\psi(W;\theta_0,\eta_0)[\eta-\eta_0]=0. \tag{2}$$

The proofs of the general results show that this term's vanishing is a key to establishing good behavior of an estimator for $\theta_0$ (Chernozhukov et al. 2018). This property is referred as "Neyman orthogonality" and to $\psi$ as the Neyman orthogonal score function.

## 3.3 Sample Splitting

In the Double Machine Learning framework, sample splitting serves as a fundamental technique to enhance model robustness and mitigate overfitting. Alongside employing orthogonal score functions for accurate identification and employing machine learning methods to estimate nuisance parameters, sample splitting adds an additional layer of rigor. This approach helps in reducing potential biases introduced by overfitting and ensures that the estimation of causal effects or predictive parameters is both reliable and generalizable (Athey & Imbens 2019).

Sample splitting involves partitioning the data into distinct subsets to separate the learning of nuisance functions from the estimation of causal effects. This separation is crucial because it allows the model to avoid overfitting biases that might arise if the same data were used for both tasks. Specifically, one portion of the data is used to train the nuisance parameter models, while another, separate portion is utilized for estimating the causal parameter. This division ensures that the samples used for learning the nuisance functions are independent of those used for evaluating the causal parameter, thereby improving the accuracy and generalizability of the model (Friedman et al. 2001).

Cross-fitting is an efficient data-splitting technique that utilizes the entire dataset by rotating the roles of training and holdout data across multiple iterations, as it can be seen in figure 1. This method maximizes data usage while preserving the independence between training and testing phases. The steps for implementing cross-fitting, as described by Chernozhukov et al. (2018), are outlined below:

1. Begin by randomly dividing the dataset into $K$ folds. Each fold, denoted $I_k$, contains approximately $n = \frac{N}{K}$ observations, where $N$ is the total number of observations. For each fold $k$, the complementary set $I_k^c$ is defined as $I_k^c = \{1, \ldots, N\} \setminus I_k$.

2. For each fold $k$, train an ML estimator $\hat{\eta}_{0,k}$ for the nuisance parameters using the observations in the complementary set $I_k^c$. The objective is to estimate the nuisance functions $\eta_0$ based on data not used in the causal parameter estimation.

3. For each fold $k$, use the trained nuisance model $\hat{\eta}_{0,k}$ to estimate the causal parameter $\hat{\theta}_{0,k}$ based on the observations in the fold $I_k$. The final estimator $\hat{\theta}_0$ is obtained by averaging the estimates across all $K$ folds:

$$\hat{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{0,k}$$

This averaging ensures that the causal parameter estimation benefits from the variability across different folds, leading to a more robust final estimate.

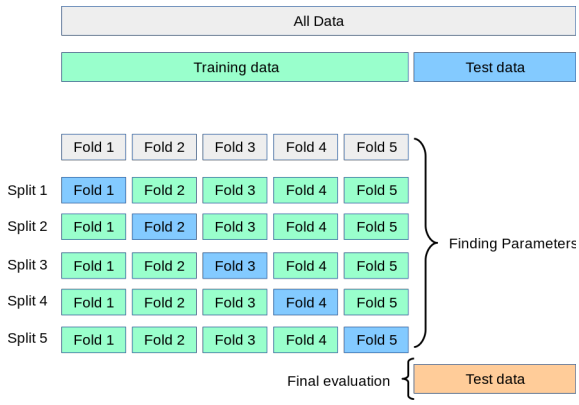


Figure 1: The cross-validation method. Based on (Pedregosa et al. 2011).

## 3.4 Hyperparameter Tuning

Identifying the best set of hyperparameters is crucial for maximizing the performance of machine learning models, particularly in the context of effect

estimation. Hyperparameters are settings or configurations that define the structure and behavior of machine learning algorithms. Unlike model parameters, which are learned during the training process (e.g., the coefficients in a linear regression), hyperparameters are set prior to training and guide the learning process itself. These values control aspects such as model complexity, learning speed, and regularization, influencing how well the algorithm generalizes to unseen data. Hyperparameter optimization involves experimenting with various combinations of hyperparameter values to find the most effective configuration (Bergstra & Bengio 2012). This process typically uses resampling techniques like cross-validation (CV) to assess the algorithm's performance based on metrics such as the RMSE. This procedure continues until a specified stopping criterion, such as a maximum number of evaluations, is met. The configuration yielding the best performance (e.g., the lowest RMSE) is then selected for training and testing the final model.

Failing to conduct hyperparameter tuning can considerably reduce the model's performance. Hyperparameters play a critical role in defining the behavior and capacity of a model. Examples include the learning rate and the number of boosting rounds in gradient boosting algorithms like XGBoost, the maximum depth of decision trees, and the minimum samples required to split a node in random forests. Without proper tuning, the model may not be able to capture the underlying patterns in the data effectively, leading to suboptimal results (Bach et al. 2024).

In the Double Machine Learning algorithm, hyperparameter tuning follows these steps (Clarke & Polselli 2024):

1. In methods like K-fold cross-validation, the training sample for fold $k$ ($W_k^c$) is used. This sample is further divided using methods like K-fold CV to create inner training and testing sets. Therefore, $k$-th CV fold acts as the test set while the remainder serves as the training set.

2. Models are tuned using a grid search method, which is a hyperparameter optimization algorithm. It evaluates the performance of base learners across various combinations of hyperparameter values. The optimizer conducts a random search over a predefined number of values for each hyperparameter and halts once a predetermined number of evaluations is reached.

3. Each assessment during the tuning process identifies the best set of hyperparameters by evaluating across all $k$ CV folds, focusing on minimizing the RMSE. Upon completion of the tuning procedure (for instance, after $j$ evaluations), the optimal set of hyperparameters—determined by the lowest RMSE—is selected from these $j$ evaluations and then applied in the DML

algorithm.

4. The best configuration is applied to the learners of the nuisance parameters. The model is trained on the complementary set for fold $k$ ($W_k^c$) and tested on $W_k$. Predictions for the nuisance functions $m$ and $l$ are stored.

For the present study, hyperparameter tuning was conducted with the following parameters for each method:

Table I: Hyperparameter Tuning

| Learner | Hyperparameters | Value of Parameter in Set | Description |
|---|---|---|---|
| Lasso | `lambda` | {0.001, 0.01, 0.1, 1, 10} | Penalty on the absolute values of the coefficients |
| Boosting | `n_estimators` | {50, 100, 200} | Number of boosting rounds (trees) in the model. |
| | `max_depth` | {10, 15, 20, 25} | Maximum depth of each tree. |
| | `learning_rate` | {0.01, 0.05, 0.1} | Step size shrinkage used in updating the weights. |
| RF | `n_estimators` | {50, 100, 200} | Number of trees in the forest. |
| | `max_features` | {50, 100, 200, 300} | Number of features to consider when looking for the best split. |
| | `max_depth` | {10, 15, 20, 25} | Maximum depth of any tree in the forest. |
| | `min_samples_leaf` | {1, 2, 4} | Minimum number of samples required to be at a leaf node. |
| | `ccp_alpha` | {0.0, 0.01, 0.05, 0.1} | Complexity parameter used for pruning. A higher value leads to more pruning, reducing overfitting. |

The selected hyperparameters for the simulation study and the empirical study can be found at tables VI, VII, VIII, IX and X.

## 3.5 Double Machine Learning

Here we extend the partially linear regression model to allow for instrumental variable (IV) identification in a panel data setup, as defined by Chernozhukov et al. (2018). Specifically, we consider the model illustrated at figure 2:

$$\begin{aligned} Y_{it} &= \theta D_{it} + g_0(X_{it}) + \mathcal{E}_{it}, \quad \mathbb{E}[\mathcal{E}_{it}|Z_{it}, X_{it}] = 0, \\ D_{it} &= m_0(X_{it}) + V_{it}, \quad \mathbb{E}[V_{it}|X_{it}] = 0, \\ Z_{it} &= \ell_o(X_{it}) + U_{it}, \quad \mathbb{E}[U_{it}|X_{it}] = 0 \end{aligned} \tag{3}$$

where $Z_{it}$ is a vector of instrumental variables, $D_{it}$ is a vector of endogenous variables, and $X_{it}$ is a vector of exogenous variables.

Additionally, $g_0(X_{it}) = \mathbb{E}[Y_{it} \mid X_{it}]$ and $m_0(X_{it}) = \mathbb{E}[D_{it} \mid X_{it}]$ and $\ell_0(X_{it}) = \mathbb{E}[Z_{it} \mid X_{it}]$. As before, the parameter of interest is $\theta$.

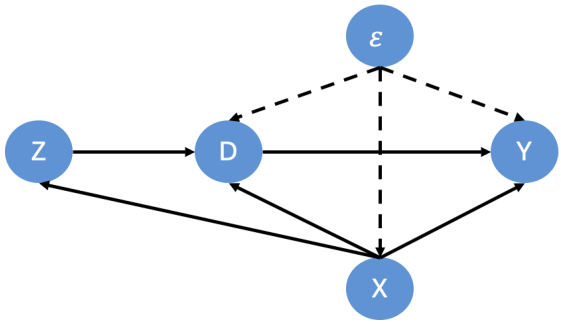


Figure 2: The causal graph representation for IV DML, based on (Shao et al. 2024).

To estimate $\theta$ and to perform inference on it, as outlined by Chernozhukov et al. (2024), the partialling-out operator is applied in order to obtain a parameter $\theta$ that is orthogonal to the nuisance components, $g_0(X_{it})$, $m_0(X_{it})$ and $\ell_o(X_{it})$. The partialling-out operator is defined by the same author as an operation where any random vector $V$ with respect to another random vector $X$ as the residual that is left after subtracting the best predictor of $V$ given $X$: $\tilde{V} = V - \mathbb{E}[V \mid X]$

Furthermore, it is possible to start with the following moment condition:

$$\mathbb{E}[(\tilde{Y_{it}} - \theta \tilde{D_{it}})\tilde{Z_{it}}] = 0 \tag{4}$$

where $\tilde{Y_{it}} = Y_{it} - g_0(X_{it})$, $\tilde{D_{it}} = D_{it} - m_0(X_{it})$ and $\tilde{Z_{it}} = Z_{it} - \ell_0(X_{it})$.

Therefore, as shown by Clarke & Polselli (2024), it is possible to derive the score function:

$$\frac{1}{N_k} \sum_{i \in \mathcal{W}_k} \psi_k^{\perp}(W_i; \theta, \widehat{\eta}_k) = 0 \tag{5}$$

$$\begin{aligned} \psi(W; \theta, \eta) &:= (Y_{it} - \theta(D_{it} - m_0(X_{it})) - g_0(X_{it}))(Z_{it} - \ell_0(X_{it})), \quad \eta = (m, g, \ell) \\ &= -(D_{it} - m_0(X_{it}))(Z_{it} - \ell_0(X_{it}))\theta + (Y_{it} - g_0(X_{it}))(Z_{it} - \ell_0(X_{it})) \\ &= \psi_a(W; \eta)\theta + \psi_b(W; \eta) \end{aligned} \tag{6}$$

where $W = (Y_{it}, D_{it}, X_{it}, Z_{it})$, $\eta = (m_0, g_0, \ell_0)$ are square-integrable functions mapping the support of $X$ to $\mathbb{R}$. Both scores satisfy the Neyman orthogonality condition, making them robust to biases in the estimation of nuisance parameters,

getting really close to the standard IV regression. Therefore, as demonstrated in Bach et al. (2022), the final estimator for $\hat{\theta}$ is:

$$\hat{\theta}_0 = \frac{\mathbb{E}_N[\psi_b(W;\eta)]}{\mathbb{E}_N[\psi_a(W;\eta)]} = \frac{\mathbb{E}[\tilde{Y}_{it}\tilde{Z}_{it}]}{\mathbb{E}[\tilde{D}_{it}\tilde{Z}_{it}]} \tag{7}$$

The final estimated target parameter $\hat{\theta}$ is the average over the $k$ folds. As illustrated by Klosin & Vilgalys (2023), to compute the asymptotic variance of the estimator, we take into account the correlation of the average derivative within panel units. The asymptotic variance, described by Clarke & Polselli (2024), is:

$$\hat{\sigma}_k^2 = \hat{J}_k^{-1} \left\{ \frac{1}{N_k} \sum_{i \in \mathcal{W}_k} \psi^{\perp}(W_{it};\theta,\hat{\eta}_k)\psi^{\perp}(W_{it};\theta,\hat{\eta}_k)' \right\} \hat{J}_k^{-1} \tag{8}$$

where $\hat{J}_k = N_k^{-1} \sum_{i \in \mathcal{W}_k} \hat{\psi}_a$ is the average derivative term for cluster $k$ in N folds.

## 3.6 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a popular regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Introduced by Tibshirani (1996), LASSO is particularly useful in scenarios where the number of predictors exceeds the number of observations or where multicollinearity is present.

LASSO regression minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients (Tibshirani, 1996). The LASSO estimate $\hat{\beta}$ is defined by:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{9}$$

where $y_i$ is the dependent variable, $x_{ij}$ are the predictor variables, $\beta_j$ are the coefficients, $n$ is the number of observations, $p$ is the number of predictors, and $\lambda$ is a non-negative tuning parameter that controls the amount of shrinkage applied to the coefficients.

The key feature of LASSO is its ability to shrink some of the coefficients to exactly zero when the tuning parameter $\lambda$ is sufficiently large. This property effectively selects a simpler model that retains only the most important predictors,

thus performing variable selection. The LASSO constraint is given by:

$$\sum_{j=1}^{p} |\beta_j| \leq t, \tag{10}$$

where $t$ is a constant that determines the amount of regularization. As $\lambda$ increases, more coefficients are shrunk to zero, leading to a more parsimonious model.

The DML framework often involves LASSO as a first-stage estimator for selecting relevant covariates and regularizing the estimation process to handle the high-dimensionality of the data (Chernozhukov et al. 2018).

## 3.7 Random Forests

Random Forest, introduced by Breiman (2001), is an ensemble learning technique designed to reduce prediction error by mitigating overfitting and lowering variance of decision trees. This is achieved by constructing a large number of de-correlated decision trees and combining their predictions. For each of the $B$ trees, a bootstrap sample of the data is taken, and a decision tree $T_b$ is built using recursive binary splitting. To ensure the trees are de-correlated, a random subset of $m$ variables is chosen from the total $p$ predictor variables, where $m \leq p$. The final prediction is made by averaging the results of all trees for regression tasks, and by majority voting for classification tasks.

In mathematical terms, for regression, the final prediction is computed as:

$$\hat{f}_{\mathrm{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x),$$

whereas for classification, the final prediction is determined by the majority vote over the class predictions of all trees (Friedman et al. 2001).

Random Forest balances the bias-variance trade-off by using a few hyperparameters, like the maximum depth of any node of the final tree and the minimal node size to be valid to split. Moreover, reducing the number of selected variables $m$ can decrease variance but may increase bias. The size of each tree is controlled by the minimum size of the terminal nodes. Based on recommendations, for regression, $m$ is typically set to $\frac{p}{3}$ and the minimum node size to $n = 5$ (Breiman 2001). For classification, $m$ is often set to $\sqrt{p}$ and $n = 1$. Random Forests can outperform overfitted single trees and generally offer comparable performance to Boosting, though they are simpler to train and tune (Friedman et al. 2001).

However, Random Forests may not perform as well as Boosting in scenarios

where there are few relevant variables amid many noisy ones. In such cases, the random selection process might overlook important variables, leading to suboptimal splits and decreased performance (Friedman et al. 2001).

## 3.8 Boosting

Boosting is an ensemble method that sequentially combines multiple decision trees to enhance predictive performance by minimizing a specified loss function. In regression settings, it usually relies on the Gaussian (squared error) loss function, whereas for classification tasks, the AdaBoost exponential loss function is commonly used (Friedman et al. 2001). The primary concept behind Boosting is to iteratively use the residuals from the previous tree as input for the next, which incrementally improves performance in areas where previous models were lacking.

To begin, an initial model $t_0$ is created by minimizing the chosen loss function $L$, that can be formally expressed as $\hat{t}(x) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$. For each subsequent tree $b = 1, \ldots, B$, pseudo-residuals are computed as the negative gradient of the loss function, evaluated at the previous tree's predictions. Mathematically, the pseudo-residuals are $r_i = -\frac{\partial}{\partial t(x_i)} L(y_i, t(x_i))$, where $t(x_i) = \hat{t}_{b-1}(x_i)$ (Friedman et al. 2001).

Using these pseudo-residuals as the dependent variable, a regression tree with $K$ splits is built using a randomly selected subset of observations, known as a bagging fraction $p$. Adjusting $K$ allows control over the complexity of each tree, with smaller values focusing learning on specific covariates with challenging split criteria. The current model is then updated by $\hat{t}_b(x) = \hat{t}_{b-1}(x) + \lambda \rho_k(x)$, with $\lambda$ being a shrinkage parameter that controls the learning rate (Friedman et al. 2001).

In practice, the training process involves creating a specified number of trees, using a certain number of splits per tree, a bagging fraction and a shrinkage parameter. For prediction, the best model is selected using hyperparameter tuning with CV to ensure robustness and prevent overfitting.

Moreover, Friedman et al. (2001) highlight that the advantages of Boosting include its high predictive accuracy and flexibility in handling various loss functions. It can effectively manage both regression and classification tasks and is particularly powerful in handling complex data structures. However, the author also mentions that it also has limitations, such as being computationally intensive and requiring careful tuning of parameters to avoid overfitting. Additionally, Boosting models can be difficult to interpret compared to simpler models. On the present work, we use the XGBoost kind of Boosting method in the simulation study and also on the empirical analysis.

# 4 Simulation Study

## 4.1 Simulation Design

The simulation illustrates the performance of the Double Machine Learning IV estimator in a simple instrumental variables model with fixed effects and many controls and few instruments. We follow the data generation processes specified by the work of Belloni et al. (2016) and Chernozhukov et al. (2015). We combine both processes by using the panel data setup of the former and the controls/instruments setup of the latter. For our simulation, we generate data as $n \times T$ from the model

$$\begin{aligned} y_{it} &= e_i + \theta d_{it} + x_{it}\beta + \varepsilon_{it}, \\ d_{it} &= f_i + x_{it}\gamma + z_{it}\delta + u_{it}, \\ z_{it} &= \Pi x_{it} + \zeta_{it}, \end{aligned} \quad \Bigg| \quad \begin{pmatrix} \varepsilon_{it} \\ u_{it} \\ \zeta_{it} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho_\nu & 0 \\ \rho_\nu & 1 & 0 \\ 0 & 0 & 0.25 I_{p_z} \end{pmatrix}\right).$$

where $I_{p_n^z}$ is a $p_n^z \times p_n^z$ identity matrix.

The variable $x_{it}$, which represents the exogenous controls, is a $p_n^x$-dimensional vector for each individual $i$ at time $t$. These controls affect both the endogenous variable $d_{it}$ and the outcome $y_{it}$. The coefficients in $\beta$, describing the relationship between $x_{it}$ and the outcome $y_{it}$, are set to $\beta_j = \frac{1}{j^2}$. The controls $x_{it}$ are generated with serial correlation and individual heterogeneity, reflecting the panel data structure. The individual heterogeneity $e_i$ is generated for $i = 1, ..., n$ as correlated normal random variables with $\mathrm{E}[e_i] = 0$, $\mathrm{Var}(e_i) = \frac{4}{T}$, and $\mathrm{Corr}(e_i, e_j) = .5^{|i-j|}$ for all $i$ and $j$. We set $e_i = f_i$.

The endogenous variable $d_{it}$ is influenced by both the controls $x_{it}$ and the instrumental variable $z_{it}$, as specified by Chernozhukov et al. (2015). Specifically, $d_{it}$ is modeled as a linear function of the fixed effects $f_i$, the controls $x_{it}$, and the instrument $z_{it}$, with disturbances $u_{it}$, which capture the unobserved shocks to the endogenous variable equation. Similarly, $\varepsilon_{it}$ represents the disturbance in the outcome equation $y_{it}$, and $\zeta_{it}$ accounts for the disturbances in the instrument equation $z_{it}$. The vector $\gamma$, linking the controls $x_{it}$ to the endogenous variable, is specified as $\gamma_j = \frac{1}{j^2}$, and the vector $\delta$, linking the instruments to the endogenous variable, with entries $\delta_j = \frac{1}{j^2}$.

The instrument $z_{it}$ plays a crucial role in addressing the endogeneity of $d_{it}$. It provides exogenous variation for $d_{it}$ and is generated as a linear combination of the controls $x_{it}$. Specifically, $z_{it} = \Pi x_{it} + \zeta_{it}$, where $\zeta_{it}$ is a normally distributed error term independent of the controls and disturbances. The matrix $\Pi$ follows the

particular structure:

$$\Pi_j = (-1)^{j-1}\left(\frac{1}{\sqrt{s}}1_{\{j\leq s\}} + \frac{1}{j^2}1_{\{j>s\}}\right), s = \frac{1}{2}n^{1/3}.$$

We generate disturbances according to:

$$\varepsilon_{it} = \rho_\varepsilon \varepsilon_{it-1} + \nu_{1,it}$$
$$u_{it} = \rho_u u_{it-1} + \nu_{2,it}$$

with initial conditions for $\varepsilon_{it}$ and $u_{it}$ being generated from their stationary distribution. This formulation represents an autoregressive process for both disturbances, where $\rho_\varepsilon$ and $\rho_u$ denote the autoregressive parameters.

The exogenous variables conditional on the fixed effects are obtained from:

$$x_{i1j} = \frac{e_i}{1-\rho_x} + \sqrt{\frac{1}{1-\rho_x^2}}\varphi_{i1j}$$
$$x_{itj} = e_i + \rho_x x_{i(t-1)j} + \varphi_{itj} \quad t > 1$$

where $\varphi_{itj}$ are normal random variables with $\mathrm{E}[\varphi_{itj}] = 0$, $\mathrm{Var}(\varphi_{itj}) = 1$, and $\mathrm{Corr}(\varphi_{itj}, \varphi_{itk}) = .5^{|j-k|}$, independent across $i$ and $t$. In all simulations, we set $\rho_\epsilon = \rho_u = \rho_x = .8$, and we set $\rho_\nu = .5$.

The disturbances $\varepsilon$ and $u$, the fixed effects, the controls, and instruments are generated at each simulation replication. The number of potential exogenous controls ($p_n^x$) is set to $n \times (T-2)$, the number of instruments ($p_n^z$) to 10 and $\theta = 0.5$. Moreover, we consider different sample sizes set to $n = 15, 50, 100, 200$, all with $T = 10$. The study was reported based on 100 simulation replications.

## 4.2 Simulation Results

In this chapter, we analyze the performance of different machine learning algorithms employing the DML framework and traditional econometrics methods applied to high-dimensional panel data. The performance is assessed using the bias, RMSE, and clustered standard errors of the estimated coefficients. Our analysis compares four DML models—LASSO, Random Forests, Boosting (XGBoost), and First Differences 2SLS—alongside the Pooled 2SLS estimator, across various panel sizes (15, 50, 100, and 200 units) and $T = 10$. The results can be seen at table II and figure 3.

According to Friedman et al. (2001), bias measures the average deviation of

the estimated coefficient from the true value, indicating systematic error but not accounting for the variability of the estimates. It shows whether the estimator consistently overestimates or underestimates the true value, but does not provide information about the spread or dispersion of these estimates.

In contrast, RMSE accounts for both bias and the variance of the estimates, offering a more comprehensive assessment of the estimator's performance. By considering the square root of the average squared differences between the estimated and true values, RMSE captures the overall accuracy, reflecting both systematic errors and random variability. By squaring the differences, RMSE penalizes larger errors more than smaller ones, making it sensitive to outliers (Friedman et al. 2001). The author also mention that RMSE is always non-negative, with larger values indicating greater total error and poorer performance, emphasizing both consistent deviations and the dispersion of estimates.

In summary, while bias reveals the direction and magnitude of systematic errors in the estimator, RMSE provides a fuller picture by incorporating both the systematic bias and the estimator's variability.

The definition of Bias and RMSE are the following:

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right) \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2} \tag{12}$$

By analyzing the results that are presented at the table II, the Pooled 2SLS estimator consistently performs the worst among the models tested. For smaller panel sizes, the estimator shows a high bias and RMSE, which remain substantial even as the panel size increases. Specifically, with a bias starting at 0.3598 and RMSE of 0.3718 for 15 units, these values peak at 0.4083 and 0.4118 respectively for 50 units, before slightly decreasing but remaining high at 0.3971 and 0.3988 for 200 units. However, the clustered standard errors were smaller comparing to the DML applications across all panel sizes. These results indicate that Pooled 2SLS struggles to provide accurate and reliable estimates in high-dimensional panel data settings, regardless of the sample size.

The First Differences 2SLS model, while still underperforming compared to DML methods, shows a slightly better performance than Pooled 2SLS. For smaller panels, the bias and RMSE are lower than those of Pooled 2SLS, with values of 0.3012 and 0.3142 respectively for 15 units. However, the bias and RMSE increase as the

panel size grows, peaking at 0.3444 and 0.3465 for 100 units, and slightly decreasing to 0.3383 and 0.3390 for 200 units. Although First Differences 2SLS outperforms Pooled 2SLS, its performance does not improve significantly with larger panel sizes, suggesting it is less effective for high-dimensional data.

Table II: Panel IV, $p = n \times (T - 2)$

| | $n = 15$ | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|---|
| | A. Bias | | | |
| Pooled 2SLS | 0.3598 | 0.4083 | 0.3924 | 0.3971 |
| First Differences 2SLS | 0.3012 | 0.3341 | 0.3444 | 0.3383 |
| DML: Random Forests | 0.1475 | 0.0656 | 0.0383 | 0.0361 |
| DML: LASSO | 0.1410 | 0.0501 | 0.0276 | 0.0183 |
| DML: Boosting | 0.1242 | 0.0624 | 0.0316 | 0.0293 |
| | B. RMSE | | | |
| Pooled 2SLS | 0.3718 | 0.4118 | 0.3949 | 0.3988 |
| First Differences 2SLS | 0.3142 | 0.3376 | 0.3465 | 0.3390 |
| DML: Random Forests | 0.1810 | 0.0872 | 0.0692 | 0.0531 |
| DML: LASSO | 0.2243 | 0.1228 | 0.0678 | 0.0517 |
| DML: Boosting | 0.2142 | 0.1026 | 0.0627 | 0.0482 |
| | C. Cluster s.e. | | | |
| Pooled 2SLS | 0.0645 | 0.0361 | 0.0264 | 0.0185 |
| First Differences 2SLS | 0.0609 | 0.0346 | 0.0254 | 0.0176 |
| DML: Random Forests | 0.1091 | 0.0691 | 0.0528 | 0.0376 |
| DML: LASSO | 0.1152 | 0.0724 | 0.0580 | 0.0401 |
| DML: Boosting | 0.1398 | 0.0801 | 0.0580 | 0.0394 |

*Notes:* Author's own elaboration. This table presents simulation results for the IV model with high dimensional controls and fixed effects. Estimators include DML estimators: Random Forests, LASSO, Boosting, First Differences, and Pooled 2SLS. Bias, RMSE, and statistical size for 5% level tests using clustered standard errors are reported based on 100 simulation replications.

Among the DML models, the DML: LASSO model shows a clear trend of improving performance as the panel size increases, reflecting its suitability for high-dimensional data. For smaller panels, the bias and RMSE are relatively high at 0.1410 and 0.2243 respectively for 15 units. As the panel size increases, the bias and RMSE decrease substantially, reaching 0.0183 and 0.0517 for 200 units. The clustered standard errors also decreases with larger panel sizes, indicating more precise estimates. The improved performance can be attributed to the sparsity assumption inherent in LASSO, which allows it to efficiently handle high-dimensional data by shrinking less important coefficients to zero.

The DML: Random Forests model also demonstrates improved performance with larger panel sizes. For smaller panels, the model shows higher bias and RMSE at 0.1475 and 0.1810 respectively for 15 units. As the panel size increases, these values decrease to 0.0361 and 0.0531 for 200 units. The clustered standard errors also follows a decreasing trend, highlighting the model's robustness across different panel sizes. Although Random Forests is slightly less precise than LASSO at larger panel sizes, it still provides reliable estimates.

The DML: Boosting model performs well across different panel sizes, showing improved accuracy with increasing panel sizes. For smaller panels, the bias and RMSE are 0.1242 and 0.2142 respectively for 15 units, which are smaller than those of both LASSO and Random Forests. As the panel size increases to 50 units, Boosting maintains a competitive edge with a bias of 0.0624 and RMSE of 0.1026. With 100 units, the bias and RMSE further reduce to 0.0316 and 0.0627, respectively, and for 200 units, these values are 0.0293 and 0.0482. Boosting performs better than Random Forests, particularly in panel sizes of 50 and 100 units, where it shows lower bias. However, the RMSE is bigger than Random Forest's RMSE, indicating that the Boosting algorithm generate estimates with more variance.

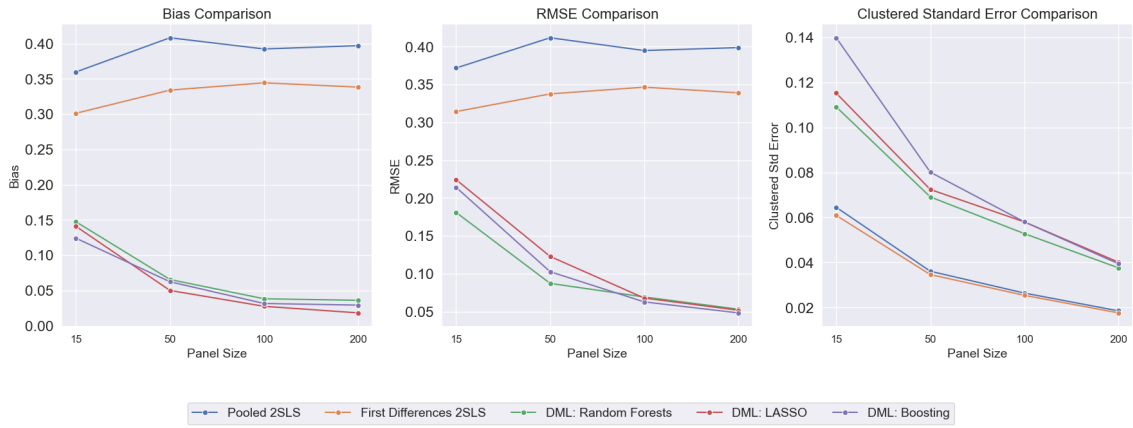


Figure 3: Simulation metrics for the different estimators. Author's own elaboration.

As we can see in figure 3, in terms of smaller panel sizes, the DML models generally outperform traditional methods like Pooled 2SLS and First Differences 2SLS. Among the DML models, Boosting shows the lowest bias for smaller panels, followed by LASSO and Random Forests. Specifically, for 15 units, Boosting's bias is 0.1242, which is better than Random Forests at 0.1475 and LASSO at 0.1410. For RMSE, the Random Forest algorithm provided the best estimates for smaller panels with an RMSE of 0.1810, compared to Boosting's 0.2142 and LASSO's 0.2243. These

results indicate that even with smaller sample sizes, DML methods provide more reliable estimates compared to traditional methods, with Random Forests offering superior RMSE performance in smaller panels. More results can be found in figures 6 and 7.

Comparing across all models, it is evident that DML methods—particularly LASSO, Random Forests, and Boosting—significantly outperform traditional methods like First Differences 2SLS and Pooled 2SLS in terms of bias, RMSE, but not regarding clustered standard errors: the traditional econometric approaches returned smaller clustered standard errors among all panel sizes. Among the DML methods, LASSO generally provides the most precise estimates, especially for larger panel sizes, followed closely by Boosting and Random Forests; that can be due to the linear and sparse structure of the data generation process that has been studied. The substantial improvement in performance with increasing panel sizes highlights the importance of adequate sample sizes when applying DML techniques to high-dimensional panel data.

Overall, the results suggest that when dealing with high-dimensional panel data, DML methods were reaching results closer to the true $\theta$ comparing to traditional econometric approaches. The Random Forest applied with DML framework, in particular, is highly effective in reducing estimation errors and providing more reliable coefficient estimates even in smaller sample sizes, and also improving as the sample size increases. This highlights the advantage of leveraging machine learning techniques in econometric analysis of complex datasets.

# 5 Empirical Analysis

## 5.1 Data

The data for this empirical analysis is sourced from the publicly available dataset provided by Joshua Angrist at the MIT Archive. The original study by Angrist & Kugler (2003) examines the effects of immigration on labor market outcomes in European countries, considering the interaction between immigration and labor market institutions.

The dataset comprises a panel of 15 European countries observed over the period from 1983 to 1999. The authors examined various demographic groups, including males and females, both above and below 40 years old, across different countries. In our study, we specifically examine the impact of immigration on the labor market outcomes for males under 40 years old. This demographic is particularly relevant

as it represents a significant portion of the labor force that might be directly affected by immigration. After processing the data and removing observations with empty information, the sample size consists in a total of 167 observations, with 101 variables. The panel structure allows for the analysis of both cross-country and over-time variations in the data.

The main groups of explanatory variables in the dataset are diverse. Country demographics include variables such as population, immigrant population, and the population of working age, which are essential for understanding the labor market dynamics and the potential impact of immigration.

To account for temporal changes that might affect labor market outcomes, the dataset includes year-specific trends for each country. Key macroeconomic indicators included in the dataset are GDP, interest rates, and whether the country is a member of the Schengen Area, which help control for the broader economic environment in which labor markets operate.

Institutional variables are also critical for analyzing how labor market regulations and policies influence the impact of immigration. These include measures of labor standards, barriers to entrepreneurship, and replacement rates. The study uses instruments such as the distance from former Yugoslavian cities (Sarajevo and Pristina) to European capitals to address potential endogeneity in immigration flows. These distances serve as exogenous sources of variation in immigration patterns.

The dataset used in this study provides a comprehensive view of the labor market conditions and institutional settings across European countries over a significant period. By including a wide range of variables, the analysis can investigate the complex interactions between immigration, labor market outcomes, and institutional factors.

## 5.2 Two Stage Least Squares Regression Model

In their paper, Angrist & Kugler (2003) propose two distinct regression models to investigate the impact of non-EU immigrants on European countries. Their analysis aims to determine not only the isolated effects of immigration on various economic outcomes but also how institutional factors might exacerbate these effects. The authors' approach provides insights into the interaction between immigration and labor market rigidity.

The first model examines the direct impact of immigrants without considering the interaction with institutional variables. The second model incorporates these interactions to assess whether institutional factors, such as labor market rigidity, amplify the effects of immigration.

The first regression model, which does not include interactions with institutional variables, is specified as follows:

$$\ln(y_{it}) = \mu_i + \delta_t + \theta \ln(s_{it}) + \varepsilon_{it} \tag{13}$$

In this model:

- $y_{it}$ represents the employment rate for country $i$ at time $t$.
- $\mu_i$ captures country-specific fixed effects, accounting for time-invariant characteristics of each country that could influence the employment outcome.
- $\delta_t$ represents time fixed effects, which control for global shocks or time trends that affect all countries equally.
- $\theta$ denotes the elasticity of a percentage change in native employment levels in response to a one percent change in the immigrant share.
- $s_{it}$ is the size of the immigrant share in the working population on the employment rate, where $s_{it}$ is the number of non-EU immigrants in country $i$ at time $t$. The coefficient $\theta$ indicates how changes in the immigrant population as a fraction of the working population influence the employment rate in country $i$.
- $\varepsilon_{it}$ is the error term, capturing unobserved factors affecting the outcome.

The second model extends the first by including interaction terms between the share of immigrants and institutional variables, allowing for the assessment of how these institutions influence the impact of immigration:

$$\ln(y_{it}) = \mu_i + \delta_t + (\theta_0 + \theta_1 \tilde{x}_i) \ln(s_{it}) + v_{it} \tag{14}$$

where:

- $y_{it}$ retains the same meaning as in the first model.
- $\mu_i$ and $\delta_t$ are similarly defined, controlling for country-specific and time effects, respectively.
- The terms $\tilde{x}_i$ and $s_{it}$ captures the interaction between the size of the immigrant population and institutional factors. Here, $\tilde{x}_i$ represents a vector of institutional variables(e.g., a measure of labor market rigidity) specific to country $i$.

  - $\theta_0$ is the baseline effect of immigration on the economic outcome, as mentioned earlier.
  - $\theta_1$ represents how the effect of immigration varies with the institutional variable $\tilde{x}_i$. This coefficient assesses whether and how the interaction between immigration and institutional factors amplifies or moderates the impact of immigration on the economic outcome.

- $v_{it}$ is the error term for this model, capturing any unobserved factors affecting the economic outcome not included in the regression.

According to Angrist & Kugler (2003), OLS estimates may be biased upwards by the endogeneity of the term $s_{it}$, where immigrants choosing to locate where their employment prospects are best, being a source of reverse causality. Furthermore, omitted factors such as local economic conditions, policies, or historical migration patterns may simultaneously influence both the immigrant share and the labor market outcomes, leading to biased estimates in OLS regressions. The authors also mention that the most important ommited variables are time-varying productivity or labour demand shocks correlated with both immigrant shares and native employment.

The choice of instruments was motivated by the sharp large increase in the number of Yugoslavs among European immigrants in the early and late 1990s. The authors suggest that distance from the Yugoslav conflict should be a good predictor of the foreign share in the 1990s. The first step of the IV strategy proposed by Angrist & Kugler (2003) is:

$$\ln(s_{it}) = \pi_i + \gamma_t + b_{it}\beta_1 + n_{it}\beta_2 + k_{it}\beta_3 + \eta_{it} \tag{15}$$

where:

- $\pi_i$ and $\gamma_t$ are similarly defined as before, controlling for country-specific and time effects, respectively.
- $b_{it}$ represents the distance from Sarajevo multiplied by a dummy variable for the years 1991-1995 (Bosnia War years).
- $n_{it}$ represents the distance from Sarajevo multiplied by a dummy variable for the years 1996-1997 (inter-war years).
- $k_{it}$ represents the distance from Pristina multiplied by a dummy variable for the years 1998-1999 (Kosovo War years).

The authors also mention that the essence of the IV strategy is to look for a break in the time-series behavior of employment rates for countries relatively close to Yugoslavia. The strategy is also inspired by the study of Card (1990) for the impact of the Mariel Boatlift.

## 5.3 Double Machine Learning Regression Model

Following the equations specified by Angrist & Kugler (2003), we allow the construction of nuisance parameters to incorporte non-linearities and to deal with the high-dimensionality of the model specifications (including more explanatory variables in comparison to the original study). Since both $y_{it}$ and $s_{it}$ are expressed in logarithmic form, the estimated coefficients represent elasticities, meaning they capture the percentage change in $y_{it}$ in response to a one-percent change in $s_{it}$.

The equations can be shown below:

$$
\begin{aligned}
ln(y_{it}) &= \theta ln(s_{it}) + g_0(x_{it}) + \varepsilon_{it}, \\
ln(s_{it}) &= m_0(x_{it}) + v_{it}, \\
z_{it} &= \ell_0(x_{it}) + u_{it}
\end{aligned}
$$

The confounding exogenous factors $x_{it}$ affect the policy variable via the function $m_0(x_{it})$, the dependent variable via the function $g_0(x_{it})$ and the instrumental variable via the function $\ell_0(x_{it})$, as specified by Chernozhukov et al. (2018). Year dummies are also part of $x_{it}$, to take into account possible time fixed effects. Since the exact form of these functions is unknown, high-dimensional and could be highly nonlinear, we use machine learning techniques to learn the nuisance functions $g_0$, $m_0$ and $\ell_0$, since they are well-suited to model their complexity.

In order to remove country fixed effects, we estimate the model with the first difference method. Therefore, we have the following model:

$$
\begin{aligned}
\Delta ln(y_{it}) &= \theta \Delta ln(s_{it}) + g_0(\Delta x_{it}) + \Delta\varepsilon_{it}, \\
\Delta ln(s_{it}) &= m_0(\Delta x_{it}) + \Delta v_{it}, \\
\Delta z_{it} &= \ell_0(\Delta x_{it}) + \Delta u_{it}
\end{aligned}
$$

Moreover, the final score function (also known as moment equation) can be

derived from the residualized version of the regular IV moment condition:

$$\mathbb{E}[(\Delta \ln(\tilde{y_{it}}) - \theta \Delta \ln(\tilde{s_{it}}))\Delta \tilde{z_{it}}] = 0$$
$$\mathbb{E}[(\Delta \ln(y_{it}) - g_0(\Delta x_{it}) - \theta(\Delta \ln(s_{it}) - m_0(\Delta x_{it})))(\Delta z_{it} - \ell_0(x_{it}))] = 0$$

Where the Neyman orthogonal score is defined below and $\eta = (m_0, g_0, \ell_0)$:

$$\psi(W; \theta, \eta) = (\Delta \ln(y_{it}) - g_0(\Delta x_{it}) - \theta(\Delta \ln(s_{it}) - m_0(\Delta x_{it})))(\Delta z_{it} - \ell_0(\Delta x_{it})) \tag{16}$$

Therefore, we can retrieve the final estimate:

$$\hat{\theta}_k = \left( \frac{1}{N_k} \sum_{i \in W_k} \Delta \ln(\tilde{s_{it}}) \Delta \tilde{z_{it}} \right)^{-1} \frac{1}{N_k} \sum_{i \in W_k} \Delta \ln(\tilde{y_{it}}) \Delta \tilde{z_{it}} \tag{17}$$

Where:

$$\begin{aligned} \Delta \ln(\tilde{y_{it}}) &= \Delta \ln(y_{it}) - g_0(\Delta \hat{x}_{it}) \\ \Delta \ln(\tilde{s_{it}}) &= \Delta \ln(s_{it}) - m_0(\Delta \hat{x}_{it}) \\ \Delta \tilde{z_{it}} &= \Delta z_{it} - \ell_0(\Delta \hat{x}_{it}) \end{aligned} \tag{18}$$

Moreover, $\hat{\theta}_k$ is the average over the k folds and $W_k$ is the k-th estimation sample.

## 5.4 Labor market effects of immigration in EU countries

The empirical investigation explores the impact of non-EU immigration on the employment of EU natives using a Double Machine Learning (DML) approach. This method is applied in the context of high-dimensional panel data models, which accommodates a large number of regressors relative to the sample size and incorporates unobserved individual-specific heterogeneity. The DML techniques employed include LASSO, Random Forests, and Boosting, with results compared against traditional econometric estimation methods from the original study. The analysis is divided into two panels: without interactions with institutions (Panel A) and with interactions with institutions (Panel B). The estimates can be seen at table III. For the empirical analysis of the immigration, we follow the structural equation form of equations 16, 17 and 18. As mentioned earlier, since our dependent and endogenous variable are in logarithmic form, the estimated coefficients represent elasticities. Thus, they indicate the percentage change in native employment levels in response to a one percent change in the immigrant share.

Table III: Effect estimates of the non-EU immigrants on employment of EU natives

| | (1) DML: LASSO | (2) DML: Random Forests | (3) DML: Boosting | (4) Original estimates |
|---|---|---|---|---|
| *Panel A: Without interactions with institutions* | | | | |
| Main effect | 0.0173 | 0.0395** | 0.0423*** | -0.042 |
| | (0.0096) | (0.0178) | (0.0180) | (0.031) |
| Observations | 167 | 167 | 167 | 167 |
| Raw covariates | 98 | 98 | 98 | 68 |
| *Panel B: With interactions with institutions* | | | | |
| Main effect | 0.0028 | 0.0621*** | 0.0697*** | -0.102** |
| | (0.0027) | (0.0142) | (0.0203) | (0.044) |
| Barriers to entrepreneurship | 0.0013 | 0.0019 | -0.0048 | -0.124*** |
| | (0.0035) | (0.0031) | (0.0038) | (0.043) |
| Labor Standards | 0.0018 | 0.0062** | 0.0072** | 0.029 |
| | (0.0032) | (0.0027) | (0.0032) | (0.026) |
| Replacement Rate | -0.0006 | -0.0030 | -0.0043 | 0.006 |
| | (0.0023) | (0.0023) | (0.0030) | (0.018) |
| Observations | 167 | 167 | 167 | 167 |
| Raw covariates | 101 | 101 | 101 | 71 |

*Notes:* Author's own elaboration. Column (4) reports the original paper estimates. Standard errors are reported in parentheses. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Standard errors adjusted for clustering at the country level are reported in column 4.

In Panel A (that can also be seen in figure 4), the main effect of non-EU immigration on EU native employment is positive across all DML models, although not statistically significant in DML: LASSO. Specifically, the DML: LASSO model estimates an elasticity of 0.0170% with a standard error of 0.0096, indicating a positive but imprecise effect. Similarly, the DML: Boosting model provides a coefficient of 0.0423% with a standard error of 0.0180, which is significant at 1%. The DML: Random Forests model also stands out with a statistically significant positive effect estimate of 0.0395% (standard error of 0.0178).

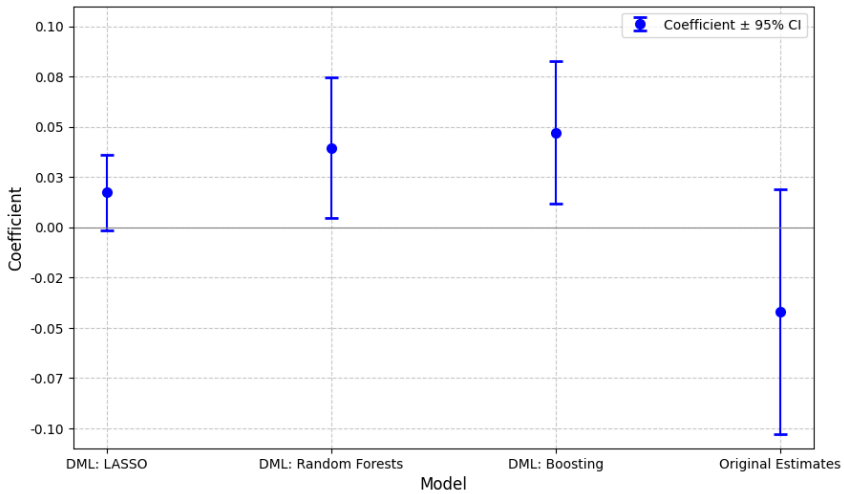


Figure 4: Panel A: Without interaction with institutions. Author's own elaboration.

This suggests that, in the absence of institutional interactions, non-EU immigration may have a modest positive impact on the employment of EU natives. On the other hand, the original estimates report a negative main effect of -0.042% (standard error of 0.031), which is not statistically significant.

Panel B (that can also be seen in figure 5) extends the analysis by incorporating interactions with institutional variables such as barriers to entrepreneurship, labor standards, and replacement rates. The main effect of non-EU immigration on EU native employment becomes more pronounced on DML: Boosting and DML: Random Forests. The DML: LASSO model estimates a coefficient of 0.0028% (standard error of 0.0027), which is not statistically significant. The DML: Random Forests model shows an even larger effect of 0.0621% with a standard error of 0.0142, significant at the 1% level. The DML: Boosting model again provides another positive estimate, with a coefficient of 0.0697% and a standard error of 0.0203, significant at the 1% level. These results suggest that when accounting for institutional interactions, non-EU immigration has a significantly but small positive impact on EU native employment. The original estimates in this context indicate a substantial negative effect of -0.102% (standard error of 0.044), significant at the 5% level. This is also contrasted with our DML estimates, also on the sign of the coefficient.

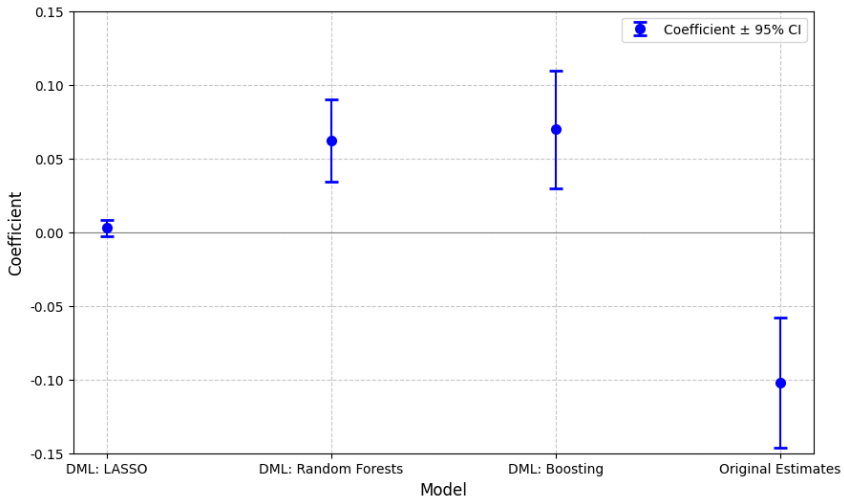


Figure 5: Panel B: With interaction with institutions. Author's own elaboration.

The interaction terms with institutional variables provide additional insights. The effect of barriers to entrepreneurship is statistically insignificant across all DML models, with coefficients close to zero. This contrasts with the original estimate, which suggests a significant negative impact. This indicates that the regulatory environment for starting and operating businesses does not significantly influence the employment impact of non-EU immigrants on EU natives when assessed using DML techniques.

The labor standards variable shows a positive and significant effect at 5% in the DML: Random Forests and DML: Boosting. On the other hand, the effects are positive but insignificant in DML: LASSO model and also at the original estimates. This suggests that stricter labor standards might positively interact with immigration to enhance native employment.

The replacement rate interaction is statistically insignificant at every DML model and also at the original estimates, indicating that higher replacement rates may not influence employment effects of immigration.

The RMSE results (that are shown on figure 8 at the appendices) highlight interesting differences in model performance for predicting both $g_0(x)$ and $m_0(x)$. For $g_0(x)$, LASSO demonstrated the best performance, achieving the lowest RMSE out-ot-sample, followed by Boosting, and then Random Forest, which had the highest error. This suggests that $g_0(x)$ might be closer to a linear or sparsely linear relationship, which LASSO, being a linear model with regularization, is well-suited to capture. On the other hand, for $m_0(x)$, Random Forest yielded the lowest RMSE, indicating that it better captured the underlying structure in the data, with Boosting coming in second and LASSO producing the largest errors. This suggests that the

relationship between the explanatory variables and immigrant allocation, $m_0(x)$, may be more complex and non-linear in nature, which is why more flexible models like Random Forest and Boosting perform better. These results indicate that while $g_0(x)$ might be closer to a linear specification, the process behind $m_0(x)$ could involve more complicated interactions, better suited for highly non-linear models.

Overall, the application of DML estimators to the analysis of immigration effects on labor markets provides distinct but complementary insights compared to traditional methods. The DML: Boosting model consistently shows the most significant positive effects, particularly when accounting for institutional interactions. These findings highlight the potential of machine learning techniques in economic research, especially in handling high-dimensional non-linear data and capturing complex relationships that traditional methods might miss. The contrasting results with the original estimates may show the importance of testing the new methodological advancements in accurately assessing policy impacts.

# 6 Conclusion

This thesis examines the causal effects of immigration on labor market outcomes, specifically focusing on employment rates using advanced machine learning techniques within the DML framework. The research involved a Monte Carlo simulation study and an empirical analysis, revisiting the work of Angrist & Kugler (2003) and extending it to incorporate high-dimensional data.

The simulation results indicate that DML estimators, particularly those leveraging Random Forests, Boosting, and LASSO, provide robust and reliable estimates even in the presence of high-dimensional controls and endogeneity. These methods outperform traditional econometric techniques in terms of handling complex data structures. The simulation follows the data generation processes specified by Belloni et al. (2016) and Chernozhukov et al. (2015), combining these approaches to test the performance of DML estimators for an IV panel data setup with few instruments, fixed effects and many controls.

In the empirical analysis, the study revisits the labor market effects of immigration on native employment rates across 15 European countries, as originally investigated by Angrist & Kugler (2003). The findings from our empirical analysis using DML techniques provide new insights into the complex relationship between immigration and native employment in European countries. Our results generally align with the more recent literature that suggests positive effects of immigration on native employment, while also highlighting the importance of considering institutional factors.

Our analysis shows a modest positive impact of non-EU immigration on EU native employment, particularly when accounting for institutional interactions. This aligns with the findings of Moreno-Galbis & Tritah (2016), who found that natives' employment rates increase in occupations and sectors receiving more immigrants. Our results, showing an increase in native employment rates by 0.4% up to 0.7% for a 10 percent increase in the foreign share, support their conclusion that immigrants can improve employment prospects for natives. The positive effect we observe is consistent with the occupational upgrading mechanism described by D'Amuri & Peri (2014). While we didn't directly measure job complexity, the positive employment effect could be indicative of natives moving towards more complex jobs as immigrants fill manual-routine occupations. This aligns with their finding of a 0.7% increase in native wages for a doubling of the immigrant share.

Our results also resonate with Ortega & Peri (2009) findings of no evidence of crowding-out of native workers. In fact, our analysis suggests that immigration could

even increase employment in receiving countries, supporting their conclusions.

However, our findings regarding institutional interactions present a more nuanced picture than some previous studies. While Angrist & Kugler (2003) found that labor market rigidities and high business entry costs exacerbate negative impacts of immigration, our results show less clear evidence for these effects. The insignificant effect of barriers to entrepreneurship across all DML models suggests that the regulatory environment for businesses may not significantly influence the employment impact of immigration, contrary to some previous findings. The mixed results we found for the effects of labor standards on the immigration-employment relationship highlight the complexity of these interactions. This complexity was also noted by D'Amuri & Peri (2014), who found that the positive reallocation process was stronger in more flexible labor markets.

In conclusion, our empirical analysis, using DML techniques, generally supports the more optimistic view of immigration's impact on native employment found in recent literature. However, it also underscores the need for further research to fully understand the complex interactions between immigration, employment, and labor market institutions. These findings contribute to the ongoing debate on immigration policies and their economic impacts in European countries.

Future research could build on this work by exploring several avenues. First, expanding the dataset to include more recent data and additional countries would provide a broader context and potentially more generalizable results. Second, integrating other advanced machine learning techniques, such as neural networks, could further enhance the predictive power and robustness of the models. Third, examining other labor market outcomes, such as wages, would provide a more comprehensive understanding of the impact of immigration. Lastly, investigating the role of different types of immigration (e.g., skilled vs. unskilled) and their specific effects on various labor market segments could offer more targeted policy recommendations. Lastly, incorporating dynamic panel models within the DML framework could address potential issues related to time dynamics and unobserved heterogeneity, providing an even deeper understanding of the labor market impacts of immigration.

# References

Angrist, J. D. & Frandsen, B. (2022), 'Machine labor', *Journal of Labor Economics* **40**(S1).

Angrist, J. & Kugler, A. (2003), 'Protective or counter-productive? labour market institutions and the effect of immigration on eu natives', *The Economic Journal* **113**(488), F302–F331.
**URL:** *https://doi.org/10.1111/1468-0297.00136*

Athey, S. & Imbens, G. W. (2019), 'Machine learning methods that economists should know about', *Annual Review of Economics* **11**(Volume 11, 2019), 685–725.
**URL:** *https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080217-053433*

Athey, S., Imbens, G. W. & Wager, S. (2018), 'Approximate residual balancing: Debiased inference of average treatment effects in high dimensions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(4), 597–623.

Bach, P., Chernozhukov, V., Kurz, M. S. & Spindler, M. (2022), 'DoubleML – An object-oriented implementation of double machine learning in Python', *Journal of Machine Learning Research* **23**(53), 1–6.
**URL:** *http://jmlr.org/papers/v23/21-0862.html*

Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S. & Spindler, M. (2024), 'Hyperparameter tuning for causal inference with double machine learning: A simulation study'.
**URL:** *https://arxiv.org/abs/2402.04674*

Baiardi, A. & Naghi, A. A. (2021), 'The value added of machine learning to causal inference: Evidence from revisited studies'.
**URL:** *https://arxiv.org/abs/2101.00878*

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

Belloni, A., Chernozhukov, V. & Hansen, C. (2011), 'Lasso methods for gaussian instrumental variables models'.
**URL:** *https://arxiv.org/abs/1012.1297*

Belloni, A., Chernozhukov, V. & Hansen, C. (2014*a*), 'High-dimensional methods and inference on structural and treatment effects', *Journal of Economic Perspectives* **28**(2), 29–50.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014*b*), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**(2), 608–650.
**URL:** *https://doi.org/10.1093/restud/rdt044*

Belloni, A., Chernozhukov, V., Hansen, C. & Kozbur, D. (2016), 'Inference in high-dimensional panel models with an application to gun control', *Journal of Business & Economic Statistics* **34**(4), 590–605.
**URL:** *https://doi.org/10.1080/07350015.2015.1102733*

Bergstra, J. & Bengio, Y. (2012), 'Random search for hyper-parameter optimization', *The Journal of Machine Learning Research* **13**(null), 281–305.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Card, D. (1990), 'The impact of the mariel boatlift on the miami labor market', *ILR Review* **43**(2), 245–257.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M. & Syrgkanis, V. (2024), 'Applied causal inference powered by ml and ai'.
**URL:** *https://arxiv.org/abs/2403.02467*

Chernozhukov, V., Hansen, C. & Spindler, M. (2015), 'Post-selection and post-regularization inference in linear models with many controls and instruments', *American Economic Review* **105**(5), 486–490.

Clarke, P. & Polselli, A. (2024), 'Double machine learning for static panel models with fixed effects'.
**URL:** *https://arxiv.org/abs/2312.08174*

D'Amuri, F. & Peri, G. (2014), 'Immigration, jobs, and employment protection: Evidence from europe before and during the great recession', *Journal of the European Economic Association* **12**(2), 432–464.
**URL:** *https://doi.org/10.1111/jeea.12040*

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1 of *Springer series in statistics*, Springer, New York.

Frisch, R. & Waugh, F. V. (1933), 'Partial time regressions as compared with individual trends', *Econometrica: Journal of the Econometric Society* pp. 387–401.

Klosin, S. & Vilgalys, M. (2023), 'Estimating continuous treatment effects in panel data using machine learning with a climate application'.
**URL:** *https://arxiv.org/abs/2207.08789*

Kock, A. B. & Tang, H. (2019), 'Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects', *Econometric Theory* **35**(2), 295–359.

Lovell, M. C. (1963), 'Seasonal adjustment of economic time series and multiple regression analysis', *Journal of the American Statistical Association* **58**(304), 993–1010.

Mackey, L., Syrgkanis, V. & Zadik, I. (2018), 'Orthogonal machine learning: Power and limitations'.
**URL:** *https://arxiv.org/abs/1711.00342*

Moreno-Galbis, E. & Tritah, A. (2016), 'The effects of immigration in frictional labor markets: Theory and empirical evidence from eu countries', *European Economic Review* **84**, 76–98.
**URL:** *https://doi.org/10.1016/j.euroecorev.2015.06.006*

Neyman, J. (1979), 'C() tests and their use', *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **41**(1/2), 1–21.

Ortega, F. & Peri, G. (2009), 'The causes and effects of international migrations: Evidence from oecd countries 1980-2005', (14833).
**URL:** *http://www.nber.org/papers/w14833*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Robinson, P. M. (1988), 'Root-n-consistent semiparametric regression', *Econometrica* **56**(4), 931–954.

Semenova, V., Goldman, M., Chernozhukov, V. & Taddy, M. (2022), 'Estimation and inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence'.
**URL:** *https://arxiv.org/abs/1712.09988*

Shao, D., Soleymani, A., Quinzan, F. & Kwiatkowska, M. (2024), 'Learning decision policies with instrumental variables through double machine learning'.
**URL:** *https://arxiv.org/abs/2405.08498*

Sheppard, K., Ro, J., bot, S., Lewis, B., Clauss, C., Guangyi, Jeff, Yu, J. Q., Jiageng, Wilson, K., Migrator, L., Thrasibule, WilliamRoyNelson, RENE-CORAIL, X. & vikjam (2024), 'bashtage/linearmodels: Version 6.1 (v6.1)'.
**URL:** *https://doi.org/10.5281/zenodo.13832604*

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.
**URL:** *http://www.jstor.org/stable/j.ctt5hhcfr*

# A Appendices

## A.1 Descriptive statistics

Table IV: Descriptive Statistics by Country

| | Employment | | Non-EU Immigrant Share | | EU Immigrant Share | | Total Population (000s) | |
|---|---|---|---|---|---|---|---|---|
| Country | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| AT | 0.8645 | 0.0090 | 0.0848 | 0.0024 | 0.0130 | 0.0016 | 4518.6143 | 18.8131 |
| BE | 0.8291 | 0.0136 | 0.0227 | 0.0024 | 0.0556 | 0.0033 | 5424.5679 | 107.6793 |
| CH | 0.9113 | 0.0138 | 0.0558 | 0.0023 | 0.1695 | 0.0063 | 4034.8794 | 10.6092 |
| DE | 0.8415 | 0.0110 | 0.0540 | 0.0058 | 0.0301 | 0.0030 | 40837.4922 | 5794.6924 |
| DK | 0.8540 | 0.0263 | 0.0122 | 0.0027 | 0.0085 | 0.0008 | 2842.1067 | 101.5041 |
| ES | 0.7095 | 0.0384 | 0.0038 | 0.0023 | 0.0025 | 0.0009 | 20183.9902 | 676.1739 |
| FI | 0.7590 | 0.0674 | 0.0087 | 0.0032 | 0.0015 | 0.0001 | 2825.3701 | 48.0095 |
| FR | 0.8253 | 0.0395 | 0.0365 | 0.0017 | 0.0278 | 0.0036 | 29456.0039 | 1101.7648 |
| IE | 0.7741 | 0.0291 | 0.0062 | 0.0018 | 0.0239 | 0.0026 | 1735.8468 | 129.6851 |
| IT | 0.7377 | 0.0221 | 0.0061 | 0.0019 | 0.0014 | 0.0007 | 31731.9102 | 447.8229 |
| NL | 0.8585 | 0.0304 | 0.0224 | 0.0022 | 0.0159 | 0.0014 | 8538.1602 | 432.6641 |
| NO | 0.8482 | 0.0284 | 0.0183 | 0.0011 | 0.0104 | 0.0008 | 2386.1411 | 31.4627 |
| PT | 0.8229 | 0.0360 | 0.0086 | 0.0021 | 0.0026 | 0.0009 | 5375.5083 | 77.1725 |
| SE | 0.7612 | 0.0168 | 0.0315 | 0.0062 | 0.0135 | 0.0050 | 4732.9248 | 16.6726 |
| UK | 0.8406 | 0.0215 | 0.0293 | 0.0079 | 0.0169 | 0.0010 | 30387.0996 | 1011.7522 |

*Notes:* This table presents descriptive statistics for employment, non-EU immigrant share, EU immigrant share, and total population across different countries.

Table V: Descriptive Statistics for Barriers to Entrepreneurship, Labor Standards, and Replacement Rate

| | Barriers to Entrepreneurship | | Labor Standards | | Replacement Rate | |
|---|---|---|---|---|---|---|
| Country | Mean | Std | Mean | Std | Mean | Std |
| AT | 16 | 0 | 5 | 0 | 50 | 0 |
| BE | 17 | 0 | 4 | 0 | 60 | 0 |
| CH | 6 | 0 | 3 | 0 | 70 | 0 |
| DE | 15 | 0 | 6 | 0 | 63 | 0 |
| DK | 5 | 0 | 2 | 0 | 90 | 0 |
| ES | 19 | 0 | 7 | 0 | 70 | 0 |
| FI | 10 | 0 | 5 | 0 | 63 | 0 |
| FR | 14 | 0 | 6 | 0 | 57 | 0 |
| IE | 12 | 0 | 4 | 0 | 37 | 0 |
| IT | 20 | 0 | 7 | 0 | 20 | 0 |
| NL | 9 | 0 | 5 | 0 | 70 | 0 |
| NO | 11 | 0 | 5 | 0 | 65 | 0 |
| PT | 18 | 0 | 4 | 0 | 65 | 0 |
| SE | 13 | 0 | 7 | 0 | 80 | 0 |
| UK | 7 | 0 | 0 | 0 | 38 | 0 |

*Notes:* This table presents descriptive statistics for Barriers to Entrepreneurship, Labor Standards, and Replacement Rate across different countries.

## A.2 Hyperparameters configuration

Table VI: Selected Hyperparameters for Models (Empirical Analysis)

| Model | Configuration | Hyperparameters |
|---|---|---|
| LASSO | $g_0(x)$ | alpha: 0.0005 |
| | $m_0(x)$ | alpha: 0.001 |
| Boosting | $g_0(x)$ | learning_rate: 0.1<br>max_depth: 3<br>n_estimators: 200 |
| | $m_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 200 |
| Random Forest | $g_0(x)$ | ccp_alpha: 0.0<br>max_depth: 20<br>max_features: 50<br>min_samples_leaf: 1<br>n_estimators: 200 |
| | $m_0(x)$ | ccp_alpha: 0.0<br>max_depth: 20<br>max_features: 20<br>min_samples_leaf: 1<br>n_estimators: 200 |

*Notes:* Author's own elaboration. This table presents the selected hyperparameters for LASSO, Boosting, and Random Forest models. Each configuration is specified alongside its respective hyperparameters.

Table VII: Selected Hyperparameters for Simulation Study (N = 15)

| Model | Configuration | Hyperparameters |
|---|---|---|
| LASSO | $g_0(x)$ | alpha: 0.1 |
| | $m_0(x)$ | alpha: 0.1 |
| Boosting | $g_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 50 |
| | $m_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 50 |
| Random Forest | $g_0(x)$ | ccp_alpha: 0.01<br>max_depth: 3<br>max_features: 100<br>min_samples_leaf: 4<br>n_estimators: 50 |
| | $m_0(x)$ | ccp_alpha: 0.01<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 2<br>n_estimators: 200 |

*Notes:* Author's own elaboration. This table presents the updated hyperparameters for LASSO, Boosting, and Random Forest models. Each configuration is specified alongside its respective hyperparameters.

Table VIII: Selected Hyperparameters for Simulation Study (N = 50)

| Model | Configuration | Hyperparameters |
|---|---|---|
| LASSO | $g_0(x)$ | alpha: 0.1 |
| | $m_0(x)$ | alpha: 0.1 |
| Boosting | $g_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 100 |
| | $m_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 100 |
| Random Forest | $g_0(x)$ | ccp_alpha: 0.01<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 1<br>n_estimators: 200 |
| | $m_0(x)$ | ccp_alpha: 0.01<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 1<br>n_estimators: 200 |

*Notes:* Author's own elaboration. This table presents the updated hyperparameters for LASSO, Boosting, and Random Forest models. Each configuration is specified alongside its respective hyperparameters.

Table IX: Selected Hyperparameters for Simulation Study (N = 100)

| Model | Configuration | Hyperparameters |
|---|---|---|
| LASSO | $g_0(x)$ | alpha: 0.1 |
| | $m_0(x)$ | alpha: 0.1 |
| Boosting | $g_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 200 |
| | $m_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 100 |
| Random Forest | $g_0(x)$ | ccp_alpha: 0.01<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 1<br>n_estimators: 100 |
| | $m_0(x)$ | ccp_alpha: 0.0<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 1<br>n_estimators: 100 |

*Notes:* Author's own elaboration. This table presents the updated hyperparameters for LASSO, Boosting, and Random Forest models. Each configuration is specified alongside its respective hyperparameters.

Table X: Selected Hyperparameters for Simulation Study (N = 200)

| Model | Configuration | Hyperparameters |
|---|---|---|
| LASSO | $g_0(x)$ | alpha: 0.1 |
| | $m_0(x)$ | alpha: 0.05 |
| Boosting | $g_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 100 |
| | $m_0(x)$ | learning_rate: 0.05<br>max_depth: 3<br>n_estimators: 100 |
| Random Forest | $g_0(x)$ | ccp_alpha: 0.0<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 1<br>n_estimators: 50 |
| | $m_0(x)$ | ccp_alpha: 0.01<br>max_depth: 5<br>max_features: 100<br>min_samples_leaf: 4<br>n_estimators: 100 |

*Notes:* Author's own elaboration. This table presents the updated hyperparameters for LASSO, Boosting, and Random Forest models. Each configuration is specified alongside its respective hyperparameters.

## A.3 Simulation study model comparison

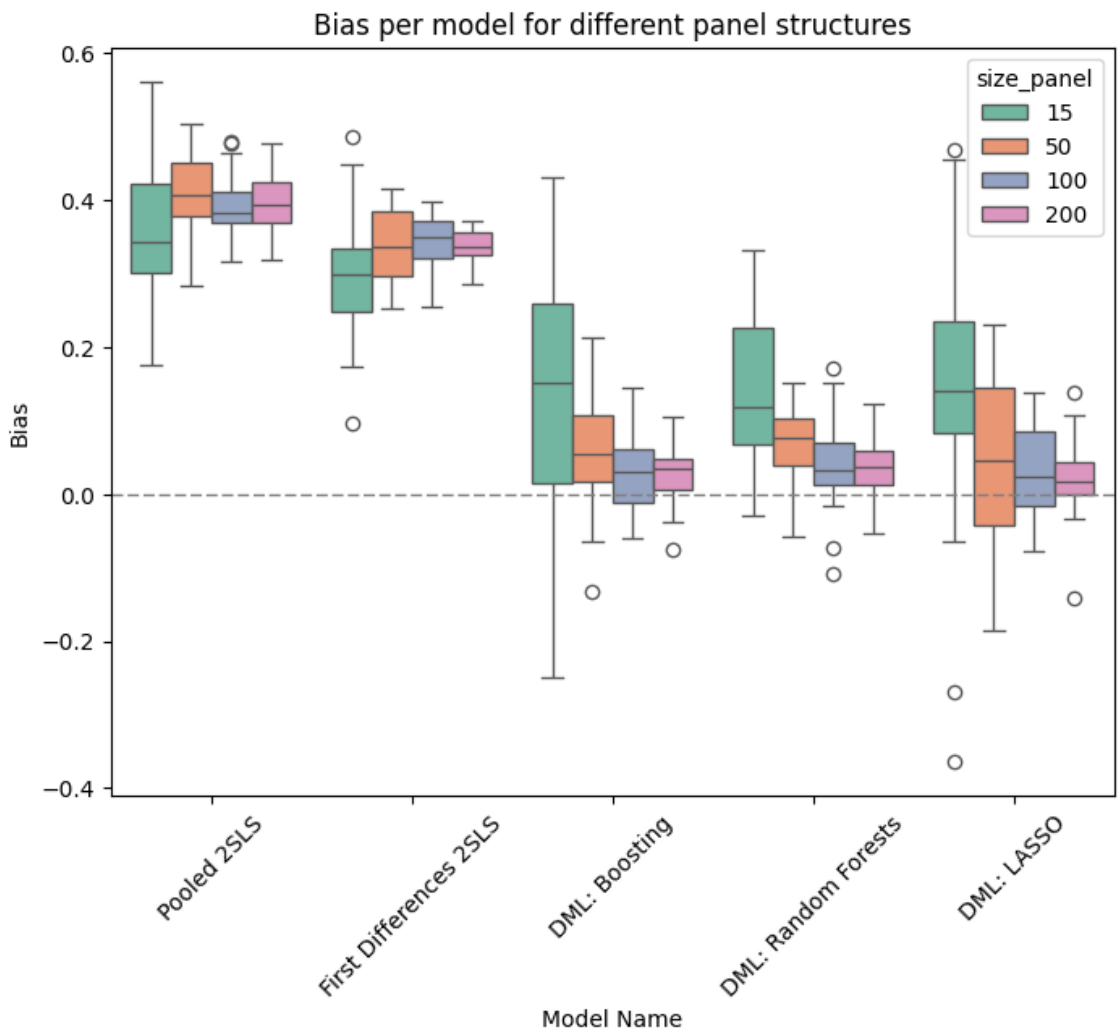


Figure 6: Simulation bias for different panel specifications. Author's own elaboration.

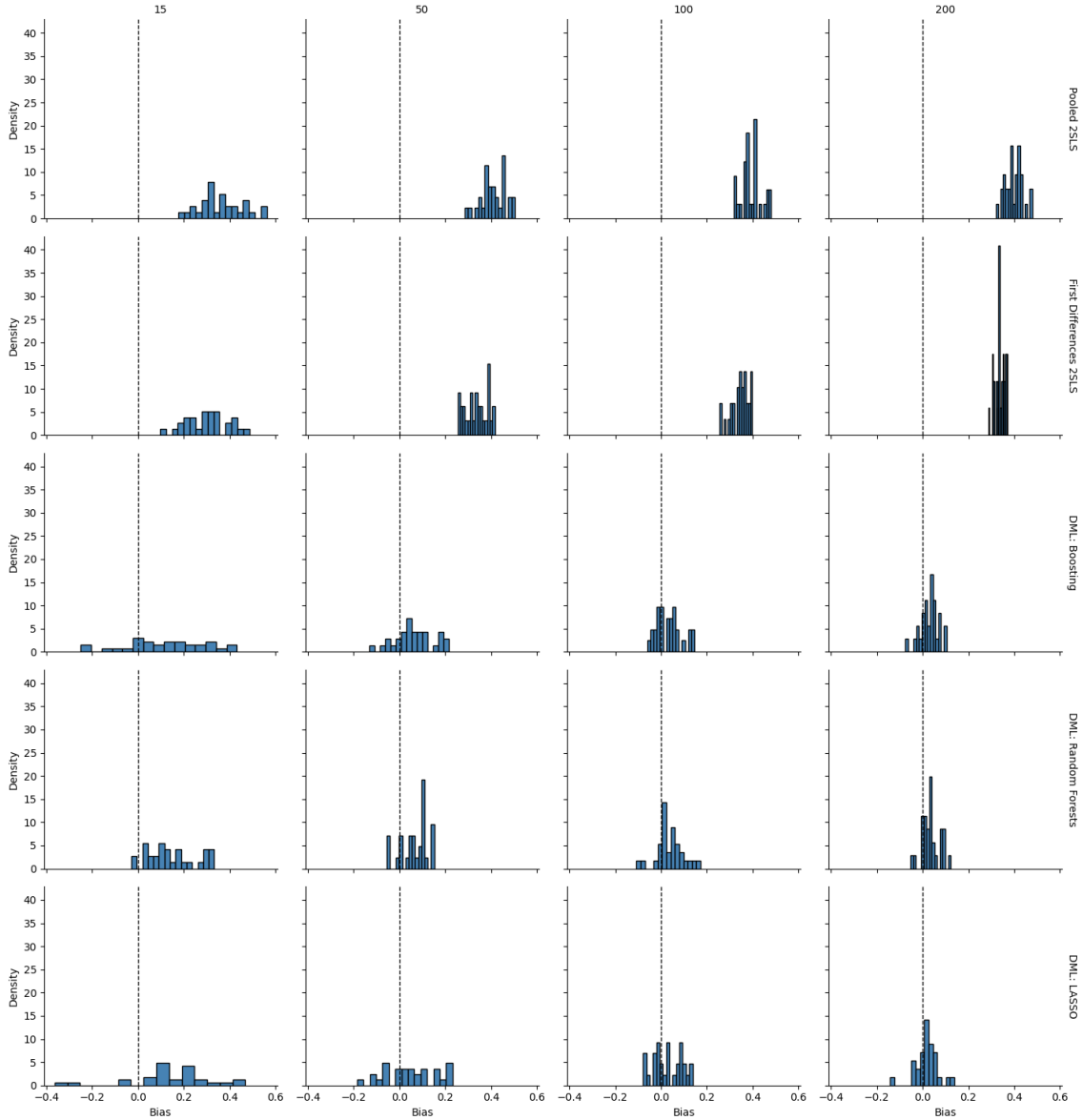


Figure 7: Simulation distributions for different panel specifications (N = 15, 50, 100, 200). Author's own elaboration.

## A.4 Empirical study model comparison

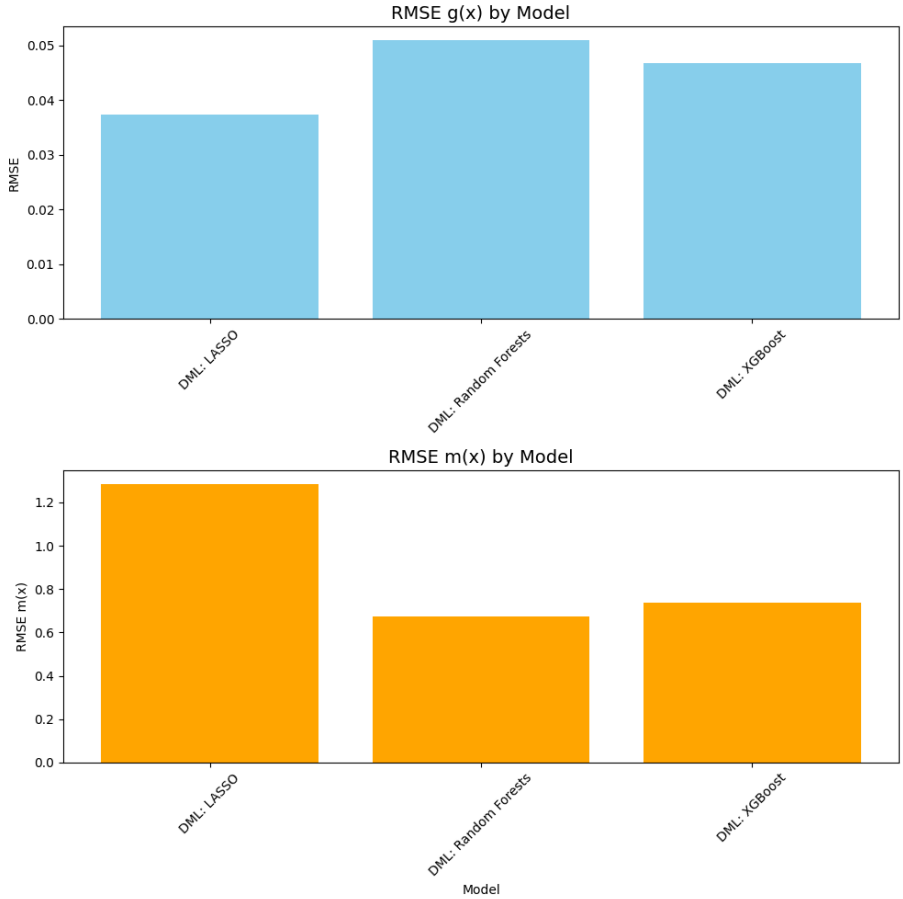


Figure 8: Notes: Author's own elaboration. RMSE results for the three different ML model specifications

## A.5 Implementation

In this study, the analysis was conducted using Python as the primary programming language. The implementation of machine learning models and causal inference techniques was facilitated through various libraries, notably DoubleML (Bach et al. 2022), scikit-learn (Pedregosa et al. 2011) for additional machine learning functionalities and linearmodels (Sheppard et al. 2024) for econometric models. These libraries provided a framework for model building, hyperparameter tuning, and evaluation, enabling a comprehensive analysis of the labor market effects of immigration.