Beyond Traditional Causal Inference: A Causal Forest Approach to Estimating the Heterogeneous Effects of EU Regional Subsidies

Eleni Vossou

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

A dissertation submitted to the Athens University of Economics and Business
in partial fulfilment of the requirements for the attainment of the
MSc degree in Business Economics with Analytics

School of Economic Sciences
Department of Economics

February 2025

Abstract

Can a single policy truly serve the diverse needs of Europe's regions, or does its impact unfold within the hidden nexus of place and time? The asymmetrical distribution of economic growth and development across European regions has been a persistent challenge for the European Union, despite decades of regional subsidy programs aimed at fostering convergence. The average treatment effect (ATE) of these have traditionally been studied through econometric approaches, such as as Difference-in-Differences (DiD) and Two-Way Fixed Effects (TWFE). However, these methods not only suffer from bias in staggered adoption settings or when treatment effects vary over time, but also they fail to capture heterogeneous treatment effects. This thesis aims to address these limitations by evaluating policy implementation through Generalized Synthetic Control (GSC) and Causal Forests. First GSC validates the global effect of EU regional subsidies, providing a flexible counterfactual framework to assess overall program success. Complementing this analysis, Causal Forests are employed to uncover the significant variation in those effects across regions, identifying which regions benefit the most and which remain unaffected. The empirical results reveal substantial heterogeneity in the impact of EU regional subsidies. Causal Forests demonstrate different levels of treatment effects across regions, despite the positive average effects, as some regions experience great benefits while others show minimal responses. These insights shift the focus to understanding for whom and under what conditions the subsidies are most effective, rather than solely evaluating the average impact of the subsidy. In addition, this thesis contributes to the methodological literature by illustrating how machine learning methods can complement traditional econometrics, to address long-standing issues in causal inference in high-dimensional and noisy data environments, while improving precision policy-making, for more equitable and impactful policies across EU regions that respond to their unique realities.

Keywords: Causal Inference, Machine Learning, Generalized Synthetic Control Method, Causal Forest, Heterogeneous Treatment Effects

"All events seem entirely loose and separate. One event follows another; but we never can observe any tie between them. They seem conjoined, but never connected."

— *David Hume*

"It is far more important to know what person the disease has than what disease the person has"

— *Hippocrates*

Acknowledgments

I am deeply grateful to my supervisor, Prof. Angelos Alexopoulos, for his guidance and suggestions throughout this research journey. I would also like to express my gratitude to the members of my thesis committee, Prof. Spyridon Pagratis and Prof. Fabio Antoniou, for their time and effort in reviewing my work.

I sincerely thank Dr. George Melios for his advice on refining my writing throughout this thesis. His constructive feedback was greatly appreciated.

Special thanks to my family and friends for their unwavering support throughout this personal journey.

A heartfelt thank you to my partner, for always believing in me—your love, patience throughout this entire Master's journey and enthusiasm (whether real or expertly faked) for my research, have meant the world to me. I couldn't have done it without you.

Contents

List of Figures

List of Tables

1 Introduction

Few people are eager to embrace determinism, as it challenges the comforting idea of free will. Our natural instinct is to resist it. Determinism argues that everything in the world happens through a chain of cause and effect, governed by universal natural laws. For every cause, there is an effect and the better we understand these causes and their interconnections, the more accurately we can predict what happens next (Hoefer, 2003). Probability, on the other hand, is our tool to deal with the uncertainty that arises when we don't know all the causes (Hacking, 2006). To the world, things either happen or they don't—there's no chance. But what about our choices? Determinism suggests they are also part of this causal chain, an illusion of free will. This concept of determinism has shaped philosophical thought for centuries, influencing Galileo, Descartes, Newton, Spinoza and Laplace (Drake, 1978; Miller, 1983; Smith, 2007; Kisner, 2001; Laplace, 2012). Laplace famously proposed that if we knew the complete state of the universe at any given moment, we could predict everything that follows. But reality is far more complex. We rarely, if ever, have complete information, which is why probability and statistics have become such crucial tools in understanding cause and effect. David Hume questioned whether we could ever truly observe a cause, or whether we were simply observing patterns of regularity (Hume, 1748). John Stuart Mill, meanwhile, focused on methods for identifying causes through counterfactual reasoning—a cornerstone of modern causal analysis (Mill, 1843). As these philosophical ideas evolved, so did the tools for investigating them scientifically. By the early 20th century, Pierre-Simon Laplace, Francis Galton and Karl Pearson were formalizing probability theory and laying the groundwork for correlation and regression analysis (Laplace, 1820; Galton, 1886; Pearson, 1896). These methods helped us understand relationships between variables, but they weren't enough to definitively establish causality. Correlation alone doesn't imply causation—a distinction that has haunted statistics ever since (Pearl, 2018).

Thus, what does this all mean for our contemporary understanding of choice and causality? It is rather true that these foundational ideas have shaped econometrics through statistical models focused on identifying correlations. However, policy evaluation requires methods that go beyond correlation into establishing causal relationships. Understanding whether a policy intervention truly drives economic change, rather than merely coinciding with it, is at the heart of empirical economics. For decades, policymakers have relied on econometric models, such as Difference-in-Differences (DiD) and Two-Way Fixed Effects (TWFE) (Angrist & Pischke, 2009), to estimate those policy effects. These models provide reliable estimates of average treatment effects (ATE), offering insights into overall program success (Imbens & Rubin, 2015). However, these traditional approaches rely on the homogeneity assumption, which implies that policy effects are uniform across treated regions. While useful, this assumption often fails to account for region-specific economic structures, institutional differences and varying degrees of policy exposure—factors that naturally differ between regions and influence the effectiveness of EU structural funds. In the 21st century, an emerging consensus acknowledges that the effects of subsidies/treatment are far from homogeneous and are, in fact, inherently heterogeneous and should not be considered uniform for successful policy design and implementation (Athey & Imbens, 2016; Fratesi & Perucca, 2023; Destefanis & Di Giacinto, 2024). Just as no two regions are identical in their economic structure, institutional capacity or historical trajectory, their responses to external interventions are equally diverse.

This dual need—capturing both the global effects of interventions and the heterogeneous effects experienced by different regions—has led to the development of advanced methods, such as Generalized Synthetic Controls (Xu, 2017) that improves upon traditional methods by providing more flexible counterfactuals to estimate global effects and Causal Forests (Wager & Athey, 2018) that estimate the heterogeneous treatment effects by dynamically partitioning data into subgroups with similar responses. While traditional methods rely on strong assumptions and limited data, machine learning algorithms like Causal Forests can model complex systems, even with high-dimensional and noisy data more flexibly. Together, these methods offer a more comprehensive framework for evaluating both the overall impact of regional subsidies and their varying effects across different economic contexts.

One area where these advanced techniques prove particularly valuable is in assessing the impact of large-scale policy interventions (Athey & Imbens, 2017). The European Union's (EU) mission of economic and social convergence (Dinan, 2005) epitomizes the aspiration of creating a more equitable EU by narrowing economic and social disparities between member states and regions, ensuring income levels, employment opportunities and equal access to quality education, healthcare and social protections (European Commission, 2023). EU has long relied on regional subsidies as a key tool

for fostering economic growth and promoting convergence in its less developed regions, supporting innovation, infrastructure, and job creation (European Commission, 2024). These subsidies, primarily delivered through the Structural and Cohesion Funds, represent a significant component of the EU's strategy to reduce economic disparities among member states. This vision deepens the belief that shared prosperity reinforces the European project, amplifying solidarity, resilience and stability when facing global challenges (Rodríguez-Pose, 2018). However, despite the substantial financial resources dedicated to these programs, the true causal impact of these policies on economic outcomes remains the subject of ongoing debate (Crescenzi & Giua, 2020).

In this dissertation, I aim to address this causal inference challenge for EU subsidies, by critically examining Generalized Synthetic Controls for the overall effect of subsidies, while incorporating Causal Forests to capture the heterogeneous treatment effects across regions. By leveraging and comparing these methods, this research provides an integrated framework that goes beyond average treatment effects to more precise estimation of how these interventions affect growth and convergence, offering deeper insights into their actual benefits. This approach not only advances the methodological literature on causal inference but also contributes to the design of more successful, region-specific policies that promote sustainable growth and convergence.

2 Review of Existing Literature

According to Garg and Fetzer's "Causal Claims in Economics", published in November 2024, over the course of the past few decades, economics has gradually prioritized establishing credible causal relationships, indicating a shift towards more precise empirical methods, the so-called "credibility revolution". Since the adoption of advanced empirical methodologies the proportion of research papers with explicit causal claims rose significantly from 4% in 1990 to 28% in 2020. However the growing methodological complexity raises concerns about the transparency and replicability in modern economic research. The clear trade-off between complex causal narratives and transparency underscores how the incentives in academic publishing may shape research priorities, favoring depth and complexity over accessibility and general relevance. Recognizing this, I aim to outline fundamental and advanced methodologies in causal inference for policy evaluation in economics, where randomization is not feasible, as well as providing innovative solutions for causal inference while maintaining a commitment to transparency. My goal is to provide a comprehensive overview of traditional methodologies while highlighting their key strengths and limitations, that naturally lead to the progression from Difference-in-Differences to Machine Learning for Causal Inference, in order to handle increasingly complex data structures, relax restrictive assumptions (like parallel trends) and better account for unobserved heterogeneity and staggered treatments. In doing so, I hope to demonstrate that it is achievable to obtain methodological sophistication without sacrificing the principles of scientific interpretability.

2.1 Difference-in-Differences

In its simplest form, the Difference-in-Differences (DiD) method is very intuitive and transparent. But as it's used in more complex research settings, it becomes a bit more complicated. DiD became well-known after David Card and Alan Krueger's 1994 study, where they applied it to assess the impact of economic policies. It's a quasi-experimental method that estimates causal effects by comparing changes over time between a treated group and a control group. The key idea is that the control group reflects what would have happened to the treated group if the treatment hadn't occurred.

The idea is to use the control group as a counterfactual—a way to project what the outcomes for the treated group would have been in the absence of the treatment. From this, we can estimate the treatment effect, often called the Average Treatment Effect on the Treated (ATT)

For this method to work, we rely on several important assumptions. The most critical is the parallel trends assumption, which says that without the treatment, the outcomes for both the treated and control groups would have followed the same path over time (Lechner, 2011; Angrist & Pischke, 2009). This assumption is crucial but untestable. We also assume there are no spillover effects between groups, meaning the treatment in one group doesn't affect the outcomes of the control group (Imbens & Rubin, 2015).

DiD is popular in policy evaluation because it simplifies estimating causal effects by controlling for factors that don't change over time and comparing outcome changes between treated and control

groups (Angrist & Pischke, 2009). However, it heavily depends on the parallel trends assumption, which can introduce bias if trends differ for reasons unrelated to the treatment. Other assumptions, like no anticipation of the treatment and no interference between units (SUTVA) (Imbens & Rubin, 2015), are also needed to ensure valid results.

While effective in simple, two-period settings, DiD struggles with staggered treatments or time-varying effects, where uniform treatment effects are harder to assume (Goodman-Bacon, 2021).

To address these issues, extensions to the DiD model include using two-way fixed effects (TWFE), which control for unobserved, time-invariant differences across groups and time periods. However, TWFE models can introduce bias in staggered adoption settings (Goodman-Bacon, 2021). Staggered adoption allows us to handle cases where treatments are implemented at different times, and dynamic treatment effects, through leads and lags, help capture how the effects evolve over time (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021).

2.1.1 Two-Way Fixed Effects Extension

The Two-Way Fixed Effects (TWFE) model builds on the traditional DiD, which makes it more suitable for panel data with multiple units and time periods. By controlling for unobserved, time-invariant factors within units and for common shocks that affect all units simultaneously (like macroeconomic changes), TWFE provides flexibility in complex data settings (Angrist & Pischke, 2009; Wooldridge, 2010). The strongest and often unrealistic (due to the real-world heterogeneity of treatment effects) assumption is the homogeneity of treatment effects across units and over time, where the treatment effect is not only constant but contemporaneous, meaning that the model does not allow for today's treatment to affect future outcome. Like DiD, TWFE relies on parallel trends assumption as the most central assumption and on no anticipation assumption. Strict exogeneity implies that treatment assignment is independent of past, present and future outcomes and supports that the treatment is assigned in one shot, thus implying that the treatment is orthogonal to the two potential outcomes.

A conceptual issue is that strict exogeneity is more demanding than we often acknowledged. In 2019 Imai and Kim showed that strict exogeneity assumption can be "decomposed" into these following assumptions; there is no unobserved time-varying confounder, past outcomes don't directly affect current outcome (no LDV), past treatments don't direclty affect current outcome (no "carryover effect") and the most important assumption; past outcomes don't affect current treatment (no "feedback"). Feedback effects or carryover effects from past treatments, can bias TWFE estimates. In order to relax the no LDV and no "carryover effect" assumption, we control for past treatments and for the no "feedback" assumption, we need instrument variables, according to Arellano and Bond in 1991.

TWFE estimates can be biased due to presence of time-varying confounders, feedback from past outcome and, critically, heterogeneous treatment effects. In the next extension we will see that TWFE can lead to biased estimates in staggered adoption designs. In these designs, different units receive treatment at different times and TWFE, due to the homogeneity in the treatment effect assumption, uses already-treated units as controls for later-treated units. Naturally this leads to biased estimates, because the model cannot properly account for the fact that the treatment effect may vary over time or across units.

2.1.2 Staggered Adoption Extension

In more realistic frameworks, treatment can be adopted at different times across units. For these scenarios we use staggered adoption extension, a weighted average of multiple smaller DiDs. This DiD extension acknowledges that different units, such as states or regions, might adopt treatments at various points in time, rather than simultaneously.

Goodman-Bacon (2021) decomposed the TWFE model into smaller, well define[d] DiD models. The TWFE estimator under staggered adoption in Goodman-Bacon's paper is a weig[hted] average of all possible 2x2 DiD estimators that compare different timing groups to each other. He argues that the weights on the 2x2 DiDs are proportional to timing group sizes and the varianc[e] of the treatment dummy in each pair of groups, which is the highest for units treated in the midd[le] of the panel. In essence, units treated in the middle get more weight as treated and units treate[d] at the beginning or towards the ends get more weight as controls. He introduces three units; the n[ever-treated group,] the early adopters group and the late adopters group. By applying a TWFE mo[del to this type of] data, we are essentially estimating a weighted average of four smaller 2x2 DiDs. T[hese are the early]

adopters vs. never-treated group, the late adopters vs. never-treated group, the early adopters vs. late adopters (as control for early adopters) and the late adopters vs. early adopters (as control for late adopters). The last DiD is problematic because the treatment effects don't remain constant over time within each unit. Due to the time-varying treatment effects, in this fourth group, early adopters have already experienced this "evolution" of the treatment effect, essentially making them a poor control group and resulting in biased estimates for late adopters.

It is rather important to note that each 2x2 DiD in Goodman-Bacon's decomposition estimates a Local Average Treatment Effect (LATE) for the specific timing groups being compared (e.g., early adopters vs. never-treated). The TWFE estimate is a weighted average of these local effects, which be different across groups and time periods. As a consequence, this estimator may not represent a global treatment effect, especially in heterogeneous treatment effect settings.

In 2020, De Chaisemartin and D'Haultfœuille showed that the weight assigned to each treatment effect, τ_{it}, is a weighted average that depends on the residual, with higher weights assigned when there is greater variation in the treatment dummy, similar to what Goodman-Bacon argued in his paper. Smaller weights are given to periods where more units are treated and to units with more treated periods. If staggered adoption occurs and the proportion of treated units is non-increasing in time, later periods end up with smaller or even negative weights. We observe this in the fourth DiD example, where the weights can be negative, even if all τ_{it} are positive, resulting to negative TWFE estimator.

Given this limitation of TWFE's handling heterogeneous treatment effects in staggered adoption, Callaway and Sant'Anna in 2021 addressed this by estimating group-time average treatment effects (ATT(g,t)), that isolated treatment effects for each group and time period, leading eventually to more reliable causal estimates. While their approach offered ways to overcome the biases inherent in TWFE models, such as negative weighting and contamination of treatment effects across groups and time periods, under staggered adoption, it also underscored broader limitations of TWFE when heterogeneous treatment effects are present.

2.1.3 Dynamic Treatment Effects Extension

Another important extension is dynamic treatment effects, which allow us to explore how the impact of a treatment changes over time. Rather than assuming a constant treatment effect, dynamic models use leads and lags to capture how the outcome evolves both before and after the treatment is applied. This approach is crucial for understanding the full scope of the treatment's impact, as effects may not be immediate or uniform across all periods.

The conceptual foundation for studying time-varying causal effects emerged in the 1980s, particularly in the work of Heckman and Robb (1985), who emphasized the importance of accounting for how treatment effects evolve over time, even though the focus was not explicitly on the DiD framework. Recently, this approach was integrated by Callaway and Sant'Anna in 2021 to address treatment heterogeneity across groups and time, in the context of staggered adoption settings, for more precise causal inferences.

2.2 Synthetic Control

Athey and Imbens in 2017 in their paper argued that Synthetic Control (SC) method is the most important innovation in the policy evaluation field in the past 15 years. SC is a method of causal inference that offers a revolutionary, powerful approach for situations where there is only a single treated unit and more control units resulting in traditional methods like DiD to fall short. Causal inference is, in essence, a problem of "missing" information. This method, introduced by Abadie and co-authors in seminal papers like Abadie & Gardeazabal (2003) and Abadie et al. (2010), uses the information of the pre-treatment control unit (donor pool), to predict the counterfactual in the post-treatment period for the treatment unit. The way this method does this is by assigning a weight, a set of fixed numbers that add up to 1, to control units. Once this is done, the synthetic control unit, that approximates a counterfactual for the treated unit, is constructed as a weighted combination of untreated units. This unit replicates a synthetic control trajectory such in pre-treatment period the synthetic control and treated unit show similar trends, but in post-treatment period being convex combination of the control units, serves as a prediction of a counterfactual for the treated unit representing what would have happened to the treated unit in absence of the treatment.

Despite its innovative advantages, in 2018 Bouttell et al. argue that in order for the results to be unbiased, SC depends on finding a well-matched donor pool for the treated unit. Additionally from an algorithmic point of view, SC only handles one treated unit and one outcome at a time and as discussed in a 2022 article by McClelland and Mucciolo, extensions to multiple treated units have been proposed to enhance its applicability. According to Abadie this can be seen as a safeguard, making SC a conservative method that reduces the chance of making large errors by avoiding extrapolation, though this limits its flexibility.

2.2.1 Penalized Synthetic Control

In 2019, Abadie and L'Hour addressed the balance of making large errors while limiting flexibility, by proposing a penalized synthetic control estimator to handle this challenge, associated with disaggregated data. To ensure unique and stable weight estimation, their approach introduces a penalty term that aims to balance the fit between the treated unit and its synthetic control against the discrepancies between the treated unit and each contributing control unit. In essence, Abadie and L'Hour's estimator minimizes an objective function that is a combination of two components. The first component calculates the overall difference between the treated units' and its synthetic controls' characteristics, while the second component penalizes, with the penalization parameter $\lambda > 0$, the differences between the treated unit and each unit that contributes to its synthetic control. As $\lambda \to 0$ the method closely aligns with the traditional SC as it focuses on matching the treated unit's characteristics with the weighted combination of control units, whereas as $\lambda \to \infty$ the method focuses on minimizing the differences between the treated unit and each unit that contributes to its synthetic control, making it resemble more the nearest neighbor matching. This penalization approach improved Abadie's original framework as it aimed to reduce interpolation biases.

2.2.2 Augmented Synthetic Control

To address the dependence of SC into finding a well-matched donor pool for the treated unit and thus its requirement for excellent pre-treatment fit, in 2020 Ben-Michael, Ferell and Rothstein introduced the Augmented Synthetic Control (ASC) method, as an refining extension to the widely used SC. ASC addresses this limitation by combining SC with an outcome model for bias-correction. The outcome model is ridge regression because of its ability to control overfitting through regularization and therefore to improve pre-treatment fit and reduce estimation bias in settings with poor pre-treatment fit. The ridge-ASC improves the synthetic control by allowing the use of negative weights on control units, as a last resort to improve the fit, while penalizing excessive deviations from the original SC weights to maintain interpretability. Then with cross-validation the regularization parameter is determined, in order to capture the essence of the improvement of the pre-treatment fit (reducing bias) while avoiding overfitting to noise (controlling variance). Unlike the SC, ASC is capable of addressing situations, where the treated unit falls outside the convex hull of control units or when auxiliary covariates are critical for estimation, while in high-dimensional or noisy data, ASC's augmentation ensures better balance between the treated unit and the synthetic control, improving the validity of causal estimates.

While the ASC is useful in settings with poor pre-treatment fit, it assumes that all factors influencing the outcome are either observed or uncorrelated with the treatment assignment (strong ignorability assumption) and assumes no interference or spillovers across units, which can be challenging to hold in real-world scenarios. Due to the reliance on regularization, its effectiveness is sensitive to the choice of the regularization parameter (λ_i). A small λ_i can lead to overfitting to noise and therefore to biased and unstable estimates, due to its flexibility in weights adjustment while a large λ_i can result in biased estimates because it heavily penalizes deviations from the standard SC weights, which reduces overfitting but might fail to sufficiently improve pre-treatment fit. To strike a balance between precision and generalizability—avoiding both under-regularization and over-regularization— cross-validation is used to select the appropriate λ_i that minimizes the prediction error on the validation set by testing different λ_i.

Although this new approach is not a full ML model, it introduces a framework for the integration of SC with ML models, which could potentially lead to extensions of hybrid methodologies.

2.2.3 Generalized Synthetic Control

Many interesting, real-world policies are implemented in multiple regions, possibly at different time points, with either simultaneous adoption or more commonly, with staggered adoption.

Generalized Synthetic Control (GSC) method, an extension of the classic SC, was provided by Xu in 2017 as a way to handle these limitations with methods and algorithms that generalize the synthetic control method to cases with multiple units. For cases with multiple treated units, time-varying effects and non-parallel trends, GSC incorporates the Interactive Fixed Effects (IFE) model to capture unobserved time-varying heterogeneity. IFE model accounts for hidden factors that vary between units and over time, offering a flexible counterfactual prediction, important, when units do not share parallel trends, as assumed in SC method. The way it does this, is by estimating these hidden factors from the control units in the pre-treatment period, to represent the unobserved patterns that affect all units over time, and then the model projects treated units onto these estimated factors to predict the counterfactual in the post-treatment period. The GSC model becomes robust with the cross-validation approach that manages to select an optimal number of hidden factors, eventually reducing the risk of overfitting.

With great strengths comes great trade-offs; In his paper, Xu argues the model demands larger time periods (T) and unit dimensions (N) for reliable estimation, because a more limited number of periods or units can reduce the model's effectiveness and precision in estimating latent factors f_t and factor loadings λ_i. GSC also requires significant computational resources due to the complexity of factor estimation and cross-validation and careful selection of number of latent factors for the stability of the estimates (Athey et al., 2021; Gobillon & Magnac, 2016). When these challenges are particularly pronounced, machine learning methods offer a robust and compelling alternative, driven by their ability to relax strict parametric assumptions and improve predictive accuracy (Baiardi & Naghi, 2024).

2.3 Machine Learning for Causal Inference

In an econometrics framework, aiming to understand causal and non-causal relationships between variables, we tend to interpret regression as the question "What is going to happen if we hold every variable fixed and only change one?". However, in a Machine Learning (ML) framework, we adopt a more prediction-focused perspective, wondering "Given a dataset with outcomes and independent variables, could we build a reliable model that predicts outcome from new, unknown independent variables?" (Varian, 2014). Unlike econometrics models, ML predictive models don't focus on assumptions, mechanisms and causality, but rather on the predictive accuracy, robustness and reliability of the models.

The lack of reproducibility has been persistently holding back the credibility of causal estimation in policy problems. To make more informed policy evaluations, we want to ensure that the results are representing accurately the reality. Systematic robustness checks are always essential but rare in econometrics. Some of the main benefits of ML lie in its capacity to construct granular statistical models, generate more precise counterfactual predictions about counterfactuals and to control for confounders without relying on prespecified functional form assumptions. Unlike traditional methods such as Ordinary Least Squares (OLS), which assumes treatment effects vary linearly with X_i covariates, ML algorithms can capture complex, nonlinear relationships. Additionally, as the number of covariates increases—a situation commonly encountered in high-dimensional datasets—the performance of OLS deteriorates due to the curse of dimensionality. In contrast, ML methods are specifically designed to handle high-dimensional data more accurately and effectively (Mullainathan & Spiess, 2017). These characteristics of ML allow for a reproducible and systematic approach to address causal inference challenges (Baiardi & Naghi, 2024). Additionally in econometrics we favor unbiased results (as OLS is BLUE) (Wooldridge, 2016), whereas in ML settings, where the goal of the model is to generalize well to new covariates, we explicitly allow for bias to reduce overfitting and improve predictive accuracy due to the bias-variance tradeoff in predictions (Geman, Bienenstock & Doursat, 1992).

Despite their core differences, these frameworks can be applied in causal inference by decomposing them into predictive and causal components. Using ML to understand the sources of heterogeneity in treatment effects can allows us to better target treatments to groups that would benefit more from the treatment (Lechner, 2023). By leveraging flexible, data-driven models, ML is able to work with high-dimensional data, select the best model for these data, and evaluate model to not only fit the data well, but also support its ability to do inference (Mullainathan & Spiess,

2017). ML methods perform exceptionally in prediction tasks but due to their ability to evaluate their predictions outcomes via a test set, they don't achieve well-established statistical properties (Athey & Imbens, 2017). Unlike ML, for causal inference the ground truth for each observation is unknown, as it is unknown what would have happened in the absence of the treatment for the treatment group and if the control group had taken the treatment. By introducing ML methods into causal inference, we aim to decompose the challenge (Imbens & Rubin, 2015) and identify the components that ML optimization can be applied without compromising the primary objective of obtaining accurate estimands of causal effects (Athey, 2015).

Therefore, the question lies in how we can obtain and combine the best of both worlds.

2.3.1 Double Machine Learning

In the context of estimating causal effects in high-dimensional settings, in 2018 Chernozhukov et. al. introduce the combination of ML methods with traditional statistical methods. Regularization, while helpful for preventing overfitting, can introduce bias, underscoring the need for debiasing techniques. The Double Machine Learning (DML) framework operates in two stages and flexibly leverages ML techniques, like lasso or random forests, to debias estimators, ensuring reliable causal inference even when dealing with large and complex datasets. It eliminates the bias introduced from regularization in ML models through orthogonalization of score functions and cross-fitting. Neyman orthogonal scores make the process less sensitive to ML-errors from the first stage when estimating the causal parameter, while splitting the data into two parts—one for ML predictions and the other for estimating the parameter of interest—finally alternating their roles and combining the results, eliminates overfitting biases.

By estimating these nuisance parameters, we can remove their effects and focus on the main effect of interest. To do that, in the first stage, a ML model predicts the outcome Y from a set of covariates X, and another model predicts the treatment T from the same covariates. These models yield predicted values $\hat{Y}$ and $\hat{T}$. In the second stage, residuals are calculated by subtracting these predicted values from the actual outcomes and treatment values. This process, called partialling out, effectively removes the influence of the covariates X, isolating the variations in the Y and T that are independent of the X. The residuals are then used in an orthogonalized regression framework where the causal effect of T on Y is estimated with Neyman orthogonality ensuring that errors in the first-stage nuisance parameter estimation do not introduce bias in the estimation of the causal effects.

DML theoretically offers consistency and efficiency but handling complex and high-dimensional datasets can be computationally intensive, and the quality of the nuisance parameter estimation highly depends on ML algorithms and their tuning. Additionally, like most causal inference methods, DML assumes unconfoundedness which might not hold in real-world scenarios. Despite these, as an advanced method, DML manages to effectively bridge the gap between predictive modeling and causal analysis through the combination of ML with econometric models.

2.3.2 Causal Forests

Causal forests, introduced by Wager and Athey in 2018, come to build on the strengths of decision trees and Random Forests while adapting on causal inference frameworks. By averaging over many trees, Causal Forests provide a more robust framework for estimating heterogeneous treatment effects across different subpopulations. In policy-making because of the treatment effects being rarely uniform, this flexibility makes them extremely useful. The Random Forests technique, introduced by Breiman in 2001, is widely used to predict an outcome as a function of the independent variables. A random forest is a collection of trees. To introduce variability and improve generalization, the forest is constructed by taking different random sub-samples of the data, with replacement, and building the optimal tree for each sample (bootstrapping). For each tree build, we generate a prediction for each observation, ending up with multiple predictions for each observation—one from each tree. By averaging all of these predictions we get a final aggregated prediction for each observation. The use of bootstrapping and aggregation is essential because it reduces the risk of overfitting, while improving the reliability of the estimates. Causal Forests are the combination of this ML technique with causal inference. In Causal Forests we essentially build Random Forests but instead of predicting the outcome, the focus shifts on predicting each observation's conditional average treatment effect (CATE), as a function of independent variables:

$$\mathbb{E}[\tau_i|X_i] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

Conditional Average Treatment Effect (CATE) is the key to precision policy-making as it measures how the effect of the treatment changes based on specific characteristics, giving the ability to policy-makers to identify which countries benefit from the treatment and which don't. Unlike standard Random Forests, Causal Forests select splits to maximize differences in treatment effects between subgroups, using the criterion:

$$\max \sum_{i=1}^{n} \tau(X_i)^2$$

To avoid overfitting and ensure more unbiased results, Causal Forests also employ honest splitting, a technique that separates the data in two groups. One group is used for building trees and determine the optimal splits, while the other group is used to accurately estimate treatment effects within each leaf. Finally, the CATE for each observation is obtained by averaging the treatment effect estimates from all trees, ensuring a stable and unbiased prediction.

2.4 Heterogeneous Treatment Effects: A Methodological Gap

While existing literature on policy evaluation has extensively explored average treatment effects of EU subsidies by leveraging Difference-in-Differences and Two-Way Fixed Effects, there is a significant gap in understanding the heterogeneous treatment effects, as those methods often overlook these effects especially in settings with staggered adoption and time-varying impacts. This research addresses this gap by explicitly focusing on estimating heterogeneous treatment effects in EU regional subsidies. By comparing Generalized Synthetic Control, which validates global treatment effects, and Causal Forests, which uncover subgroup-specific effects, this study explores the underutilized potential of Causal Forests in identifying important EU regional variations in policy impact and effectively contributing to more context-sensitive policy design.

3 Addressing the Challenges of Causal Inference: From DiD to Causal Forests

Earlier i mentioned that causal questions are crucial when evaluating the impact of policies or interventions. But how is this best approached? In the real world, something either happens or it doesn't. Therefore, how can one compares what actually happened (the factual) with what would have happened if a specific cause hadn't occurred (the counterfactual)?

This "what if?" scenario lets us imagine how things might have unfolded under a different decision. It's a simple but fundamental question in the social sciences, and lies at the heart of the potential outcomes framework—a concept first introduced by Neyman in 1923. This framework provides a structured way to think about causality by assigning two potential outcomes to each individual or unit: one if the treatment occurs and one if it does not (the control). The causal effect is the difference between these two potential outcomes for a given individual or unit.

In the Potential Outcomes Framework, a binary treatment variable D_i for each individual or unit i is considered, where:

$$D_i = \begin{cases} 1 & \text{if the individual receives the treatment,} \\ 0 & \text{if the individual does not receive the treatment.} \end{cases}$$

For each individual i, the two potential outcomes are:

- $Y_i(1)$: the outcome for individual i if they receive the treatment ($D_i = 1$).
- $Y_i(0)$: the outcome for individual i if they do not receive the treatment ($D_i = 0$)

and the causal effect for individual i is:

$$\delta_i = Y_1 i - Y_0 i$$

Due to the Fundamental Problem of Causal Inference—both versions of potential outcomes cannot be observed at the same time, it is impossible to directly measure individual causal effects (Holland, 1986).

Holland emphasized a statistical solution to this; to use randomization in order to leverage information from other groups to get an idea of an average counterfactual. The idea was that, even if we cannot directly observe the true counterfactual, in a Randomized Control Trial the control group's average (Average Treatment Effects (ATE)) could serve as an estimate for the average counterfactual:

$$\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

A different approach to the Potential Outcomes Framework is the Structural Causal Model (Pearl, 2000), that exploits invariance assumptions of the model arguing that by manipulating the model under different interventions, in the long-run the model itself can serve as a counterfactual and the deviations from that model can be attributed to either random noise or the treatment itself.

Estimating causal effects in practice is further complicated by selection bias and confounding. Selection bias arises when units are not randomly selected for treatment and control groups, causing the groups to be different in ways that can influence the outcome. As a result, any observed differences in outcomes may be driven by these pre-existing characteristics rather than the treatment itself, making it difficult to determine whether the treatment is truly responsible for the observed effects and eventually leading to biased conclusions about the causal relationship. Therefore, in practical terms, the estimation involves comparing the average outcomes of treated groups with those of untreated groups, which provides the average treatment effect among the treated (ATT), including a selection bias term, as treatment assignment is not random:

$$\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] = \underbrace{\mathbb{E}[Y_1 - Y_0|D=1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0]}_{\text{Selection Bias}}$$

Confounding occurs when an external variable, the confounder, due to its relation to the treatment and the outcome influences both of them, resulting in biased estimates and misleading conclusions of the treatment effect (Greenland & Robins, 2009).

To address for these biases we rely on strong assumptions, like Conditional Ignorability, that assumes that treatment assignment is independent of potential outcomes, given observed covariates (Chernozhukov et. al., 2024). Traditional methods like DiD and TWFE models depend heavily on this assumption. However, newer approaches, such as GSC that uses advanced regularization techniques and Causal Forests that leverage machine learning are used to overcome the limitations of traditional methods by handling heterogeneity and complex data more effectively.

In the following sections, I will explore how these advanced techniques handle the challenges of causal inference—particularly in the context of real-world EU subsidies data—and offer more robust solutions for estimating treatment effects.

3.1 Conditional Ignorability and Confounders

Conditional Ignorability, also known as conditional exogeneity or assumption of no unmeasured confounding, is a key concept in causal inference. It assumes that, given a set of covariates X, the treatment assignment D is independent of the potential outcomes $Y(1)$ and $Y(0)$. This assumption implies that, once we account for the relevant covariates, the treatment can be assumed to be randomly assigned, allowing us to estimate causal effects without bias from unmeasured confounders (Chernozhukov et. al., 2024).

Formally, this assumption is written as:

$$Y_i(d) \perp D_i \,|\, X_i \quad \text{for all } d \in \{0,1\}$$

It's important to note that Conditional Ignorability is fundamentally untestable; it is a theoretical assumption based on domain knowledge and is often represented visually through causal diagrams (Directed Acyclic Graphs or DAGs) (Tennant et. al., 2021).

In both traditional and ML approaches, choosing the right covariates is essential to ensure that our estimates aren't biased by selection bias or confounding. By adjusting for these relevant covariates, we reduce bias and improve the accuracy of our causal estimates.

Advanced methods like GSC and Causal Forests build on this concept by offering more flexible ways to account for complex data structures and heterogeneous treatment effects. GSC extends traditional synthetic controls by using interactive fixed effects, incorporating hidden factors to address time-varying unobserved heterogeneity. In contrast, Causal Forests provide a non-parametric method for estimating heterogeneous treatment effects across different subpopulations. Relying on decision trees and honest splitting, they adjust for covariates while avoiding confounding.

3.1.1 Causal Diagrams and Confounders

Causal diagrams, especially DAGs, are essential tools for visualizing the assumptions in causal models. A DAG helps to clearly represent the relationships between variables, making it easier to identify confounders—variables that affect both the treatment and the outcome. In order to avoid bias in estimating the true treatment effects we have to control for these confounders.

$$D \leftarrow X \rightarrow Y$$

In the diagram above, X is a confounder that influences both the treatment D and the outcome Y, potentially creating a spurious relationship between them. Failing to account for X, our estimates of the causal effect of D on Y will be biased, as the variation in Y could be partly due to the influence of X. By conditioning on X—essentially controlling for it—we "close the backdoor" path and isolate the causal effect of D on Y, ensuring that they no longer share a common cause.

DAGs also help identify other sources of bias, such as collider bias and selection bias. Collider bias occurs when a variable (S) is influenced by both the treatment and the outcome. Unlike confounders, conditioning on a collider introduces a spurious association/dependence between D and Y.

$$D \rightarrow S \leftarrow Y$$

When conditioning on S (for instance, by including it in our analysis), a backdoor path opens between D and Y, introducing a spurious relationship—a relationship between them that didn't exist—that distorts the true causal effect (Hernán & Robins, 2020).

When selection bias occurs in analysis due, to focusing on a subgroup of the population that does not accurately represent the population it can skew the relationships being estimated. Sometimes it can be seen as a type of collider bias from conditioning on a common effect of two causes (D and Y). In essence, selection bias arises from factors, like random sampling or omitting particular groups based on their treatment or outcome status. By accounting for these potential biases and adjusting for all relevant confounders, we can minimize biased selections and generate more precise evaluations of the treatments impact (Hernán & Robins, 2020).

3.2 One Size Doesn't Fit All: Heterogeneous Treatment Effects

While estimating the ATE can provide valuable information, for more precise policy-making, the critical point is to understand how treatment effects vary across different sub-populations and ensure that findings are accurate and not due to sampling variation. Sampling variation can be challenging for targeted policy because of the uniqueness of each unit. Due to this uniqueness, the data to predict exactly the effects of a treatment for any particular unit, are limited. Taking it a step further, the question is not only "for whom" the treatment is effective, but also "why" it is. By identifying groups that treatment has different effects, hypotheses can be generated about the underlying mechanisms that drive these differences. In essence, understanding which groups benefit more from the treatment can lead to policy reformulation and more improved treatments.

3.3 DiD and TWFE Limitations

In this section, I will outline the main limitations of both DiD and TWFE. While DiD is [illegible] in many contexts, it runs into challenges when there is only one treated unit or [illegible] a small number of treated units. In those cases, it is difficult to find an appropriate control group, and the estimates become unreliable (Callaway & Sant'Anna, 2021). DiD also depends heavily on the parallel trends assumption—the idea that the treated and control groups would have followed the same trend if there

were no treatment. Under this assumption, the goal of the DiD estimator is to estimate the unbiased ATT:

$$\text{ATT}_{\text{DiD}} = \mathbb{E}[Y_{it}(1) - Y_{it}(0)|D_i = 1]$$

But when that assumption does not hold, the DiD estimator can produce biased results, as illustrated by the bias formula below:

$$\text{Bias}_{\text{DiD}} = \mathbb{E}[(Y_{it}(0) - Y_{i,t-1}(0))|D_i = 1] - \mathbb{E}[(Y_{it}(0) - Y_{i,t-1}(0))|D_i = 0]$$

With TWFE, the big limitation shows up in staggered adoption settings, where units are treated at different times. TWFE can produce biased estimates by combining positive and unintended negative weights, as a result of early and late adopters comparisons, distorting the treatment effect. This issue gets even more problematic when treatment effects are dynamic-changing over time. Goodman-Bacon's decomposition illustrates this problem that leads to misleading estimates of the treatment's actual impact:

$$\text{TWFE}_{\text{ATT}} = \sum_{(g,g')} \omega_{g,g'} \cdot \text{ATT}_{g,g'}$$

here (g, g') is each unique two-group comparison, g and g' are the groups that receive treatment at different times, $\text{ATT}_{g,g'}$ is the DiD estimate of the ATT for the comparison between g and g', capturing the difference in outcomes under the assumption that the trends of g and g' would be the same in the absence of treatment and $\omega_{g,g'}$ are the weights for each comparison, which can be negative in early and late adopters comparison.

3.3.1 DiD and Parallel Trends

As mentioned earlier, the most critical assumption in a DiD model is the parallel trends assumption, which is crucial to ensure that any observed differences in outcomes after treatment are due to the treatment itself, rather than some other factor. When this assumption is violated, the treatment effect estimates may be biased, eventually leading to incorrect conclusions about the intervention's impact. To address confounders, selection bias, or time-varying factors, several initial methods were introduced to approximate parallel trends (balance treated and control groups on observed covariates).

Confounders and selection bias can weaken the parallel trends assumption by introducing variables that affect both the treatment assignment and the outcome. In panel data settings, the unobserved factors that influence both treatment and outcome can cause the treated and control groups to follow different trends without treatment. For instance, if the treated group is more exposed to specific economic conditions or demographic factors, the differences in outcomes might reflect those confounders and not the treatment itself. Selection bias can also come into play when groups are non-randomly selected, but are chosen based on characteristics tied to both the treatment and the outcome. This bias also happens when conditioning on colliders, which opens backdoor paths between treatment and outcome, eventually distorting the causal estimates. The problem is further amplified when there are time-varying confounders, making the parallel trends assumption particularly fragile. One way to address this important violation and approximate the parallel trends assumption is by including covariates in the regression model to account for differences between the treated and control groups.

$$Y_{it} = \alpha + \delta\,\text{Treatment}_i + \gamma_t + \theta(\text{Treatment}_i \times \text{Post}_t) + \beta X_{it} + \epsilon_{it}$$

here Y_{it} is the outcome variable for unit i at time t, Treatment_i is the indicator for the treated group, θ is the treatment effect estimate, Post_t is the indicator for the post-treatment period, β are the coefficients for the covariates and X_{it} is the vector of covariates that account for confounding factors.

But this approach has its own challenges. While it helps address confounding, adding too many covariates can introduce overfitting or mis-specification, which further complicates estimation and interpretation, leading to biased treatment effect estimates if these relationships are not correctly modeled (Angrist & Pischke, 2009).

To address selection bias, Inverse Probability Weighting (IPW) is used (Rosenbaum and Rubin, 1983). Each observation is weighted by the inverse probability of receiving treatment, estimated via a propensity score. In the standard DiD model with a treatment effect:

$$Y_{it} = \alpha + \theta(\text{Treatment}_i \times \text{Post}_t) + \epsilon_{it}$$

observations are weighted as:

$$w_i = \frac{D_i}{p(X_i)} + \frac{1 - D_i}{1 - p(X_i)}$$

with D_i, $i \in \{0, 1\}$, being the indicator for treatment and $p(X_i)$ the propensity score (the probability of treatment given covariates X_i).

IPW's role is to make the treated and control groups more comparable on observed covariates by ensuring that every observation contributes to the analysis based on how likely it was to receive the treatment and therefore reducing selection bias. However it cannot address bias due to unobserved confounders.

Combining these two approaches, the Doubly Robust (DR) estimator offers additional robustness to selection bias (Sant'Anna & Zhao, 2020):

$$Y_{it} = \alpha + \theta(\text{Treatment}_i \times \text{Post}_t) + \beta X_{it} + \epsilon_{it}$$

The goal of this approach is to balance treated and control groups on observed covariates while using covariate adjustment to adjust for residual differences resulting in selection bias reduction due to observed factors. However due to the sensitivity to unobserved confounders, to produce unbiased estimates it heavily relies on specifying either the propensity score model or the covariate adjustment model correctly.

3.3.2 TWFE Limitations

The TWFE estimator is widely used in DiD models to estimate changes in outcomes by applying time and group fixed effects.

$$Y_{it} = \alpha_i + \gamma_t + \theta(\text{Treatment}_i \times \text{Post}_t) + \epsilon_{it}$$

In the standard TWFE DiD model α_i is the individual fixed effects that capture time-invariant unobserved characteristics for each unit and γ_t are the time fixed effects that capture any time-specific shocks affecting all units.

However this method faces significant challenges in staggered adoption settings, where the timing of treatment varies across units (Goodman-Bacon, 2021). In these cases, TWFE estimates a variance-weighted average of treatment effects, and some of those weights end up being negative. That happens when early treatment adopters are used as controls for groups that are treated later, which introduces bias, especially when treatment effects change over time. The estimated treatment effects are distorted because those earlier units are still being affected by the treatment, leading to invalid comparisons and less reliable, biased estimates.

Additionally if treatment effects vary over time, TWFE produces a biased estimate because in essence it assumes a constant treatment effect θ across all units and time periods (Callaway & Sant'Anna, 2021).

$$Y_{it} = \alpha_i + \gamma_t + \theta_{it}\,\text{Treatment}_{it} + \epsilon_{it}$$

In the dynamic setting with time-varying treatment effects above, θ_{it} is the time-varying treatment effect specific to unit i and time t, that TWFE cannot directly estimate. Instead it averages these time-varying effects across units and periods, introducing bias in scenarios where treatment effects are heterogeneous across units or vary over time (De Chaisemartin & D'Haultfœuille, 2020).

3.4 Synthetic Control

To reduce the biases that TWFE introduces in staggered adoption settings, more flexible models like SC have been developed, to provide a more reliable counterfactual for the treated unit. SC methods handle unobserved heterogeneity across units and over time by creating a synthetic control group which is a weighted combination of untreated units with the weights vector $W = (w_1, w_2, \ldots)$ that produces this by solving the following objective function:

$$\min_{W} \sum_{k=1}^{K} v_k \left(X_{1k} - \sum_{j=2}^{J+1} w_j X_{jk} \right)^2$$

here X_{1k} is the value of pre-treatment covariate k for the treated unit, X_{jk} is the value of pre-treatment covariate k for control unit j, v_k is a weight reflecting the importance of covariate k in the matching process and $W = (w_2, \ldots, w_{J+1})$ are the weights for the control units, constrained such that $w_j \geq 0$ and $\sum_{j=2}^{J+1} w_j = 1$ (Abadie, Diamond & Hainmueller, 2010).

When the synthetic control group is constructed, the treatment effect for the treated unit $i = 1$ at time t is estimated as the difference between the outcome for the treated unit and the outcome for the synthetic control (Abadie & Gardeazabal, 2003):

$$\text{Treatment Effect}_{it} = Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}$$

here Y_{1t} is the observed outcome for the treated unit at time t and $\sum_{j=2}^{J+1} w_j Y_{jt}$ is the estimated counterfactual outcome for the treated unit, constructed from the weighted outcomes of control units.

3.5 Generalized Synthetic Control

Expanding the innovative idea of replicating SC trajectories based on control units' weights for a single treated unit, GSC takes this idea a step further. As discussed in the literature review section, GSC method improves SC by handling multiple treated units, time-varying effects and non-parallel trends.

In his paper, Xu uses a factor-augmented model to motivate the SC method to include both observed and unobserved factors:

$$Y_{it}(0) = \theta_t' Z_i + \xi_t + \lambda_i' f_t + \varepsilon_{it}$$

where the counterfactual outcome is $Y_{it}(0)$, the observed covariates are Z_i, the time-fixed component ξ_t, the hidden factors f_t with unit-specific factor loadings λ_i. The essence of the GSC, the IFE component ($\lambda_i' f_t$) captures the unobserved heterogeneity that vary over time. Although GSC relaxes the parallel trends assumption for observed outcomes, for consistent counterfactual predictions it assumes parallel trends in latent factors. Thus, in the absence of treatment, the latent factors ($\lambda_i' f_t$) unfold similarly across treated and control units:

$$\Delta F_t = F_{t+1} - F_t$$

For valid counterfactual construction, based on the overlap assumption, treated units should obtain adequately similar control units for both observed covariates (Z_i) and latent factors (λ_i):

$$P(D_i = 1 \mid Z_i, \lambda_i) > 0 \quad \text{and} \quad P(D_i = 0 \mid Z_i, \lambda_i) > 0$$

for all i in the sample. Under this assumption no treated unit lies outside the support of the control units, which prevents extrapolation beyond the observed data. It is also crucial to assume that in the pre-treatment period ($t < T^*$), the treatment effect is assumed to be zero to ensure that the behavior of treated units in the pre-treatment period is an unbiased baseline to estimate the counterfactuals (no anticipation of treatment):

$$\delta_{it} = 0 \quad \forall t < T^*$$

Moreover, in order to ensure valid causal inference and confirm that unobserved factors influencing the outcome are not methodically associated to the treatment or observed variables, the model assumes strict exogeneity of the error term, which means that the error term ϵ_{it} is uncorrelated with the treatment assignment, covariates and latent factors:

$$\mathbb{E}[\epsilon_{it} \mid W_{it}, X_{it}, \lambda_i, F_t] = 0$$

It also assumes that the treatment of one unit does not influence the outcomes of other units (no interference or spillover effects/dependencies between treated and control units - SUTVA):

$$Y_{it}(d_i, d_j) = Y_{it}(d_i) \quad \forall j \neq i$$

Here $Y_{it}(d_i, d_j)$ is the outcome of unit i under its own treatment status d_i and the treatment status of other units d_j. To prevent spurious correlations and make certain that the estimation of latent factors is consistent, the assumption of weak serial dependence of errors ensures that the error term ϵ_{it} does not show strong autocorrelation over time:

$$\text{Cov}(\epsilon_{it}, \epsilon_{is}) \to 0 \quad \text{as } |t - s| \to \infty$$

Xu's innovation now builds on the SC method, by incorporating a IFE model, eventually creating the GSC method. In this method, a three-step approach is utilized based on a latent factor model. To guarantee proper identification of the latent factors and factor loadings, while preventing overfitting the factor model incorporates two identification conditions. The fist condition is the orthogonality and normalization of latent factors to ensure that the factors are uncorrelated and have unit variance, resolving scale and rotational indeterminacy:

$$\frac{1}{T} F'F = I_r$$

Here F is a $T \times r$ matrix of latent factors, T is the number of time periods, r is the number of factors and I_r is the $r \times r$ identity matrix. The second one is the orthogonality of factor loadings that ensures that the factor loadings corresponding to different factors are uncorrelated, allowing for unique decomposition of the model's covariance structure:

$$\Lambda'\Lambda = \text{diagonal matrix}$$

Here Λ is an $N \times r$ matrix of factor loadings and N is the number of cross-sectional units.

For the first foundational model, the IFE uses data from the control units, during the pre-treatment period to estimate, the hidden factors f_t that represent the unobserved patterns that affect treated and untreated units over time and how much each control unit is affected from these factors λ_i'.

$$\textbf{Control:} \quad Y_{it}(0) = X_{it}'\beta + \alpha_i + \xi_t + \lambda_i' f_t + \varepsilon_{it}$$

Here, $X_{it}'\beta$ are the observed covariates, α_i are the unit-specific fixed effects, ξ_t are time-specific effects, $\lambda_i' f_t$ the hidden factors that vary over time and across units and capture the unobserved heterogeneity and the error ε_{it}. Additionally, to ensure unbiased and efficient estimation, the model makes the assumption that errors are independent across units and have constant variance (cross-sectional independence and homoscedasticity of the error term):

$$\mathbb{E}[\epsilon_{it}\epsilon_{js}] = 0 \quad \forall i \neq j$$

$$\text{Var}(\epsilon_{it}) = \sigma^2$$

Second, the same model structure is used for ensuring that in a "no-treatment" scenario the treated units would have similar patterns to control units. For the treated units, the pre-treatment period provides a baseline for observing whether the latent factor structure holds for treated units as well.

$$\textbf{Treatment (pre-treatment):} \quad Y_{it}(0) = X_{it}'\beta + \alpha_i + \xi_t + \lambda_i' f_t + \varepsilon_{it}$$

Then, in the post-treatment period, the model projects the treated units onto the factors to create the counterfactual prediction for the treated units.

$$\textbf{Treatment (post-treatment):} \quad Y_{it}(1) = X_{it}'\beta + \alpha_i + \xi_t + \lambda_i' f_t + \varepsilon_{it} + \tau_{it}$$

Here, $Y_{it}(1)$ is the actual outcome and τ_{it} treatment effect.

This way, by first estimating hidden factors in the control group and then projecting treated units onto these factors, GSC generates counterfactuals without the need for parallel trends. Unlike SC which constructs non-parametric weighted average of control units, GSC is a more model-based approach utilizing the IFE model.

3.5.1 Limitations of GSC

Hidden factors and loadings require careful specification making GSC's structure more challenging than the straightforward SC. Because of GSC not relying on non-parametric weights like the classic SC, but rather on parametric factor model, it ends up with interpretation challenges (Xu, 2017). Additionally, in GSC a small number of time periods (T) and few control units (N) can make the model unable to identify the actual trends and that could lead to overfitting or unstable predictions. While cross-validation can help to determine the right number of factors, with a smaller dataset, there is a higher risk of picking a number that doesn't fully represent the true underlying trends, affecting the model's accuracy. Cross-validation also increases the computation intensity of the method, as it demands significant computational resources for choosing the optimal number of factors.

3.6 Addressing Challenges of Causal Inference with Machine Learning

After discussing the broader issue of historical Causal Inference challenges, i will attempt to address some of the challenges with Machine Learning, particularly with Causal Forests. In the previous chapter I wonder how ML can be leveraged to obtain data-driven identification of treatment effect heterogeneity while maintaining reproducibility and allowing for valid statistical inference. The answer is, through Causal Forests. Moving away from the "one-size-fits-all" world towards a world with customize solutions, the promise of Causal Forests is precision policy-making, by providing reliable estimates of which countries seem to benefit from a treatment and which don't (Athey, 2019).

As discussed, Causal Forests are a combination of Random Forests with causal inference to estimate CATE. In 2018 Wager and Athey provided the first formal proof that Random Forests can be used not only for predicting outcomes, but also for estimating treatment effects with confidence that the inferences are statistically valid in large samples. In essence they proved how to modify the Random Forests algorithms to make them useful for inference.

3.6.1 Mechanics and Challenges of Causal Forests

Causal Forests are constructed from multiple causal trees. A causal tree begins with one "question" that partition the data into two branches based on the answer and recursively each branch gets partitioned further based on following series of "questions", eventually creating a tree structure. At the end of each branch, in each leaf, each tree contains a group of units with similar characteristics of the factors used to partition the data. Then the causal tree estimates the treatment effect for each leaf, assuming consistent treatment effect within each one (Wager & Athey, 2018).

Causal Forests rely on unconfoundedness assumptions and honesty in the tree-splitting process, to provide asymptotically unbiased treatment effect estimates and avoid overfitting. Under unconfoundedness assumption treatment assignment D_i (1 for treated and 0 for control) is assumed to be independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on the covariates X_i to ensure that X_i will capture all confounders affecting treatment and outcome:

$$Y_i(1), Y_i(0) \perp D_i \mid X_i.$$

Additionally, under the overlap assumption there must be a non-zero probability of receiving both treatment and control, for all values of X:

$$0 < P(D = 1 \mid X) < 1$$

Moreover, SUTVA makes certain that there is no interference between units and ensures consistency in the definition and application of treatment. In addition to the structural assumptio[illegible]s of Causal Forests above, the process of tree construction relies to a great extent on the honesty pr[illegible]nciple. Honesty is crucial in the construction of the trees, in order to not allow for bias into the analy[illegible]sis. It strategically prevents the trees from using the outcome data to decide where to partition the data, rese[illegible]arch [illegible] estimate treatment effects exclusively after the partition process is completed (Wa[illegible]ger & Athey, 20[illegible]). An essential question is how to enforce honesty in tree construction. One appr[illegible]oach is the double-sample trees, where the data are separated into two parts. The first part is used to [illegible] the tree by partitioning the data into groups and the second part is used to estimate the treatment effect for each group. To identify the groups and ensure that the tree stays on track, Causal Forests introduce

controlled randomness with random feature selection. In each split, the tree randomly chooses a subset of features to examine to ensure that no single feature, even if it is irrelevant, can hijack the splitting process, in order to keep the focus on the most important factors (Breiman, 2001). Additionally, Causal Forests, instead of just grouping units with similar outcomes or based on prediction accuracy that hides the effect of the treatment, use a splitting rule that focuses on maximizing the differences in treatment effects between the groups created by each partition. The greater the difference, the more opportunity there is to observe how the treatment affects different units differently.

However, how do these special splitting rules are implemented? The second approach is propensity trees, which ensure the trees are honest without actually splitting the data into two parts. They use only the treatment assignment to decide how to partition the data, without allowing for the outcome data to bias the partitioning decisions but only use them to estimate the treatment effect, once the structure of the tree is completed (Wager & Athey, 2018). Wager and Athey's formula below, computes the treatment effect for each subgroup, using the estimation sample to avoid overfitting.

$$\hat{\tau}(x) = \frac{1}{n} \sum_{i \in \text{Estimation Sample}} \left(\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{e}(X_i)} \right)$$

where D_i is the treatment indicator for unit i, Y_i is the outcome for unit i, $\hat{e}(X_i)$ is the estimated propensity score, or the probability of receiving treatment given covariates X_i

Once the honest tree is build, with sample splitting or propensity tree, we can use the correct data to estimate the treatment effect for each group. Building an honest structure and then estimating treatment effects within that structure is a key element of what makes Causal Forests so effective.

Causal Forests are build by averaging multiple honest trees.

$$RF(x; Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_s \le n} \mathbb{E}_{\xi \sim \mathcal{D}} \left[T(x; \xi, Z_{i_1}, \ldots, Z_{i_s}) \right]$$

here $T(x; \xi, Z_{i_1}, \ldots, Z_{i_s})$ is the prediction from a single tree, and ξ is the randomness in the tree-building process.

As the sample size increases, the treatment effect estimates are asymptotically normal (Gaussian) and centered around the true treatment effect.

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2(x) \xrightarrow{p} 0$$

This asymptotic normality allow for constructing confidence intervals around the estimates, instead of point estimates, providing a range for uncertainty in the estimates (Wager & Athey, 2018). To quantify this uncertainty, the variance around the treatment effect is computed, to allow for precision measurement. Initially, an estimate of this variance is calculated across subgroups based on the splitting of data, capturing the variability of treatment effect estimates across subgroups:

$$\hat{\sigma}^2_{\tau(x)} = \text{Var}(\hat{\tau}(x) | X = x)$$

For a more robust estimate, Wager and Athey suggest the infinitesimal jackknife variance estimator that measures the variation of the treatment effect estimates across all the trees in the causal forest. In the variance formula below, by capturing this variation, the confidence intervals around the treatment effect are more reliable and less likely to be too narrow:

$$\hat{V}_{IJ}(x) = \frac{n - 1}{n} \left(\frac{n}{n - s} \right)^2 \sum_{i=1}^{n} \text{Cov}^* \left[\hat{\tau}^*_b(x), N^*_{ib} \right]^2$$

here $\hat{\tau}^*_b(x)$ is the estimate from the b-th tree, and N^*_{ib} shows if the i-th sample was included in the b-th tree.

The confidence intervals for the treatment effect are calculated as:

$$\text{CI} = \hat{\tau}(X_i) \pm 1.96 \sqrt{\hat{V}_{IJ}(X_i)}$$

Traditional methods that was used until now, such as k-NN neighbours and propensity score matching, aim to find matches based on similarity but in high-dimensional datasets identifying meaningful

similarity among units is rather impossible. In such situations, Causal Forests shine and improve. The more complex and high-dimensional the data, the more effectively they filter out irrelevant noise of the data and focus on accurate results. When adding more noise, with random feature selection, Causal Forests disregard all the irrelevant information by examining only a subset of features at each split. By focusing on details and not overall trends, Causal Forests capture the treatment variation effectively. They adaptively dive into targeted policy-making by addressing not only if the treatment works but for whom does it is more effective. This is valuable not only for tailoring interventions but also for understanding why different units respond differently to treatment. By understanding what makes units different, we can gain insights into how treatments actually work and how to improve them.

Causal Forests face challenges too. One issue is controlling bias, especially at the edges of the data, where the estimates may be less accurate, due to limited observations. As an ML method, they rely on training data and when predicting outside from training range, boundary bias can occur. Another challenge is developing more robust methods to estimate variance, especially when data are limited or there is a large number of features. Additionally, parameter tuning is crucial for Causal Forests, as finding the balance between bias and variance can be challenging (Hastie, Tibshirani, & Friedman, 2009).

3.7 Structured vs. Flexible Causal Inference: GSC vs. Causal Forests

Extensive research and simulations comparing k-NN with Causal Forests have been already conducted by Wager and Athey in 2018; however, in this paper, i will attempt to compare them with the GSC method.

Xu's SC extension, GSC is typically more structured, by using parametric factor models like the IFE model, incorporating latent factors to capture unobserved time-varying heterogeneity across units over time. It is designed for panel data with observable time patterns in treatment effects across groups. However, it may not be as flexible as the fully non-parametric method of Causal Forests. The non-parametric Causal Forests extends Random Forest, and more flexibly allows for highly granular estimates of heterogeneity in treatment effects across high-dimensional covariates, based on individual characteristics.

Because of the IFE model incorporation, GSC handle more effectively, settings with multiple treated units and time-varying trends making it generally robust in scenarios with staggered adoption. The IFE model also allows GSC to flexibly model counterfactual outcomes when complex, time-varying unobserved factors are present, adapting well to situations where trends are not parallel between treated and control units. For further robustness, GSC use cross-validation to optimize the number of latent factors, reducing the risk of overfitting. The reliable estimation of latent factors and factor loadings is possible as long as there is a sufficient number of time periods (T) and units (N). Additionally, cross-validation and latent factor estimation are computationally intensive tasks and unlike simpler SC models, GSC's parametric factor model adds complexity, making interpretation challenging.

On the other hand, Causal Forests are designed for high-dimensional data because of their ability to handle many covariates effectively and model complex interactions. In individual-level predictions, for more precise policy-making, this method thrive due to the identification of the treatment effects for subgroups. Additionally by providing asymptotic confidence intervals for treatment effects they support hypothesis testing and inference in complex observational data settings. However, they are sensitive to data structure as they perform best with large sample size and balanced distribution of treated and control units, but if the data are limited or unbalanced it may not capture accurately the treatment heterogeneity. They also rely on unconfoundedness and if confounders are not fully accounted for, there could be biased estimated treatment effects.

Naturally, GSC is more appropriate for policy evaluations in structured settings when the treatments are introduced at different times across groups and parallel trend assumptions don't hold. Whereas Causal Forests are unmatched in precise policy-making where the goal is to estimate group treatment effects and especially in high-dimensional, non-linear settings with complex covariate interactions. Both of them provide answers, but the final choice depends on the specific research question one seeks to get answers for.

4 Methodologies: Bridging Theory and Practice

The following section contains the empirical application of GSC and Causal Forests methods and comparison of their results, on real data from the EU in order to observe the strengths and areas of improvement for each method. These methods are tested on how they handle real-world complexities and are compared to highlight their complementary roles in providing a more robust and comprehensive framework for policy evaluation.

The main objective of this approach is to evaluate the strengths and limitations of each method to answering a natural question emerging from the growing academic interest in understanding the causal impact of public investment programs on regional economic outcomes; "What is the effect of EU funding on the regions' Gross Value Added?".

5 Empirical Data Analysis and Discussion

Reality is far from optimal and controlled. It is common for real data to contain complexities like unobserved confounders, measurement errors or correlations between covariates and treatment, and irregularities. Yet, it is precisely these challenges that make the use of real data critical for generating insights that have an actual meaning in complex economic realities of the systems, connecting analysis to relevant, reliable and actionable conclusions. For the purpose of this analysis, i will leverage a dataset that captures information from key economic variables, crucial for exploring the impact of EU's subsidies in EU regions. These regions' eligibility for EU funding programs that mostly target less-developed regions, determines whether they receive treatment or not. The eligibility is determined based on predefined criteria, like gross value added (GVA) being below a specific threshold. For this reason, in the dataset, a binary dummy variable D indicates whether a region received the EU subsidy ($D = 1$) or did not receive funding or were not eligible to receive it yet during the observation period ($D = 0$). Some regions being treated earlier than others, due to them being qualified sooner based on economic metrics, indicate staggered adoption of EU funding programs. The variable *treat_year* records the year that a region received the subsidy, while *rel_year* calculates the number of years until and since the treatment for each region, for dynamic analysis of effects.

To identify initial patterns, understand baseline differences and pre-treatment trends in the dataset and generate hypotheses that will guide the next steps in the analysis, I perform an Exploratory Data Analysis to describes the data. The data is a panel with 8,456 observations, for 228 NUTS2 regions, within 26 countries (see Table 11, Appendix), over time. As illustrated in Figure 1, out of these regions, 96 NUTS2 regions received the treatment.

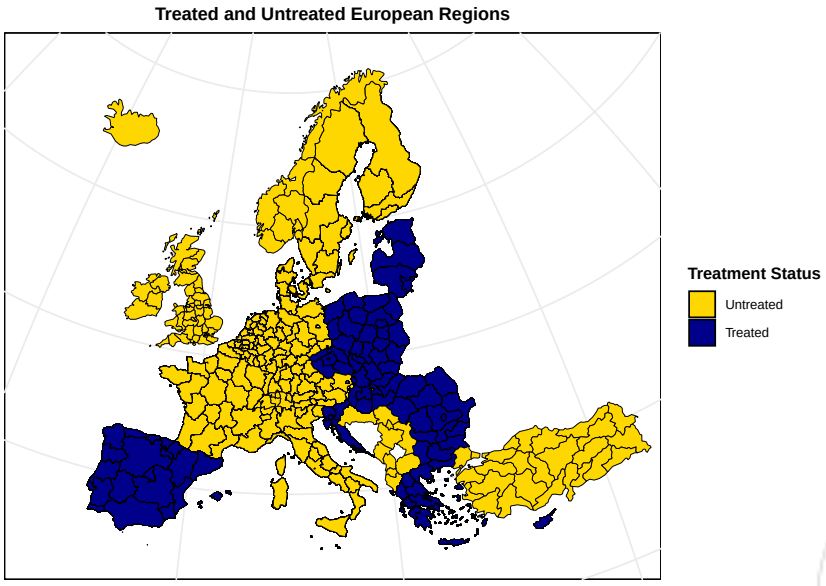

Figure 1: European Regions by Treatment Status

The inclusion of NUTS2 regions in Europe, representing subnational administrative units, is ideal for a detailed analysis while preserving comparability across countries, as EU subsidies are often implemented and monitored at this level of granularity. Spanning from 1980 to 2022, the dataset includes periods such as the biggest enlargement of the EU from 2004-2007, the 2008 financial crisis and the recovery post-COVID-19 period. EU funding targets infrastructure, productivity and investment making GVA a great tool for measuring the regional economic performance, due to its ability to provide insightful information about the standard of living and the productivity within EU regions. By measuring the value of goods and services produced, accounting for the value of intermediate consumption, GVA is particularly suitable for a target variable in forecasting models. For instance, in 2013 Lehmann and Wohlrabe employed an autoregressive distributed lag model aiming to forecast the total and sectoral GVA in the German economy, highlighting its predictive capabilities. Therefore, representing the inflation-adjusted output of a region's economy per individual, the outcome variable Y is decided to be the real GVA per capita. Measuring changes in GVA will help to evaluate whether these subsidies are achieving the economic impact that they were designed to achieve.

Labour appears to be one of the main inputs influencing economic output, productivity and growth alongside capital, according to economic theory. For this reason, employment rate, reflecting the share of employed people relative to the total population contributing to the economy, capital stock per capita representing individual's average value of physical assets and relative gross fixed capital formation measuring economy's capital investment (buildings, machinery and infrastructure), will serve as the predictors X in this analysis. Including these predictors will also help to control for pre-existing differences in regions, such as infrastructure levels that pre-existed and could influence both treatment assignment and economic outcomes, better governed regions that could secure more funding leading to higher growth or economic shocks that may affect treated and untreated regions differently.

Due to relationships in economics being frequently multiplicative rather than additive, the variables will be log-transformed. The log-transformed variables will reduce skewness of variables with long tails, normalize their distributions, ensure homoscedasticity by stabilizing the variance and compresses the scale of outliers. This transformation aligns better with economic theory because the variables will better capture nonlinear relationships, while assume proportional relationships between variables.

To understand economic conditions of regions and the challenges in fostering equitable growth, Table 1 offers critical insights into the distribution, variability and patterns in the data. The overall average log GVA per capita is 9.82, with a relatively low standard deviation of 0.61, indicates moderate variability in economic performance across all regions. The distribution is slight left-skewed (-0.81) that suggests that high-performing regions are fewer relative to low-performing ones. Delving further into the regions, treated regions appear to have a lower average GVA (9.28) than the untreated ones (10.02), aligning well with the idea that EU subsidies are designed for regions that are economically weaker, with lower GVA. The slight smaller standard deviation (0.53 for treated vs. 0.51 for not treated) could indicate a homogeneity in GVA of the untreated, possibly suggesting more common characteristics between those more-developed regions. The high kurtosis (3.65) for untreated regions indicates more extreme outcomes than the treated regions. Moving to the overall average log employment rate -0.85 suggests that a large proportion of regions have low employment rates relative to their population, while the slight positive skew (0.067) reflects the few regions that have relatively high employment rates. The slightly lower average employment rates (-0.87) of the treated regions compared to the untreated regions' rates (-0.83) could indicate that lower employment rates are one of the eligibility criteria for receiving EU subsidies. The overall capital stock per capita shows an asymmetric negative skew (-1.30), that highlights that many regions have low levels of physical capital. Confirming the targeting of EU subsidies toward less developed regions with insufficient infrastructure, treated regions appear to have lower average capital stock (10.29) compared to the untreated ones (11.22). Untreated regions exhibit high kurtosis (8.14) signaling the presence of outliers, likely because of significant industrial or financial hubs. Finally the overall negative skew of gross fixed capital formation (-0.33) in[illegible]dicates that only a small number of regions have relatively higher investment levels in infrastructure a[illegible]nd physical assets. Treated regions are falling behind in investment in buildings, machinery and infr[illegible]astructure, because of the lower average gross fixed capital formation (8.08) compared to the untrea[illegible]ted regions (8.[illegible]). The slight positive skew of the treated regions (0.16) suggest that despite the low l[illegible]evels in infrastructure there are some treated regions with significant investments. These results clearly [illegible]indicate that employment rates and capital stock on average for treated regions are lower com[illegible]pared to untreated regions', which makes sense because by design these regions, have lower economi[illegible]c performance and

investment levels, highlighting their eligibility for funding. In all variables, untreated regions exhibit higher kurtosis in general, that highlights the existing outliers because of the inclusion of both highly developed regions and those just above the funding eligibility threshold. The Jarque–Bera (JB) that measures deviations from normality in the distribution of each variable indicates the presence of strong outliers, especially in the untreated regions underscoring the need to account for non-normality and outliers in the dataset.

Table 1: Summary Statistics for Key Variables by Treatment Status

Variable	Group	Mean	Median	SD	Min	Max	Skew	Kurt	JB
log_rgva_pc	Overall	9.8185	9.9593	0.6105	7.6802	11.4334	-0.8080	3.3638	966.5421
	Treated (D=1)	9.2815	9.3232	0.5257	7.7203	10.5496	-0.2446	2.4995	46.8662
	Not Treated (D=0)	10.0187	10.0899	0.5105	7.6802	11.4334	-1.3514	3.6550	4269.3312
log_emp_pop	Overall	-0.8456	-0.8456	0.1742	-1.4206	-0.2643	0.0670	3.0154	6.4141
	Treated (D=1)	-0.8709	-0.8796	0.1605	-1.3008	-0.2701	0.3660	2.7791	167.9257
	Not Treated (D=0)	-0.8362	-0.8309	0.1781	-1.4206	-0.2643	-0.0427	2.7905	13.1366
log_kstock_pc	Overall	10.9742	11.2014	0.8025	7.3071	12.3624	-1.3027	4.3868	3069.2920
	Treated (D=1)	10.2938	10.4474	0.8453	7.3071	11.9101	-0.5192	2.5604	121.6068
	Not Treated (D=0)	11.2279	11.3333	0.6172	8.1069	12.3624	-1.9253	8.1411	10589.4755
log_rgfcf	Overall	8.6524	8.7381	1.0743	4.1542	12.0934	-0.3316	3.2207	172.1391
	Treated (D=1)	8.0815	8.0953	0.9139	5.4653	10.7549	0.1683	2.3500	16.8155
	Not Treated (D=0)	8.8658	8.9901	1.0504	4.1542	12.0934	-0.6207	3.9393	622.0192

To explore further the relationships between the variables, Figure 2 offers a visual depiction of them, grouped by treatment status. The upper panels display the correlations between the variables, the diagonal panels illustrate the density plots for individual variables and the lower panels display the scatterplots providing initial insights about the relationship between two variables.

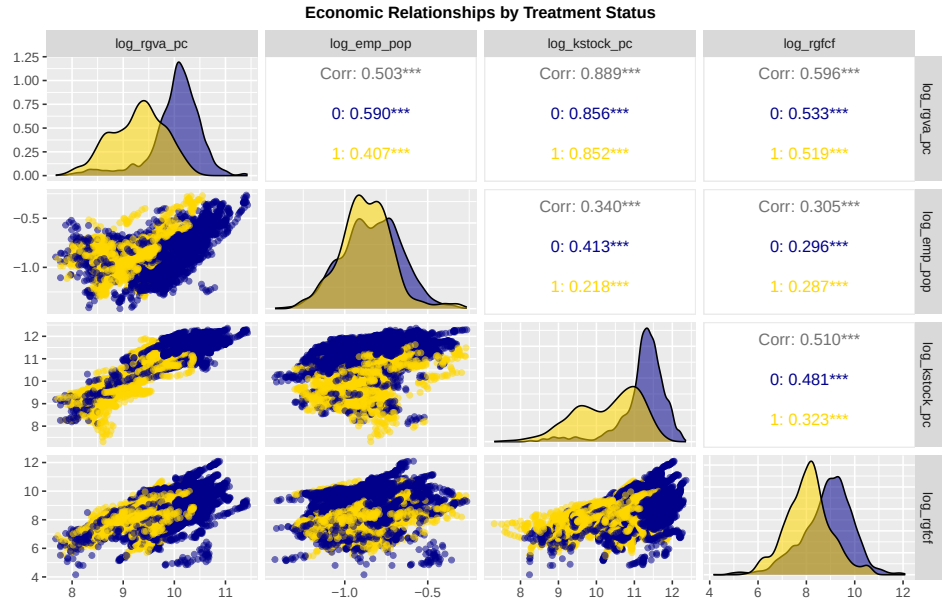

Figure 2: Relationships Between Key Economic Variables by Treatment Status

Starting from the upper panels of Figure 2, the correlation values suggest that employment rates, capital stock and gross fixed capital formation are critical drivers of economic pe rformance. Digging into the groups' results a strong positive association between GVA per capita and employment rates (0.59), capital stock per capita (0.86) and gross fixed capital formation (0.53) is sugg ested for untreated regions. In more well-functioning, untreated regions, an increase in employment rates, capital stock and investments levels will lead to GVA rising, possibly because these variables ar e tightly connected to economic performance. For treated regions, while correlations are also positiv e between GVA per capita and the three variables, the slightly lower correlations (0.4 for GV A and employment rates, 0.85 for GVA and capital stock, 0.52 for GVA and gross fixed capital form ation) may reveal

economic, structural and temporal complexities introduced by the treatment process. The lower correlations can be caused from the underdevelopment in treated regions, possible distortions of the natural economic relationships may be introduced by the subsidies, time lag in the actual treatment effects, completely different economic structures (reliance in low-productivity or high-productivity sectors) between treated and untreated regions, heterogeneity in the treatment impact and spillover effects being "exported" from the treated regions to neighboring untreated ones. Moving to the diagonal panels, the first density plot, illustrating the distribution of GVA for both treated and untreated regions, shows that untreated regions have a higher mean and a narrower distribution compared to the treated regions, that appears more spread out. This happens possibly because the untreated group includes more developed, with higher GVA levels, while the treated group includes regions with heterogeneous responses to treatment programs or underlying regional disparities. The distributions of capital stock, confirms that narrative as untreated regions, having a higher mean and a narrower and taller distribution, represent more capital-intensive economies, whereas the flatter and wider distribution with a long tail of the treated regions indicate that regions respond differently to subsidies. Interestingly, the distributions of gross fixed capital formation overlap significantly, while the untreated regions have a slightly higher mean, due to stable and mature economic systems supporting consistent investment levels. In this plot the treated regions seem to be distributed normally, suggesting that the funding led to more uniform investment behaviors, while the left skewed distribution of the untreated regions indicate a mix of developed and a few regions struggling with investment levels. The difference in the treated group's distributions in the capital stock and gross fixed capital formation plots is clear and is due to capital stock being a long-term measure, reflecting the cumulative impact of investments over time, while gross fixed capital formation, capturing the current investment flows, is a short-term measure. Due to treated regions being less developed initially, they could have started with lower baseline levels of capital stock, but funding could have allowed them to catch up in terms of gross fixed capital formation. The similarity of the treated and untreated groups' distributions for gross fixed capital formation could have happened because treated regions might prioritize current investments to address immediate development needs, while the untreated, more developed regions focus on more stable investment patterns. Additionally this convergence, in short-term investment flows, could be caused by treated regions' spillover effects to nearby untreated regions, while still exhibiting differences in long-term capital accumulation. Finally the lower panel scatter plots are used to observe the bivariate relationships between the variables for both groups. The scatter plot of GVA with employment rates validates the finding on stronger correlation in untreated regions with the clustering implying that employment has a more consistent effect on economic output for those regions. In the scatter plot for GVA and capital stock, while the strong positive relationship between those variables is validated in both groups, treated regions appear more spread out, especially in lower levels of capital stock and untreated regions display more linear and tighter clustering indicating more well-functioning markets. Finally in the scatter plot depicting the relationship between GVA and gross fixed capital formation, stability and effectiveness in investment for untreated regions is reflected from the tighter clustering, while for treated regions it shows more variability possibly due to inefficiencies in translating investments into productivity.

Convergence vs. divergence is a constant battle in EU. The goal of European countries with historically low GVA per capita is to narrow the gap with higher-income economies. Divergence occurs when these countries fail to catch up with the higher-income ones. Figure 3 illustrates countries' deviations from the median GVA, showing how they perform economically for each year relative to the other countries. Economic outliers appear persistently above or below the median along the years. Luxembourg as a high-income economy, consistently outperforms the median, while Bulgaria and Romania as lower-income economies show negative deviations for many years. In general, Eastern European countries like Bulgaria, Romania, Hungary, Poland, Slovakia and Lithuania exhibit delayed economic convergence with Western Europe as their large negative deviations indic ate, while economic stable countries like Sweden, Germany, Belgium and France are closer to the m edian. Taking consideration that Eastern European countries were centrally planned economies i the Eastern Bl with weaker infrastructure, lower levels of investment and limited access to global mark s compared to Western Europe, this delayed convergence seems reasonable. The 2008 fina cial crisis affe deviations for European countries with weaker economies like Greece that saw wi ing negative deviations due to its structural weaknesses, while stronger economies like Luxembou rg saw widened the positive deviation because of its robust financial system. While Poland or Esto ia s ow signs of

convergence, the historical persistent economic divide between Western and Eastern Europe remains visible.

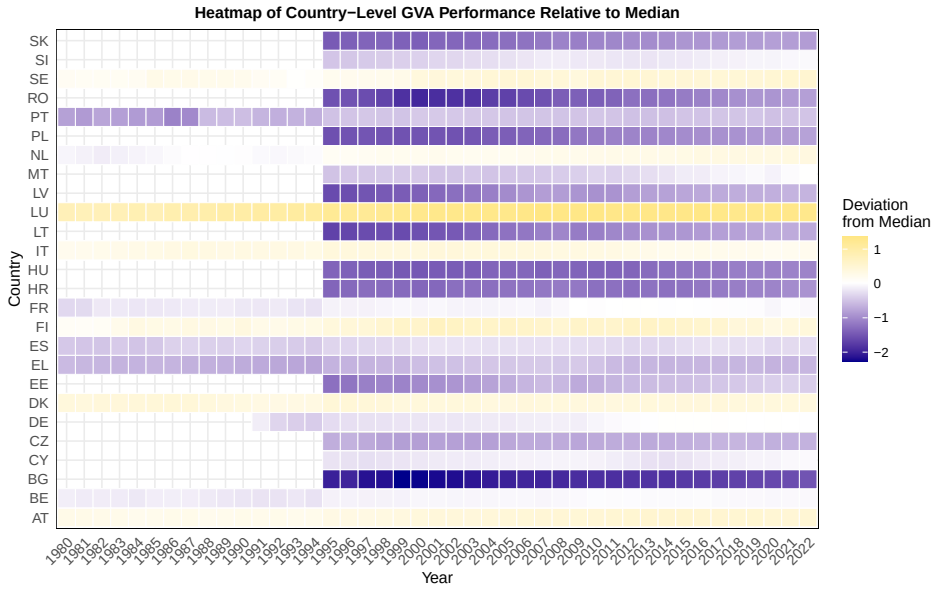

Figure 3: Historical Deviations from the GVA-Median per Country (1980–2022)

Diving into the regional analysis, Figure 4 illustrates the dynamic evolution of GVA in regional economies across six time intervals post-treatment. The goal is dual; to visually examine the potential for long-term convergence and observe possible persistent barriers faced by underdeveloped NUTS2 regions. The first plot marks log GVA at the year of the treatment for each of the treated regions. Countries like Poland, Romania, Slovakia, Bulgaria, etc., appear to have lower GVA values that Spain, Greece or Croatia. After this year, regions exhibit gradual upward movement in GVA, indicating partial economic convergence possibly as a result of the funding. If this is the case, then the graph clearly highlights heterogeneity in the treatment's impact because the rate of improvement appears uneven or stagnant. Examining the last two plots of the graph also underscores that impact accumulates and becomes more visible over the long term because structural changes in infrastructure or productivity take time to materialize and affect the economies in real time. Additionally, the fact that high-performing regions, like regions in Spain, in the treatment year maintain their relative advantage, suggests that pre-existing economic conditions play a central role in post-treatment trajectories. These plots effectively visualize the initial idea of the need for precise policy approaches and interventions, because an "one-size-fits-all" approach is not applicable in this case.

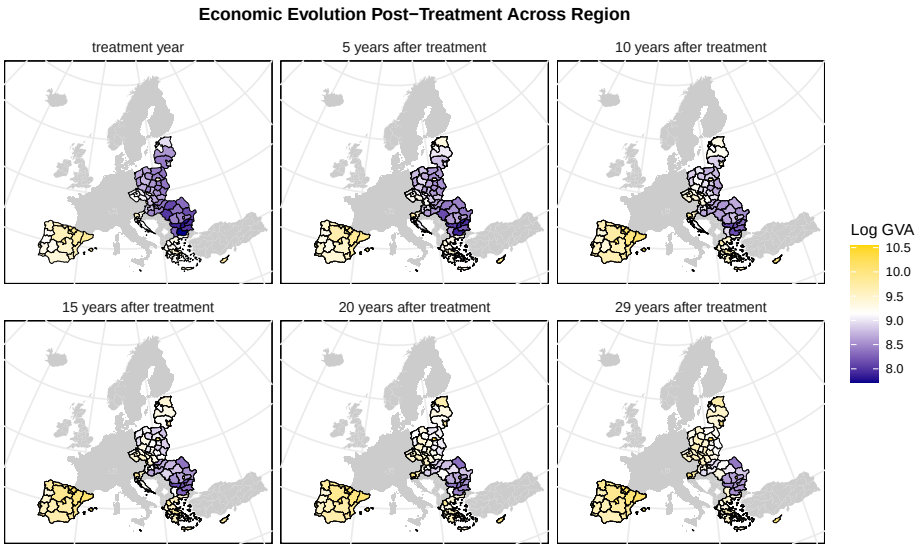

Figure 4: Log GVA Per Capita Across Six Time Intervals

5.1 GSC application

Traditional causal inference methods like DiD rely on the assumption of parallel pre-treatment trends between treated and control regions. However, in this specific dataset, due to the absence of valid *rel_year* in control regions, as they are never treated, preventing a pre-treatment trend analysis and because treated regions adopt treatment at varying times, further complicating trend alignment, I will employ the GSC method to allow for non-parallel pre-treatment trends and staggered treatment adoption. By constructing synthetic controls for each treated region and flexibly weight control regions and covariates I anticipate robust counterfactual estimates in such cases that traditional methods fail to align regions' timelines.

By applying GSC in the panel data with repeated observations for treated and control regions over time, I will attempt to discover the impact of the intervention on log GVA per capita by estimating the counterfactual trajectory for treated regions using the untreated donor pool, while highlighting possible limitations of the method for this specific dataset. The key components for this method are the *log_rgva_pc*, serving as the outcome variable, the treatment *D_cf* indicator and the covariates employment rate, capital stock per capita and gross fixed capital formation. Moreover, including unit fixed effects, accounts for unit-specific heterogeneity and time fixed effects accounts for time shocks common across units. Cross-validations ensures that the model has selected the best number of latent factors (r) within a specific range, avoiding overfitting and underfitting, while to also provide confidence intervals for the estimated treatment effect the standard errors are being bootstrapped. To evaluate this method, I will compare the observed and synthetic counterfactual outcomes, calculate the ATT that represents the causal effect of treatment:

$$\text{ATT} = \frac{1}{N_T} \sum_{i \in \text{Treated}} \left[Y_i^{\text{Observed}} - \hat{Y}_i^{\text{Counterfactual}} \right],$$

where N_T is the number of treated units, Y_i^{Observed} is the observed outcome for treated unit i after the treatment and $\hat{Y}_i^{\text{Counterfactual}}$ is the counterfactual outcome for treated unit i after the treatment, estimated using the pre-treatment data and control units, and the Mean Squared Prediction Error (MSPE):

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

to quantify how well the synthetic counterfactual replicates the observed outcomes in the pre-treatment period, conduct Sensitivity Analysis and Placebo Tests. It is crucial to ensure that the GSC treatment synthetic counterfactuals are reliable. This provides a solid foundation for calculating ATT robustness.

by sensitivity analysis to test whether the ATT is actually robust to model assumptions. Afterwards placebo tests will assess whether ATT is not spurious.

5.2 GSC Results

The first step of the model evaluation is to assess the goodness-of-fit, ensuring that the chosen model can reliably estimate counterfactuals for treated units. To validate the model with a rank range from 0 to 5 and seven minimum pre-treatment periods, I will plot the observed vs. predicted outcomes for treated regions. Rank (r) determines the range of numbers for latent factors or dimensions that are being selected for the matrix factorization, to approximate the unobserved heterogeneity in the data, while the cross-validated rank $r*$ identifies the optimal number of factors based on a balance between model complexity and predictive accuracy. As illustrated in Figure 5, in the pre-treatment period the synthetic counterfactual closely tracks the observed treated average, demonstrating that before the treatment, the model has effectively captured the underlying trends of the treated regions. This pre-treatment alignment implies that the covariates and latent factors are effective predictors of *log_rgva_pc*. In the post-treatment period the observed treated average and the synthetic counterfactual appear to diverge. The upward shift of the treated average compared to the estimated one, indicates a positive and persistent ATT over time, highlighting the lasting impact of the the treatment on the *log_rgva_pc*. Moreover it appears to exist some degree of variability in the observed outcomes of individual treated regions in the plot. These variations among those regions indicate the need for deeper exploration for potential heterogeneity in treatment effects. For this reason, Causal Forests could be used to help to investigate whether specific subgroups of treated regions experience different effects systematically. Another interesting observation is that the divergence between the observed and counterfactual averages after the treatment initially grows, reaches a peak and stabilizes afterwards, indicating that treatment effect might take time to occur but eventually it stabilizes in the long term.

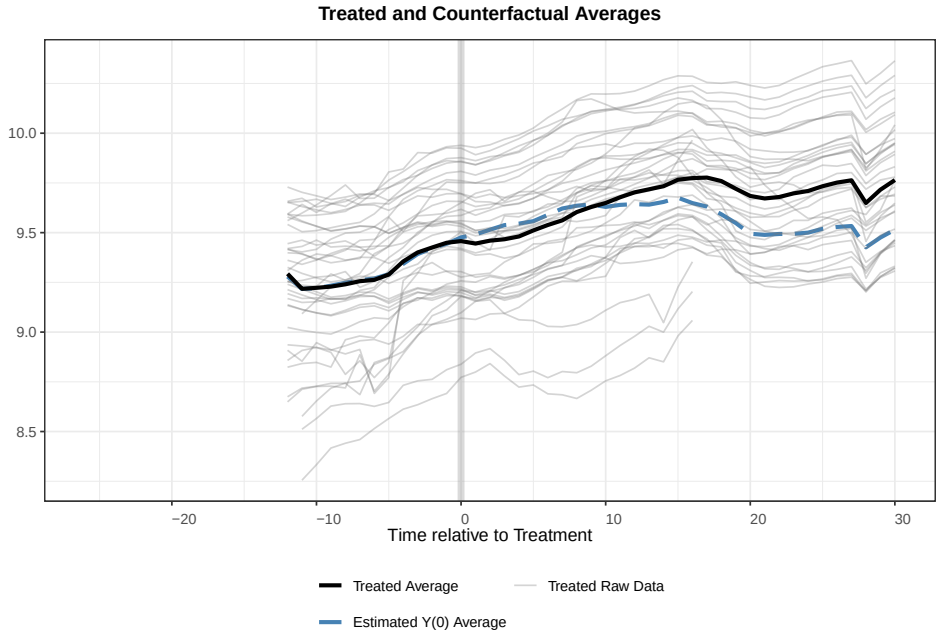

Figure 5: Treated and Counterfactual Averages Over Time Relative to T reatment

For further assessment of the pre-treatment fit, I proceed to an examinatio n of how well predicted values align with observed values before the treatment. Figure 6 provide s a visual for the residuals of the model over time for the pre-treatment period. They are mainly c ntered around zero, indicating a model that effectively captures the trends in *log_rgva_pc* before the treatment. the synthetic counterfactuals seem to replicate reasonably well the observed outcomes. Additionally the lack of systematic patterns implies that unobserved factors do not introduce bias in the treatment predictions. Despite the presence of few residuals exceeding 0.1 and -0.1 in earlier time periods, there

is no evidence of heteroscedasticity over time, suggesting that the model does not require further modifications to account for variance heterogeneity. As a consequence, the reliability of the synthetic counterfactuals and therefore the good pre-treatment fit are validated from the overall pattern of the residuals. This indicates that the model adequately captures the dynamics of the treated regions.

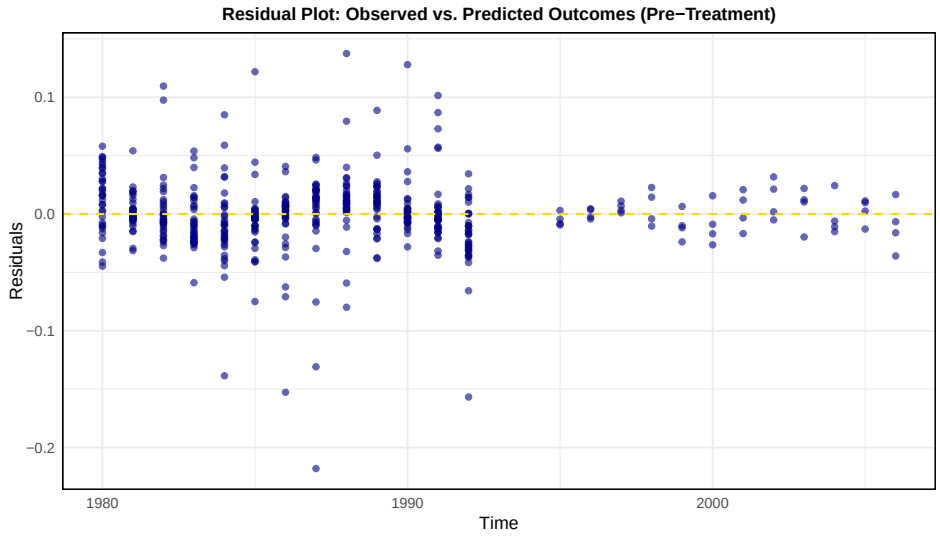

Figure 6: Residual Plot of Observed vs. Predicted Outcomes in the Pre-Treatment Period

While this strong pre-treatment fit validates the construction of synthetic counterfactuals for a single model, I wonder whether these results hold under alternative plausible model setups. For this reason, sensitivity analysis is used to ensure that the results from the model are robust to modeling assumptions when the range of latent factors is changed. This way allows to check the sensitivity of ATT when complexity of the model is changing. The results for different GSC models are displayed in Table 2, and were consistent across 100 runs to reinforce confidence in the robustness of these findings.

Table 2: Model Comparison for Varying Rank Range and Pre-Treatment Periods

Model #	Rank Range (r)	Min. Pre-Treatment Periods ($min.T0$)	Cross-Validated Rank (r^*)	MSPE	ATT Avg (Estimate)	P-Value
1	0–2	2	1	0.0049	0.1317	0.009
2	0–3	5	1	0.0049	0.1317	0.009
3	0–5	7	2	0.0018	0.0908	0.210
4	0–3	7	2	0.0018	0.0908	0.213

The thought process for comparing different models is that different rank range (r) and minimum long enough pre-treatment periods (min.T0) to model counterfactual trends will offer different model results, adapting differently to different situations. Selecting an appropriate rank range is critical because lower ranks are easier in terms of interpretabilty but they could miss some of the underlying complexity, while higher ranks can capture more nuanced patterns in the data, sacrificing some interpretability. By requiring a minimum number of pre-treatment periods for treated units the min.T0 parameter ensures robust estimation. Setting this parameter too low could lead to underfitting because of possible inclusion of insufficient pre-treatment data, while a high min.T0 could exclude valuable treated units.

For the first model, with a rank range from 0 to 2 and two minimum pre-treatment periods required, the average ATT estimate is positive (0.1317), having a statistically significant p[illegible]sitive effect on log GVA per capita (p-value = 0.009 < 0.05), meaning that the treatment is associa[illegible]ted with a[illegible] increase in the log GVA per capita. The first plot of the Figure 7 illustrates t[illegible]his estimated [illegible] before and after the treatment. In pre-treatment periods, with p-values higher than [illegible] close to zero and not statistically significant, suggesting no strong evidence of a[illegible]nticipatory effe[illegible] This means that there is no strong evidence that log GVA's path has started to change before the treatment implementation (dashed vertical line at 0), indicating that the parallel trends assumption might hold. Anticipatory effects can make it unclear whether observed post-treatment effects are due to the treatment itself or the pre-treatment adjustments made in anticipation of i[illegible]. After treatment,

ATT becomes increasingly significant and positive, before experiencing a decline at period 23. This gradual improvement over time underscores the delayed benefits of the treatment, but the widening confidence intervals indicate either great variability or limited data for those periods.

Moving forward to the second model, with a rank range from 0 to 3 and five minimum pre-treatment periods required, similarly the average ATT estimate is positive (0.1317), with a statistically significant effect on log GVA per capita (p-value = $0.009 < 0.05$), suggesting that the treatment is similarly to the first model associated with a 13.17% increase in log GVA per capita. The second plot in Figure 7 illustrates this estimated ATT. In the pre-treatment period, ATT remains close to zero, with p-values higher than 0.05, indicating no statistically significant effects prior to treatment, providing further evidence that the parallel trends assumption is reasonable for this specification (no indication of anticipatory effects) and the treatment itself, rather than pre-treatment adjustments, is likely the primary driver of the observed post-treatment changes. After the treatment, ATT becomes positive and statistically significant, indicating a sustained improvement in log GVA per capita over time. Interestingly, compared to the first model, the confidence intervals in the long term become slightly narrower, suggesting that the increased number of pre-treatment periods from 2 to 5, contributes to more precise estimates. Despite this, there is a similar decline in ATT is observed around period 23, accompanied by widening confidence intervals, probably underscoring some variability or limited observations in the long term.

Observing the third and default GSC model, with a rank range from 0 to 5 and seven minimum pre-treatment periods required, the average ATT estimate is lower (0.0898) and not statistically significant (p-value = $0.210 > 0.05$). For this model, while the treatment is associated with an increase in log GVA per capita, the evidence appears to be weaker than in the first two models. In Figure 7 plot, in the pre-treatment period, ATT remains close to zero reasserting the absence of anticipatory effects. As evidenced by the lower MSPE ($0.0018 < 0.0049$), the longer pre-treatment period (from 5 to 7) achieves to improve the reliability of the synthetic counterfactual, strengthening the argument that the parallel trends assumption is satisfied. After the treatment, ATT decreases and exhibit negative values for the first 9 periods. This decline could happen because many interventions have lagged effects (reallocating resources or learning curves) or because treated regions may be more vulnerable to external shocks leading to synthetic counterfactual overestimating what would have happened in the absence of treatment. Additionally in case of significant heterogeneity, the ATT estimates for early post-treatment periods might be less stable, resulting to those negative values. Afterwards, the ATT gradually increases and while it remains positive and stabilizes over time, it is not statistically significant. This could happen because over time, treatment benefits accumulate and outweigh initial adjustment costs or due to the model adjusting to new post-treatment equilibrium and synthetic counterfactual aligns better with the observed outcomes as illustrated in Figure 7. For this model the average ATT is lower and more conservative than the first models ($0.0908 < 0.1317$), likely due to the inclusion of a longer pre-treatment period. Confidence intervals appear to widen considerably in the long term, probably indicating increased variation or limited precision in those long term estimates.

Finally for the fourth and last model, with a rank range from 0 to 3 and seven minimum pre-treatment periods required, the average ATT estimate is also 0.0898 and not statistically significant (p-value = $0.213 > 0.05$). In the pre-treatment period, ATT exhibits no significant deviations from 0, supporting the validity of the parallel trends assumption. The similarly low MSPE (0.0018) suggests a reliable synthetic counterfactual, while compared to the previous model, the lower rank range introduces less complexity with a similar level of robustness. Post-treatment, ATT exhibits a similar pattern over time with the previous model, as it remains negative for the first 9 periods, becomes positive but insignificant and stabilizes in the long term. The confidence intervals are slightly narrower in the early post-treatment periods, indicating improved precision, probably because of the reduced model complexity. Despite this early trend, the intervals become wider in later periods, highlighting greater variability or limited data.

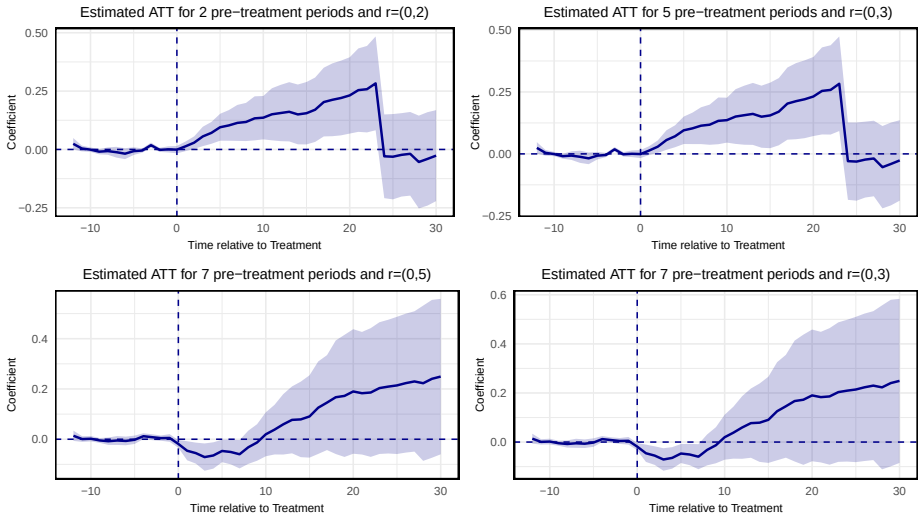

Figure 7: GSC models: Average Treatment effect on the Treated by Pre-treatment Periods and Rank Ranges

Therefore, the sensitivity analysis results highlight that for the first two models the confidence intervals suggest decent precision combined with greater uncertainty in later periods, whereas the longer pre-treatment periods of the last two models provide more conservative and robust estimates, even though confidence intervals become wider especially in later periods. However the consistent positive ATT across all four models indicates that the treatment effect is robust to variations in pre-treatment periods and rank ranges. Models 1 and 2, with $r* = 1$, represent simpler structures, while models 3 and 4, with $r* = 2$, capture additional variability, enabling the assessment of whether increasing model complexity enhances predictive accuracy without overfitting, as shown by their respective MSPE values. Notably, comparing model's 3 and 4 MSPE values (0.0018) to models' 1 and 2 MSPE results (0.0049) suggests that more pre-treatment periods or more expanded rank range improve the model's fit. The considerable improvement in MSPE when increasing $min.T0$ from 5 to 7 highlights not only the is sensitivity of the MSPE to the pre-treatment period length compared to small changes in the rank range, but also the importance of a sufficient number of pre-treatment periods for more robust inference. Hence, model 3 is preferred due to obtaining the same high level of predictive accuracy as Model 4 (MSPE = 0.0018) and presenting greater flexibility for capturing complex data dynamics with a wider rank range (0-5 > 0-3).

Synthetic control model's central assumption is that the synthetic counterfactual created for treated regions mimics their unobserved outcomes in the absence of treatment. Failing to meet this assumption can lead to spurious treatment effects. For determining whether the observed ATT reflects the true causal impact of the treatment or whether it occurred due to random noise, overfitting, violations of assumptions or unobserved confounders, placebo tests with "fake" treatments are applied. The process of creating placebo datasets is repeated 100 times to reduce the impact of random variability in placebo treatment assignments, giving more stable and reliable results. If placebo tests result in significant ATT estimates, there is evidence that the true ATT is likely to be spurious, casting doubts on the validity of the estimated causal effect. Since the placebo-treated regions were not really treated, the model should estimate little to no treatment effect for them. Therefore, ideally, the ATT averages are expected to be centered around zero. In Figure 8, I illustrate the magnitude of pla ebo ATT averages as well as their comparison to the real ATT, as a means of validating the GSC mo del's reliability the strength of the causal conclusions. As it is expected due to the randomly assig ed treatments, the generated ATT distribution in Figure 8 is centered around zero with a narrow sp read for the placebo results across placebo tests. The robustness of the placebo test is also supporte d from the lack of extreme outliers. The real ATT (0.0908) lies far to the right of the placebo distri bution, indicating that the treatment effect that is observed in the real data is remarkably differen from the placebo effects and is unlikely to be due to chance, validating that the observed ATT is not spurious and it likely reflects an actual treatment effect.

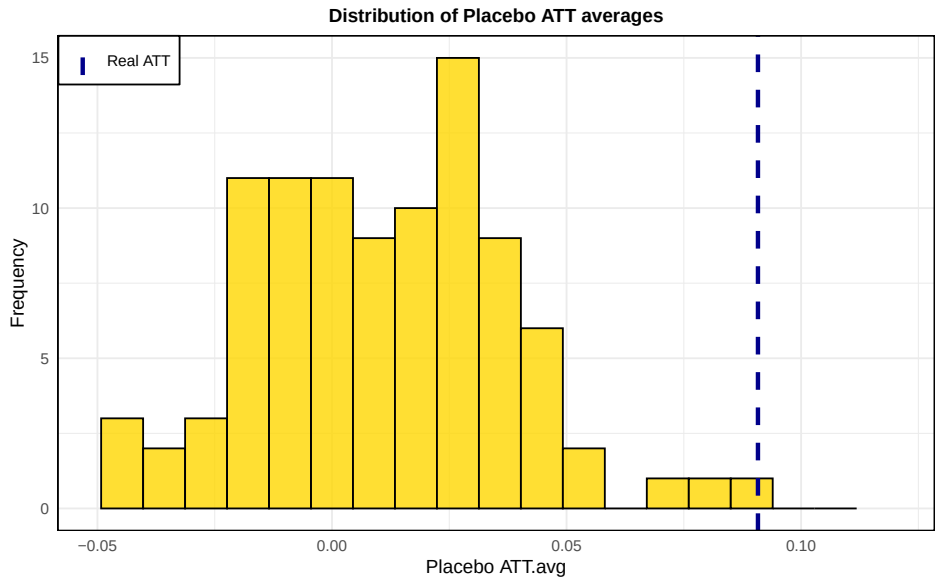

Figure 8: Distribution of Average Treatment effect on the Treated averages for all placebo-treated regions with Real Average Treatment effect on the Treated

Additionally, the comparison of the MSPE in Figure 9 of 100 placebo iterations with the real MSPE will be used to evaluate model's robustness and the credibility of the ATT. As a general rule, for the model to capture the treatment effect accurately and not overfit to noise, the real MSPE is expected to be lower than the MSPE of most placebo-treated regions. Remarkably, Figure 9 shows that the real MSPE (0.0018), serving as the benchmark, lies slightly above the general distribution of placebo MSPEs that range from approximately 0.0009 to 0.0021. This wide range may reflect heterogeneity in the untreated regions originating from unobserved factors not sufficiently captured by the synthetic counterfactual. Although the real MSPE is not distinctly lower than the majority of placebo MSPEs, the absence of extreme deviations in most of the cases suggests that the GSC model has a reasonable level of fit in capturing the underlying data patterns.

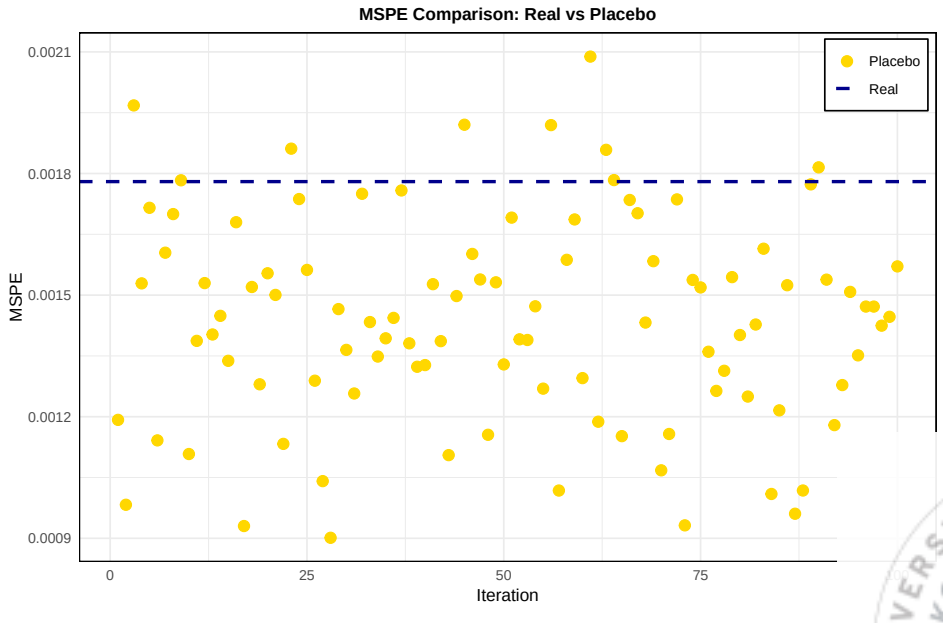

Figure 9: MSPE Values for Placebo Scenarios and Real MSPE

However, given that GSC models are not explicitly designed to account for potential heterogeneity in treatment effects, combined with the wide range of MSPE values in the placebo tests (0.0009 to 0.0021), that signals the presence of unobserved factors influencing outcomes, their reliance on adequate pre-treatment periods for constructing reliable counterfactuals, that limits generalizability, and their inability to capture potential nonlinear interactions between covariates and treatment effects, the incorporation of Causal Forests in the analysis could offer a more comprehensive understanding of the treatment effects.

5.3 Causal Forests application

Starting from the assumption that treatment effects are not uniform but differ based on observed covariates, the objective of this application is to identify whether treatment effects actually vary across subgroups and answer to the question "Does the impact of the EU funding differ by region?". The dataset for this analysis has already received meticulous preparation as part of the GSC modelling process and examination. Log-transformed real GVA per capita serves as the outcome variable, log-transformed employment rate, log-transformed capital stock per capita and log-transformed gross fixed capital formation as the covariates while the binary variable D_cf reflects whether a region received funding or remained untreated. To achieve both robust model evaluation, by testing on unseen data and reliable treatment effect estimation, I will apply a stratified splitting by treatment status. The rationale behind this is to ensure that any imbalance in treatment assignment does not skew results in either the training or testing set. Moreover similar treatment proportions can improve the ability of the causal forest model to generalize better to new data and confirm that treatment effect estimates' heterogeneity is not biased by differences in treatment-control ratios between splits. Consequently I will partition the dataset into two subsets ($D_cf = 1$ and $D_cf = 0$) and within each of these subsets, I will randomly assign observations to training and testing sets based on a 80/20 splitting rule (80% training and 20% testing). As a final step, in order to create the final datasets, I will merge the stratified training and testing splits. As depicted in Table 3, the original dataset's proportions are 72.8% for treated and 27.2% for control group and in both training and testing datasets, these proportions are successfully maintained, reflecting that the stratified splitting has preserved the treatment-control balance.

Table 3: Proportions of Treatment and Control Groups in Original, Training and Testing Datasets

Dataset	D_cf = 0 (Control)	D_cf = 1 (Treated)	Total Observations
Original	0.728	0.272	8456
Training	0.728	0.272	6765
Testing	0.729	0.271	1691

Furthermore, to avoid overfitting or underfitting when capturing the heterogeneity in treatment effects, it is crucial to properly tune the hyperparameters. The size of the forest is controlled by the number of trees. As the number increases, the stability improves (less variance) but computational time also increases. The minimum number of observations per leaf is specified by the minimum node size. Small node sizes allow finer subgroup splits but may lead to overfitting. Maximum depth limits the depth that a tree is able to grow. As depth increases, more complex splits are allowed but there is a risk of overfitting. Finally, the sample fraction is the dataset's proportion that is used to grow each tree, ensuring adequate diversity among trees. To select the optimal values for hyperparameters that both minimize the error in treatment effect estimates and maximize generalizability, I will leverage cross-validation. After specifying a range for the number of trees (500, 1000, 1500, 2000, 4000), for the minimum number of observations in each leaf (5, 10, 20) and for the proporti on of data used for each tree (0.2, 0.3, 0.4, 0.5), I will separate the training set into 10 equal-sized f olds and use the folds to train the model and the remaining fold for validation. On the validation fold I will predict treatment effects and compute the MSE of the estimated treatment effects. Final y, after calculating the average MSE across all folds for every hyperparameter combination, the hyperp arameters with the lowest average MSE will be chosen. For these hyperparameter combinations, the r ange of the average MSE values is between 99.46778 and 99.71335. The lowest MSE is achieved with 10 00 number of trees, 10 observations in each leaf and 0.5 of data used for each tree.

Inference in Causal Forests is facilitated by the use of sample-splitting and a symptotic approxi-

mations. In reference to the previous discussion on Causal Forest application, the algorithm expands the Random Forest methodology (minimizing prediction error) to identifying splits that optimize the heterogeneity of the CATE within each node, by constructing a group of honest, randomized decision trees. For unbiased estimates of $\tau(X)$, honest splitting separates the dataset (Y, X, D) into two disjoint parts: one to determine the splitting structure of the trees and another to estimate treatment effects within the leaves. Causal Forest algorithm builds on the concept of recursive partitioning. Each tree in the forest is grown recursively by partitioning the data based on covariates X aiming to maximize the heterogeneity in treatment effects across the resulting child nodes and to minimize the within-leaf variance of $\hat{\tau}$ for improving the precision in the treatment effect estimation. Afterwards, for each leaf L of a tree, the treatment effect $\hat{\tau}_L$ is estimated as:

$$\hat{\tau}_L = \frac{1}{|L_{\text{treated}}|} \sum_{i \in L_{\text{treated}}} Y_i - \frac{1}{|L_{\text{control}}|} \sum_{i \in L_{\text{control}}} Y_i$$

where L_{treated} and L_{control} are the sets of treated and control units in the leaf, respectively. Eventually, once the forest is constructed, the final estimate of $\tau(X)$ for a new observation X is acquired by averaging $\hat{\tau}_L$ over all trees in the forest:

$$\hat{\tau}(X) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(X)$$

where B is the number of trees.

A core requirement for reliable inference is the estimation of the variance of treatment effect estimates. For CATEs, the infinitesimal jackknife (IJ), previously introduced, is used to estimate the variance and construct confidence intervals. Since causal forests rely on bootstrapped samples for tree construction, variance estimation captures the randomness introduced by bootstrapping and aggregates the variability across trees. The IJ successfully quantifies the uncertainty in treatment effect estimates, while ensuring that, even in high-dimensional settings, confidence intervals adequately reflect variability.

GSC is primarily designed to capture time-relative patterns and provide robust estimates of ATT across time periods. However, due to Causal Forest's inherent design for static treatments, to assess the effectiveness model, I will calculate ATE and CATE to analyse heterogeneity and visualize treatment effects by subgroups. Although Causal Forests are not ideal for modelling time-varying treatment effects since the temporal structure isn't explicitly modelled, I will aggregate individual CATEs for multiple time periods and multiple regions to calculate the ATT across time periods based on regions' individual characteristics.

For each time period, each ATT is calculated as:

$$\text{ATT} = \frac{1}{N} \sum_{i=1}^{N} \text{CATE}_i$$

where N is the number of treated units ($D_{cf} = 1$) and CATE_i is the Conditional Average Treatment Effect (predicted by the causal forest) for treated unit i in rel_year $= j$.

The ATE is the average of the treatment effects across the entire population (both treated and untreated units):

$$\text{ATE} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_i,$$

where N is the total number of units (treated + untreated) and $\hat{\tau}_i$ is the estimated treatment effect for individual i.

As formerly discussed, CATE measures the treatment effect conditional on a set of covariates and is defined as:

$$\text{CATE}(X = x) = \mathbb{E}[\tau \mid X = x],$$

where τ is the individual treatment effect.

5.4 Causal Forests Results

5.4.1 Evaluating Overlap Assumption

As previously mentioned, the overlap assumption for the covariate distributions of treated and control groups is critical in causal inference. The overlap is evaluated by fitting a logistic regression model for propensity scores, that capture the likelihood of receiving treatment based on observed covariates. For further improvement of the specification of the propensity score model while addressing potential non-linearities, a quadratic term for *log_kstock_pc_pre* and is included. To ensure that treatment effect estimation is performed within areas with adequate overlap, observations that were outside of a common support region (94.19%) were trimmed. Figure 10 reveals limited overlap between treated and control groups before and after trimming, as the control group is overrepresented in the lower propensity score range indicating that wealthier regions were systematically less likely to receive treatment, resulting in limited generalizability of the estimated treatment effects.

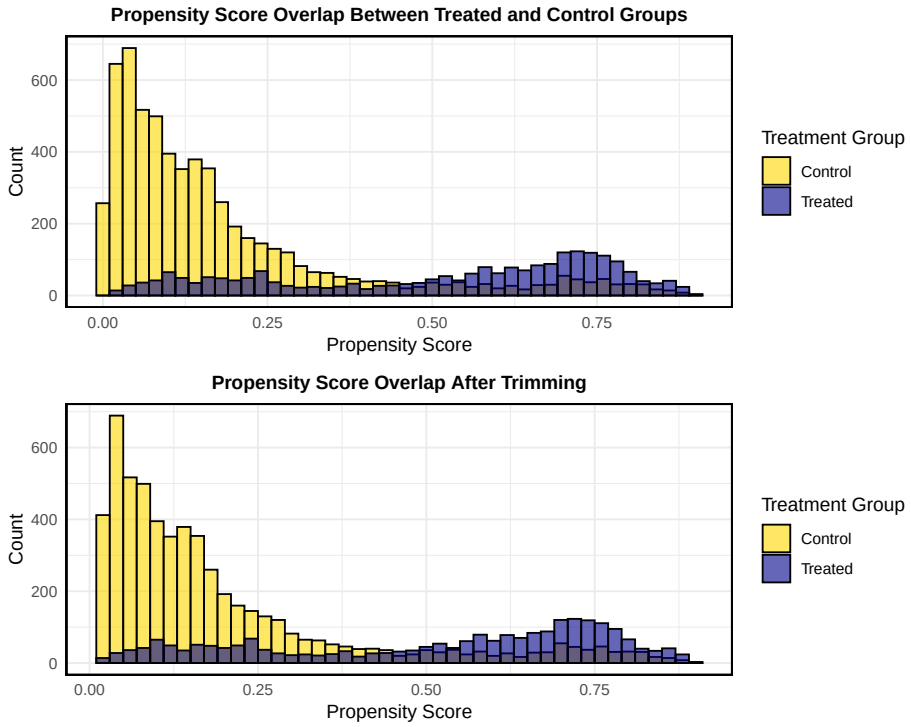

Figure 10: Propensity Scores Distributions Before and After Trimming

The model enabled the formation of matched groups with improved covariate balance. After this matching, covariate balance is used to ensure the comparability between treated and control groups. Standardized Mean Differences (SMDs) quantify whether treated and control groups have similar distributions for the covariates. Table 4 shows that the balance (SMD) is improved for most variables, with *log_emp_pop_pre* achieving balance 0.0698. However, moderate imbalances persist for *log_kstock_pc_pre* with balance -0.5105 and its quadratic term's balance -0.5188, suggesting that there are inherent structural differences between treated and control groups, that reflect the natural systematic under-representation of wealthier regions in treated group.

Table 4: Covariate Balance Before and After Matching

Covariate	Means (Treated)	Means (Control)	SMD (Before)	SMD (Afte
log_emp_pop_pre	-0.8709	-0.8821	-0.2165	0.0698
log_kstock_pc_pre	10.2938	10.7253	-1.1051	-0.5105
log_rgfcf_pre	8.0815	8.3234	-0.8533	-0.2634
I(log_kstock_pc_preˆ2)	106.6756	115.5215	-1.1595	-0.5188

However, due to the nature of the data a modification with aggressive trimming (

sity scores between 0.2 and 0.8) or employing additional covariates masks important heterogeneity in treatment effects, especially in cases where inherent differences are deep-rooted to the population. Therefore, my analysis proceeds, cautiously, with the application of the causal forest model to the entire dataset aiming to assess its ability to handle regions with imperfect overlap and inherent covariate imbalances. Validating this decision, Figure 11 reveals that CATE estimates increase gradually across the overlap groups, while exhibiting no erratic behaviour, no excessive variation or instability in low-overlap groups. Hence, this serves as an indication that the observed heterogeneity is relevant rather than an extrapolation bias' outcome. Moreover, as a final robustness check, in Figure 27 (Appendix) I assess whether CATE estimates manifest extreme variation in low and high-propensity score regions and demonstrate that treatment effects do not suffer from extreme extrapolation in the tails of the propensity score distribution.

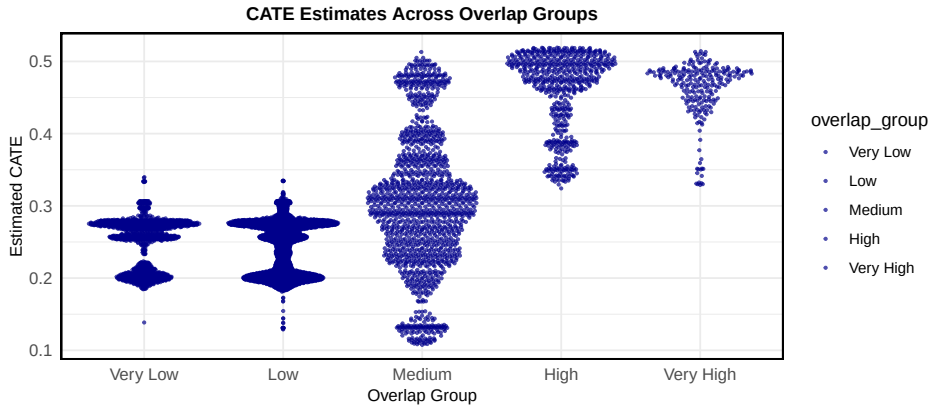

Figure 11: Conditional Average Treatment Effects Estimates Across Overlap Groups

In essence, instead of attempting to eliminate these differences, the model is employed to capture treatment effect heterogeneity within the observed data. This procedure aims on insightful knowledge into how effects vary across subpopulations with well-defined covariate profiles rather than making generalized causal claims. Although the findings have limited applicability to the full population as heterogeneity estimates may not be reliable for subgroups that are not adequately represented, the focal point of my approach predominantly lies on demonstrating the applicability of ML for causal inference in real-world datasets where methodological assumptions, such as perfect overlap, are rarely satisfied.

5.4.2 Training Causal Forests

Since the focus is on reliable and interpretable results about heterogeneity, I will leverage a parsimonious and simple model, using only pre-treatment covariates. First and foremost, the inclusion of post-treatment covariates could induce post-treatment bias to the model because those covariates are essentially affected by the treatment. Furthermore, in Causal Forests, pre-treatment covariates are guiding the partitioning process, ensuring that the model is well-calibrated and avoids overfitting. These covariates provide the baseline characteristics of the regions and without them the model lacks the necessary structure to successfully identify treatment heterogeneity. The selection of covariates in the pre-treatment period to train the Causal Forest model follows a structured econometric thought process that is likely to explain treatment effect heterogeneity while maintaining interpretability and robustness. After training the model, I will evaluate the Out-of-Bag (OOB) CATE estimates distribution for the training set, for an initial heterogeneity inspection, as it yields engaging insights about the treatment effects heterogeneity. The OOB CATE predictions are considered "honest" estimates that are derived during model training. The histogram in Figure 12 shows a right-skewed distribution of the estimated CATEs, while most of the observations are predominantly concentrated between 0.2 and 0.3. The model suggests that treatment effects are not uniform across regions which indicates that the majority of regions experience moderate treatment effects and only a smaller subset of "high responders" experience stronger effects. This variation emphasizes that the treatment impact depends on covariates and raises the profound question "Is the observed heterogeneity genuine?

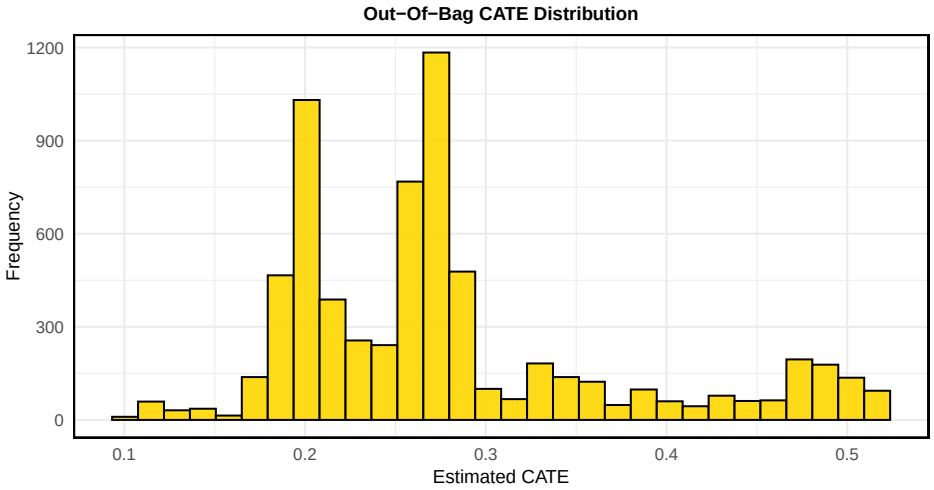

Figure 12: Out-of-Bag Conditional Average Treatment Effects for the training dataset

5.4.3 Quantifying Treatment Heterogeneity

Firstly, in order to evaluate the model, Best Linear Projection (BLP) is leveraged. As a linear regression of the CATEs on covariates, it estimates how well those covariates explain the variation in treatment effects. The goal is to identify drivers of treatment effect heterogeneity. Table 5, provides insights about the contributions of those covariates and the quality of the model's predictions. The $log\_emp\_pop\_pre$, $log\_kstock\_pc\_pre$ and log_rgfcf_pre variables are statistically significant (p-value $<$ 0.001) and stand out as important contributors as they significantly explain treatment effect heterogeneity, with estimates 0.2522, -0.1388 and 0.0392, respectively. These estimates indicate that higher employment levels experience stronger positive treatment effects, whereas the negative coefficient for $log\_kstock\_pc\_pre$ reflects the average decline in treatment effect for every 1-unit increase in the capital stock per capita. These results could reveal diminishing marginal impact of interventions in wealthier or better-capitalized regions, where infrastructure that already pre-exists, reduces the scope for additional benefits. Additionally, the positive log_rgfcf_pre's coefficient suggests that investments in infrastructure and fixed capital are associated with greater treatment effects.

Table 5: Best Linear Projection of the Conditional Average Treatment Effect (CACE)

Variable	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1.6558	0.1177	14.065	$<$ 2.2e-16***
log_emp_pop_pre	0.2522	0.0524	4.8111	1.540e-06***
log_kstock_pc_pre	-0.1388	0.0102	-13.582	$<$ 2.2e-16***
log_rgfcf_pre	0.0392	0.0089	4.3844	1.184e-05***

Note: *** 0.001, ** 0.01, * 0.05, · 0.1, 1. CI are cluster- and heteroskedasticity-robust (HC3)

While BLP provides insights into the linear relationship between covariates and treatment heterogeneity, the variable importance scores, visualized in Figure 13, for each variable will operate as a quantitative assessment metric measuring the magnitude of the non-linear contributions of each covariate to the model's ability to predict treatment effects. However, it is critical to highlight that these scores do not reflect the causal impact of the covariates on treatment effects but rather the relative contribution of each covariate to splitting the data during tree construction. The log_kstock_pc is the most important variable, since it contributes approximately 64.9% of the total importance, reinforcing the BLP result while reflecting that regional disparities in capital stock heavily influence heterogeneity in treatment effects. While employment and infrastructure are influential, they play a secondary role as they collectively account for the remaining total importance.

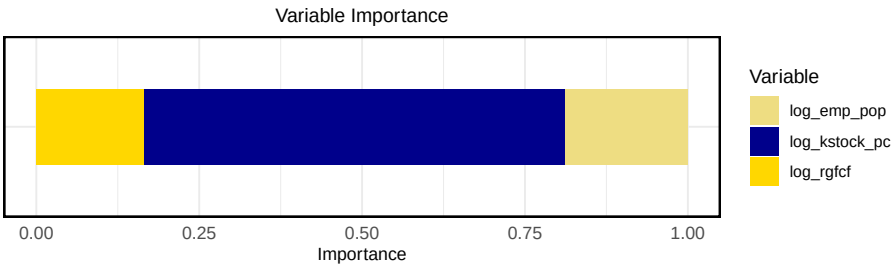

Figure 13: Variable Importance for Causal Forest model

5.4.4 Model Calibration

To validate the reliability of the entire model, I will employ calibrated orthogonal regression with Best Linear Predictor method that was formalized in Chernozhukov et. al. in 2018, to evaluate whether the mean forest prediction and the differential prediction are well-calibrated with the observed treatment effects. Mean forest prediction is the average predicted treatment effect (ATE) for all units and represent the global treatment effect, whereas differential forest prediction is each unit's CATE deviation from the mean forest prediction (CATE − ATE) that captures the heterogeneity in treatment effects across units. For a forest to be considered well-calibrated the coefficients for the predictions should be close to 1. The calibration process involves regressing the estimated CATEs on transformed outcomes that represent actual treatment effects from held-out data, in order to evaluate how well those estimated CATEs align with observed treatment heterogeneity:

$$\text{CATE}_i \sim \alpha\bar{\tau} + \beta\left(\hat{\tau}(X_i) - \bar{\tau}\right)$$

where, α is the mean forest prediction estimate and β is the differential forest prediction estimate. Table 6 shows the calibration results for the trained model. The mean forest prediction coefficient is very close to 1 (0.994), which indicates that the model captures adequately the overall ATE across the data and is able to provide accurate predictions of treatment effects on average, while the small standard error (0.02578) further reinforces confidence in the precision of this estimate. Additionally, confirming that the model effectively detects and captures treatment effect heterogeneity, the differential coefficient for the CATEs is 1.150 and statistically significant at 0.001 level. The value's divergence from zero, provides adequate evidence that there is a positive correlation between the estimated CATEs and the true treatment effects, leading to the conclusion that as the estimated CATEs increase, the actual treatment effects also increase. Thus, these results, not only provide robust evidence that treatment effect heterogeneity exists ($\beta > 0$), but also that the model is capable of identifying it and differentiating between treated regions with diverse responses to the treatment.

Table 6: Calibration Results for Causal Forest Model

CF Model	**Estimate**	**Std. Error**	**t value**	**Pr(>t)**
Mean Forest Prediction	0.99404	0.02578	38.547	< 2.2e-16***
Differential Forest Prediction	1.15020	0.05731	20.068	< 2.2e-16***

Note: *** 0.001, ** 0.01, * 0.05, · 0.1, 1

5.4.5 Testing Causal Forests

To ensure that the model not only performs well on the train data but also it does not fail to generalize well to new, unseen data, I will validate the results on the test set before proceeding to the final heterogeneity analysis. This methodical process ensures that the heterogeneity in treatment effects that was estimated, is not an artifact of overfitting but instead it reflects authentic patterns within the underlying data. First and foremost, I apply the trained Causal Forest model to the test data to leverage the already estimated heterogeneity structure that was earlier learned from the training phase. Consistency is ensured by mirroring the structure used during training with the specific relevant pre-treatment covariates for prediction. Afterwards each observation on the test set is assigned with a

predicted treatment effect. For the validation process I leverage the Kolmogorov-Smirnov (KS) test as well as descriptive statistics of the CATE distributions across the training and test sets. The KS test, introduced by Kolmogorov in 1933 and extended by Smirnov in 1948, is a non-parametric statistical test for assessing whether two empirical distributions differ significantly, by measuring the maximum absolute difference between the cumulative distribution functions (CDFs) of two samples. In this test, the null hypothesis assumes that the two samples come from the same distribution, whereas the alternative hypothesis indicates that the distributions statistically differ. In this case, the test will be leveraged for the comparison of the distributions of predicted CATEs in the training and test sets. The KS statistic (D) is 0.090 and it indicates that the maximum absolute difference between the training and test CATE distributions is 9%, whereas the significant p-value ($9.683e-12$) rejects the null hypothesis that the two distributions are identical and indicates that this difference is unlikely due to random chance. However, while statistical significance is important, it does not imply practical significance. The density plot of CATE distributions in Figure 14 reveal a different viewpoint. Despite the existence of slight variations, the overall shape of the two distributions is largely maintained, indicating that the model preserves reasonable generalization in spite of the formal rejection of the null hypothesis.

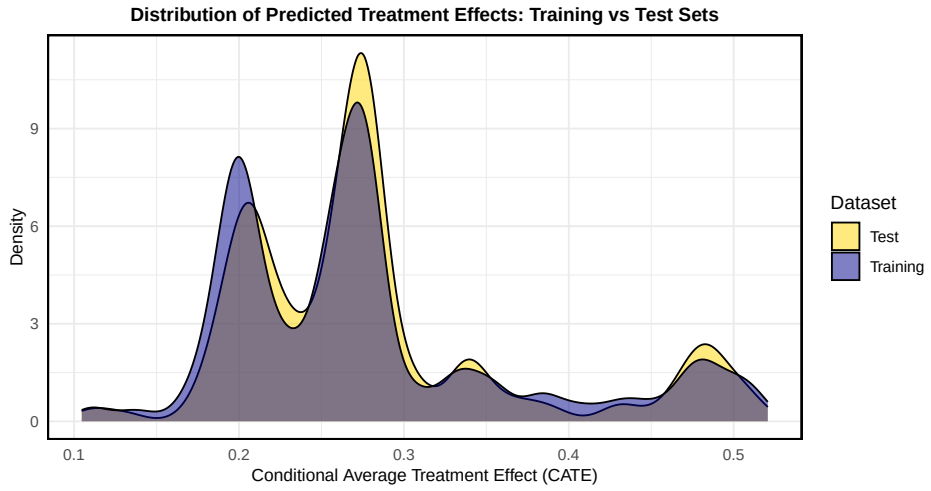

Figure 14: Distributions of Conditional Average Treatment Effects across the training and test sets

Complementing the KS test, Table 7 presents the descriptive statistics for additional intuition into the differences in training and test predicted CATEs. The mean CATE values appear almost identical for training (0.2782) and test (0.2827) sets which indicates that the model does not over-/under-estimate treatment effects in new data, while the slightly lower standard deviation in the test set ($0.0876 < 0.0911$) suggests that the model produces more dense treatment effect estimates in out-of-sample predictions, to some extent. Moreover, the higher skewness for the test set suggests that extreme positive CATEs occur more frequently in that set, while kurtosis being also higher for the test set indicates heavier tails potentially due to great presence of extreme treatment effect estimates.

Table 7: Descriptive statistics for the predicted CATE in the training and test datasets

Dataset	Mean	SD	Skewness	Kurtosis
Training	0.2782	0.0911	1.1014	3.5473
Test	0.2827	0.0876	1.1667	3.8135

Additionally, the covariate balance in Table 8 appears to be also well preserved, indicating that any observed differences in CATE predictions reflect genuine heterogeneity in treatment effects and are not driven by changes in underlying features, confirming that the test set contains observations in similar regions with the training set.

Table 8: Covariate Balance Between Training and Test Sets

Variable	Training Mean	Test Mean
log_emp_pop_pre	-0.8600	-0.8589
log_kstock_pc_pre	10.866	10.881
log_rgfcf_pre	8.5178	8.5207

To verify that the heterogeneous treatment effects are meaningful in the test set for predicting the outcome, I run the simple validation regression of $\log\_rgva\_pc = \beta_0 + \beta_1 \times \text{CATE} + \varepsilon$. Table 9 exhibits the results of the regression, where the coefficient of CATE (β_1) is strongly significant as p-value is smaller than 2.2e-16. This statistical significance suggests that there is a highly systematic relationship between estimated treatment effects and *log_rgva_pc*. The magnitude of the effect (-4.2535) reveals that a one-unit increase in predicted treatment effects corresponds to a significant decrease in *log_rgva_pc*. The significance and size of the coefficient confirms that the model captures relevant heterogeneity in treatment effects, even in a limited-overlap data set-up.

Table 9: Regression Results: Predicting log_rgva_pc using CATE

Variable	Estimate	Std. Error	t-Value	Pr(>t)
Intercept	11.0194	0.0423	260.77	< 2.2e-16***
CATE	-4.2535	0.1443	-29.49	< 2.2e-16***

Note: *** 0.001, ** 0.01, * 0.05, · 0.1, 1

Finally, as preliminary step before the following heterogeneity analysis, by grouping the observations of the training and test datasets into quartiles based on their CATEs, I aim to initially examine how treatment effects vary across subpopulations. This approach also serves as a robustness check to evaluate whether the estimated treatment effects are persistent in an independent sample. In Figure 15 the first quartile contains the observations with the smallest predicted CATEs and the fourth quartile hold the observations with the largest predicted CATEs. For each quartile I estimate the ATE as the mean of the predicted CATEs and the standard errors using the IJ variance estimates. The closely aligned train-test ATE estimates and the overlapping confidence intervals, validate that the estimated heterogeneity is unlikely due to overfitting and that the model generalizes well to new, unseen data. Notably, higher quartiles exhibit larger estimates due to the model successfully capturing treatment heterogeneity. The observed monotonic increase in ATEs across the quartiles confirms the expectation of stronger treatment effects being associated with higher predicted CATEs.

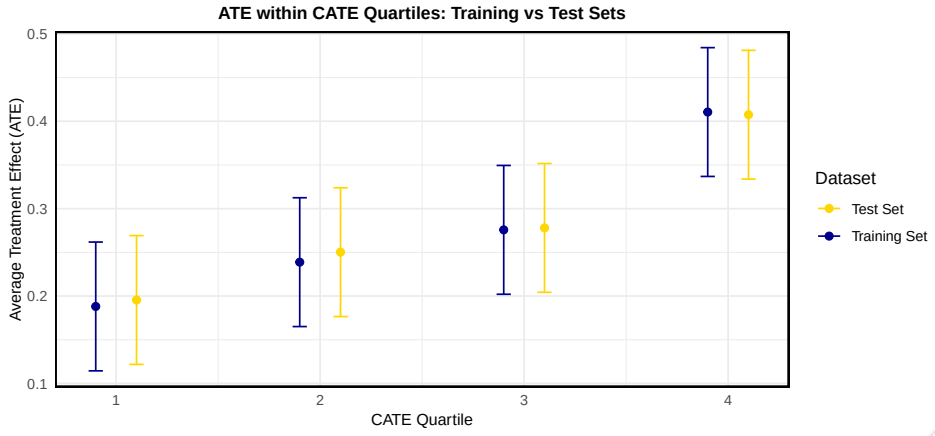

Figure 15: Average Treatment Effects within Conditional Average Treatment Effects Quartiles for both training and test sets

5.4.6 Hyperparameter Sensitivity Analysis

In reference to the previous selection of optimal hyperparameters through cross-validation aiming to minimize the MSE, it is highly important to verify the robustness of the ATE estimates and ensure that they are not overly sensitive to specific parameter choices. High sensitivity would mean that the observed results are artifacts of specific parameter settings and that they do not reflect actual relationships. Therefore, I conducted a sensitivity analysis by varying key hyperparameters, such as the minimum node size (5, 10, 20) and the number of trees (500, 1000, 1500, 2000, 4000), using the same range of values tested earlier for minimizing the MSE. With variation no greater than ± 0.02, the results in Figure 16 validate the stability of ATE estimates across different configurations, indicating that the findings are consistent, reliable and not driven by specific model configuration settings. Specifically, the ATE estimates are fairly stable across different minimum node sizes, especially for minimum node size greater than 10, striking a balance between capturing local heterogeneity and ensuring reliable average estimates.

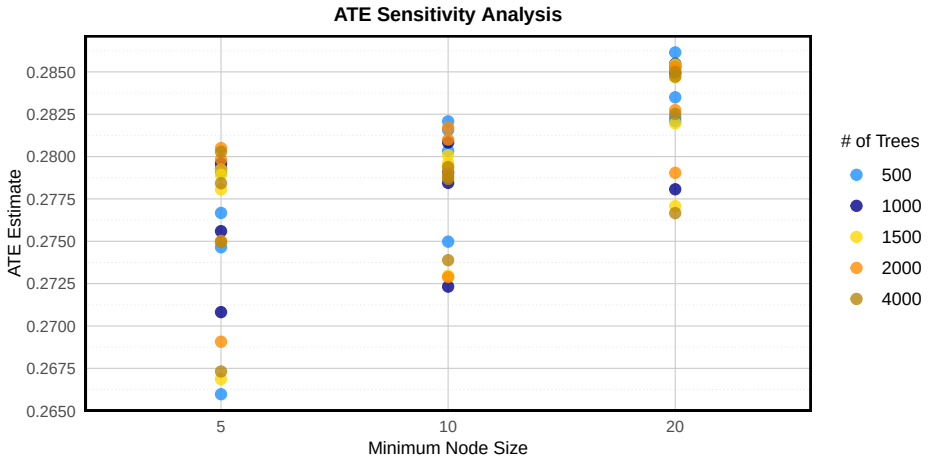

Figure 16: Average Treatment Effects Sensitivity to Hyperparameters

5.4.7 Advanced Heterogeneity Analysis: Interactions and Subgroup Effects

While the average treatment effect provides a broad overview, it fails to capture the rich complexity of how different regions respond to the same intervention. Some regions could prosper as they turn targeted funding into significant growth, while others could stagnate and get trapped in structural limitations dampening their progress potential. Therefore, for more equitable and effective interventions, it is critical to embrace the phenomenon of non-uniform treatment responses.

As previously discussed, a single-peaked, symmetric distribution, centred around the average effects is expected in cases where treatment effects are homogeneous across regions. However, the histogram in Figure 17 reveals a bimodal distribution of CATEs, with two distinguishable peaks, verifying that the treatment effects are not homogeneously distributed across regions. The first peak on 0.2 suggests that this subset of regions shows moderate effects while the second peak on 0.27 indicates a subset with slightly higher effects. Therefore, these peaks reveal the existence of two distinct subpopulations that experience different levels of treatment effects. Moreover, the presence of high responders who benefit more than average (CATE > 0.4) is confirmed from the noticeable right-skewness of the distribution (mean is 0.279 and median is 0.264) and a smaller peak around 0.48. This further supports that there are indeed systemic differences among treated regions, as some of them experience remarkably large treatment effects, while most regions experience moderate treatment effects. Additionally, minimum 0.108 and maximum 0.516 that indicate a broad range of estimated effects, suggesting that one-size-fits-all policy approach may not be ideal.

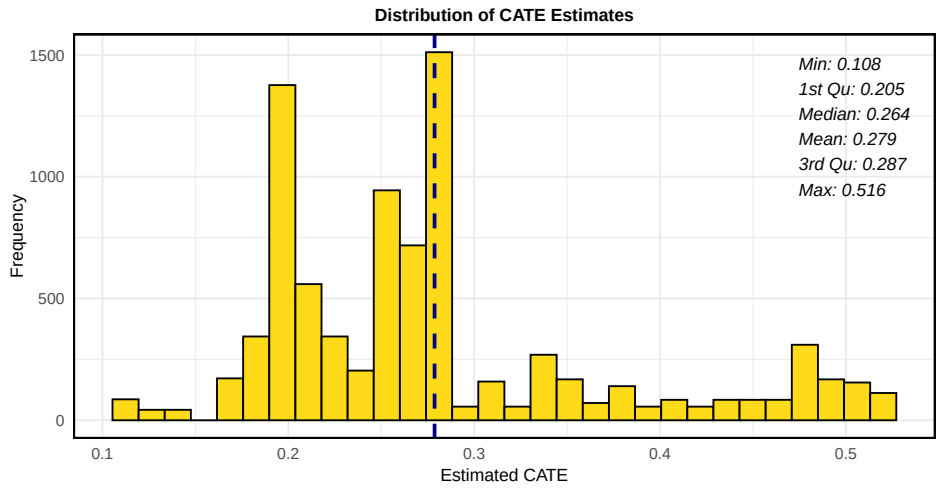

Figure 17: Overall Distribution of Conditional Average Treatment Effects

For further investigation of the heterogeneity that became evident from the above plot and to gain a more intuitive sight of the distribution, Figure 18 presents the grouping results of hierarchical clustering applied to the estimated CATEs. Prior research has demonstrated the efficacy of hierarchical clustering in regional policy evaluation in a panel data setting (Altuntas, Selim, & Altuntas, 2022). Hierarchical clustering is an unsupervised machine learning technique that groups observations into clusters based on similarity, particularly useful for identifying latent group structures in high-dimensional data (Hastie, Tibshirani, & Friedman, 2009). In the context of my analysis, I will apply hierarchical clustering to the estimated CATEs for the identification of meaningful subgroups of regions. Due to the ability of hierarchical methods to uncover natural groupings within economic data, the three different response groups that are constructed will be leveraged to identify heterogeneity patterns within the heterogeneous distribution of CATEs. At first, a distance matrix is constructed, using the Euclidean distance between CATE values. Following that, to ensure compact and well-separated clusters, Ward's method minimizes the variance within each cluster while iteratively merges clusters, creating eventually three clusters/subgroups. Low responders with values from 0.1 to 0.25 create a clear left-skewed cluster, indicating a group of regions that benefit less from treatment. The neutral responders from 0.26 to 0.31, is the group with moderate treatment effects, centered around the mean CATE. On the right, the group of high responders (0.32-0.53) appear to be the most dispersed one, while it experiences strong positive treatment effects, suggesting that the intervention is exceptionally beneficial for this subset of regions.

In essence, while there are regions that treatment has a moderate but positive effect, the presence of outliers on either side of the distribution reflects the necessity of targeted analysis. Low responders could be regions where intervention was ineffective due to structural barriers, such as capital saturation leading to low capital absorption, as described in the diminishing returns framework (Solow, 1956). To benefit from interventions, these regions may require pre-intervention support, such as training programs. In contrast, high responders could include regions with a strong entrepreneurial culture and high employment elasticity, allowing firms and workers to respond more effectively to interventions, making them ideal candidates for scaling current interventions.

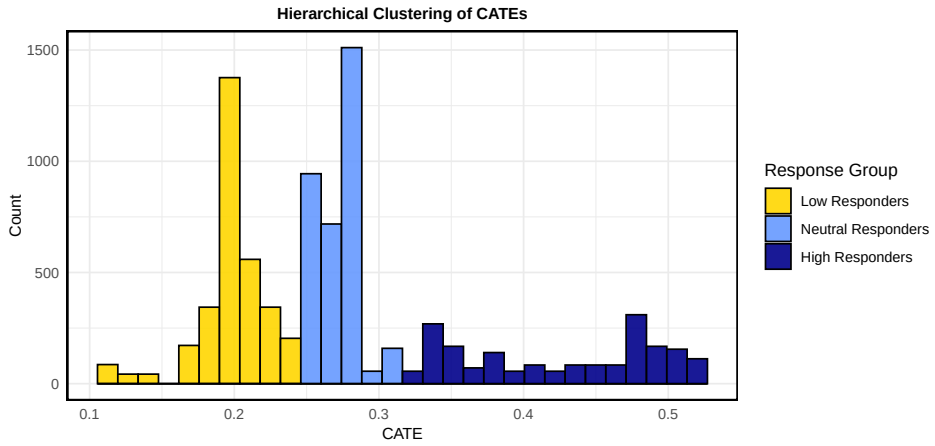

Figure 18: Histogram of Conditional Average Treatment Effects clustered into three response categories

As a consequence, this considerable variation in treatment effects naturally raises the question; "What determines whether a region is a high or low responder?". For a more detailed examination and a granular view I examined the dataset by CATE deciles, with 846 observations per decile. By quantifying the observed heterogeneity in Table 10, I aim to investigate whether treatment effects are driven by economic structure. If this is the case, systematic differences in covariates across CATE deciles, should be apparent for the three variables. The table shows the mean values for each variable, across the 1st and 10th CATE deciles, with the 1st decile representing regions with the weakest treatment effects and the 10th decile regions with the strongest. This comparison highlights the distinct economic characteristics of regions at the extremes of treatment effect heterogeneity. Indeed, the reported means (with the standard deviations in the parentheses) indicate systematic differences in economic characteristics between regions with low and high CATEs and the p-values from the two-sample t-tests confirm the statistical significance of those differences. Notably, capital stock exhibits the largest difference, as regions with low estimated CATEs tend to have significantly higher capital stock per capita (10.9), while regions with high CATEs tend to have lower capital stock per capita (9.20). This finding is consistent with the diminishing marginal returns to capital accumulation. For the employment rate, the less pronounced difference between the 1st decile mean (-0.95) and the 10th (-0.84) suggests that treatment effects are stronger in regions with higher labour mobility and market flexibility, allowing quicker adjustment to new economic opportunities, potentially due to greater labour market dynamism or absorptive capacity. Regarding gross fixed capital formation, the difference in magnitude between the 1st decile mean (7.87) and the 10th (7.59) likely reflects investment saturation or inefficiencies in capital allocation in regions with low treatment effects. These results highlight the importance of designing different, more tailored, policy tools, that address the different economic profiles of regions. For instance, capital-saturated regions with considerable capital stock may require incentives that prioritize innovation and productivity over capital accumulation, whereas for regions that exhibit lower levels of capital but higher employment elasticity, direct labour market reforms and training programs could be more effective.

Table 10: Comparison of Mean Differences Across Conditional Average Treatment Effects Deciles

	1st Decile (N=846)	10th Decile (N=846)	Overall p-value (two-sample t-test)
Employment Rate	-0.95 (0.13)	-0.84 (0.11)	<0.001
Capital Stock	10.9 (0.35)	9.20 (0.49)	<0.001
GFCF	7.87 (1.29)	7.59 (0.51)	<0.001

Motivated to deepen the analysis on the observed heterogeneity I will exam[illegible] continuo[illegible] marginal relationship between the covariates and CATE using Partial Depende[illegible]nce Plots (PDPs) shown in Figure 19. In contrast to the simpler, decile-based comparisons, PDPs all[illegible]w for continuous and granular assessment of how treatment effects evolve across the entire distributio[illegible]n of each predictor. By holding other covariates constant PDPs isolate the marginal effect of each pred[illegible]ictor on CATE[illegible]

anticipated, these PDPs visually confirm nonlinear patterns in treatment effect heterogeneity, within a 95% confidence interval. For employment rate the first plot illustrates the non-monotonic relationship that Table 10 above cannot capture, mainly because the latter compares only the most extreme groups. In this PDP the treatment effects remain low for regions with higher unemployment. Notably, beyond a certain threshold, CATE increases and remains stable at lower unemployment levels, likely reflecting that regions with higher labour market participation are prone to more responsive behaviour to interventions. Consistent with previous findings, for capital stock, lower values are associated with higher CATEs, while at stock levels around 10-11, CATE declines sharply as the stock increases and its additional contribution to productivity and treatment effects reduces. Eventually CATEs are stabilized in lower levels, for regions with higher levels of capital stock. Finally, complementing the table's results for gross fixed capital formation that associated higher investment with lower CATEs, as additional investment yields smaller incremental benefits, PDP reveals their complex relationship. For low investment levels the treatment effects are low, but after a certain threshold a positive slope indicates that an increase in investment levels increases the CATEs. However, after this steep increase, the effect flattens at higher levels validating the notion that past investment patterns shape the effectiveness of future interventions.

Hence, in presence of non-monotonic relationships, both continuous and discrete approaches are necessary to assess treatment heterogeneity. PDPs provide a continuous, smoothed estimate of the relationship between covariates and CATEs while decile-based comparisons still serve as an important complementary tool for validating general trends and providing a cleared view on treatment effect variation across distinct segments. For completeness, Figures 28, 29 and 30 in the Appendix depict the mean CATE within deciles of the covariates for dual purpose; not only to provide a clear perspective on which variables vary strongly with treatment effects, but also to confirm the general patterns observed in the PDPs but with greater variability due to discretization.

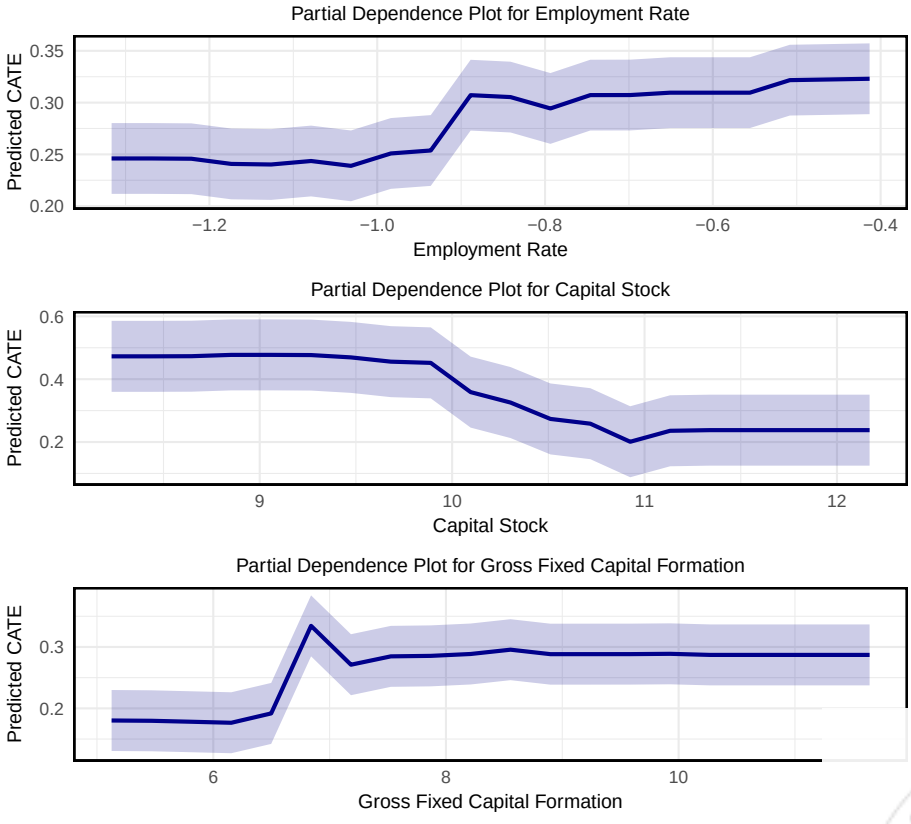

Figure 19: Partial Dependence Plots Across Covariates

While marginal PDPs are extremely useful for isolating the individual effect of each covariate on CATE, they may fail to observe key interaction effects between covariates. In cases where economic

variables are correlated, the independent effect assumptions of marginal PDPs may not hold. While there is a low-to-moderate positive correlation for the pairs of employment rate & capital stock (0.321) and employment rate & gross fixed capital formation (0.309), there is a moderate-to-high positive correlation between capital stock and gross fixed capital formation (0.624) (see Figure 31, Appendix). Therefore, building on the insights from the marginal PDPs, I examine how these covariates interact to jointly shape CATEs. In Figures 20, 21 and 22 I display the two-dimensional PDPs for a visual investigation of how combinations of multiple economic profiles influence the predicted treatment effects.

Figure 20 reveals a clear non-linear pattern between employment rate and capital stock. For regions with lower capital stock levels, as the employment rate increases there is a significant positive effect on CATE (0.40-0.50). In contrast, for regions with high capital stock levels, the predicted CATE remains at low levels (0.20-0.35), regardless of the employment rate, indicating that interventions that are likely aimed at capital accumulation are no longer effective. This pattern indicates that early-stage interventions, targeting regions with low capital stock and employment growth, yield stronger positive treatment effects. These regions are more responsive to the subsidies compared to regions with low capital stock and low employment rate. Unlike those regions, regions with high capital stock and low employment rate are the least responsive to treatment, while regions with high capital stock show a modest improvement in responsiveness as employment rates increase. The horizontal blue band observed in this plot, particularly among regions with high capital stock and high employment rates, reflects a clear pattern of limited treatment responsiveness (0.20), suggesting that certain regions within this group exhibit weaker responses to the intervention, compared to the rest regions of the group (0.30).

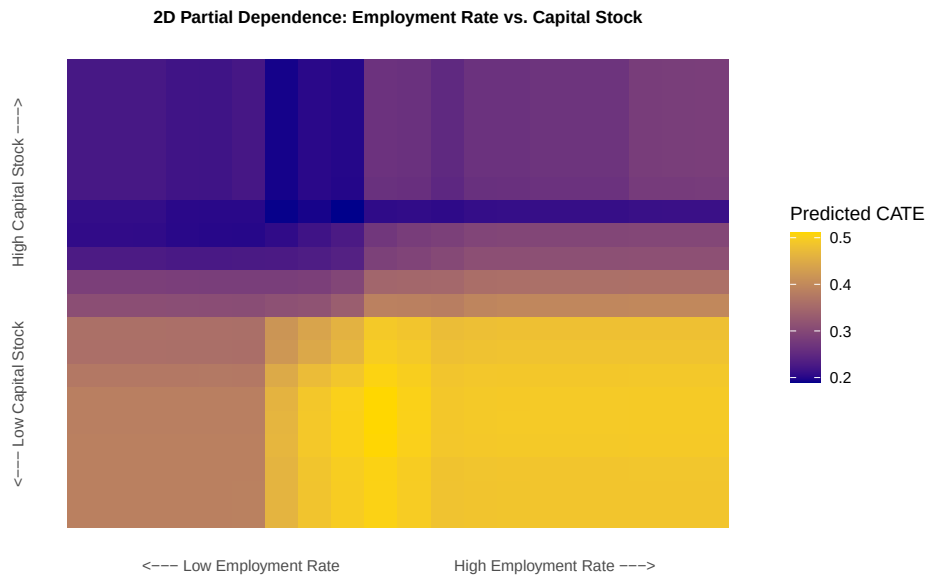

Figure 20: Conditional Average Treatment Effects Heatmap – Employment Rate vs. Capital Stock

The non-linear interaction of employment rate and gross fixed capital formation is illustrated in Figure 21, where the low gross fixed capital formation levels are associated with low CATEs (0.20-0.25) regardless of employment rate, suggesting limited responsiveness of interventions at early stages of investment. Interestingly, there is a horizontal yellow line that appears just b[illegible] between low and high GFCF levels stands out as an anomaly compared to the re[illegible] Conceptually, this unexpected pattern could indicate a "transition zone" where [illegible] the cusp of moving from low to high GFCF levels experience disproportionately [illegible] subsidies. Additionally this pattern may be attributed to a nonlinear interaction b[illegible] and GFCF. This could suggest that when GFCF is near a crtical threshold, incr[illegible] unlocks productivity levels that are not achievable in regions further below the th[illegible] capital shortages. Notably, higher gross fixed capital formation levels combined wit[illegible] rates correspond to increased CATEs (0.30-0.36), indicating that subsidies are m[illegible]

dynamic labour markets where resources are more effectively allocated. Therefore, to maximize their impact, subsidies should be prioritized in regions with high GFCF, particularly those exhibiting high employment rates. In cases where subsidies are directed toward regions with low GFCF, targeting those with relatively higher employment rates within this group is likely to yield more substantial treatment effects ($0.23 > 0.17$).

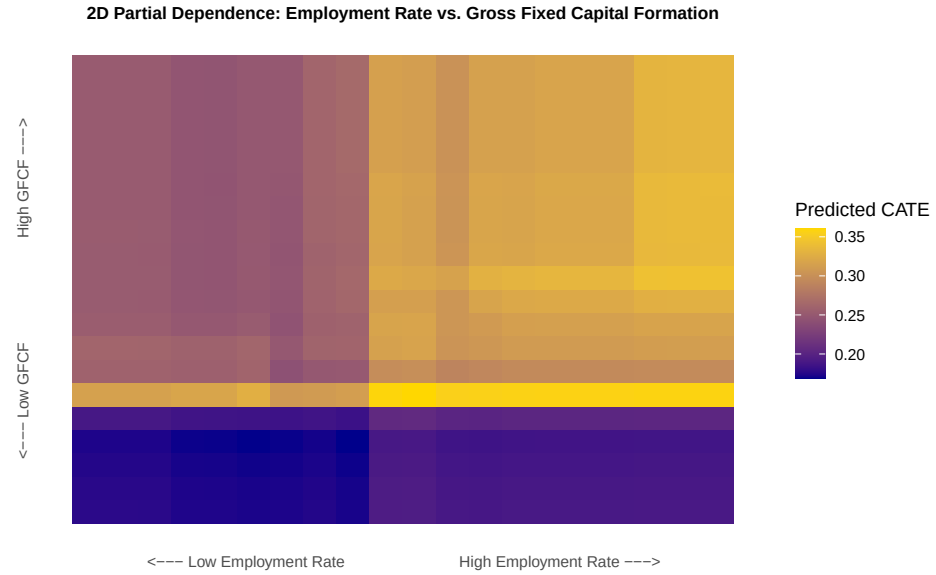

Figure 21: Conditional Average Treatment Effects Heatmap – Employment Rate vs. Gross Fixed Capital Formation

Next, Figure 22, highlights the combined effect of capital stock and gross fixed capital formation on CATE. When both capital stock and investment levels are low, the relatively high predicted CATEs (0.40), indicate that early-stage interventions are highly beneficial for regions with limited initial resources. As expected, when capital stock increases, CATE values severely decline below 0.20. Interestingly, regions with both high capital stock and high gross fixed capital formation levels show extremely high responsiveness to treatment (0.40-0.49) compared to those with high capital stock & low investment (> 0.20) and low capital stock & high investment (0.20-0.30). These findings suggest that the interventions should be targeted in low capital stock regions, combined with high investment levels, as they have more noticeable effects. Conversely, in regions characterized by high capital stock, subsidies should be directed toward those with elevated investment levels, as they are more likely to yield substantial benefits compared to regions with lower investment levels. The presence of a horizontal yellow band in this plot, mirroring the pattern observed in the previous figure, suggests that regions with low GFCF but nearing the threshold for higher investment levels exhibit notably positive treatment effects (0.40). This finding indicates that certain regions with high capital stock and relatively low GFCF still respond positively to the treatment, challenging the assumption that high capital stock universally corresponds to low responsiveness. If policymakers were to assume that all high-capital-stock regions exhibit limited treatment effects, they would risk overlooking these responsive areas, thereby misallocating resources. Additionally, the presence of a darker blue band among regions characterized by both high GFCF and high capital stock signals that these regions exhibit notably lower responsiveness to the intervention (0.20). While the broader category of high-GFCF & high-capital-stock regions generally demonstrates low responsiveness, the specific subset within this darker band appears to experience even weaker treatment effects, falling below the 0.20 threshold. This suggests that within this category, there exists further heterogeneity in policy responsiveness, reinforcing the need for a more granular approach to policy targeting.

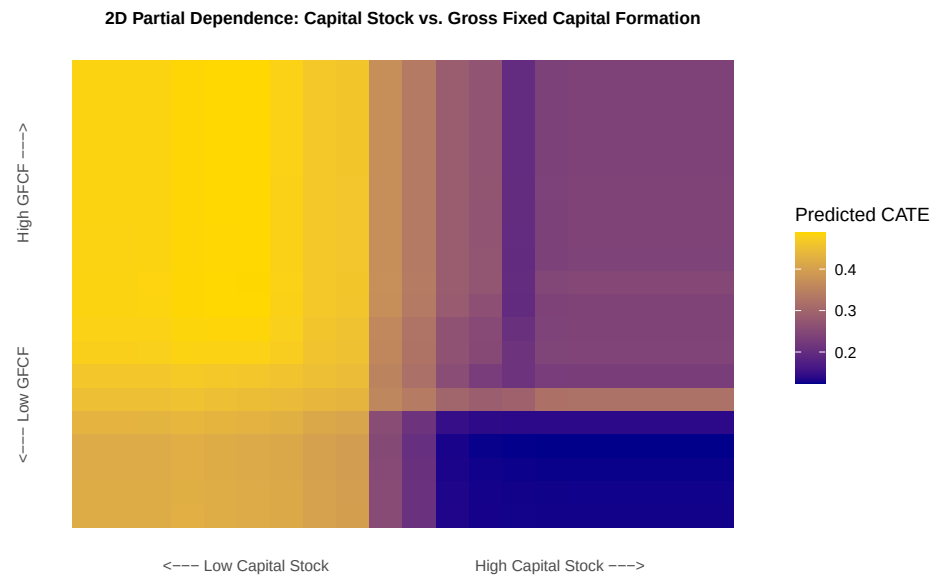

Figure 22: Conditional Average Treatment Effects Heatmap – Capital Stock vs. Gross Fixed Capital Formation

After establishing that the treatment effect varies across economic conditions, a natural question emerges; "Which are those regions that require targeted policy interventions?". As a matter of course, in Figure 23 my focus shifts from examining how the effect varies across regions to contextualizing this heterogeneity geographically on a spatial map.

In my analysis, the maps present CATE estimates for both treated and control regions, offering a complete geographical perspective on potential treatment responses. However, for control regions, CATEs represent counterfactual estimates—hypothetical values indicating how these regions would have responded if treated. This extrapolation of treatment effects to regions outside the treated sample is a significant strength of causal forests, as they leverage complex patterns in covariate relationships. Since this analysis does not impose strict overlap conditions the maps serve as a valuable exploratory tool to highlight broader trends. While this flexibility aligns well with real-world scenarios where perfect overlap is indeed rare, caution is warranted for regions with high covariate divergence, as these estimates may be less reliable. However, this reflects the reality of policy evaluation, where overlapping covariate profiles between treated and control units are often imperfect, making the capacity to explore heterogeneity in counterfactual responses particularly valuable.

With that premise established, I begin my geographical analysis with Figure's 23 spatial map that clusters the 228 European regions into three previously discussed groups based on their estimated CATEs: low responders, neutral responders, and high responders. The clustering pattern indicates that high responders are predominantly concentrated in Eastern Europe, while low responders dominate Western and Southern Europe. I cannot assume that all European regions are a tabula rasa, equally capable of absorbing a given treatment and producing identical outcomes. Each region is inherently different due to its distinct historical, political and economic context. As Giannetti (2002) argued, historical disparities between European regions are deeply rooted in these differences. For Western and Northern regions, the early industrialization, stronger institutions and stable political environments contributed to their current economic success. In contrast, Southern regions are often characterized by their lower productivity and less diversified economies. Central and Eastern reg[illegible]ons are historically shaped by the political and economic isolation during the communist era, burde[illegible]ed to continue a legacy of underinvestment and weaker institutions, even after the transition to a market [illegible]. These disparities affect regions' capacity to "accept" and respond to the treatmen[illegible].

As evident from the map, regions in Southern Europe (parts of Italy, Spai[illegible] and France) considered low responders to treatment due to structural challenges like lower pr[illegible]oductivity and diversified economies or lagging innovation ecosystems and bureaucratic hurdles, [illegible] typical barriers in Southern European countries. The parts of Western Europe (e.g., most regions in [illegible]France and Germany) that exhibit low responsiveness likely reflect regions with saturated ca[illegible]pital [illegible]unimpeded

marginal returns from additional capital investments, highlighting that interventions should be targeted at innovative, high-value-added sectors such as green energy, digital transformation or research and development (R&D) ecosystems. Having started from a lower baseline and striving to converge to the already developed regions, Eastern regions in countries like Poland, Slovakia and Hungary, create a cluster of high responders to treatment, reflecting their lower capital saturation and stronger marginal impact of capital investments. These regions have been major beneficiaries of EU cohesion funds that target infrastructure development and economic modernization. Finally, for the most regions in Northern Europe (Sweden, Finland, Denmark) and in Central Europe (Austria, Slovenia, most parts of Germany), due to the well-functioning and robust institutions, capable of effectively implementing interventions, regions' responses are less dramatic but still distinguishable. Their high levels of capital accumulation and already working-properly labour markets may limit the potential for additional gains from new interventions, but targeted interventions in emerging sectors (e.g., green energy) or research facilities can still deliver remarkable returns.

These findings confirm that heterogeneous treatment effects are regionally structured and vary significantly in magnitude.

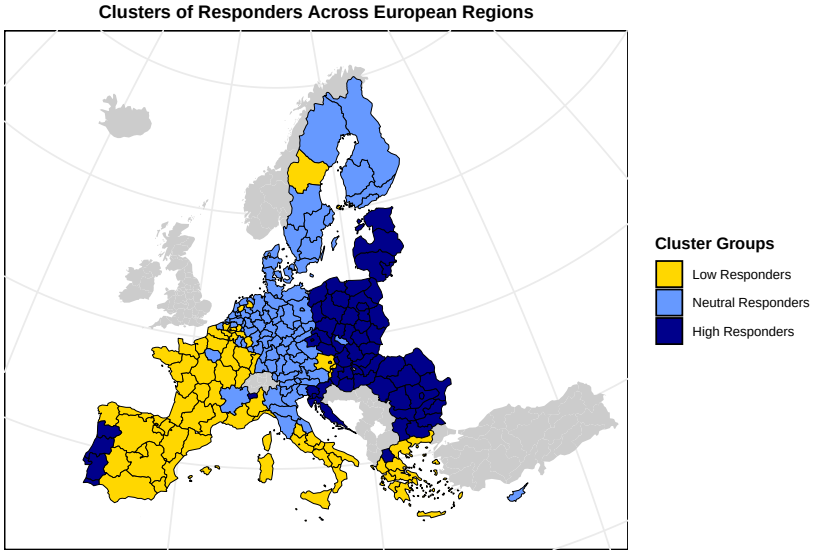

Figure 23: Map of policy response clusters based on mean Conditional Average Treatment Effects

Moving beyond clusters, the map in Figure 24 reveals the actual magnitude of treatment effects across all years per European region. It presents a continuous gradient of mean CATEs, ranging from approximately 0.1 to 0.5. This substantial variation of CATEs suggests that some regions, mainly concentrated in Eastern Europe, experience more than double the magnitude of impact (CATEs $>$ 0.4), as they are in their catch-up phase, compared to Western and Southern ones. As a matter of fact Southern regions display consistently low CATEs (< 0.3) raising concerns about persistent divergence patterns. These high CATEs in Eastern regions suggest that EU cohesion policy has been successful in fostering regional convergence, especially in infrastructure and industrial development.

However, this raises an intriguing counterfactual question; "What would the treatment effect map look like if wealthier regions—those with higher initial capital and development levels— had been targeted for intervention?". In line with economic theory, wealthier regions in Western and Northern Europe would likely exhibit generally lower mean CATEs due to diminishing marg[illegible]nal returns. Addi[illegible]t effects would[illegible] be uniform. Urban centres and innovation hubs' areas, such as Lombardy in Nort[illegible]hern Italy or Pa[illegible] in France, might still exhibit moderate treatment effects but rural and peripheral [illegible]areas like [illegible] Basilicata, and parts of Sicily in Southern Italy or Aquitaine in Western France, [illegible]would [illegible] adequate responsiveness. Notably, regions with economies heavily reliant on to[illegible]urism[illegible] seas[illegible] industries like Andalusia in Spain (CATE < 0.3), Algarve in Portugal (CATE < 0.3[illegible]), [illegible]ia in [illegible] (CATE < 0.3), Crete, Ionian Islands and South Aegean in Greece (CATEs: 0.3–[illegible]0.35[illegible]) [illegible]mixed CATEs. This occurs as a result of capital investments actually improving physical [illegible]infrastructure[illegible]t.

without economic diversification, their long-term economic impact becomes limited.

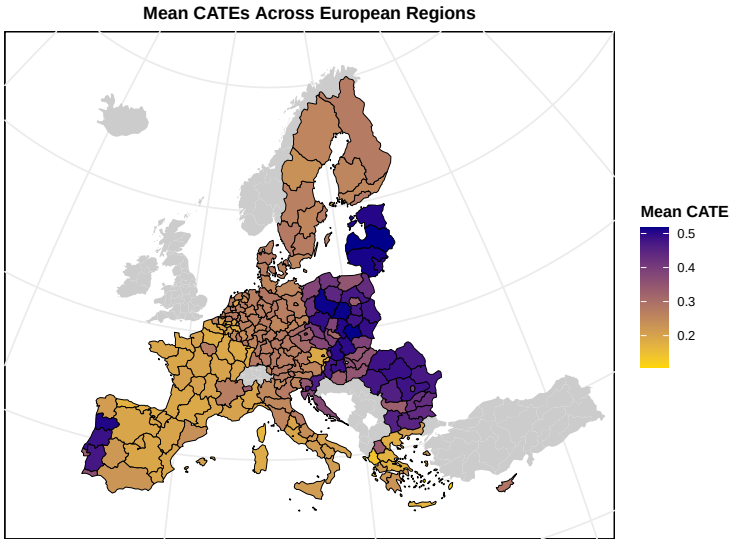

Figure 24: Map of mean Conditional Average Treatment Effects by region

At this stage, having established beyond doubt a clear causal link and identified significant heterogeneity in treatment effects, my analysis is ready to answer the core research question; "What is the effect of EU funding on the regions' Gross Value Added?". However, just as importantly, it dives deeper into the exploration of the accompanying question; "How does the effect of EU funding differ across regions?".

As a consequence, Figure 25 illustrates the intersection of CATEs and GVA values across European regions, for a dynamic view of how policy effectiveness is causally connected to regional GVA. The dark-green regions (high CATEs & high GVA) epitomize "success stories". For these regions, parts of Germany, Austria, Luxemburg and Southern Sweden, the targeted investments effectively translate into major economic gains, likely reinforcing regional confidence and forward-looking expectations. In these regions, with high-governance settings, policy interventions have significant positive impacts on GVA. In contrast, the light-green (low CATEs & low GVA) regions appear more structurally challenged as interventions have shown limited effectiveness, highlighting a persistent trend of economic stagnation. Economic growth cannot be driven by money alone, in the absence of good governance. For regions in parts of Southern Italy and Spain or Greece, the results could stem from mistrust in institutions, due to corruption or political clientelism, and short-term thinking shaped by decades of persistent economic underperformance that fostered a climate of reduced willingness to take risks or invest in new opportunities (Martín-Fernández et al., 2021). The golden regions in Central and Eastern Europe (Poland, Hungary, the Baltics or Slovakia) with high CATEs & low GVA exhibit an unexpressed potential, shaped by their high responsiveness which indicates that they are on the verge of growth but face short-term challenges, such as potential mismatch between labour market readiness and funded projects. The fewer blue regions (low CATEs & high GVA) in Western Germany, Italy and Sweden and the light-blue regions (low CATEs & moderate GVA) are portraying self-sufficient regions, where their endogenous growth mechanisms navigate their economic performance, independent of external interventions. Interestingly, the previously identified low-responder regions, exhibit distinct growth patterns. Most Greek regions, Southern Spain, Northern Portugal and a region in Southern Italy appear to have low GVA levels indicating their limited capacity for growth, but for most French, Southern Italian and Northern Spanish regions the underlying factor behind their low responsiveness is their moderate GVA values, and therefore reduced necessity for reliance on external investments. Moreover, the spatial pattern of CATEs reflects spillover effects between neighbouring regions. For instance, regions in Italy, Sweden and Germany that are closely tied to high-growth neighbours may see some uplift, while those near to heavily funded regions in Eastern Europe could be missing talent investment.

Geography has the ability to amplify or dampen a region's response to the policy. Regions with

the economic "core" of Europe (central locations with good market access) can better exploit new investments, whereas peripheral and rural regions (remote, mountainous or far from major markets), even with the same investment, might see a smaller GVA bump due to smaller labour pools, out-migration, or weak access to supply chains.

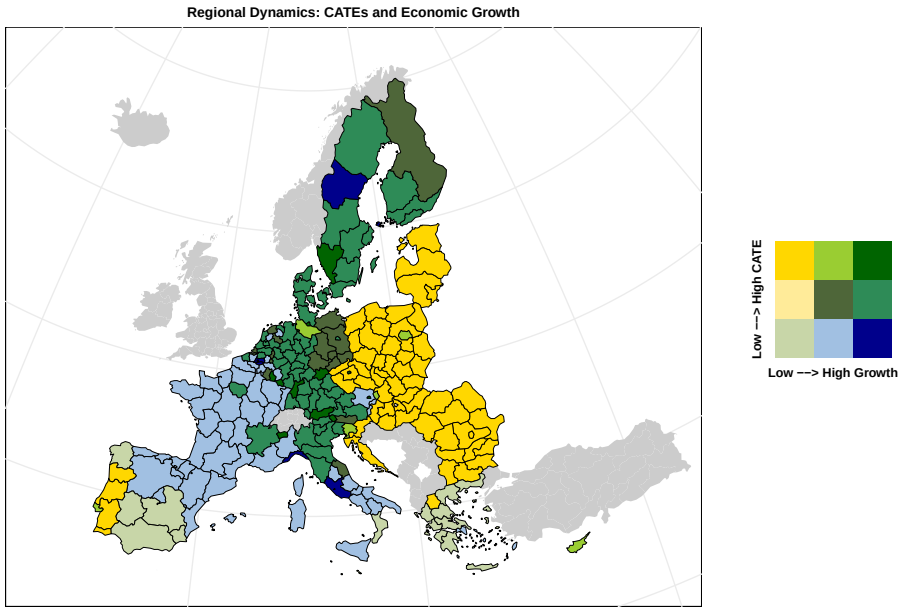

Figure 25: Intersection of Conditional Average Treatment Effects and Economic Growth

5.5 Empirical Benchmarking: GSC vs. Causal Forests

Although GSC is more interpretable for panel data and pre-/post-treatment comparisons and Causal Forests provide plentiful insights into treatment effect heterogeneity, by observing the single and dynamic ATT estimates for both methods, I aim to evaluate their consistency. While plotting ATT for Causal Forests is not yet as standard as it is for GSC, it is increasingly being adopted as a practice in empirical research. The model estimates the CATE ($\hat{\tau}(X_i)$) for each unit in the dataset. In order to compute the ATT as the average of the CATE values for all treated units, the process focuses only on treated units. By plotting GSC's and Causal Forest's dynamic ATT in Figure 26, I aim to observe the average treatment effect per time period. For GSC, dynamic ATT is computed directly as the average unit-level effects for each time period after the treatment, whereas for Causal Forests, the ATT values can be approximated by grouping treated units by *rel_year* and averaging their CATEs for each *rel_year*, leveraging the heterogeneity captured by covariate-based predictions. Finally, when the comparison focuses on unit-specific treatment effects, each method has a different approach for the estimation. GSC constructs counterfactual outcomes for each unit in the post-treatment period by assuming parallel trends and leveraging control units' pre-treatment trends, in order to estimate unit-specific treatment effects over time, while Causal Forests use covariates to predict the unit-level $\hat{\tau}(X_i)$ that captures heterogeneity across units.

Differences in ATT trends between GSC (rooted in a parametric approach) and Causal Forests (a non-parametric ML method) are not only feasible but rather expected due to the fundamental different approaches of the methods. Despite the methods providing complementary insights, they are not directly comparable. However, in view of comparing them, in Figure 26, GSC's dynamic ATT provides estimates of the dynamic treatment effect over time, offering clarity on time-varying treatment effects, whereas Causal Forest's manually-constructed from estimated CATEs dynamic ATT, captures the covariate-driven heterogeneity over time. For the GSC model, the more conservative average ATT of 0.0908, indicates a small average positive treatment effect over time periods. The increasing positive dynamic ATT trend suggests a delayed or cumulative treatment effect for treated regions over time. Moreover, as time progresses the widening post-treatment confidence intervals suggest increasing uncertainty in the treatment effect estimates. However due to GSC's simplicity, these results

may oversimplify the treatment effects by ignoring covariate-driven heterogeneity or dynamic patterns. The Causal Forest ATT plot presents a unique perspective, with a strong positive average ATT (0.3471) suggesting a positive treatment effect on average for the treated regions. As Causal Forests emphasize heterogeneity across covariates the dynamic positive results reflect stronger and more stable initial treatment effects over time compared to GSC results, despite the decline in later periods. This positive dynamic trend serves as an indication that subgroups with specific covariate profiles, are driving higher treatment effects as they experience disproportionately greater post-treatment benefits compared to others. The consistent narrow confidence intervals across time, suggest high precision in the ATT estimates. Although such behaviour is typical across ML models as it reflects the natural tendency of forests to capture fine-grained patterns in the data.

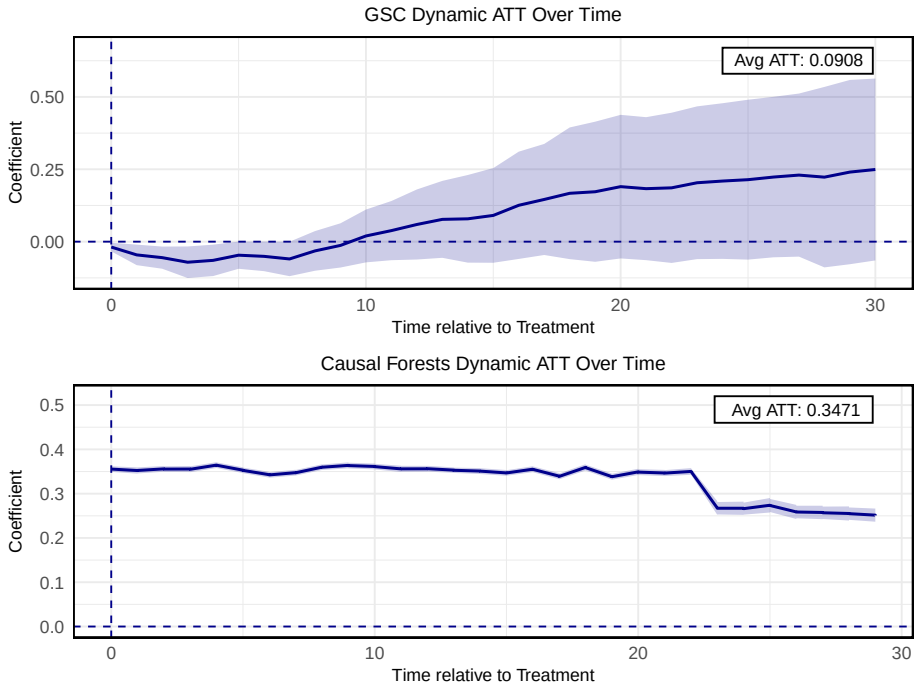

Figure 26: Dynamic and single Average Treatment effect on the Treated values for GSC and Causal Forests over time

This comparison is not intended to determine which method is superior but rather to clarify the specific contexts and research objectives for which each method is best suited and to underscore the demand for synergistic insights of both methods. After establishing this, GSC appears to be better suited for reliably estimating the overall treatment effect while Causal Forests are leveraged to explain heterogeneity and dynamic patterns, allowing policymakers to combine these approaches to identify overall program success and adapt regional strategies based on subgroup-specific outcomes. In this case, GSC empirically validates the initial assumption that global effect of subsidies are on average positive, with the estimation of a significant average treatment effect across all treated regions (0.0908). Causal Forests extend this analysis by revealing how the magnitude of this positive effect varies across different EU regions, with regions in the top deciles experiencing significantly higher treatment effects, while others in the bottom benefit only marginally. This provides a solid foundation for more targeted policy interventions that address the specific needs of each region.

6 Conclusion

Within this research I aimed to bridge the gap between traditional econometric models and machine learning methodologies by depicting their complementary strengths and their relevance for policy evaluation. This study contributes to a growing body of literature that seeks to refine the

ways for estimating heterogeneous treatment effects, especially in EU regional subsidies' context, by methodically evaluating Generalized Synthetic Controls (GSC) and Causal Forests.

As George E. P. Box famously argued "All models are wrong, but some are useful". A key takeaway is that no single model can adequately capture the complexity of real-world policy impacts. Traditional models, such as Difference-in-Differences (DiD) and Two-Way Fixed Effects (TWFE) are useful but limited, especially in staggered adoption settings or settings with time-varying treatment effects. These limitations can distort policy conclusion, leading to biased results and suboptimal decisions. Although, some of these shortcomings are addressed with GSC due to its flexibility in handling multiple treated units and time-varying treatment effects, it introduces challenges related to factor estimation and computational demands. Meanwhile, by embracing machine learning's strengths, Causal Forests offer a robust alternative for precision policy-making as they focus on treatment effect heterogeneity.

The findings of EU regional subsidies' empirical application confirm the existence of heterogeneity in treatment effects across regions. Understanding regions' differences is critical and they highlight the need for a more context-sensitive policy design that tailors interventions to the specific characteristics of each region. Additionally, this application reveals that Causal Forests outperform traditional methods in identifying these heterogeneous effects, making them irreplaceable for economists that aim to go beyond global average treatment effects into subgroup-specific treatment effects. In cases that treatment assignment is not random or when regions do not follow parallel trends, GSC offers a flexible approach than DiD and TWFE, but its success heavily relies on cautious selection of latent factors and adequate data availability.

Throughout this research journey, while focusing on my primary research question, I also sought to address additional critical questions regarding the inherent differences and complexities of European regions. Going forward, this research underscores the importance of methodological pluralism in causal inference. It highlights that policymakers should move away from one-size-fits-all approaches not just to improve policy evaluation but to allocate resources more effectively and better address the diverse needs of different regions. Moving forward, further research should explore how the integration of ML techniques in causal inference is becoming a promising direction for future research, while the ongoing collaboration between economists and data scientists will be crucial to developing robust, scalable models that are both theoretically and practically relevant for actually promoting growth and reducing inequalities across Europe. The future of policy evaluation does not lie in a choice between tradition and innovation, but rather in harnessing their combined strengths to shape policies that are not only effective but truly inclusive.

Bibliography

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021. doi: 10.1080/01621459.2021.1971535.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Serkan Altuntas, Sibel Selim, and Fatma Altuntas. A hierarchical clustering based panel data approach: A case study of regional incentives. *International Journal of Information Management Data Insights*, 2(2):100098, 2022. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2022.100098. URL `https://www.sciencedirect.com/science/article/pii/S2667096822000416`.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2):277–297, 1991.

Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 5–6, 2015.

Susan Athey. The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

Susan Athey, Guido Imbens, and Vikas Ramachandra. Machine learning methods for estimating heterogeneous causal effects. *Annual Review of Economics*, 04 2015.

Anna Baiardi and Andrea A Naghi. The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal*, page utae004, 2024.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

Janet Bouttell, Peter Craig, James Lewsey, Mark Robinson, and Frank Popham. Synthetic control methodology as a tool for evaluating population-level health interventions. *J Epidemiol Community Health*, 72(8):673–678, 2018.

George EP Box and Norman R Draper. *Empirical model-building and response surfaces.* John Wiley & Sons, 1987.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Elena Calegari, Antonella Rita Ferrara, Marzia Freo, and Aura Reggiani. The heterogeneous effect of european union cohesion policy on regional well-being. *European Urban and Regional Studies* (4):311–318, 2023.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993. URL `http://www.nber.org/papers/w4509`.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper 24678, National Bureau of Economic Research, June 2018b. URL `http://www.nber.org/papers/w24678`.

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*, 2024.

Riccardo Crescenzi and Mara Giua. One or many cohesion policies of the european union? on the differential economic impacts of cohesion policy across member states. *Regional Studies*, 54(1):10–20, 2020.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, September 2020.

Sergio Destefanis and Valter Di Giacinto. The heterogeneous effects of eu structural funds: A spatial var approach. *Journal of Regional Science*, 2024.

Desmond Dinan. *Ever closer union: an introduction to European integration*. Lynne Rienner Publishers, 2005.

Stillman Drake. *Galileo at work: His scientific biography*. Courier Corporation, 2003.

European Commission. Regional trends for growth and convergence in the european union, 2023. URL `https://ec.europa.eu/regional_policy/information-sources/publications/reports/2023/regional-trends-for-growth-and-convergence-in-the-european-union_en`.

European Commission. Cohesion policy benefits eu's economy and regions, 2024. URL `https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/cohesion-policy-benefits-eus-economy-and-regions-2024-04-11_en`.

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Prashant Garg and Thiemo Fetzer. Causal claims in economics. Discussion Paper 183, Institute for Replication (I4R), 2024. URL `https://hdl.handle.net/10419/306280`. I4R Discussion Paper Series, No. 183.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Mariassunta Giannetti. The effects of integration on regional disparities: Convergence, divergence or both? *European Economic Review*, 46(3):539–567, 2002.

Laurent Gobillon and Thierry Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, 98(3):535–551, 2016.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Sander Greenland and James M Robins. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6:1–9, 2009.

Ian Hacking. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2006.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

James J. Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1-2):239–267, 1985.

M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2020. ISBN 9781420076165.

Carl Hoefer. Causal determinism. 2003.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81 (396):945–960, 1986.

David Hume. An enquiry concerning human understanding and other writings. 2007.

Kosuke Imai and In Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63, 03 2019.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Matthew J Kisner. *Spinoza on human freedom: Reason, autonomy and the good life*. Cambridge University Press, 2011.

Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

Pierre-Simon Laplace. *Pierre-Simon Laplace philosophical essay on probabilities: translated from the fifth french edition of 1825 with notes by the translator*, volume 13. Springer Science & Business Media, 2012.

Michael Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4:165–224, 2011.

Michael Lechner. Causal machine learning and its use for public policy. *Swiss Journal of Economics and Statistics*, 159(1):8, 2023.

Robert Lehmann and Klaus Wohlrabe. Forecasting gross value-added at the regional level: are sectoral disaggregated predictions superior to direct ones? *Review of Regional Research*, 34:61–90, 2014. doi: 10.1007/s10037-013-0083-8.

Jesús Martín-Fernández, Ángel López-Nicolás, Juan Oliva-Moreno, Héctor Medina-Palomino, Elena Polentinos-Castro, and Gloria Ariza-Cardiel. Risk aversion, trust in institutions and contingent valuation of healthcare services: trying to explain the wta-wtp gap in the dutch population. *Cost Effectiveness and Resource Allocation*, 19(1):27, 2021.

Robert McClelland and Livia Mucciolo. An update on the synthetic control method as a tool to understand state policy. *Tax Policy Center*, 2022.

John Stuart Mill. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Longmans, green, and Company, 1916.

RP Miller. *René Descartes: Principles of Philosophy: Translated, with Explanatory Notes*, volume 24. Springer Science & Business Media, 1984.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.

Andrés Rodríguez-Pose. The revenge of the places that don't matter (and what to do about it). *Cambridge journal of regions, economy and society*, 11(1):189–209, 2018.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020.

N. Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279 – 281, 1948.

George Smith. Newton's philosophiae naturalis principia mathematica. 2007.

Robert M Solow. A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1):65–94, 1956.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Peter WG Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, et al. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2):620–632, 2021.

Hal R Varian. Big data: New tricks for econometrics. *Journal of economic perspectives*, 28(2):3–28, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

Jeffrey M Wooldridge. Introductory econometrics: A modern approach 6rd ed., 2016.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

7 Appendix

Table 11: List of Country Codes and Country Names

Country Code	Country Name
AT	Austria
BE	Belgium
BG	Bulgaria
CY	Cyprus
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
EL	Greece
ES	Spain
FI	Finland
FR	France
HR	Croatia
HU	Hungary
IT	Italy
LT	Lithuania
LU	Luxembourg
LV	Latvia
MT	Malta
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SE	Sweden
SI	Slovenia
SK	Slovakia

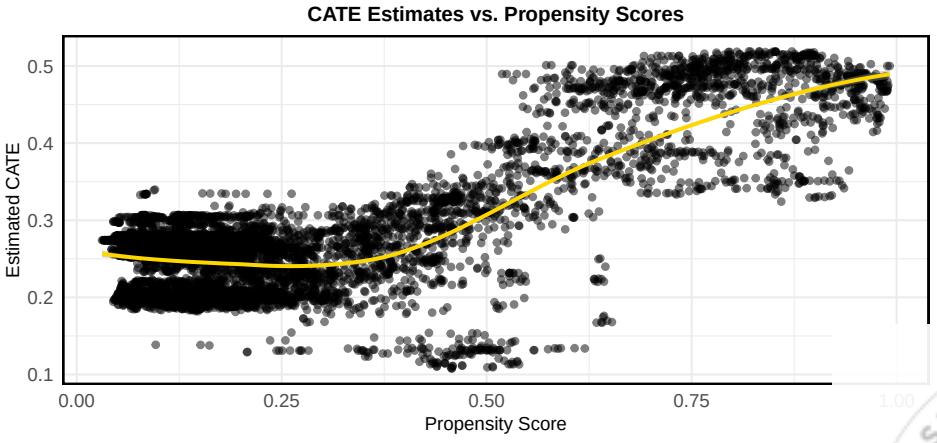

Figure 27: Conditional Average Treatment Effects estimates in low-propensity and high-propensity score regions

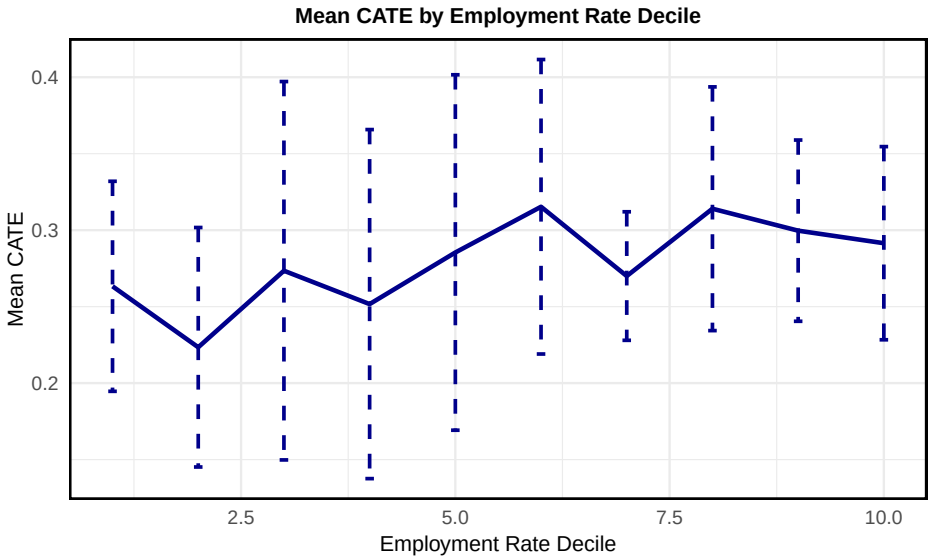

Figure 28: Conditional Average Treatment Effects Heterogeneity by Employment Rate Deciles

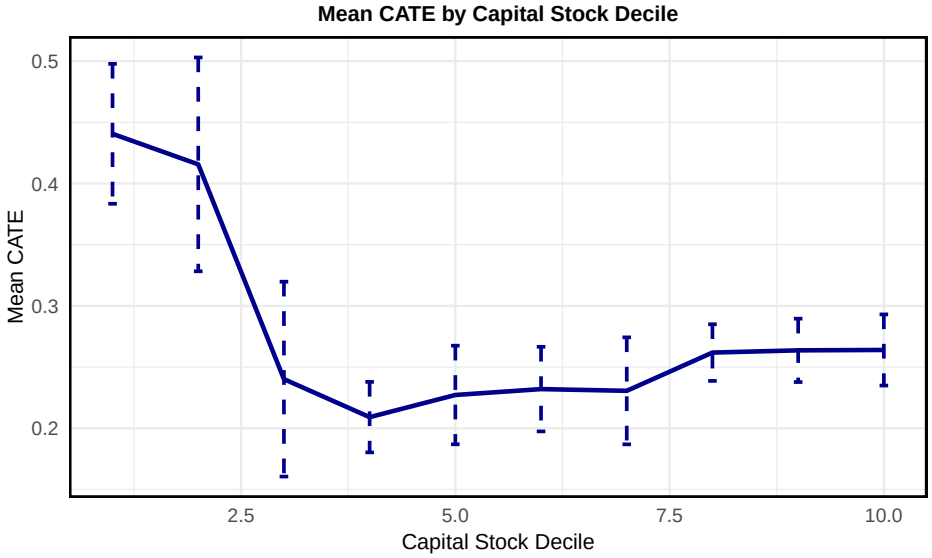

Figure 29: Conditional Average Treatment Effects Heterogeneity by Capital Stock Deciles

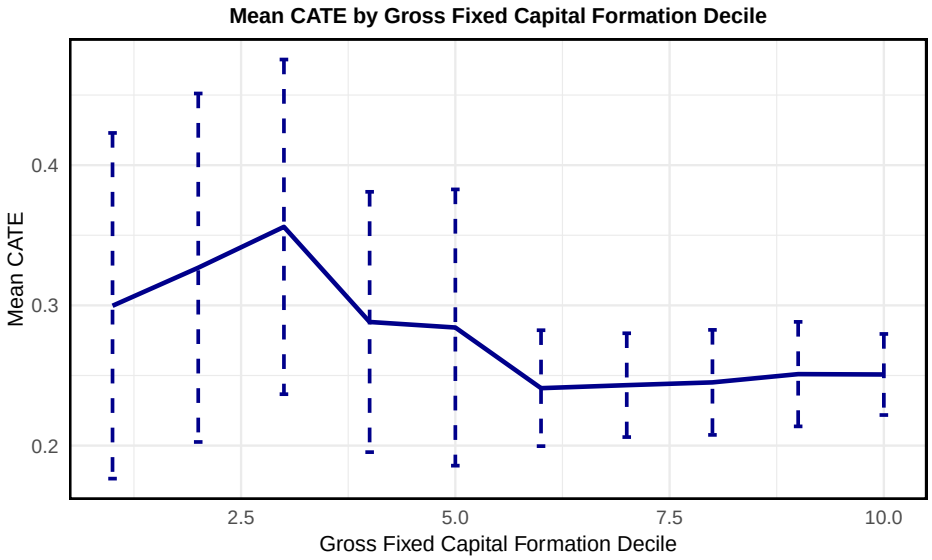

Figure 30: Conditional Average Treatment Effects Heterogeneity by Gross Fixed Capital Formation Deciles

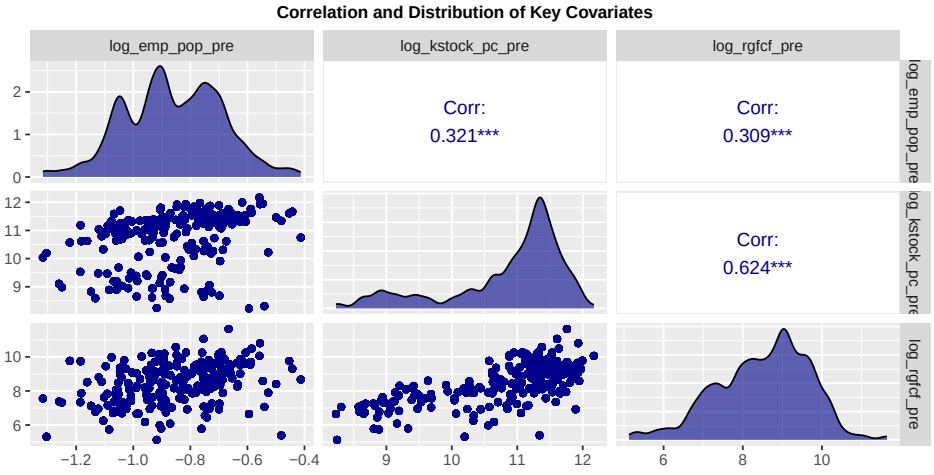

Figure 31: Correlation and Distribution of Covariates