UNIVERSITY
OF TRENTO

Doctoral School of Social Sciences

Doctoral programme in Economics and Management

Estimating causal impacts under complex conditions: Two applications in presence of multiple fixed effects and continuous multidimensional treatments

a dissertation submitted in partial fulfillment of the requirements for the Doctoral degree (Ph.D.) doctoral programme in Economics and Management

Enrico
Cristofoletti

Academic Year 2019-2020

Supervisors

– Prof. Roberto Gabriele, University of Trento, Italy

– Prof. Edoardo Gaffeo, University of Trento, Italy

Doctoral Committee

– Prof. Luciano Andreozzi, University of Trento, Italy

– Prof. Emanuele Taufer, University of Trento, Italy

– Prof. Oksana Tokarchuk, University of Trento, Italy

Acknowledgments

I thank my supervisors Edoardo Gaffeo and Roberto Gabriele. I am grateful to Edoardo Gaffeo for his optimism and support during the initial years of my Ph.D. I am grateful to Roberto Gabriele for believing in me, sharing ideas, and his tireless support even during the pandemic.

I thank Sujiao Zhao and all members of Banco de Portugal Microdata Research Laboratory. Without their work, the first chapter of the thesis would not have been possible.

I also thank my fellow students Austeja Kazemekaityte, Lucia Pederiva, and Valeria Fangella, who shared the ups and downs of the courses with me,

I am grateful to Giacomo Caterini, Davide Bernardini, and Vittorio Guida for the many stimulating conversations.

I thank Annalisa Caloffi and Giovanni Cerulli for their valuable comments on a previous version of the thesis.

I thank Fabio Zona, a person who will always make me feel extremely grateful and ungrateful.

I am grateful to my family for always being there when needed and my girlfriend for sharing her life with me.

Abstract

This thesis is a collection of three essays in causal evaluation.

The first chapter investigates the effects of formal ties between firms and banks on the amount of credit received. I focus on the micro-effects of ties (bank-firm level) and how they reverberate at the macro level. Results are consistent with the literature considering links as a source of favoritism. However, efficient firms are more likely to be connected to banks, thus benefiting more often than less efficient firms from connections. The comparison of Portugal's GDP in 2017 with that produced under a hypothetical scenario where every tie was severed shows that severing links results in virtually no changes in GDP. I interpret the result as evidence that the different likelihood of being connected experienced by efficient and not efficient firms counterbalances the misallocating potential of connections.

The second chapter introduces a novel Stata implementation of Egger and von Ehrlich's (2013) econometric framework for the estimation of treatment effect when the treatment is continuous and multidimensional. After the illustration of the package, I present a simple simulation to show the capability of the method to overcome bias.

The third chapter consists of an evaluation of European regional policy. It analyzes how different mixes of investments in infrastructure and productive investments affect regions' growth rate. The main results are that allocations in infrastructure foster growth only when coupled with expenditures in productive investments. Moreover, the highest growth is obtained when investments have high intensity in both dimensions. By generating two hypothetical scenarios, I investigate how the allocation of funds can be improved. The results show that regions could allocate more efficiently. However, the actual transfer intensity is not enough to choose the mix that would globally maximize growth. The findings are consistent with Becker et al. (2012) since enforcing common support restricts the analysis to regions with low transfer intensity.

Keywords: bank-firm ties, interlocking directorates, misallocation, continuous multiple treatments, policy evaluation, EU regional policy, regional growth.

Contents

List of Tables

Introduction

Finding the proper identification strategy for the focal research question or the evaluation of a policy is the characteristic task of every applied economist aiming to give a causal interpretation to her results. It includes employing the appropriate data sources and adopting techniques of analysis that are consistent with the object of the research and capable of accounting for potential confounding factors. This thesis is a collection of three essays concerning causal evaluation. It consists of two empirical essays (the first and the third chapters) and a methodological essay (the second chapter). The first chapter examines the effects of bank-firms connections on the amount of credit received by a firm. The second chapter introduces an implementation of a novel econometric framework for the estimation of causal impact when the treatment is continuous and multidimensional. The third chapter leverages the implementation developed in the second chapter for estimating the causal impact of the European regional policy focusing on different policy mixes.

The first chapter is entitled "The Importance of Being Connected: an Assessment of the Effect of Interlocking directorates on the Allocation of Resources". This essay is rooted in two theoretical strands. The first includes the literature concerning the positive effects of connections on a firm's availability of credit. The second consists of the literature underscoring the importance of misallocation in explaining the income per capita variations across countries. In particular, the essay aims to assess whether the positive effects of interlocking directorates (when a bank and a firm share an exponent) on the availability of credit results in a misallocation of capital. Answering appropriately to the research question requires high-quality data on credit at the bank-firm level. Bank of Portugal allows me to have surrogate access to data from Portugal Central Credit Register (Banco de Portugal's Central Credit Register). I further had access to data on firms' balance sheets (Central de Balancos). I instead retrieve data on exponents from ORBIS-Bureau van Dijk. To measure the firms' efficiency and aggregate the results, I rely on Dias et al.'s (2014) theoretical model.

The assessment of causal effects when the unit of analysis is dyadic is complicated by three sources of endogeneity: the nodes and the relationship itself.

Thanks to the data quality, the econometric models adopted accounts for both demand effects, supply effects, and the fixed effects of the relationship. In particular, I use the least square dummy variable estimator controlling for bank fixed effects and relationship lending measures (Sette and Gobbi 2015). I control for demand effect by adding a set of dummies for clusters defined as in Degryse et al. (2019).

Contrarily to the previous literature, the analysis does not focus solely on direct links between banks and firms. I broaden the view by taking into account the entire network generated by sharing exponents. In particular, I consider the effect of being closer to a bank on the amount of credit. I define closeness as the geodesic distance from the bank measured on the network of shared exponents.

The empirical results are consistent with the literature considering links as a source of favoritism: banks do not use connections as an additional screening device capable of improving allocation. However, I found that efficient firms are more likely to be closer to banks. Therefore, efficient firms are also those that benefit more likely from the favoritism due to the connections. I finally compare the actual GDP with that of a hypothetical scenario in which ties are removed. Not finding differences, I conclude that the differential in the probability of being connected with banks between efficient and non-efficient firms counterbalances the misallocating potential of connections in Portugal.

The second chapter is entitled "An implementation for the estimation of the dose-response function when the treatment is multidimensional: a Stata package". The chapter aims to illustrate Egger and von Ehrlich's (2013) econometric framework for estimating causal effect when the treatment is continuous and multidimensional as well as its Stata implementation. In its essence, the framework consists in a generalization of the propensity score approach. Propensity score methods are wildly used techniques for the evaluation of causal effects in observational studies.

While Rosenbaum and Rubin's (1983) original article focused solely on binary treatments, further studies generalize the methods to multi-valued treatments, continuous treatments, and multidimensional continuous treatments. Despite its potential, Stata offers plenty of packages for all the cases but the last one. It is a pity because often, as in the case of European Regional Policy, the treatment is appropriately seen as an aggregation of treatments that occurred over several dimensions. Its impact should, thus, be assessed by considering the complex interplay between its different dimensions.

This chapter illustrates the econometric framework, its proofs, and the commands implemented. I finally perform a simple simulation to show the commands and the method's capability to overcome bias by comparing the dose-response estimated with the propensity score method to those estimated with an OLS regression rightly specified according to the data generating process.

The last chapter is entitled "Choosing the right expenditure mix: An evaluation of the EU's regional policy using generalized propensity scores for multiple continuous treatments". This essay exploits the novel econometric framework introduced in the second chapter to investigate how different mixes of investments in infrastructure and productive investments impact regions' growth rate. Indeed, the evaluation of European regional policy has mainly focused on the overall effectiveness of the policy, thus neglecting the heterogeneous effects due to different policy mixes. I focus on the programming period 2007-2013. Data on Allocation are retrieved from European Commission-DG REGIO and consist in the package "Integrated database of allocations and expenditure for 2000-2006/2007–2013" which includes consolidated data at NUTS2 level for ERDF and CF allocations from 2000 and 2014. Data for the outcome (per capita GDP growth rate) and the propensity score are collected from the regional databases of Cambridge Econometrics and Eurostat.

Results depict a four-class typology based on the allocation intensity in the two dimensions. The main result is that allocation in infrastructure has a positive effect only if associated with productive investments. Moreover, I find that the maximal impact on growth is obtained by a policy allocating with high intensity in both dimensions. I finally assess whether regions allocate efficiently the funds received by generating two scenarios. In the first one, every region chooses the best mix available under the constraint of the actual funds received. In the second, each region was unconstrained. The scenarios enable me to investigate how the effectiveness of Regional policy can be improved. If regions could improve the allocation of the actual amount of financing, stricter controls on the policy implementation could help the policymaker choose the more worthy investments. Suppose instead that regions do not receive enough money to take advantage of Regional Policy maximally. In that case, it should be examined whether the distribution of financings between regions is consistent with the policy’s primary goals. The comparison with the actual allocation shows that, although the regions can allocate more efficiently, the observed transfer intensity is not enough to choose the mix that would globally maximize growth. Results are consistent with Becker et al. (2012) since enforcing common support restricts the analysis to regions with low transfer intensity.

References

Becker, Sascha O., Peter H. Egger, and Maximilian von Ehrlich. 2012. ‘Too Much of a Good Thing? On the Growth Effects of the EU’s Regional Policy’. *European Economic Review* 56 (4): 648–68.

Degryse, Hans, Olivier De Jonghe, Sanja Jakovljević, Klaas Mulier, and Glenn Schepens. 2019. ‘Identifying Credit Supply Shocks with Bank-Firm Data: Methods and Applications’. *Journal of Financial Intermediation*, April.

Dias, Daniel A., Christine Richmond, and Carlos Robalo Marques. 2014. ‘Misallocation and Productivity in The Lead Up to The Eurozone Crisis’. WORKING PAPERS Lisbon Banco de Portugal.

Egger, Peter H., and Maximilian von Ehrlich. 2013. ‘Generalized Propensity Scores for Multiple Continuous Treatment Variables’. *Economics Letters* 119 (1): 32–34.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. ‘The Central Role of the Propensity Score in Observational Studies for Causal Effects’. *Biometrika* 70 (1). Oxford University Press: 41–55.

Sette, Enrico, and Giorgio Gobbi. 2015. ‘Relationship Lending During a Financial Crisis’. *Journal of the European Economic Association* 13 (3): 453–81.

1 The importance of being Connected: An assessment of the effect of interlocking directorates on the allocation of resources[1]

Abstract

In the last few years, several studies have focused on the beneficial effects of interlocking directorates (e.g., when a bank and a firm share one or more exponent) in firms' credit availability. Although important from a managerial point of view, these studies leave unanswered questions about the macro-effects of formal ties. It is unclear whether bank-firm connections facilitate or are detrimental to the optimal allocation of financings between firms. This is an important issue since misallocation of funds penalizes efficient firms, thus dampening a country's overall growth. Using Portugal data, the present chapter consists of assessing whether the positive effects of interlocking directorates on credit availability result in a misallocation of capital. Empirical results are consistent with the literature considering links as a source of favoritism. However, I found that efficient firms are more likely to be connected with banks. In a final hypothetical exercise, I compare the actual GDP with that of a scenario in which ties were severed. Not finding differences, I conclude that non-efficient firms' higher distance from banks counterbalances the misallocating potential of connections.

Keywords: bank lending, interlocking directorates, bank-firm ties, shared exponents, misallocation, misallocation micro-determinants.

1.1 Introduction

In the last few years, there has been a renewed interest in the effect of formal (i.e., organizations sharing exponents)[2] and informal bank-firm ties (e.g., executives having attended the same school) on the availability of credit (Sisli-Ciamarra 2012; Engelberg, Gao, and Parsons 2012; M. A. Ferreira and Matos 2012; Barone, Mirenda, and Mocetti 2017; Braun, Briones, and Islas 2018; Karolyi 2018; Coin et al. 2011). Virtually every study has found a positive effect of these ties on the access to credit and the amount of credit received. However, it is not clear what implications this effect has on the misallocation of funds and how this effect reverberates on the economic system

[1] I am grateful to all members of Banco de Portugal Microdata Research Laboratory – BPLIM and in particular to Sujiao Zhao for making the data available and all the valuable suggestions. I also want to thank Domenico Delli Gatti for the help in retrieving data on exponents.

[2] A vast literature refers to formal ties as interlocking directorates — see e.g. Mizruchi (1996) and Ratcliff (1980). I will use interlocking directorates and formal bank-firm ties interchangeably in the chapter.

as a whole. The literature is not unanimous and reports findings and rationales coherent with both positive and negative effects. This is not a trivial issue since misallocation of financings can dampen efficient firms growth (Banerjee and Moll 2010; Schivardi, Sette, and Tabellini 2017) and have pernicious effects on the overall productivity of a country (Restuccia and Rogerson 2017; Restuccia and Rogerson 2013; Banerjee and Duflo 2005; Kalemli-Ozcan and Sorensen 2012). Thus, the present chapter aims to assess whether bank-firm ties are detrimental to the efficient allocation of resources and to estimate the consequence of this effect at the macro level. The chapter, therefore, contributes to two strands of literature: the one concerning banks-firms relationships in finance and the one concerning the micro-determinants of misallocation.

From an empirical perspective, investigating the misallocation of credit due to interlocks is not straightforward: it involves three main issues. Measuring misallocation requires measuring the marginal product of capital, which is not directly observable. Moreover, fine-grained data on firms, banks, and credit are needed to cope with the three primary sources of endogeneity: demand, supply effects, and what I call the fixed effects of the relationship between a bank and a firm. Finally, since the focus is on the macro consequences of misallocation, I need a model for aggregating the results obtained at the micro-level.

To cope with these issues, I exploit the richness of the Bank of Portugal central credit register. Balance sheet data come from Portugal Central Credit do Balancos, while data on exponents come from ORBIS Bureau Van Dijk. Moreover, I employ Dias et al.' (2014) model on misallocation both in the empirical analysis and in the final hypothetical exercise. Since the Dias et al.' model does not account for the extensive margin of credit, I focus on the intensive margin of credit in the chapter.

It is worth noticing that, contrarily to the extant literature on corporate ties in credit availability, I broaden the definition of ties considering not only direct links but also the effects generated by indirect connections. In particular, I will focus on the distance a firm has from a bank considering the entire network generated by shared exponents (Stefano Battiston, Weisbuch, and Bonabeau 2003; Heemskerk, Daolio, and Tomassini 2013).

Empirical results about the effect of ties on the amount of credit the firm borrows are consistent with the literature considering links as a source of favoritism.

Being connected with a bank increases the amount of credit received irrespective of the firm's efficiency. Nevertheless, since efficient firms show a higher probability of being connected with banks, the effects of ties on GDP remain unclear. I, therefore, compare the actual GDP with that of a counterfactual scenario where I severed every tie in the system. Removing links results in virtually no changes. I interpret the results as evidence that the different probability of connections with banks characterizing efficient and non-efficient firms counterbalances the misallocating effects of ties.

The remainder of this chapter is structured as follows: the second paragraph reviews the literature; the third describes the data; the fourth summarizes the theoretical model I used; the fifth reports the empirical results; the sixth describes the counterfactual exercise. The seventh concludes.

1.2 Related literature

In the last fifteen years, the literature on misallocation has gained momentum. Misallocation is deemed one of the most important causes of the different aggregate total factor productivity between countries (Hopenhayn 2014; Restuccia and Rogerson 2017). The logic is relatively straightforward, although, usually, this literature is quite mathematically demanding. In a world with perfect allocation, the marginal product of capital and labor would be equal between firms[3]. Instead, when there are some frictions, factors of production are not optimally allocated, and more efficient firms can grow less than their potential. As a consequence, misallocation lowers aggregate total factor productivity and thus country GDP.

Hsieh and Klenow (2009) measured the extent of misallocation in the U.S.A., India, and China in different years. They found that eliminating misallocation would increment the aggregate total factor productivity for the 86%-115% in China, 100%-128% in India, and 30%-43% in the United States (see also Restuccia and Rogerson (2008), Midrigan and Xu (2014) and Inklaar et al. (2017) for similar studies). Making the same reallocation exercise, Calligaris (2015) found that Italy would experience a TFP gain ranging from 58% and 80% over the period 1993-2011 (for the Italian case,

[3] The intuition is the following. A firm that would earn more than another from having more capital — that is: its marginal product of capital is higher — would also "pay" more than the other firm for an investment — if the firm maximizes, the marginal product of capital is equated to one plus the interest rate. The investor would thus choose to invest in the firm with higher marginal product until the marginal product of capitals — and thus the payable interest rate — were equalized across the whole set of firms.

see also Calligaris et al. (2018)). In Portugal, misallocation is deemed one of the leading causes of the slump between 2000 and 2007 (Reis 2013). Dias et al. (2014) reported evidence that between the years 1996 and 2011, gross output level would have been from 17% to 28% higher than the actual gross output level in the absence of distortions.

Bleck and Liu (2018), Borio et al. (2016) studied how credit expansion impacts on misallocation (see also Gopinath et al. (2017) and Daniel A. Dias et al. (2014)). In particular, Bleck and Liu (2018) found crowding out effects subsequently to a credit expansion, while Borio et al. (2016) found that the misallocation of labor following a credit expansion decreases productivity and dampens the recovery after a crisis. Kalemli-Ozcan et al. (2012), using African data, find evidence of the association between institutional factors, namely property rights, and misallocation.

Banerjee and Duflo (2005) Banerjee and Moll (2010) underlined the importance of financial friction in the misallocation of capital. In particular, financial frictions can affect efficient firms through two channels: reducing credit availability and enabling inefficient firms to compete in the product and input market (Banerjee and Moll 2010; Schivardi, Sette, and Tabellini 2017). Moreover, the increasing cost of entry can discourage new firms, even if they had a more efficient technology, from entering the market (Caballero, Hoshi, and Kashyap 2008). Misallocation thus affects both the growth rate of efficient firms ─ the intensive margin ─ and the surviving and replacing of the firm ─ the extensive margin.

Although the importance of misallocation in explaining country development is well established in the literature, it is not clear whether bank-firm ties have a role in generating inefficiencies.

The literature on relationship lending predicts that a lasting relationship between a firm and a bank increases the sharing of soft information and, thus, the availability of funds (Petersen and Rajan 1994; Berger and Udell 1995; Boot 2000; Ongenah and Smith 2000). By building on this literature, works on interlocks and informal bank-firms relations claims that the effect of bank-firm ties is due to the reduction in the informational asymmetry (M. A. Ferreira and Matos 2012; Engelberg, Gao, and Parsons 2012; Karolyi 2018; Coin et al. 2011)[4]. Following this strand of literature, a banal extension[5] of Tirole's barebones model of adverse selection (Tirole 2010) can show that, if relationships are independent of the borrowing type, and if banks financings are independent between each other, in a context without credit rationing, bank-firm ties increase the overall quality of credit in comparison to a world with complete informational asymmetry. If this were the case, relationships would positively affect the allocation of financings.

However, relationships are hardly random. Seminal managerial insights have shown that personal relationships can arise from the firm's desire to co-opt resources (Pfeffer and Salancik 1978; Stearns and Mizruchi 1993; Hillman and Dalziel 2003) and that firm's needs vary with their life stage (Bonn and Pettigrew 2009) and with changes in the environment (Hillman, Cannella, and Paetzold 2000). Again, peculiar features of the legal system can cause the self-selection of bankers on corporate boards (Kroszner and Strahan 2001). For example, several authors have found that, in the U.S.A., banks interlock more frequently with larger and more stable firms with low or medium volatility (Kroszner and Strahan 2001; Byrd and Mizruchi 2005; Sisli-Ciamarra 2012). Kroszner and Strahan (2001) claim that the reason is the U.S.A. doctrine of "lender liability". In the USA, when an interlocked bank is found guilty of abusing its board position or of using inside information to change its transactions to shift the risk on other creditors, it loses the priority of its debt claims in case of bankruptcy. In order to

[4] Sharing an exponent is considered an extension of relationship lending (Sisli-Ciamarra 2012). It, indeed, enables to limit the extent of both ex ante informational asymmetry and moral hazard.

[5] Consider a setting where there is a population of potential borrowers $\mathcal{P} = \alpha + \beta$ where α and β are the shares of the population with projects having respectively probability p and q of success. Assume $p > q$. There is competition between lenders so that they lend at the break even. To make the things easier all projects have positive NPV, then: $p(R) - I > q(R) - I > 0$, where R is the return of the project and I the investment needed. There is only one type of project and the entrepreneurs do not own funds. Lenders can finance two projects. The most efficient allocation is that consisting in financing two projects from the α share of the population. In the first setting there is total informational asymmetry. Then the lender probability of success is: $m = \alpha p + \beta q$ and the payoff of the borrower is $R_b = R - \frac{1}{m} I$. The two investments are assumed independent and thus the probability of the most efficient allocation is: $P(MEA_1) = \left(\frac{\alpha}{\alpha+\beta}\right)^2$. We can think at a context in which there are personal relationships as a context in which the banks know the possible borrower of the first financing, which is then paid as it worth, but not the second, which is then penalized by the adverse selection. If being known is independent from the borrowing type and assuming again independence between the two financings, then the probability of the most efficient allocation is $P(MEA_2) = P(MEA_1)$. However, the sum of the payoffs for the borrowers is now: $R - \frac{1}{p} I + R - \frac{1}{m} I > R - \frac{1}{m} I + R - \frac{1}{m} I$.

avoid possible charges, banks would interlock with firms where the risk of conflict between shareholders and creditors[6] is less pronounced (Kroszner and Strahan 2001).

Studies on the venture capital (VC) market provide further evidence of strategies behind bank-firm connections. Hellmann et al. (2008), using U.S.A. data, found that banks tend to invest more in high-debt industries in order to build a relationship with future clients. In a VC industry less populated by independent VC firms like Japan − where VC firms are usually subsidiary of a bank − banks interlocks with firms lacking financial expertise to support and invest in them (Takahashi 2015).

Another strand of literature argues that managers seek to be appointed in prestigious companies to improve their careers (Zajac 1988). Consequently, the most prestigious and efficient companies would show a higher centrality in the interlocking network (Wasserman and Faust 1994). The most prestigious firms would thus be more likely closer to banks, which are historically highly central (Davis, Yoo, and Baker 2003; Dagnes 2014; Heemskerk, Daolio, and Tomassini 2013).

Dubious random assignment aside, the effect of being linked itself could represent a potential source of misallocation. Bank-firm ties can, indeed, generate favoritism (Barone, Mirenda, and Mocetti 2017) or even be a source of looting (La Porta, Lopez-de-Silanes, and Zamarripa 2003). Therefore, the effect of ties would be that banks direct funds towards "friends" rather than worthy firms. In this way, connected but scarcely efficient firms could obtain financings they would not have received without having ties. Barone et al. (2017) found, using panel data with banks put under special administration by the supervisory authority, that losing a link increases the probability of default. They attribute this effect to the excessively favorable lending conditions obtained when banks and firms were connected (Barone, Mirenda, and Mocetti 2017). La Porta et al. found instead evidence of looting in Mexico, especially when the shared director is also the firm's owner (La Porta, Lopez-de-Silanes, and Zamarripa 2003). Similarly, by using Japanese data, Peek and Rosengren (2005) (see also Morck and Nakamura (1999)) reported evidence of a higher likelihood of evergreening practices — extensions of credit to non-viable firms — between members of the same keiretsu.

[6] Which is the conflict arising from the different risk preferences.

Another concern is that the shared exponent may enact in the interests of creditors rather than in the interest of the firm's shareholders (Güner, Malmendier, and Tate 2008; Dittmann, Maug, and Schneider 2010). More specifically, if the firm has a low risk of default, the director can push the firm to subscribe a loan even if it is not needed or finance bad investments. Güner et al. (2008), using data on publicly traded firms in the U.S.A., found evidence that having a commercial banker as a director decreases the investment-cash flow sensitivity and increases debt in non financially constrained firms even if the firm is less profitable and with worse investment opportunities. Dittmann et al. (2010) reported similar findings with data from Germany.

Summarizing: to the extent the probability of connections is biased in favor of under-the-average borrowers, an allocation with complete informational asymmetry should be more efficient; otherwise, the opposite is true. We will, moreover, observe misallocation if the ties involve favoritism and resources cooptation instead of better screening. The mix of these two mechanisms determines the overall misallocation.

In the rest of the chapter, before accounting for the combined effect the two mechanisms have on GDP, I will investigate the two mechanisms separately. At first, I will estimate the impact of bank-firm distance on the intensive margin of credit. A full set of interactions will enable me to assess whether efficient and non-efficient firms benefit from links differently. Then, I will estimate the probability of a connection in order to determine whether there are some differences between efficient and non-efficient firms. Finally, I will provide an assessment of the impact the two mechanisms together have on national GDP. In particular, by using the model elaborated by Dias et al. (2014) and the results obtained during the empirical estimations, I will compare the actual situation with a scenario in which every link between banks and firms is severed.

Differently from most of the previous literature, which only focuses on the direct link between firms and banks, I will assume a broader perspective by taking into account the entire network generated by sharing exponents (Heemskerk, Daolio, and Tomassini 2013; S. Battiston and Catanzaro 2004; Vitali, Glattfelder, and Battiston 2011; Stefano Battiston 2004; Davis, Yoo, and Baker 2003). In fact, I believe that information can flow over the network (Stefano Battiston, Weisbuch, and Bonabeau 2003; Granovetter 2005). Therefore, the true effect of exponent ties can be assessed only by broadening the analysis beyond the bank-firm dyads. The chapter will focus on a firm's distance from a bank measured over the network generated by exponent sharing. More technically, I will consider the geodesic distances between banks and firms in the network. Exponents, interacting, share information about people and practices (Davis 1991; Davis and Greve 1997; Haunschild and Beckman 1998). Thus, links can be valuable tools for reducing the transaction cost of gathering soft information (Liberti and Petersen 2019). The implication is that, by leveraging interlocks (asking for information from connected exponents), banks can collect information about closer firms more easily than from distant ones. In this case, closeness in the interlocking network works similarly to geographical proximity (Degryse and Ongena 2005; Degryse, Kim, and Ongena 2009; Berger et al. 2005; Petersen and Rajan 2002). However, proximity with a firm can also increase the probability of the "basking-in-reflect-glory" (Kilduff and Krackhardt 1994) when the bank evaluates a shared connection with a firm as a signal of positive quality. Moreover, closeness increases the probability of a firm being perceived as a "friend" (or a "friend of a friend"), thus possibly causing favoritism (Barone, Mirenda, and Mocetti 2016).

1.3 Efficiency: the model and measurement

An efficient allocation of financings implies lending to the firms with the highest marginal revenue return of capital, which is not a quantity directly observable from balance sheet data. Therefore, in order to measure the marginal revenue return of capital for each firm, I rely on the model of Dias et al. (2014)[7]. Since both the empirical part and the construction of the counterfactual scenario where connections are removed employ their work, this section provides a brief sketch of the model. It is worth noting that this model enables us to focus on the within-industry misallocation and not on the inter-industry one.

[7] Another way to proceed could have been to estimate efficiency by Data envelopment analysis (DEA)(Cooper, Seiford, and Tone (auth.) 2006). However, I tried several packages (Badunenko and Mozharovskyi 2016; Ji and Lee 2010), but the sample size made the computation infeasible. Moreover, model assumptions enable me to correct the fact that the observed measure of factors is endogenous to the misallocation in the system. Corrections would not have been possible within a DEA approach. A further issue would have been the aggregation of the results. For these reasons, I decided to follow a more "model-based" approach. Nevertheless, it is worth noting that the approach I followed in the chapter strongly relies on model assumptions (e.g., Cobb Douglas, perfect competition…), and this could have implications for the results obtained.

The system includes S industries. Moreover, it is assumed to produce a single final good Y in a competitive market with an aggregate production function that takes the form of a Cobb Douglas with constant return to scale:

$$Y = \prod_{s=1}^{S} Y_s^{\theta_s} ; \sum_{s=1}^{S} \theta_s = 1 \tag{1}$$

Where Y_s the gross output in industry s. Cost minimization and the assumption of competitive markets implies that Cobb Douglas exponents are equal to the factor shares[8]:

$$\theta_s = \frac{P_s Y_s}{PY} \tag{2}$$

The final good is assumed to be the nummeraire and we set $P = 1$. At the industry level, the gross output is produced by the following CES production function:

$$Y_s = \left(\sum_{i=1}^{M_s} (Y_{si})^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \tag{3}$$

Where Y_{si} is the gross output of firm i in sector s, M_s the total number of firms in the industry, and σ is the elasticity of substitution. The assumption of free entry and monopolistic competition determine the following inverse demand function for firm i in the in industry s (Brakman, Garretsen, and Van Marrewijk 2001):

$$P_{si} = Y_s^{\frac{1}{\sigma}} P_s (Y_{si})^{-1/\sigma} \tag{4}$$

[8] I calculate these coefficients by summing the turnover in the system and by summing the turnover within the sectors (not the value added since the system produces by using the total output of every industry). I use data from Central Credit do Balancos.

We assume that the number of firms in a sector is so large that $\frac{\partial Y_s}{\partial Y_{si}} \cong 0$. As a consequence, $Y_s^{\frac{1}{\sigma}} P_s = \gamma_s$ appears in the maximization simply as a scale factor that is common for each firm in a sector.

The single firm produces using a Cobb Douglas technology with constant return to scale:

$$Y_{si} = A_{si}\, K_{si}^{a_s}\, H_{si}^{b_s}\, Q_{si}^{1-a_s-b_s} \tag{5}$$

Where K_{si}, H_{si}, Q_{si} are respectively capital, labor, and intermediate products of firm i in sector s. As customary in the literature on misallocation (e.g., Hsieh and Klenow (2009)), the profit function is defined as:

$$\pi_{si} = \left(1 - \tau_{Y_{si}}\right) P_{si} Y_{si} - \left(1 + \tau_{K_{si}}\right) R_s K_{si} - \left(1 + \tau_{H_{si}}\right) W_s H_{si} - Z_s Q_{si} \tag{6}$$

Where the various $\tau_.$ are the wedges that determine the misallocation. They represent any distortion present in the economy at a given time. $\tau_{Y_{si}}$ represents any distortion on output, $\tau_{K_{si}}$ the distortions on capital and $\tau_{H_{si}}$ includes labor distortions. As an example for $\tau_{Y_{si}}$, we can think of a heterogeneous taxation on revenues while an example for $\tau_{H_{si}}$ can be a tax on labor costs. It is worth noting that wedges are specific to a single firm. Sharing an exponent with a bank will work through these wedges – more precisely through the wedge on capital $\tau_{K_{si}}$ – by increasing or decreasing the firm's cost of capital. In particular, $\tau_{K_{si}}$ will be high for more opaque firms, and it will reduce if banks can more easily get access to soft information through common ties or because of "basking-in-reflect-glory". Similarly, closer firms can see the cost of capital reduced because perceived as "friends", and this is modeled as a $\tau_{K_{si}}$ lower than similar but distant firms.

Profit maximization with the assumption on the inverse demand function leads to the conditions:

$$\left(1+\tau_{K_{si}}\right) = \frac{a_s}{1-a_s-b_s}\frac{Z_sQ_{si}}{R_sK_{si}} \tag{7}$$

$$\left(1-\tau_{H_{si}}\right) = \frac{b_s}{1-a_s-b_s}\frac{Z_sQ_{si}}{W_sH_{si}} \tag{8}$$

$$\left(1-\tau_{Y_{si}}\right) = \frac{\sigma}{\sigma-1}\frac{1}{(1-a_s-b_s)}\frac{Z_sQ_{si}}{P_{si}Y_{si}} \tag{9}$$

The marginal revenue return of capital for a single firm is, finally, obtained as follows. Let the revenue be $P_{si}Y_{si}$, then, by (4) and the assumption on the form of firm's production function:

$$P_{si}Y_{si} = \gamma_s(Y_{si})^{-\frac{1}{\sigma}}Y_{si} = \gamma_s(Y_{si})^{1-\frac{1}{\sigma}} = \gamma_s(Y_{si})^{\frac{\sigma-1}{\sigma}} = \gamma_s\left(A_{si}K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}\right)^{\frac{\sigma-1}{\sigma}}$$

If we take the derivatives with respect to capital:

$$\frac{\partial P_{si}Y_{si}}{\partial K_{si}} = \gamma_s\frac{\sigma-1}{\sigma}\left(A_{si}K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}\right)^{\frac{-1}{\sigma}}\alpha_sK_{si}^{\alpha_s-1}A_{si}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s} \tag{10}$$

Unfortunately, since we do not observe the firm's gross output, we cannot directly use this formula. We need to exploit the assumptions. Noting that by (4), $Y_{si} = \left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}$, by defining the total factor productivity quantity as $TFPQ_{si} = A_{si} = \frac{Y_{si}}{K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}}$, (10) becomes:

$$\frac{\partial P_{si}Y_{si}}{\partial K_{si}} \tag{11}$$

$$= \frac{\sigma - 1}{\sigma}\left(\frac{\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}}{K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}}K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}\right)^{\frac{-1}{\sigma}} \alpha_s K_{si}^{\alpha_s-1}\frac{\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}}{K_{si}^{\alpha_s}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta_s}}H_{si}^{\beta_s}Q_{si}^{1-\alpha_s-\beta}$$

$$= \alpha_s\frac{\sigma - 1}{\sigma}\left(\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}\right)^{\frac{-1}{\sigma}}\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}\frac{K_{si}^{\alpha_s-1}}{K_{si}^{\alpha_s}}$$

$$= \alpha_s\frac{\sigma - 1}{\sigma}\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{-1}{\sigma-1}}\left(\frac{P_{si}Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}K_{si}^{-1} = \alpha_s\frac{\sigma - 1}{\sigma}\frac{P_{si}Y_{si}}{\gamma_s K_{si}} = \frac{\alpha_s(\sigma - 1)P_{si}Y_{si}}{\sigma\gamma_s K_{si}}$$

This quantity can be observed since the balance sheet contains the revenue $P_{si}Y_{si}$. We do not observe γ_s [9]. Therefore, within the two-digit sector[10], I normalize the marginal revenue of capital for the minimum marginal revenue. The normalization is not a problem since I am interested in the misallocation within industries.

1.4 Data

Although some studies have been based entirely on balance sheet data (Stearns and Mizruchi 1993), a robust analysis of the effects of formal ties requires data on single loans and information on who lends to whom. This richness is contained only in data provided by national credit bureaus.

Consequently, I use Data from the Bank of Portugal central credit register. It is very fine-grained and includes every exposure of an amount higher than 50 euros. Although during the last decade Portugal has experienced a process of deleveraging (Bank of Portugal 2018), it is a good laboratory because of the importance of the banking system in the allocation of credit (M. Ferreira, Lacerda, and Antão 2011; Bonfim, Dai, and Franco 2009). Balance data comes from the database of Central Credit do Balancos (Sforza 2017). This dataset covers the entire population of non-financial firms in Portugal.

[9] γ_s is the reason why we will be able to focus only on within sector misallocation.

[10] NACE rev2 classification.

I retrieve data on boards from ORBIS database of Bureau Van Dijk. ORBIS has often been considered one of the best sources for research concerning corporate ties (Heemskerk et al. 2016; Garcia-Bernardo and Takes 2018). Indeed, it enables worldwide comparisons and to map interlocking ties extensively. Unfortunately, exponent data are not longitudinal in ORBIS and, when they are, they are not reliable. ORBIS constantly updates the information in the database, and, thus, data on exponents are updated at the date of retrieval. I collected data from ORBIS in January 2019. Since the primary source of data on exponents in ORBIS is the annual report[11], I assume that the information on exponents concerns the years 2017-2018.

It is worth noting that the sample includes all the exponents (not only board members) detected by ORBIS. The rationale is that if a person is linked to a firm in ORBIS, she should have a prestigious position within that firm/bank. The prestigious position implies that she can influence the behavior of the firm/bank to which she belongs. I use data on firms, banks, and exponents to generate a network with banks and firms as nodes.

Table 1-1 Geodesic distance by macro-sector

	Agriculture	Manufacturing	Services
	Freq	Freq	Freq
Distance	(Percent)	(Percent)	(Percent)
0	597,371	2.632e+06	6.161e+06
	(77.64)	(86.07)	(90.57)
1	395	947	1,483
	(0.0513)	(0.0310)	(0.0218)
2	6,308	17,655	22,267
	(0.820)	(0.577)	(0.327)
3	33,804	107,773	131,418
	(4.393)	(3.524)	(1.932)
4	66,771	170,433	230,951
	(8.678)	(5.573)	(3.395)
5	47,682	98,294	177,500
	(6.197)	(3.214)	(2.609)
6	17,123	31,048	77,593
	(2.225)	(1.015)	(1.141)
Total	769454	3.058e+06	6.802e+06

[11] There is also an update based on firms' internet websites but it is common only for listed firms that are the minority. Moreover, it is not reasonable to assume that an exponent in January 2019 was not known in 2017 by the firm.

Table 1-2 Mean Credit by macro-sector

	Agriculture	Manufacturing	Services
	Mean	Mean	Mean
Credit	(Freq)	(Freq)	(Freq)
	749.4	1,982	632.0
	(769,454)	(3.058e+06)	(6.802e+06)

Table 1-3 Mean credit by geodesic distance

	Distance 1	Distance 2	Distance 3	Distance 4	Distance 5	Distance 6	Distance ∞
	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Credit	(Freq)	(Freq)	(Freq)	(Freq)	(Freq)	(Freq)	(Freq)
	79,003	33,157	11,503	4,660	1,956	1,093	329.1
	(2,825)	(46,230)	(272,995)	(468,155)	(323,476)	(125,764)	(9.391e+06)

Firms and banks are connected in the network if they share at least an exponent. By using this dataset, for each firm, I calculate the minimum distance – the geodesic distance – from every bank[12].

Given ORBIS data limitations, I keep only the firms that were active at least in the years 2016 and 2017. In addition, I set to infinite every distance higher than six. The reason is twofold. On the one hand, it is scarcely reasonable to find an effect with higher distances, and the inclusion would then be redundant. On the other hand, and most important, measurement error can increase with distance[13].

Following Dias et al. (2014), I focus only on those observations that belong to the macro sectors: Agriculture, Manufacturing, Services[14]. The final sample includes 89328 firms and 119 Banks. ORBIS coverage is satisfactory. My check shows that banks and firms included in the ORBIS database account for 85% of the total lending in

[12] I have used the R package igraph (Csardi and Nepusz 2006)

[13] Assuming a non zero probability of measuring a tie with errors, longer path are more likely to include errors.

[14] Agriculture includes NACE REV 2 two digits sectors: 01-09. Manufacturing includes NACE REV 2 two digits sector: 10-33. Services includes NACE REV 2 two digits sector: 35-43, 49-63, 68-82, 90-96.

2017[15]. The final dataset consists of the 10,630,032 bank-firm couples generated by the Cartesian product of the banks and the firms in the sample.

As expected, most couples have infinite geodesic distance (which is recoded to 0) (Table 1-1). The service sector appears to be the macro-sector closer to the banks. However, in percentage terms, viable paths are more likely in Agriculture. This is probably due to the extended geographical coverage of the Crédito Agrícola Group, which historically pays particular attention to agriculture (Stefanelli 2010).

Regarding the credit in 2017[16] (Table 1-2), we observe a higher mean for the manufacturing sector while Agriculture and Services behave similarly. The mean of the credit received significantly decreases with distance. It goes from 72,003 euros for couples at a distance of one to 329.1 euros for couples that are not connected (Table 1-3).

I also strictly follow Dias et al. (2014) in measuring the quantities for estimating the marginal revenue return of capital (margrev2dig). I measure capital, K_{si}, as the book value of the total capital stock net of depreciation. I set the elasticity of substitution σ to 3. Labor, H_{si}, is measured as the total labor cost experienced by the firm. This means to assume that wages per worker adjust for single worker skills and hour worked (i.e. $W_s = 1$). Similarly, I measure firm's intermediate product, Q_{si}, as expenditure on intermediate products. For robustness, I also measure H_{si} as the firm's number of employees. Data are retrieved from Central Credit do Balancos.

Cobb Douglas parameters could have been estimated by sector using a constrained regression on the logarithm of the production function (Coelli et al. 2005, 220):

$$\ln(P_{si}Y_{si}) = \ln(P_{si}A_{si}\, K_{si}^{a_s}\;\, H_{si}^{b_s}\;\, Q_{si}^{1-a_s-b_s})$$

[15] Despite its reasonable coverage, ORBIS does not include every firm in Portugal. However, ORBIS tends to overrepresent the larger firms (Garcia-Bernardo and Takes 2018), which are also those that interlock more likely. Therefore, although I am not looking at the complete network, the error in the estimation should be limited.

[16] When I talk about credit, I refer to the total available credit a firm can access. It is given by the sum of regular (effective debt e.g. loans for the acquisition of financial instruments) and potential credit (e.g. lines of credit).

However, this method would not have considered that the observed measure of factors is endogenous to the misallocation in the system. Consequently, as in Dias et al. (2014) and Hsieh e Klenow (2009), I calculate the industry level factor shares[17] using data from the United States, which is considered a relatively undistorted economy.

I retrieve data from statistics of the Bureau of Economic Analysis (BEA)[18]. The concordance between NAICS 2017 and NACE rev2 has been done by following the file for the conversion provided from Eurostat[19]. The transformation is not always one to one, and, thus, for each NACE rev2 industry with multiple matches, I calculated the factor shares as the mean of the factor shares of the corresponding NAICS industries. For each sector, the final Cobb Douglas parameters are recovered by averaging the factor shares in the years from 2008 to 2016.

1.5 Results

1.5.1 The intensive margin

Table 1-4 Variable descriptives Agriculture

	mean	sd	count
Total credit in 2017	749.381	30188.69	769454
Log(Total credit in 2017)	9.666052	2.590251	5227
DicoCredit2017*	.0067931	.0821401	769454
distance	.939179	1.812572	769454
margrev2dig	10324.37	48858.11	769454
Equity / Total assets	.2663464	.9116309	769454
RelDummy_ij	.0133393	.1147232	769454
RelAmount_ij	.0064305	.0721595	769454
Distancereachable**	4.199456	.9995265	172083
Notevenreachable***	.223643	.4166858	769454
N	769454		

*A dummy equal 1 if the firm borrows from the bank in the couple
**This is the mean distance calculated only on those observations that have a path
***This is a dummy equal one if the bank in the couple is reachable and 0 otherwise

[17] Labor share: labor expenses/gross output. Intermediate factor share: intermediate inputs/gross output. Capital share: 1- Labor share- Intermediate factor share.

[18] https://www.bea.gov/ in particular https://apps.bea.gov/iTable/iTable.cfm?ReqID=51&step=1

[19] https://ec.europa.eu/eurostat/ramon/relations/index.cfm?TargetUrl=LST_REL&StrLanguageCode=EN&IntCurrentPage=11

Table 1-5 Variable descriptives Manufacturing

	mean	sd	count
Total credit in 2017	1982.405	120597.1	3058419
Log(Total credit in 2017)	9.720319	2.837247	29246
DicoCredit2017*	.0095625	.0973192	3058419
distance	.5620767	1.444007	3058419
margrev2dig	2027.14	3812.08	3058419
Equity / Total assets	.1418812	1.547398	3058419
RelDummy_ij	.0177844	.1321668	3058419
RelAmount_ij	.0068594	.071333	3058419
Distancereachable**	4.033946	.9795658	426150
Notevenreachable***	.1393367	.3462976	3058419
N	3058419		

*A dummy equal 1 if the firm borrows from the bank in the couple

**This is the mean distance calculated only on those observation that has a path

***This is a dummy equal one if the bank in the couple is reachable and 0 otherwise

Table 1-6 Variables descriptives Services

	mean	sd	count
Total credit in 2017	631.9713	98812.06	6802159
Log(Total credit in 2017)	8.559341	2.778356	38228
DicoCredit2017*	.00562	.0747556	6802159
distance	.3994516	1.278239	6802159
margrev2dig	11593.53	39295.84	6802159
Equity / Total assets	-.5125745	3.761023	6802159
RelDummy_ij	.01123	.1053748	6802159
RelAmount_ij	.0059757	.0709689	6802159
Distancereachable**	4.237496	1.033987	641212
Notevenreachable***	.094266	.2921984	6802159
N	6802159		

*A dummy equal 1 if the firm borrows from the bank in the couple

**This is the mean distance calculated only on those observation that has a path

***This is a dummy equal one if the bank in the couple is reachable and 0 otherwise

The first part of the empirical analysis assesses whether efficient and non-efficient firms differently benefit from the distance from banks. I run the following model to estimate the intensive margin:

$$
\begin{aligned}
&ln\left(Borrow_{ij}^{2017}\right) \\
&\qquad = \alpha + \sum_{l=1}^{\max dist} \beta_{1l} Dis_{ij}^{l} \ + \beta_2 EFF_i \ + \sum_{l=1}^{\max dist} \beta_{3l}(Dis_{ij}^{l} * EFF_i) \\
&\qquad + \beta_4 DEGRYSE_i + \beta_5 BA_j \ + \beta_6 RelDummy_{ij} \\
&\qquad + \beta_7 RelAmount_{ij} \ + \beta_8 \frac{Equity}{TotalAsset} + \varepsilon_{ij}
\end{aligned} \tag{12}
$$

Where $Borrow_{ij}^{2017}$ is the new credit the firm i obtained from bank j in 2017 [20]. Dis_{ij}^{l} is a dichotomic variable equal to 1 if the geodesic distance between i and j is equal i and 0 otherwise. The geodesic distance has been calculated on the network generated between banks and firms when they share an exponent. In particular, a bank or a firm is linked to another bank or a firm if there is an exponent in common. In the model, I set the reference category at the infinite distance (recoded at 0). I choose to dichotomize distances to avoid any a priori assumptions about the relationship between distance and intensive margin of credit.

EFF_i is the marginal revenue return of capital of the firm (margrev2dig). I added a full set of interactions with the dummies for the distance to account for the possible different effects of being linked between efficient and non-efficient firms. Misallocation can indeed arise from the fact that efficient and non-efficient firms similarly benefit from the connections. If it is the case, we will observe positive and significant coefficients for dummy variables indicating distance but not significant coefficients for the interactions. Otherwise, if connections help reduce informational asymmetry, we should expect positive coefficients for the interactions between the distance and the firm's efficiency.

[20] The outcome $Borrow_{ij}^{2017}$ is calculated for the 2017 by exploiting the monthly frequency of the Central credit register in the following way: $Borrow_{ij}^{2017} = \sum_{t=1}^{12} Borrow_{ijt}^{2017} - Borrow_{ij(t-1)}^{2017}$. where t is the month and $Borrow_{ij0}^{2017} \equiv Borrow_{ij12}^{2016}$.

As anticipated in the introduction, this estimate could be flawed by three factors: supply effect, demand effect, and the fixed effects of the relationship between a bank and a firm.

In order to control for the supply effect, I add bank fixed effect BA_j (Schivardi, Sette, and Tabellini 2017). This control has the form of a set of dummy variables (one for each bank). In order to control for the demand effect, I follow Degryse (2019), who finds that firm's demand is homogeneous within clusters formed by the triple (ILS): industry (firm's two digits Nace Code), location (firm's two digits postal code), and size (firm's deciles of total asset). Therefore, I add the set of "cluster" dummies to the model[21].

A problem in proxying demand effects with ILS instead of controlling for firms' fixed effect is that it does not prevent the endogeneity due to the firm's leverage. Leverage could be correlated with connections because, when lending to highly leveraged firms, the bank may want to exert more control over the borrower. Appointing a trusted exponent could be a way to increase the control over the borrowing firm. In order to avoid this potential flaw, I also control for the equity over total asset.

Since it is plausible that having a lending relationship increases the likelihood of interlocking, I control the following dimensions of relationship lending (Sette and Gobbi 2015; Degryse, Kim, and Ongena 2009): the duration of the relationship, the share of credit granted by the bank. I use the indicators suggested by Sette and Gobbi (2015) and Bharath (2009). $RelDummy_{ij}$ is a dummy equal one if the amount of loan (total credit) from bank j to firm i in the last 5 years is higher than 0 while $Rel(Amount)_{ij}$ is the fraction of the total lending firm i has obtained from bank j in the last 5 years.

I used the BACON algorithm (Billor, Hadi, and Velleman 2000; Weber 2010) on the covariates and the outcome to detect multivariate outliers. The algorithm works as follows:1) it starts by selecting a subset of observations; 2) by using this subset, it estimates the mean and the covariance matrix of the variables in the model; 3) over the total sample, it calculates the Mahalanobis distance; 4) it increases the initial subset

[21] Given the large amount of fixed effect, I used the Stata package REGHDFE for the estimation (Correia, 2015; Correia 2016).

with the observations having a distance lower than a certain threshold[22]. Then it starts again from the beginning until the subset does not grow anymore. What remains outside the subset is considered an outlier. I used the 15% threshold parameters without finding any outlier.

As extant literature suggests (Schivardi, Sette, and Tabellini 2017; Accetturo et al. 2020), confidence intervals have been double clustered (Colin Cameron and Miller 2015) at the bank and firm level.

In order to detect possible effects specific to the macro-sector, for both the analysis on the probability of a link and on the extensive margin, I run separate models for macro sectors Agriculture, Manufacturing, Services.

[22] This threshold is defined by the threshold parameter. Every iteration we keep the observations that are in the part of the distribution that have a probability (1-threshold parameter) to happen.

Table 1-7 Intensive margin Agriculture

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
1.distance	0.771	0.804	0.771	0.771	0.758	0.761
	(0.955)	(0.994)	(0.955)	(0.988)	(0.966)	(0.997)
2.distance	0.347	0.341	0.347	0.306	0.336	0.297
	(0.369)	(0.376)	(0.369)	(0.373)	(0.369)	(0.374)
3.distance	-0.253	-0.240	-0.253	-0.252	-0.261	-0.258
	(0.259)	(0.248)	(0.259)	(0.241)	(0.257)	(0.240)
4.distance	-0.191	-0.160	-0.191	-0.194	-0.199	-0.200
	(0.159)	(0.156)	(0.157)	(0.153)	(0.156)	(0.151)
5.distance	-0.343	-0.400	-0.343	-0.418	-0.338	-0.413
	(0.320)	(0.330)	(0.320)	(0.339)	(0.323)	(0.341)
6.distance	-0.145	-0.280	-0.145	-0.316	-0.120	-0.293
	(0.366)	(0.374)	(0.366)	(0.368)	(0.367)	(0.368)
margrev2dig	-1.06e-07	4.78e-08	-1.05e-07	3.39e-08	2.62e-08	1.43e-07
	(4.58e-07)	(4.66e-07)	(4.58e-07)	(4.77e-07)	(6.78e-07)	(6.58e-07)
1.distance*	-7.26e-05	-6.59e-05	-7.24e-05	-7.16e-05	-7.14e-05	-7.07e-05
margrev2dig	(0.000158)	(0.000159)	(0.000158)	(0.000161)	(0.000159)	(0.000162)
2.distance*	1.94e-05	1.67e-05	1.94e-05	1.69e-05	1.94e-05	1.70e-05
margrev2dig	(1.24e-05)	(1.18e-05)	(1.23e-05)	(1.27e-05)	(1.25e-05)	(1.29e-05)
3.distance*	6.77e-05***	6.90e-05***	6.77e-05***	7.18e-05***	6.88e-05***	7.26e-05***
margrev2dig	(1.76e-05)	(1.79e-05)	(1.78e-05)	(1.84e-05)	(1.74e-05)	(1.82e-05)
4.distance*	3.72e-05***	3.54e-05**	3.71e-05***	3.82e-05***	3.81e-05***	3.91e-05***
margrev2dig	(1.40e-05)	(1.40e-05)	(1.40e-05)	(1.40e-05)	(1.39e-05)	(1.39e-05)
5.distance*	4.19e-05**	4.54e-05**	4.19e-05**	4.93e-05**	4.21e-05**	4.93e-05**
margrev2dig	(2.07e-05)	(2.06e-05)	(2.07e-05)	(2.07e-05)	(2.07e-05)	(2.08e-05)
6.distance*	7.08e-06	1.42e-05	7.13e-06	1.29e-05	6.37e-06	1.23e-05
margrev2dig	(2.20e-05)	(2.32e-05)	(2.20e-05)	(2.47e-05)	(2.22e-05)	(2.48e-05)
RelDummy_ij		-1.011***		-1.354***		-1.337***
		(0.0925)		(0.120)		(0.118)
RelAmount_ij			-0.00839	0.620***		0.613***
			(0.120)	(0.149)		(0.149)
Equity /					0.211**	0.177**
Total assets					(0.0857)	(0.0673)
Constant	9.659***	10.52***	9.662***	10.55***	9.594***	10.48***
	(0.0344)	(0.0812)	(0.0614)	(0.0801)	(0.0522)	(0.0778)
Observations	5,137	5,137	5,137	5,137	5,137	5,137
R-squared	0.232	0.247	0.232	0.252	0.234	0.253
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1-8 Intensive margin Manufacturing

VARIABLES	(1) MODEL 1	(2) MODEL 2	(3) MODEL 3	(4) MODEL 4	(5) MODEL 5	(6) MODEL 6
1.distance	-0.0669	-0.0717	-0.0657	-0.0642	-0.0588	-0.0579
	(0.424)	(0.441)	(0.423)	(0.446)	(0.422)	(0.444)
2.distance	0.379***	0.384***	0.380***	0.394***	0.383***	0.397***
	(0.127)	(0.128)	(0.128)	(0.131)	(0.127)	(0.132)
3.distance	0.361***	0.346***	0.363***	0.356***	0.366***	0.360***
	(0.0796)	(0.0813)	(0.0804)	(0.0827)	(0.0798)	(0.0829)
4.distance	0.299***	0.295***	0.301***	0.307***	0.302***	0.310***
	(0.0873)	(0.0869)	(0.0879)	(0.0879)	(0.0877)	(0.0884)
5.distance	0.190**	0.168**	0.192**	0.177**	0.196***	0.181**
	(0.0731)	(0.0725)	(0.0749)	(0.0742)	(0.0738)	(0.0751)
6.distance	-0.0448	-0.0758	-0.0440	-0.0777	-0.0400	-0.0737
	(0.134)	(0.134)	(0.135)	(0.139)	(0.134)	(0.139)
margrev2dig	8.19e-05***	7.99e-05***	8.20e-05***	8.03e-05***	8.57e-05***	8.33e-05***
	(1.57e-05)	(1.62e-05)	(1.58e-05)	(1.61e-05)	(1.62e-05)	(1.65e-05)
1.distance*	4.12e-04***	3.90e-04**	4.13e-04***	3.95 e-04**	4.11e-04***	3.95e-04**
margrev2dig	(0.000145)	(0.000153)	(0.000147)	(0.000154)	(0.000144)	(0.000154)
2.distance*	-3.16e-05	-2.96e-05	-3.20e-05	-3.20e-05	-3.37e-05	-3.37e-05
margrev2dig	(3.50e-05)	(3.49e-05)	(3.51e-05)	(3.47e-05)	(3.51e-05)	(3.49e-05)
3.distance*	-2.20e-05	-2.06e-05	-2.23e-05	-2.25e-05	-2.46e-05	-2.45e-05
margrev2dig	(1.57e-05)	(1.57e-05)	(1.57e-05)	(1.62e-05)	(1.55e-05)	(1.59e-05)
4.distance*	3.03e-07	-1.19e-06	1.72e-07	-2.55e-06	-1.52e-06	-3.97e-06
margrev2dig	(2.36e-05)	(2.37e-05)	(2.36e-05)	(2.38e-05)	(2.34e-05)	(2.36e-05)
5.distance*	-1.06e-05	-9.36e-06	-1.06e-05	-9.12e-06	-1.27e-05	-1.08e-05
margrev2dig	(1.86e-05)	(1.79e-05)	(1.88e-05)	(1.87e-05)	(1.88e-05)	(1.87e-05)
6.distance*	-1.38e-05	-8.05e-06	-1.37e-05	-5.76e-06	-1.53e-05	-7.00e-06
margrev2dig	(3.37e-05)	(3.37e-05)	(3.36e-05)	(3.34e-05)	(3.36e-05)	(3.33e-05)
RelDummy_ij		-0.867***		-1.088***		-1.081***
		(0.104)		(0.118)		(0.118)
RelAmount_ij			0.0614	0.460***		0.455***
			(0.0916)	(0.104)		(0.103)
Equity /					0.102**	0.0807*
Total assets					(0.0415)	(0.0429)
Constant	9.512***	10.29***	9.489***	10.32***	9.476***	10.29***
	(0.0303)	(0.109)	(0.0561)	(0.111)	(0.0414)	(0.119)
Observations	28,762	28,762	28,762	28,762	28,762	28,762
R-squared	0.295	0.302	0.295	0.304	0.295	0.304
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1-9 Intensive margin Services

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
1.distance	1.348***	1.274***	1.348***	1.249***	1.349***	1.250***
	(0.417)	(0.407)	(0.413)	(0.426)	(0.416)	(0.425)
2.distance	0.457***	0.471***	0.451***	0.504***	0.462***	0.508***
	(0.134)	(0.130)	(0.135)	(0.130)	(0.135)	(0.130)
3.distance	0.654***	0.643***	0.649***	0.668***	0.656***	0.670***
	(0.133)	(0.131)	(0.134)	(0.131)	(0.133)	(0.131)
4.distance	0.377***	0.370***	0.373***	0.388***	0.379***	0.390***
	(0.0813)	(0.0768)	(0.0819)	(0.0751)	(0.0810)	(0.0750)
5.distance	0.0333	-0.00243	0.0302	0.00260	0.0171	-0.0108
	(0.101)	(0.103)	(0.102)	(0.101)	(0.107)	(0.104)
6.distance	0.183	0.119	0.180	0.116	0.186	0.119
	(0.146)	(0.150)	(0.147)	(0.149)	(0.146)	(0.149)
margrev2dig	5.22e-06**	5.01e-06**	5.18e-06**	5.10e-06**	5.56e-06**	5.40e-06**
	(2.37e-06)	(2.41e-06)	(2.39e-06)	(2.45e-06)	(2.67e-06)	(2.72e-06)
1.distance*	-9.75e-05	-9.18e-05	-9.74e-05	-9.02e-05	-9.73e-05	-9.02e-05
margrev2dig	(6.90e-05)	(7.29e-05)	(6.97e-05)	(7.07e-05)	(6.92e-05)	(7.08e-05)
2.distance*	1.67e-05	1.19e-05	1.64e-05	1.19e-05	1.65e-05	1.18e-05
margrev2dig	(1.40e-05)	(1.43e-05)	(1.40e-05)	(1.37e-05)	(1.40e-05)	(1.37e-05)
3.distance*	1.09e-05*	9.68e-06	1.07e-05*	1.01e-05	1.10e-05*	1.02e-05
margrev2dig	(5.88e-06)	(6.02e-06)	(5.82e-06)	(6.19e-06)	(5.83e-06)	(6.18e-06)
4.distance*	5.66e-06	5.21e-06	5.72e-06	4.76e-06	5.37e-06	4.53e-06
margrev2dig	(4.28e-06)	(3.68e-06)	(4.25e-06)	(3.71e-06)	(4.37e-06)	(3.81e-06)
5.distance*	1.33e-05**	1.42e-05**	1.33e-05**	1.45e-05**	1.67e-05***	1.74e-05***
margrev2dig	(6.01e-06)	(5.99e-06)	(5.97e-06)	(6.19e-06)	(5.23e-06)	(5.48e-06)
6.distance*	-1.26e-06	-8.89e-09	-1.28e-06	4.63e-07	-9.76e-07	6.89e-07
margrev2dig	(1.58e-05)	(1.62e-05)	(1.59e-05)	(1.57e-05)	(1.58e-05)	(1.57e-05)
RelDummy_ij		-1.051***		-1.376***		-1.363***
		(0.136)		(0.136)		(0.137)
RelAmount_ij			-0.104	0.527***		0.525***
			(0.1000)	(0.0661)		(0.0649)
Equity / Total assets					0.0559***	0.0474***
					(0.0103)	(0.0108)
Constant	8.443***	9.338***	8.495***	9.352***	8.442***	9.340***
	(0.0219)	(0.136)	(0.0667)	(0.137)	(0.0243)	(0.140)
Observations	37,245	37,245	37,245	37,245	37,245	37,245
R-squared	0.291	0.305	0.292	0.309	0.293	0.310
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The results (Table 1-7 to Table 1-9)[23] seem to corroborate this decision: closeness appears to have different effects in the macro sectors analyzed. For the macro sector Agriculture, all models do not show any significant effects of being linked for scarcely efficient firms. In contrast, the effects are directly proportional to the marginal revenue return of capital. In particular, my favorite specification (model 6) estimates that borrowings increase with the efficiency when the firm is at a distance of 3, 4, and 5. Although controls for the dimension of relationship lending and leverage are always significant, the estimation seems to be scarcely affected by their introduction: the magnitude of the coefficients remains virtually unaltered.

For the Manufacturing sector, results are different. Here distance matters also for non-efficient firms. Dummies for distances report positive and statistically significant coefficients from a distance 2 to 5 in any model. Only the interaction for a distance equal to one is positive and significant, suggesting that direct connections improve bank screening. Unlike in Agriculture, the firm's efficiency is always positive and significant, meaning that more efficient firms obtain a larger amount of lending with respect to less efficient firms.

The results for the service sector are similar to those for the manufacturing sector. However, now the fixed term for a distance equal to 1 is positive for all firms, and we do not observe any positive interaction with efficiency. The interaction between the dummy for a path of five steps and efficiency is instead positive. Overall, it seems that the effect of connections does not differ between efficient and non-efficient firms.

Except for the macro-sector Agriculture and direct links in Manufacturing, the results are coherent with the literature that depicts shared exponents as a source of favoritism (Barone, Mirenda, and Mocetti 2017) rather than with the literature considering interlocks as a tool reducing informational asymmetry (M. A. Ferreira and Matos 2012). Indeed, if the hypothesis about the reduction of informational asymmetry were correct, we would have observed only a positive interaction between distance measures and the efficiency of the firms. However, this is the case only for firms in agriculture and firms in the manufacturing sector directly connected to banks.

[23] The number of observations differs from the one reported in the descriptive statistics because REGHDFE package iteratively removes singletons (Correia, 2015).

1.5.2 The probability of a link

The second channel through which ties can impact misallocation is the probability of having a connection. Indeed, it is not the favoritism arising from connections per se that causes misallocation. Favoritism would generate misallocation only if the least efficient firms interlocks with banks more frequently than the efficient firms.

I order to disentangle the effect of being efficient on the probability of a link, I run the following linear regression:

$$\begin{aligned} DistanceRecoded_{\ ij} \\ &= \alpha + \ \beta_1 EFF_i \ + \beta_3 DEGRYSE_i + \ \beta_4 BA_j \ + \ \beta_5 RelDummy_{ij} \\ &+ \ \beta_6 RelAmount_{ij} \ + \beta_7 \frac{Equity}{TotalAsset} + \ \varepsilon_{ij} \end{aligned}$$

DistanceRecoded is a variable that measures the length of the geodesic path between a bank and a firm. It goes from 1 to 7, where 7 includes those couples not linked by any path (infinite distance). The other variables are the same as the model for the intensive margin of credit. As before, I cluster standard error for banks and firms and run separate models by macro-sectors.

Except for Agriculture, the marginal revenue of capital is a good predictor of closeness to a bank. For both the Manufacturing and Service sectors, the coefficient for efficiency is always negative, and the magnitude is scarcely affected by the addition of controls (Tables from 1-10 to 1-12)[24].

I obtained similar results (not reported) by regressing a dummy taking the values 1 if there is a path between the bank and the firm and 0 otherwise, on the same model.

[24] The R squared reported does not vary due to rounding.

Table 1-10 Linear Regression Distance recoded Agriculture

VARIABLES	(1) MODEL 1	(2) MODEL 2	(3) MODEL 3	(4) MODEL 4	(5) MODEL 5	(6) MODEL 6
margrev2dig	3.20e-07	3.20e-07	3.20e-07	3.20e-07	3.38e-07	3.38e-07
	(2.64e-07)	(2.64e-07)	(2.64e-07)	(2.64e-07)	(2.88e-07)	(2.88e-07)
RelDummy_ij		-0.0153		0.00342		0.00449
		(0.0302)		(0.0424)		(0.0423)
RelAmount_ij			-0.0345	-0.0385		-0.0391
			(0.0348)	(0.0460)		(0.0460)
Equity / Total assets					0.0199*	0.0199*
					(0.0119)	(0.0119)
Observations	769,454	769,454	769,454	769,454	769,454	769,454
R-squared	0.273	0.273	0.273	0.273	0.273	0.273
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1-11 Linear Regression Distance recoded Manufacturing

VARIABLES	(1) MODEL 1	(2) MODEL 2	(3) MODEL 3	(4) MODEL 4	(5) MODEL 5	(6) MODEL 6
margrev2dig	-5.21e-06***	-5.21e-06***	-5.21e-06***	-5.21e-06***	-4.84e-06***	-4.83e-06***
	(1.63e-06)	(1.63e-06)	(1.63e-06)	(1.63e-06)	(1.67e-06)	(1.67e-06)
RelDummy_ij		-0.0468**		-0.0779**		-0.0776**
		(0.0207)		(0.0313)		(0.0313)
RelAmount_ij			-0.0121	0.0813**		0.0811**
			(0.0199)	(0.0336)		(0.0335)
Equity / Total assets					0.00366	0.00363
					(0.00235)	(0.00235)
Observations	3,058,419	3,058,419	3,058,419	3,058,419	3,058,419	3,058,419
R-squared	0.449	0.449	0.449	0.449	0.449	0.449
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1-12 Linear Regression Distance recoded Services

VARIABLES	(1) MODEL 1	(2) MODEL 2	(3) MODEL 3	(4) MODEL 4	(5) MODEL 5	(6) MODEL 6
margrev2dig	-1.57e-07**	-1.57e-07**	-1.57e-07**	-1.57e-07**	-1.34e-07*	-1.34e-07*
	(7.56e-08)	(7.55e-08)	(7.56e-08)	(7.55e-08)	(7.02e-08)	(7.01e-08)
RelDummy_ij		-0.0494***		-0.110***		-0.110***
		(0.0146)		(0.0288)		(0.0288)
RelAmount_ij			-0.00915	0.114***		0.114***
			(0.0103)	(0.0272)		(0.0271)
Equity / Total assets					0.00192***	0.00192***
					(0.000448)	(0.000448)
Observations	6,802,159	6,802,159	6,802,159	6,802,159	6,802,159	6,802,159
R-squared	0.389	0.389	0.389	0.389	0.389	0.389
ILS	YES	YES	YES	YES	YES	YES
bank FE	YES	YES	YES	YES	YES	YES
Clustered errors	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank	Firm bank

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

1.6 The pruned scenario

The above empirical results suggest that exponent links may not determine misallocation for the system in Portugal. In fact. On the one hand, close ties seem to determine an increase in the amount of lending independent of the actual efficiency of the linked firm. On the other hand, efficient firms are more likely to be close to banks, thus experiencing more frequently than non-efficient firms the benefits of the connections. This section aims to investigate how these two mechanisms interact and reverberate on national GDP. In particular, I will compare the actual GDP with that observed if every tie would be severed. To generate this counterfactual scenario, I will again exploit Dias et al.'s model (2014) summarized above.

Before focusing on the macro level, I need to recover what would happen at the micro-level to lending and capital after the pruning. I define the counterfactual borrowings as follows[25]:

[25] Since I clustered by firm and bank, it is scarcely reasonable to assume homoscedasticity between observations. As a consequence, I cannot use the estimate of the untransformed scale expectation to build the counterfactual because I could not account for the residuals' variance (Duan 1983). My solution assumes that heteroscedasticity is not due to the links.

$$Borrow_{ij(C)} = \exp(X_i' \beta_{(C)} + \varepsilon_{ij})$$

Where X_i' is the vector of covariates in the econometric model, and $\beta_{(C)}$ is the vector of coefficients with any coefficient for the distance set to 0. Since the actual borrowings are $Borrow_{ij} = \exp(X_i' \beta + \varepsilon_{ij})$, I compute the counterfactual borrowings as:

$$Borrow_{ij(C)} = \frac{Borrow_{ij}}{\exp(B_{1l} Dis_{ij}^{1} + B_{3l}(Dis_{ij}^{1} * EFF_i)}$$

I use $Borrow_{ij(C)}$ in order to compute the counterfactual capital quantity the firm will have after the pruning. Then, by using (7)-(9), (5), and the fact that the capital of a firm is simply composed of the capital provided by the bank and the one that comes from other sources, I calculate the optimal quantities H_{si}^{*}, Q_{si}^{*} that would have been chosen if the quantity of borrowed capital was the counterfactual quantity K_{si}^{*} .

A possible issue in simply using $K_{si}^{*} = \sum_{j=1}^{J} Borrow_{ij(C)}$ as the counterfactual quantity is that it does not assure that the sector demand for financing remains the same. Changes in sector demand would imply a general equilibrium effect, raising the concern for the possible reallocation of financings between sectors. Since the interest of the chapter is in the misallocation within sectors, I define the counterfactual capital as follows:

$$\begin{aligned} K_{si}^{*} = K_i - \sum_{j=1}^{J} Borrow_{ij} \\ &+ max\left\{0, \sum_{j=1}^{J} Borrow_{ij(C)} \right. \\ &\left. + \frac{\sum_{k \in Cluster(i)} \left(\sum_{j=1}^{J} Borrow_{kj} - \sum_{j=1}^{J} Borrow_{kj(C)} \right)}{|ILS_i|} \right\} \end{aligned}$$

Where K_i is the actual capital, $|ILS_i|$ is number of firms belonging to cluster ILS_i. Making this computation means redistributing the sum of the credit gains or losses caused by severing links in a cluster equally in the cluster. The rationale is that freeing [immobilizing] money from [within] a line of credit increases [decreases] the money present in the market. Thus, in order for the market to clear, the price decreases [increases], thus increasing [decreasing] the quantity borrowed from firms. I redistribute only within firms with similar characteristics (the same cluster) to account for different demands. In the counterfactual exercise, I furtherly assume that the prices of labor and intermediate products remain the same[26] and that $Y_s^{\frac{1}{\sigma}}P_s = Y_s^{*\frac{1}{\sigma}}P_s^* = \gamma_s$. After recovering the counterfactual borrowings, I use Dias et al.'s model (2014) to aggregate the results.

Since pruning ties influence only the misallocation through capital, profit maximization implies:

$$\left(1+\tau_{K_{si}}^*\right) = \frac{a_s}{1-a_s-b_s}\frac{Z_s Q_{si}^*}{R_s K_{si}^*}$$

$$\left(1+\tau_{H_{si}}\right) = \frac{b_s}{1-a_s-b_s}\frac{Z_s Q_{si}^*}{W_s H_{si}^*}$$

$$\left(1-\tau_{Y_{si}}\right) = \frac{\sigma}{\sigma-1}\frac{1}{(1-a_s-b_s)}\frac{Z_s Q_{si}^*}{P_{si}^* Y_{si}^*}$$

Therefore, it must be that:

$$\begin{gathered}
\frac{Z_s Q_{si}^*}{W_s H_{si}^*} = \frac{Z_s Q_{si}}{W_s H_{si}} \\
\frac{Z_s Q_{si}^*}{P_{si}^* Y_{si}^*} = \frac{Z_s Q_{si}}{P_{si} Y_{si}} \\
Y_{si}^* = A_{si}\ (K_{si}^*)^{a_s}\ (H_{si}^*)^{b_s}\ (Q_{si}^*)^{1-a_s-b_s} \\
P_{si}^* = \gamma_s (Y_{si}^*)^{-\frac{1}{\sigma}}
\end{gathered} \tag{13}$$

Which assuming $Z_s = W_s = 1$ as before, becomes:

$$\frac{Q_{si}^*}{H_{si}^*} = \frac{Q_{si}}{H_{si}} \tag{14}$$

[26] Avoiding this assumption would have made the computations too cumbersome.

$$\frac{Q_{si}^*}{P_{si}^* Y_{si}^*} = \frac{Q_{si}}{P_{si} Y_{si}}$$

$$Y_{si}^* = A_{si}\,(K_{si}^*)^{a_s}\,(H_{si}^*)^{b_s}\,(Q_{si}^*)^{1-a_s-b_s}$$

$$P_{si}^* = \gamma_s (Y_{si}^*)^{-\frac{1}{\sigma}}$$

Solving the system results in the counterfactual output quantity:

$$Y_{si}^* = \left[A_{si}\,(K_{si}^*)^{a_s}\left(\gamma_s \frac{H_{si}}{P_{si} Y_{si}}\right)^{b_s}\left(\gamma_s \frac{Q_{si}}{P_{si} Y_{si}}\right)^{1-a_s-b_s}\right]^{\frac{\sigma}{1+(\sigma-1)a_s}} \tag{15}$$

And implies the aggregate change[27]:

$$\frac{Y^*}{Y} = \prod_{s=1}^{S}\left(\frac{Y_s^*}{Y_s}\right)^{\theta_s} = \prod_{s=1}^{S}\left(\frac{\left(\sum_{i=1}^{M_s}(Y_{si}^*)^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}}{\left(\sum_{i=1}^{M_s}(Y_{si})^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}}\right)^{\theta_s}$$

$$= \prod_{s=1}^{S}\left(\left(\frac{\sum_{i=1}^{M_s}(Y_{si}^*)^{\frac{\sigma-1}{\sigma}}}{\sum_{i=1}^{M_s}(Y_{si})^{\frac{\sigma-1}{\sigma}}}\right)^{\frac{\sigma}{\sigma-1}}\right)^{\theta_s}$$

In doing this calculation, I have also added the product of whom have not had any lendings by using the fact that $Y_{si} = \left(\frac{P_{si} Y_{si}}{\gamma_s}\right)^{\frac{\sigma}{\sigma-1}}$.

I calculate the change in the value added deriving from the reallocation as:

[27] Actually, what we can recover from the data is only ${Y^*}_{si}^{not\gamma} = \left[A_{si}\,(K_{si}^*)^{a_s}\left(\frac{H_{si}}{P_{si}Y_{si}}\right)^{b_s}\left(\frac{Q_{si}}{P_{si}Y_{si}}\right)^{1-a_s-b_s}\right]^{\frac{\sigma}{1+(\sigma-1)a_s}}$ since we do not observe γ_s. This is not a problem since we calculate the ratio between Y_{si}^* and Y_{si} and we are able to recover $Y_{si}^{not\gamma}$ through the formula (15) by using as capital the actual capital.
Similarly for the intermediate products we observe only ${Q_{si}^*}^{not\gamma} = \frac{Q_{si}}{P_{si}Y_{si}}\left(\left[A_{si}\,(K_{si}^*)^{a_s}\left(\frac{H_{si}}{P_{si}Y_{si}}\right)^{b_s}\left(\frac{Q_{si}}{P_{si}Y_{si}}\right)^{1-a_s-b_s}\right]^{\frac{\sigma}{1+(\sigma-1)a_s}}\right)^{\frac{\sigma-1}{\sigma}}$ so we need to recalculate for Q_{si}, ${Q_{si}}^{not\gamma}$.

$$\frac{V^*}{V}=\frac{\sum_{s=1}^{S}\sum_{i=1}^{Ms}P_{si}Y_{si}\left(\frac{P_{si}^*Y_{si}^*}{P_{si}Y_{si}}\right)-\sum_{s=1}^{S}\sum_{i=1}^{Ms}ZQ_{si}\left(\frac{Q_{si}^*}{Q_{si}}\right)}{\sum_{s=1}^{S}\sum_{i=1}^{Ms}P_{si}Y_{si}-\sum_{s=1}^{S}\sum_{i=1}^{Ms}ZQ_{si}}$$

Where $P_{si}^*Y_{si}^*=\gamma_s(Y_{si}^*)^{-\frac{1}{\sigma}}Y_{si}^*=\gamma_s Y_{si}^{*\frac{\sigma-1}{\sigma}}$.

In the following tables, results for both value added and GDP are reported. For GDP, I present the counterfactuals calculated by measuring labor as the total labor cost (empexp) and as the number of employees (nempl). Table 1-13 reports the results disaggregated by macro sectors. The changes we observe are too small to reasonably hypothesize an actual difference between real data and the counterfactual. I obtain the same result if the system is considered in its entirety (Table 1-14).

I interpret these results as evidence that the difference we observe in the probability of linking with a bank between efficient and non-efficient firms is large enough to counterbalance the negative effect of connections.

Table 1-13 Counterfactual gain by macro-sector

	Agriculture	Manufacturing	Services
GainGrossMacSecCF_empexp	.9965037	.9997864	.9989619
GainGrossMacSecCF_nempl	.9965315	.9997706	.9989583
GainValueAddMacSecCF	.9990232	.9999131	.998699

Table 1-14 Counterfactual gain Total

	Total
GainGrossPTCF_empexp	.9993004
GainGrossPTCF_nempl	.9992916
GainValueAddPTCF	.9991939

1.7 Conclusion

In the present chapter, I use firm-level data from Portugal to analyze the importance of interlocking for firms' borrowings. Unlike the extant literature, I investigate the consequences of relations both at the micro and macro levels. The essay is divided into two main parts. In the first one, I empirically assess the direction of two mechanisms through which connections can generate misallocation. In particular, I focus on how interlocks impact the intensive margin of credit and the difference in the probability of linking with a bank between efficient and non-efficient firms. In the second part, I assess how these two mechanisms interact and how this influences Portugal's GDP.

Analyzing the effect of ties on the intensive margin, I find evidence that connections can potentially cause misallocation of resources. The results are consistent with the literature that considers ties as a source of favoritism: overall, ties positively affect the amount of credit received irrespectively to the firm's efficiency. However, I find that efficiency negatively correlates with the distance from banks, meaning that more efficient firms are those closer to banks. This mechanism may curb the misallocating potential of connections because the firms which are more likely to take advantage of relationships would also be those that would be favored in the absence of frictions. I test this implication by setting up a counterfactual exercise to estimate the macro consequences of deleting ties for the GDP. Removing interlocks in Portugal would not result in any GDP changes. I interpret the results as evidence that the different closeness to banks experienced by efficient and non-efficient firms balances the favoritism generated by connections.

1.8 References

Accetturo, Antonio, Giorgia Barboni, Michele Cascarano, and Emilia Garcia-Appendini. 2020. ‘Cultural Proximity and the Formation of Lending Relationships’. Competitive Advantage in the Global Economy (CAGE).

Badunenko, Oleg, and Pavlo Mozharovskyi. 2016. 'Nonparametric Frontier Analysis Using Stata'. *The Stata Journal: Promoting Communications on Statistics and Stata* 16 (3): 550–89.

Banerjee, Abhijit V., and Esther Duflo. 2005. ‘Growth Theory through the Lens of Development Economics’. *Handbook of Economic Growth* 1: 473–552.

Banerjee, Abhijit V., and Benjamin Moll. 2010. ‘Why Does Misallocation Persist?’ *American Economic Journal: Macroeconomics* 2 (1): 189–206.

Bank of Portugal. 2018. ‘Economic Bulletin’.

Barone, Guglielmo, Litterio Mirenda, and Sauro Mocetti. 2017. ‘Losing My Connection: The Dark Side of Bank-Firm Interlocking Directorates’.

Battiston, S., and M. Catanzaro. 2004. ‘Statistical Properties of Corporate Board and Director Networks’. *The European Physical Journal B* 38 (2): 345–52.

Battiston, Stefano. 2004. ‘Inner Structure of Capital Control Networks’. *Physica A: Statistical Mechanics and Its Applications* 338 (1–2): 107–12.

Battiston, Stefano, Gérard Weisbuch, and Eric Bonabeau. 2003. ‘Decision Spread in the Corporate Board Network’. *Advances in Complex Systems* 6 (04). World Scientific: 631–44.

Berger, Allen N., and Gregory F. Udell. 1995. ‘Relationship Lending and Lines of Credit in Small Firm Finance’. *Journal of Business*, 351–81.

Berger, Allen N., Nathan H. Miller, Mitchell A. Petersen, Raghuram G. Rajan, and Jeremy C. Stein. 2005. ‘Does Function Follow Organizational Form? Evidence from the Lending Practices of Large and Small Banks’. *Journal of Financial Economics* 76 (2). Elsevier: 237–69.

Bharath, Sreedhar T., Sandeep Dahiya, Anthony Saunders, and Anand Srinivasan. 2009. ‘Lending Relationships and Loan Contract Terms’. *The Review of Financial Studies* 24 (4): 1141–1203.

Billor, Nedret, Ali S. Hadi, and Paul F. Velleman. 2000. ‘BACON: Blocked Adaptive Computationally Efficient Outlier Nominators’. *Computational Statistics & Data Analysis* 34 (3). Elsevier: 279–98.

Bleck, Alexander, and Xuewen Liu. 2018. ‘Credit Expansion and Credit Misallocation’. *Journal of Monetary Economics* 94 (April): 27–40.

Bonfim, Diana, Qinglei Dai, and Francesco A. Franco. 2009. ‘The Number of Bank Relationships, Borrowing Costs and Bank Competition’. *SSRN Electronic Journal*.

Bonn, Ingrid, and Andrew Pettigrew. 2009. ‘Towards a Dynamic Theory of Boards: An Organisational Life Cycle Approach’. *Journal of Management & Organization* 15 (01): 2–16.

Boot, Arnoud W. A. 2000. ‘Relationship Banking: What Do We Know?’ *Journal of Financial Intermediation* 9 (1): 7–25.

Borio, Claudio, Enisse Kharroubi, Christian Upper, and Fabrizio Zampolli. 2016. ‘Labour Reallocation and Productivity Dynamics: Financial Causes, Real Consequences’.

Brakman, Steven, Harry Garretsen, and Charles Van Marrewijk. 2001. *An Introduction to Geographical Economics: Trade, Location and Growth*. Cambridge university press.

Braun, Matías, Ignacio Briones, and Gonzalo Islas. 2018. ‘Interlocking Directorates, Access to Credit, and Business Performance in Chile during Early Industrialization’. *Journal of Business Research*.

Byrd, Daniel T., and Mark S. Mizruchi. 2005. ‘Bankers on the Board and the Debt Ratio of Firms’. *Journal of Corporate Finance* 11 (1): 129–73.

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2008. 'Zombie Lending and Depressed Restructuring in Japan'. *American Economic Review* 98 (5): 1943–77.

Calligaris, Sara. 2015. 'Misallocation and Total Factor Productivity in Italy: Evidence from Firm-Level Data'. *LABOUR* 29 (4): 367–93.

Calligaris, Sara, Massimo Del Gatto, Fadi Hassan, Gianmarco I. P. Ottaviano, and Fabiano Schivardi. 2018. 'The Productivity Puzzle and Misallocation: An Italian Perspective'. *Economic Policy* 33 (96): 635–84.

Coelli, Timothy J., Dodla Sai Prasada Rao, Christopher J. O'Donnell, and George Edward Battese. 2005. *An Introduction to Efficiency and Productivity Analysis*. Springer Science & Business Media.

Coin, D., L. Infante, G. Nuzzo, and M. Piazza. 2011. 'A Known Unknown? Networks of Firms and Access to Credit in Italy'. https://www.researchgate.net/profile/Luigi_Infante/publication/265632147_A_known_unknown_Networks_of_firms_and_access_to_credit_in_Italy/links/55253d3c0cf201667be69d45.pdf.

Colin Cameron, A., and Douglas L. Miller. 2015. 'A Practitioner's Guide to Cluster-Robust Inference'. *Journal of Human Resources* 50 (2): 317–72.

Cooper, William W., Lawrence M. Seiford, and Kaoru Tone . 2006. *Introduction to Data Envelopment Analysis and Its Uses: With DEA-Solver Software and References*. 1st ed. Springer US.

Correia, Sergio. 2016. 'Reghdfe: Estimating Linear Models with Multi-Way Fixed Effects'. In *2016 Stata Conference*. Stata Users Group.

———. 2015. 'Singletons, Cluster-Robust Standard Errors and Fixed Effects: A Bad Mix', 7.

Csardi, Gabor, and Tamas Nepusz. 2006. 'The Igraph Software Package for Complex Network Research'. *InterJournal, Complex Systems* 1695 (5): 1–9.

Dagnes, Joselle. 2014. 'Struttura e Dinamica Dei Legami Inter-Organizzativi Nel Capitalismo Finanziario Italiano'. *Stato e Mercato*, no. 2/2014.

Davis, Gerald F., Mina Yoo, and Wayne E. Baker. 2003. 'The Small World of the American Corporate Elite, 1982-2001'. *Strategic Organization* 1 (3): 301–26.

Degryse, Hans, Olivier De Jonghe, Sanja Jakovljević, Klaas Mulier, and Glenn Schepens. 2019. 'Identifying Credit Supply Shocks with Bank-Firm Data: Methods and Applications'. *Journal of Financial Intermediation*, April.

Degryse, Hans, Moshe Kim, and Steven Ongena. 2009. *Microeconometrics of Banking: Methods, Applications, and Results*. Oxford University Press.

Davis, Gerald F. 1991. 'Agents without Principles? The Spread of the Poison Pill through the Intercorporate Network'. *Administrative Science Quarterly*, 583–613.

Davis, Gerald F., and Henrich R. Greve. 1997. 'Corporate Elite Networks and Governance Changes in the 1980s'. *American Journal of Sociology* 103 (1): 1–37.

Degryse, Hans, Moshe Kim, and Steven Ongena. 2009. *Microeconometrics of Banking: Methods, Applications, and Results*. Oxford University Press.

Degryse, Hans, and Steven Ongena. 2005. 'Distance, Lending Relationships, and Competition'. *The Journal of Finance* 60 (1): 231–66.

Dias, Daniel A., Christine Richmond, and Carlos Robalo Marques. 2014. 'Misallocation and Productivity in The Lead Up to The Eurozone Crisis'. WORKING PAPERS Lisbon Banco de Portugal.

Dittmann, Ingolf, Ernst Maug, and Christoph Schneider. 2010. 'Bankers on the Boards of German Firms: What They Do, What They Are Worth, and Why They Are (Still) There'. *Review of Finance* 14 (1): 35–71.

Duan, Naihua. 1983. 'Smearing Estimate: A Nonparametric Retransformation Method'. *Journal of the American Statistical Association* 78 (383): 605–10.

Engelberg, Joseph, Pengjie Gao, and Christopher A. Parsons. 2012. 'Friends with Money'. *Journal of Financial Economics* 103 (1): 169–88.

Ferreira, Miguel A., and Pedro Matos. 2012. 'Universal Banks and Corporate Control: Evidence from the Global Syndicated Loan Market'. *The Review of Financial Studies* 25 (9): 2703–44.

Ferreira, Miguel, Ana Lacerda, and Paula Antão. 2011. 'Bank Loans and Banks' Corporate Control: Evidence for Portugal'. *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*.

Garcia-Bernardo, Javier, and Frank W. Takes. 2018. 'The Effects of Data Quality on the Analysis of Corporate Board Interlock Networks'. *Information Systems* 78 (November): 164–72.

Gopinath, Gita, Şebnem Kalemli-Özcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. 2017. 'Capital Allocation and Productivity in South Europe'. *The Quarterly Journal of Economics* 132 (4): 1915–67.

Granovetter, Mark. 2005. 'The Impact of Social Structure on Economic Outcomes'. *Journal of Economic Perspectives* 19 (1): 33–50.

Güner, A. Burak, Ulrike Malmendier, and Geoffrey Tate. 2008. 'Financial Expertise of Directors'. *Journal of Financial Economics* 88 (2): 323–54.

Haunschild, P.R., and C.M. Beckman. 1998. 'When Do Interlocks Matter?: Alternate Sources of Information and Interlock Influence'. *Administrative Science Quarterly* 43 (4): 815–44.

Heemskerk, Eelke M., Fabio Daolio, and Marco Tomassini. 2013. 'The Community Structure of the European Network of Interlocking Directorates 2005 – 2010'. *PLoS ONE* 8 (7).

Heemskerk, Eelke M., Frank W. Takes, Javier Garcia-Bernardo, and M. Jouke Huijzer. 2016. 'Where Is the Global Corporate Elite? A Large-Scale Network Study of Local and Nonlocal Interlocking Directorates'. *ArXiv Preprint ArXiv:1604.04722*.

Hellmann, Thomas, Laura Lindsey, and Manju Puri. 2008. 'Building Relationships Early: Banks in Venture Capital'. *The Review of Financial Studies* 21 (2). Oxford Academic: 513–41.

Hillman, Amy J., Albert A. Cannella, and Ramona L. Paetzold. 2000. 'The Resource Dependence Role of Corporate Directors: Strategic Adaptation of Board Composition in Response to Environmental Change'. *Journal of Management Studies* 37 (2): 235–56.

Hillman, Amy J., and Thomas Dalziel. 2003. 'Boards of Directors and Firm Performance: Integrating Agency and Resource Dependence Perspectives'. *Academy of Management Review* 28 (3): 383–96.

Hopenhayn, Hugo A. 2014. 'Firms, Misallocation, and Aggregate Productivity: A Review'. *Annual Review of Economics* 6 (1): 735–70.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. 'Misallocation and Manufacturing TFP in China and India'. *The Quarterly Journal of Economics* 124 (4): 1403–48.

Inklaar, Robert, Addisu A. Lashitew, and Marcel P. Timmer. 2017. 'The Role of Resource Misallocation in Cross-Country Differences in Manufacturing Productivity'. *Macroeconomic Dynamics* 21 (3): 733–56.

Ji, Yong-bae, and Choonjoo Lee. 2010. 'Data Envelopment Analysis'. The Stata Journal 10 (2): 267–80.

Kalemli-Ozcan, Sebnem, and Bent E Sorensen. 2012. 'Misallocation, Property Rights, and Access to Finance: Evidence from Within and Across Africa'. Working Paper 18030. National Bureau of Economic Research.

Karolyi, Stephen Adam. 2018. 'Personal Lending Relationships'. *The Journal of Finance* 73 (1): 5–49.

Kilduff, Martin, and David Krackhardt. 1994. 'Bringing the Individual Back in: A Structural Analysis of the Internal Market for Reputation in Organizations'. *Academy of Management Journal* 37 (1). Academy of Management Briarcliff Manor, NY 10510: 87–108.

Kroszner, Randall S, and Philip E Strahan. 2001. 'Bankers on Boards: Monitoring, Conflicts of Interest, and Lender Liability'. *Journal of Financial Economics* 62 (3): 415–52.

La Porta, Rafael, Florencio Lopez-de-Silanes, and Guillermo Zamarripa. 2003. 'Related Lending'. *The Quarterly Journal of Economics* 118 (1): 231–68.

Liberti, José María, and Mitchell A Petersen. 2019. 'Information: Hard and Soft'. *The Review of Corporate Finance Studies* 8 (1): 1–41.

Midrigan, Virgiliu, and Daniel Yi Xu. 2014. ‘Finance and Misallocation: Evidence from Plant-Level Data’. *American Economic Review* 104 (2): 422–58.

Mizruchi, Mark S. 1996. ‘What Do Interlocks Do? An Analysis, Critique, and Assessment of Research on Interlocking Directorates’. *Annual Review of Sociology* 22: 271–98.

Morck, Randall, and Masao Nakamura. 1999. ‘Banks and Corporate Control in Japan’. *The Journal of Finance* 54 (1): 319–39.

Ongenah, Steven, and David C. Smith. 2000. ‘Bank Relationships: A Review’. In *Performance of Financial Institutions: Efficiency, Innovation, Regulation*, 221.

Peek, Joe, and Eric S. Rosengren. 2005. ‘Unnatural Selection: Perverse Incentives and the Misallocation of Credit in Japan’. *American Economic Review* 95 (4): 1144–66.

Petersen, Mitchell A., and Raghuram G. Rajan. 1994. ‘The Benefits of Lending Relationships: Evidence from Small Business Data’. *The Journal of Finance* 49 (1): 3–37.

———. 2002. ‘Does Distance Still Matter? The Information Revolution in Small Business Lending’. *Journal of Finance* 57 (6): 2533–70.

Pfeffer, Jeffrey, and Gerald R. Salancik. 1978. *The External Control of Organizations: A Resource Dependence Perspective*. Harper & How, Publishers. Inc.

Ratcliff, Richard E. 1980. ‘Banks and Corporate Lending: An Analysis of the Impact of the Internal Structure of the Capitalist Class on The Lending Behavior of Banks’. *American Sociological Review* 45 (4): 553–70.

Reis, Ricardo. 2013. ‘The Portuguese Slump and Crash and the Euro Crisis’. *Brookings Papers on Economic Activity* 2013 (1): 143–210.

Restuccia, Diego, and Richard Rogerson. 2008. ‘Policy Distortions and Aggregate Productivity with Heterogeneous Establishments’. *Review of Economic Dynamics* 11 (4): 707–20.

———. 2013. ‘Misallocation and Productivity’. *Review of Economic Dynamics*, Special issue: Misallocation and Productivity, 16 (1): 1–10.

———. 2017. ‘The Causes and Costs of Misallocation’. *Journal of Economic Perspectives* 31 (3): 151–74.

Schivardi, Fabiano, Enrico Sette, and Guido Tabellini. 2017. ‘Credit Misallocation during the European Financial Crisis’. *CESifo Working Paper, No. 6406, Center for Economic Studies and Ifo Institute (CESifo)*.

Sette, Enrico, and Giorgio Gobbi. 2015. ‘Relationship Lending During a Financial Crisis’. *Journal of the European Economic Association* 13 (3): 453–81.

Sforza, Alessandro. 2017. ‘Shocks and the Organization of the Firm: Who Pays the Bill?’ http://www.csef.it/IMG/pdf/sforza_jmp.pdf.

Sisli-Ciamarra, Elif. 2012. ‘Monitoring by Affiliated Bankers on Board of Directors: Evidence from Corporate Financing Outcomes’. *Financial Management* 41 (3): 665–702.

Stearns, Linda Brewster, and Mark S. Mizruchi. 1993. ‘Board Composition and Corporate Financing: The Impact of Financial Institution Representation on Borrowing’. *Academy of Management Journal* 36 (3): 603–18.

Stefanelli, Valeria. 2010. ‘The Cooperative Banking System in Portugal: The Case of Credito Agricola Mutuo Group’. In *Cooperative Banking in Europe*, 7–22. Springer.

Takahashi, Hidenori. 2015. ‘Dynamics of Bank Relationships in Entrepreneurial Finance’. *Journal of Corporate Finance* 34 (October): 23–31.

Tirole, Jean. 2010. *The Theory of Corporate Finance*. Princeton University Press.

Vitali, Stefania, James B. Glattfelder, and Stefano Battiston. 2011. ‘The Network of Global Corporate Control’. *PLoS ONE* 6 (10): e25995.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge university press.

Weber, Sylvain. 2010. ‘Bacon: An Effective Way to Detect Outliers in Multivariate Data Using Stata (and Mata)’. *The Stata Journal* 10 (3). SAGE Publications Sage CA: Los Angeles, CA: 331–38.

Zajac, Edward J. 1988. ‘Interlocking Directorates as an Interorganizational Strategy: A Test of Critical Assumptions’. *Academy of Management Journal* 31 (2): 428–38.

2 An implementation for the estimation of the dose-response function when the treatment is multidimensional: a Stata package.

Abstract

Propensity score methods are wildly used techniques for the evaluation of causal effects in observational studies. Although Rosenbaum and Rubin's (1983) original article focused solely on binary treatments, further studies generalize the methods to multi-valued treatments, continuous treatments, and multidimensional continuous treatments. Despite its potential, Stata offers plenty of packages for all the cases but the last one. This chapter aims to introduce a new Stata package that implements the propensity score generalization to multidimensional continuous treatment developed by Egger and von Ehrlich (2013). The chapter illustrates the econometric framework and presents the commands implemented. I finally perform a simple simulation to show the commands and the method's capability to overcome bias.

Keywords: continuous multiple treatments, GPSMD, dose-response, generalized propensity score.

2.1 Introduction

Since Rosenbaum and Rubin's (1983) groundbreaking article, propensity score (PS) methods have become widely used instruments in the evaluation of causal effects in observational studies (Pearl 2010; King and Nielsen 2019). Unlike in (ideal) randomized experiments − where the treatment is randomly assigned to different groups and thus exposed and unexposed can be considered exchangeable (Hernán and Robins 2006) −, in observational studies, the treatment assignment is not governed by the researcher and, therefore, exchangeability is an issue. The consequence is that in the ideal case of random experiments, the effect of exposure can be recovered by simply using association measures; in fact, the groups differ only for the reception of the treatment. In observational studies, the indiscriminate use of association measures gives a misleading estimation of the causal effect because different groups likely have different compositions (Hernán and Robins 2006). PS methods are born to reestablish the balance between exposed and non-exposed groups so that the causal effect can be identifiable. In particular, Rosenbaum and Rubin (1983) show that, in settings where selection on observables undermines balance (Cerulli 2015), conditioning on the propensity score (i.e., the probability of being treated conditional on observable covariates) succeeds in restoring the balance between groups.

Originally, Rosenbaum and Rubin (1983) developed PS for settings with binary treatments. More recent extensions include the generalized propensity score (GPS) for multi-valued treatments (Imbens 2000) as well as continuous treatments (Hirano and Imbens 2004; Imai and Van Dyk 2004). Egger and von Ehrlich (2013) provide an extension of Hirano and Imbens (2004) to the case in which the treatment is composed of more than one continuous dimension. Stata makes available a large set of packages for the estimation of propensity score – see Guo and Fraser (2015) and Caliendo and Kopeinig (2008) for an overview mainly devoted to binary treatments, and Bia and Mattei (2008), Guardabascio and Ventura (2014), Bia et al. (2014) for continuous treatments. Nevertheless, none of them enables analyses adopting the multidimensional framework. This is probably why only a handful of studies (Peter Hannes Egger and Egger 2016; Peter H. Egger and Lassmann 2018; Peter H. Egger, Ehrlich, and Nelson 2020; Erhardt 2017) adopts the multidimensional framework despite its potential usefulness. Many policies are better conceptualized as a multidimensional treatment rather than binary or mono-dimensional. For example, the effect of financial aids can differ depending on the type of investments the financial aids trigger. Moreover, different policies can be contemporarily adopted, and their effects can be seen only by considering them together.

This chapter aims to present a Stata package for the estimation of dose-response function in the presence of multidimensional treatment when the dimensions are continuous. The package has been named GPSMD. The structure of the chapter is the following: first, I will describe Egger and von Ehrlich's econometric framework; second, I will describe the different commands; third, I will present a simple example by using simulated data to show the commands and the performance of the method. The last section concludes and introduces some further modifications that I mean to implement.

2.2 The econometric outline

In this section is summarized the econometric framework defined by Egger and von Ehrlich (2013).

The actual treatment level experienced by the i^{th} observational unit is a random vector $\boldsymbol{T_i} = (T_{1i}, \dots, T_{Mi})'$ where M is the number of dimensions of the treatment. There are N observational units in the sample.

The outcome Y_i is assumed to be a function of the treatment $Y_i(\boldsymbol{T_i}) = h(T_{1i}, \dots, T_{Mi})$. The potential outcome is defined as: $Y_i(\boldsymbol{t}), \boldsymbol{t} \in \Im$. Where $\Im$ is the set of all potential treatments. The average dose-response function is defined as:

$$\mu(\boldsymbol{t}) \equiv E[Y_i(\boldsymbol{t})]$$

An m-equation structural model determines the level of the treatment. The reduced equations are defined as:

$$T_{mi} = f(\boldsymbol{Z_i}, \gamma_m) + v_{mi}\,, m = 1, \dots, M$$

Where $\boldsymbol{Z_i}$ is a vector formed by the union of the exogenous variables X_{mi} and possibly their interactive terms[28].

For the identification, weak unconfoundedness is assumed, that is:

$$Y_i(\boldsymbol{t}) \perp \boldsymbol{T_i} | \boldsymbol{Z_i} \; \forall \boldsymbol{t} \in \Im$$

This means that the potential outcome at level $\boldsymbol{t}$ of treatment is independent of the actual treatment status $\boldsymbol{T_i}$ when we condition on the covariates $\boldsymbol{Z_i}$ and that this is true for all treatments.[29] Simply controlling for covariates can induce a problem concerning the dimensionality of the model: the solution is conditioning on the propensity score.

The density of observing the treatment $\boldsymbol{T_i} = \boldsymbol{t} = (t_{1i}, \dots t_{Mi})'$ conditional on the exogenous variables

[28] The choice of using the same set of variables in estimating the reduced equations can be scarcely parsimonious. I am planning to implement also the possibility of using different sets of covariates. This will require the implementation of the SUR estimator in order to increase efficiency (Cameron and Trivedi 2009). I thanks Giovanni Cerulli for the suggestion.

[29] This imply that we can estimate the average treatment effect by using those that actually have received the treatment: $E[Y_i(\boldsymbol{t})|\boldsymbol{Z_i}] = E[Y_i(\boldsymbol{t})|\boldsymbol{T_i} = \boldsymbol{t}',\ \boldsymbol{Z}_i] = E[Y_i(\boldsymbol{t})|\boldsymbol{T_i} = \boldsymbol{t},\ \boldsymbol{Z}_i] = E[Y_i|\boldsymbol{T_i} = \boldsymbol{t},\ \boldsymbol{Z}_i]$. (Imbens 2000)

$$g(\boldsymbol{t},\boldsymbol{z}) \equiv f_{\boldsymbol{T}_i|\boldsymbol{Z}_i}(\boldsymbol{T}_i = \boldsymbol{t}|\boldsymbol{Z}_i = \boldsymbol{z})$$

Then the propensity score is defined as $G_i = g(\boldsymbol{T}_i, \boldsymbol{Z}_i)$. It is worth noting that the generalized propensity score defines the random variable $G_i = g(\boldsymbol{T}_i, \boldsymbol{Z}_i)$, as a transformation of both $\boldsymbol{T}_i$ and $\boldsymbol{Z}_i$, as well as the family of random variables indexed by $\boldsymbol{t}$, $g(\boldsymbol{t}, \boldsymbol{Z}_i)$ (Imbens 2000).

By construction, the generalized propensity score implies the balance property. Loosely speaking:

$$\boldsymbol{Z}_i \perp 1\{\boldsymbol{T}_i = \boldsymbol{t}\}|g(\boldsymbol{t},\boldsymbol{Z}_i)\ \forall\boldsymbol{t} \in \Im$$

This means that, once controlled for $g(\boldsymbol{t}, \boldsymbol{Z}_i)$, $\boldsymbol{Z}_i$ and the treatment are independent (see Appendix 1 for the proof).

As a consequence, under weak unconfoundedness, it can be shown (Peter H. Egger and von Ehrlich 2013) (see Theorem 1 in Appendix 1) that, once conditioned on the propensity score, the potential outcome is independent of the treatment:

$$Y_i(\boldsymbol{t}) \perp \boldsymbol{T}_i|g(\boldsymbol{t},\boldsymbol{Z}_i), \forall\boldsymbol{t} \in \Im$$

And, thus, that conditioning on $g(\boldsymbol{t}, \boldsymbol{Z}_i)$ is like if we conditioned on covariates[30]. This implies the following (see Theorem 2 in Egger and von Ehrlich (2013)):

$$E[Y_i|\boldsymbol{T}_i = \boldsymbol{t}, g(\boldsymbol{T}_i,\boldsymbol{Z}_i)] = E[Y_i(\boldsymbol{t})|T_i = \boldsymbol{t}, g(\boldsymbol{t},\boldsymbol{Z}_i)] = E[Y_i(\boldsymbol{t})|g(\boldsymbol{t},\boldsymbol{Z}_i)]$$

And, then, that, by using the law of iterated expectation, the average dose-response function can be retrieved:

$$\mu(t) \equiv E[Y_i(\boldsymbol{t})] = E\big[E[Y_i(\boldsymbol{t})|g(\boldsymbol{t},\boldsymbol{Z}_i)]\big]$$

[30] It is worth noting that independence, as well as the balancing property, hold within strata of propensity score calculated at a given treatment t, $g(\boldsymbol{t},\boldsymbol{Z}_i)$, not within strata of the propensity score $g(\boldsymbol{T}_i,\boldsymbol{Z}_i)$.

The practical implementation consists in estimating $E[Y_i|\boldsymbol{T_i} = \boldsymbol{t}, g(\boldsymbol{T_i}, \boldsymbol{Z_i})]$ by a flexible polynomial function with $g(\boldsymbol{T_i}, \boldsymbol{Z_i})$ as a covariate. Then, we average over the propensity score for retrieving the dose-response function. If the flexible polynomial function is correctly specified, then GPSMD should reduce the bias of the estimation.

Summarizing, similarly to the case of mono-dimensional continuous treatment, the steps involved in the implementation of the GPS method are:

1. Estimating the Generalized propensity score (command: `gpsMD`)
2. Enforcing the common support if needed (command: `CommSupp`)
3. Testing the balancing property (command: `gpsMDbal`)
4. Estimating the dose-response function (command: `gpsMDPolEs`)

In the following, I describe the commands and algorithms for implementing each step. Before starting, I briefly summarize the structure of commands in Stata and their outputs.

To invoke a program, a Stata user must type the command's name followed by several arguments. The number and nature of arguments are peculiar to each command and constitute the command syntax. Every command can have (loosely) two types of output: printed and not printed. Printed outputs are those outputs that Stata displays in the "Stata's Result window" when the command is invoked. Non-printed outputs are of two kinds: the results of the program stored in memory that will be removed if another command of the same type is invoked; the objects stored in memory whose life is independent of the commands that follow.

Regarding the first type, programs differ in Stata depending on where the results are stored in memory. The following programs but `ComSupp` are e-class commands. This means that Stata stores saved results in `e()`. For this reason, I will refer to e-class objects when I describe the programs' stored results. `ComSupp` is an n-class command, and, as a consequence, it does not store anything in spaces like `e()`. Like all programs, however, it can create objects that will last independently of further program invocations.

In every section where I present a command, I will describe the algorithm, the syntax of the program, and the objects and e-class objects generated by the program. I will discuss the printed output in the section with the application.

2.3 gpsMD

The first step in the application of the GPS method is the estimation of the propensity score. I provide the command: gpsMD. In the following, I describe the algorithm as well as the command syntax.

Recalling that the reduced equations are defined as:

$$T_{mi} = f(\boldsymbol{Z_i}, \boldsymbol{\gamma_m}) + v_{mi}\ , m = 1, \dots, M$$

Where $\boldsymbol{Z_i}$ is the vector including the union of the exogenous variables X_{mi} and possibly their interactive terms. In the implementation of gpsMD, I assume errors having a multivariate normal distribution:

$$\boldsymbol{v_i} = (v_{1i}, \dots, v_{Mi})' \sim \mathcal{N}(\ \boldsymbol{0_m}, \boldsymbol{\Sigma})$$

The assumption on the errors implies that $\boldsymbol{T}_i|\boldsymbol{Z}_i$ is distributed: $\boldsymbol{T}_i|\boldsymbol{Z}_i \sim \mathcal{N}(\ \boldsymbol{f}(\boldsymbol{Z_i}, \boldsymbol{\gamma_m}), \boldsymbol{\Sigma})$.

In order to estimate the propensity score we proceed in 4 steps:

1. The m reduced equations are estimated by OLS;
2. The m vectors of residuals are predicted;
3. By using the residuals, the variance-covariance matrix, $\boldsymbol{\Sigma}$, is estimated: $\boldsymbol{\Sigma} = Cov(\boldsymbol{v_1}, \dots, \boldsymbol{v_M})$ where $\boldsymbol{v_m} = [v_{1m}, \dots, v_{Nm}]$;
4. Finally, the generalized propensity score for each observational unit is estimated by:
$$G_i = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\ \boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\{-\frac{1}{2}\boldsymbol{v}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{v}_i\}$$
This implies the following formula for the vector $\boldsymbol{G} = (G_1, \dots, G_N)'$:
$$\boldsymbol{G} = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\ \boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\{-\frac{1}{2} diagonal(\boldsymbol{v}'\ \ \boldsymbol{\Sigma}^{-1}\boldsymbol{v}\ \ \ )\}$$
Where exp(.) is now a function for the elementwise exponentiation of a matrix, $\boldsymbol{v}'\ \ = \begin{bmatrix} \boldsymbol{v}_1{}' \\ \dots \\ \boldsymbol{v}_N{}' \end{bmatrix}$, and N is the number of observational units in the sample.

If the treatment is deemed to follow a multivariate log-normal distribution, the program can calculate (option `ln(varlist)`) the conditional density using the log transformation of the dimensions of the treatment in the list. The program then divides the resulting conditional density by the dimensions of the treatment that have been transformed to recover the propensity score of the untransformed treatment (Dobrow 2013, para. 6.6).

To select the model the Akaike's Information Criterion (Cavanaugh and Neath 2019) could be used. The command Akaikemax, which identifies the model which minimizes the Akaike's Information Criterion is described in Appendix 2.2.

2.3.1 Syntax (gpsMD)

```
gpsMD varlist(min=1) , exogenous(varlist) gpsMD(string)
      [chosenpoint(string) ln(varlist)]
```

`varlist(min=1)` : the dimensions of the treatment.

`exogenous(varlist)`: the list of the exogenous variable and their possible interactions and powers, depending on the model the user has in mind.

`gpsMD(string)`: the name for the variable where the generated propensity score will be stored.

`chosenpoint(string)`: the name of the Stata column vector with the point at which we want to calculate the propensity score (mostly for programs). It is an option that enables the user to generate the propensity score calculated at a given point. It generates $g(\boldsymbol{t}, \boldsymbol{Z_i})$ instead of $g(\boldsymbol{T_i}, \boldsymbol{Z_i})$.

`ln(varlist)` : the treatment dimensions that have to be log-transformed.

2.3.2 Variables generated (gpsMD)

The variable named as specified in `gpsMD(string)` with the estimated propensity score.

2.3.3 E-class objects generated (gpsMD)

Macros:

`e(gpsMDvar)` : macro with the string in the option `gpsMDvar`.

`e(Exogenous)`: macro with the varlist in the option `Exogenous`.

`e(Dimensions)`: macro with the varlist of the dimensions of the treatment.

`e(cmdline#)`: macro with the cdmline of the reduced equation for dimension #. The user may want to run again only one of the regressions and focus on those results. This macro enables to do it easily.

`e(cmd)`: macro with the name of the command just invoked (gpsMD).

`e(cmdline)`: macro with the cdmline. This macro reports the command just invoked, including options and specifications.

`e(chosenpoint)`: macro with the name of the column vector with the chosen point.

`e(LNVarCreated)` : if the `ln(varlist)` option is specified, the program generates variables named `LN_`*`var`* consisting of the logarithmic transformation of the variables in the varlist. `e(LNVarCreated)` contains the list of the variable generated.

`e(DimensionsFS)` : macro with the name of the dimension used in calculating the propensity score. It differs from `e(Dimensions)` only if the `ln(varlist)` option is used.

Matrices:

`e(VarCov)` : the estimated variance-covariance matrix, **Σ**.

2.4 ComSupp

Egger et al. (Peter Hannes Egger and Egger 2016; Peter H. Egger, Ehrlich, and Nelson 2020) generalize to the multidimensional case Flores et al.'s (2012) methodology to identify the common support in the case of continuous treatment. The rationale is similar to that of the "minima and maxima comparison" for binary treatments (Caliendo and Kopeinig 2008). "Minima and maxima comparison" consists in reducing the sample to those observations whose propensity score is higher than the maximum of the minimums of treated and controls group, as well as lower than the minimum of the maximums of treated and controls group. The main difference is that continuous dimensions must be discretized to apply a similar criterion: we do not have "treated" and "not treated" anymore. Moreover, differently to the binary case, for each discrete treatment, it must be chosen a treatment point that represents the discrete set and at which the GPS will be calculated[31]. The outline of the algorithm follows.

We partition each dimension of the treatment $T_{mi}\ , m = 1, \ldots, M$ in L sets such that:

$$range(T_{mi}) = \bigcup_{l}^{L} T_{mi}^{l} \;\; m = 1, \ldots, M$$

T is then discretized in $M \times L$ sets defined by $\times_{m=1}^{M} \{T_m^1, \ldots, T_m^L\}_m$. Denote this set as $\mathcal{J}$.

Then for each discrete subset of the treatment $T^D \in \mathcal{J}$:

1. we chose a representative point (e.g., mean, median), $\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}$;
2. we evaluate the GPS at that point, $g_i(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})$, for each observation in the sample;
3. we keep only those observations such that their GPS calculated at point 2 satisfies:

$$g_i(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i}) \in \left[\max\left\{\mathop{min}_{j\in T^D}\left(g_j(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_j})\right), \mathop{min}_{j\notin T^D}\left(g_j(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_j})\right)\right\}, \min\left\{\max_{j\in T^D}\left(g_j(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_j})\right), \max_{j\notin T^D}\left(g_j(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_j})\right)\right\}\right]$$

The command does not eliminate observations. It only generates a variable so that the user is free to inspect the characteristics of observations inside and outside the common support.

2.4.1 Syntax (ComSupp)

```
ComSupp varlist(min=1) , exogenous(varlist) index(string)
       cutpoints(numlist integer max=1) obs_notsup(string)
       [testing(numlist integer max=1) ln(varlist)]
```

`varlist(min=1)` : the treatment dimensions in the same order as in the gpsMD command.

[31] As underscored above, an analogy exists between the propensity score for binary treatment and the GPS calculated at a given treatment point, not between the propensity score for binary treatment and the GPS as such.

`exogenous(varlist)`: exogenous variables in the same order that in the gpsMD command.

`index(string)`: the point $\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}$ where the user wants to calculate the GPS. It can be "mean" or "p50": "mean" for the mean, and "p50" for the median.

`cutpoints(numlist integer max=1)`: the number of discrete intervals of the dimensions of the treatment.

`obs_notsup(string)`: the name for the dummy variable that takes value 1 if the observation is outside the common support and 0 if the observation is inside the common support.

`testing(numlist integer max=1)`: the user may want to inspect the distribution of the GPS calculated at the representative point of the discrete subsets of the treatment, $g_i(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})$. If `testing` is set to 1, the program generates one variable for each discrete subset of the treatment. This variable stores, for all observations, the GPS calculated at the representative point of that discrete subset of the treatment. These variables are named `obs_notsup#` where `obs_notsup` is the name specified in `obs_notsup(string)` and `#` stands for the number of the discrete subset. The dummy variable indicating whether the observation is inside the common support is named simply as in `obs_notsup(string)`.

`ln(varlist)` : the treatment dimensions that have to be log-transformed.

2.4.2 Variables generated (ComSupp)

As explained above, if `testing` is different from 1, the program generates a variable named as in `obs_notsup` that takes value 1 if the observation is outside the common support and value 0 if the observation is within the common support.

Note that `ComSupp` is a nclass command, so you can invoke it after running `gpsMD` and before invoking `gpsMDBal` without incurring an error.

2.5 gpsMDbal

The gpsMDbal command tests if the balancing property of the GPS holds. A strategy similar to Bia e Mattei is implemented (Bia and Mattei 2012; Bia and Mattei

2008; Peter H Egger and Erhardt 2014; Hirano and Imbens 2004; Guo and Fraser 2015) [32].

Recalling that the balancing property is loosely defined as:

$$\boldsymbol{Z_i} \perp 1\{\boldsymbol{T_i} = \boldsymbol{t}\}|g(\boldsymbol{t}, \boldsymbol{Z_i}), \forall \boldsymbol{t} \in \Im$$

The command tests whether conditioning on the propensity score effectively removes the differences in respect to the mean of the exogenous covariates between groups treated with different doses. The procedure is the multidimensional analog to the t-tests for equality of means before and after matching implemented in `pstest` for binary treatments (Leuven and Sianesi 2003). The test consists of the following steps:

1. We partition each dimension of the treatment T_{mi} , $m = 1, \dots, M$ in L sets such that:

$$range(T_{mi}) = \bigcup_{l}^{L} T_{mi}^{l} \;\; m = 1, \dots, M$$

2. T is then discretized in $M \times L$ sets defined by $\times_{m=1}^{M} \{T_m^1, \dots, T_{mi}^L\}_m$. Denote this set as $\mathcal{J}$ and a single discrete treatment as T^D.
3. As customary, at first, for each exogenous covariate Z and for each T^D, independence is assessed before conditioning on the propensity score. That is, we check whether:

$$\bar{Z}_{T^D} = \bar{Z}_{T^{D'}} , T^D \neq T^{D'}; T^D, T^{D'} \in \mathcal{J}$$

   Or

$$\bar{Z}_{T^D} = \bar{Z}_{T^{-D}} , T^D \in \mathcal{J}; T^{-D} = \bigcup \mathcal{J} \setminus T^D$$

   Where $\bar{Z}_{T^D}$ is the sample mean for a given exogenous covariate calculated using the observations in T^D, $\bar{Z}_{T^{D'}}$ is the sample mean for a given exogenous covariate calculated using the observations in a discrete set different from T^D, and $\bar{Z}_{T^{-D}}$ is the sample mean for a given exogenous covariate calculated using the observations in the union of the discrete sets different from T^D.
   (In the package it is implemented the second test through the command `ttest` for unpaired two-sample).
4. Then, the program tests whether differences remain if we condition on the propensity score. Therefore, it stratifies the sample by propensity score and, within homogeneous propensity score strata, estimates the differences in the mean between groups with different treatments.

[32] Although the balancing property is a statement about distribution, as customary, the implementation focus only on the first moment (Lee 2013).

Following Bia (2008), the following algorithm is iterated over each discrete set of the treatment T^D:

a. A representative point $\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}$ is chosen (e.g., mean, median);
b. The generalized propensity score $g_i(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})$ is calculated for each observational unit.
c. The propensity score is then discretized in a finite number of intervals. Denote a single discrete interval as $g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})^D$.
d. Then for each variable in the propensity score, the program tests if the following average is different from 0.

$$\frac{1}{N}\sum_{g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D} N_{g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D}\left(\bar{Z}_{T^D\, g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D} - \bar{Z}_{T^{-D} g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D}\right)$$

Where the sum is over the intervals of the propensity score, N is the number of observational units in the sample, and $N_{g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D}$ is the number of observations in a given interval of the propensity score. $\bar{Z}_{T^D g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D}$ and $\bar{Z}_{T^{-D} g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}},\boldsymbol{Z_i})^D}$ are the sample means of the exogenous variable for those observations that belong to the set $T^D \cap g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})^D$ and $T^{-D} \cap g(\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}, \boldsymbol{Z_i})^D$ respectively (in Appendix 2.3 the test statistic is derived).

2.5.1 Syntax (gpsMDbal)

```
gpsMDbal varlist (min=1), cutpoints(numlist max=1) index(string)
          nq_gpsMD(numlist max=1) discrTreat(string) [ptile(string)
          obs_notsup(string) gpsMDTequalt(string) ln(varlist) ]
```

`varlist(min=1)`: The variables for which the balancing property has to be assessed.

`cutpoints(numlist max=1)`: the number of discrete intervals of the dimensions of the treatment (min 2).

`index(string)`: the point $\bar{\boldsymbol{t}}_{\boldsymbol{T^D}}$ where the user wants to calculate the GPS. It can be "mean" or "p50": "mean" for the mean, and "p50" for the median.

`nq_gpsMD(numlist max=1)`: the number of discrete subsets of the GPS.

`discrTreat(string)`: the program discretizes the treatment in a user-defined number of subsets. It also generates a variable storing the information about the discrete subset to which an observation belongs. In `discrTreat(string)` the user must specify the name of this variable.

`obs_notsup(string)`: the string with the name of the dummy variable generated by the command `ComSupp`. The variable indicates whether the observation is included in the common support or not. If specified, `ComSupp` must have been run before invoking gpsMDbal. If it is not specified, gpsMDbal will perform the analysis using the entire sample.

ptile(string) : the program generates the discrete subsets of the treatment by calculating the cartesian product of the discrete intervals of the dimensions. In addition, the program generates a variable for each dimension where it stores the discrete subset of the dimension to which an observation belongs. In ptile(string) the user must specify the incipit for the name of these variables.

gpsMDTequalt(string) : the user may want to inspect the distribution of the GPS calculated at the representative point of the discrete subsets of the treatment, $g_i(\bar{\boldsymbol{t}}_{\boldsymbol{T}^D}, \boldsymbol{Z}_i)$. When gpsMDTequalt(string) is specified, the program generates one variable for each discrete subset of the treatment storing the GPS calculated at the representative point of that discrete subset. These variables are named gpsMDTequalt# where gpsMDTequalt is the name specified in gpsMDTequalt(string) and # stands for the number of the discrete subset. By default, the program does not generate these variables.

ln(varlist) : the treatment dimensions that have to be log-transformed.

level(numlist max=1) : the program prints the table with the adjusted and unadjusted differences in means (e(NofDiscTreat)) both entirely and setting to missing cells whose p-value is higher than a certain threshold (specified in level()). The default is 0.05.

2.5.2 Variables generated (gpsMDbal)

One variable named as specified in discrTreat. This variable reports, for each observation, the discrete set of the treatment to which the observation is assigned.

One variable named `ptile'# for every dimension # of the treatment. the `ptile'# variable stores the number of the discrete subset of the #dimension to which the observation is assigned. discrTreat is generated as the Cartesian product of `ptile'#.

2.5.3 E-class object generated

Some objects are simply copied from the gpsMD results. These are results the user could find helpful to have access also after running gpsMDbal.

Macros:

e(NofDiscTreat): a macro with the number of discrete treatments.

e(cmd): a macro with the command.

e(cmdline): a macro with cdmline. This macro reports the command just invoked. It includes options and specifications.

e(DimensionsFS): macro with the name of the dimension used in calculating the propensity score. It differs from e(Dimensions) only if the ln(varlist) option is used.

e(LNVarCreated): if the ln(varlist) option is used, e(LNVarCreated) contains the list of the variables generated by gpsMD.

Matrices:

e(TabellaImpRes): matrix having one row for each variable the user wanted to test and two columns for each discrete treatment. In the cells, the program reports the p-value of the test before and after adjusting for the GPS.

e(ResultAdj#): for each discrete subset of the treatment, the program generates a matrix reporting the results of the t-test for the adjusted mean. The first column reports the t statistic, the second column reports the p-value, and the third column reports the degree of freedom. There is one row for each variable that the user wanted to test.

e(Result#): the program generates a matrix reporting the results of the t-test for discrete subset # before the adjustment. There is one column for every r-class object of ttest plus one for the estimated difference. There is one row for each variable that the user wanted to test. e(ResultAdj#) and e(Result#) are somehow redundant objects. e(TabellaImpRes) already includes the essential information.

e(Chosenpoint#): The program reports, for each discrete subset of the treatment #, a matrix storing the representative treatment vector $\bar{t}_{T^D}$ chosen.

2.6 gpsMDPolEst

This command implements the estimation of the outcome given the treatment and the propensity score. The command estimate models of the form:

$$
\begin{aligned}
&E[Y_i|\boldsymbol{T_i}, G_i(\boldsymbol{T_i}, \boldsymbol{Z_i})] \\
&\qquad = \alpha_0 \\
&\qquad + \sum_{m=1}^{M}\sum_{j=1}^{k}\big(\alpha_{mj}T_{mi}^{j} + \alpha_{Gj}G_i^{j} + \alpha_{mGj}(G_i \cdot T_{mi})^{j} + \alpha_{\ln G_i}(\ln G_i)^{j} \\
&\qquad + \alpha_{m\ln G_i}(\ln G_i \cdot T_{mi})^{j}\big)
\end{aligned}
$$

Recalling that $E[Y_i|\boldsymbol{T_i}, G_i(\boldsymbol{T_i}, \boldsymbol{Z_i})] = E[Y_i|\boldsymbol{T_i} = \boldsymbol{t}, G_i(\boldsymbol{T_i} = \boldsymbol{t}, \boldsymbol{Z_i})]$= $E[Y_i|g(\boldsymbol{t}, \boldsymbol{Z_i})]$. Now the problem is that we must calculate $E\big[E[Y_i|g(\boldsymbol{t}, \boldsymbol{Z_i})]\big]$. Since we do not observe for each $\boldsymbol{t}$ the entire distribution of $g(\boldsymbol{t}, \boldsymbol{Z_i})$ and that the distribution of $g(\boldsymbol{t}, \boldsymbol{Z_i})$ – and, thus, of $E[Y_i|g(\boldsymbol{t}, \boldsymbol{Z_i})]$ – depends entirely on the distribution of $\boldsymbol{Z_i}$, we can use the sample distribution of $\boldsymbol{Z_i}$ to predict $\hat{g}(\boldsymbol{t}, \boldsymbol{Z_i})$ and, then, we can use the polynomial just estimated to predict $E[Y_\iota|\widehat{\boldsymbol{t}, g(\boldsymbol{t}}, \boldsymbol{Z_\iota})]$.

Therefore, the average response function is estimated for a given $\boldsymbol{t}$ as:

$$
E\big[\widehat{Y(\boldsymbol{t})}\big] = \frac{1}{N}\sum_{i=1}^{N} E[Y_\iota|\widehat{\boldsymbol{t}, g(\boldsymbol{t}}, \boldsymbol{Z_\iota})]
$$

In addition to the dose-response function, the program `gpsMDPolEst` estimates the partial derivatives for each dimension of the treatment.

In estimating confidence intervals, the t-approximation is not reliable because the model includes the propensity score, which is a generated regressor (Wooldridge 2010). Confidence intervals for the response function are thus estimated by bootstrap. Although the bias-corrected accelerated method (BCa) would have been more reliable, it would have been too computationally demanding. Therefore, the program calculates the Bias corrected method (BC) (Carpenter and Bithell 2000; Efron and Tibshirani 1994).

More specifically, for B bootstrap samples, the entire procedure for obtaining the response function and the partial derivatives is replicated – starting from the estimation of the reduced equations. The confidence interval at the $1-\alpha$ level is calculated as:

$$\left(F_{\theta^*}^{-1}\left(\Phi\left(2b+z_{\frac{\alpha}{2}}\right)\right), F_{\theta^*}^{-1}\left(\Phi\left(2b-z_{\frac{\alpha}{2}}\right)\right)\right)$$

Where $F_{\theta^*}^{-1}$ is the inverse of the bootstrap empirical distribution function, $\Phi(.)$ the cumulative distribution function of the normal distribution, and $z_{\frac{\alpha}{2}} = \Phi^{-1}\left(\frac{\alpha}{2}\right)$. $b = \Phi^{-1}\left(\frac{\#\theta^* \leq \hat{\theta}}{B}\right)$ is the correction for the bias, and $\hat{\theta}$ is the parameter estimated from the original sample.

The quantiles are calculated according to Carpenter and Bithell (2000): $Q = (B+1)\Phi\left(2b \pm z_{\frac{\alpha}{2}}\right)$. If needed, the results are rounded, and if $Q > B$, Q is set to B. Similarly, if $Q = 0$, Q is set to 1.

The program does not produce any graph. Drawing more than two-dimensional graphs in Stata is not easy, and any graph would require some adjustments. The program's output consists of several datasets that the user can furtherly process with programs like `graph3d` (Rostam-Afschar and Jessen 2014), `plotmatrix` (Mander 2019; Präg 2019), or `graph twoway contour`.

2.6.1 Syntax (gpsMDPolEst)

```
gpsMDPolEst outcome treatment dimensions, gpsmd(string)
      model(string) exogenous(varlist) file_pred(string)
      numboot(numlist integer max=1) [dividingint(numlist integer
      max=1) matrtreat(string)level(numlist max=1)
      cutpoints(numlist integer max=1) index(string) ln(varlist)
      matrixwithresults(string)]
```

`model(string)` : a string with the right side of the model. The right side of the model must be explicitely written due to how program parse inputs (e.g. "T1 + T2 + gps + T1*gps + T2*gps + T1^(2) + T2^2 + (gps^2) + ((T1*gps)^(2)) + (T2*gps)^2 + ln(gps) + (ln(gps))^2 + (T2*ln(gps))^2 + T2*ln(gps) ")

`dividingint(numlist integer max=1)`: if `matrtreat(string)` is not specified the program generates a matrix by dividing the dimensions in `dividingint` number of intervals. The Cartesian product of the extremes of the intervals in the different dimensions constitutes the set of treatment points for which the program estimates the response. The set of treatment points will be stored in `e(matrtreat)` as a matrix with $(dividingint + 1)^{number\ of\ dimensions}$ rows and columns equal to the number of dimensions.

`matrtreat(string)` : The user can specify the treatment points for which she is interested in estimating the response. Treatment points must be stored in a Stata matrix named as specified in `matrtreat(string)`. The matrix must have one column for each treatment dimension. A row of the matrix identifies a single point. The user can specify only one option between `dividingint` and `matrtreat`.

`exogenous(varlist)` : the exogenous variables the user wants to use in the reduced equations.

`file_pred(string)` : As explained above, the program does not generate any graphs but only various datasets (see below for a more detailed description of the files generated) with the necessary information for the user to generate the desired graphs. In `file_pred(string)`, the user must specify the incipit of the name for the files .dta storing the results.

`level(numlist max=1)` : the confidence level for the confidence intervals (default 0.05).

`numboot(numlist integer max=1)` : the number of bootstrap samples. Since bootstrapping is the only way to obtain the confidence intervals, this is not an optional argument.

`cutpoints(numlist integer max=1)` : the number of discrete intervals of the dimensions of the treatment when you calculate the common support. It is worth noticing that when common support is required, the program estimates the dose-response function by using only those observations that lie on the common support. I suggest using the same number used to calculate the common support.

`index(string)`: the point $\bar{\boldsymbol{t}}_{\boldsymbol{TD}}$ where the user wants to calculate the GPS. It can be "mean" or "p50": "mean" for the mean, and "p50" for the median. `ln(varlist)` : the treatment dimensions that have to be log-transformed.

matrixwithresults(string): if the argument is “T”, the program returns a matrix called e(returnresults) that includes all the results as well as the chosen doses. The default is “T”. If “F” the matrix is not generated. This option can be helpful when the number of treatment points exceeds Stata matrix limits.

2.6.2 Variables generated (gpsMDPolEst)

The command generates the variables specified in model but the treatment dimensions and the GPS. All the variables are named starting with I_, P_, or LN_. I_ represent interaction variables and P_ variables with power (variables like ((T1*GPS)^(2)) are named with both, e.g., P_2_I_T1_GPS). LN_ represents logarithmic transformation. The user should check if in her dataset there are variables with these incipits. If it is the case, it is preferable to change their names before running the program.

2.6.3 Dataset generated (gpsMDPolEst)

One dataset named as specified in file_pred(). It includes one row for each treatment point. The columns store the response, the partial derivatives, and the upper and lower bound of the confidence intervals.

If the dimensions are two, the program also generates nine datasets where every result is presented in matrix form.

The results estimated from the sample has the names: `file_pred'Mat_response.dta, `file_pred'Mat_PD_` Dim1'.dta, `file_pred'Mat_PD_` Dim2'.dta

The names for the datasets that include the bootstrap results in matrix form are similar: `file_pred'Mat_BootL_PD_`Dim1', `file_pred'Mat_BootL_PD_`Dim2', if the matrix stores the results for the lower bound; `file_pred'Mat_BootH_PD_`Dim1' `file_pred'Mat_BootH_PD_`Dim2', if the matrix stores the results for the upper bound.

The matrixes include two columns with the names of the rows and columns for the easy implementation of plotmatrix.

2.6.4 E-class objects generated (gpsMDPolEst)

Macros:

`e(gpsmd)` : the name of the variable with the GPS estimates.

`e(exogenous)` : the exogenous variables for the reduced equation estimation.

`e(Dimensions)` : the dimensions of the treatment.

`e(listgenvar)`: the program generates the variables as specified in `model(string)`. This macro reports the list of the variables generated.

`e(regmodel)` : the command for the regression for the polynomial estimation.

`e(cmd)` : macro with the command.

`e(cmdline)`: a macro with `cdmline`. This macro reports the command just invoked. It includes options and specifications.

`e(Outcome)` : a macro containing the name of the outcome variable.

matrices:

`e(matrtreat)` : a matrix with the treatment points for which the dose-response has been estimated.

`e(returnresults)`: if `matrixwithresults(T)` the program return a matrix with the same information included in `` `file_pred'.dta``.

2.7 A simple simulation for investigating the GPS method performance

In the rest of the chapter, I provide an example of the application of the package. Instead of proposing an application to a real dataset, I will show how the package works by using a generated dataset. I will finally investigate the performance of the GPS method by comparing the estimates obtained with GPS with those obtained using a rightly specified linear regression. The first step is, then, to generate the data. I set the following data generating process. The exogenous covariates are seven, $X_l \sim N(0,1); l = 1, \dots, 7$, while the treatment dimensions are two, T_1 and T_2. The reduced equations are:

$$T_1 = 1 \cdot X_1 + \ 0.5 \cdot X_2 + 1 \cdot X_3 + 0.5 \cdot X_4 + 1 \cdot X_5 + 0.5 \cdot X_6 + 1 \cdot X_7 + \varepsilon_1$$
$$T_2 = 0.5 \cdot X_1 + \ 1 \cdot X_2 + 0.5 \cdot X_3 + 1 \cdot X_4 + 0.5 \cdot X_5 + 1 \cdot X_6 + 0.5 \cdot X_7 + \varepsilon_2$$

Where $\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}$ follows a multivariate normal distribution:

$$\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix} \sim \mathcal{MN}(\mathbf{0_2}, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{bmatrix}25 & 2\\ 2 & 25\end{bmatrix}$$

The variance-covariance matrix chosen, $\mathbf{\Sigma}$, results in an R squared of 0.80 and 0.75 when estimating the reduced equations for T_1 and T_2 respectively. The following code generates the data.

```
. clear all
. set more off
. set type double
. set matsize 11000
. *I set the seed.
. set seed 13131
. *I generate the observations.
. set obs 1200
. *I generate the exogenous covariates X_m m=1,…,7.
. gen X1 = 1 * rnormal(0,1)
. gen X2 = 2 * rnormal(0,1)
. gen X3 = 3 * rnormal(0,1)
. gen X4 = 4 * rnormal(0,1)
. gen X5 = 5 * rnormal(0,1)
. gen X6 = 6 * rnormal(0,1)
. gen X7 = 7 * rnormal(0,1)
. *I generate the residuals in the reduced equations and the treatments.
. *I define the matrix of correlation.
. matrix R = (25, 2 \2, 25)
. *I generate residuals from the multivariate normal.
. drawnorm V1 V2, cov(R)
. *I generate the treatments
. gen T1= 1*X1 + .5*X2 + 1*X3 + .5*X4 + 1*X5 + .5*X6 + 1*X7 + V1
. gen T2= .5*X1 + 1*X2 + .5*X3 + 1*X4 + .5*X5 + 1*X6 + .5*X7 + V2
```

The outcome follows the model:

$$Y = 1 + 2 \cdot T_1 + \ 1.5 \cdot T_2 + \ 1 \cdot X_1 + \ 1.5 \cdot X_2 + 2 \cdot X_3 + 1 \cdot X_4 + 1.5 \cdot X_5 + 2 \cdot X_6 \\ + 1 \cdot X_7 + \eta$$

Where $\eta \sim \mathcal{N}(0{,}25)$. The R squared of this model, when estimated with a rightly specified linear regression, is 0.58.

```
. *I generate the outcome
. gen res= rnormal(0, 25)
. gen Y = 1+ 2*T1 + 1.5*T2 + 1*X1 + 1.5*X2 + 2*X3 + 1*X4 + 1.5*X5 + 2*X6 + 1*X7 + res
```

The first step involved in the implementation of the GPS method is to estimate the generalized propensity score $G_i(\boldsymbol{T}_i, \boldsymbol{X}_i)$. The command is `gpsMD`.

```
. * gpsMD
. gpsMD T1 T2, exogenous(X1 X2 X3 X4 X5 X6 X7) gpsMD(GPS)
****************
The regression for dimension: T1
****************

      Source |       SS           df       MS      Number of obs   =     1,200
-------------+----------------------------------   F(7, 1192)      =    716.05
       Model |  123285.356         7  17612.1938   Prob > F        =    0.0000
    Residual |  29318.8675     1,192  24.5963654   R-squared       =    0.8079
-------------+----------------------------------   Adj R-squared   =    0.8067
       Total |  152604.224     1,199   127.27625   Root MSE        =    4.9595

------------------------------------------------------------------------------
          T1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |   1.260802   .1459665     8.64   0.000     .9744223    1.547182
          X2 |   .5192245   .0692367     7.50   0.000     .3833851    .6550639
          X3 |   .9630556   .0476939    20.19   0.000     .8694823    1.056629
          X4 |   .4361229   .0365548    11.93   0.000      .364404    .5078417
          X5 |   1.002808   .0285938    35.07   0.000     .9467078    1.058907
          X6 |   .4845244   .0240968    20.11   0.000     .4372475    .5318013
          X7 |   1.004617   .0203212    49.44   0.000     .9647476    1.044486
       _cons |  -.1534133   .1439645    -1.07   0.287    -.4358655    .1290388
------------------------------------------------------------------------------

****************
The regression for dimension: T2
****************

      Source |       SS           df       MS      Number of obs   =     1,200
-------------+----------------------------------   F(7, 1192)      =    524.03
       Model |  96776.5041         7  13825.2149   Prob > F        =    0.0000
    Residual |  31448.1284     1,192  26.3826581   R-squared       =    0.7547
-------------+----------------------------------   Adj R-squared   =    0.7533
       Total |  128224.632     1,199   106.94298   Root MSE        =    5.1364

------------------------------------------------------------------------------
          T2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |   .6615431   .1511739     4.38   0.000     .3649465    .9581397
          X2 |   .8111875   .0717068    11.31   0.000     .6705019    .9518731
          X3 |    .509123   .0493954    10.31   0.000     .4122114    .6060346
          X4 |   .9728744   .0378589    25.70   0.000     .8985969    1.047152
          X5 |   .5054108   .0296139    17.07   0.000     .4473096     .563512
          X6 |   1.036732   .0249565    41.54   0.000     .9877685    1.085696
          X7 |   .4989273   .0210461    23.71   0.000     .4576357    .5402189
       _cons |   .0161213   .1491006     0.11   0.914    -.2764075    .3086501
------------------------------------------------------------------------------

****************
The Variance Covariance Matrix:
****************

             |     T1Res      T2Res
-------------+----------------------
       T1Res |  24.45277
       T2Res |  1.308303   26.22863
```

The command generates a variable named “GPS” and a printed output consisting of the outputs for the regressions estimating the reduced equations and the variance-covariance matrix.

The second step consists in identifying the observations outside the common support. The command is `ComSupp`. The output consists of a new dummy variable taking value 1 if the observation is outside the common support and 0 otherwise. The one-way table of frequencies of this variable is displayed.

The choice of the `cutpoints` is somehow arbitrary and affects the number of observations in the common support. I chose to discretize each dimension in two intervals corresponding to four discrete treatments.

```
. * ComSupp
. ComSupp T1 T2, exogenous(X1 X2 X3 X4 X5 X6 X7) index("p50") cutpoints(2)
obs_notsup(Commonsupport)

****************
COMMON SUPPORT (variable: "Commonsupport")
1 correspond to observations outside the common support
0 correspond to observations inside the common support
****************


Commonsuppo |
         rt |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        739       61.58       61.58
          1 |        461       38.42      100.00
------------+-----------------------------------
      Total |      1,200      100.00
```

The reported results show that only 61.58% of the sample is within the common support region. Unreported simulations resulted in restrictions with a similar magnitude (see also the restriction applied in Egger et al.(2020)). It seems that finding a broad common support is more demanding with multidimensional treatments than with mono-dimensional treatments. Although restricting the sample to the common support increases the consistency and credibility of the estimates, it can also be problematic when important observations are excluded (Lechner 2008). In the conclusions, I will propose a strategy that, arguably, can be adopted to reduce this flaw.

The third step consists in evaluating the balancing property. I restrict the evaluation to the common support region specifying the option `obs_notsup`. As in the case of `ComSupp,` I divide the treatments into four sets. The propensity score is instead divided into four intervals (`nq_gpsMD(4)`).

For simplifying reading the table, the command prints the output with and without omitting cells whose p-value is higher than the threshold decided by the user (in this case, the default: 0.05). In this case, we see that before adjusting for GPS only X_1, X_2, X_3 were reasonably balanced. Adjusting for GPS removes the unbalance in 13 groups over 17 and, thus, increases the dataset's balance quite strongly.

```
. * gpsMDbal
. gpsMDbal X1 X2 X3 X4 X5 X6 X7, index("p50") cutpoints(2) nq_gpsMD(4)
discrTreat(Discretetreat) obs_notsup(Commonsupport)

****************
In the following TabellaImpRes is reported
****************


   |     1r(p)      2r(p)      3r(p)      4r(p)  1Adj_r(p)  2Adj_r(p)  3Adj_r(p)  4Adj_r(p)
---+---------------------------------------------------------------------------------------
-
 X1|  .7471008   .4675596   .4154943   .0852119   .4991558    .442324    .509242   .6052064
 X2|  .6966121    .245041   .1762446   .8669508   .5684971   .1077818   .2513263   .5270022
 X3|  .0285343   .7127349   .0844789   .3174637   .6071563    .945155   .3662211   .6655259
 X4|  .0000645   .0018252   .0000178   6.01e-08   .7163549   .9850062   .0163841   .2335904
 X5|  4.58e-07   .0217688    .001727   .0000149    .913471   .4456956   .8739467   .5670758
 X6|  2.48e-07   2.13e-16   4.97e-12   3.55e-06    .866194   .0007445   .0081649   .3424754
 X7|  6.48e-09   9.54e-09   5.47e-10   1.21e-07   .1351072   .5067833   .0243766   .6991704

****************
In the following TabellaImpRes is reported but p-values higher than the threshold specified
in level(string) are omitted
****************


   |     1r(p)      2r(p)      3r(p)      4r(p)  1Adj_r(p)  2Adj_r(p)  3Adj_r(p)  4Adj_r(p)
---+---------------------------------------------------------------------------------------
-
 X1|         .          .          .          .          .          .          .          .
 X2|         .          .          .          .          .          .          .          .
 X3|  .0285343          .          .          .          .          .          .          .
 X4|  .0000645   .0018252   .0000178   6.01e-08          .          .   .0163841          .
 X5|  4.58e-07   .0217688    .001727   .0000149          .          .          .          .
 X6|  2.48e-07   2.13e-16   4.97e-12   3.55e-06          .   .0007445   .0081649          .
 X7|  6.48e-09   9.54e-09   5.47e-10   1.21e-07          .          .   .0243766          .
```

The last step involves the estimation of the control function and the response corresponding to the chosen doses. I restrict the estimation to the common support region also here. I decided not to define any particular matrix of the doses for which I intend to estimate the corresponding response. I specify `dividingint(3)` instead. If this option is invoked, the command selects the doses by performing the Cartesian product of the extremes of the intervals obtained by dividing the range of every dimension into three intervals. With two dimensions and by choosing `dividingint(3)`, we end up with 16 doses. It is worth noticing that when only the common support region is considered (as in this example), the program chooses the doses considering only the observations on the common support. As suggested in these cases (Efron and Tibshirani 1994), to obtain the confidence intervals for the estimated response, I bootstrap each estimation 1000 times. I specify both `index` and `cutpoints` for constraining the estimation to the common support region.

The command prints the result of the regression estimation. However, it is worth noticing that coefficients have no causal interpretation (Hirano and Imbens 2004; Bia and Mattei 2008), and the various tests reported are biased because GPS is a generated regressor (Wooldridge 2010). Only testing if terms including GPS are jointly different from 0 would be informative since it could be considered a test of whether exogenous covariates introduce bias (Hirano and Imbens 2004; Bia and Mattei 2008).

```
. * gpsMDPolEst
. gpsMDPolEst Y T1 T2, gpsmd(GPS)  exogenous(X1 X2 X3 X4 X5 X6 X7) ///
> model("T1 + T2  + GPS + T1*GPS + T2*GPS") ///
> file_pred(ExampleStata) numboot(1000) dividingint(3) index("p50") cutpoints(2)

****************
The regression estimating the dose-response function is calculated only on the common
support. The output is the following:
****************


      Source |       SS           df       MS      Number of obs   =       739
-------------+----------------------------------   F(5, 733)       =    182.20
       Model |  601889.434         5  120377.887   Prob > F        =    0.0000
    Residual |   484292.11       733  660.698649   R-squared       =    0.5541
-------------+----------------------------------   Adj R-squared   =    0.5511
       Total |  1086181.54       738  1471.79071   Root MSE        =    25.704

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          T1 |   2.564899   .2137103    12.00   0.000     2.145341    2.984456
          T2 |   1.667426   .2154419     7.74   0.000     1.244469    2.090383
         GPS |  -243.7405   519.4568    -0.47   0.639    -1263.541      776.06
    I_T1_GPS |  -47.61917   73.55481    -0.65   0.518    -192.0224    96.78404
    I_T2_GPS |   246.3664   75.36452     3.27   0.001     98.41031    394.3224
       _cons |   1.819601   1.901145     0.96   0.339    -1.912738     5.55194
------------------------------------------------------------------------------

****************
It starts the bootstrap, it may take a while
****************


****************
Bootstrap ended
****************
```

I now compare the result of the GPS method with those obtained by using a rightly specified linear regression. I constrain the sample to the common support region. As specified above, I required estimating the responses corresponding to 16 doses. Figure 2-1 reports the responses and the upper and lower bounds for the estimations obtained with the OLS linear regression and the GPS method. Confidence intervals are at the 0.95 level.

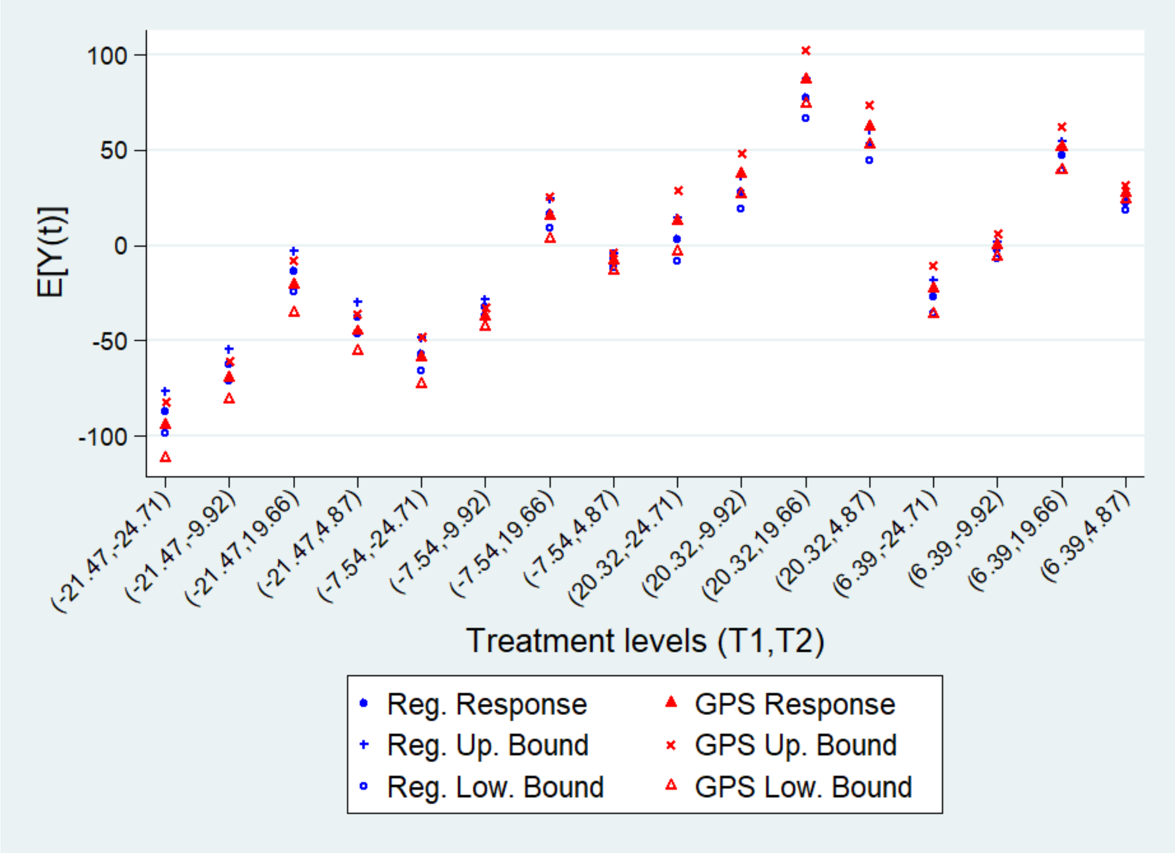
The estimated responses obtained by using the GPS method are qualitatively and quantitatively similar to those obtained by using a rightly specified regression. Indeed, they are statistically equal. This suggests that the GPS method succeeds in reducing bias.

```
. *comparison with a regression
. mat define matrtreat= e(matrtreat)
. mat define returnresults=e(returnresults)
. qui: reg c.Y c.T1 c.T2 c.X1 c.X2 c.X3 c.X4 c.X5 c.X6 c.X7 if Commonsupport==0
. qui: margins, at( T1= `=matrtreat[1,1]' T2= `=matrtreat[1,2]' )
. mat define temp=r(table)
. mat define resreg=(`=matrtreat[1,1]', `=matrtreat[1,2]', temp[5,1], temp[1,1], temp[6,1])
. forvalues i=2(1)`=rowsof(matrtreat)' {
 2.     qui: margins, at( T1= `=matrtreat[`i',1]' T2= `=matrtreat[`i',2]' )
 3.     mat define temp=r(table)
 4.     mat define resreg=(resreg \ (`=matrtreat[`i',1]', `=matrtreat[`i',2]', temp[5,1],
        temp[1,1], temp[6,1]))
 5. }
. mat colnames resreg= T1 T2 LBReg ResponseReg UBReg
. mat define resreg=(resreg, returnresults[1...,9], returnresults[1...,3]
,returnresults[1...,6])

. *I generate the graph
. clear
. set obs 16
. svmat resreg, names(col)
. gen treatment_levels="(" + string(T1, "%9.2f") + "," + string(T2, "%9.2f") + ")"
. encode treatment_levels, gen("id")
. label variable id "Treatment levels"
. label variable ResponseReg "Reg. Response"
. label variable LBReg "Reg. Low. Bound"
. label variable UBReg "Reg. Up. Bound"
. label variable Response "GPS. Response"
. label variable BootL_response "GPS. Low. Bound"
. label variable BootH_response "GPS. Up. Bound"

. graph twoway (scatter ResponseReg id, mcolor(blue) msymbol(circle) msize(small)) ///
> (scatter UBReg id , mcolor(blue) msymbol(plus) msize(small)) ///
> (scatter LBReg id , mcolor(blue) msymbol(circle_hollow) msize(small)) ///
> (scatter Response id , mcolor(red) msymbol(triangle) msize(small)) ///
> (scatter BootH_response id , mcolor(red) msymbol(x) msize(medium)) ///
> (scatter BootL_response id , mcolor(red) msymbol(triangle_hollow ) msize(small)), ///
> title("Regression and GPS estimations correspondence") xtitle("Treatment levels (T1,T2)")
///
> ytitle("E[Y(t)]", size(med)) ///
> xlabel(1(1)16 , labsize(small) valuelabel angle(45)) ///
> ylabel(, labsize(small) valuelabel angle(0)) ///
> legend(colfirst) ///
> name(CompGPSReg, replace)
```

Figure 2-1 Regression and GPS estimations. For both regression estimates and GPS estimates: point estimate, upper bound, lower bound. Bounds are calculated at the 0.95 confidence level.



2.8 Conclusions and a proposal for overcoming small common support

Causal evaluation has generally focused on binary or continuous treatments. In this chapter, I present a novel Stata package, gpsMD, which implements Egger and von Ehrlich's (2013)' extension of the GPS method to the cases when treatment is multidimensional and continuous. After having summarized the econometric framework and described the commands, I present a simple simulated dataset for the reader to familiarize herself with the commands. Moreover, I compare the results obtained with the GPS method with those obtained employing a rightly specified linear regression. As expected, the dose-response estimated with the GPS method is not statistically different from those obtained with the regression, thus suggesting that the GPS method effectively estimates causal parameters in the case of multidimensional continuous treatments. Nevertheless, the implementation also indicates a potential problem when applying the GPS method: the lack of common support. Although restricting the sample improves the reliability of the estimation, it may exclude from the analysis interesting observations.

The criterion used to define the common support exploits the entire range of the treatment observed in the sample. At first, we divide the treatment into an arbitrary number of subsets. Then, assuming that every arbitrary subset can be meaningfully represented by a point (mean or median), we transform the continuous treatment into a multi-valued treatment. Second, for each value of the treatment, we calculate its density given the covariates. We remove observations with a pattern of covariates whose corresponding density is not – in the sample – sufficiently represented in both the treated (the set of observation that has a treatment falling in the interval represented by the value) and the control (all the other observations) group. The observations that remain constitute the common support region, and our estimate of the dose-response function is reliable for this group only. Generally, the analysis is also constrained to the treatment range in the resulting sample.

In these conclusions, I propose that we could act iteratively to include the observations that are not in the common support in the analysis. Once we have made the first selection, we can use the range of the treatment of the highest [lowest] part of the distribution (which is a subset of the starting range) and finding a common support considering only that range. The further analysis can only generalize to the resulting commons support regions and the resulting range of the treatment. It is worth noting that although a restriction on the treatment informs the selection, ultimately, the sample is selected considering the exogenous covariates. The rationale of the criterion proposed is that the observations outside the common support region calculated in the first step present patterns of observables that theoretically almost impede (while empirically simply impede) to have a treatment belonging to a given interval. Therefore, it would not be *empirically* unreasonable to treat them as a sample from a population on his own. In other words, the continuous treatment would be divided into a family of intervals empirically considered qualitatively different. For any of these treatments, only a subset of the original population would be considered eligible.

2.9 Appendix 2.1

2.9.1 Proof of the Balancing property, Theorem 1 and 2 (Peter H. Egger and von Ehrlich 2013)

Before the various proofs, we report Theorem 2.1.8 (Casella and Berger 2002, 2:53) about the transformation of random variables. Proofs assume the propensity score is a function satisfying the premises of Theorem 2.1.8.

Theorem 2.1.8 (Casella and Berger 2002, 2:53): Let X have a pdf $f_X(x)$, let $Y = g(X)$, and define the sample space $\mathcal{X}$ as $\mathcal{X} = \{x: f_X(x) > 0\}$. Suppose there exists a partition, $A_0, A_1, \dots, A_k$, of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exist functions $l_1(x), \dots, l_k(x)$ defined on $A_1, \dots, A_k$, respectively satisfying:

i. $l(x) = l_i(x)$, for $x \in A_i$

ii. $l_i(x)$ is monotone on A_i

iii. the set $\mathcal{Y} = \{y: y = l_i(x)\ for\ some\ x \in A_i\}$ is the same for each $i = 1, \dots, k$

and

iv. $l_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, for each $i = 1, \dots, k$

then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\left(l_i^{-1}(y)\right)\left|\frac{dl_i^{-1}(y)}{dy}\right| & y \in \mathcal{Y} \\ 0\ otherwise & \end{cases}$$

To simplify the notation, I suppress the subscript for the individuals in this section. I denote probability density functions as $f(.)$. I suppress the subscripts indicating the random variable from the notation because it is obvious from the context (e.g. $f_Y(y) = f(y)$).

Lemma 1: $f(\boldsymbol{t}|f(\boldsymbol{t}|\boldsymbol{Z})=k)=k$

Proof

$$f(\boldsymbol{t}|f\ (\boldsymbol{t}|\boldsymbol{Z})=k)=\frac{f\big(\boldsymbol{t},f(\boldsymbol{t}|\boldsymbol{Z})\big)}{f\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}=\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\left(\boldsymbol{t}\,,l_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)\right)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\left(l_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)\right)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$=\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\left(\boldsymbol{t}\mid l_i^{-1}\big(f(\boldsymbol{T}|\boldsymbol{Z})\big)\right)f\big(l_i^{-1}\big(f(\boldsymbol{T}|\boldsymbol{Z})\big)\ \big)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\left(l_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)\right)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$=k\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\big(l_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)\ \big)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k}f\left(l_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)\right)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}=k$$

Balancing property: $f(t|\boldsymbol{Z},g(\boldsymbol{t},\boldsymbol{Z}))=f(\boldsymbol{t}|\ g(\boldsymbol{t},\boldsymbol{Z}))$

Proof

$$f(\boldsymbol{t}|\boldsymbol{Z},g(\boldsymbol{t},\boldsymbol{Z}))=f(\boldsymbol{t}|\boldsymbol{Z})=f\big(\boldsymbol{t}\big|f(\boldsymbol{t}|\boldsymbol{Z})\big)=f\big(\boldsymbol{t}\big|\ g(\boldsymbol{t},\boldsymbol{Z})\big)$$

Theorem 1: $Y_i(\boldsymbol{t})\perp\boldsymbol{T_i}|g(\boldsymbol{t},\boldsymbol{Z_i}),\forall\boldsymbol{t}\in\mathfrak{I}$

Proof

We need to show that $f\big(\boldsymbol{t}\big|g(\boldsymbol{t},\boldsymbol{Z}),Y(\boldsymbol{t})\big)=f\big(\boldsymbol{t}\big|g(\boldsymbol{t},\boldsymbol{Z})\big)$. Then we show that both sides are equal to $g(\boldsymbol{t},\boldsymbol{Z})=k$

a)

$$f\left(\boldsymbol{t}|g(\boldsymbol{t},\boldsymbol{Z})=k,Y(\boldsymbol{t})\right)=\frac{f\left(\boldsymbol{t},f(\boldsymbol{t}|\boldsymbol{Z}),Y(\boldsymbol{t})\right)}{f\left(f(\boldsymbol{t}|\boldsymbol{Z}),Y(\boldsymbol{t})\right)}$$

$$=\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(\boldsymbol{t}\,,l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(l_i^{-1}\left(f(t|Z)\right),Y(t)\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$=\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(\boldsymbol{t}\mid l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right) f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$=\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(\boldsymbol{t}\mid l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)\right) f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$=k\frac{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}{\sum_{i:f(\boldsymbol{t}|\boldsymbol{Z})=k} f\left(l_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\left(f(\boldsymbol{t}|\boldsymbol{Z})\right)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}=k$$

b)

$$f(\boldsymbol{t}|g(\boldsymbol{t},\boldsymbol{Z})=k)=f(\boldsymbol{t}|\boldsymbol{Z},g(\boldsymbol{t},\boldsymbol{Z}))=f(\boldsymbol{t}|\boldsymbol{Z})=k$$

Theorem 2 (Peter H. Egger and von Ehrlich 2013): Denote $\mu(\boldsymbol{t},\boldsymbol{g})\equiv E[Y(\boldsymbol{t})|g(\boldsymbol{t},\boldsymbol{Z})]$. Under weak unconfoundedness:

i. $\mu(\boldsymbol{t},\boldsymbol{g})\equiv E[Y(\boldsymbol{t})|g(\boldsymbol{t},\boldsymbol{Z})=k]=E[Y|\boldsymbol{T}=\boldsymbol{t},G=k]$
ii. $\mu(\boldsymbol{t})=E_{g(\boldsymbol{t},\boldsymbol{Z})}\left[\mu\left(\boldsymbol{t},g(\boldsymbol{t},\boldsymbol{Z})\right)\right]$

Proof

See Egger and von Ehrlich (2013)

2.9.2 Proof that a balancing function works well as the propensity score

Let $b(\boldsymbol{T},\boldsymbol{Z})$ a function such that the balancing property is satisfied. This means:

$$f(\boldsymbol{t}|\boldsymbol{Z},b(\boldsymbol{t},\boldsymbol{Z}))=f(\boldsymbol{t}|\,b(\boldsymbol{t},\boldsymbol{Z}))$$

Lemma 1 – Balancing function: $b(\boldsymbol{t},\boldsymbol{z}) = b(\boldsymbol{t},\boldsymbol{m}) \Rightarrow g(\boldsymbol{t},\boldsymbol{z}) = g(\boldsymbol{t},\boldsymbol{m})$

Proof

Similarly to Theorem 2 in Rosenbaum and Rubin (1983).

By absurd, if $g(\boldsymbol{t},\boldsymbol{z}) \neq g(\boldsymbol{t},\boldsymbol{m})$ and $b(\boldsymbol{t},\boldsymbol{z}) = b(\boldsymbol{t},\boldsymbol{m})$, then $f(\boldsymbol{t}|\boldsymbol{z}) \neq f(\boldsymbol{t}|\boldsymbol{m})$ and $b(\boldsymbol{t},\boldsymbol{Z})$ cannot balance.

Indeed if $b(\boldsymbol{t},\boldsymbol{m})$ = $b(\boldsymbol{t},\boldsymbol{z})$ and $b(\boldsymbol{t},\boldsymbol{Z})$ is a balancing function it cannot be that:

$$f(\boldsymbol{t}|b(\boldsymbol{t},\boldsymbol{z})\,) = f(\boldsymbol{t}|\boldsymbol{z},b(\boldsymbol{t},\boldsymbol{z})\,) = f(\boldsymbol{t}|\boldsymbol{z}) = g(\boldsymbol{t},\boldsymbol{z}) \;\neq\; g(\boldsymbol{t},\boldsymbol{m}) = f(\boldsymbol{t}|\boldsymbol{m})$$

$$= f(t|\boldsymbol{m},b(\boldsymbol{t},\boldsymbol{m})) = f(\boldsymbol{t}|b(\boldsymbol{t},\boldsymbol{m}))$$

With a balancing function, the theorems become:

Theorem 1 - balancing function: $Y_i(\boldsymbol{t}) \perp \boldsymbol{T_i}|b(\boldsymbol{t},\boldsymbol{Z_i}), \forall \boldsymbol{t} \in \Im$

Proof

We need to show that $f(\boldsymbol{t}|\ b(\boldsymbol{t},\boldsymbol{Z}),Y(\boldsymbol{t})) = f(\boldsymbol{t}|\ b(\boldsymbol{t},\boldsymbol{Z}))$. Then we show that both sides are equal to $g(\boldsymbol{t},\boldsymbol{Z}) = g$.

a)

$$f\big(\boldsymbol{t}\big|b(\boldsymbol{t},\boldsymbol{Z}) = k, Y(\boldsymbol{t})\big) = \frac{f\big(\boldsymbol{t},b(\boldsymbol{t},\boldsymbol{Z}),Y(\boldsymbol{t})\big)}{f\big(b(\boldsymbol{t},\boldsymbol{Z}),Y(\boldsymbol{t})\big)}$$

$$= \frac{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f\left(\boldsymbol{t}\,,l_i^{-1}(b(\boldsymbol{t},\boldsymbol{Z})),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big)}{db(\boldsymbol{t},\boldsymbol{Z})}\right|}{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f\left(l_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$= \frac{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f\left(\boldsymbol{t}\mid l_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\, f\left(l_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}(b(\boldsymbol{t},\boldsymbol{Z}))}{db(\boldsymbol{t},\boldsymbol{Z})}\right|}{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f\left(l_i^{-1}\big(b(\boldsymbol{t}|\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}(b(\boldsymbol{t},\boldsymbol{Z}))}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|}$$

$$= g\,\frac{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f\left(l_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\big(b(\boldsymbol{t},\boldsymbol{Z})\big)}{db(\boldsymbol{t},\boldsymbol{Z})}\right|}{\sum_{i:b(\boldsymbol{t},\boldsymbol{Z})=k} f_Z\left(l_i^{-1}\big(b(\boldsymbol{t}|\boldsymbol{Z})\big),Y(\boldsymbol{t})\right)\left|\frac{dl_i^{-1}\big(f(\boldsymbol{t}|\boldsymbol{Z})\big)}{df(\boldsymbol{t}|\boldsymbol{Z})}\right|} = g$$

b)

$$f(\boldsymbol{t}|b(\boldsymbol{t},\boldsymbol{Z}) = k) = f(\boldsymbol{t}|\boldsymbol{Z},b(\boldsymbol{t},\boldsymbol{Z})) = f(\boldsymbol{t}|\boldsymbol{Z}) = g$$

Theorem 2 is similar to Theorem 2 for the propensity score.

2.10 Appendix 2.2

2.10.1 Akaikemax

The command identifies the combination of variables that minimizes the chosen information criterion. Only linear regression models are supported. It works simply by brutal force. Given the rules specified in the options, the list of all the possible combinations of variables is generated. Then, by using each combination of variables, a regression is run and the information criterion computed. Finally, the model that minimizes the criterion is chosen.

2.10.2 The Syntax

```
Akaikemax, outcome(varlist max=1) power(numlist integer max=1)
       [baseVar(varlist min=1) controls(varlist)
       noIntNoPow(varlist) aloneandpow(varlist) aloneVars(varlist)
       ic(string) stopiflarge(string) reg_opt(string)]
```

`outcome (varlist max=1)`: the outcome variable.

`baseVar (varlist min=1)`: the variables that it is required to be always in the model. These variables are also interacted with controls and/or exponentiated as indicated in power.

`power (numlist integer max=1)`: the power to which interactions (if needed) and basevars will be exponentiated.

`controls (varlist)`: the variables that will be interacted and exponentiated with BaseVars (if the user also wants the variables alone, she should specify them in noIntNoPow or aloneandpow or aloneVars). If specified, also `baseVar` must be specified.

`aloneandpow(varlist)`: Variables that will enter alone and exponentiated in the model.

`noIntNoPow (varlist)`: Variables that will enter in the combinations simply as they are (no interacted nor exponentiated).

`aloneVars (varlist)`: Variables that will always enter in the model simply as they are (neither interacted nor exponentiated).

`ic(string)` : the information criteria AIC (default) or BIC.

`stopiflarge(string)`: the combinations can be unexpectedly many. The default is that the program stops if the number of combinations is higher than 2^20. If set to "`N`" the program does not stop, and only the pc of the user or Stata limits themselves have value.

`reg_opt(string)`: a string with all the options the user would put in regress (e.g., no constant).

2.10.3 Variables generated

The command generates variables for the interaction and the exponentiation (Only those variables belonging to the preferred model will be kept). They all start with `I_` or `P_`.

2.10.4 e-class object

Macro

`e(bestmodel)`: the command for the linear regression with the minimum information criterion.

`e(ic)` : the information criterion chosen.

`e(cmd)`: the name of the command just launched.

`e(cmdline)`: The command just launched.

`e(NAttempt)`: The number of combinations tested.

`e(NewVarInBest)`: a list with the variables in the best model that the program generates.

Scalar

`e(minIC)`: scalar with the Information criterion value of the best model.

2.11 Appendix 2.3

2.11.1 Derivation of the t-test for the balancing property

The sample quantity is:

$$\frac{1}{N}\sum_{g\left(\bar{t}_{T^D},Z\right)^D} N_{g\left(\bar{t}_{T^D},Z\right)^D}\left(\bar{Z}_{T^D\, g\left(\bar{t}_{T^D},Z\right)^D} - \bar{Z}_{T^{-D} g\left(\bar{t}_{T^D},Z\right)^D}\right)$$

With a change of notation, it becomes:

$$\frac{1}{N}\sum_{j}^{K} N_j(\bar{X}_j - \bar{Y}_j)$$

Where $\bar{Z}_{T^D g\left(\bar{t}_{T^D},Z\right)^D} = \bar{X}_j, \bar{Z}_{T^{-D} g\left(\bar{t}_{T^D},Z\right)^D} = \bar{Y}_j$, $N_{g\left(\bar{t}_{T^D},Z\right)^D} = N_j$ for $j = 1, \dots, K$ the index for the intervals of the generalized propensity score. Define N_{X_j}, N_{Y_j} respectively the number of units in the sets X_j and Y_j. By the central limit theorem, see Ross (2004), for large enough N_{X_j}, N_{Y_j} the following holds:

$$\bar{X}_j \sim N\left(\mu_{X_j}, \frac{\sigma_{X_j}^2}{N_{X_j}}\right), j = 1, \dots, K$$

$$\bar{Y}_j \sim N\left(\mu_{Y_j}, \frac{\sigma_{Y_j}^2}{N_{Y_j}}\right), j = 1, \dots, K$$

Then, since sets are assumed independent:

$$\frac{N_j}{N}\bar{X}_j \sim N\left(\frac{N_j}{N}\mu_{X_j}, \left(\frac{N_j}{N}\right)^2 \frac{\sigma_{X_j}^2}{N_{X_j}}\right), j = 1, \dots, K$$

$$\frac{N_j}{N}\bar{Y}_j \sim N\left(\frac{N_j}{N}\mu_{Y_j}, \left(\frac{N_j}{N}\right)^2 \frac{\sigma_{Y_j}^2}{N_{Y_j}}\right), j = 1, \dots, K$$

It follows that the distribution of the sum of the K couples is then:

$$\sum_{j}^{K}\frac{N_j}{N}(\bar{X}_j-\bar{Y}_j)\sim N\left(\sum_{j}^{K}\frac{N_j}{N}\left(\mu_{X_j}-\mu_{Y_j}\right),\sum_{j}^{K}\left[\left(\frac{N_j}{N}\right)^2\left(\frac{\sigma_{X_j}^2}{N_{X_j}}+\frac{\sigma_{Y_j}^2}{N_{Y_j}}\right)\right]\right)$$

And thus:

$$\frac{\sum_j^K\frac{N_j}{N}(\bar{X}_j-\bar{Y}_j)-\sum_j^K\frac{N_j}{N}\left(\mu_{X_j}-\mu_{Y_j}\right)}{\sqrt{\sum_j^K\left[\left(\frac{N_j}{N}\right)^2\left(\frac{\sigma_{X_j}^2}{N_{X_j}}+\frac{\sigma_{Y_j}^2}{N_{Y_j}}\right)\right]}}\sim N(0\,,1)$$

The estimators for the variances $\sigma_{\cdot}^2$ are

$$S_{X_j}^2=\frac{1}{N_{X_j}-1}\sum_{i=1}^{N_{X_j}}\left(X_{ji}-\bar{X}_j\right)^2,j=1,\dots,K$$

$$S_{Y_j}^2=\frac{1}{N_{Y_j}-1}\sum_{i=1}^{N_{Y_j}}\left(Y_{ji}-\bar{Y}_j\right)^2,j=1,\dots,K$$

The well-known result (Ross 2004) is:

$$\frac{\left(N_{X_j}-1\right)}{\sigma_{X_j}^2}S_{X_j}^2\sim\chi_{N_{X_j}-1}^2\,,j=1,\dots,K$$

$$\frac{\left(N_{Y_j}-1\right)}{\sigma_{Y_j}^2}S_{Y_j}^2\sim\chi_{N_{Y_j}-1}^2\,,j=1,\dots,K$$

Then

$$\sum_{j}^{K}\left(\frac{\left(N_{X_j}-1\right)}{\sigma_{X_j}^2}S_{X_j}^2+\frac{\left(N_{Y_j}-1\right)}{\sigma_{Y_j}^2}S_{Y_j}^2\right)\sim\chi_{N-2K}^2$$

If we assume that $\sigma^2_{X_j} = \sigma^2_{X_{j'}} = \sigma^2_{Y_j} = \sigma^2_{Y_{j'}} = \sigma^2$:

$$\frac{\sum_j^K \frac{N_j}{N}(\bar{X}_j - \bar{Y}_j) - \sum_j^K \frac{N_j}{N}\left(\mu_{X_j} - \mu_{Y_j}\right)}{\sqrt{\sum_j^K \left[\left(\frac{N_j}{N}\right)^2 \left(\frac{\sigma^2_{X_j}}{N_{X_j}} + \frac{\sigma^2_{Y_j}}{N_{Y_j}}\right)\right]}} = \frac{\frac{1}{N}\sum_j^K N_j(\bar{X}_j - \bar{Y}_j) - \sum_j^K N_j\left(\mu_{X_j} - \mu_{Y_j}\right)}{\frac{\sigma}{N}\sqrt{\sum_j^K \left[{N_j}^2\left(\frac{1}{N_{X_j}} + \frac{1}{N_{Y_j}}\right)\right]}}$$

$$\sum_j^K \left(\frac{\left(N_{X_j} - 1\right)}{\sigma^2_{X_j}} S^2_{X_j} + \frac{\left(N_{Y_j} - 1\right)}{\sigma^2_{Y_j}} S^2_{Y_j}\right) = \frac{1}{\sigma^2}\sum_j^K \left[\left(N_{X_j} - 1\right) S^2_{X_j} + \left(N_{Y_j} - 1\right) S^2_{Y_j}\right]$$

Then:

$$\frac{\sum_j^K N_j\left(\bar{X}_j - \bar{Y}_j\right) - \sum_j^K N_j\left(\mu_{X_j} - \mu_{Y_j}\right)}{\sigma\sqrt{\sum_j^K \left[{N_j}^2\left(\frac{1}{N_{X_j}} + \frac{1}{N_{Y_j}}\right)\right]}} \cdot \sqrt{\frac{\sigma^2(N - 2K)}{\sum_j^K \left[\left(N_{X_j} - 1\right) S^2_{X_j} + \left(N_{Y_j} - 1\right) S^2_{Y_j}\right]}} =$$

$$\frac{\sum_j^K N_j(\bar{X}_j - \bar{Y}_j) - \sum_j^K N_j\left(\mu_{X_j} - \mu_{Y_j}\right)}{\sqrt{\sum_j^K \left[{N_j}^2\left(\frac{1}{N_{X_j}} + \frac{1}{N_{Y_j}}\right)\right]}} \cdot \sqrt{\frac{(N - 2K)}{\sum_j^K \left[\left(N_{X_j} - 1\right) S^2_{X_j} + \left(N_{Y_j} - 1\right) S^2_{Y_j}\right]}} \sim t_{N-2K}$$

The p-value is then[33]

$$\text{pvalue} = 2 \cdot \left\{ 1 - F_{t_{N-2K}} \left(\left| \frac{\sum_j^K N_j\left(\bar{X}_j - \bar{Y}_j\right) - 0}{\sqrt{\sum_j^K \left[{N_j}^2 \left(\frac{1}{N_{X_j}} + \frac{1}{N_{Y_j}} \right) \right]}} \cdot \sqrt{\frac{(N - 2K)}{\sum_j^K \left[\left(N_{X_j} - 1 \right) S_{X_j}^2 + \left(N_{Y_j} - 1 \right) S_{Y_j}^2 \right]}} \right| \right) \right\}$$

[33] In order to avoid precision problems, the actual formula exploits the fact that the t distribution is symmetric (Gould 2006).

2.12 References

Bia, Michela, Carlos A. Flores, Alfonso Flores-Lagunes, and Alessandra Mattei. 2014. 'A Stata Package for the Application of Semiparametric Estimators of Dose–Response Functions'. *The Stata Journal* 14 (3). SAGE Publications Sage CA: Los Angeles, CA: 580–604.

Bia, Michela, and Alessandra Mattei. 2008. 'A Stata Package for the Estimation of the Dose-Response Function through Adjustment for the Generalized Propensity Score': *The Stata Journal*, September.

———. 2012. 'Assessing the Effect of the Amount of Financial Aids to Piedmont Firms Using the Generalized Propensity Score'. *Statistical Methods & Applications* 21 (4): 485–516.

Caliendo, Marco, and Sabine Kopeinig. 2008. 'Some Practical Guidance for the Implementation of Propensity Score Matching'. *Journal of Economic Surveys* 22 (1): 31–72.

Cameron, Adrian Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata*. Vol. 5. Stata press College Station, TX.

Carpenter, James, and John Bithell. 2000. 'Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians'. *Statistics in Medicine* 19 (9). Wiley Online Library: 1141–64.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. Vol. 2. Duxbury Pacific Grove, CA.

Cavanaugh, Joseph E., and Andrew A. Neath. 2019. 'The Akaike Information Criterion: Background, Derivation, Properties, Application, Interpretation, and Refinements'. *WIREs Computational Statistics* 11 (3): e1460.

Cerulli, Giovanni. 2015. *Econometric Evaluation of Socio-Economic Programs*. Vol. 49. Advanced Studies in Theoretical and Applied Econometrics. Berlin, Heidelberg: Springer Berlin Heidelberg.

Dobrow, Robert P. 2013. *Probability: With Applications and R*. John Wiley & Sons.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Egger, Peter H., Maximilian v. Ehrlich, and Douglas R. Nelson. 2020. 'The Trade Effects of Skilled versus Unskilled Migration'. *Journal of Comparative Economics*, January, S0147596718304189.

Egger, Peter H, and Katharina Erhardt. 2014. 'Determinants of Firm-Level Investment and Exporting'. *Manuscript, ETH Zurich*, 35.

Egger, Peter H., and Andrea Lassmann. 2018. 'The Impact of Common Native Language and Immigration on Imports'. *The World Economy* 41 (7): 1903–16.

Egger, Peter H., and Maximilian von Ehrlich. 2013. 'Generalized Propensity Scores for Multiple Continuous Treatment Variables'. *Economics Letters* 119 (1): 32–34.

Egger, Peter Hannes, and Peter Egger. 2016. 'Heterogeneous Effects of Tariff and Nontariff Policy Barriers in General Equilibrium'. Kiel und Hamburg: ZBW-Deutsche Zentralbibliothek für ….

Erhardt, Katharina. 2017. 'On the Heterogeneous Effects of Trade an Fiscal Policy'. PDF. ETH Zurich.

Flores, Carlos A., Alfonso Flores-Lagunes, Arturo Gonzalez, and Todd C. Neumann. 2012. 'Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps'. *Review of Economics and Statistics* 94 (1). MIT Press: 153–71.

Gould, William. 2006. 'Mata Matters: Precision'. *The Stata Journal* 6 (4). SAGE Publications Sage CA: Los Angeles, CA: 550–60.

Guardabascio, Barbara, and Marco Ventura. 2014. 'Estimating the Dose–Response Function through a Generalized Linear Model Approach'. *The Stata Journal: Promoting Communications on Statistics and Stata* 14 (1): 141–58.

Guo, Shenyang, and Mark W. Fraser. 2015. *Propensity Score Analysis*. SAGE.

Hernán, Miguel A, and James M Robins. 2006. 'Estimating Causal Effects from Epidemiological Data'. *Journal of Epidemiology and Community Health* 60 (7): 578–86.

Hirano, Keisuke, and Guido W. Imbens. 2004. ‘The Propensity Score with Continuous Treatments’. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164:73–84.

Imai, Kosuke, and David A. Van Dyk. 2004. ‘Causal Inference with General Treatment Regimes: Generalizing the Propensity Score’. *Journal of the American Statistical Association* 99 (467). Taylor & Francis: 854–66.

Imbens, Guido W. 2000. ‘The Role of the Propensity Score in Estimating Dose-Response Functions’. *Biometrika* 87 (3): 706–10.

King, Gary, and Richard Nielsen. 2019. ‘Why Propensity Scores Should Not Be Used for Matching’. *Political Analysis* 27 (4).

Lechner, Michael. 2008. ‘A Note on the Common Support Problem in Applied Evaluation Studies’. *Annales d’Économie et de Statistique*, no. 91/92. [GENES, ADRES]: 217–35.

Lee, Wang-Sheng. 2013. ‘Propensity Score Matching and Variations on the Balancing Test’. *Empirical Economics* 44 (1). Springer: 47–80.

Leuven, Edwin, and Barbara Sianesi. 2003. ‘PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing’.

Mander, Adrian. 2019. ‘PLOTMATRIX: Stata Module to Plot Values of a Matrix as Different Coloured Blocks’. Boston College Department of Economics.

Pearl, Judea. 2010. ‘The Foundations of Causal Inference’. *Sociological Methodology* 40 (1). Wiley Online Library: 75–149.

Präg, Patrick. 2019. ‘Visualizing Individual Outcomes of Social Mobility Using Heatmaps’. *Socius* 5 (January). SAGE Publications: 2378023119855486.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. ‘The Central Role of the Propensity Score in Observational Studies for Causal Effects’. *Biometrika* 70 (1). Oxford University Press: 41–55.

Ross, Sheldon M. 2004. *Introduction to Probability and Statistics for Engineers and Scientists*. 3rd ed. Academic Press.

Rostam-Afschar, Davud, and Robin Jessen. 2014. ‘GRAPH3D: Stata Module to Draw Colored, Scalable, Rotatable 3D Plots’. Boston College Department of Economics.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

3 Choosing the right expenditure mix: An evaluation of the EU's regional policy using generalized propensity scores for multiple continuous treatments

Abstract

The evaluation of European regional policy has mainly focused on the overall effectiveness of the policy, thus neglecting the heterogeneous effects due to different policy mixes. This chapter exploits the novel econometric framework proposed by Egger and von Ehrlich (2013) to investigate how different combinations of infrastructure and productive investments impact regions' growth rates. Results depict a four-class typology based on the allocation intensity in the two dimensions. The main results are that allocation in infrastructure has a positive effect only if it is associated with expenditures in productive investments and that the maximal impact on growth is obtained by a policy allocating with high intensity in both dimensions. The extent of misallocation is then assessed by generating two scenarios. In the first one, every region chooses the best mix available under the constraint of the actual funds received. In the second, each region was unconstrained. The comparison with the actual allocation shows that, although the regions can allocate more efficiently, the actual transfer intensity is not enough to choose the mix that would globally maximize growth. Results are consistent with Becker et al. (2012) since enforcing common support restricts the analysis to regions with low transfer intensity.

Keywords: continuous multiple treatments, policy evaluation, EU regional policy, optimal policy mix, regional growth

3.1 Introduction

Cohesion policy is an extensive investment program representing roughly one-third of the European budget. It mainly aims to reduce disparities between European regions and foster growth and convergence (Brunazzo 2016). The expenditures are articulated in 7 years duration planning periods. The regional policy mainly exploits three financial instruments (Stephenson 2016; Olsen 2020, chap. 3): 1) The European Regional Development Fund (ERDF), 2) The European Social Fund (ESF), The Cohesion Fund. Every fund has its own priorities. ERDF mainly aims to reduce disparities within the EU. It primarily focuses on innovation and research, support for SMEs, and the development of infrastructure. ESF is, instead, finalized to interventions that promote employment and educational opportunities. Finally, Cohesion Fund has been mainly designed for promoting sustainable development in lagging countries. Together, fund priorities define a classification of the projects which can be financed[34].

The extant literature provides comprehensive studies about the possible impact of the EU regional policy (Crescenzi and Giua 2016). Nonetheless, evaluation exercises have mainly focused on the overall effectiveness of the investments in enhancing regional growth and have not fully investigated the heterogeneity in effect due to different expenditure strategies or regional peculiarities (Percoco 2017). The present study aims to provide insights into the heterogeneity of the effects that policymakers could obtain by mixing different investment aids in EU regions.

The study focuses on 240 EU regions during the programming period 2007-2013. The analysis is restricted to ERDF and CF funds. In contrast with previous evaluations, which focused on the total amount of funds received by a region, I focus on how the funds are allocated between two categories of expenditures. The treatment I consider consists of the mix of investments in infrastructure and productive investments and technical assistance (from now on called simply “productive investments”) decided in a region during the programming period. Technically, the treatment is conceived as a bi-dimensional vector whose components consist of the two categories. The focus on these two categories is motivated by several strands of literature on the local development determinants, according to which infrastructural investments are not sufficient to guarantee growth. Complementary measures supporting the intangible ecosystem mechanisms, and institutional efforts are needed to ensure economic development (Capello 2010; Pike, Rodríguez-Pose, and Tomaney 2016; Rodríguez-Pose 2020; Hassink, Isaksen, and Trippl 2019).

The empirical strategy employed in this study leverages the method proposed by Egger and von Ehrlich (2013) to evaluate causal impacts in the presence of multidimensional treatments. The proposed framework extends Hirano and Imbens’s (2004) generalized propensity score method for continuous treatments to multiple continuous dimensions. It includes three steps: first, a set of two reduced equations is estimated; second, the generalized propensity score is estimated by assuming

[34] See Molle (2007) for a less concise treatment of the topic.

residuals having a multivariate normal distribution; third, the generalized propensity score is used in a flexible control function (a polynomial with interaction and power up to two) to estimate the dose-response function.

Results show that the two categories cause different impacts on the outcome variable. In particular, I can distinguish four cases based on the allocation intensity in the two dimensions. The main results are that allocation in infrastructure has a positive effect only if associated with a large amount of productive investments. Moreover, although productive investments generally impact growth positively, the maximal impact on growth is obtained by a policy allocating with high intensity in both dimensions.

Using the dose-response function estimated, I finally generate two hypothetical scenarios to evaluate the inefficiencies in regions' allocation of funds and the distribution of funds between regions. In the first scenario, every region optimally allocates the received funds under the constraint of the actual funds received. In the second scenario, regions were instead unconstrained. The comparison with the actual allocation shows that, although regions can improve their ability to allocate the received funds, they generally do not receive enough financings to choose the global best allocation mix. Therefore, regions could benefit from additional funds.

It is worth noting that because of the common support restriction, the sample has been restricted to the NUTS2 regions showing the highest value of GDP, employment rate, and the lowest degree of transfer intensity between the 240 European regions present in the dataset. Therefore present results cannot be generalized to the most lagging regions. Nevertheless, findings are consistent with Becker et al. (2012), who found that regions with lower transfer intensity, contrarily to regions with high transfer intensity, could benefit from additional funds.

The chapter is structured as follows: the first section presents the relevant literature; the second section introduces the method; the third section includes the description of the dataset; the fourth section presents the results of the estimation of the dose-response function; the fifth presents the analysis of the counterfactual scenarios; the sixth section concludes.

3.2 Literature

Although the chapter's aim is straightforward, it lies at the intersection of several strands of literature. The literature review is thus organized as follows: at first, I review literature about Cohesion policy evaluation. Then, since I will evaluate the impact of policy mixes based on two dimensions, namely investment in infrastructure and productive investments, I will review the mechanisms through which these policies should work.

The choice of focusing on only two broad dimensions is motivated by the sample size which does not enable further breakdown. I aggregate the funds in such a way that the first dimension accounts for the modifications of the tangible assets of a regions while the second dimension – which includes support to enterprise, RND, human resources as well as technical assistance for improving region's institutional quality – accounts for all the attempts to modify the absortive capacity, institutions and the intangible envirorment of a region.

3.2.1 ***Cohesion policy evaluation***

Given the resources spent and the importance of the goal, over the last twenty years, scholars have devoted many efforts to the evaluation of European regional policy (Crescenzi and Giua 2016). Overall, the results are positive, especially in recent works (Fratesi 2016). Becker et al. (2010), considering all of Europe, found that having the Objective 1 status raises GDP per capita by 1.6%. Pellegrini et al. (2013) found similar results. Positive effects on employment have been documented in Giua (2017) and (although limitedly) in Ciani et al. (2015). Nevertheless, regions whose absorptive capacity (i.e., quality of institutions and human capital in a region) is lower benefit less from the programs (Becker, Egger, and Von Ehrlich 2013; Andrés Rodríguez-Pose and Garcilazo 2015; Ederveen, Groot, and Nahuis 2006). Evidence from the south of Italy seems to corroborate these findings (Accetturo, Blasio, and Rossi 2019), and Albanese et al. (2020) found that cohesion funds positively affect firms' total factor productivity but that weak institutions impair it. Other causes of heterogeneity are the level of urbanization (Gagliardi and Percoco 2017) and the service sector's development (Percoco 2017). Moreover, the effect on GDP growth is not linear but strictly concave (Becker, Egger, and von Ehrlich 2012; Cerqua and Pellegrini 2018). To summarize, there has been significant attention to the overall effect of cohesion policy and in what regional features moderate the effect – Crescenzi and Giua (2016) refer to these two streams as “identification” and “contextualization.”

However, as Percoco (2017) noticed, evidence about the different mix of expenditures is still lacking. Extant literature, indeed, considers the treatment as the amount of money received – or merely the status of Objective 1 region. The actual allocation is not considered (to the best of my knowledge, the only one is Percoco (2013))[35], and it is therefore implicitly assumed homogeneous between regions. The lack of understanding about the effects of mixing different types of allocations is not only a problem for the police maker, who would surely benefit from knowing how to optimally set the expenditures (Berkowitz, Monfort, and Pieńkowski 2020). Assuming homogeneity of the treatment could also impair identification. Regional development strategies certainly vary enormously. By using the period 2000-2006, Percoco (2013), with a data-driven approach, founds that Objective 1 regions can be divided into three homogeneous clusters of expenditures, and the same is true for Objective 2 regions (see also Caloffi et al. (2018; 2013)). The focus on regional differences in the expenditure mix seems to be even more important given the “place-based” (Neumark and Simpson 2015) nature of the Cohesion policy (Barca, McCann, and Rodríguez-Pose 2012)[36].

Methodologically, the first wave of studies employed macro theoretical model of growth (de la Fuente 1997) to evaluate cohesion policy (Percoco 2005; Aiello and Pupo 2012; Rodrìguez-Pose and Fratesi 2004; Ederveen, Groot, and Nahuis 2006; Esposti 2007; Dall’Erba and Le Gallo 2008; De la Fuente 2002) – for a meta-analysis see Dall’Erba and Fang (2017). Most of the papers rely on extensions of the neoclassical growth model (de la Fuente 1997; Mankiw, Romer, and Weil 1992; Barro and Sala-i-Martin 1990) and attempt to measure the effect of cohesion funds by adding a covariate measuring cohesion funds to a convergence equation. Apart from the criticism underscored by Esposti (2007) and Berkowitz et al. (2020), one of the potential flaws of this kind of literature is that virtually every study uses a Cobb Douglas functional form for the country production function (Dall’Erba and Fang 2017).

[35] Other studies has focused only on single dimensions (see for instance Ferrara et al. (2017), Albanese et al. (2020)).

[36] Although the economic logic of cohesion policy has changed over time(Iain Begg 2016), after the Barca report (Barca 2009), it has taken the form of place based policies(Crescenzi and Giua 2019).

In general, the use of a Cobb Douglas is legit because, under the assumptions of constant factor shares and constant exponential growth rate of profit and wages, it results from the income account identity $Q_t = W_t + \Pi_t$, where Q_t, W_t, Π_t, are respectively the aggregate value added, the aggregate wages, the aggregate profit at a given time (Felipe and Adams 2005). The problem in using Cobb Douglas specification is twofold. On one side, and this independently from cohesion funds, it is arguable that factor shares do not change over time (Zuleta 2012; Karabarbounis and Neiman 2014; Blanchard, Nordhaus, and Phelps 1997). On the other side, it is questionable that a stream of funds does not change factor shares in a country. If we imagine that funds can be used to increase automation, the assumption of constant factor share seems doubtful (Aghion, Jones, and Jones 2017; Zeira 1998). Moreover, it is reasonable that funds can take the form of subsidies to firms. If it were the case, by increasing the survival rate of the subsidized firms, subsidies would affect the market power distribution. Changes in the market power distribution would in turn affect the labor shares (De Loecker, Eeckhout, and Unger 2020).

A second strand of literature, and to which the present work belongs, remains agnostic in respect of the structural model generating data and adopts instead the technics born within the counterfactual framework (Wooldridge 2010, chap. 18; Morgan and Winship 2007). An example of this approach can be found in Becker et al. (2010), who exploits the threshold of regional per capita GDP lower than 75% of EU per capita GDP[37] to apply a regression discontinuity design (Lee and Lemieux 2010). Giua (2017), within the same framework, leverages the spatial discontinuity of the treatment. Others, as Becker et al. (2012), adopt a matching design for continuous treatment (Hirano and Imbens 2004). Methodologically, Becker et al.'s (2012) paper is the closest to mine. Indeed, my identification strategy represents an extension of propensity score matching for multidimensional treatment (Peter H. Egger and von Ehrlich 2013). In the next section, I will outline the econometric framework used in the chapter.

[37] The threshold that distinguishes objective 1 regions from the others.

3.2.2 Investment in infrastructure

Infrastructure includes transport supply, information, and communication technology (ICT), and energy infrastructure (Bröcker and Rietveld 2010). According to Rietveld and Bruinsma[38] (2012, chap. 3), we should distinguish effects according to two dimensions: temporary-non temporary effects, demand-supply effects. Temporary Demand-side effects concern the increase in the demand and the crowding-out effect generated by the increase in public expenses due to the construction of new transportation infrastructure. Non-temporary effects include the maintenance of the infrastructures. Supply effects are only non-temporary and concern the productivity and location of the firms (Rietveld and Bruinsma 2012).

Regarding transport infrastructure, literature does not agree about its supply effects on regional development (Puga 2002; Crescenzi and Rodríguez-Pose 2012; Rietveld and Bruinsma 2012; Elburz, Nijkamp, and Pels 2017). The basic building block is that transport infrastructures reduce transportation costs but what the reduction in transportation costs generates is not so clear.

One strand of literature is based on an extended aggregate production function able to accommodate the infrastructures. Here infrastructures are seen as a factor of production that influences aggregate total factor productivity (Aschauer 1989). Although some studies relying on this perspective find positive effects of infrastructure on productivity, further works have cast some doubts about the direction of causality (Gramlich 1994). Moreover, additional empirical works have not confirmed the results (Holtz-Eakin and Schwartz 1995; Crescenzi and Rodríguez-Pose 2012).

Another strand of literature, traceable to the works on new economic geography (Puga 2002), argues that decreasing transportation costs influence the process of agglomeration and, in turn, regional disparities. Moderate costs of trade can push firms to cluster together to benefits from economies of scale, thus generating a core-periphery structure (Krugman and Venables 1990). Moreover, decreasing transportation costs can favor the already developed regions by allowing their firms to penetrate the market of peripheral and less developed regions from afar. This would, thus, dampen the development in lagging areas (Puga 2002; Vickerman, Spiekermann, and Wegener 1999).

[38] The typology originally refer to transportation infrastructure, but I believe that the extension to all the infrastructures is painless.

The highly stylized nature of these perspectives has left behind other factors that can influence the effectiveness of infrastructural development. A third strand of literature underscores that infrastructure per se has not a large impact (if any) on development unless some preconditions are met (Banister and Berechman 2001; Crescenzi and Rodríguez-Pose 2012; Andrěs Rodríguez-Pose 1999). Infrastructure investments are seen as complementary to the quality of the labor force and socio-economic institutions. If, for instance, political conditions are met, but the quality of the labor force remains unvaried, investing in infrastructure will increase accessibility. Nevertheless, it will only have the effect of redistributing in the space the actual economic development: the additional growth will be undermined (Banister and Berechman 2001)

Finally, not all infrastructural changes have the same impact. Indeed the return of investment in infrastructure decreases as the network gets more developed (Vickerman, Spiekermann, and Wegener 1999; Rietveld and Bruinsma 2012). Also, Vickerman et al. (1999) show how Europe's high-speed rail network's development has increased the divergence between regions in terms of accessibility[39] due to its strong nodal aspect (Puga 2002)[40]. Therefore, high-speed rail tends to dampen growth in non-nodal places, which will end up being less coveted firm locations (Puga 2002).

Crescenzi et al. (2016) reported evidence that interregional motorway development scarcely affects growth. Secondary roads seem to be more effective since they are less object to corruption and foster links between key local economic actors. Martin and Rogers (1995) found similar results.

Moreover, it is worth noting that infrastructure modifications change not only the transportation costs but also the logistical strategies of the firm. For example, the adoption of the just-in-time principle is conditioned by the availability of transport infrastructure (Rietveld and Bruinsma 2012; Andersson, Anderstig, and Haarsman 1990).

In respect to ICT development, Vu (2011) argues that – by easing communications and the diffusion of knowledge (Tranos 2012) – it can foster growth through three channels: facilitating learning, the creation of new knowledge and the

[39] A general definition as well as a review of operational definition is given in See Rietveld and Bruinsma (2012, chap. 2). The general definition of accessibility is: “it is the potential of opportunities for interaction” (2012, 33).

[40] See also Vickerman (2018) for an updated review on the topic.

adoption of innovations; reducing production costs; improving decision making of both firms and household thus decreasing the misallocation of resources. Empirical research at the macro-level (Vu 2011; Madden and Savage 1998; Del Bo and Florio 2012; Greenstein and McDevitt 2009) and micro-level (Grimes, Ren, and Stevens 2012) confirms the positive impact. Besidesurz (2017) finds in his meta-analysis that ICT seems to be the most effective infrastructural investment. In addition, improving telecommunication in disadvantaged areas should foster convergence in the long run because of the increased marginal product of capital (Cieślik and Kaniewska 2004).

However, the effect is not linear. Koutroumpis (2009)(see also Roller and Waverman (2001)), focusing on broadband adoption in European countries, shows that major gains are obtained when the broadband penetration – the percentage of the population who have adopted a broadband connection – is higher than 20%.

Finally, also for investment in ICT, there is evidence of "conditional effectiveness". In particular, Forman et al. (2012), referring to the years 1995 to 2000 in the USA, found that only counties with the highest presence of high-skilled workers, the highest income, population, and the highest number of IT-intensive firms experienced wage and occupation growth as a consequence of an increase in advanced internet investments. Akerman et al.'s (2015) supplement these results finding that the firm's broadband adoption positively affects high-skilled workers while the effect is negative for a low-skilled worker. The authors explained that new ICT acts as a complement for high skilled workers because they are more likely to perform "non-routine abstract tasks". Low-skilled workers are instead more likely to perform routine tasks, and, in these cases, ICT improvements result in a substitution of human labor (Akerman, Gaarder, and Mogstad 2015). Using USA data, Kolko (2012) similarly found that broadband expansion mainly influences total employment in counties with IT-intensive firms. However, he did not find increases in the employment rate and suggested that the ICT improvement generated a demand for workers that was not satisfied locally.

Further evidence concerning "conditional effectiveness" is provided by Tranos (2012). He evaluates the effect of degree centrality within the international intercity Internet backbone network on GDP growth. He found that only in 28 out of 48 European city-regions degree centrality granger causes GDP growth. The effect is positively correlated (at the city-region level) with the percentage of tertiary sector in the economy, the number of registered patents in high technology, and the number of students[41]. The explanation reported is that ICT is a necessary but not sufficient condition for development: ability and know-how are crucial for exploiting the new technology[42]. Therefore, he argues that regional policy should jointly focus on these capabilities to foster growth (see Capello and Nijkamp (1996) for similar results).

I will not review the relations between energy infrastructures and economic growth extensively. Indeed, changes in energy infrastructures, at least in developed countries, could be qualitatively interpreted in the general frame of reducing the costs of a factor of production (Toman and Jemelkova 2003; Isaksson 2010; Stern, Burke, and Bruns 2019)[43].

3.2.3 ***Productive investment***

According to Berkowitz et al. (2020), it can be identified direct and indirect channels through which support to enterprise and R&D innovation affect growth. Concerning support to enterprise, policies in the form of subsidies, grants, and tax credits can directly enhance growth by decreasing the cost of capital, thus boosting investments. Boosting investments would, in turn, increase the competitiveness of the targeted firms (Berkowitz, Monfort, and Pieńkowski 2020) with the indirect effect of technology spillovers (Berkowitz, Monfort, and Pieńkowski 2020). Possible unexpected consequences can be easing the survival of inefficient firms, thus discouraging new and more technologically advanced firms from entering the market (Caballero, Hoshi, and Kashyap 2008).

A more structural policy can be oriented in developing clusters (Wolman and Hincapie 2015). This policy could take (although not only) the form of encouraging collaboration between enterprises and technology transfer, organizing workshops, offering managerial, technical assistance, and creating sector-specific technology centers (Garone et al. 2015). The agglomeration should increase labor market pooling, reduce transportation costs, and increase technological spillovers (Garone et al. 2015; Wolman and Hincapie 2015). However, the benefits deriving from clusters may

[41] These are respectively proxies for the intensity of the service economy, the innovative regional environment and the knowledge intensity.

[42] Tranos (2012), echoing Cohen and Levintal (1990), calls this ability "absorptive capacity".

[43] However, things could change if we look at the long run development. See Ayres (2013) and Stern (2010) for theoretical discussions on the role of energy in production.

depend on the industry type and stage in the life circle (Mcdonald et al. 2007; Duschl et al. 2015; Potter and Watts 2011).

R&D support can include measures similar to those aimed to develop clusters (cluster development policy includes policies aimed to boost R&D). In this section, I focus on innovation incentives. Overall, the evidence is quite mixed. Di Blasio et al. (2015), evaluating the effectiveness of the Fund for Technological Innovation in Italy, found no evidence of additionality (firms would have spent in R&D the same amount also without perceiving subsidies). However, the fund aimed to subsidize applied innovation. Howell (2017), looking at firms at the early stages of new technology development, found positive results of subsidies on patents and the growth of the firms. Bronzini et al. (2014) evaluate a program implemented in Emilia Romagna (a region in the North of Italy) and find evidence of remarkable additionality, although only for small firms. The heterogeneous effect, biased towards small firms, is also confirmed by the evaluation of Di Gennaro et al. (2019). The findings are probably due to the liquidity constraint that affects small firms (Bronzini and Iachini 2014; Howell 2017).

As in the case of infrastructure, also in R&D investments, socio-economics factors moderate the effectiveness of investments. Andrěs Rodríguez-Pose (1999) calls these factors the region's "social filter" and argues that it determines the permeability of the region to R&D investment and, consequently, the territorial capability to trigger the adoption and the development of innovations. It could be argued that the social filter corresponds to the determinants of the "absorptive capacity" (Tranos 2012) of a region. Crescenzi et al. (2013) distinguish "social filter" in three broad domains that act simultaneously: educational achievement, the structure of productive resources, demography. Both Andrěs Rodríguez-Pose (1999) and Crescenzi et al. (2013) found evidence of the importance of social filter for innovation (see also Xiong et al. (2020)).

The debate about regional development is increasingly moving from a firm-centered perspective toward a multi-actor perspective (Hassink, Isaksen, and Trippl 2019), bringing to the fore the role institutions and administrations have in determining regional growth (Rodríguez-Pose 2020; Capello 2010; Hassink, Isaksen, and Trippl 2019).

Institutions are not only important for the administration of funds (Surubaru 2017). They also condition the regional innovative performance and regional productivity by contributing to determining the economic uncertainty in a region (Rodríguez-Pose and Di Cataldo 2015; Rodríguez-Pose and Ganau 2021; Mokyr 2010). However, institutions and, in particular, administrations are, as firms, important actors in the process that leads to the identification and undertaking changes necessary to new path developments (Hassink, Isaksen, and Trippl 2019; Dawley 2014; MacKinnon et al. 2019; Jolly, Grillitsch, and Hansen 2020). Modifying and innovating the administrations (De Vries, Bekkers, and Tummers 2016) will improve their capability of acting as place leaders (Sotarauta 2017; Beer et al. 2019; Jolly, Grillitsch, and Hansen 2020), their deployment of dynamic capabilities (Labory and Bianchi 2021), and the ability to elaborate well designed and comprehensive policies truly devoted to development instead of the maintenance of the status quo (Bianchi and Labory 2019; Jolly, Grillitsch, and Hansen 2020; Bianchi and Labory 2018).

3.3 Methodology

The methodology employed in the chapter consists of the extension to multiple continuous dimensions of the generalized propensity score (Hirano and Imbens 2004) proposed by Egger and von Ehrlich (2013). The econometric framework is the following.

Given a sample of N units, $\boldsymbol{T}_i = (T_{1i}, \dots, T_{Mi})'$ is the random variable concerning the treatment-experienced the i^{th} unit. M is the number of dimensions of the treatment. In our case $= 2$: investment in infrastructure and productive investments.

The level of the treatment is defined by an m-equation structural model where the reduced equations are defined as:

$$T_{mi} = f(\boldsymbol{Z}_i, \boldsymbol{\gamma}_m) + v_{mi}\, , m = 1, \dots, M \tag{1}$$

$\boldsymbol{Z}_i = \bigcup_m^M X_{mi}$ is the union of the exogenous variables X_{mi} and possibly their interaction terms.

We are interested in the average dose-response function:

$$\mu(\boldsymbol{t}) \equiv E[Y_i(\boldsymbol{t})]$$

Where $Y_i(\boldsymbol{t})$ is the potential outcome for the i^{th} unit when treated with $\boldsymbol{t} \in \Im$. $\Im$ is the set of all possible treatments. For the dose-response function to be identifiable weak unconfoundedness must hold:

$$Y_i(\boldsymbol{t}) \perp \boldsymbol{T}_i | \boldsymbol{Z}_i \ \forall \boldsymbol{t} \in \Im$$

It means that once conditioned on $\boldsymbol{Z}_i$, the potential outcome and the experienced treatment are independent.

The conditional density function of the treatment given the covariates is defined as:

$$g(\boldsymbol{t}, \boldsymbol{z}) \equiv f_{\boldsymbol{T}_i|\boldsymbol{Z}_i}(\boldsymbol{T}_i = \boldsymbol{t} | \boldsymbol{Z}_i = \boldsymbol{z})$$

The generalized propensity score is instead defined as the random variable:

$$\boldsymbol{G}_i = g(\boldsymbol{T}_i, \boldsymbol{Z}_i)$$

The propensity score generates a family of random variables $(\boldsymbol{t}, \boldsymbol{Z}_i)$, $\forall \boldsymbol{t} \in \Im$.

We assume that $\boldsymbol{T}_i | \boldsymbol{Z}_i \sim \mathcal{N}(\ \boldsymbol{f}(\boldsymbol{Z}_i, \boldsymbol{\gamma}_{\boldsymbol{m}}), \boldsymbol{\Sigma})$, $m = 1, \dots, M$, that is: the conditional distribution of the treatment given the covariates is a multivariate normal distribution with constant between observations variance-covariance matrix. This implies that $\boldsymbol{v}_{\boldsymbol{i}} = (v_{1i}, \dots, v_{Mi})' \sim \mathcal{N}(\ \boldsymbol{0}_{\boldsymbol{m}}, \boldsymbol{\Sigma})$ and that the variance-covariance matrix $\boldsymbol{\Sigma}$ is equal $Cov(\boldsymbol{v}_{\boldsymbol{1}}, \dots, \boldsymbol{v}_{\boldsymbol{M}})$ where $\boldsymbol{v}_{\boldsymbol{m}} = (v_{1m}, \dots, v_{Nm})$.

The generalized propensity score for the i^{th} unit is then:

$$G_i = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\ \Sigma)^{\frac{1}{2}}} \exp\{-\frac{1}{2} \boldsymbol{v}_i' \Sigma^{-1} \boldsymbol{v}_i\}$$

While the estimated one is:

$$\hat{G}_i = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\ \hat{\Sigma})^{\frac{1}{2}}} \exp\{-\frac{1}{2} \hat{\boldsymbol{v}}_i' \hat{\Sigma}^{-1} \hat{\boldsymbol{v}}_i\}$$

The estimated quantities are obtained by estimating (1) by OLS.

The propensity score satisfies by construction the balancing property, loosely speaking:

$$\boldsymbol{Z_i} \perp 1\{\boldsymbol{T}_i = \boldsymbol{t}\} | g(\boldsymbol{t}, \boldsymbol{Z}_i) \ \forall \boldsymbol{t} \in \mathfrak{I}$$

It can be shown (Peter H. Egger and von Ehrlich 2013) that weak unconfoundedness and balancing property imply:

$$Y_i(\boldsymbol{t}) \perp \boldsymbol{T}_i | g(\boldsymbol{t}, \boldsymbol{Z_i}), \forall \boldsymbol{t} \in \mathfrak{I}$$

That is: the potential outcome is independent of the treatment once we have conditioned on the propensity score calculated at $\boldsymbol{t}$.

Therefore

$$E[Y_i | \boldsymbol{T}_i = \boldsymbol{t}, g(\boldsymbol{T}_i, \boldsymbol{Z}_i)] = E[Y_i(\boldsymbol{t}) | \boldsymbol{T}_i = \boldsymbol{t}, g(\boldsymbol{t}, \boldsymbol{Z}_i)] = E[Y_i(\boldsymbol{t}) | g(\boldsymbol{t}, \boldsymbol{Z}_i)]$$

And

$$\mu(\boldsymbol{t}) \equiv E[Y_i(\boldsymbol{t})] = E_g\big[E[Y_i(\boldsymbol{t}) | g(\boldsymbol{t}, \boldsymbol{Z}_i)]\big] = E_g\big[E[Y_i | \boldsymbol{T}_i = \boldsymbol{t}, g(\boldsymbol{T}_i, \boldsymbol{Z}_i)]\big]$$

This means that we can retrieve the dose-response function for $\boldsymbol{t}$ by estimating $E[Y_i | \boldsymbol{T}_i, g(\boldsymbol{T}_i, \boldsymbol{Z}_i)]$ with a flexible polynomial of $\boldsymbol{T}_i$ and $g(\boldsymbol{T}_i, \boldsymbol{Z}_i)$. Then, we can use the resulting coefficients to predict $E[Y_i(\boldsymbol{t}) | g(\boldsymbol{t}, \boldsymbol{Z}_i)]$ for each i. Finally, by taking the average of the predictions, we recover $E\widehat{[Y_i(\boldsymbol{t})}]$.

I now summarize how to identify the common support and test the balancing property.

According to Flores et al. (Flores et al. 2012) and Egger's generalization to a multidimensional treatment(Peter Hannes Egger and Egger 2016; Peter H. Egger, Ehrlich, and Nelson 2020), the common support can be selected by partitioning the treatment in an arbitrary number of subsets (two in the present exercise). Then for each discrete subset T^D we chose a representative point $\overline{\boldsymbol{t_D}}$. There, we calculate the propensity score, $G_i(\boldsymbol{t_D}) = g(\boldsymbol{t_D}, \boldsymbol{Z}_i)$, for each observation in the sample.

We then keep all the observations such that:

$$G_i(\overline{\boldsymbol{t_D}}) \in \left[\max\left\{\underset{j\in T^D}{min}\left(G_j(\overline{\boldsymbol{t_D}})\right), \underset{j\notin T^D}{min}\left(G_j(\overline{\boldsymbol{t_D}})\right)\right\}, \min\left\{\max_{j\in T^D}\left(G_j(\overline{\boldsymbol{t_D}})\right), \max_{j\notin T^D}\left(G_j(\overline{\boldsymbol{t_D}})\right)\right\}\right]; for\ D = 1, \dots$$

The balancing property is tested in a similar manner – the procedure is similar to the one in Bia (2008) for the case one-dimensional continuous case. The treatment is partitioned into an arbitrary number of subsets. Then, for each subset, we do the following. A representative point is chosen, and the propensity score at that point is calculated for each unit. The calculated propensity scores are also partitioned in an arbitrary number of subsets. For each exogenous variable, we compute the weighted average of the differences in the mean between the focal subset of the treatment and the others within the same subset of the propensity score. This generates the following test statistics:

$$\frac{1}{N}\sum_{g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D} N_{g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D}\left(\bar{Z}_{T^D g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D} - \bar{Z}_{T^{-D} g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D}\right)$$

Where N is the number of observational units in the sample, and $N_{g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D}$ is the number of observations in a given interval of the propensity score $g(\bar{\boldsymbol{t}}, \boldsymbol{Z})^D$. $\bar{Z}_{T^D g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D}$ is the sample mean of the exogenous variable for those observations that belong to the intersection between the subset of the treatment T^D and interval of the propensity score $g(\bar{\boldsymbol{t}}, \boldsymbol{Z})^D$. $\bar{Z}_{T^{-D} g(\bar{\boldsymbol{t}},\boldsymbol{Z})^D}$ is instead the sample mean of the exogenous variable for those observations that belong to the intersection between the subset of the treatment T^D and interval of the propensity score $g(\bar{\boldsymbol{t}}, \boldsymbol{Z})^D$. A t-test evaluates if the test statistics are different from 0.

3.4 Data

The present exercise exploits different data sources. Data on funds are retrieved from European Commission-DG REGIO and consist in the package "Integrated database of allocations and expenditure for 2000-2006/2007–2013"(IDE) (Ciffolilli et al. 2015). The package includes consolidated data at the NUTS2 level for ERDF and CF allocations from 2000 and 2014. In the data, funds allocations and expenditures are disaggregated according to a classification of funds priorities. The exercise focuses only on the programming period 2007-2013[44]. The dataset presents a classification of financings that includes 13 categories. These can be aggregated in three macro-categories (Ciffolilli et al. 2015)[45]: Investment in infrastructure, Productive investments, Environment. I will consider only Investments in infrastructure and Productive investments. Investment in infrastructure includes priorities about transport, IT, energy, and social infrastructure, while productive investments include priorities about business support, RTD, human resources and technical support (Ciffolilli et al. 2015).

Monies are generally not paid in advance (Stephenson 2016), and the reception of funds is faster in well-managed regions (Crescenzi and Giua 2016). According to Giua (2016), this could entail a mechanical correlation between region institutional quality and recorded expenses. Therefore, in the present study, I prefer to consider the allocation of funds rather than the funds that have been actually paid off. However, I am aware that there may be a mismatch between money actually spent and the money allocated (Crescenzi and Giua 2016).

Data for the outcome –objective variable– (per capita GDP growth rate) and covariates are retrieved from the regional databases of Cambridge Econometrics and Eurostat. To link the datasets, I needed to make the NUTS classification uniform. While IDE classification of regions mixes NUTS 2003 and NUTS 2006 codes, Cambridge Econometrics and Eurostat use the NUTS 2016 classification. Since most of the data comes from the Cambridge database, to minimize the error due to the absence of a

[44] Although the package would have been well suited for a comparison between programming period 2000-2006 and 2007-2013, I have been unable to find a proper common support for the 2000-2006 period.

[45] Ciffolilli et al.'s (2015) aggregation did not originally includes the technical support priority in "productive investments". Since treating this priority as an autonomous dimension would have required a larger sample, I added investments in technical support to the productive investments.

complete one-to-one match between classifications, I decided to convert all data to the NUTS 2016 classification[46].

The final dataset includes 240 European regions. I correct the outcome variable and financings for both inflation (ref.: the year 2015) and differences in purchasing power parities (ref.: EU 28)[47]. I retrieve multipliers from Eurostat and normalize funds by region population[48]. I decline the outcome variable in one window: it consists of the average per capita growth rate during 2007-2016 (Window 1). Similarly to Crescenzi and Giua (2019), I define the range of the outcome in order to account for the programming period (2007-2013), the fact that expenditures are allowed two years after the programming period, and that regions concentrate investments in the last years available frequently (2014-2015) (Stephenson 2016; Crescenzi and Giua 2019). Finally, I added one further year to account that short-run effects need some time to emerge.

As a robustness check and to account for the medium-run, I replicate the analysis using the period 2007-2019 (Window 2), finding similar results (Appendix 3.2).

Table 3-1 shows descriptive statistics for both the dimensions of the treatment and outcome variables. The difference between the mean and the median suggests a distribution of per capita financings skewed toward zero, with the bulk of the regions receiving less than 700 euros per capita from the regional policy. On average, regions allocate more than 45% of the total to productive investments. Investments in infrastructure are instead roughly 30% of the total on average. The average growth rate is 0.015 for the period 2007-2016.

[46] The rule was the following: if a region has been divided in the new classification, founds have been divided too; if instead, in the new classification, two regions have been merged, then funds of the old regions have been summed.

[47] Although improper, in the rest of the chapter when I talk about outcomes and treatment I will simply talk of per capita outcome and treatment.

[48] I tried also a different normalization, namely for the region GDP, but both the assumption of multivariate normality and the balancing property were poorly satisfied.

Tabella 3-1 Outcome and Treatment descriptive statistics

	mean	p50	p75	sd	max	min
Total funds	633.021	113.041	669.881	1251.498	10293.781	1.431
Infrastructure	267.062	31.862	266.319	593.487	4934.567	0.003
Productive investment	207.497	56.425	180.349	666.517	9540.332	0.913
Share of Infrastructure allocation	0.298	0.282	0.447	0.187	0.751	0.000
Share of Productive investment allocation	0.461	0.475	0.613	0.223	0.953	0.049
Window1	0.015	0.016	0.026	0.017	0.065	-0.032
Observations	240					

Total funds, Infrastructure, Productive investment are per capita and corrected for inflation and purchasing power parity differences.

Central to the empirical strategy adopted is the estimation of the propensity score. Particularly important at this stage is to control for each potential confounder (Austin 2011). Similarly to Becker et al. (2012), I include the average per capita GDP in the five years before the programming period. Controlling for GDP is necessary because it is a variable that the EU considers when allocates funds. Moreover, the effect of financing may depend on the development of a country. Other variables (see Table 3-2) have been chosen because they are deemed to proxy for the region's economic structure. The region's economic structure is considered important in both eligibility for regional transfers and the effectiveness of financings (Becker et al. 2012; Hassink, Isaksen, and Trippl 2019). To mirror the industrial structure, I control for the shares of employed in agriculture, industry (no construction), and in financial and business services. In order to mirror the labor market, I consider the employment rate and the total employment in the region. The overall presence of factors is then proxied by gross fixed capital formation and compensation of employees. All the variables, but employment rate, are calculated as the average in the five years preceding the programming period. I measure the employment rate in the year 2006. Finally, I add to the covariates the European funds (normalized for population) allocated neither in infrastructure nor in productive investment to avoid the mechanical correlation with the treatment due to the total amount of financing received by the region.

I also tried to control for more direct measures of infrastructures (namely Kilometres of motorways per thousand square kilometers and percentage of households with broadband access) and R&D investments (Gross domestic expenditure on R&D as a percentage of GDP). Unfortunately, data provided by Eurostat includes many missings, and the estimation would not have been feasible. Nevertheless, I will provide some statistics in the results section that can be helpful for the interpretation of the results.

3.5 Results

In the following, I estimate the causal effect of different allocations mixes. As previously stated, the outcomes considered are the average growth rate for the periods 2007-2016 (Window 1). The treatment corresponds to the regional mix of allocations in Investment in infrastructure and Productive investments.

The first step consists of the estimation of the propensity score.

To specify the functional form, I run several models, and I selected the one which performs better when testing the balancing property. I identified one outlier – NUTS region ES63 – by performing the BACON procedure[49] with threshold parameter 0.05 (Weber 2010; Billor, Hadi, and Velleman 2000). One outlier is identified by the inspection of the distribution of financings – NUTS region SK01. I removed both regions, thus remaining with a final sample of 238 observations. All variables but those related to employment are corrected for inflation and purchasing power parity differences. All the variables but employment rate (which corresponds to the year 2006) and per capita remaining funds consist in the average for the variable calculated in the years 2002-2006.

[49] The Bacon algorithm consists in the following: 1) selecting a subset of observations; 2) by using this subset, estimating the mean and the covariance matrix of the variables in the model; 3) using the quantities estimated in the subset calculating the Mahalanobis distance in the entire sample; 4) adding all the observations within a certain distance to the initial subset. The procedure is iterated until the starting subset does not grow anymore. The observations that are not included are the outliers.

Table 3-1: Set of variables for the estimation of the propensity score

	mean	p50	p75	sd	max	min
Per capita GDP	19594.427	21392.556	27250.245	11416.412	53066.062	30.080
Per capita GDP * Share Services	2753.019	2500.677	3694.748	2384.543	13838.586	1.537
(Per capita GDP)^2	5.137e+08	4.576e+08	7.426e+08	4.608e+08	2.816e+09	904.785
(Per capita GDP * Share Agriculture)^2	1.614e+06	354129.952	1.348e+06	3.610e+06	3.023e+07	8.079
(Per capita GDP * Share Industry)^2	1.532e+07	1.049e+07	2.230e+07	1.707e+07	1.008e+08	50.428
(Per capita GDP * Share Services)^2	1.324e+07	6.253e+06	1.365e+07	2.381e+07	1.915e+08	2.362
(Per capita GDP)^3	1.493e+13	9.790e+12	2.024e+13	2.023e+13	1.494e+14	27215.619
(Per capita GDP * Share Agriculture)^3	4.454e+09	2.107e+08	1.564e+09	1.589e+10	1.662e+11	22.963
(Per capita GDP * Share Industry)^3	8.508e+10	3.397e+10	1.053e+11	1.329e+11	1.012e+12	358.099
(Per capita GDP * Share Services)^3	8.753e+10	1.564e+10	5.044e+10	2.730e+11	2.650e+12	3.630
Share of employed in Agriculture	0.074	0.038	0.092	0.090	0.542	0.000
Total employment	801814.839	610455.300	1.023e+06	712651.275	5.887e+06	25638.800
Employment rate	63.956	65.200	69.600	7.441	79.200	41.700
Gross fixed capital formation	8.130e+09	5.858e+09	1.057e+10	1.053e+10	9.615e+10	8.805e+06
Compensation of employee	1.779e+10	1.224e+10	2.318e+10	2.392e+10	2.446e+11	1.504e+07
Per capita remaining funds	155.875	27.112	131.725	294.313	2055.546	0.007
Share of employed in Industry (excluding construction)	0.182	0.173	0.235	0.073	0.376	0.029
Observations	238					

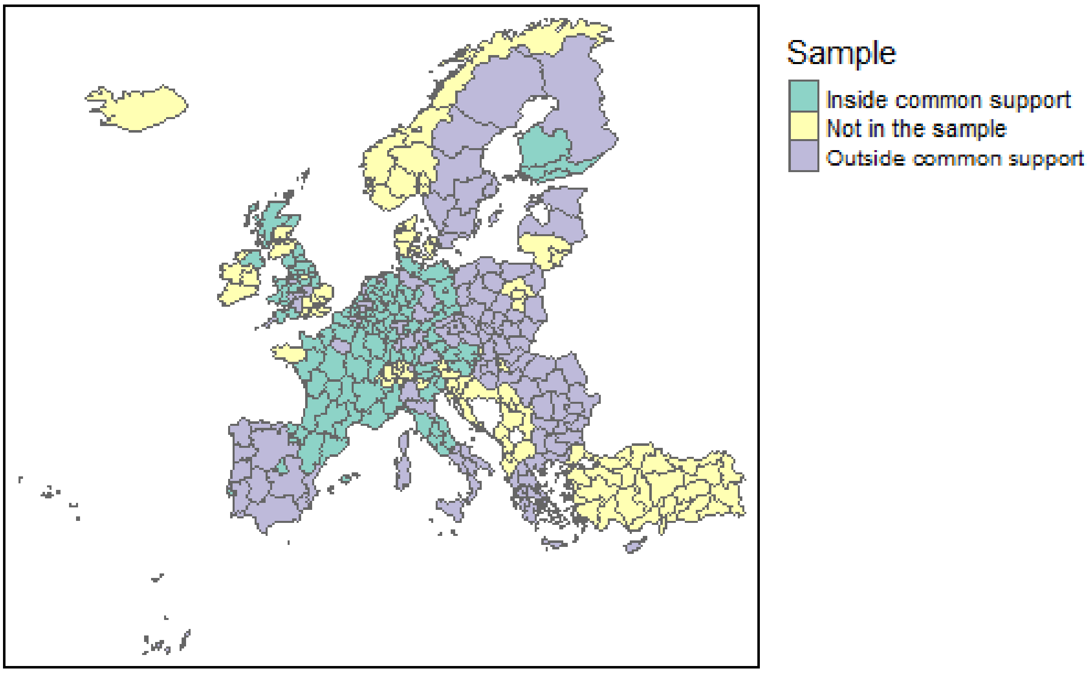
Table 3-2: First stage for allocation in infrastructure and productive investments

	(1)	(2)
VARIABLES	Infrastructure	Productive investments
Per capita GDP	.000635239***	.000577401***
	(.000139653)	(8.21e-05)
Per capita GDP * Share Services	-.0014047**	-.0010723***
	(.000654234)	(.000384414)
(Per capita GDP)^2	-2.33e-08***	-1.94e-08***
	(5.71e-09)	(3.35e-09)
(Per capita GDP * Share Agriculture)^2	-1.46e-07	-1.13e-07
	(1.40e-07)	(8.22e-08)
(Per capita GDP * Share Industry)^2	-1.11e-07**	-6.61e-08**
	(4.92e-08)	(2.89e-08)
(Per capita GDP * Share Services)^2	2.11e-07**	1.39e-07**
	(1.03e-07)	(6.08e-08)
(Per capita GDP)^3	3.22986e-13***	2.30091e-13***
	(8.18560e-14)	(4.80969e-14)
(Per capita GDP * Share Agriculture)^3	2.81780e-11	1.66460e-11
	(2.70626e-11)	(1.59015e-11)
(Per capita GDP * Share Industry)^3	5.44817e-12	3.76096e-12
	(4.88679e-12)	(2.87138e-12)
(Per capita GDP * Share Services)^3	-1.33090e-11***	-7.34561e-12**
	(5.07035e-12)	(2.97923e-12)
Share of employed in Agriculture	4.22**	2.79***
	(1.72)	(1.01)
Total employment	8.09e-07*	7.20e-07***
	(4.58e-07)	(2.69e-07)
Employment rate	-.09466***	-.02523***
	(.0165)	(.0096944)
Gross fixed capital formation	4.44705e-11	-3.38150e-11
	(4.33402e-11)	(2.54658e-11)
Compensation of employee	-4.14641e-11**	-8.55092e-12
	(1.84026e-11)	(1.08130e-11)
Per capita remaining funds	.002517***	.001528***
	(.000513225)	(.00030156)
Share of employed in Industry (excluding construction)	9.09***	3.9**
	(2.82)	(1.66)
Constant	5.23***	2.27***
	(1.44)	(.849)
Observations	238	238
R-squared	.679	.675
Adj. R squared	0.654	0.650

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

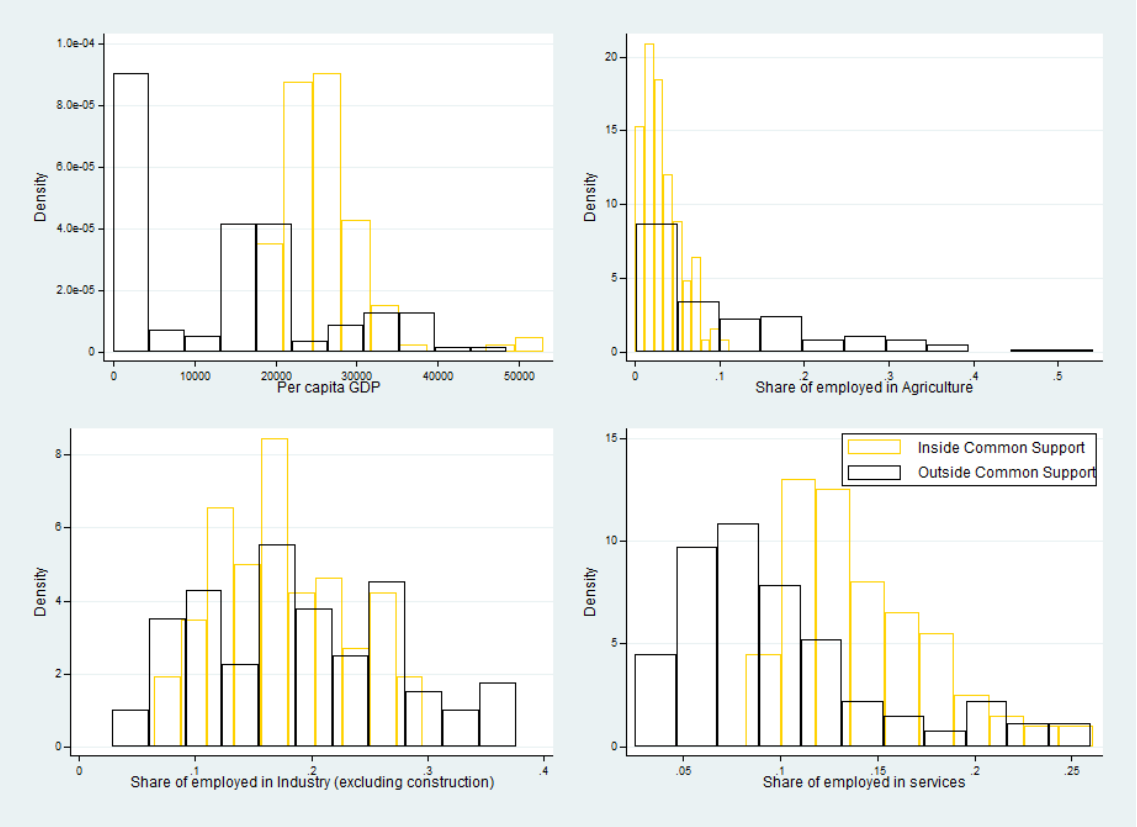
The first stage in the estimation consists in regressing each dimension of the treatment on all the variables deemed important in the reduced equations. Before selecting the model that showed better balancing properties, I made an initial selection of the variables by estimating several forms of the reduced equations and keeping those variables that belonged to the models minimizing the Akaike information criterion (Cavanaugh and Neath 2019). The first stage of the estimation of the propensity score is shown in Table 3-2. For both dimensions, the adjusted R squared is around 65%, thus suggesting a satisfying explanatory power of the chosen models.

Figure 3-1 Observations in the sample before and after common support enforcement. Note: Overseas departments of France have been excluded from the figure. All of them but Mayotte are not inside the common support. Mayotte is not in the sample.



The second step consists of evaluating observations common support. Reducing the sample to the regions that lie on the common support region leaves us with 112 observations (47.06 % of the total)(Figure 3-1), accounting for 14% of the total allocation in the sample (roughly 25 of the 182 billion inflation corrected PPP allocated). Figures 3-2 and 3-3 show the distribution of the different variables for the regions inside common support (yellow) and regions outside common support (white). The common support includes regions with a higher per capita GDP, lower occupation in Agriculture, and higher occupation in the service sector. While the total employment is similar for the two groups, the employment rate tends to be higher for the regions on the common support. Factors are comparable.

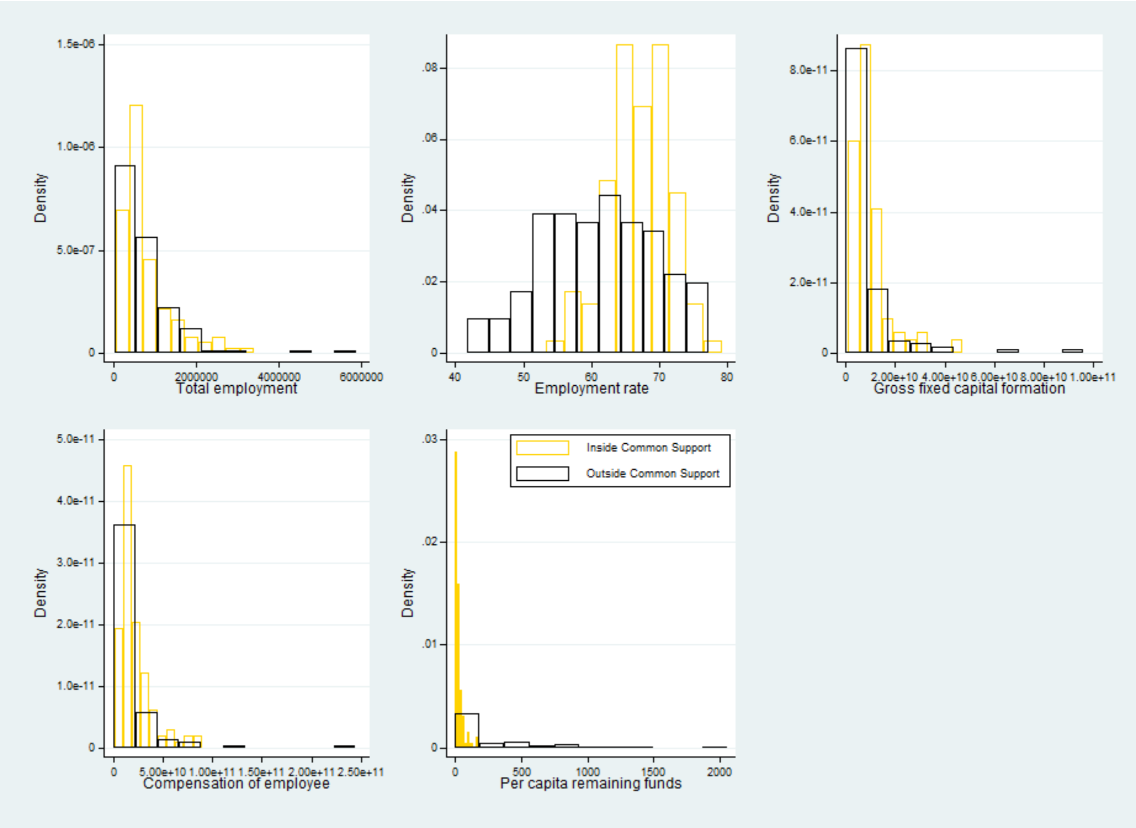
Figure 3-2: Histogram of per capita GDP, shares of employed in Agriculture, Industry, Services for regions inside common support and outside common support.



Regions on common support and outside common support also differ for both the Gross domestic expenditure on R&D as a percentage of GDP and Kilometres of motorways per thousand square kilometers. In particular, the difference in mean between regions on the common support and outside common support is 19.2 Kilometres of motorways per thousand square kilometers (p-value t-test <0.05) and 130.6 euros (p-value t-test <0.05) of Gross domestic expenditure on R&D as a percentage of GDP[50]. Although these statistics are flawed by a significant number of missing values (when considering only observations with non-missing values in both variables, the sample reduces from 238 to 160 observations)[51], they suggest that the infrastructural endowment of regions on the common support is significantly higher than those outside the common support. Moreover, the regions inside the common support are also those that spend more on R&D.

Overall, the characteristics of common support regions cast some doubts about generalizing the results of the present paper to the most lagging areas.

Figure 3-3: Histogram of Total employment, Employment rate, Gross fixed capital formation, Compensation of employee and remaining funds for regions inside common support and outside common support.



I test the balancing property by splitting the two dimensions at the median, thus generating four groups. GPS score is evaluated at the median of the four subsets and,

[50] Differences of medians are even larger.

[51] It is worth noting that missing values for data on Gross domestic expenditure on R&D as a percentage of GDP are not independentely distributed in respect to being inside common support (p-value Chi2-test <0.01).

always at the median, discretized. Table 3-3 and 3-4 shows the balancing between groups for each covariate before and after adjusting for the propensity score. If we consider a significant level of 0.05, we obtain a satisfying reduction of bias by controlling for GPS score: the significant tests are reduced from 18 to 4. The tests have been computed using only the observations on the common support.

Table 3-3: Balancing test for each covariate (p-value)

	1r(p)	2r(p)	3r(p)	4r(p)	1Adj r(p)	2Adj r(p)	3Adj r(p)	4Adj r(p)
Per capita GDP	**.029**	.694	.838	**.013**	.686	.869	.865	.317
Per capita GDP * Share Services	**.002**	.15	.244	.096	.055	.351	.25	.515
(Per capita GDP)^2	**.045**	.621	.79	**.042**	.611	.999	1	.372
(Per capita GDP * Share Agriculture)^2	**.008**	**.048**	.118	.949	**.004**	.209	.535	.417
(Per capita GDP * Share Industry)^2	**.029**	.888	.096	.065	.304	.605	.188	.322
(Per capita GDP * Share Services)^2	**.015**	.255	.319	.215	.119	.422	.346	.521
(Per capita GDP)^3	.066	.555	.743	.104	.524	.851	.878	.43
(Per capita GDP * Share Agriculture)^3	**.004**	**.028**	.186	.734	**.001**	.132	.632	.674
(Per capita GDP * Share Industry)^3	**.029**	.931	.09	.077	.266	.602	.179	.312
(Per capita GDP * Share Services)^3	.057	.344	.441	.359	.2	.459	.479	.568
Share of employed in Agriculture	**.034**	.208	.099	.918	.117	.676	.501	.201
Total employment	**.044**	.369	.157	.559	.056	.475	.114	.358
Employment rate	.189	.775	.509	.098	.94	.98	.663	.288
Gross fixed capital formation	**.025**	.184	.246	.548	.056	.269	.187	.621
Compensation of employee	**.042**	.497	.142	.409	.088	.691	.116	.488
Per capita remaining funds	**0**	.656	.77	**0**	**0**	.806	.798	**0**
Per capita GDP	.287	.727	.104	.58	.408	.814	.114	.711

Table 3-4 Balancing test for each covariate (mean difference and adjusted weighted mean difference)

	1 mean. diff.	2 mean. diff.	3 mean. diff.	4 mean. diff.	1 weighted mean. diff.	2 weighted mean. diff.	3 weighted mean. diff.	4 weighted mean. diff.
Per capita GDP	**-2369.81**	516.57	375.42	**3810.67**	-398.31	-210.15	-347.94	1714.63
Per capita GDP * Share Services	**-1022.42**	575.88	652.00	789.75	-606.19	361.39	721.47	362.84
(Per capita GDP)^2	**-1.45E+08**	4.31E+07	3.25E+07	**2.09E+08**	-3.52E+07	1.58E+05	-2.85E+04	1.06E+08
(Per capita GDP * Share Agriculture)^2	**629653.7**	**-566810.6**	-628217.2	-21977.35	**729788.7**	-357795.4	-273737.7	346566.5
(Per capita GDP * Share Industry)^2	**-6.19E+06**	-4.84E+05	7.96E+06	7.49E+06	-2.91E+06	-1.80E+06	6.99E+06	4.51E+06
(Per capita GDP * Share Services)^2	**-1.12E+07**	6.32E+06	7.74E+06	8.17E+06	-7.23E+06	4.45E+06	8.21E+06	5.12E+06
(Per capita GDP)^3	-7.34E+12	2.84E+12	2.21E+12	9.29E+12	-2.54E+12	9.01E+11	1.16E+12	5.35E+12
(Per capita GDP * Share Agriculture)^3	**1.50E+09**	**-1.38E+09**	-1.17E+09	-2.55E+08	**1.84E+09**	-9.44E+08	-4.67E+08	3.95E+08
(Per capita GDP * Share Industry)^3	**-5.00E+10**	-2.40E+09	6.51E+10	5.78E+10	-2.56E+10	-1.44E+10	5.74E+10	3.72E+10
(Per capita GDP * Share Services)^3	-1.16E+11	6.96E+10	7.94E+10	8.03E+10	-8.12E+10	5.54E+10	8.17E+10	6.15E+10
Share of employed in Agriculture	**0.01**	-0.01	-0.01	0.00	0.01	0.00	-0.01	0.01
Total employment	**-2.36E+05**	1.27E+05	2.79E+05	9.82E+04	-2.37E+05	1.01E+05	3.46E+05	1.92E+05
Employment rate	-1.14	-0.30	0.97	2.05	-0.07	0.03	-0.64	1.47
Gross fixed capital formation	**-3.31E+09**	2.37E+09	2.89E+09	1.28E+09	-2.98E+09	1.95E+09	3.61E+09	1.31E+09
Compensation of employee	**-6.76E+09**	2.73E+09	8.21E+09	3.94E+09	-5.99E+09	1.59E+09	9.81E+09	4.14E+09
Per capita remaining funds	**24.37**	-3.10	2.85	**-47.16**	**21.06**	-1.77	2.78	**-30.00**
Share of employed in Industry (excluding construction)	-0.01	0.00	0.03	0.01	-0.01	0.00	0.03	0.01

Table 3-5 reports the results for the estimation of the flexible control function. In order not to saturate the model, I decided to estimate a polynomial with the dimensions of the treatment, the propensity score, and interactions up to degree two[52]. The parameters have been estimated by OLS. According to Hirano and Imbens (2004), the model does not have a causal interpretation; the only valuable information is whether the GPS terms are jointly significant. If it is the case, the observable covariates matter for selection into treatment intensities. I conclude that it is the case since we observe GPS interaction terms with Productive investments significant at the 10% level.

Table 3-5: Regression with a flexible polynomial for Window 1.

VARIABLES	(1) Window1
Infrastructure	-.000134536**
	(6.40e-05)
Productive investments	-6.79e-05
	(5.27e-05)
GPS	-1.52
	(4.02)
Infrastructure * GPS	-2.37
	(2.08)
Productive investments * GPS	.47*
	(.257)
(Infrastructure)^2	3.56e-07**
	(1.66e-07)
(Productive investments)^2	3.21e-07**
	(1.29e-07)
(GPS)^2	549
	(412)
(Infrastructure * GPS)^2	145
	(290)
(Productive investments * GPS)^2	-5.32*
	(2.86)
Constant	.01672***
	(.0046929)
Observations	112
R-squared	.278

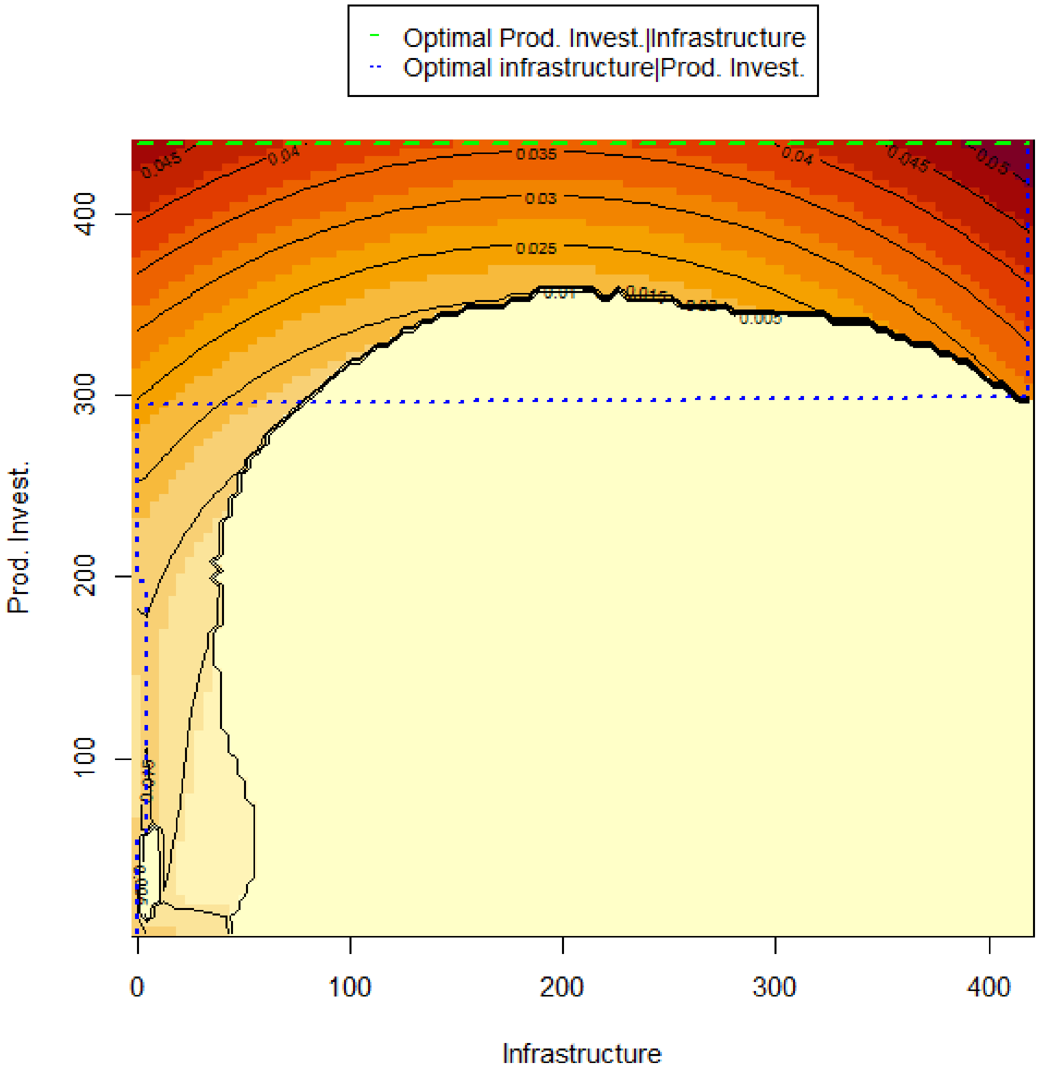
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 3-4 shows the estimated dose-response function. In particular, the heatmap is built as follows. For each dimension, I selected 100 equidistant points

[52] As robustness check, in Appendix 3.3, I present results where I have added terms up to power three. Results are consistent.

within its range. The Cartesian product of these two sets of points defines the set of treatments for which I estimate the dose. Bias corrected method (BC) confidence intervals (Carpenter and Bithell 2000; Efron and Tibshirani 1994) are then computed for each estimated response using 1000 bootstrap samples. The figure reports the point estimates for the treatments, which are significantly different from 0. Figures 3-5 to 3-8 report for each dimension univariate dose-response function keeping the other dimension fixed at points representing its range.

Figure 3-4: Window 1 heatmap and optimal allocation by dimension



According to my estimate, the mix between productive investments and infrastructure investments is crucial for determining growth. Regions obtain the maximum gain when per capita investments correspond roughly to 400€ in both dimensions. Allocation in infrastructures has a limited effect on growth (when not undermining) unless the allocation is high in magnitude and combines with medium or high magnitude productive investments. Productive investments have an appreciable impact on growth, even if when small in volume. However, they reach maximal efficacy when coupled with high investment in infrastructure.

Figure 3-5 Window1. Dose-response function keeping Prod. Inv. fixed

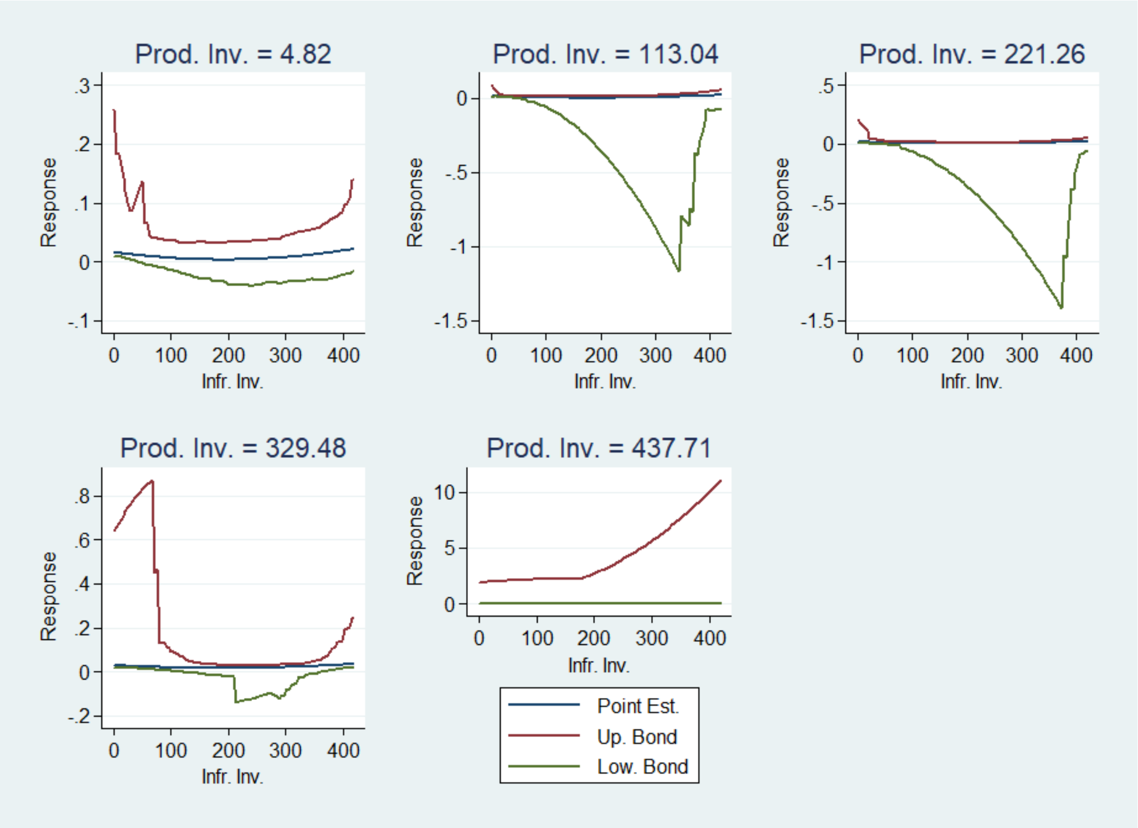


Figure 3-6 Window1. Dose-response function keeping Infr. Inv. fixed

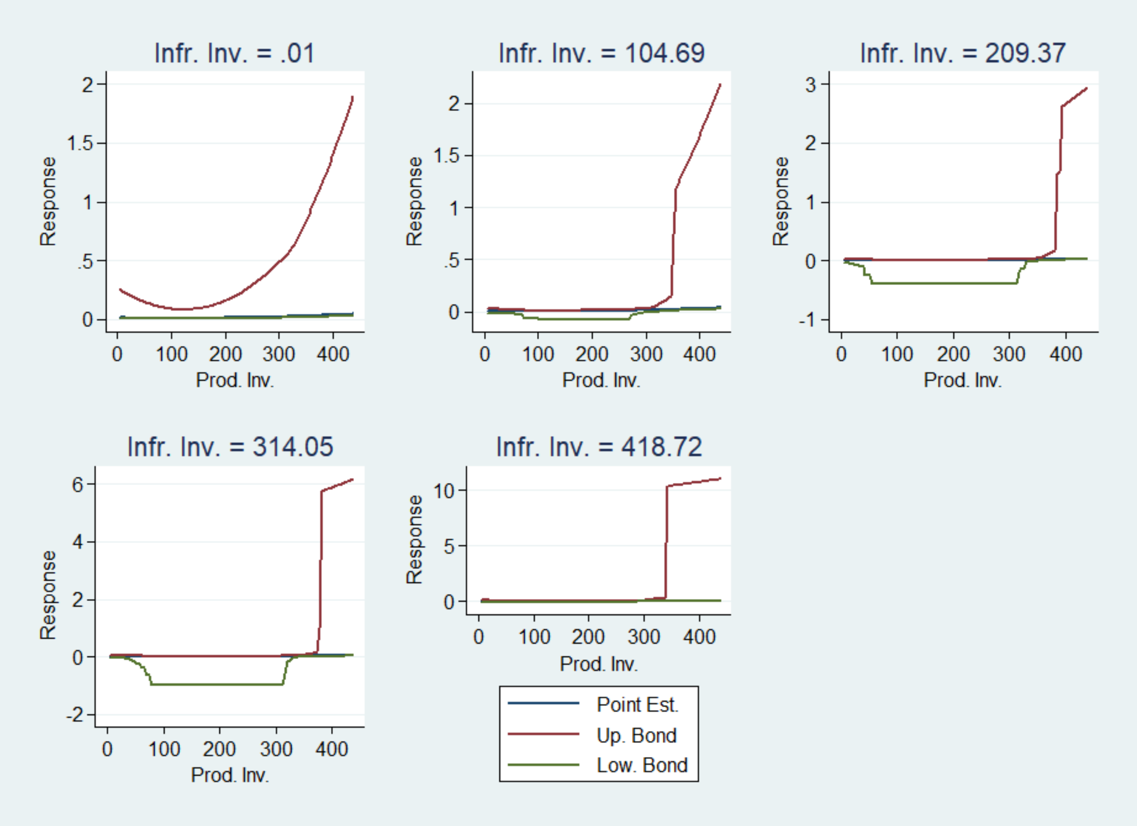


Figure 3-7 Window1. Dose-response function keeping Prod. Inv. fixed

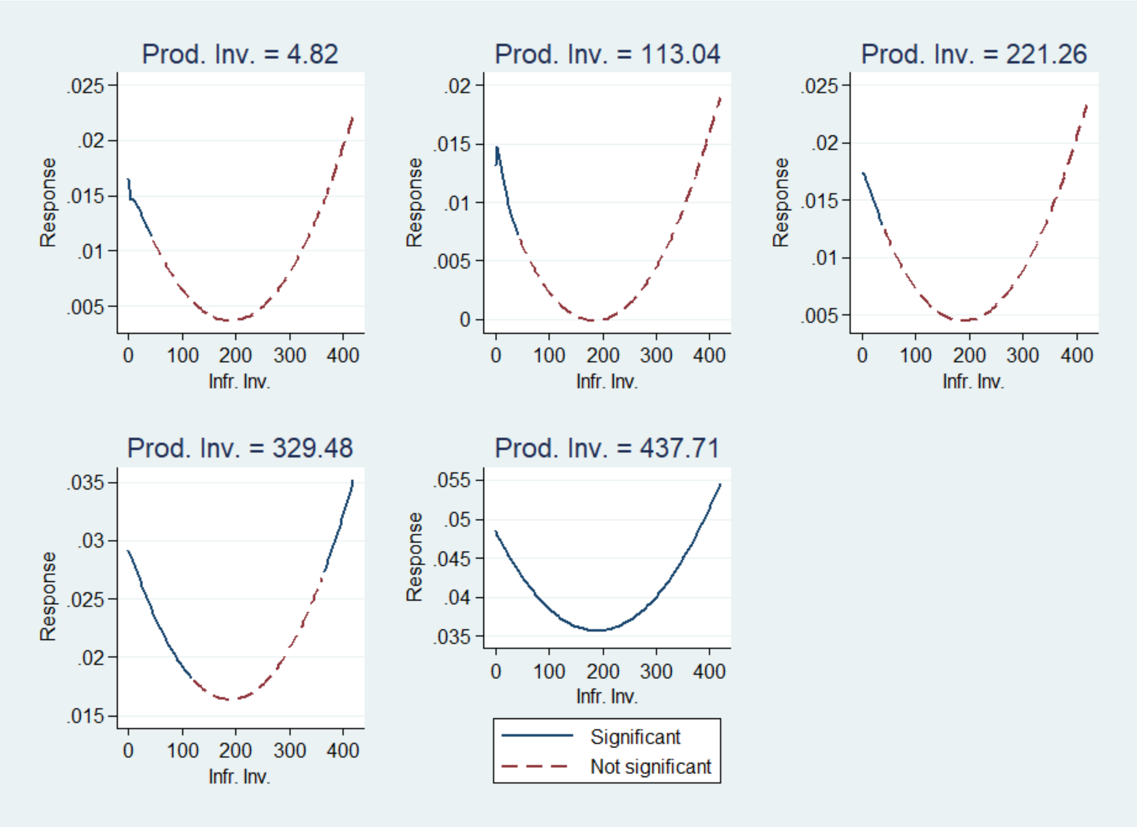
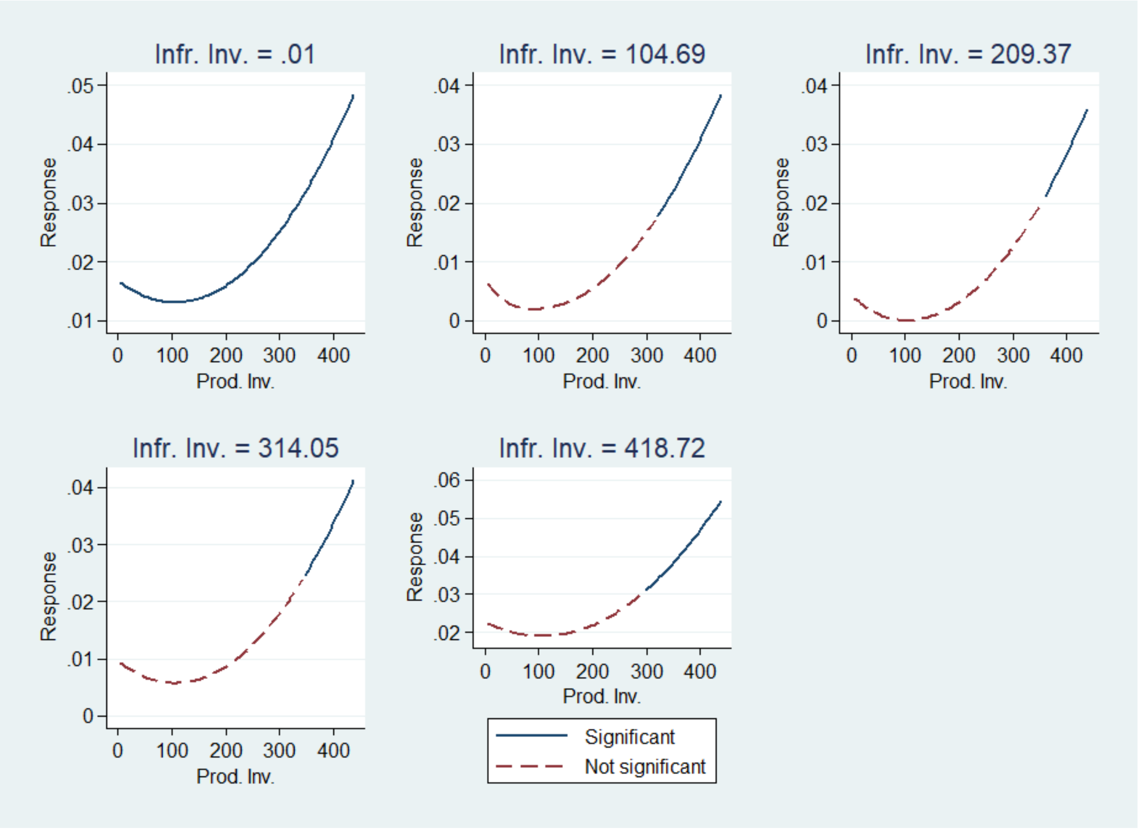


Figure 3-8 Window1. Dose-response function keeping Infr. Inv. fixed



My interpretation of the results[53] follows a four-fold typology of investment mixes (Table 3-6). When allocations on Infrastructure are low, we observe a detrimental effect on growth because of the crowding-out effect (Rietveld and Bruinsma 2012, chap. 3). Moreover, the impact on growth can be irrelevant because the overall infrastructural network is unlikely affected by small projects (Roller and Waverman 2001; Vickerman, Spiekermann, and Wegener 1999; Rietveld and Bruinsma 2012). The positive effects (Berkowitz, Monfort, and Pieńkowski 2020) of high magnitude productive investments can overcome these adverse mechanisms.

When the allocations in infrastructure have a high magnitude but productive investments are scarce, the negative impact of decreased transportation costs and the lack of "absorptive capacity" to manage the change dominate (Puga 2002; Tranos 2012). When both productive investments and investments in infrastructure have a

[53] The interpretation of results is based on the literature reviewed above. It is, however, worth noticing that because of data limitations I am not able to actually test which of the mechanisms are responsible for the observed results. Further research are needed to address the issue.

high intensity, the positive effects of both kinds of investments magnify. In particular, productive investments help firms exploit the new possibilities enabled by the infrastructure change by decreasing the cost of capital and favoring contexts more permeable to innovation (Berkowitz, Monfort, and Pieńkowski 2020; Andrěs Rodríguez-Pose 1999; Garone et al. 2015). Moreover, by increasing firms' competitiveness, productive investment can mitigate the shortcomings of reducing transportation costs (Berkowitz, Monfort, and Pieńkowski 2020).

It is important to stress that these results cannot be easily generalized to all the regions irrespective of their characteristics. Results, in particular the weak impact of infrastructure alone, may be driven by the fact that the common support includes primarily highly developed regions already endowed with infrastructures. The present analysis cannot exclude that different types of regions (less developed and lacking infrastructures) would benefit more, and irrespective of the intensity, from investments in infrastructure.

The large infrastructural endowments of common support regions could also explain why the gains obtained coupling high-intensity investments in infrastructure and productive investments are comparable to those obtained with high-intensity investments in productive investments alone (mean growth rate of 5.4% against 4.8%)[54]. When a region is already provided with sufficient infrastructures, the bulk of the gains depends on using those infrastructures efficiently and developing an ecosystem capable of extracting value from the existing asset rather than constructing new assets per se.

Table 3-6: Typology of allocations.

			Infrastructure	
			High	Low
Productive investments		High	Positive	Positive
		Low	Not appreciable	Not appreciable

[54] I thank Annalisa Caloffi for stressing this point.

3.6 Two hypothetical scenarios: investigating the reallocation of funds

The final part of the chapter is devoted to measuring the extent of misallocation in the deployment of regional funds. I will focus on two possible sources of misallocation. The first one pertains to the degree of efficiency and efficacy the regions have in managing regional funds. It is possible indeed that, with the resources received, regions could have chosen a better allocation mix in respect to the actual. The better mix could be more effective in fostering growth, but it could also simply require a smaller amount of financings for obtaining the same result. The second source of misallocation involves the allocation of funds as managed by the European Union. It is indeed possible that the amount of money received by the regions in the sample is not enough to trigger growth maximally. Therefore, in the rest of the section, by leveraging the counterfactual analysis above, I will try to answer two questions: 1) Do the regions choose the right expenditure mix? 2) Can they select the mix that has the best effect on growth?

Both are important questions, and the answers can require different policy responses. Suppose regions could have allocated the actual amount of financings better. In that case, it can be argued that stricter controls on the policy implementation could help the policymaker choose the more worthy investments. If, instead, regions do not receive enough money to take advantage of regional policy maximally, then, at the EU level, it should be examined whether the distribution of financings between regions is consistent with the policy's primary goals.

To answer these questions, I generate two hypothetical scenarios. In the first one, I investigate what would happen if each region spent the money received in the most efficient and effective way. In particular, I look to a scenario where each region chooses the best allocation mix available given the amount of financing it actually received (it is worth noting that I do not constrain regions to spend all the money they receive). In the second scenario, I allow regions to choose the optimal allocation mix without any constraints.

In order to generate these two scenarios, I will leverage the dose-response function estimated above, which enables me to calculate the growth a region would experience if endowed with a different amount of financing or if it used the received funds differently.

However, since the econometric model adopted does not generate a dose-response function for each dose but only for a finite set, we cannot directly compare the optimal allocations with the real ones. Indeed, the method approximates a continuous dose-response function by means of a discrete set of points extracted from the range of the treatment. In the present chapter, I have estimated the response corresponding to 10000 couples (see above). In order to represent the range of the treatment uniformly, these couples are results of the Cartesian product between 100 equidistant points chosen from the investment in infrastructure dimension and 100 equidistant points chosen from the productive investment dimension. Consequently, it is not guaranteed that the chosen couples correspond to the actual allocation a region experienced.

Therefore, for all the regions, when needed, I approximate the actual response with the closest dose (in terms of Euclidean distance) for which I have estimated the response. I called this quantity the estimated actual allocation. The figures from 3-9 to 3-15 show the comparisons between the optimal allocations and the estimated actual allocations.

Before commenting on the figures, a last clarification is in order. According to the previous estimation, the maximum growth (mean growth rate of 5.4%) is obtained coupling 418 PPP per capita investments in infrastructure with 437 PPP per capita in productive investment. In the second scenario, each region will therefore choose this mix. However, it is arguable that this allocation mix is genuinely the best. Indeed, as stressed above, regions could obtain comparable results (mean growth rate of 4.8%) by spending roughly half of the money (keeping 437 PPP per capita of productive investment while avoiding investing in infrastructure almost completely). Therefore, in appendix 3.4, I report a third scenario where the optimal allocation mix is considered the couple (0.045 PPP, 437 PPP). However, the results are qualitatively consistent.

Figure 3-9: Difference from optimal allocation (regional expenditures constraints). Per capita difference disaggregated by dimension.

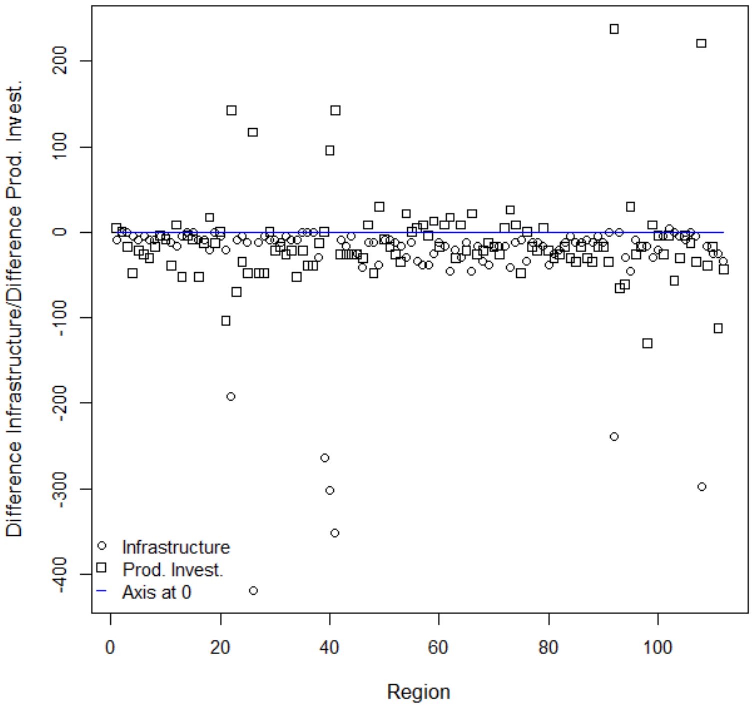


Figure 3-10: Difference from optimal allocation (regional expenditures constraints). Per capita total difference.

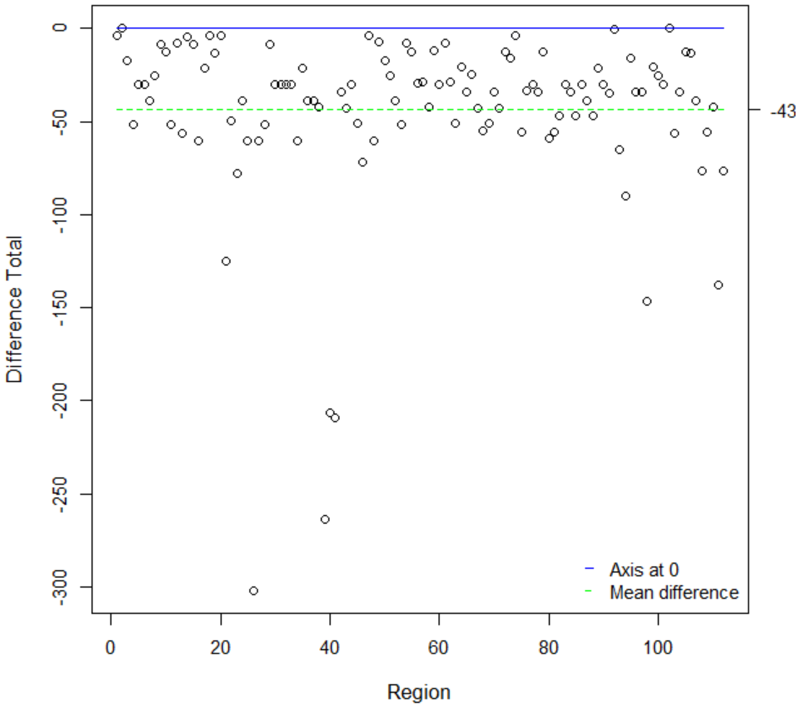


Figure 3-9 reports the per capita differences between the optimal constrained allocation and the estimated actual allocation for productive investments and infrastructure investments. Overall, the constrained optimal allocation has a lower amount than the actual. The largest savings would be on the investments in the infrastructure side. By choosing the constrained optimal allocation, every region would have saved a mean of 80 million euros (Figure 3-11), and the mean of the average per capita growth rate during the years 2007-2016 would have doubled (0.008 to 0.019) (Figure 3-12). The results suggest that regions should manage more carefully the funds received to obtain higher growth.

Figure 3-11: Difference from optimal allocation (regional expenditures constraints). Absolute Total difference.

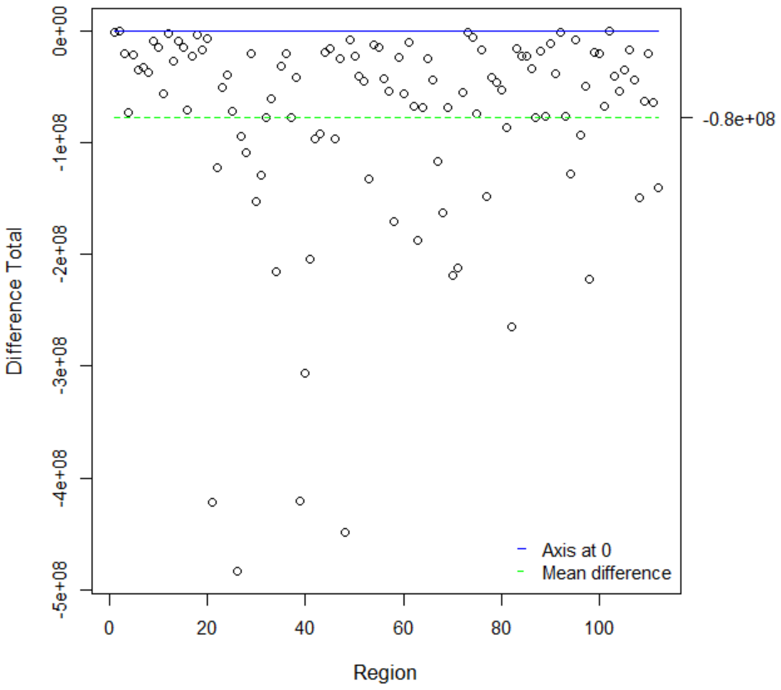
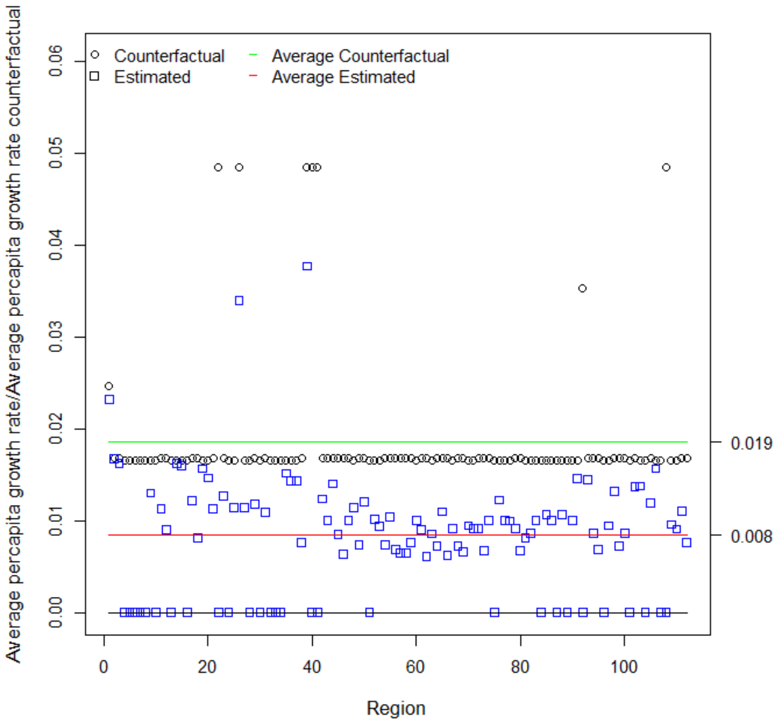


Figure 3-12: GDP gain (regional expenditures constraint)



The unconstrained scenario depicts quite different results. The dose-response function estimated indicates that regions would maximize growth by coupling 418 PPP per capita investments in infrastructure with 437 PPP per capita in productive investment. None of the regions included in the sample receives enough funds to choose this allocation (Figure 3-13). In particular, regions would need a mean of 1.4 billion euros more than the amount received (Figure 3-14). The reallocation exercise shows the growth experienced by the regions if they could choose the allocation mix that guarantees the maximal growth (Figure 3-15). If all regions had chosen the unconstrained optimal allocation, the mean of the average per capita growth rate during the years 2007-2016 would have been 5.4% rather than 0.8%. Considering that the common support mainly includes regions with lower transfer intensity, the results are consistent with Becker et al. (2012), who found that regions with lower transfer intensity would benefit from additional funds, contrarily to the regions with high transfer intensity.

Figure 3-13: Difference from optimal allocation (no constraints). Per capita and disaggregated by dimension.

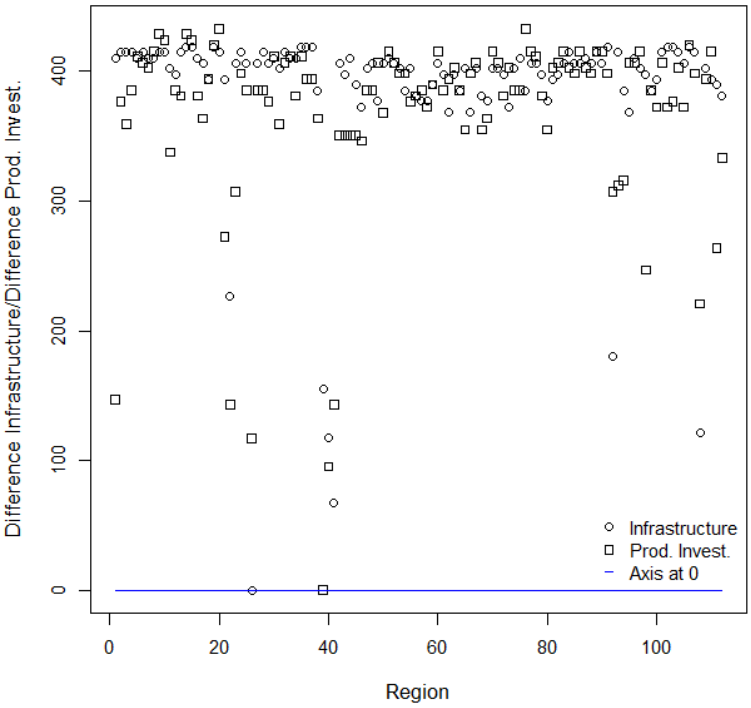


Figure 3-14: Difference from optimal allocation (no constraints). Absolute Total difference.

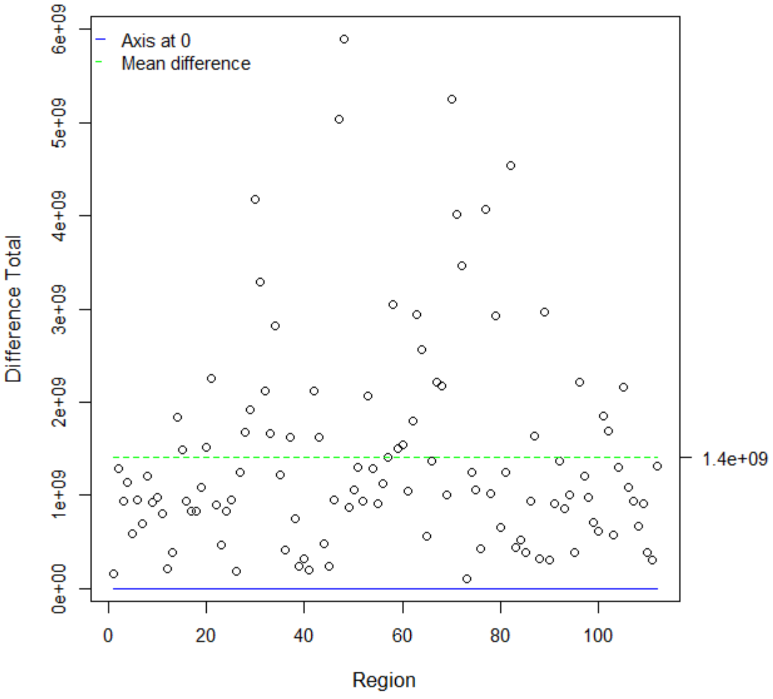
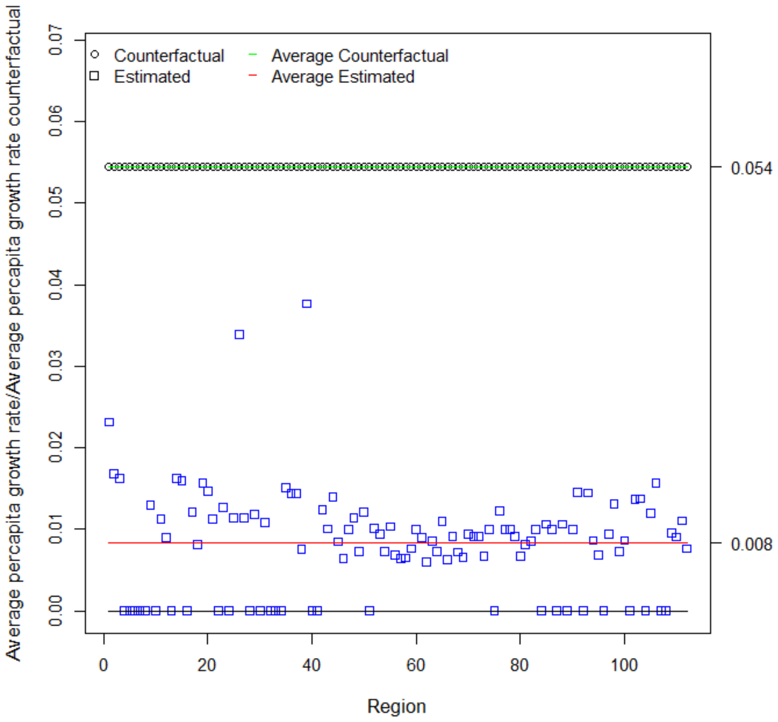


Figure 3-15: GDP gain no constraints



3.7 Conclusions

The chapter presents an evaluation of the European regional policy 2007-2013 employing a novel method, namely the generalized propensity score for multidimensional continuous treatment (Peter H. Egger and von Ehrlich 2013). In particular, I have focused on the heterogeneous effects of policies differently mixing infrastructural investments and productive investments. The optimal allocation is the one combining high funding intensity in both dimensions. In particular, I observe the maximum growth at 418 PPP per capita in Infrastructure and 437 PPP per capita in productive investment. If not mixed with productive investments, investments in infrastructure tend to be detrimental to growth. High funding intensity in productive investments is associated with growth irrespective of investments in infrastructure.

However, these results cannot be easily generalized to all the regions irrespective of their characteristics. The common support enforcement restricted the analysis to the regions comparatively more developed, with lower occupation in agriculture and higher occupation in the service sector. They were also the regions already endowed with a developed system of infrastructures, thus corresponding to regions with low transfer intensity (Becker et al. 2012).

To assess the extent of inefficiencies in the deployment and distribution of funds, I have presented two counterfactual scenarios. In the first one, every region chooses the best allocation possible with the constraint of the actual amount of received funds. In the second, I do not impose any budget constraints. The final results are that although regions could improve their allocations (they could spend the amount received better), when unconstrained, they do not receive enough money to benefit from financings maximally.

The results are consistent with Becker et al. (2012), who found that regions with lower transfer intensity would benefit from additional funds, contrarily to the regions with high transfer intensity.

Extant literature on European regional policy generally focuses on a highly aggregated measure of financings – the overall stream of funds or the Objective 1 status. The present study is one of the first attempts at opening the black box. It investigates the impact of different investment mixes on GDP growth, thus providing policymakers with one more tool to fine-tune their interventions. However, the results can be generalized only to the most developed regions. Moreover, sample size restrictions impede testing the impact of more complex and realistic financing mixes by increasing the number of treatment dimensions. As a consequence, the present study only partially succeeds in filling the gap present in the literature. In the future developments of the study, I will try to improve the methodology to avoid these pitfalls.

3.8 Appendix 3.1

The convergence equation is mainly an equation that is used to estimate β-convergence. β is substantially the coefficient of a first-order Taylor approximation in a ball of the steady-state. Therefore it is a measure of the speed to which a given system (e.g., a country, a region) reaches the steady-state (i.e., the "balance growth path").

The exposition follows de la Fuente (2000). The Solow model assumes a system producing at time t an outcome, Y, by the following production function:

$$Y(t) = K(t)^{\alpha}\big(A(t)L(t)\big)^{1-\alpha}$$

Where K, A, L, are respectively aggregate capital, total factor productivity, and aggregate labor. A and L evolve in time at a constant rate:

$$L(t) = L(0)e^{nt}$$
$$A(t) = A(0)e^{gt}$$

The derivative of K in respect to time is set as:

$$\dot{K} = sY - \delta K$$

Where s is the share of total output invested in the next period, and δ is the rate at which capital depreciates.

I derive now the consequence of this setting.

Define $k = \frac{K}{AL}$, which is the capital per effective unit of labor, and $y = \frac{Y}{AL} = k^{\alpha}$.

If we take the log of k:

$$\frac{\partial \ln k}{\partial t} = \frac{\dot{k}}{k} = \frac{\partial \ln K}{\partial t} - \frac{\partial \ln AL}{\partial t} = K^{-1}\dot{K} - (AL)^{-1}\big(\dot{A}L + A\dot{L}\big)$$
$$= \frac{sY}{K} - (\delta + n + g)$$

This implies

$$\dot{k} = \frac{sY}{AL} - (\delta + n + g)\frac{K}{AL} = sk^{\alpha} - (\delta + n + g)k$$

The steady-state is reached when $\dot{k}(k) = 0$ that is when:

$$k^* = \left(\frac{s}{(\delta + n + g)}\right)^{\frac{1}{1-\alpha}}$$

Once the steady-state is reached, the outcome per capita becomes:

$$\frac{Y(t)}{L(t)} = A(0)e^{gt}k^{*\alpha}$$

And the "balanced growth path" begins. Per capita growth is not determined by capital anymore but only by technological improvements g.

The speed of convergence to the steady-state is determined by the speed with which capital reaches k^*. If we linearize the behavior of $\dot{k}$ with a Taylor first-order approximation, we obtain:

$$\dot{k} \cong \dot{k}(k^*) + \left(\frac{\partial \dot{k}(k)}{\partial k}\bigg|_{k=k^*}\right)(k - k^*) \cong -(1-\alpha)(\delta + n + g)(k - k^*)$$

We set $(1-\alpha)(\delta + n + g) = \beta$. The convergence rate is the same for both the outcome per effective unit of labor and capital (Romer 2017; Islam 2003). It can be shown by also approximating y. Recalling that $y = \frac{Y}{AL}$=k^α. Then, the first-order approximation of y at $y^*(k^*)$ is $y \cong y^*(k^*) + \frac{\partial y(k^*)}{\partial k}(k - k^*)$. If we derive for time, we obtain $\dot{y} \cong \frac{\partial y(k^*)}{\partial k}\dot{k}$. Then, since $\frac{\partial y(k^*)}{\partial k} \cong \frac{y-y(k^*)}{(k-k^*)}$: $\dot{y} \cong \frac{y-y(k^*)}{(k-k^*)}\dot{k} = -\beta\big(y - y(k^*)\big)$.

In order to estimate β, it can be employed the trick in de la Fuente (2000), which exploits the fact that $\ln(k)$ behave similarly to k. Then, recalling that $Q(t) = \frac{Y(t)}{L(t)}$=$A(t)k(t)^\alpha$, after some manipulations, we can derive[55]:

[55] Actually the β of beta convergence is referring to the estimated equation. Nevertheless, since the beta estimated is a function only of $\left(\frac{\partial \dot{k}(k)}{\partial k}\Big|_{k=k^*}\right)$ I refer to $\left(\frac{\partial \dot{k}(k)}{\partial k}\Big|_{k=k^*}\right)$ as beta.

$$\frac{\ln\big(Q(t+h)\big)-\ln\big(Q(t)\big)}{h} = g+\frac{\left(1-e^{-\beta h}\right)}{h}\left(\frac{\alpha}{1-\alpha}\ln\frac{s}{(\delta+n+g)}-\ln Q_t+\ln A_t\right)$$

Having data on factors and output per capita, the equation can be estimated.

3.9 Appendix 3.2

Table 3-7: Regression with a flexible polynomial for Window 2.

VARIABLES	(1) Window2
Infrastructure	-.00014037**
	(6.77e-05)
Productive investments	-5.25e-05
	(5.58e-05)
GPS	-3.61
	(4.25)
Infrastructure * GPS	-.266
	(2.2)
Productive investments * GPS	.547**
	(.272)
(Infrastructure)^2	3.72e-07**
	(1.76e-07)
(Productive investments)^2	2.89e-07**
	(1.36e-07)
(GPS)^2	749*
	(436)
(Infrastructure * GPS)^2	-96
	(307)
(Productive investments * GPS)^2	-5.93*
	(3.03)
Constant	.01515***
	(.0049664)
Observations	112
R-squared	.233

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 3-16: Window 2 heatmap and optimal allocation by dimension

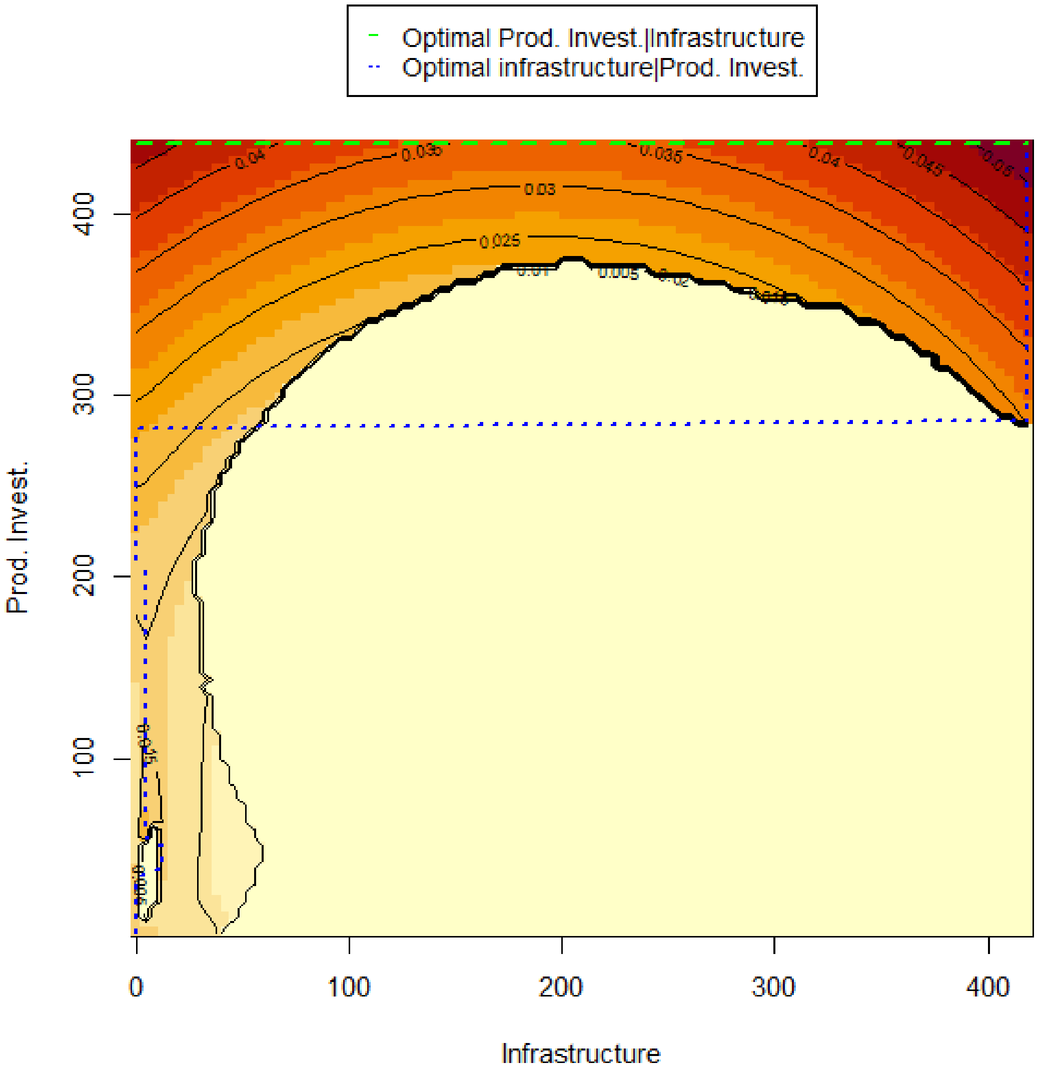


Figure 3-17: Window 2 GDP gain (regional allocation constraints)

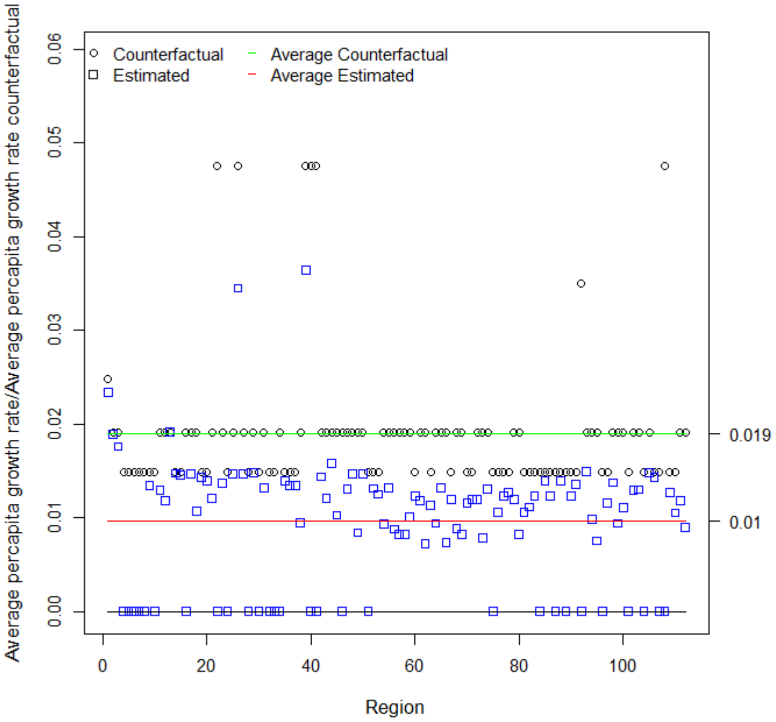
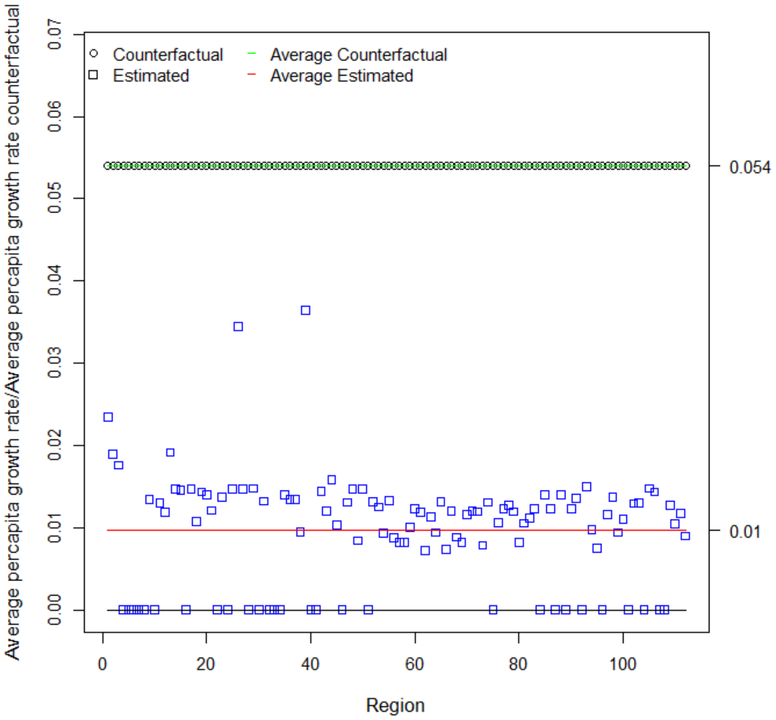


Figure 3-18: Window 2 GDP gain (no constraints)



3.10 Appendix 3.3

To test the robustness of the estimation, I add terms at power three in the control function. That is I estimate:

$$E[Y|X,T] = \sum_{p=1}^{3} \alpha_p GPS^p + \sum_{p=1}^{3} \alpha_{1p} Infrastructure^p$$
$$+ \sum_{p=1}^{3} \alpha_{2p} Productive\_investment^p + \sum_{p=1}^{3} \alpha_{tp} (Infrastructure * GPS)^p$$
$$+ \sum_{p=1}^{3} \alpha_{tp} (Productive\_investment * GPS)^p$$

Although the variance of the estimation is higher than the estimation obtained with a polynomial of degree 2 [56], the results are similar. A high response is obtained only if there are high-intensity investments in productive investments. Interestingly the global maximum is now when there is an investment of 372.77 PPP per capita in Productive investments and 0.015 PPP per capita in investment in infrastructure. It means that the maximum is when expenditures in Infrastructure are low. Nevertheless, the average per capita growth rate during the years 2007-2016 estimated for higher investment in infrastructure – the couple (372.77, 314.05) – is 3.3%, and it is similar to the global maximum of 3.5%.

[56] See the difference between Figure 3-14, the point estimate, and Figure 3-15 where non-significant estimates are set to 0.

Figure 3-19 Heatmap response model power 3.

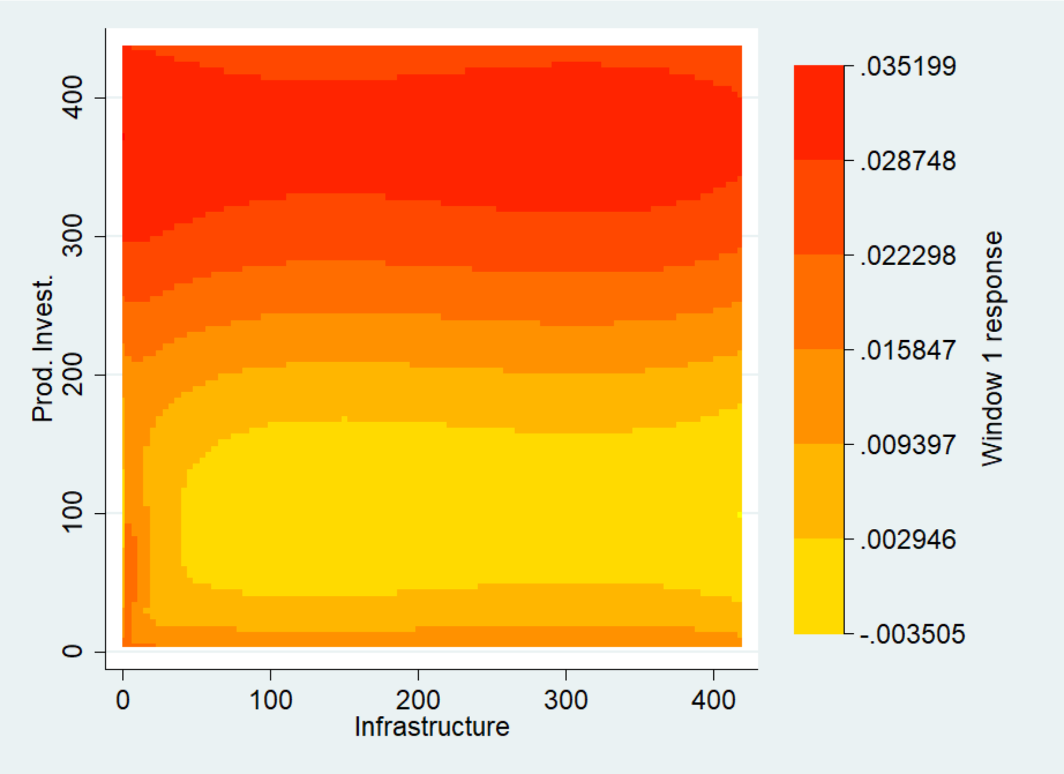
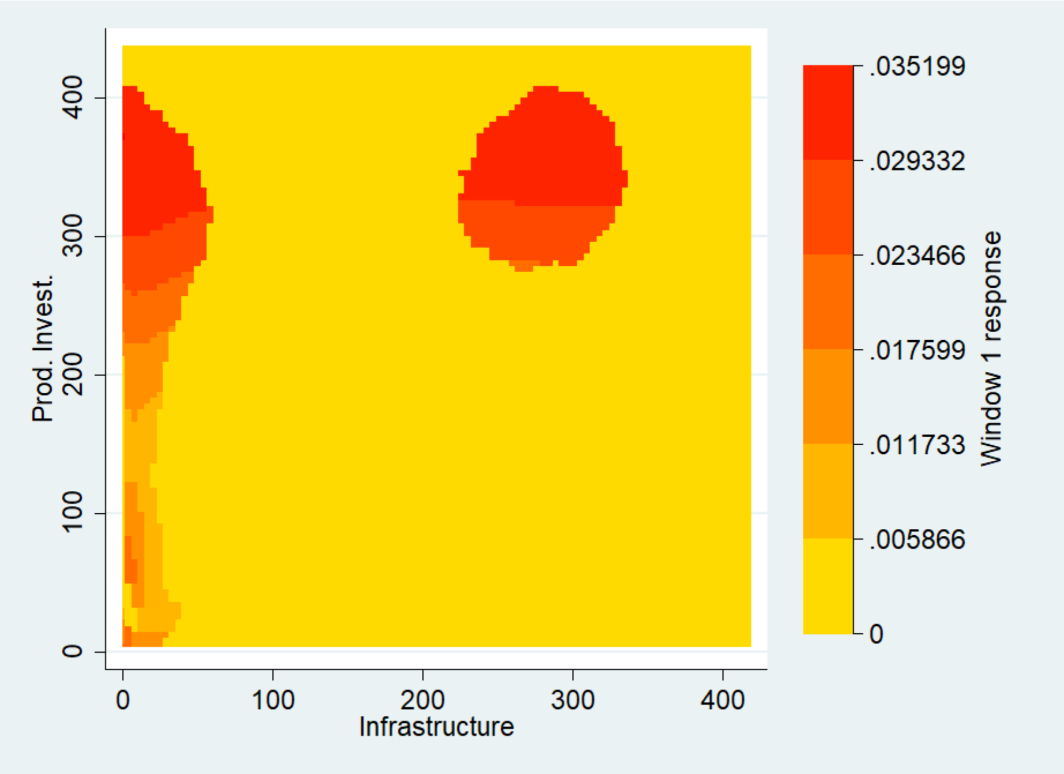


Figure 3-20 Heatmap response model power 3. Not significant responses set to 0



3.11 Appendix 3.4

Choosing the couple 418 PPP per capita investments in infrastructure with 437 PPP per capita in productive investment as the optimal allocation in the second scenario was arguable. Indeed with half of the investment, the policymaker could obtain similar results. Therefore, this appendix will compare the estimated actual allocation with a scenario in which regions choose the allocation mix composed of 0.045 PPP per capita investments in infrastructure and 437 PPP per capita in productive investment.

Consistent with the previous scenario, most regions do not receive enough money to afford the allocation mix. In order to choose the mix, regions would need a mean of 0.6 billion euros more than the amount received.

Figure 3-21 Difference from allocation (0.045, 437). Absolute Total difference.

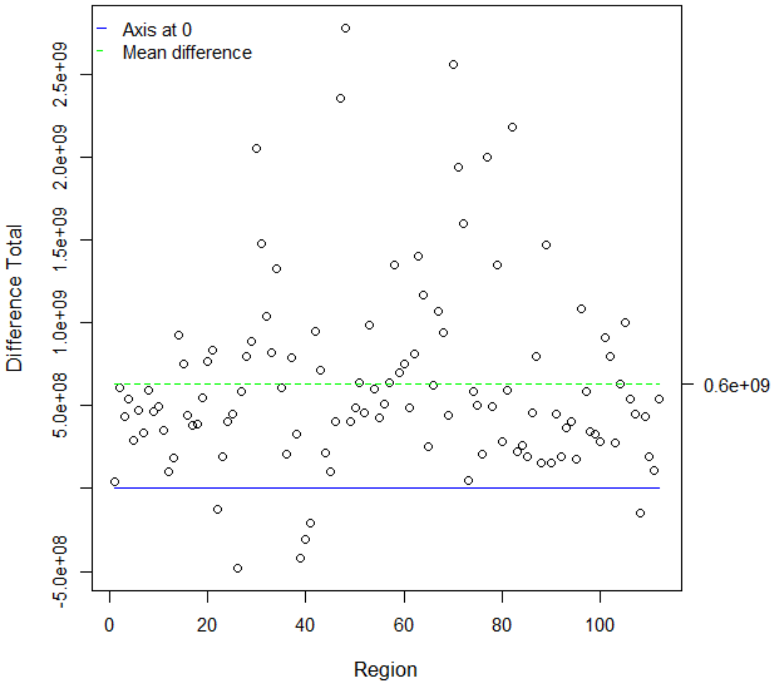
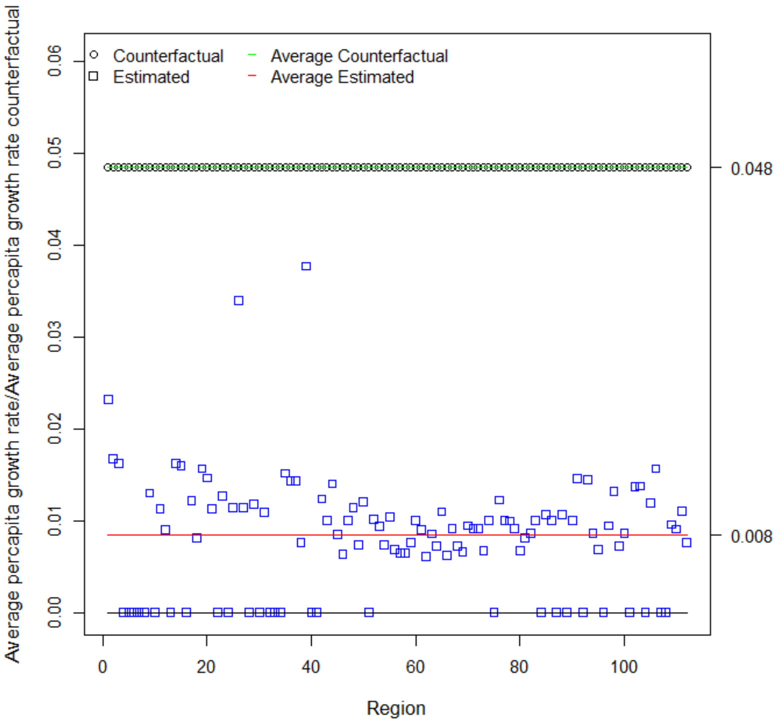


Figure 3-22 GDP gain allocation (0.045, 437)



3.12 References

Accetturo, Antonio, Guido de Blasio, and Nicola Rossi. 2019. *Morire di aiuti: I fallimenti delle politiche per il Sud*. IBL Libri.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones. 2017. 'Artificial Intelligence and Economic Growth'. National Bureau of Economic Research.

Aiello, Francesco, and Valeria Pupo. 2012. 'Structural Funds and the Economic Divide in Italy'. *Journal of Policy Modeling* 34 (3): 403–18.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. 2015. 'The Skill Complementarity of Broadband Internet'. *The Quarterly Journal of Economics* 130 (4). MIT Press: 1781–1824.

Albanese, Giuseppe, Guido de Blasio, and Andrea Locatelli. 2020. 'Does EU Regional Policy Promote Local TFP Growth? Evidence from the Italian Mezzogiorno'. *Papers in Regional Science* n/a (n/a).

Andersson, \AAke E., Christer Anderstig, and Björn Haarsman. 1990. 'Knowledge and Communications Infrastructure and Regional Economic Change'. *Regional Science and Urban Economics* 20 (3). Elsevier: 359–76.

Aschauer, David Alan. 1989. 'Is Public Expenditure Productive?' *Journal of Monetary Economics* 23 (2). Elsevier: 177–200.

Austin, Peter C. 2011. 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies'. *Multivariate Behavioral Research* 46 (3): 399–424.

Ayres, Robert U., Jeroen CJM Van den Bergh, Dietmar Lindenberger, and Benjamin Warr. 2013. 'The Underestimated Contribution of Energy to Economic Growth'. *Structural Change and Economic Dynamics* 27. Elsevier: 79–88.

Banister, David, and Yossi Berechman. 2001. 'Transport Investment and the Promotion of Economic Growth'. *Journal of Transport Geography*, Mobility and Spatial Dynamics, 9 (3): 209–18.

Barca, Fabrizio. 2009. *Agenda for a Reformed Cohesion Policy*. European Communities Brussels.

Barca, Fabrizio, Philip McCann, and Andrés Rodríguez-Pose. 2012. 'The Case for Regional Development Intervention: Place-Based Versus Place-Neutral Approaches*'. *Journal of Regional Science* 52 (1): 134–52.

Barro, Robert J., and Xavier Sala-i-Martin. 1990. 'Economic Growth and Convergence across the United States'. National Bureau of Economic Research.

Becker, Sascha O., Peter H. Egger, and Maximilian von Ehrlich. 2010. 'Going NUTS: The Effect of EU Structural Funds on Regional Performance'. *Journal of Public Economics* 94 (9): 578–90.

———. 2012. 'Too Much of a Good Thing? On the Growth Effects of the EU's Regional Policy'. *European Economic Review* 56 (4): 648–68.

Becker, Sascha O., Peter H. Egger, and Maximilian Von Ehrlich. 2013. 'Absorptive Capacity and the Growth and Investment Effects of Regional Transfers: A Regression Discontinuity Design with Heterogeneous Treatment Effects'. *American Economic Journal: Economic Policy* 5 (4): 29–77.

Beer, Andrew, Sarah Ayres, Terry Clower, Fabian Faller, Alessandro Sancino, and Markku Sotarauta. 2019. 'Place Leadership and Regional Economic Development: A Framework for Cross-Regional Analysis'. *Regional Studies* 53 (2). Routledge: 171–82.

Berkowitz, Peter, Philippe Monfort, and Jerzy Pieńkowski. 2020. 'Unpacking the Growth Impacts of European Union Cohesion Policy: Transmission Channels from Cohesion Policy into Economic Growth'. *Regional Studies* 54 (1). Routledge: 60–71.

Bia, Michela, and Alessandra Mattei. 2008. 'A Stata Package for the Estimation of the Dose-Response Function through Adjustment for the Generalized Propensity Score': *The Stata Journal*, September.

Bianchi, Patrizio, and Sandrine Labory. 2018. *Industrial Policy for the Manufacturing Revolution: Perspectives on Digital Globalisation*. Edward Elgar Publishing.

———. 2019. ‘Regional Industrial Policy for the Manufacturing Revolution: Enabling Conditions for Complex Transformations’. *Cambridge Journal of Regions, Economy and Society* 12 (2). Oxford University Press UK: 233–49.

Billor, Nedret, Ali S. Hadi, and Paul F. Velleman. 2000. ‘BACON: Blocked Adaptive Computationally Efficient Outlier Nominators’. *Computational Statistics & Data Analysis* 34 (3). Elsevier: 279–98.

Blanchard, Olivier J., William D. Nordhaus, and Edmund S. Phelps. 1997. ‘The Medium Run’. *Brookings Papers on Economic Activity* 1997 (2). JSTOR: 89–158.

Bröcker, Johannes, and Piet Rietveld. 2010. ‘Infrastructure and Regional Development’. In *Handbook of Regional Growth and Development Theories*, edited by Roberta Capello and Peter Nijkamp, 152–81. Edward Elgar Publishing.

Bronzini, Raffaello, and Eleonora Iachini. 2014. ‘Are Incentives for R&D Effective? Evidence from a Regression Discontinuity Approach’. *American Economic Journal: Economic Policy* 6 (4): 100–134.

Brunazzo, Marco. 2016. ‘The History and Evolution of Cohesion Policy’. In *Handbook on Cohesion Policy in the EU*, edited by Simona Piattoni and Laura Polverari, 17–35. Edward Elgar Publishing.

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap. 2008. ‘Zombie Lending and Depressed Restructuring in Japan’. *American Economic Review* 98 (5): 1943–77.

Caloffi, Annalisa, and Marco Mariani. 2018. ‘Regional Policy Mixes for Enterprise and Innovation: A Fuzzy-Set Clustering Approach’. *Environment and Planning C: Politics and Space* 36 (1): 28–46.

Caloffi, Annalisa, Marco Mariani, and Luca Rulli. 2013. ‘Le Politiche per Le Imprese e l’innovazione in Italia: Le Scelte Delle Regioni’. *La Finanza Territoriale in Italia: Rapporto 2013*.

Capello, Roberta. 2010. ‘Space, Growth and Development’. In *Handbook of Regional Growth and Development Theories*, edited by Roberta Capello and Peter Nijkamp, 33–52. Edward Elgar Publishing.

Capello, Roberta, and Peter Nijkamp. 1996. ‘Regional Variations in Production Network Externalities’. *Regional Studies* 30 (3). Taylor & Francis: 225–37.

Carpenter, James, and John Bithell. 2000. ‘Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians’. *Statistics in Medicine* 19 (9). Wiley Online Library: 1141–64.

Cavanaugh, Joseph E., and Andrew A. Neath. 2019. ‘The Akaike Information Criterion: Background, Derivation, Properties, Application, Interpretation, and Refinements’. *WIREs Computational Statistics* 11 (3): e1460.

Cerqua, Augusto, and Guido Pellegrini. 2018. ‘Are We Spending Too Much to Grow? The Case of Structural Funds’. *Journal of Regional Science* 58 (3): 535–63.

Ciani, Emanuele, and Guido De Blasio. 2015. ‘European Structural Funds during the Crisis: Evidence from Southern Italy’. *IZA Journal of Labor Policy* 4 (1): 20.

Cieślik, Andrzej, and Magdalena Kaniewska. 2004. ‘Telecommunications Infrastructure and Regional Economic Development: The Case of Poland’. *Regional Studies* 38 (6): 713–25.

Ciffolilli, A., S. Condello, M. Pompili, and R. Roemish. 2015. *Geography of Expenditure. Final Report. Work Package 13. Ex Post Evaluation of Cohesion Policy Programmes 2007–2013, Focusing on the European Regional Development Fund (ERDF) and the Cohesion Fund (CF)*. European Commission.

Cohen, Wesley M., and Daniel A. Levinthal. 1990. ‘Absorptive Capacity: A New Perspective on Learning and Innovation’. *Administrative Science Quarterly* 35 (1): 128.

Crescenzi, Riccardo, Marco Di Cataldo, and Andrés Rodríguez-Pose. 2016. ‘Government Quality and the Economic Returns of Transport Infrastructure Investment in European Regions’. *Journal of Regional Science* 56 (4): 555–82.

Crescenzi, Riccardo, and Mara Giua. 2016. ‘Different Approaches to the Analysis of EU Cohesion Policy: Leveraging Complementarities for Evidence-Based Policy Learning’. In *EU Cohesion Policy (Open Access)*, 21–32. Routledge.

———. 2019. ‘One or Many Cohesion Policies of the European Union? On the Differential Economic Impacts of Cohesion Policy across Member States’. *Regional Studies* 0 (0): 1–11.

Crescenzi, Riccardo, and Andrés Rodríguez-Pose. 2012. ‘Infrastructure and Regional Growth in the European Union’. *Papers in Regional Science* 91 (3). Wiley Online Library: 487–513.

———. 2013. ‘R & D, Socio-Economic Conditions, and Regional Innovation in the U. S’. *Growth and Change* 44 (2). Wiley Online Library: 287–320.

Dall’Erba, Sandy, and Fang Fang. 2017. ‘Meta-Analysis of the Impact of European Union Structural Funds on Regional Growth’. *Regional Studies* 51 (6): 822–32.

Dall’Erba, Sandy, and Julie Le Gallo. 2008. ‘Regional Convergence and the Impact of European Structural Funds over 1989–1999: A Spatial Econometric Analysis’. *Papers in Regional Science* 87 (2). Wiley Online Library: 219–44.

Dawley, Stuart. 2014. ‘Creating New Paths? Offshore Wind, Policy Activism, and Peripheral Region Development’. *Economic Geography* 90 (1). Taylor & Francis: 91–112.

De Blasio, Guido, Davide Fantino, and Guido Pellegrini. 2015. ‘Evaluating the Impact of Innovation Incentives: Evidence from an Unexpected Shortage of Funds’. *Industrial and Corporate Change* 24 (6). Oxford University Press: 1285–1314.

De la Fuente, Angel. 1997. ‘The Empirics of Growth and Convergence: A Selective Review’. *Journal of Economic Dynamics and Control* 21 (1): 23–73.

———. 2002. ‘On the Sources of Convergence: A Close Look at the Spanish Regions’. *European Economic Review* 46 (3): 569–99.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. ‘The Rise of Market Power and the Macroeconomic Implications’. *The Quarterly Journal of Economics* 135 (2). Oxford Academic: 561–644.

De Vries, Hanna, Victor Bekkers, and Lars Tummers. 2016. ‘Innovation in the Public Sector: A Systematic Review and Future Research Agenda’. *Public Administration* 94 (1). Wiley Online Library: 146–66.

Del Bo, Chiara F., and Massimo Florio. 2012. ‘Infrastructure and Growth in a Spatial Framework: Evidence from the EU Regions’. *European Planning Studies* 20 (8): 1393–1414.

Di Gennaro, Daniele, and Guido Pellegrini. 2019. ‘Are Regional Policies Effective? An Empirical Evaluation on the Diffusion of the Effects of R&d Incentives’. *Politica Economica* 35 (1). Società editrice il Mulino: 3–26.

Duschl, Matthias, Tobias Scholl, Thomas Brenner, Dennis Luxen, and Falk Raschke. 2015. ‘Industry-Specific Firm Growth and Agglomeration’. *Regional Studies* 49 (11). Routledge: 1822–39.

Ederveen, Sjef, Henri L. F. de Groot, and Richard Nahuis. 2006. ‘Fertile Soil for Structural Funds?A Panel Data Analysis of the Conditional Effectiveness of European Cohesion Policy’. *Kyklos* 59 (1): 17–42.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Egger, Peter H., Maximilian v. Ehrlich, and Douglas R. Nelson. 2020. ‘The Trade Effects of Skilled versus Unskilled Migration’. *Journal of Comparative Economics*, January, S0147596718304189.

Egger, Peter H., and Maximilian von Ehrlich. 2013. ‘Generalized Propensity Scores for Multiple Continuous Treatment Variables’. *Economics Letters* 119 (1): 32–34.

Egger, Peter Hannes, and Peter Egger. 2016. ‘Heterogeneous Effects of Tariff and Nontariff Policy Barriers in General Equilibrium’. Kiel und Hamburg: ZBW-Deutsche Zentralbibliothek für ….

Elburz, Zeynep, Peter Nijkamp, and Eric Pels. 2017. ‘Public Infrastructure and Regional Growth: Lessons from Meta-Analysis’. *Journal of Transport Geography* 58 (January): 1–8.

Esposti, Roberto. 2007. ‘Regional Growth and Policies in the European Union: Does the Common Agricultural Policy Have a Counter-Treatment Effect?’ *American Journal of Agricultural Economics* 89 (1): 116–34.

Felipe, Jesus, and F Gerard Adams. 2005. '"A Theory Of Production" The Estimation Of The Cobb-Douglas Function: A Retrospective View'. *Eastern Economic Journal* 31 (3): 20.

Ferrara, Antonella Rita, Philip McCann, Guido Pellegrini, Dirk Stelder, and Flavia Terribile. 2017. 'Assessing the Impacts of Cohesion Policy on EU Regions: A Non-Parametric Analysis on Interventions Promoting Research and Innovation and Transport Accessibility'. *Papers in Regional Science* 96 (4). Wiley Online Library: 817–41.

Flores, Carlos A., Alfonso Flores-Lagunes, Arturo Gonzalez, and Todd C. Neumann. 2012. 'Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps'. *Review of Economics and Statistics* 94 (1). MIT Press: 153–71.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2012. 'The Internet and Local Wages: A Puzzle'. *American Economic Review* 102 (1): 556–75.

Fratesi, Ugo. 2016. 'Impact Assessment of EU Cohesion Policy: Theoretical and Empirical Issues'. In *Handbook on Cohesion Policy in the EU*, edited by Simona Piattoni and Laura Polverari, 443–60. Edward Elgar Publishing.

Fuente, Angel de la. 2000. *Mathematical Methods and Models for Economists*. Cambridge University Press.

Gagliardi, Luisa, and Marco Percoco. 2017. 'The Impact of European Cohesion Policy in Urban and Rural Regions'. *Regional Studies* 51 (6): 857–68.

Garone, Lucas Figal, Alessandro Maffioli, Joao Alberto de Negri, Cesar M. Rodriguez, and Gonzalo Vázquez-Baré. 2015. 'Cluster Development Policy, SME's Performance, and Spillovers: Evidence from Brazil'. *Small Business Economics* 44 (4). Springer: 925–48.

Giua, Mara. 2017. 'Spatial Discontinuity for the Impact Assessment of the EU Regional Policy: The Case of Italian Objective 1 Regions'. *Journal of Regional Science* 57 (1): 109–31.

Gramlich, Edward M. 1994. 'Infrastructure Investment: A Review Essay'. *Journal of Economic Literature* 32 (3). JSTOR: 1176–96.

Greenstein, Shane, and Ryan C. McDevitt. 2009. 'The Broadband Bonus: Accounting for Broadband Internet's Impact on US GDP'. National Bureau of Economic Research.

Grimes, Arthur, Cleo Ren, and Philip Stevens. 2012. 'The Need for Speed: Impacts of Internet Connectivity on Firm Productivity'. *Journal of Productivity Analysis* 37 (2). Springer: 187–201.

Hassink, Robert, Arne Isaksen, and Michaela Trippl. 2019. 'Towards a Comprehensive Understanding of New Regional Industrial Path Development'. *Regional Studies* 53 (11). Routledge: 1636–45.

Hirano, Keisuke, and Guido W. Imbens. 2004. 'The Propensity Score with Continuous Treatments'. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164:73–84.

Holtz-Eakin, Douglas, and Amy Ellen Schwartz. 1995. 'Infrastructure in a Structural Model of Economic Growth'. *Regional Science and Urban Economics* 25 (2). Elsevier: 131–51.

Howell, Sabrina T. 2017. 'Financing Innovation: Evidence from R&D Grants'. *American Economic Review* 107 (4): 1136–64.

Iain Begg. 2016. 'The Economic Theory of Cohesion Policy'. In *Handbook on Cohesion Policy in the EU*, edited by Simona Piattoni and Laura Polverari, 50–64. Edward Elgar Publishing.

Isaksson, Anders. 2010. *Energy Infrastructure and Industrial Development*. United Nations Industrial Development Organization.

Islam, Nazrul. 2003. 'What Have We Learnt from the Convergence Debate?' *Journal of Economic Surveys* 17 (3). Wiley Online Library: 309–62.

Jolly, Suyash, Markus Grillitsch, and Teis Hansen. 2020. 'Agency and Actors in Regional Industrial Path Development. A Framework and Longitudinal Analysis'. *Geoforum* 111: 176–88.

Karabarbounis, Loukas, and Brent Neiman. 2014. 'The Global Decline of the Labor Share'. *The Quarterly Journal of Economics* 129 (1). Oxford Academic: 61–103.

Kolko, Jed. 2012. 'Broadband and Local Growth'. *Journal of Urban Economics* 71 (1). Elsevier: 100–113.

Koutroumpis, Pantelis. 2009. ‘The Economic Impact of Broadband on Growth: A Simultaneous Approach’. *Telecommunications Policy* 33 (9). Elsevier: 471–85.

Krugman, Paul, and Anthony J. Venables. 1990. ‘Integration and the Competitiveness of Peripheral Industry.’ 363.

Labory, Sandrine, and Patrizio Bianchi. 2021. ‘Regional Industrial Policy in Times of Big Disruption: Building Dynamic Capabilities in Regions’. *Regional Studies*. Taylor & Francis, 1–10.

Lee, David S., and Thomas Lemieux. 2010. ‘Regression Discontinuity Designs in Economics’. *Journal of Economic Literature* 48 (2): 281–355.

MacKinnon, Danny, Stuart Dawley, Andy Pike, and Andrew Cumbers. 2019. ‘Rethinking Path Creation: A Geographical Political Economy Approach’. *Economic Geography* 95 (2). Taylor & Francis: 113–35.

Madden, Gary, and Scott J Savage. 1998. ‘CEE Telecommunications Investment and Economic Growth’. *Information Economics and Policy* 10 (2): 173–95.

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. ‘A Contribution to the Empirics of Economic Growth’. *The Quarterly Journal of Economics* 107 (2). MIT Press: 407–37.

Martin, Philippe, and Carol Ann Rogers. 1995. ‘Industrial Location and Public Infrastructure’. *Journal of International Economics* 39 (3–4). Elsevier: 335–51.

Mcdonald, Frank, Qihai Huang, Dimitrios Tsagdis, and Heinz Josef Tüselmann. 2007. ‘Is There Evidence to Support Porter-Type Cluster Policies?’ *Regional Studies* 41 (1). Routledge: 39–49.

Mokyr, Joel. 2010. ‘The Contribution of Economic History to the Study of Innovation and Technical Change: 1750–1914’. *Handbook of the Economics of Innovation* 1. Elsevier: 11–50.

Molle, Willem. 2007. *European Cohesion Policy*. Routledge.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. 1st ed.

Neumark, David, and Helen Simpson. 2015. ‘Chapter 18 - Place-Based Policies’. In *Handbook of Regional and Urban Economics*, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 5:1197–1287. Handbook of Regional and Urban Economics. Elsevier.

Olsen, Jonathan. 2020. *The European Union: Politics and Policies*. Routledge.

Pellegrini, Guido, Flavia Terribile, Ornella Tarola, Teo Muccigrosso, and Federica Busillo. 2013. ‘Measuring the Effects of European Regional Policy on Economic Growth: A Regression Discontinuity Approach’. *Papers in Regional Science* 92 (1): 217–33.

Percoco, Marco. 2005. ‘The Impact of Structural Funds on the Italian Mezzogiorno, 1994-1999’. *Région et Développement* 21: 141–52.

———. 2013. ‘Strategies of Regional Development in European Regions: Are They Efficient?’ *Cambridge Journal of Regions, Economy and Society* 6 (2): 303–18.

———. 2017. ‘Impact of European Cohesion Policy on Regional Growth: Does Local Economic Structure Matter?’ *Regional Studies* 51 (6): 833–43.

Pike, Andy, Andrés Rodríguez-Pose, and John Tomaney. 2016. *Local and Regional Development*. Routledge.

Potter, Antony, and H. Doug Watts. 2011. ‘Evolutionary Agglomeration Theory: Increasing Returns, Diminishing Returns, and the Industry Life Cycle’. *Journal of Economic Geography* 11 (3). Oxford University Press: 417–55.

Puga, Diego. 2002. ‘European Regional Policies in Light of Recent Location Theories’. *Journal of Economic Geography* 2 (4). Oxford University Press: 373–406.

Rietveld, Piet, and Frank Bruinsma. 2012. *Is Transport Infrastructure Effective?: Transport Infrastructure and Accessibility: Impacts on the Space Economy*. Springer Science & Business Media.

Rodríguez-Pose, Andrěs. 1999. ‘Innovation Prone and Innovation Averse Societies: Economic Performance in Europe’. *Growth and Change* 30 (1). Wiley Online Library: 75–105.

———. 2020. 'Institutions and the Fortunes of Territories'. *Regional Science Policy & Practice* 12 (3). Wiley Online Library: 371–86.

Rodríguez-Pose, Andrés, and Marco Di Cataldo. 2015. 'Quality of Government and Innovative Performance in the Regions of Europe'. *Journal of Economic Geography* 15 (4). Oxford University Press: 673–706.

Rodriguez-Pose, Andrés, and Ugo Fratesi. 2004. 'Between Development and Social Policies: The Impact of European Structural Funds in Objective 1 Regions'. *Regional Studies* 38 (1). Routledge: 97–113.

Rodríguez-Pose, Andrés, and Roberto Ganau. 2021. 'Institutions and the Productivity Challenge for European Regions'. *Journal of Economic Geography*, no. lbab003 (June).

Rodríguez-Pose, Andrés, and Enrique Garcilazo. 2015. 'Quality of Government and the Returns of Investment: Examining the Impact of Cohesion Expenditure in European Regions'. *Regional Studies* 49 (8): 1274–90.

Roller, Lars-Hendrik, and Leonard Waverman. 2001. 'Telecommunications Infrastructure and Economic Development: A Simultaneous Approach'. *American Economic Review* 91 (4): 909–23.

Romer. 2017. *Advanced Macroeconomics*.

Sotarauta, Markku. 2017. 'Making Sense of Leadership in Urban and Regional Development'. *Regional Studies*, 8.

Stephenson, Paul. 2016. 'The Institutions and Procedures of Cohesion Policy'. In *Handbook on Cohesion Policy in the EU*, edited by Simona Piattoni and Laura Polverari, 36–49. Edward Elgar Publishing.

Stern, David I. 2010. 'The Role of Energy in Economic Growth'. *USAEE-IAEE Working Paper*, no. 10–055.

Stern, David I., Paul J. Burke, and Stephan B. Bruns. 2019. 'The Impact of Electricity on Economic Development: A Macroeconomic Perspective', May.

Surubaru, Neculai-Cristian. 2017. 'Administrative Capacity or Quality of Political Governance? EU Cohesion Policy in the New Europe, 2007–13'. *Regional Studies* 51 (6): 844–56.

Toman, Michael T., and Barbora Jemelkova. 2003. 'Energy and Economic Development: An Assessment of the State of Knowledge'. *The Energy Journal* 24 (4). International Association for Energy Economics.

Tranos, Emmanouil. 2012. 'The Causal Effect of the Internet Infrastructure on the Economic Development of European City Regions'. *Spatial Economic Analysis* 7 (3). Routledge: 319–37.

Vickerman, Roger. 2018. 'Can High-Speed Rail Have a Transformative Effect on the Economy?' *Transport Policy*, Selected papers presented at the 14th World Conference of Transport Research under Topic Area E: Transport Economics and Finance, 62 (February): 31–37.

Vickerman, Roger, Klaus Spiekermann, and Michael Wegener. 1999. 'Accessibility and Economic Development in Europe'. *Regional Studies* 33 (1). Taylor & Francis Group: 1–15.

Vu, Khuong M. 2011. 'ICT as a Source of Economic Growth in the Information Age: Empirical Evidence from the 1996–2005 Period'. *Telecommunications Policy* 35 (4). Elsevier: 357–72.

Weber, Sylvain. 2010. 'Bacon: An Effective Way to Detect Outliers in Multivariate Data Using Stata (and Mata)'. *The Stata Journal* 10 (3). SAGE Publications Sage CA: Los Angeles, CA: 331–38.

Wolman, Harold, and Diana Hincapie. 2015. 'Clusters and Cluster-Based Development Policy'. *Economic Development Quarterly* 29 (2). SAGE Publications Sage CA: Los Angeles, CA: 135–49.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Xiong, Ailun, Senmao Xia, Zhen Peter Ye, Dongmei Cao, Yanguo Jing, and Hongyi Li. 2020. 'Can Innovation Really Bring Economic Growth? The Role of Social Filter in China'. *Structural Change and Economic Dynamics* 53 (June): 50–61.

Zeira, Joseph. 1998. ‘Workers, Machines, and Economic Growth’. *The Quarterly Journal of Economics* 113 (4). MIT Press: 1091–1117.

Zuleta, Hernando. 2012. ‘Variable Factor Shares, Measurement and Growth Accounting’. *Economics Letters* 114 (1): 91–93.