UNIVERSITÉ DU LUXEMBOURG

PhD-FDEF-2022-007
The Faculty of Law, Economics and Finance

DISSERTATION

Presented on 01/06/2022 in Luxembourg
to obtain the degree of

Docteur de l'université du Luxembourg
en sciences économiques

by

Alessio Monetti
Born on 28 August 1985 in Naples (Italy)

Applying matching models with imperfect transferable utility to diverse markets

Dissertation Defense Committee

Prof.Dr.Arnaud Dupuy	(Supervisor, University of Luxembourg)
Prof.Dr.Skerdilajda Zanaj	(Chair, University of Luxembourg)
Prof.Dr.Louis Chauvel	(Vice-Chair, University of Luxembourg)
Prof.Dr.Simon Weber	(Member, University of York)
Dr.Audrey Bousselin	(Member, LISER)

Acknowledgements

I am extremely grateful to my supervisor, Prof.Dr.Arnaud DUPUY for his invaluable advice, continuous support, patience, kindness, and human attitude during my PhD study. His experience and expertise has encouraged me at all times of my research, and I learned a lot from discussions with him. He has always been thoughtful and has always put much time into my research-related issues. I could not have completed this thesis save for his help.

Contents

Introduction

Evolution of the matching models

At the early stage the theory of matching model has been developed and adopted to investigate the marriage market. Indeed the first attempt has been proposed by Gale and Shapley (1962) who conceived the marriage market as matching game in which each individual of heterogeneous sex decide to pair with the intent to reach a higher well-being than they wouldn't get by remaining singles. This model assumes that the outcome generated by the marriage for each couple is fully determined by the attributes of the individuals forming the couple, therefore none partner can transfer utility to the other partner as a form of compensation whether it exhibits deficiencies in its attributes. In the context of the marriage market this assumption appears to be unrealistic and challenging to maintain given the fact that some commodities are privately consumed and then one partner may decide to decrease private consumption for the partner's benefit.

Becker (1973), Becker (1991) makes the transfer of resources feasible. However differently from the previous model the allocation of the outcome generated by the couple is now determined endogenously. Becker's contribution is mostly related to the study of the homogamy pattern (higher types of one sex tend to match with higher types of the opposite sex). In the attempt to explain this phenomenon he developed a model of positive assortative mating (PAM) in which the types of partners are one-dimensional and are complementary in producing the marital surplus. In that model Becker created a convenient index which embedded the socio-economic features relevant in the marriage market, and the prediction of that model indicates that men and women matching are as close as possible regarding this index.

Nonetheless as Chiappori (2020) underlined two potential issues arise with this approach: individuals with different socio-economic characteristics may have the same index making them perfect substitutes and more the implications of the Becker's model are quite unrealistic given that the matches observed in the data may occur between partners exhibiting different features. This implicitly recognizes that the matching is multidimensional and other characteristics not known to the econometricians may be relevant.

Choo and Siow (2006) own the merit to have introduced a convenient "sympathy shock" associated to the unobserved heterogeneities entering the surplus function additively and distributed as type-I extreme value distribution. A no negligible aspect of this framework is that the individual's choice is now constrained by the choices of the other individuals and that the marriage market must clear. Along this literature Dupuy and Galichon (2014) extended the work of Choo and Siow (2006) to the multidimensional case enabling to consider discrete as well as continuous observables. The first proposing this extension was Dagsvik (1994), who assumed that within a market each agent is facing only a countable set of choices, associated to a random

Poisson process whose intensity is ruled by type-I extreme value distribution. The merit of Dupuy and Galichon (2014) have been to extend to the matching framework this process. The appealing of this approach is that the probability that an agent matches another agent can be directly identified by the logit formula. Lastly, Dupuy and Galichon (2014) also present a convenient bilinear parametrization allowing to estimate the complementarity and substitutability between the agents' characteristics (*affinity matrix*).

From Becker on the matching models developed enable the transfer of utility between agents. In most of the cases, the transfer technology is such that one individual gives up one unit of his/her utility to increase by one unit the utility of the partner (this refers to the Transferable utility setting (TU)). However as recognized in Galichon et al. (2017) when the transfer is under the form of favor exchange (rather than a monetary transfer) even in the marriage market the cost/benefit may not exactly match. In that case the transfer of utility is possible but with friction (this refers to the Imperfect Transferable utility setting (ITU)).

Indeed several attempts have been proposed in order to deal with the cases in which frictions occur in the transfer of utility between agents: Legros and Newman (2007) introduce the conditions underlying the assortative pairing in ITU case. They employ those conditions to the marriage markets in which partners with different risk attitudes share exogenous uninsurable risks and incomes. They find out a negative assortativeness when men and women decide to match, meaning that the most risk averse male is paired with the least risk adverse female; Chiappori and Reny (2016) extend this work by considering general risk averse preferences. They find out similar results.

Along that strand of literature, a general and systematic framework to deal with imperfect transferable utility setting (ITU) has been proposed in Galichon et al. (2017). The seminal paper discloses the conditions making the matching stable and feasible in TU and ITU case. Furthermore they prove the existence and uniqueness of the equilibrium matching when assuming that agents have heterogeneous tastes and propose a convenient and faster procedure to compute the agents' utilities at equilibrium, through the iterative projection fitting procedure (IPFP). Lastly they provide a suitable approach to identify the feasible set of solutions in the ITU case. The challenging arises given that the straight line representing the Pareto frontier in the TU case (delimiting the feasible set of solutions) is replaced by the set of bargaining frontiers in the ITU case which envelope generates the feasible set of solutions. Specifically, they show that the intersection and union of bargaining frontiers in the agents' utilities space is equivalent to compute the minimum and maximum of the corresponding matching functions.

The short introduction above provides a brief digression concerning the evolution of the

matching models and presents the key assumptions which have been constantly employed in the thesis for constructing matching models with imperfect transferable utility applied to the diverse markets proposed.

Chapter 1[1] deals with the decision processes of the agents forming the Luxembourg childcare market, namely households and childcare providers. The construction of the matching model is complicated by the presence of two types of childcare providers (public and private) and the Luxembourg system of prices which introduces an important element of friction (price is regulated in the public sector and subsidies are offered to households purchasing private care). By employing administrative data of 2016 from the Ministry of Education, Childhood and Youth (MECY) and the National Insurance System (IGSS) we find out the preferences of the agents forming the Luxembourg childcare market. Afterwards the estimated preferences are deployed to simulate a counterfactual scenario by measuring the impact of a different quality standards on the welfare of the households.

Chapter 2 measures the preferences of the agents forming the labor market, namely workers and firms. Therefore we build a matching model with imperfect transferable utility occurring between worker and firm by emphasizing the frictional role played by the taxation. By employing the 2015 march supplement of the Current Population Survey (ASEC) administered by the US Census, we investigate the effect of taxation on the jobs mismatch which we define through a normative approach as the discrepancy between the level of education of the worker and the job qualification requirements (Berlingieri and Erdsiek (2012); Stoevska (2017))). The findings clearly indicate that the taxation is able to reduce dramatically the capacity of the firm to compensate for the disutility of the worker. The taxation may therefore trigger two contrasting effects on the jobs mismatch depending on the worker education: highly educated workers would tend to respond to the taxation by choosing jobs for which they are overeducated (increasing the jobs mismatch) while workers owning at most the high school diploma would react to taxation by choosing jobs more appropriate with their education level (decreasing the jobs mismatch).

Chapter 3 investigates the decision mechanisms underlying the fruitful collaborations between university department and firm. Particularly, we focus on studying the interplay between the geographic proximity (representing the spatial closeness of the agents) and the cognitive proximity (representing the similarity of the knowledge base of the agents) in driving the partnerships between university department and firm. In that case we build a matching model where the R&D tax credit the company benefits represents the element making the transfer between the

[1]This chaper is based on the working paper "Childcare, Quality and Location: A Hedonic Approach", written by Audrey Bousselin, Arnaud Dupuy, and Alessio Monetti. I am indebted to Prof. Dupuy and Dr. Bousselin for letting me using the paper as chapter 1 of my PhD thesis

university department and firm imperfect. Unfortunately we do not have at our disposal a sample disclosing the observed matching between university departments and firms, and hence we are compelled to create a dataset ad hoc (by merging a sample of US firms from Compustat and a sample of US university departments from National Science Foundation (NSF)). We therefore propose two simulated scenarios: one obtained by varying the level of preferences attached to the geographic proximity, the other obtained by varying the level of preferences attached to the cognitive proximity. The latter can only be measured ex-post (after simulation), moreover we vary the level of preferences associated to the interaction of the firm R&D intensity with the quality of university department (assuming that the complementarity of these variables represents a good indicator of the cognitive similarity). Overall, the findings reveal the potential substitutability between cognitive and geographic proximity.

Details on the estimation strategy

In those markets the matchings and the transfers between the agents are generally observed. Moreover the preferences of the agents are derived by maximizing a likelihood function which it is made of two parts: one identifying the likelihood of observing the matchings and the other identifying the likelihood of observing the transfers. The essential elements forming those expressions are the parametrized preferences of the agents (which are the objects of our estimation) and the utilities of the agents (which are solutions of the agent problems).

Optimization

The estimation strategy adopted makes large use of second-order optimization algorithm (BFGS), therefore it relies on the second-order condition for obtaining the minimum of the likelihood function (as typically done in optimization we simply turn a maximization into a minimization problem). The crucial condition of the Hessian is that it needs to be positive definite. However the BFGS method (which belongs to the family of Quasi-Newton methods) does not require to compute and to store the Hessian at each step of the optimization procedure. Indeed the BFGS method computes an approximate Hessian at each step by making use of the information concerning the gradient of the likelihood function (which it is analytically supplied) in the previous step. Two conditions are then employed to achieve this computation:

- the secant equation which approximates the Hessian by making use of the gradient
- the curvature condition which guarantees the convexity of the approximate Hessian

Two choices are possible to initialize the BFGS method:

- approximate the Hessian by the identity matrix, therefore the first step of the optimization is a gradient descent
- compute the true Hessian and its inverse

The end of the optimization supplies the Hessian which allows to compute its inverse. The latter (which coincides with the Fisher information) is employed to derive the standard errors (King (1998)).

According to Tan and Lim (2019), the advantage of using Quasi-Newton method are multiples such as it does not need:

- to compute the Hessian matrix which is a time-consuming practice
- to store the Hessian which requires large memory

IPFP algorithm

Within the optimization process an important part is devoted to the computation of the agent utilities, which are identified by the amount of utility accruing to each agent from the pairing. Those utilities are obtained by fulfilling the feasibility constraints which simply state that any given agent of one side of the market can be paired to one agent of the other side at most (Chiappori and Salanié (2016)). Therefore the feasibility conditions coincide with the solution of a system of non-linear equations implying mathematically that the mass of pairing of agent types (x, y) $(\mu(x, y))$ with agents of type x needs to be equal to the mass of agent type x $(m(x))$ and that the mass of pairing of agent types (x, y) $(\mu(x, y))$ with agents of type y needs to be equal to the mass of agent type y $(n(y))$:

$$\begin{aligned} \sum_{y} \mu(x, y) &= m(x) \\ \sum_{x} \mu(x, y) &= n(y), \end{aligned}$$

where $\mu(x, y)$ represents the matching function containing $u(x)$ and $v(y)$, the utilities of the agents.

Since the system involves dependent equations, it is necessary to impose a normalization such that the solution of the system is unique (Dupuy and Galichon (2015)). A widely adopted methodology to solve this system hinges on the iterative proportional fitting procedure (IPFP) (Galichon and Salanié (2015); Galichon et al. (2015); Galichon et al. (2017); Dupuy et al. (2017)). The IPFP allows to solve the system by iterating the expression of $u(x)$ from $v(y)$, and vice-versa

(Galichon et al. (2017)). Therefore it allows to recast $u(x)$ and $v(y)$ as an implicit function of one another. The core steps of the algorithm are the followings:

IPFP

1-Step 0 $\rightarrow$ Fix the initial value $v^0(y)$ at $v^0(y) = m(y)$

2-Step t $\rightarrow$ Replace $v^0(y)$ in $\mu(x, y)$ and solve $\sum_y \mu(x, y) = m(x)$ for $u^t(x)$

3-Step t+1 $\rightarrow$ Replace $u^t(x)$ in $\mu(x, y)$ and solve $\sum_x \mu(x, y) = n(y)$ for $v^{t+1}(y)$

Repeat step 2 and 3 until the requirements of the terminal condition are met, i.e after T periods $sup_x|u^T(x) - u^{T-1}(x)| < \epsilon$ and $sup_y|v^T(y) - v^{T-1}(y)| < \epsilon$

The use of this algorithm is particularly suitable for large markets (more than 100 observations) and may be easily adjusted to accommodate parallel computation (Chen et al. (2019))

MATLAB

The type of matching model proposed enables a natural formulation based on the matrix algebra, therefore to implement the optimization procedure the MATLAB environment appears to be particularly suitable (The MathWorks (2019)). Indeed MATLAB allows to perform matrix operations with ease since it embeds several built-in mathematical functions which can be readily applied to matrices or arrays. Specifically, the vectorization which facilitates the combination of arrays is implemented through one command, despite the use of for or while loop as required by other softwares such as R, fortran and C (M.T.Gastner (2019)). This is an essential property since the code proposed involves the regular combination of arrays and matrices. Furthermore MATLAB has also developed several tools for parallel computing, which enables to easily parallelize the part of the code associated with the computation of the feasibility constraints. Lastly, MATLAB consents to run the code by employing an interactive or batch mode. The latter feature is of primary importance since the code has been parallelized and launched on the High Performance Computing (HPC) of the university of Luxembourg. The use of the PCs cluster is further beneficial since the appropriate levels of randomness to be included in the agent problems (the agents maximize random utility functions) are inferred from a grid search procedure, which may then be implemented simultaneously for different combinations of those levels with a remarkable saving of time.

Bibliography

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political economy*, 81(4):813–846.

Becker, G. S. (1991). A treatise on the family (revised and enlarged edition).

Berlingieri, F. and Erdsiek, D. (2012). How relevant is job mismatch for german graduates? *ZEW-Centre for European Economic Research Discussion Paper*, (12-075).

Chen, L., Choo, E. S. Y., Galichon, A., and Weber, S. (2019). Matching function equilibria with partial assignment: Existence, uniqueness and estimation. *Uniqueness and Estimation (May 13, 2019)*.

Chiappori, P.-A. (2020). The theory and empirics of the marriage market. *Annual Review of Economics*, 12:547–578.

Chiappori, P.-A. and Reny, P. J. (2016). Matching to share risk. *Theoretical Economics*, 11(1):227–251.

Chiappori, P.-A. and Salanié, B. (2016). The econometrics of matching models. *Journal of Economic Literature*, 54(3):832–61.

Choo, E. and Siow, A. (2006). Who marries whom and why. *Journal of political Economy*, 114(1):175–201.

Dagsvik, J. K. (1994). Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica: Journal of the Econometric Society*, pages 1179–1205.

Dupuy, A. and Galichon, A. (2014). Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319.

Dupuy, A. and Galichon, A. (2015). A note on the estimation of job amenities and labor productivity.

Dupuy, A., Galichon, A., Jaffe, S., and Kominers, S. D. (2017). Taxation in matching markets.

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

Galichon, A., Kominers, S., and Weber, S. (2017). Costly concessions: An empirical framework for matching with imperfectly transferable utility.

Galichon, A., Kominers, S. D., and Weber, S. (2015). The nonlinear bernstein-schrodinger equation in economics. In *International Conference on Networked Geometric Science of Information*, pages 51–59. Springer.

Galichon, A. and Salanié, B. (2015). Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*.

King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

Legros, P. and Newman, A. F. (2007). Beauty is a beast, frog is a prince: Assortative matching with nontransferabilities. *Econometrica*, 75(4):1073–1102.

M.T.Gastner (2019). Data analysis and visualization with r.

Stoevska, V. (2017). Qualification and skill mismatch: Concepts and measurement. *ILO*.

Tan, H. H. and Lim, K. H. (2019). Review of second-order optimization techniques in artificial neural networks backpropagation. In *IOP Conference Series: Materials Science and Engineering*, volume 495, page 012003. IOP Publishing.

The MathWorks, I. (2019). *Symbolic Math Toolbox*. Natick, Massachusetts, United State.

Chapter 1

Childcare, quality and location: a hedonic approach

1 Introduction

Child care provision is at the heart of numerous important societal questions by having crucial implications on the current and future generations. First, child care provision affects many important and related household decisions, including fertility, labour supply, residential and job location. While the question of the link between child care provision and maternal labour supply is well documented in the literature (Kalb (2009); Blau and Currie (2010)), the link between residential mobility and child care provision has been so far overlooked, due to methodological as well as data limitations, with the exception of Compton and Pollak (2014) who focus on informal care. Second, child care provision is of crucial importance in the building of cohesive and skilled societies. Recent research shows that early childhood interventions foster the development of cognitive and non-cognitive skills, promote educational achievement and ensure better health, social and professional outcomes later in life (Heckman et al. (2013)). Third, the benefits of child care provision go well beyond private returns, by generating positive externalities, through the reduction of inequalities of opportunities and greater social cohesion, which motivates public intervention. Consequently, over the last decades, both academics and policy makers have paid growing attention to the implications of the observed shift towards external child care provision and the accessibility and quality of this provision.

This paper aims at providing a structural analysis involving both the supply and demand sides of the childcare market, by means of a unifying matching-hedonic model, which will be developed and tested for Luxembourg. This method will provide structural estimates of households' willingness to pay for child care quality and will allow us to evaluate the impact of various child care policies such as quality standards and vouchers through counterfactual exercises.

The success of policy interventions aimed at improving child care quality depends crucially on households' willingness to purchase and providers' willingness to supply higher quality of care. Identifying and estimating both requires considering child care as a differentiated product sold by care providers and purchased by households, hence applying a structural hedonic model to the market of child care quality. Only few attempts along this line have been done in the literature, i.e. Blau and Hagy (1998); Hagy (1998) to the extent of our knowledge. Both papers use estimation methods that require multimarket data to identify and estimate households' marginal willingness to pay for quality in a classical hedonic model (Rosen, 1974). The key identification strategy is that there exists distinct markets with no mobility possible across markets, preferences and technology are the same in all markets but the distribution of households types and providers types are different. In this setting, Blau and Hagy (1998); Hagy (1998) consistently find that while income effects on the demand for quality are small, price effects are non-negligible. Their estimation

approach has the following weaknesses: i) it does not account for mobility of households and providers across markets (households must choose a providers within their regional market and providers must locate their business in their regional market), ii) it does not allow for governmental regulation by assuming perfectly transferable utility that is the price paid by households is equal to the payment received by providers.

While these shortcomings are generic to the methods developed in the 1980s, e.g. Bartik (1987); Epple (1987), and not specific to Kahn and Lang (1988), recent breakthroughs have been achieved in the several related strands of the literature that open up new avenues to tackle the three issues mentioned above. First, Choo and Siow (2006); Dupuy and Galichon (2014); Dupuy (2018) show how perfectly transferable utility matching and hedonic models with unobserved heterogeneity can be written as two-sided discrete choice problems and identified using a single market. Second, Galichon et al. (2017); Dupuy et al. (2017) developed extensions of the previous models to settings in which utility is imperfectly transferable. In this paper, we build on this literature and show how the childcare market, can be seen as a (hedonic) matching market with piecewise linear transfers as a result of governmental regulations.

This model allows us to understand parents' decisions regarding the use of formal childcare taking into account the supply of care, the quality of care and its price. In this application of the unified model, households are characterized by a set of observable (to the econometrician) attributes, as well as an unobserved (to the econometrician) heterogeneity. Households have decided where to live and where to work in a previous step; they have now to decide where to put their children cared for.

On the other side of the market, potential childcare providers are differentiated by their observable type, capturing various dimensions associated with the quality of care provided, and unobservable (to the econometrician) heterogeneity. Providers must choose in which locality to establish their business.

In this model, parents will choose the childcare option that maximizes their utility among all alternatives. Similarly, childcare providers choose the locality to establish their business in order to maximize their utility. The price function is determined in equilibrium such as to equate, at each locality, the households' demand for quality to providers' supply of quality.

The remainder of the chapter is organized as follows: Section 2 presents the economic model, with the equilibrium matching and price equations; Section 3 provides the estimation strategy; Section 4 presents the administrative dataset used to derive the preferences of households and childcare provider for the Luxembourg childcare market; Section 5 presents the empirical results; Section 6 discusses the simulation; Section 7 concludes.

2 Economic model

The formal childcare market is characterized by:

- a highly differentiated supply of care quality with public and private providers competing.
- a highly differentiated demand for care quality with households of various composition (single-parent, high/low education, native/migrants, (un)available grandparents etc.).

Two important questions (for e.g. policy makers) arise:

- What is households' marginal willingness to pay (MWTP) for quality of care?
- What is formal providers' marginal willingness to accept (MWTA) providing higher quality of care?

Identifying and estimating both requires considering childcare as a differentiated product sold by heterogeneous care providers and purchased by heterogeneous households. Our aim is to uncover structural estimates of households' MWTP for and providers' MWTA childcare quality using data from a single market.

The methodology used is this paper draws from the model developed in Dupuy (2018) that encompasses the Tinbergen/Rosen hedonic model and the Becker matching model. This framework is suited for hedonic/matching markets with (perfectly) transferable utility (TU). In TU hedonic models, the price paid by the consumer corresponds exactly to the price charged by the producer. There are no "frictions" on transfers. In contrast, in regulated markets, frictions on the transfer break the symmetry between the price charged and the price paid. This is the case in markets regulated by the government through taxes, subventions and subsidies. The market for childcare is to a large extend regulated by the government through price regulation of public providers and subsidies granted to households using private care providers. In this paper, we therefore extend the encompassing model to the case of imperfect transferable utility (ITU). This extension to the ITU case draws from the recent work by Galichon et al. (2017) and in particular its piecewise linear application in Dupuy et al. (2017). Indeed, as it turns out, subsidies in the private sector and price regulation in the public sector are such that frictions on the transfers can be seen as (negative) taxes. In that respect, the childcare market can be thought of as an encompassing hedonic market with (piecewise) linear taxation (not necessarily convex) on transfers.

2.1 Childcare location

We introduce different locations because space plays an important role in shaping the marginal willingness to pay for and the marginal willingness to provide childcare. Regarding families, access (distance) to childcare providers is often a key determinant of childcare choices, especially for low income families (Herbst and Barnow (2008)). First, because it lowers commuting costs, it is more convenient to use childcare services located close to family's home. Second, families have better access to information on available slots when they live close to childcare services. Third, living close to family may increase families' trust in these services. Introducing different locations allows to account for benefits from local amenities into households mobility choices (Bayer and McMillan (2010)).

Regarding providers, childcare providers are sensitive to local demand conditions, such as a high prevalence of families with young children or high income families: they will tend to locate their business where the local demand conditions are favourable (Owens and Rennhoff (2014))

Let $\mathcal{Z}$ be the (finite) set of locations. To each location corresponds a local childcare market. We allow (costly) mobility across markets.

2.2 Household problem

Households are assumed to maximize their (unitary) utility. They do so by deciding where to put their child for day care $z \in \mathcal{Z}$ and for what quality $y \in \mathcal{Y}$.[1]

Households are heterogenous and we denote $x_i \in \mathcal{X} = \mathbb{R}^{d_x}$ the vector of observable attributes of a household i. The vector x_i includes for instance information about the place of residence of the household, income. There is a mass $m(x)$ of households of observed type x.

We denote $\alpha(x, z, y)$ the utility derived by an household i of observable attributes $x_i = x$ choosing a care provider at location z and providing care quality y. This utility is systematic for all households of type x. However, the model allows for unobserved (to the econometrician) heterogeneity in preferences captured by terms $\varepsilon(i, z, y)$ indicating the idiosyncratic utility household i derives from choosing a care provider at location z and of quality y. Each household i of type x draws a vector $\varepsilon(i, z, y)$ from a known distribution S_x.

We denote $p^H(x, z, y)$ the (net of subsidy)[2] price paid by a household of type x choosing a formal care provider of type (z, y).

[1] We do not model the choice of where to live $z_L \in \mathcal{Z}$ and where to work and assume these have been made prior to their childcare decision.

[2] We will see below that government may offer subsidy to some households. The net price is therefore the price paid to the care provider less the subsidy received from the government.

Hence, if household i of type x chooses day care (z, y) it enjoys utility

$$\alpha(x, z, y) - p^H(x, z, y) + \sigma_1 \varepsilon(i, z, y)$$

where σ_1 is the scaling factor of unobserved taste of household i, i.e. $\varepsilon(i, .)$.

It is further assumed that households maximize their utility. An household i of type $x_i = x$ that maximizes utility therefore solves

$$u(i) = \max_{(z,y)} (\alpha(x, z, y) - p^H(x, z, y) + \sigma_1 \varepsilon(i, z, y)). \tag{1}$$

In other words, households maximizing utility solve a discrete choice problem.

2.3 Childcare problem

The formal sector for childcare consists of public and private care providers among which households can freely choose.

Let $y_j \in \mathcal{Y} = \mathbb{R}^{d_y}$ denote the observable type of care provider j. The vector y typically contains information about whether the provider is public $y_j \in \mathcal{Y}^{PUB} \subseteq \mathcal{Y}$, or private $y_j \in \mathcal{Y} \backslash \mathcal{Y}^{PUB}$ and the quality of care with measures such as the educational level of staff, child/staff ratio etc.. There is a mass $n(y)$ of providers of observed type y.

A care provider must decide:

1. the location and,
2. the type of child (household) to care for

We denote $p^C(x, z, y)$ the price charged by provider y at location z for a child from household x.

The output of a day care provider j of type $y_j = y$ deciding to locate at z and accepting child from household x reads as

$$\gamma(x, z, y) + \sigma_2 \eta(j, x, z)$$

where $\gamma(x, z, y)$ is the systematic output for a care provider of type y located at z when caring for a child of household type x and $\eta(j, x, z)$ is the unobserved idiosyncratic output drawn from a known distribution T_y.

We consider that γ differs with regard to the type of provider. For private providers, γ is the profit because they are expected to behave like other private providers of services. For public

providers, γ is the child well-being produced through the provision of high quality of care. We assume that a care provider j of type y does

$$v(j) = \max_{(x,z)} (\gamma(x,z,y) + p^C(x,z,y) + \sigma_2 \eta(j,x,z)). \tag{2}$$

For private care providers, since the price charged is endogenously determined at equilibrium, this boils down to profits maximization. In contrast, as we explain below, for public care providers the price is set exogenously by the government. This program consists then merely in output maximization.

The hedonic market therefore boils down to a set of two discrete choice problems related to each other through the relation between p^C and p^H and to the fact that if one observes a match (i,j) then j solves problem (1) and i solves problem (2). The exact relationship will be made clear below.

2.4 Regulations in the childcare market

In a classical hedonic model the price paid equates the price charged, i.e. $p^C(.) = p^H(.)$, and is determined in equilibrium such as to equate, at each location, the households' demand for quality to providers' supply of quality. However, an application to the market for child care necessitates to take into account exogenous regulations from the government that affect the price paid by households, i.e. $p^H(.)$, and the price charged by providers, i.e. $p^C(.)$.

Regulations are very often such that the price charged by public providers and the price paid by households using public providers are set exogenously by the government and the latter is function of households characteristics x. We take this into account by denoting $\bar{p}(x)$ the price paid by households using public providers and $\bar{p}$ the price charged by public care providers as set by the government, assuming it is exogenous to the model. As a result, the net price faced by a household choosing a care provider of type (z,y) and the price charged by a care provider of type (z,y) are given as

$$\begin{aligned} p^H(x,z,y) &= \bar{p}(x) \\ p^C(z,y) &= \bar{p}, \end{aligned}$$

if $y \in \mathcal{Y}^{PUB}$.

We take into account the fact that governments also often offer subsidies to households opting for private care providers. We model this by letting the price charged by private providers

being determined in equilibrium but introducing a government subsidy, denoted $s(x,z,y)$, to households of type x choosing a private day care of type (y,z). As a result, whereas households using daycare providers do pay the charged price $p^C(x,z,y)$, their net cost is

$$p^H(x,z,y) = p^C(x,z,y) - s(x,z,y),$$

if $y \in \mathcal{Y} \backslash \mathcal{Y}^{PUB}$.

In particular, in the case of Luxembourg, government subsidies for household opting for private care providers, $y \in \mathcal{Y} \backslash \mathcal{Y}^{PUB}$, are calculated as

$$s(x,z,y) = \max\left(\min\left(p^C(x,z,y) - \bar{p}(x), \bar{p} - \bar{p}(x)\right), 0\right),$$

where $\bar{p} = \max_x \bar{p}(x)$ is the price reference used by the government as a subsidy cap.

As shown in figure 1, when the price charged by a private provider exceeds what the household of type x should pay at a public provider, the government subsidies the difference $p^C(x,z,y) - \bar{p}(x) > 0$ so that the net price is $p^H(x,z,y) = \bar{p}(x)$. However, this subsidy is subject to a maximum of $\bar{p} - \bar{p}(x)$. As a result, if the price charged by a private provider $p^C(x,z,y)$ exceeds $\bar{p}$, the household pays $p^C(x,z,y)$ but receives $\bar{p} - \bar{p}(x)$ as subsidy making up to a net price of

$$p^H(x,z,y) = p^C(x,z,y) - \bar{p} + \bar{p}(x).$$

There are two cases where the government offers no subsidy to household using private care such that $p^H(x,z,y) = p^C(x,z,y)$. The first case arises when the price charged by a private provider $p^C(x,z,y)$ is lower than what the household would pay at a public provider $\bar{p}(x)$. The second case arises for households of type x such that $\bar{p}(x) = \bar{p}$ (typically high income households).

The following lemma summarizes these regulations.

Lemma 1. *The net (of subsidy) price paid by households of type x purchasing care quality y at location z is piecewise linear in the price charged by the care provider (z,y) and reads as*

$$p^H(x,z,y) = (1-\lambda_y)\bar{p}(x) + \lambda_y\left(p^C(x,z,y) - \max\left(\min\left(p^C(x,z,y) - \bar{p}(x), \bar{p} - \bar{p}(x)\right), 0\right)\right)$$

where $\lambda_y = 1\left(y \in \mathcal{Y} \backslash \mathcal{Y}^{PUB}\right)$.

Proof. The net price paid by households of type x at a public provider y, $y \in \mathcal{Y}^{PUB}$, in location

z is given as

$$p^H(x,z,y) = \bar{p}(x).$$

The net price paid by households of type x at a private provider y, $y \in \mathcal{Y}\backslash\mathcal{Y}^{PUB}$, is given as

$$p^H(x,z,y) = p^C(x,z,y) - \max\left(\min\left(p^C(x,z,y) - \bar{p}(x), \bar{p} - \bar{p}(x)\right), 0\right).$$

Hence, letting $\lambda_y = 1\left(y \in \mathcal{Y}\backslash\mathcal{Y}^{PUB}\right)$ one can write the net price paid by households of type x at care providers of type $y \in \mathcal{Y}$ in location z as

$$p^H(x,z,y) = (1-\lambda_y)\bar{p}(x) + \lambda_y\left(p^C(x,z,y) - \max\left(\min\left(p^C(x,z,y) - \bar{p}(x), \bar{p} - \bar{p}(x)\right), 0\right)\right).$$

■

2.5 Equilibrium

Agents in this economy maximize their utilities, i.e. households solve pb. (1) and providers solve pb. (2). Assuming a large market, the expected utility of a household of type x is therefore

$$u(x) := \mathbb{E}_{S_x}\left[\max_{z,y} U(x,z,y) + \sigma_1\varepsilon(i,z,y)\right] \tag{3}$$

whereas the expected utility of provider y is

$$v(y) := \mathbb{E}_{T_y}\left[\max_{x,z} V(x,z,y) + \sigma_2\eta(j,x,z)\right], \tag{4}$$

where $U(x,z,y) = \alpha(x,z,y) - p^H(x,z,y)$ and $V(x,z,y) = \gamma(x,z,y) + p^C(x,z,y)$, are the net systematic utilities of households and providers respectively.

An equilibrium outcome can therefore be defined as in the following definition.

Definition 1. *A pair $\left(\mu, p^C\right)$ consisting of a matching $\mu \geq 0$ and a price function p^C is an equilibrium outcome if and only if μ is*

1) feasible meaning that it satisfies the accounting constraints

$$\begin{aligned}\sum_{xz}\mu(x,z,y) &= n(y)\\ \sum_{zy}\mu(x,z,y) &= m(x),\end{aligned}$$

and

2) solution to both pb. (3) and pb. (4).

As stated in lemma 1 above, the frictions on transfers are piecewise linear in transfers $p^C(x,z,y)$. As it turns out, the Pareto frontier associated with each pair of household and provider can be obtained as the intersection and union of three linear frontiers as depicted in figure 2. Our model is therefore shares similarities with the matching model with a convex tax schedule a la Dupuy et al. (2017), with the difference that in our case the "tax" schedule is not convex.

Since the frontier can be expressed as the union and intersection of elementary frontiers, it follows that Corollary 1 in Galichon et al. (2017) applies.

Corollary 1. *Galichon et al. (2017) There exists a unique equilibrium outcome $\left(\mu, p^C\right)$ in the childcare hedonic market studied in this paper.*

For the sake of notational convenience let

$$\begin{aligned}\tilde{\alpha}(x,z,y) &= \alpha(x,z,y) + \sigma_1 \log m(x), \\ \tilde{\gamma}(x,z,y) &= \gamma(x,z,y) + \sigma_2 \log n(y).\end{aligned}$$

Assuming that the unobserved heterogeneity on both side of the market follow $(0,1)-$Gumbel type I distributions, the equilibrium matching can be written as stated in Theorem 1 below.

Theorem 1. *In the (childcare) hedonic market depicted above, when idiosyncratic shocks $\varepsilon()$ and $\eta()$ are i.i.d. $(0,1)-$Gumbel type I distributed, the equilibrium matching $\mu(x,z,y)$ is given as*

$$\mu(x,z,y) = \tag{5}$$

$$(1-\lambda_y) M^0(u,v) + \lambda_y \max\left[M^1(u,v), \min\left(M^2(u,v), M^3(u,v)\right)\right]$$

where

$$M^0(u,v) := \min\left(\exp\left(\frac{\tilde{\alpha}(x,z,y) - \bar{p}(x) - u(x)}{\sigma_1}\right), \exp\left(\frac{\tilde{\gamma}(x,z,y) + \bar{p} - v(y)}{\sigma_2}\right)\right),$$

$$M^1(u,v) := \exp\left(\frac{\tilde{\alpha}(x,z,y) - u(x) + \tilde{\gamma}(x,z,y) - v(y)}{\sigma_1 + \sigma_2}\right),$$

$$M^2(u,v) := \exp\left(\frac{\tilde{\alpha}(x,z,y) - \bar{p}(x) - u(x)}{\sigma_1}\right),$$

$$M^3(u,v) := \exp\left(\frac{\tilde{\alpha}(x,z,y) - u(x) + \bar{p} - \bar{p}(x) + \tilde{\gamma}(x,z,y) - v(y)}{\sigma_1 + \sigma_2}\right).$$

Note that $u(x)$ *and* $v(y)$ *are such that*

$$\begin{aligned}\sum_{xz}\mu(x,z,y) &= n(y)\\ \sum_{zy}\mu(x,z,y) &= m(x).\end{aligned}$$

The equilibrium price paid and price charged are given respectively as:

$$\begin{aligned}p^H(x,z,y) &= \bar{p}(x),\\ p^C(x,z,y) &= \bar{p},\end{aligned}$$

if $\mu(x,z,y) = M^0(u,v)$,

$$\begin{aligned}p^H(x,z,y) &= p^C(x,z,y)\\ &= \frac{\sigma_2}{\sigma_1+\sigma_2}\left(\tilde{\alpha}(x,z,y) - u(x)\right) - \frac{\sigma_1}{\sigma_1+\sigma_2}\left(\tilde{\gamma}(x,z,y) - v(y)\right),\end{aligned}$$

if $\mu(x,z,y) = M^1(u,v)$,

$$\begin{aligned}p^H(x,z,y) &= \bar{p}(x),\\ p^C(x,z,y) &= \frac{\sigma_2}{\sigma_1}\left(\tilde{\alpha}(x,z,y) - \bar{p}(x) - u(x)\right) - \left(\tilde{\gamma}(x,z,y) - v(y)\right),\end{aligned}$$

if $\mu(x,z,y) = M^2(u,v)$ *and*

$$\begin{aligned}p^H(x,z,y) &= p^C(x,z,y) - \bar{p} + \bar{p}(x),\\ p^C(x,z,y) &= \frac{\sigma_2}{\sigma_1+\sigma_2}\left(\tilde{\alpha}(x,z,y) - u(x) + \bar{p} - \bar{p}(x)\right) - \frac{\sigma_1}{\sigma_1+\sigma_2}\left(\tilde{\gamma}(x,z,y) - v(y)\right),\end{aligned}$$

if $\mu(x,z,y) = M^3(u,v)$.

Proof. If $\lambda_y = 0$, the frontier is a NTU frontier whose optimal point is located at

$$\left(\tilde{\gamma}(x,z,y) + \bar{p}, \tilde{\alpha}(x,z,y) - \bar{p}(x)\right)$$

instead of point $\left(\tilde{\gamma}(x,z,y), \tilde{\alpha}(x,z,y)\right)$. The expression of the equilibrium matching for this

frontier is obtained by first noting that

$$\begin{aligned}\frac{\tilde{\alpha}(x,z,y)-u(x)-\bar{p}(x)}{\sigma_1} &= \log\mu(x,z,y)\\ \frac{\tilde{\gamma}(x,z,y)-v(y)+\bar{p}}{\sigma_2} &= \log\mu(x,z,y).\end{aligned}$$

Solving the system for $\mu(x,z,y)$ one has

$$\mu(x,z,y)=M^0(u(x),v(y)):=\min\left(\exp\left(\frac{\alpha(x,z,y)-u(x)-\bar{p}(x)}{\sigma_1}\right),\exp\left(\frac{\tilde{\gamma}(x,z,y)-v(y)+\bar{p}}{\sigma_2}\right)\right). \tag{6}$$

Note that in this case, the equilibrium price is simply

$$\begin{aligned}p^H(x,z,y) &= \bar{p}(x),\\ p^C(x,z,y) &= \bar{p},\end{aligned}$$

If $\lambda_y=1$, the frontier can be constructed as the union and intersection of three "elementary" frontiers.

The first elementary frontier is the one defined when $p^H(x,z,y)=p^C(x,z,y)$. This is the classical TU frontier (slope -1 passing through point $(\tilde{\gamma}(x,z,y),\tilde{\alpha}(x,z,y))$. In this situation, one has

$$\tilde{\alpha}(x,z,y)-p^C(x,z,y) = \sigma_1\log\mu(x,z,y)+u(x) \tag{7}$$

$$\tilde{\gamma}(x,z,y)+p^C(x,z,y) = \sigma_2\log\mu(x,z,y)+v(y). \tag{8}$$

Adding and solving for $\mu(x,z,y)$ yields

$$\mu(x,z,y)=M^1(u(x),v(y)):=\exp\left(\frac{\tilde{\alpha}(x,z,y)-u(x)+\tilde{\gamma}(x,z,y)-v(y)}{\sigma_1+\sigma_2}\right). \tag{9}$$

The equilibrium price in this case can be recovered by plugging equation 9 into, for instance, equation 8 to obtain

$$p^C(x,z,y)=\frac{\sigma_2}{\sigma_1+\sigma_2}(\tilde{\alpha}(x,z,y)-u(x))-\frac{\sigma_1}{\sigma_1+\sigma_2}(\tilde{\gamma}(x,z,y)-v(y)).$$

The second elementary frontier is the one where $p^H(x,z,y) = \bar{p}(x)$. For this frontier one has

$$\begin{aligned} \tilde{\alpha}(x,z,y) - \bar{p}(x) &= \sigma_1 \log \mu(x,z,y) + u(x) \\ \tilde{\gamma}(x,z,y) + p^C(x,z,y) &= \sigma_2 \log \mu(x,z,y) + v(y). \end{aligned}$$

Using the first equation one obtains

$$\mu(x,z,y) = M^2(u(x), v(y)) := \exp\left(\frac{\tilde{\alpha}(x,z,y) - \bar{p}(x) - u(x)}{\sigma_1}\right). \tag{10}$$

Using the second equation

$$p^C(x,z,y) = \frac{\sigma_2}{\sigma_1}\left(\tilde{\alpha}(x,z,y) - \bar{p}(x) - u(x)\right) - \left(\tilde{\gamma}(x,z,y) - v(y)\right).$$

Finally, the third elementary frontier is the one where $p^H(x,z,y) = p^C(x,z,y) - \bar{p} + \bar{p}(x)$. This is once again a TU frontier (slope -1) but shifted to the right by $\bar{p} - \bar{p}(x)$, i.e. passing through point $(\tilde{\gamma}(x,z,y), \tilde{\alpha}(x,z,y) + \bar{p} - \bar{p}(x))$. One then has

$$\tilde{\alpha}(x,z,y) + \bar{p} - \bar{p}(x) - p^C(x,z,y) = \sigma_1 \log \mu(x,z,y) + u(x) \tag{11}$$

$$\tilde{\gamma}(x,z,y) + p^C(x,z,y) = \sigma_2 \log \mu(x,z,y) + v(y). \tag{12}$$

Adding both and solving for $\mu(x,z,y)$ yields

$$\mu(x,z,y) = M^3(u(x), v(y)) := \exp\left(\frac{\tilde{\alpha}(x,z,y) + \bar{p} - \bar{p}(x) - u(x) + \tilde{\gamma}(x,z,y) - v(y)}{\sigma_1 + \sigma_2}\right). \tag{13}$$

The equilibrium price in this case can be recovered by plugging equation 13 into equation 12 to obtain

$$p^C(x,z,y) = \frac{\sigma_2}{\sigma_1 + \sigma_2}\left(\tilde{\alpha}(x,z,y) + \bar{p} - \bar{p}(x) - u(x)\right) - \frac{\sigma_1}{\sigma_1 + \sigma_2}\left(\tilde{\gamma}(x,z,y) - v(y)\right).$$

Applying Lemma 2 in Galichon, Kominers and Weber (2018), the equilibrium matching function can then be derived from the matching functions $M^k(u,v)$, $k = 0,1,2,3$ to obtain equation (5). ■

3 Estimation strategy

We assume from now on that the analyst has access to data about a single hedonic market. The data consist of a random sample of the population of matches of households and care providers. For each match i, the analyst observes the characteristics X_i of the household, the location of care Z_i, the care quality Y_i, the price charged by the provider P_i^C. A sample is therefore a collection $\left\{\left(X_i, Z_i, Y_i, P_i^C\right), i = 1, ..., N\right\}$ where N is the sample size. It is also assumed that the analyst knows the rules associated with the government regulation binding in the market, namely he is aware of the public pricing function $\bar{p}()$.

It is further assumed that P_i^C the observed price charged by provider i is a noisy measure of the true unobserved price P_i^{C*}. We therefore adopt a latent framework relating observed price to true price as follows

$$P_i^C = P_i^{C*} + e_i$$

where e_i are measurement errors following a $N\left(0, s^2\right)$ distribution independent of (X_i, Z_i, Y_i) and the latent price P_i^{C*} is given by the model as

$$P_i^{C*} = p^C\left(X_i, Z_i, Y_i\right).$$

where $p^C()$ is defined as in Theorem (1).

3.1 Parametrization

Let the vectors $A \in \mathbb{R}^K$ and $G \in \mathbb{R}^K$ be the respective parameters of the pre-transfer systematic utilities $\alpha(x, z, y; A)$ and $\gamma(x, z, y; G)$. In particular, let $\varphi_k(x, z, y)$ be non trivial basis functions. The pre-transfer utilities are then assumed to be linear in parameters and given as

$$\begin{aligned}
\alpha(x, z, y; A) &= \sum_{k=1}^{K} A_k \varphi_k(x, z, y) = x^T A_0 y + A_1^T y \\
\gamma(x, z, y; G) &= \sum_{k=1}^{K} G_k \varphi_k(x, z, y) = x^T \Gamma_0 y + \Gamma_1^T x
\end{aligned}$$

where the matrices of parameters A_0 and Γ_0 (*affinity matrix*) indicate the level (intensity) of complementarity or substitutability between the observables, A_1 and Γ_1 indicate the vectors of parameters assessing the direct effect of the observables.

3.2 Maximum Likelihood Estimation

Note that the log-likelihood of observing a match (x_i, z_i, y_i) given parameters (A, G) is simply

$$\begin{aligned} & \log \mu\left(x_i, z_i, y_i; A, G\right) \\ = \quad & \left(1-\lambda_y\right) M^0\left(u_i, v_i\right)+\lambda_y \max\left[M^1\left(u_i, v_i\right), \min\left(M^2\left(u_i, v_i\right); M^3\left(u_i, v_i\right)\right)\right], \end{aligned}$$

Note that $u_i = u\left(x_i; A, G\right)$ and $v_i = v\left(y_i; A, G\right)$ are obtained using the sample counterpart of the accounting constraint system

$$\begin{aligned} \sum_{i=1}^{N} \mu\left(x_i, z_j, y_j; A, G\right) &= \frac{1}{N} \\ \sum_{j=1}^{N} \mu\left(x_i, z_j, y_j; A, G\right) &= \frac{1}{N} \end{aligned}$$

with the normalization $u_1 = u\left(x_1; A, G\right) = 0$.

Moreover, from the normality and independence assumptions introduced above, the log-likelihood of observing price P_i^C given parameters (A, G) reads as

$$-\frac{\left(P_i^C - p^C\left(x_i, z_i, y_i; A, G\right)\right)^2}{2s^2} - \frac{1}{2}\log s^2.$$

It follows that the log-likelihood of a sample $\left\{\left(X_i, Z_i, Y_i, P_i^C\right), i = 1, ..., N\right\}$ given parameters $\left(A, G, s^2\right)$, reads as

$$\log L\left(A, G, s^2\right) = \sum_{i=1}^{N} \log \mu\left(x_i, z_i, y_i; A, G\right) - \sum_{i=1}^{N} \frac{\left(P_i^C - p^C\left(x_i, z_i, y_i; A, G\right)\right)^2}{2s^2} - \frac{N}{2}\log s^2.$$

4 Dataset

In this paper, we use administrative files from the Luxembourgish Ministry of Education, Childhood and Youth and from the national Insurance System, combined to spatial data. In this section, we describe these different data source, we define our final estimation sample and we provide descriptive statistics.

4.1 Administrative data on children and daycare providers

Data from the Ministry of Education, Childhood and Youth (MECY) provides information on children that are cared for in a daycare center or an official childminder and on childcare providers.

An interesting feature of this data for our analysis is that it contains an unique childcare provider identifier that allow us to link information of child i and her family to that of childcare provider j that takes care for child i. Doing that, we get a child-provider dataset that contains the following information:
Child and household characteristics:

- Age of the child
- Birth order of the child
- Household's income (see table 3 for more details about this variable)
- Locality of residence and zip code
- Identifier of the childcare provider
- Number of hours of childcare per month in provider j
- Price paid by the parents
- Price paid by the government (subsidies)

Childcare provider:

- Identifier of the childcare provider
- Type of childcare provider (public or private).
- Number of children cared for
- Number of hours provided, with respect to the different price set (see hereafter more information on the price setting)
- Total price charged to the parents
- Total subsidies from the government
- Locality of the childcare provider, zip code of the provider

We need further detailed information on childcare providers to understand the choice of provider j rather than provider k. We thus complete the child-provider dataset by information on employees from the National Insurance System (IGSS). The administrative files from IGSS provide information on all the employees in the childcare sector. For each employee, we know the identifier of her employer, the number of worked hours, earnings, tenure, the type of employer

(public or private daycare), the locality and the zip code of the employer. We thus combine the child-provider data with IGSS file on childcare sector using the name of daycare providers and zip code .

We use data from November 2016, which is the most recent data available. Our sample contains 4004 parents with children up to 3 years old and 388 childcare providers

4.2 Spatial Data

We also use geocode to compute the distance (in travel time) between children and all childcare providers. In a first step, we create a dataset containing the zip codes of each childcare provider in Luxembourg. Then, the location of childcare providers is geocoded, using the zip code to assign latitude and longitude coordinates to each childcare provider. The next step is to compute the distances in travel time by car between the location of the families and the location of each childcare provider. Travel time distances are the shortest routes between the zip codes. They are computed using Arcgis software.

The distance measure is based on zip code to satisfy with anonymity requirements. Note that in Luxembourg, zip codes correspond to a street. It means that even we do not have information on the exact addresses of child and provider, our measure of distances based on zip codes is very precise.

4.3 Price variables

From the family side, for each child, we know the price paid by the parents and the amount of the subsidies, given his characteristics. From the provider side, for each provider, we know the price charged to each child cared for, given his characteristics. Using these prices information, we compute the following variables:

- p_x = price paid in the public taken into account the characteristics of the child and of her family
- p^c = price charged by the provider
- $\bar{p} = max(p^c)$

Table 1 describes how is computed the price paid by the parents (after the subsidy). There are different prices in the voucher scheme: childcare voucher, socio-family and full prices, depending on the characteristics of the child and her family and on the quantity of care (number of hours of care per week).

In particular, for children living in household with the minimum income or at risk of exclusion, the first 25 hours are free of charge; then from the 26th to the 60th hours of care, the price will be the childcare voucher price (which is maximum 3 euros per hour); after the 60th hour of care, the price will be the full price (7.5 euros per hour). Among each categories of prices, the actual price paid by the family is set according household income and the birth order of the child (see table 2 for the details).

For the other children, the first three hours are free of charge; then from the 4th to the 24th hours of care, the price will be the childcare voucher price (maximum 3 euros per hour); from the 25th to the 60th, the price will be the socio family price (maximum 7.5 euros per hour); after the 60th hour of care, the price will be the full price (7.5 euros per hour).

4.4 Final sample

We focus on children aged 0-3 years old for the two reasons. First, compulsory schooling starts at the age of four in Luxembourg. Children under four years old are not yet enrolled in school and thus they need to be cared for all the time their parents are active on the labour market. Doing this selection, we are sure to analyze homogeneous needs of childcare. Second, this group of age is of great political concern. Since the work of Heckman and coauthors, it is widely acknowledged that early interventions have higher rates of return than later one and that it would be more efficient to promote access to quality childcare for children before they enter school.

The sample used in the analysis is made by 4004 parents with children up to 3 years old and 388 childcare providers.

Table 4 presents summary statistics of the dataset. The average household is the one exhibiting income up to 2 MW (income) choosing care for a children of about 1.5 years old (age). The average type of childcare provider is the private one (the private providers outnumber the public providers) which displays a pupil-teacher ratio of about 2 and with 36 children cared (size). The average teacher of childcare providers is female (male ratio) owning a permanent contract (contract) and receiving an hourly wage of 17 euros (wage).

5 Empirical results

According the estimation strategy presented in section 3 we proceed by estimating the preferences of households and childcare providers for the Luxembourg childcare market. As preliminary step we start by proposing a parsimonious model by selecting the more relevant attributes qualifying households and childcare providers among the ones presented. Afterwards we provide a measure

of the direct as well as the interaction effects of the chosen attributes on the preferences of the agents representing the childcare market. According to the literature on childhood education the wage of the teachers as well as the pupil-teacher ratio appear to be of primary significance when evaluating the goodness of the childcare market. Indeed the teachers' wage has been recognized as one of the most important predictor of the quality of care, directly affecting the stability of the childcare provider (potentially inflating the turnover of the employees), which may have adverse consequences on the child development (Whitebook et al. (2014)). The pupil-teacher ratio has been extensively documented to have a profond impact on the child future achievements and socio-emotional behavior (Blatchford et al. (2011); Finn et al. (2003)). Therefore we decide to construct our benchmarck model by using these two observables as portraying the quality of care.

All the variables are standardized such that we measure the impact of a one standard deviation change in a variable on the welfare of households and providers, presented in euros.

We perform a grid search procedure with the intent to introduce the appropriate level of randomness to help rationalise the data. Moreover we retain the combination of sigmas ($\sigma_1 = 0.266, \sigma_2 = 0.5$) generating the largest value of the likelihood funtion while limiting in the price predictions the negative values for the out-of-sample observations. Our estimation fits the observed prices mildly well ($R^2 = 0.0127$). Table 5 summarizes the results of our estimation by presenting the direct and interacted effects of household and childcare provider features on the formation of their preferences. Concerning the households there is a clear unwilingness-to-pay for one standard deviation increase in the pupil-teacher ratio, which it is even more relevant for the wealthier households. Both these effects are significant at 1%. The direct effect of teachers' wage and its interaction with the income class of the households appear to be reasonable regarding their impact on the household preferences, suggesting a willingness-to-pay for one standard deviation increase in the teachers' wage. Nonetheless both these effects are not significant. Concerning the childcare providers there is a unambiguous unwillingness-to-accept children when the households' income increases by one standard deviation and this appears to be even more relevant for childcare providers exhibiting a pupil-teacher ratio which it is one standard deviation above the average. Both these effects are significant at 10%.

Our findings are in line with the literature which attributes to the increment of the pupil-teacher ratio a clear indication of the potential deterioration of the class environment and a decline of the effectiveness in promoting early learning (Bowne et al. (2017)). Indeed smaller pupil-teacher ratio yields to better pupil-teacher interactions and fruitfull cognitive stimulations (Phillipsen et al. (1997)). Furthermore the lower and higher income households evaluate differently the childcare provider characteristics, the former focusing on its pricing and accessibility while the

latter showing interest in the learning process and the school environment (Herbst et al. (2020)). Indeed higher income households are more proned to appraise quality features as pupil-teacher ratio, teachers' education and the childs' interactions with teachers (Gordon and Högnäs (2006); Peyton et al. (2001); Kensinger Rose and Elicker (2008)). Therefore the preference of the childcare provider unwilling to accept children from households owning higher income may be explained by the greater expectation and the potential pressure higher income households place on the quality of care supplied. The interpretation is confirmed by the fact that this preference is substantial when the pupil-teacher ratio rises, which it is a evident sign of the worsening quality of care.

6 Simulation

The goodness of our model hinges on its flexibility, which given the estimated parameters it allows to produce counterfactual matching by proposing diverse policy either acting on the quality standards or introducing subsidies based on quality. In our case we present an alternative childcare policy which raises the quality standard by imposing a unique value to the teachers' hourly wage. Therefore we equalize the teachers' hourly wage across the childcare providers to the highest wage in the sample (€55.93). Hence, we are implicitly costraining the households to choose the childcare providers based "only" on the pupil-teacher ratio feature.

To ease the interpretation and the presentation of the results achieved with the simulation we grouped the pupil-teacher ratios in 10 categories, each one containing a specific number of these ratios (see table 6). The last categories have more ratios than the first ones since we have little or none observations for some of the highest values of the pupil-teacher ratio (see table 7). Furthermore we keep the categorical variable associated with income class of the households as it is presented in table 3.

Figure 3 presents the variation of the number of couples by income classes and pupil-teacher ratio categories, the variation is produced by the different matching frequencies retrieved from the equilibrium and counterfactual matching distribution. Unsurprisingly, most of the variation occurs for the pairs where the households belong to the 1 and 7 income classes and the childcare providers to the first three categories of the pupil-teacher ratio (for those pairs the number of observations is remarkably high, see table 7). In accordance with the estimated preferences, the figure reveals a positive variation of couples for low values of the pupil-teacher ratio, and this is particularly significant for the wealthier households.

Furthermore we compare the welfare achieved by the households through the equilibrium matching with the one produced through the counterfactual matching (see figure 4). This

implies to assess the difference in the alpha (α) generated respectively under the equilibrium and conterfactual matching. Unsurprisingly, the households income and pupil-teacher ratio category experiencing a increment in the number of couples are concurrenty experiencing a rise in the welfare. This behavior has a straightforward interpretation: since the teachers' wage is uniformly raised to the highest standard, the households have only one possibility to improve their welfare given their preferences, which is to pair with providers displaying low values of the pupil-teacher ratio. Naturally we cannot expect massive variation of the welfare due to feeble value of the parameters, and indeed the positive and negative variations are of the order of few cents.

However irrespective of the magnitude of the welfares' variation the simulation outlines perfectly the behavior of the households, which given their preferences would pair with care providers presenting tiny pupil-teacher ratio values.

7 Conclusion

This paper has provided a first structural analysis involving both the supply and demand sides of the childcare market. We use a unifying matching-hedonic model, which was developed and tested for Luxembourg.

In line with the literature our findings suggest that the welfare of the households would decrease when the pupil-teacher ratio rises and the effect is substantially pronounced for high income households. Furthermore the welfare of the childcare providers decreases when the household income class increases, and this appears to be particularly relevant if the childcare providers exhibit a increment in the pupil-teacher ratio. This effect has simple explanation: the higher income households are likely to pay attention to features affecting the learning environment and pupil-teacher interactions, and therefore it may potentially rise the pressure on the childcare providers concerning the expected quality to be delivered.

Afterwards we exploit the flexibility of the model to simulate a counterfactual matching. Indeed by equalizing the teachers' wage across the childcare providers we force the households to choose based "only" on pupil-teacher ratio feature. In line with their preferences the redistribution of matchings indicates that the households privilege childcare providers with low pupil-teacher ratio, which it is followed by an increase of their welfare. Unsurprisingly, this effect is noticeably prominent for higher income households.

It is worth to note that the results obtained here should be interpreted as preliminary and non conclusive concernig the preferences of households and childcare providers for the Luxembourg childcare market. Therefore a more exhaustive grid search procedure needs to be conducted in

order to retrieve the optimal level of randomness to be included in the model, which in turn may provide a more robust prediction of the prices.

8 Appendix

The childcare system in Luxembourg

- Paid and compulsory maternity leave (16 weeks)
- Paid parental leave, mainly taken by mothers (the take-up rate among eligible fathers is less than 10%)
- Universal child care system, with mixed provision of childcare
- Highly subsidized public childcare, but rationing
- Places available in the private sector but at higher prices
- Introduction of a child care voucher in 2009 to improve access to affordable childcare
- All children in daycare are eligible
- For the parents, the voucher equalizes the price between the public and the for-profit sector
- the actual price paid by the parents depends on household income and the number of children in the household
- For-profit providers: (still) free to set their own prices but there is a cap to limit the government subsidies
- Substantial price reduction with the voucher:
  - Before the reform:
    - 2.5 euros/hour in the public
    - 4.9 euros/hour in the for-profit sector
  - After the reform: 1.4 euros/hour
- Overall daycare attendance have increased following the introduction of the voucher
- The number of slots in for-profit have been multiplied by 4 in 10 years
- The size of the queue has been dramatically reduced
- Now, after having reduced the rationing problem, the government aims at increasing the quality of the care.

Bibliography

Bartik, T. J. (1987). The estimation of demand parameters in hedonic price models. *Journal of Political Economy*, 95:81–88.

Bayer, P. J. and McMillan, R. (2010). Tiebout sorting and neighborhood stratification. Working Papers 10-48, Duke University, Department of Economics.

Blatchford, P., Bassett, P., and Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and instruction*, 21(6):715–730.

Blau, D. and Currie, J. (2010). *Pre-School, Day Care, and After-School Care: Who's Minding the Kids?*, volume 2 of *Handbook of the Economics of Education*, chapter 20, pages 1163–1278. Elsevier.

Blau, D. M. and Hagy, A. P. (1998). The Demand for Quality in Child Care. *Journal of Political Economy*, 106(1):104–146.

Bowne, J. B., Magnuson, K. A., Schindler, H. S., Duncan, G. J., and Yoshikawa, H. (2017). A meta-analysis of class sizes and ratios in early childhood education programs: Are thresholds of quality associated with greater impacts on cognitive, achievement, and socioemotional outcomes? *Educational Evaluation and Policy Analysis*, 39(3):407–428.

Choo, E. and Siow, A. (2006). Who Marries Whom and Why. *Journal of Political Economy*, 114(1):175–201.

Compton, J. and Pollak, R. A. (2014). Family proximity, childcare, and women's labor force attachment. *Journal of Urban Economics*, 79(C):72–90.

Dupuy, Galichon, Jaffe, and Kominers (2017). On the incidence of taxation in matching markets. Technical report, working paper.

Dupuy, A. (2018). Migration in china: to work or to wed? Technical report, Mimeo.

Dupuy, A. and Galichon, A. (2014). Personality traits and the marriage market. *Journal of Political Economy*, 122:1271–1319.

Epple, D. (1987). Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products. *The Journal of Political Economy*, 95:59–80.

Finn, J. D., Pannozzo, G. M., and Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research*, 73(3):321–368.

Galichon, Kominers, and Weber (2017). Costly concessions: An empirical framework for matching with imperfectly transferable utility. Technical report, Working paper.

Gordon, R. A. and Högnäs, R. S. (2006). The best laid plans: Expectations, preferences, and stability of child-care arrangements. *Journal of Marriage and Family*, 68(2):373–393.

Hagy, A. P. (1998). The Demand for Child Care Quality: An Hedonic Price Theory Approach. *Journal of Human Resources*, 33(3):683–710.

Heckman, J., Pinto, R., and Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Herbst, C. M. and Barnow, B. S. (2008). Close to home: A simultaneous equations model of the relationship between child care accessibility and female labor force participation. *Journal of Family and Economic Issues*, 29(1):128–151.

Herbst, C. M., Desouza, K. C., Al-Ashri, S., Kandala, S. S., Khullar, M., and Bajaj, V. (2020). What do parents value in a child care provider? evidence from yelp consumer reviews. *Early Childhood Research Quarterly*, 51:288–306.

Kahn, S. and Lang, K. (1988). Efficient estimation of structural hedonic systems. *International Economic Review*, 29:157–166.

Kalb, G. (2009). Children, Labour Supply and Child Care: Challenges for Empirical Analysis. *Australian Economic Review*, 42(3):276–299.

Kensinger Rose, K. and Elicker, J. (2008). Parental decision making about child care. *Journal of Family Issues*, 29(9):1161–1184.

Owens, M. F. and Rennhoff, A. D. (2014). Provision and price of child care services: For-profits and nonprofits. *Journal of Urban Economics*, 84:40–51.

Peyton, V., Jacobs, A., O'Brien, M., and Roy, C. (2001). Reasons for choosing child care: Associations with family factors, quality, and satisfaction. *Early childhood research quarterly*, 16(2):191–208.

Phillipsen, L. C., Burchinal, M. R., Howes, C., and Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early childhood research quarterly*, 12(3):281–303.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1):34–55.

Whitebook, M., Phillips, D., and Howes, C. (2014). Worthy work, still unlivable wages: The early childhood workforce 25 years after the national child care staffing study.

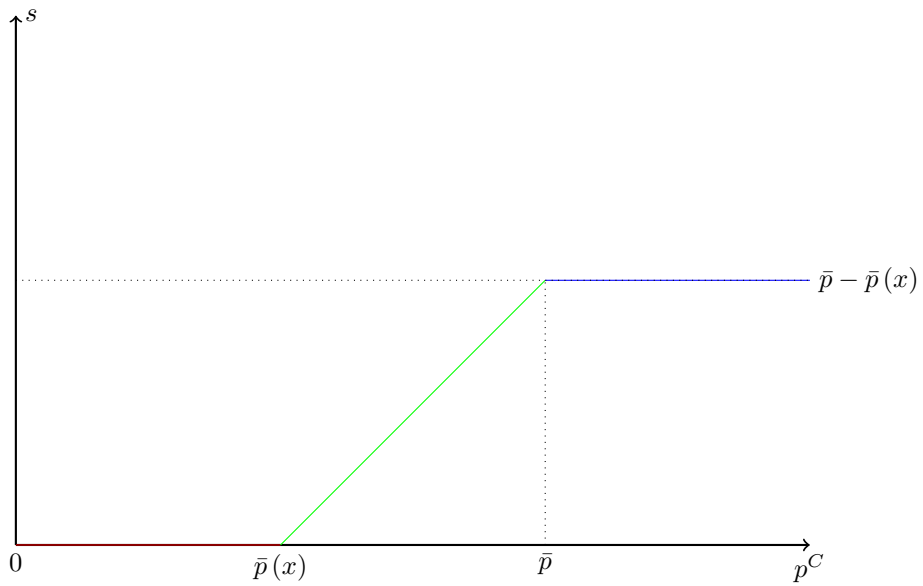


Figure 1: presents the government subsidy to households using private providers in Luxembourg.

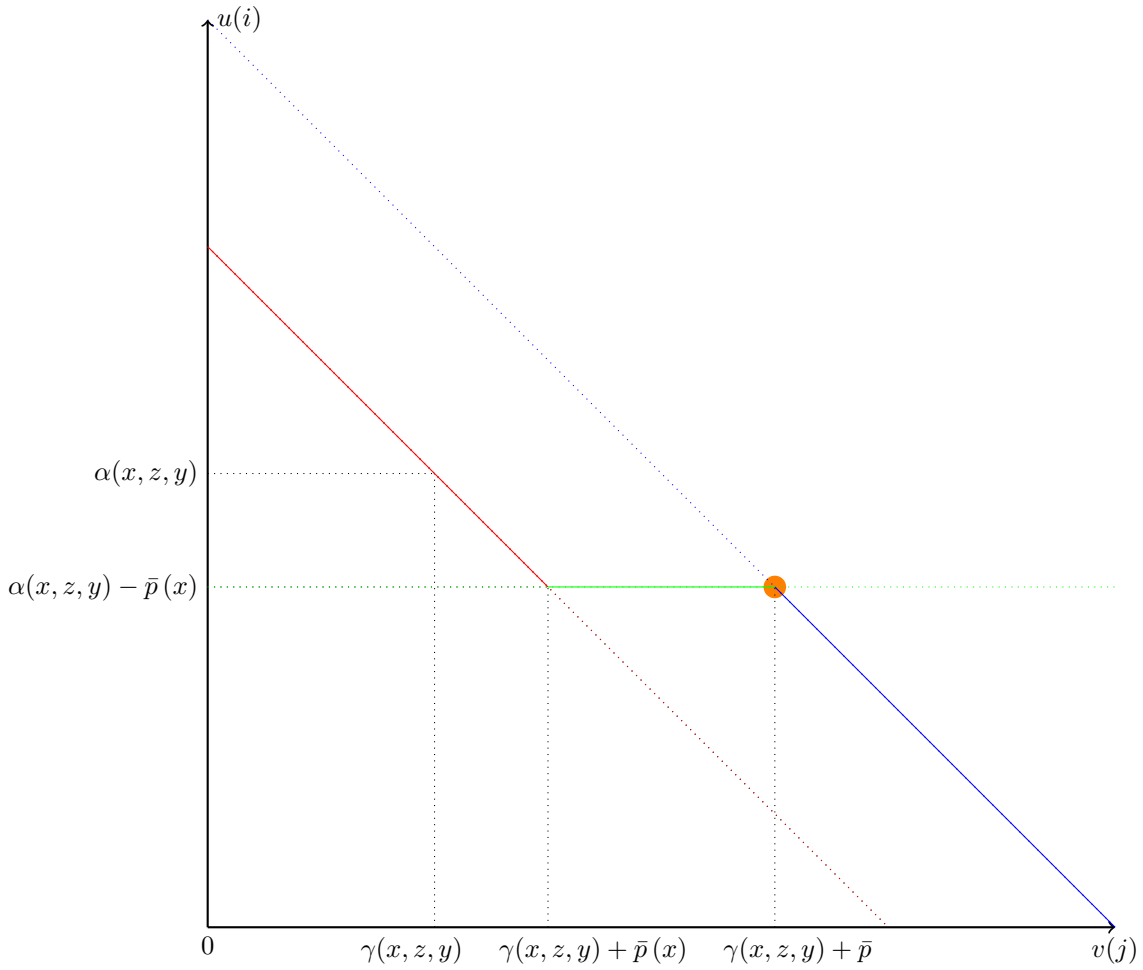


Figure 2: depicts the Pareto frontier in Luxembourg.

Table 1: presents the childcare voucher scheme

	Hourly child care prices with regard to the number of hours:			
Children:	**Free**	**max. 3 euros**	**max. 7.5 euros**	**7.5 euros and +**
With the minimum income /at risk of social exclusion	0-25 hours	26-60 hours	61 hours and +	
Other child	0-3 hours	4-24 hours	25-60 hours	61 hours and +

Source: Ministry of Education, Luxembourg

Table 2: presents the parents' financial contribution by household income in 2016

	Household income class	Child's birth order	Childcare voucher price	Socio-family price	Full price
0	minimum income or	1	0.50	-	7.50
	at risk of social exclusion	2	0.30	-	7.50
		3	0.15	-	7.50
		4 and +	FOC	-	7.50
1	<1.5 MW	1	0.50	0.50	7.50
		2	0.30	0.30	7.50
		3	0.15	0.15	7.50
		4 and +	FOC	FOC	7.50
2	< 2 MW	1	1.00	1.50	7.50
		2	0.70	1.10	7.50
		3	0.35	0.55	7.50
		4 and +	FOC	FOC	7.50
3	< 2.5 MW	1	1.50	2.50	7.50
		2	1.10	1.80	7.50
		3	0.55	0.90	7.50
		4 and +	FOC	FOC	7.50
4	< 3 MW	1	2.00	3.50	7.50
		2	1.50	2.60	7.50
		3	0.75	1.30	7.50
		4 and +	FOC	FOC	7.50
5	<3.5 MW	1	2.50	4.50	7.50
		2	1.80	3.30	7.50
		3	0.90	1.65	7.50
		4 and +	FOC	FOC	7.50
6	<4 MW	1	3.00	5.50	7.50
		2	2.20	4.10	7.50
		3	1.10	2.05	7.50
		4 and +	FOC	FOC	7.50
7	<4.5 MW	1	3.00	6.50	7.50
		2	2.20	4.80	7.50
		3	1.10	2.40	7.50
		4 and +	FOC	FOC	7.50
8	$\geq$ 4.5 MW or no data on household income	1	3.00	7.50	7.50
		2	2.20	5.60	7.50
		3	1.10	2.80	7.50
		4 and +	FOC	FOC	7.50

MW: Minimum Wage-1922.96 euros/month, FOC: free of charge
Source: Ministry of Education, Luxembourg

Table 3: presents the household income classes

Income class	Lower bound	Upper bound
0	0	1 MW
1	1 MW	1.5 MW
2	1.5 MW	2 MW
3	2 MW	2.5 MW
4	2.5 MW S	3 MW
6	3.5 MW	4 MW
7	4 MW	4.5 MW
8	4.5 MW	

MW: Minimum Wage-1 922.96 euros/month

Table 4: presents the summary statistics

	Mean	S.d.	Min	Max
Household				
income (class 0-7)	2.73	2.28	0	7
Childcare provider				
types (binary 0 or 1)	0.17	0.38	0	1
size (# children)	36.79	44.70	2	267
pupils-teacher ratio (pupils/teachers)	2.32	1.66	0.005	15.5
seniority (av. seniority/ years of business)	0.34	0.14	0.046	0.98
contract (share of permanent contract)	0.91	0.09	0.33	1
male ratio (male/female)	0.08	0.06	0	0.33
employee (share of native employee)	0.32	0.25	0	1
wage (hourly wage in €)	17.64	4.60	12.08	55.93

Table 5: reports the main and interaction effects of household and childcare features on the formation of their preferences. The covariates are standardized. In parentheses the standard errors are presented as computed from the Fisher information.

	gamma (main effect)	pupil-teacher ratio	wage
alpha (main effect)		-0.00505	0.00025689
		(0.00063761)	(0.00069928)
income		-0.023548	0.0039447
		(0.0024771)	(0.0024827)
income	-0.0084279	-0.015106	-0.0039359
	(0.0043695)	(0.0085111)	(0.0073108)

Table 6: presents the pupil-teacher ratio categories

pupil-teacher ratio category	pupil-teacher ratio
1	$\cdot \leq$ 1:1
2	1:1 $< \cdot \leq$ 2:1
3	2:1 $< \cdot \leq$ 3:1
4	3:1 $< \cdot \leq$ 4:1
5	4:1 $< \cdot \leq$ 5:1
6	5:1 $< \cdot \leq$ 6:1
7	6:1 $< \cdot \leq$ 7:1
8	7:1 $< \cdot \leq$ 8:1
9	8:1 $< \cdot \leq$ 11:1
10	11:1 $< \cdot \leq$ 16:1

Table 7: reports the number of observations by income class and pupil-teacher ratio category

income class	pupil-teacher ratio category									
	1	2	3	4	5	6	7	8	9	10
0	36	44	41	38	16	1	2	3	0	0
1	352	408	386	320	207	26	26	49	0	2
2	106	133	107	70	49	7	5	10	0	1
3	92	77	58	58	44	4	1	1	1	1
4	76	56	49	47	20	2	3	0	0	1
5	67	63	40	32	13	3	3	2	0	0
6	42	34	26	21	15	7	2	0	1	0
7	218	118	126	71	53	8	3	0	0	0

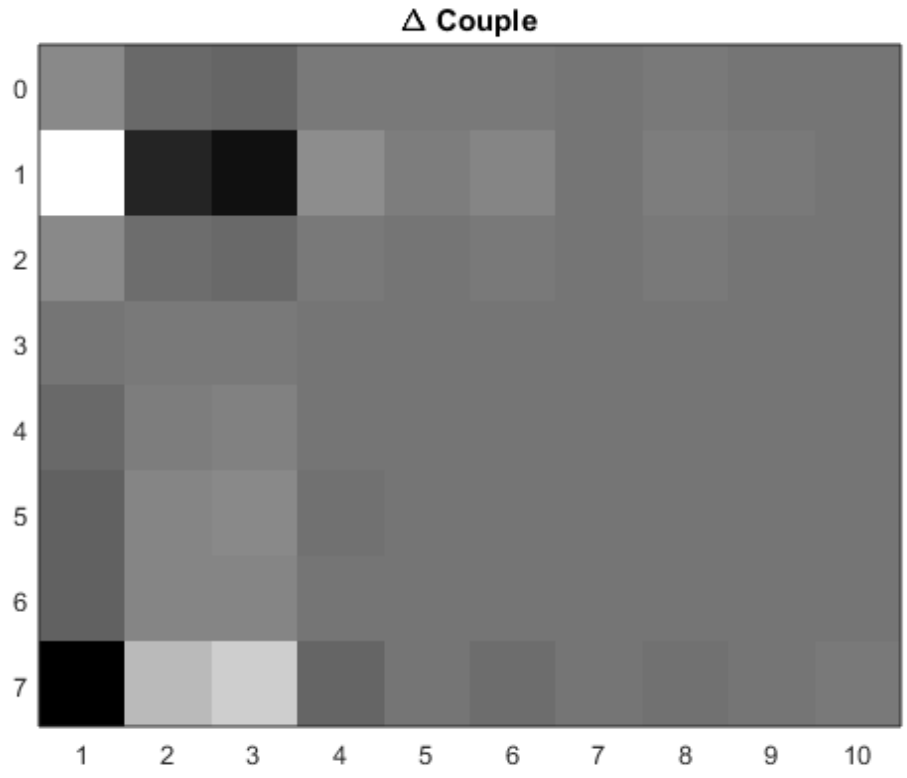


Figure 3: depicts the variation of the number of couples produced by the difference of the matching frequencies expressed by the equilibrium and the counterfactual matching distribution. On the y-axis we have the income classes of the households and on the x-axis we have the pupil-teacher ratio categories. The darker blocks indicate that for a given combination of income class and pupil-teacher ratio category the number of couples in the counterfactual is larger than the one obtained in the equilibrium.

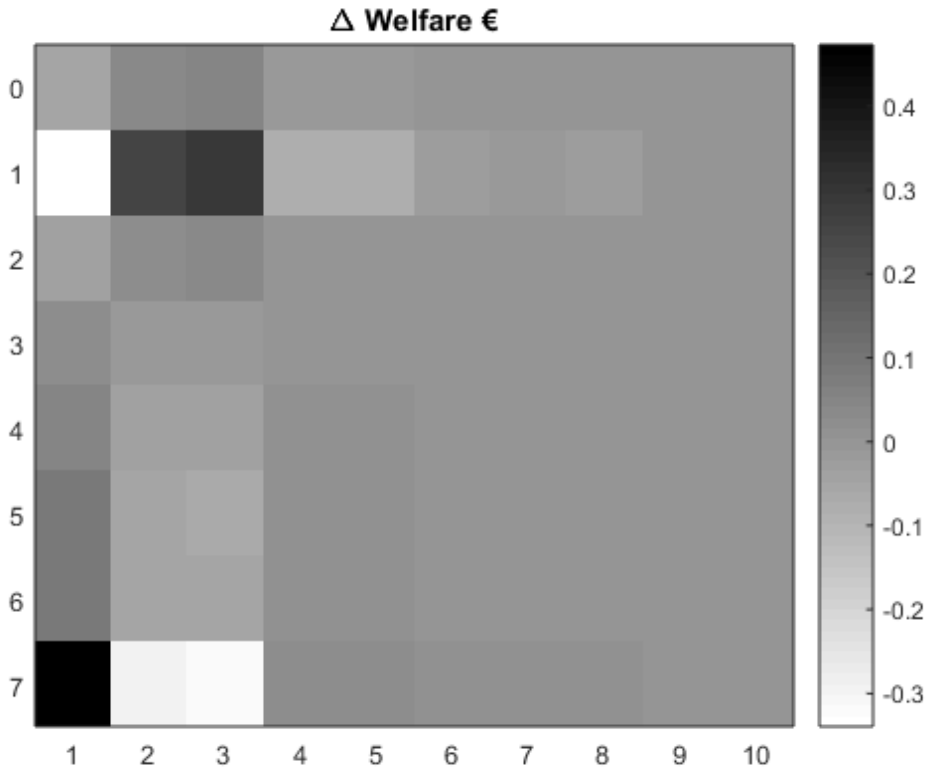


Figure 4: depicts the variation of the welfare produced by the difference of the α_s generated under respectively the equilibrium and the counterfactual matching distribution. On the y-axis we have the income classes of the households and on the x-axis we have the pupil-teacher ratio categories. The darker blocks indicate that for a given combination of income class and pupil-teacher ratio category the welfare produced by the counterfactual is larger than the one obtained from the equilibrium.

Chapter 2

Differential taxation across matching markets

1 Introduction

Recently, a growing attention has been devoted to cases where a certain degree of friction may arise when the transfer of utility occurs between agents. This represents an intermediate possibility between the two well-known extreme cases: the perfect transferable utility and the nontransferable utility framework. It introduces the Imperfect Transferable Utility (ITU) setting extensively studied by Galichon et al. (2015), Galichon et al. (2017) and Dupuy et al. (2017). In the labor market such nonlinearities produced by taxation seem to be particularly relevant where the employee does not fully receive the gross wage paid by the employer.

In our model we define the agents based on observable attributes and we allow each agent to have heterogeneous preferences and to make transfer payments to their partners. Therefore, we adopt the approach of Choo and Siow (2006), presenting the utility function of the agent featured by a random part associated with its idiosyncratic taste over the potential partners. This structure enables to identify the match values from data on matching patterns and wages.

We apply this setting to the US labor market, estimating job amenity and labor productivity from the worker-job pairs, implicitly making the assumption that the firm features refer to the job features as proposed in Wilemme (2017). We then investigate the effect of taxation on the jobs mismatch which we define as the discrepancy between the level of education of the worker and the job qualification requirements (Berlingieri and Erdsiek (2012); Stoevska (2017))). The intuition for undertaking this analysis is the following: the taxation may reduce the capacity of the firm to pay for the disutility of the worker, leaving the worker with little incentive to work on that job. Ultimately, the worker may switch seeking occupations enhancing its job's satisfaction rather than occupation where it is more productive. The consequence is a matching distortion where the taxation may play a crucial role preventing the efficient matching, agents do not pair with partners offering higher transfers but perhaps preferring matching with partners providing the desired level of amenities.

Our approach is related to the literature on the occupational choices, with a substantial difference relative to the prior literature where the matching distortion is associated to one agent of the market (Lockwood et al. (2017); Parker (2003); Sheshinski (2003); Powell and Shan (2012)). Instead our approach models the two side of the market providing a more exhaustive picture of the inefficient matching.

Moreover introducing the heterogeneity in preferences for both the agents forming the labor market enables the firm to capture some productivity surplus which it does not need to be compensated by a rise in wage. This is a clear element of difference from the Roy models which assume that the workers earn their marginal product (Rothschild and Scheuer (2013); Boadway

et al. (1991)).

Our approach is also related to the literature investigating the jobs mismatch. Firstly, we identify the jobs mismatch based on qualifications, assuming that the mismatch produced by qualification is a good proxy for measuring the gap in skills owned by the workers and required by the jobs (Quintini (2011)). Therefore we implicitly neglect the possibility that discrepancies in skills unrelated to education may compensate for differences in terms of qualifications (Hartog (2000)). Indeed we follow the literature identifying over and undereducated individuals by assessing the worker's education and the education required by the job (Duncan and Hoffman (1981)), relying on the evaluation produced by a professional job analyst and supplied by the Occupational Information Network (O*NET) (Thurow and Lucas (1972); Hartog (1980)). Secondly, we explicitly introduce heterogeneity in workers and jobs characteristics (Sattinger (1993)), and in contrast with standard neoclassical theory (generally relying on the wage and hours worked) we emphasize the relevance of non-monetary features of the job as crucially impacting the worker's decision to match (Brown (1980); Hwang et al. (1992)). Thirdly, differently from the literature on compensating wage differentials suggesting that a worker may accept a higher wage as compensation for undesirable non-monetary features, we oppose the role of the taxation able to reduce the firm's ability to offset the disutility of the worker.

Our estimation strategy hinges on the Maximum Likelihood estimation of job amenities and productivities proposed by Dupuy et al. (2017) when the matching is multidimensional, transfers are observed with noise and the heterogeneous preferences of the agents are logit.

The remainder of the paper is organized as follows: Section 2 describes the economic model with the introduction of the worker's and firm's problem; Section 3 introduces the estimation strategy; Section 4 presents the computational procedure to harmonize the US taxation, Section 5 presents the dataset; Section 6 discloses the empirical results; Section 7 underlines the effect of taxation on the jobs mismatch; Section 8 concludes.

2 Economic model

In this section we develop an ad hoc matching model to study the preferences of the agents forming the US labor market, and we therefore need to deal with the heterogeneous US taxation system. The complexity of this type of taxation regime has been managed proposing in this section a theoretic framework that ensures the tractability and applicability of the matching model by introducing a unique sequence of tax brackets (this would embed the federal and state taxations) while preserving the peculiarity of state and federal taxations by presenting a composite

computation of their marginal tax rates in that brackets. The harmonization of the taxation system (here simply intended as the process to create a common structure of brackets) developed for the US states may be readily applied to other cases, i.e to investigate the European labor market. In this case, the US states may be replaced by European countries with no substantial difference in the dissertation of the matching model. This reinforces the generality of our theoretic procedure.

The remainder of the section presents the maximization problem of the firm and worker, deriving at the equilibrium the matching function and transfer.

2.1 Unique taxation system

Let's suppose that we have s states exhibiting progressive taxation labeled with the letter $s = 1, ..., S$ and presenting different tax intervals identified by the level of transfer. In particular we interpret the heterogeneity among the various fiscal regimes as a heterogeneity in the composition of the tax intervals. For sake of convenience, the federal taxation is treated as the taxation of a fictitious state s

We associate t_s^k with the different level of transfer where s is the subscript representing the state and $k = 1, ..., K^s$ is the superscript indexing the thresholds of that state.

Definition 1. *For a generic state s, the taxation system is identified as an increasing sequence of threshold t_s^k.*

In order to get an efficient and consistent mathematical proceeding we need to play around with the concepts of intersection and union between intervals. This provides the following lemma:

Lemma 1. *Given a set of taxation systems $t_{\{s=1,..,S\}}^{\{k=1,..,K^s\}}$, the all-encompassing combinations of S taxation systems is obtained through the following expression*

$$\bigcup_{s\prime=1}^{S} \bigcup_{s=1}^{S} (1 - \delta_{ss\prime}) \left\{ \bigcup_{k=1}^{K^s} \left[(t_s^k, t_s^{k+1}) \bigcap_{k\prime=1}^{K^{s\prime}} (t_{s\prime}^{k\prime}, t_{s\prime}^{k\prime+1}) \right] \right\} \tag{1}$$

where $1 - \delta_{ss\prime}$ reads as

$$\delta_{ss\prime} = \begin{cases} 1, & \text{if } s = s\prime, \\ 0, & \text{if } s \neq s\prime. \end{cases}$$

We defer the curious reader to appendix A.1 for the details of the proof.

The expression (1) ensures that the overall structure of the intervals is obtained by merging the fiscal regimes of each state s. Indeed we have at our disposal all the possible combinations

expressing the intersections of the tax brackets in pairs between states. It is important to stress that for harmonizing the number of thresholds we need to arrange the findings of the expression (1), assuming that $t_1^1 = t_2^1 = ... = t_S^1 = 0$.

Algorithm 1 Creation of a unique sequence of thresholds

Step 1 $\rightarrow$ From (1) we gather all the combinations of the intersections of tax brackets expressed in pairs between states

Step 2 $\rightarrow$ we build the first interval using the following expression

$$\min((t_1^1, t_1^2), (t_2^1, t_2^2), ..., (t_S^1, t_S^2)) = (t_1^1, t_s^2)$$

Step 3 $\rightarrow$ we obtain the second intervals minimizing over $k\prime$ all the intersections retrieved from (1) having t_s^2 as first extreme

$$\min_{k\prime \in \{2,...,K^{s\prime}+1\}}((t_s^2, t_1^{k\prime}), (t_s^2, t_2^{k\prime}), ..., (t_s^2, t_S^{k\prime})) = (t_s^2, t_{s\prime}^{k\prime})$$

Step 4 $\rightarrow$ we update the t_s^2 with $t_{s\prime}^{k\prime}$ and repeat the step 3 finding out the minimum among all the intersections retrieved from (1) having $t_{s\prime}^{k\prime}$ as first extreme

Step 5 $\rightarrow$ we repeat step 3 and 4 until we reach the largest threshold.

For sake of tractability we assume from now on a common structure of tax brackets independently of the state s with a clear simplification concerning the dissertation of the economic model.

2.2 Set up

Consider a one-to-one matching with imperfect transferable utility. Let's assume that a worker's characteristics be identified by a vector of attributes $x \in \boldsymbol{X} = \mathbb{R}^{d_x}$ and the firm's characteristics be captured by a vector of attributes $y \in \boldsymbol{Y} = \mathbb{R}^{d_y}$. We establish a matching model assuming multidimensional types of worker and firm within a context of friction inferred from taxes. We encompass among the attributes of the worker and firm, respectively its birth state and its location.

We write down the model assuming a progressive wage schedule. Specifically, we refer to Dupuy et al. (2017) for the equilibrium matches and wages retrieved in the case of piecewise linear taxation coming from a single market.

Indeed we set the federal income tax rates which may vary by workers as $\phi^1_{fd}(x) < \phi^2_{fd}(x) <...$ $..< \phi^K_{fd}(x)$ where $\phi^1_{fd}(x)$ and $\phi^K_{fd}(x)$ are the lowest and the top federal tax rate (the subscript fd stands for federal). We then label the state income tax rates which may vary by firm as $\phi^1_{st}(y) < \phi^2_{st}(y) <.....< \phi^K_{st}(y)$ where $\phi^1_{st}(y)$ and $\phi^K_{st}(y)$ are the lowest and the top tax rate by state (the subscript st stands for state). The generics $\phi^k_{fd}(x)$ and $\phi^k_{st}(y)$ represent respectively the federal tax rate and the state tax rate above the threshold t^k where $k = 1, .., K$ identifies the unique sequence of thresholds with $t^1 = 0$. It is worth to note that we keep track of the differential magnitude regarding the tax rates of the federal taxation and state taxations, and more we let the location of the firm determining the state taxation, preserving the marginal tax rates of each state.

For the sake of completeness we embed the payroll tax denoted by $\tau(y)$. The payroll tax is essentially flat regardless of the income bracket k and may depend on the type of firm y.

2.3 Worker problem

Let the worker's characteristics be identified by a vector of observable attributes $x \in \boldsymbol{X} = \mathbb{R}^{d_x}$. Moreover we acknowledge the existence of an heterogeneous part of the utility function unknown to econometrician (expressing the unobserved preferences of the workers).

The workers maximize their utility deciding in which type of firm y they want to work. We define $\alpha(x, y)$ the utility derived by a worker i of observable attributes $x_i = x$ choosing a firm of type $y \in \boldsymbol{Y}$. Furthermore we model the unobserved heterogeneity by introducing the term $\epsilon(i, y)$ capturing the idiosyncratic part of the worker's utility when choosing a firm of type y.

Let's introduce $\alpha^k(x, y)$ the utility derived by worker of type x, earning gross wage t^k, choosing firm of type y

$$\alpha^k(x, y) = \alpha^{k-1}(x, y) + (1 - \phi^{k-1}_{fd}(x))(1 - \phi^{k-1}_{st}(y))(t^k - t^{k-1})$$

with $k = 1, ..., K$ and assuming $\alpha^1(x, y) = \alpha(x, y)$ the pre-transfer level of amenities derived by worker of type x when choosing firm of type y.

Then the utility of the worker i earning gross wage $w(x, y)$ can read as

$$\alpha^k(x, y) + (1 - \phi^k_{fd}(x))(1 - \phi^k_{st}(y))(w(x, y) - t^k) + \sigma^W \epsilon(i, y)$$

where σ^W is the scaling factor associated with the unobserved taste of worker i (the superscript W stands for worker).

The systematic part of the utility $U(x, y)$ can then be read as

$$U(x,y) = \min_{k \in \{1,\dots,K\}} (\alpha^k(x,y) + (1 - \phi^k_{fd}(x))(1 - \phi^k_{st}(y))(w(x,y) - t^k)$$

Hence, a worker i of type x maximizes the utility function

$$u(i) = \max_y \{U(x,y) + \sigma^W \epsilon(i,y)\} \tag{2}$$

2.4 Firm problem

Let's assume that firm's characteristics be identified by a vector of observable attributes $y \in \boldsymbol{Y} = \mathbb{R}^{d_y}$. More we acknowledge the existence of an heterogeneous part of the utility function (unknown to econometrician) expressing the unobservable preferences of the firms.

The firms maximize their utility deciding the type of worker x they want to hire. We define $\gamma(x, y)$ the utility derived by a firm j of observable attributes $y_j = y$ choosing worker of type $x \in \boldsymbol{X}$. Furthermore we model the unobserved heterogeneity by introducing the term $\eta(x, j)$ capturing the idiosyncratic part of the firm's utility when choosing worker of type x.

Hence, the firm j of type y chooses workers of type x enjoys utility

$$\gamma(x,y) - (1 + \tau(y))w(x,y) + \sigma^F \eta(x,j)$$

where σ^F is the scaling factor associated with the unobserved taste of firm j (the superscript F stands for worker).

The systematic utility of the firm is identified by the profit being the difference between the productivity of the worker of type x working for firm of type y and the cost of labor.

$$V(x,y) = \gamma(x,y) - (1 + \tau(y))w(x,y)$$

Hence firm j of type y maximizes the utility function

$$v(j) = \max_x \{V(x,y) + \sigma^F \eta(x,j)\} \tag{3}$$

2.5 Equilibrium

In our economy the worker maximizes the utility given in (2) and the firm maximizes the utility given in (3) so the expected utility of the worker, assuming a large market, is

$$G_x(U(x,y)) = E_{S_x}\left[\max_y\{U(x,y) + \sigma^W \epsilon(i,y)\}|x_i = x\right] \tag{4}$$

and the expected utility of the firm

$$H_y(V(x,y)) = E_{T_y}\left[\max_x\{V(x,y)) + \sigma^F \eta(x,j)\}|y_j = y\right] \tag{5}$$

S_x and T_y are the known distributions of the unobserved heterogeneity of workers and firms respectively.

From Daly-Zachary-Williams theorem[1] we know that for each couple (x,y) we have

$$\frac{\partial G_x(U(x,y))}{\partial U(x,y)} = \mu(y|x) = \frac{\mu(x,y)}{n(x)}$$
$$\frac{\partial H_y(V(x,y))}{\partial V(x,y)} = \mu(x|y) = \frac{\mu(x,y)}{m(y)}$$

where $\mu(y|x)$ represents the mass of workers type x demanding firm of type y while the $\mu(x|y)$ represents the mass of firms y demanding worker of type x.

Definition 2. *A pair (μ, w) consisting of a matching $\mu > 0$ and a transfer $w \in \mathbb{R}$ is an equilibrium outcome if and only if μ is*

- *feasible meaning that it satisfies the accounting constraints*

$$\begin{aligned}\sum_y \mu(x,y) &= m(x)\\ \sum_x \mu(x,y) &= n(y),\end{aligned}$$

- *the market clearing condition*

$$m(x)\frac{\partial G_x(U(x,y))}{\partial U(x,y)} = n(y)\frac{\partial H_y(V(x,y))}{\partial V(x,y)}$$

  ,

- *and it is solution to both problems* (4) *and* (5)

[1] The theorem states that the partial derivatives of the expected achieved utility is equal to the choice probabilities in an additive random utility model.

As underlined by Galichon and Salanié (2015) the large market assumption mitigates the concerns about the misrepresentation of the agent characteristics' [2]. The feasibility constraints are needed to ensure from the worker perspective that the mass of pairs firm-worker with workers of type x coincides with the mass of workers of type x and from the firm perspective that the mass of pairs firm-worker with firms of type y coincides with the mass of firms of type y. Finally, the market clearing condition satisfies the condition that at the equilibrium the mass of workers of type x demanding firms of type y coincides with the mass of firms of type y demanding workers of type x.

Assuming a Gumbel type I distribution for the idiosyncratic shocks, we know from McFadden et al. (1973) that the probability choice is logit

$$G_x(U(x,y)) = \sigma^W \log\left(\sum_y \exp \frac{U(x,y)}{\sigma^W}\right)$$

and,

$$\frac{\partial G_x(U(x,y))}{\partial U(x,y)} = \frac{\exp \frac{(U(x,y))}{\sigma^W}}{\sum_y \exp \frac{U(x,y)}{\sigma^W}}$$

According to Dupuy and Galichon (2015) we can then express the conditional probability of choosing firm y given a worker of type x as

$$\mu(y|x) = \exp \frac{U(x,y) - u(x)}{\sigma^W} \tag{6}$$

Similarly we can express the conditional probability of choosing worker x given a firm of type y

$$\mu(x|y) = \exp \frac{V(x,y) - v(y)}{\sigma^F} \tag{7}$$

we can now write the joint probabilities of (6) and (7) at the equilibrium respectively as

$$\mu(x,y) = m(x)\exp\left(\min_{k\in\{1,\dots,K\}} \frac{\alpha^k(x,y) + (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))(w(x,y)-t^k) - u(x)}{\sigma^W}\right) \tag{8}$$

$$\mu(x,y) = n(y)\exp\left(\frac{\gamma(x,y) - (1+\tau(y))w(x,y) - v(y)}{\sigma^F}\right) \tag{9}$$

Now manipulating the last expressions, we obtain the matching function (Galichon et al. (2017))

[2] In finite population there is always a profitable deviation which may complicate the analysis of the existence of a stable equilibrium. In practice we ensure a continuum of individuals of each type by assuming that the proportion of individuals of each type in finite population is consistent with the proportion of individuals in an infinite population (Chan et al. (2019))

which can read as

$$\mu(x,y) = \min_{k\in\{1,\dots,K\}} \mu^k(x,y) = \min_{k\in\{1,\dots,K\}} \Bigg[m(x) \frac{\xi^{W,k}(x)\sigma^W}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W} n(y) \frac{\lambda^{F,k}(y)\sigma^F}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W} \exp \frac{\xi^{W,k}(x)(\alpha^k(x,y) - u(x)) + \lambda^{F,k}(y)(\gamma^k(x,y) - v(y))}{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F} \Bigg] \quad (10)$$

where $\gamma^k(x,y) = \gamma(x,y) - (1+\tau(y))t^k$, $\xi^{W,k}(x) = \frac{1}{1-\phi^k_{fd}(x)}$ and $\lambda^{F,k}(y) = \frac{1-\phi^k_{st}(y)}{1+\tau(y)}$

The last two factors used to rescale the unobserved heterogeneity in the model express the heteroskedastic behavior of the distributions of the unobserved heterogeneity.

The minimization over $k = 1, ..., K$ is due to the lemma 2 [3] of Galichon et al. (2017). Furthermore from equation (8) and (9) we can also get the equilibrium transfer which can read as

$$w(x,y) = \min_{k\in\{1,\dots,K\}} \frac{1}{1-\phi^k_{st}(y)} \Bigg[\frac{\xi^{W,k}(x)\lambda^{F,k}(y)\sigma^W(\gamma^k(x,y) - v(y)) - \xi^{W,k}(x)\lambda^{F,k}(y)\sigma^F(\alpha^k(x,y) - u(x))}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W} + \frac{\xi^{W,k}(x)\lambda^{F,k}(y)\sigma^W\sigma^F \log\left(\frac{n(y)}{m(x)}\right)}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W} + (1-\phi^k_{st}(y))t^k \Bigg] \quad (11)$$

The details concerning the computations of the equilibrium matching and transfer are presented in appendix A.2.

3 Estimation strategy

Generally, we have access to a random sample of population of matches between firms and workers. For each match i we observe the attributes of the workers x_i, the attributes of the firms y_i and the transfer w_i. Therefore a match i identified by $\{(X_i, Y_i, W_i), i = 1, ..., N\}$ contains the observable characteristics of the worker X_i, the observable characteristics of the firm Y_i and the wage W_i.

We parametrize the function of worker's amenities α and the function of firm's productivity γ as follows:

$$\alpha(x,y;A) = x^T A_0 y + A_1^T y$$

[3] The lemma states that the equilibrium matching may be obtained minimizing the matching functions belonging to the different bargaining sets identified by $k = 1, ..., K$. Moreover the procedure is geometrically equivalent to intersect the linear frontiers.

$$\gamma(x, y; \Gamma) = x^T \Gamma_0 y + \Gamma_1^T x$$

the matrices of parameters A_0 and Γ_0 (*affinity matrix*) indicate the level (intensity) of complementarity or substitutability between the observables while A_1 and Γ_1 the vectors of parameters assess the direct effect of the observables.

The estimation strategy is based on Dupuy and Galichon (2015) using the Maximum Likelihood method which is able to capture the likelihood of the observed matches and the likelihood of the observed wages[4].

The log-likelihood of the observed matches is

$$\begin{aligned}
&\ln L_1(x_i, y_i; A, \Gamma, c) = \\
&\ln \left\{ \prod_{i \in I} \left[\min_k \left(\exp \frac{\xi^{W,k}(x_i)(\alpha^k(x_i, y_i; A) - u(x_i; A, \Gamma)) + \lambda^{F,k}(y_i)(\gamma^k(x_i, y_i; \Gamma) - v(y_i; A, \Gamma))}{\xi^{W,k}(x_i)\sigma^W + \lambda^{F,k}(y_i)\sigma^F} \right. \right. \right. \\
&\qquad \left. \left. \left. + c \, \frac{(\lambda^{F,k}(y_i) - \xi^{W,k}(x_i))}{\xi^{W,k}(x_i)\sigma^W + \lambda^{F,k}(y_i)\sigma^F} \right) \right] \right\}
\end{aligned}$$

where the term in square bracket is nothing but the equilibrium matching. Recasting the above expression we get

$$\begin{aligned}
&\ln \left\{ \prod_{i \in I} \exp \left[\frac{\xi^W(x_i)(\alpha(x_i, y_i; A) - u(x_i; A, \Gamma)) + \lambda^F(y_i)(\gamma(x_i, y_i; \Gamma) - v(y_i; A, \Gamma))}{\xi^W(x_i)\sigma^W + \lambda^F(y_i)\sigma^F} \right. \right. \\
&\qquad \left. \left. + c \, \frac{(\lambda^F(y_i) - \xi^W(x_i))}{\xi^W(x_i)\sigma^W + \lambda^F(y_i)\sigma^F} \right] \right\} = \\
&\qquad \sum_{i \in I} \left[\frac{\xi^W(x_i)(\alpha(x_i, y_i; A) - u(x_i; A, \Gamma)) + \lambda^F(y_i)(\gamma(x_i, y_i; \Gamma) - v(y_i; A, \Gamma))}{\xi^W(x_i)\sigma^W + \lambda^F(y_i)\sigma^F} \right. \\
&\qquad\qquad \left. + c \, \frac{(\lambda^F(y_i) - \xi^W(x_i))}{\xi^W(x_i)\sigma^W + \lambda^F(y_i)\sigma^F} \right]
\end{aligned}$$

where the last step has been done by applying the property of the logarithm (the logarithm of a product is the sum of the logarithms).

$u(x_i; A, \Gamma)$ and $v(y_i; A, \Gamma)$ are the expected indirect utilities of workers and firms obtained

[4] As recognized in Dupuy et al. (2017) we partially observe the labor market in the data, indeed we lack the observations identifying unemployed workers and inactive firms. Therefore we need to introduce a constant c which differently from Dupuy et al. (2017) it appears in the matching as well as in the wage equation given the heteroskedastic behavior of the agents' preferences

computing the sample analog of the accounting constraints

$$\sum_{j=1}^{N} \mu(x_i, y_j; A, \Gamma, c) = \frac{1}{N}$$
$$\sum_{i=1}^{N} \mu(x_i, y_j; A, \Gamma, c) = \frac{1}{N}$$

where the system consists of dependent nonlinear equations, therefore we can obtain a unique solution up to a normalization, i.e. $u(x_1) = u(x_1; A, \Gamma) = 0$.

We assume that the transfer observed is a noisy measure of the true unobserved transfer

$$W_i = W_i{}^* + \delta_i \tag{12}$$

where δ_i is the measurement error which follows a $\mathcal{N}(0, s^2)$ and W_i^* is assumed to be the equilibrium transfer retrieved from the model. Therefore, the log-likelihood of the observed wages is

$$\ln L_2(x_i, y_i; A, \Gamma, c, s^2) = -\sum_{i=1}^{N} \Big(W_i - w(x_i, y_i; A, \Gamma, c) \Big)^2 \frac{1}{2s^2} - \frac{N}{2} \ln s^2 \tag{13}$$

where $w(x_i, y_i; A, \Gamma, c)$ is the transfer identified by the model (delivered by equation (11)) with the additional part concerning the constant c

$$w(x_i, y_i; A, \Gamma, c) =$$
$$\min_{k \in \{1,\dots,K\}} \frac{1}{(1 - \phi_{st}^k(y_i))}$$
$$\Bigg[\frac{\xi^{W,k}(x_i)\lambda^{F,k}(y_i)\sigma^W(\gamma^k(x_i, y_i; \Gamma) - v(y_i; A, \Gamma)) - \xi^{W,k}(x_i)\lambda^{F,k}(y_i)\sigma^F(\alpha^k(x_i, y_i; A) - u(x_i; A, \Gamma))}{\lambda^{F,k}(y_i)\sigma^F + \xi^{W,k}(x_i)\sigma^W}$$
$$+ (1 - \phi_{st}^k(y_i))t^k + \frac{\xi^{W,k}(x_i)\lambda^{F,k}(y_i)\sigma^W\sigma^F \log\Big(\frac{n(y)}{m(x)}\Big)}{\lambda^{F,k}(y_i)\sigma^F + \xi^{W,k}(x_i)\sigma^W} + c\,\frac{\xi(x_i)^{W,k}\lambda^{F,k}(y_i)(\sigma^W + \sigma^F)}{\xi^{W,k}(x_i)\sigma^W + \lambda^{F,k}(y_i)\sigma^F} \Bigg]$$

4 Harmonisation procedure

In the section 2 we have introduced a theoretic framework to create a unique sequence of thresholds, starting from different state taxations s with $s = 1, ..., S$.

However applying the theoretic procedure introduced in the economic model section when dealing with a copious number of states may create a sizable sequence of thresholds that could make unfeasible the computation of the model.

The US labor market presents two levels of taxation on individual income: federal and state.

The federal taxation is unique and progressive. The state taxations could be progressive, flat or absent. According to Tax Foundation (2014) we have 34 states with progressive taxation and 17 presenting flat or none taxation.

Our idea is to define a practical procedure to encompass the two levels of US taxation, harmonizing at first the US states with progressive taxation through the cluster analysis (presented in section 4.1) and then determining the unique taxation by picking the one providing the greater closeness between the computed and observed marginal tax rates at state and federal level (presented in section 4.2). The thresholds of the unique taxation are derived by the appropriate combination of the federal taxation and the synthetic state taxation (the unique sequence of thresholds retrieved from cluster analysis).

4.1 Cluster analysis

Initially, we proceed by applying a hierarchical clustering to the 34 US states exhibiting a progressive taxation in 2014. The intent with this multilevel cluster algorithm is to generate an overview of the potential cluster solutions, based on alternative choices of grouping the US states.

The decision regarding how to group the US states might be highly subjective, so to remove or reduce this potential issue we exploit the information of the US Census concerning the marginal tax rates attributed to the median wage ($26,989) in 2014 for each US states, excluding solely the District of Columbia from the cluster analysis for which that information is missing. The US Census reference provides an external validity criteria to corroborate the differential grouping of the US states. Moreover, these marginal tax rates produce a natural division of the US states and the hierarchical clustering confirms this partition. In the parlance of cluster analysis the partition is confirmed by the dendrogram reporting the cophenetic distances, the latter providing information concerning the degree of dissimilarity at which two states or group of states are combined into a single cluster. According Uragun and Rajan (2013) we choose the dendrogram which reports the highest value of the cophenetic correlation, the latter expressing the closeness between the original pairwise distances between the US states and the cophenetic distances.

The hierarchical clustering highlights three distinctive group of states, corresponding respectively to the US states exhibiting low, medium and high marginal tax rates associated with the median wage in 2014 (figure 1).

The subsequent step requires the application of the k-means clustering to the three different group of states. This algorithm represents a partitioning method as the hierarchical clustering, but producing a single level clustering. K-means assigns deterministically each threshold to a specific cluster in order to form k mutually and exclusive clusters. In this phase of our procedure,

we lose the dimension of state treating separately each group of states as sequence of thresholds.

For each group of states, k-means method would proceed selecting uniformly among the thresholds the number of centroids (representing the arithmetic mean of the thresholds in a cluster) corresponding to the number of clusters imposed at the beginning of the procedure and then reassigning all at once the thresholds to the closest centroid with the necessary recalculation of the cluster centroids. In the attempt to further minimize the total sum of the within cluster distances, the procedure would reassign individually the thresholds to a different cluster whether the reassignment would reduce the within sum distance in this cluster and subsequently it would recalculate the cluster centroid. The method repeats this procedure until the cluster assignments do not change further.

One critical aspect of this procedure is associated with the choice of the optimal number of clusters. As mentioned above at the beginning of the procedure we dictate the number of clusters, making a subjective choice. Therefore, we need an impartial procedure to define the optimal number of clusters for each group of states. The construction of the silhouette index is the popular method used to identify the optimal number of clusters.

The silhouette method evaluates the intra-cluster and inter-cluster distances both converging in the definition of the silhouette index, which it is used to compare different cluster configurations. For each group of states, we follow the recommendations suggested in the literature on cluster analysis using ten different cluster configurations (starting from one cluster in which all the thresholds belong to the same cluster until reaching ten where the thresholds split among ten clusters).

Figure 2 shows that the cluster configuration reporting the highest silhouette index is identified as the optimal cluster configuration for each group of states (more details of the silhouette method are available in the appendix B.1 while an alternative method based on Bayesian statistics for deriving the optimal cluster configuration for each group of states is proposed in appendix B.2).

In conclusion grouping the optimal cluster configurations produced by the three group of states provides a unique state taxation presenting eight thresholds. At this number of thresholds we need to add the zero as starting threshold that it was excluded from the cluster procedure as it represents a common element among the 33 US state taxations.

4.2 Percent error algorithm

We end the harmonization procedure section by explaining in detail the last algorithm we developed to generate the conclusive and unique sequence of thresholds. This application is motivated by the fact that the federal taxation and the synthetic state taxation (resulting from

the cluster analysis) present cumulatively 16 thresholds. The computational cost of running a model with that taxation system would be too expensive. This requires a further reduction of the size of the taxation system. Moreover we introduce the following algorithm in a general way and then return to our case.

For the generic state s, let's define $\phi_s^1, ..., \phi_s^K$ as the marginal tax rates, and $w_s^1, ..., w_s^K$ as the weights of the state s owning $t_s^1, ..., t_s^K$ number of thresholds. The weights are computed as the ratio of the mass of observations in each interval to the total mass of observations.

Then, for a generic interval $[t_s^k, t_s^{k+1}]$, the weight w_s^k is defined as:

$$w_s^k = \frac{m_s^k}{\sum_{i=1}^K m_s^i} \tag{14}$$

where m_s^k is the mass of observations in $[t_s^k, t_s^{k+1}]$ and $\sum_{i=1}^K m_s^i$ is the total mass of observations. The core formulas of the algorithm are the following:

$$\overline{\phi}_s^k = \frac{w_s^k \phi_s^k + w_s^{k+1} \phi_s^{k+1}}{w_s^k + w_s^{k+1}}$$

$$\Delta_s^k = \left[100\left(\frac{|\overline{\phi}_s^k - \phi_s^k|}{\phi_s^k}\right)\right] w_s^k + \left[100\left(\frac{|\overline{\phi}_s^k - \phi_s^{k+1}|}{\phi_s^{k+1}}\right)\right] w_s^{k+1}$$

Hence, $\overline{\phi}_s^k$ represents the average marginal tax rate provided by the generic intervals $[t_s^k, t_s^{k+1}]$ and $[t_s^{k+1}, t_s^{k+2}]$ while Δ_s^k represents the percent error you would commit by replacing the tax rates in the intervals $[t_s^k, t_s^{k+1}]$ and $[t_s^{k+1}, t_s^{k+2}]$ with the average tax rate $(\overline{\phi}_s^k)$ in the interval $[t_s^k, t_s^{k+2}]$. These formulas need to be derived for all intervals of the state s, creating a percent error vector. A crucial step of the algorithm entails the computation of the minimum of the percent error vector and the subsequent comparison with a tolerance error. At this point, there are two cases:

- First case: the minimum of the percent error vector is smaller than a tolerance error. Hence, we can replace the original tax rates ϕ_s^k and ϕ_s^{k+1} respectively in the intervals $[t_s^k, t_s^{k+1}]$ and $[t_s^{k+1}, t_s^{k+2}]$ with the average tax rate $(\overline{\phi}_s^k)$. Further we replace the masses of observations m_s^k and m_s^{k+1} respectively in $[t_s^k, t_s^{k+1}]$ and $[t_s^{k+1}, t_s^{k+2}]$ by their sum, and we delete the threshold t_s^{k+1}. Then we recompute the average tax rates and the percent error vector using the updated taxation. Afterwards, we restart the process computing the minimum of the percent error vector with the subsequent comparison with a tolerance error.

- Second case: the minimum of the percent error vector is larger than a tolerance error. Hence, the algorithm stops

In our case we have at our disposal the federal taxation and the synthetic state taxation. For convenience, we treat the federal taxation and the synthetic state taxation as if they were two fictitious states ($s = 1, 2$) with their proper taxations. Moreover we separately apply the algorithm to both level of taxation and we merge the residual thresholds.

In our case we obtain a final taxation with nine thresholds (we defer the curious reader to appendix B.3 for the details concerning the choice of the tolerance error and the marginal tax rate assignment).

We provide a visual representation of the precision reached by the final taxation for the 34 US states and for the federal taxation respectively in figure 3 and figure 4. It is worth noting that the accuracy of our approximation in computing the marginal tax rate of a bracket is strictly related to the mass of observations in that bracket. The methodology clearly privileges brackets with higher density of observations.

5 Dataset

We adopt the supplement CPS survey for our application to the US labor market. The supplement CPS (ASEC) questionnaire is generally conducted in march of each year. The ASEC dataset provides rich and extensive information on work experience, income, noncash benefits and migration relative to the monthly CPS. It is worth noting that data on income and employment refer to the previous year. We choose for a specific reason the march 2015 to test our model. According to the IPUMS database, the march ASEC 2015 is the only dataset released by the US Census containing the information regarding the residence 5 years ago at least from 2009. This is used as proxy in order to supply the missing information concerning the birth state of the individuals.

We follow the vast literature using CPS for investigating the real wage and inequality trend in US economy to construct our subsample.

We refer to the definition of the full-time and full year worker provided by David et al. (2005) and Lemieux (2010). Indeed the full-year and full-time worker is the individual who worked respectively in the previous year at least 40 weeks and 35 hours per week.

Moreover, we select individuals with age between 16 and 64 years old defining for those the potential experience, adjusted using the assumption that an individual cannot begin work before the age of 16 (the legal age to work in US) and is always non-negative (Acemoglu and Autor (2011)). The potential experience has been measured using the classic formula subtracting to the age, the years of schooling and six, namely the age when the school starts in US. The years

of schooling have been drawn from the categorization of Park et al. (1994), previously used by David et al. (2005).

We then proceed to create a unique variable concerning the yearly earning. We cannot rely on the gross income provided by the Census because our analysis excludes the secondary source of labor and unearned money as rental income, interests, dividends and alimony. We construct our yearly earning picking only wage and salary from the primary source of labor identifying the longest job reported in the ASEC file through the Census occupation code. This is motivated by the lack of information regarding the occupation of the second source of labor. We select individuals whose longest job coincides with the unique job performed along the whole year in 2014.

Then, we proceed to replace the yearly earning of the observations presenting the internal censoring threshold ($1,099,999) with the average earning defined by the shape parameter of the Pareto distribution. In particular the shape parameter is computed by adopting the closed formulation supplied by Armour et al. (2016) allowing to exploit the information between the public threshold ($280,000) and the internal censoring point. Moreover the 2015 ASEC makes appropriate to borrow this procedure to compute the average earning for the observations presenting the internal censoring point compared to the procedure requiring the multiplication of the topcoded earning by a factor 1.4 or 1.5 (Autor et al. (2008); Juhn et al. (1993); Lemieux (2006))(further explanations are provided in appendix C.1).

Furthermore, our definition of the local labor market implies that we label individuals who do not move or who move between five and one years ago relative to the year of the survey but within the same county, in a different county but within the same state as nomovers and we finger as movers individuals who change US state during that period. We exclude individuals moving outside the US territory. Indeed we assume a broad definition of local labor market delimited by the state boundaries. This is motivated by the fact that unlikely the metropolitan statistical area (MSA) the state boundaries are stable over time and they would not suffer from misclassification as it may occur to commuting zones and public use microdata area (PUMA) concepts (Molloy et al. (2011)). Our identification strategy associates the birth state of individual with the state residence five years ago and the actual residence with the state residence one years ago.

Afterward, we use the Occupational Information Network (O*NET) to define the features of the job. The O*NET is an impressive database containing quantitatively descriptors of the occupational characteristics, produced by the United States Department of Labor. It has been designed to replace the Dictionary of Occupational Titles (DOT). The O*NET database allows to retrieve informations concerning cognitive, interpersonal, physical skill requirements as well

as working conditions with the direct answer by a large representation of workers to questions included in four different surveys (Skills, Abilities, Work activities and Work context). These four surveys investigate different domains concerning the characteristics of the job. According to Peterson et al. (2001), the Skills survey indicates the level of mastery needed to perform a task, the Abilities survey expresses the basic capacities for completing a wide range of task, the Work activities and Work context surveys indicate the core of job activities and behaviors for completing the major job functions, including the social-psychological and physical conditions under this work is performed. Each survey is a two-part item defined by the importance and the level of the item for each single job. In that context we adopt the procedure developed in Acemoglu and Autor (2011) to create the job characteristics, designed as composite measures of items from O*NET Skills, Abilities, Work activities and Work context surveys. As underlined by Sanders (2011), the importance and level of each item although potentially revealing interesting informations are in practice almost perfectly correlated. This explains the choice of Acemoglu and Autor (2011) to use only the importance as score for each item.

5.1 Worker variables

The educational attainment is measured as the highest level of education achieved by the individuals. The latter have the possibility to choose among 16 categories, ranging from less than the 1^{st} grade to the doctorate degree. We transform this categorical variable in a continuous variable expressing the years of schooling. According to Park et al. (1994) we associate the different level of education with the years of schooling in the following way:

- 3 years of schooling from the 1^{st} to the 4^{th} grade
- 7 years of schooling from the 5^{th} to the 8^{th} grade
- 9 years of schooling for the 9^{th} grade
- 10 years of schooling for the 10^{th} grade
- 11 years of schooling for the 11^{th} grade
- 12 years of schooling for high school with no diploma and high school graduate
- 13.5 years of schooling for some college but not degree
- 14 years of schooling for associate degree
- 16 years of schooling for bachelor degree

- 17.5 years of schooling for the master degree
- 18 years of schooling for the doctorate degree

The age of individuals is obtained by direct response of individuals to the question during the survey. Nonetheless we need to subtract one to the response reminding that our analysis refers to the previous calendar year. We exclude from the characteristics of the worker the potential experience and decide to use the years of schooling and age for the concrete issue of collinearity when including these three variables.

5.2 Firm variables

In Acemoglu and Autor (2011) we retrieve the procedure to create the variables of interest characterizing the jobs. These variables may include several items as shown below:

- Non-routine cognitive: Analytical

  Analyzing data/information

  Thinking creatively

  Interpreting information for others

- Non-routine cognitive: Interpersonal

  Establishing and maintaining personal relationships

  Guiding, directing and motivating subordinates

  Coaching/developing others

- Routine cognitive

  Importance of repeating the same tasks

  Importance of being exact or accurate

  Structured versus Unstructured work

- Routine manual

  Pace determined by speed of equipment

  Controlling machines and processes

  Spend time making repetitive motions

- Non-routine manual physical

  Operating vehicles, mechanized devices, or equipment

  Spend time using hands to handle, control or feel objects, tools or controls

  Manual dexterity

- Offshorability

  Face to face discussions

  Assisting and Caring for Others

  Performing for or Working Directly with the Public

  Inspecting Equipment, Structures, or Material

  Handling and Moving Objects

  Repairing and Maintaining Mechanical Equipment

  Repairing and Maintaining Electronic Equipment

5.3 Final sample

Our original sample is made by almost 15,000 observations. Therefore we decide to select a random subsample made of 1,000 observations to derive our estimations and bootstrap 500 times that subsample to construct robust standard errors. This choice hinges on the motivation that implementing our estimation strategy on the whole sample would be extremely time-consuming.

Table 1 presents the summary statistics of the subsample used in our analysis. The average worker has 38 years old and presents 13 years of schooling. The firm features have been standardized as resulting from the Acemoglu and Autor (2011) procedure. The average yearly revenue is 0.046M$ presenting a moderate dispersion of 0.034M$.

6 Empirical results

According to the estimation strategy presented in section 3 we proceed by estimating the preferences of workers and firms for the US labor market. We propose as relevant the model with the following specification: age, years of education and age squared concerning the attributes of the worker and non-routine cognitive analytical, routine cognitive and routine manual concerning the attributes of the firm (the above specification has been proposed by assessing the significance of the likelihood ratio estimator). We measure the direct effects as well as the interaction effects of the worker and firm attributes on the job amenity and labor productivity.

All the variables are standardized such that we measure the impact of a one standard deviation change in a variable on the preferences of workers and firms, presented in M$.

We perform a grid search procedure with the intent to introduce the appropriate level of randomness to help rationalise the data. Moreover we retain the combination of sigmas ($\sigma_1 = 0.01; \sigma_2 = 0.1$) generating the largest value of the likelihood function while limiting in the wage predictions the negative values for the out-of-sample observations. Our estimation fits the observed wages relatively well ($R^2 = 0.36$). Table 2 summarizes the results of our estimation by presenting the direct and the interacted effects of worker and firm attributes on the job amenity and productivity.

We found out that a one standard deviation increase in the non-routine cognitive analytical feature of the job decreases the willingness workers like working by 0.0079 M$ (significant at 1%). The same effect is amplified when it is complemented by a rise of one standard deviation of the worker's age (0.0079+0.0029=0.0108M$), which it is significant at 1%. Similar effects are retrieved when rising by one standard deviation the routine manual attribute. Indeed this would cause a decrement of the job's enjoyment by 0.0025 M$ (significant at 1%), which worsens when we account for one standard deviation increase in the years of education (0.0025+0.0039=0.0064M$)(significant at 1%). We interpret the disutility to work in jobs dominated by non-routine cognitive tasks by the higher mental effort those jobs would require, and the advancement of the age would contribute to exacerbate the effort. Indeed the mental strain demanded by the execution of non-routine cognitive tasks may pose a concrete threat to the well-being of the individuals, which may engage in a stressful mental process to keep pace (Converso et al. (2018)). This can be detrimental for the job's satisfaction particularly for older workers (Meyer and Hünefeld (2018)). Similarly the physical effort imposed by performing routine manual jobs would affect the job's satisfaction of the workers, which deteriorates even more when workers are better educated.

Unsurprisingly, one standard deviation increase in education would produce a job productivity enhancement of 0.010 M$ which rises when combined with a one standard deviation increase in non-routine cognitive attribute by reaching a overall effect of 0.033 M$, both significant at 1%. Conversely a one standard deviation increment in routine manual task would produce a sharp decline in productivity when paired with worker exhibiting years of education one standard deviation above the average (0.010-0.0197=-0.0097M$)(significant at 1%). These effects are explained by the fact that additional years of education would certainly contribute to the increment of the productivity, particularly in jobs requiring creative and critical thinking. The results are in line with the human capital theory from Becker (1964) on, which advocates that education facilitates the development of skills that make workers more productive.

Lastly, we evaluate the effect of age and its square on productivity. Therefore we provide the impact of age on productivity by manipulating the following expression:

$$\theta\left(\frac{age-\mu}{\sigma}\right)+\nu\left(\frac{age-\mu}{\sigma}\right)^2$$

where μ and σ are respectively the mean and standard deviation of the age variable, θ and ν are respectively the main effect of the age and its square on productivity as obtained in table 2. Then the effect on γ would be

$$\gamma=\frac{\theta}{\sigma}age-\frac{\theta\mu}{\sigma}+\frac{\nu}{\sigma^2}age^2-\frac{2\nu\mu}{\sigma^2}age+\frac{\nu\mu^2}{\sigma^2}$$

then rearranging the expression

$$\gamma=\left(\frac{\theta}{\sigma}-\frac{2\nu\mu}{\sigma^2}\right)age+\frac{\nu}{\sigma^2}age^2-\frac{\theta\mu}{\sigma}+\frac{\nu\mu^2}{\sigma^2} \tag{15}$$

We now compute the age at which the productivity starts declining by applying the derivative to the expression (15) w.r.t to age and equalizing it to zero:

$$\frac{dy}{dage}=\frac{\theta}{\sigma}+\frac{2\nu}{\sigma^2}age-\frac{2\nu\mu}{\sigma^2}$$

$$\frac{\theta}{\sigma}+\frac{2\nu}{\sigma^2}age-\frac{2\nu\mu}{\sigma^2}=0$$

$$\gamma_{tp}=\mu-\frac{\sigma\theta}{2\nu}=49.5$$

The effect of age on γ is

$$\gamma_{age}=(\frac{\theta}{\sigma}-\frac{2\nu\mu}{\sigma^2})age=(0.0026M\$)age$$

We can now assess the level-level effect of age on productivity: this means that one additional year of age would increase the productivity by 0.0026 M\$ (γ_{age}). In line with the literature we found out that the productivity follows an inverted U-shape profile (see figure 5) with a significant decrease in productivity beginning at the age of 50 years old (γ_{tp}) (Skirbekk (2004); De Hek and van Vuuren (2011)).

7 Taxation on jobs mismatch

We identify the jobs mismatch following a normative approach based on qualification (Berlingieri and Erdsiek (2012); Stoevska (2017)). Therefore we rely on a classification elaborated by a professional jobs analyst concerning the level of education required in each occupation. The classification is supplied by O*NET. The severity of the jobs mismatch is measured by the incidence of overeducated or undereducated workers whether the educational attainment (observed) is respectively above or below the reference level provided by the O*NET in each occupation. We present the jobs mismatch by education and job category pairs (see table 3 for more details on the education and job categories). According to Acemoglu and Autor (2011) the job categories 1 to 10 associated to managerial, technical and professional occupations are specialized in the execution of non-routine cognitive tasks, the jobs belonging to categories 11 to 15 associated to service occupations are dominated by the execution of non-routine manual tasks, the jobs included in categories 16 and 17 associated with clerical, administrative and sales occupations are focused in performing routine cognitive tasks, and the jobs embedded in categories 18 to 22 associated to production and operative occupations are fully dominated by routine manual tasks.

Figure 6a displays the required level of education by job categories as expressed by the O*NET classification (we have 22 job categories referring to the two-digits of the SOC code) while figure 6b shows the observed percentage level of education for job categories (the percentage implicitly reveals the severity of the jobs mismatch by identifying the missing levels of education in figure 6a). In our case the jobs mismatch can be summarized by presenting two main patterns: the presence of undereducated workers in the job categories from 1 to 10 and the presence of overeducated workers in the residual job categories.

Apparently the productivity loss coincides with the jobs mismatch (see figure 7b). Indeed the workers holding at least the bachelor degree are highly productive in performing jobs requiring non-routine cognitive tasks, little productive in jobs requiring the execution of routine cognitive tasks and in some jobs requiring non-routine manual tasks (job categories 11 and 12), and unproductive in jobs specialized in routine manual tasks. The workers holding at most the high school diploma are quite productive in jobs categories ruled by routine manual tasks and in some job categories requiring non-routine manual tasks (job categories 13 and 14) while they are unproductive in jobs requiring the execution of routine and non-routine cognitive tasks. Lastly the workers owning some years of college seem to be unproductive independently of the job categories.

Conversely the job amenity manifest different paths vis-à-vis the jobs mismatch depending on the level of education of the workers. Generally, the workers holding at least the bachelor

degree present a clear disutility for working in job categories dominated by non-routine cognitive (unique exceptions job categories 6 and 8) and routine manual tasks while they enjoy working in jobs involving non-routine manual and routine cognitive tasks. Instead the workers occupying the remaining level of education dislike working in jobs implicating the execution of non-routine cognitive tasks while they enjoy working in the residual job categories (see figure 7a).

Unsurprisingly the well educated workers are also the ones capturing the higher gross wages independently of the job categories, particularly exorbitant in jobs performing non-routine cognitive tasks while they are substantially lower in jobs performing routine cognitive and non-routine manual tasks. The same pattern is followed by the workers holding at most some years of college which present overall similar level of wages in each job category, with a consistent rise for jobs performing non-routine cognitive tasks (see figure 8a). As expected the workers capturing the higher gross wages are also the ones experiencing the major reduction in after-tax wages, and this appears to be particularly relevant for workers performing jobs dominated by non-routine cognitive tasks (see figure 8b).

At this point it is interesting to illustrate the effect of the after-tax wage on the workers' preferences (see figure 9). By education level, the workers with at least the bachelor degree capture similar level of utility after-tax by performing jobs requiring the execution of non-routine, routine cognitive and non-routine manual tasks. Indeed, for the workers holding at least the master degree the difference between the maximum value of after-tax amenity expressed by jobs specialized in non-routine cognitive tasks and the minimum value of after-tax amenity provided by jobs dominated by routine cognitive tasks is just 0.0084 M$, for workers holding the bachelor degree the same difference is even less (0.0046 M$). By repeating (for those workers) the same computation between jobs specialized in non-routine cognitive and non-routine manual tasks produces similar values. Conversely, the workers with at most the high school degree generate similar level of utility after-tax by performing jobs specialized in routine and non-routine manual tasks (limited to the job categories 13 and 14) while the level of utility after-tax decreases in jobs dominated by non-routine cognitive tasks. Lastly, the workers holding some years of college produce a level of utility after-tax which is essentially flat across job categories.

This pattern clearly indicates that the higher gross wages offered to workers independently of the educational attainment to perform jobs requiring the execution of non-routine cognitive tasks can barely compensate for their disutility when taxation is considered, creating a latent incentive for workers to switch toward jobs which enhance their individual preferences more than remaining in jobs yielding higher wages. Therefore the overall impact of the after-tax wage on the workers' preferences is noticeable, enhancing the jobs mismatch for the workers holding at least a

bachelor degree by encouraging those workers towards jobs dominated by routine cognitive and non-routine manual tasks and reducing the jobs mismatch for workers holding at most the high school diploma by pushing them toward jobs which better reflect their educational attainment.

8 Conclusion

In the context of the US labor market we propose a model of matching able to deal with the imperfect transferable utility occurring between worker and firm when the friction caused by taxation is considered. In building our model we suggest a theoretic as well as practical approach to harmonize the US taxation system, assumed as necessary step for deriving the economic model and for estimating the worker and firm preferences. The cluster analysis combined with the algorithm proposed in section 4.2 seem to approximate reasonably well the US taxation, privileging the tax brackets where the mass of observations is relevant. The goodness of our approach hinges on its generality and indeed it may be readily applied to other cases, i.e the European labor market.

Thereafter we estimate the job amenity and productivity which reveal a clear dissatisfaction to work in jobs requiring the execution non-routine cognitive and manual tasks, and both increase respectively with age and years of education. Conversely the productivity is positively related to age and years of education.

Afterwards we identify the jobs mismatch based on qualifications, illustrating its implication for the job amenity and productivity.

Apparently, the productivity loss coincides with the jobs mismatch. Indeed the well educated workers are poorly productive in jobs dominated by routine cognitive and non-routine manual tasks and unproductive in jobs requiring the execution of manual tasks while they are productive in jobs demanding the implementation of non-routine cognitive tasks. The workers holding at most the high school diploma are quite productive in jobs performing non-routine manual and routine manual tasks and unproductive in jobs specialized in non-routine and routine cognitive tasks. Ultimately, the workers with some years of college are scarcely productive if not unproductive in almost all job categories.

Conversely the job amenity manifest different paths vis-à-vis the jobs mismatch depending on the level of education of the workers. In contrast with the productivity pattern the workers with at least the bachelor degree manifest a clear satisfaction in implementing non-routine manual and routine cognitive tasks while in line with their productivity workers holding at most the high school diploma clearly enjoy working in jobs implicating the execution of non-routine and routine

manual tasks as well as routine cognitive tasks.

Lastly we assess the effect of the after-tax wage on the workers' preferences in order to figure out its potential role in stimulating or discouraging the jobs mismatch. The combined effect of after-tax wage and workers' preferences definitely reveal two main patterns: the enhancement of the jobs mismatch for those workers manifesting a superior education (at least the bachelor degree) by driving their choices from jobs specialized in the implementation of non-routine cognitive tasks toward jobs focused on the execution of routine cognitive and non-routine manual tasks and the decline of the jobs mismatch for workers holding at most a high school diploma by pushing those workers toward jobs more consistent with their level of education.

9 Appendix

A. Proofs

A.1. Proof of Lemma1

Proof. The interval of the generic state s is identified by (t_s^k, t_s^{k+1}) with $s = 1, .., S$ and $k = 1, .., K^s$. Hence the proceeding consists in evaluating for each pair of states s and $s\prime$ all the intersections of type $(t_s^k, t_s^{k+1}) \cap (t_{s\prime}^{k\prime}, t_{s\prime}^{k\prime+1})$ with $k\prime = 1, .., K^{s\prime}$ and $s\prime = 1, ..., S$. It is intuitive to recognize that the interval $(t_s^{K^s}, t_s^{K^s+1})$ for the generic state s coincides with the interval above the last threshold $t_s^{K^s}$, then for this reason the expression $(t_s^K, t_s^{K^s+1})$ is replaced by the expression $(> t_s^{K^s})$ in the proof.

Indeed, for simplicity, let's start by $s = 1$ in that case the intersections are

for $k = 1$

$$(t_1^1, t_1^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^1, t_1^2) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^1, t_1^2) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = 2$

$$(t_1^2, t_1^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^2, t_1^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^2, t_1^3) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^1 - 1$

$$(t_1^{K^1-1}, t_1^{K^1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^{K^1-1}, t_1^{K^1}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^{K^1-1}, t_1^{K^1}) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^1$

$$(> t_1^{K^1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (> t_1^{K^1}) \cap (t_{s\prime}^{K^{c\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_1^{K^1}) \cap (> t_{s\prime}^{K^{s\prime}})$$

where $s\prime$ is introduced to identify a temporary reference state picked from the set of states and replaced by another one as soon as the s states have been compared with it.

Continuing with the country $s = 2$ the intersections are,

for $k = 1$

$$(t_2^1, t_2^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^1, t_2^2) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_2^1, t_2^2) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = 2$

$$(t_2^2, t_2^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^2, t_2^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_2^2, t_2^3) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^2 - 1$

$$(t_2^{K^2-1}, t_2^{K^2}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^{K^2-1}, t_2^{K^2}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_2^{K^2-1}, t_2^{K^2}) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^2$

$$(t_2^{K^2}, t_2^{K^2+1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^{K^2}, t_2^{K^2+1}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_2^{K^2}) \cap (> t_{s\prime}^{K^{s\prime}})$$

and for country $s = S$ the intersections are,
for $k = 1$

$$(t_S^1, t_S^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^1, t_S^2) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^1, t_S^2) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = 2$

$$(t_S^2, t_S^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^2, t_S^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^2, t_S^3) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^S - 1$

$$(t_S^{K^S-1}, t_S^{K^S}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^{K^S-1}, t_S^{K^S}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^{K^S-1}, t_S^{K^S}) \cap (> t_{s\prime}^{K^{s\prime}})$$

for $k = K^S$

$$(t_S^{K^S}, t_S^{K^S+1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^{K^S}, t_S^{K^S+1}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_S^{K^S}) \cap (> t_{s\prime}^{K^{s\prime}})$$

With these proceeding we manage to obtain the intersections evaluated in pairs of the states $s = 1, .., S$ relative to a state $s\prime$. At this point we need to compute for each state $s = 1, .., S$ the union of all these intersections.
For $s = 1$ we have

$$\begin{aligned}
\cup [&(t_1^1, t_1^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^1, t_1^1) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^1, t_1^2) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(t_1^2, t_1^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^2, t_1^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^2, t_1^3) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(t_1^{K^1-1}, t_1^{K^1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_1^{K^1-1}, t_1^{K^1}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_1^{K^1-1}, t_1^{K^1}) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(> t_1^{K^1}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (> t_1^{K^1}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_1^{K^1}) \cap (> t_{s\prime}^{K^{s\prime}})] = \mathcal{U}_{1s\prime}
\end{aligned}$$

for $s = 2$

$$\begin{aligned}
\cup [&(t_2^1, t_2^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^1, t_2^2) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_2^1, t_2^2) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(t_2^2, t_2^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^2, t_2^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{p\prime}}), (t_2^2, t_2^3) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(t_2^{K^2-1}, t_2^{K^2}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_2^{K^2-1}, t_2^{K^2}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_2^{K^2-1}, t_2^{K^2}) \cap (> t_{s\prime}^{K^{s\prime}}), \\
&(> t_2^{K^2}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (> t_2^{K^2}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_2^{K^2}) \cap (> t_{s\prime}^{K^{s\prime}})] = \mathcal{U}_{2s\prime}
\end{aligned}$$

for $s = S$ we get

$$\cup\,[(t_S^1, t_S^2) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^1, t_S^2) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^1, t_S^2) \cap (> t_{s\prime}^{K^{s\prime}}),$$
$$(t_S^2, t_S^3) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^2, t_S^3) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^2, t_S^3) \cap (> t_{s\prime}^{K^{s\prime}}),$$
$$(t_S^{K^S-1}, t_S^{K^S}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (t_S^{K^S-1}, t_S^{K^S}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (t_S^{K^S-1}, t_S^{K^S}) \cap (> t_{s\prime}^{K^{s\prime}}),$$
$$(> t_S^{K^S}) \cap (t_{s\prime}^1, t_{s\prime}^2), ..., (> t_S^{K^S}) \cap (t_{s\prime}^{K^{s\prime}-1}, t_{s\prime}^{K^{s\prime}}), (> t_S^{K^S}) \cap (> t_{s\prime}^{K^{s\prime}})] = \mathcal{U}_{Ss\prime}$$

$\mathcal{U}_{1s\prime}, \mathcal{U}_{2s\prime}, .., \mathcal{U}_{Ss\prime}$ is a convenient notation to denote the operation of union of the intersections generated in pairs between country $s = 1, .., S$ and a specific $s\prime$. At this stage we need to bring together all these unions realizing a union over $s = 1, .., S$.

$$\bigcup [\mathcal{U}_{1s\prime}, \mathcal{U}_{2s\prime}, ..., \mathcal{U}_{Ss\prime}]$$

The above exposed proceeding must be repeated using at each time a different $s\prime$ until we exhaust the list of S states. Finally we need to take the union over $s\prime = 1, .., S$.

$$\bigcup_{s\prime=1}^{S} \bigcup [\mathcal{U}_{1s\prime}, \mathcal{U}_{2s\prime}, ..., \mathcal{U}_{Ss\prime}] \tag{16}$$

The expression (16) can be written in a compact form as

$$\bigcup_{s\prime=1}^{S} \bigcup_{s=1}^{S} (1 - \delta_{ss\prime})[\mathcal{U}_{ss\prime}]$$

where $\mathcal{U}_{ss\prime}$ is

$$\bigcup_{k=1}^{K^s} \left[(t_s^k, t_s^{k+1}) \bigcap_{k\prime=1}^{K^{s\prime}} (t_{s\prime}^{k\prime}, t_{s\prime}^{k\prime+1}) \right]$$

In conclusion, the procedure presented above boils down to a general formula

$$\bigcup_{s\prime=1}^{S} \bigcup_{s=1}^{S} (1 - \delta_{ss\prime}) \left\{ \bigcup_{k=1}^{K^s} \left[(t_s^k, t_s^{k+1}) \bigcap_{k\prime=1}^{K^{s\prime}} (t_{s\prime}^{k\prime}, t_{s\prime}^{k\prime+1}) \right] \right\}$$

where the introduction of the factor $1 - \delta_{ss\prime}$ avoids to compute the intersection of the state s with itself when $s = s\prime$, so the *delta of Kronecker* reads as

$$\delta_{ss\prime} = \begin{cases} 1, & \text{if } s = s\prime, \\ 0, & \text{if } s \neq s\prime. \end{cases}$$

■

Remark 1. *By definition of intersection* $\mathcal{U}_{1s\prime} = \mathcal{U}_{s\prime 1}, \mathcal{U}_{2s\prime} = \mathcal{U}_{s\prime 2}, .., \mathcal{U}_{Ss\prime} = \mathcal{U}_{s\prime S}$.

A.2. Formulation of the matching equation and transfer

Manipulating equation (8) we get:

$$[\ln \mu(x,y) - \ln m(x)]\sigma^W = \min_{k\in\{1,\dots,K\}} \alpha^k(x,y) + (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))(w(x,y)-t^k) - u(x) \quad (17)$$

summing and substracting $(1+\tau(y))t^k$ to equation (9) we get

$$[\ln \mu(x,y) - \ln n(y)]\sigma^F = \gamma^k(x,y) - (1+\tau(y))(w(x,y)-t^k) - v(y) \quad (18)$$

We can solve for $w(x,y)$ the equation (18)

$$w(x,y) = \frac{-[\ln \mu(x,y) - \ln n(y)]\sigma^F + \gamma^k(x,y) - v(y)}{1+\tau(y)} + t^k \quad (19)$$

Replace (19) in (17)

$$[\ln \mu(x,y) - \ln m(x)]\sigma^W = \min_{k\in\{1,\dots,K\}} \alpha^k(x,y) + (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))\left(\frac{-[\ln \mu(x,y) - \ln n(y)]\sigma^F + \gamma^k(x,y) - v(y)}{1+\tau(y)} + t^k\right) - (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))t^k - u(x) \quad (20)$$

being $\xi^{W,k}(x) = \frac{1}{1-\phi^k_{fd}(x)}$ and $\lambda^{F,k}(y) = \frac{1-\phi^k_{st}(y)}{1+\tau^F(y)}$ we can rewrite equation (20)

$$[\ln \mu(x,y) - \ln m(x)]\sigma^W = \min_{k\in\{1,\dots,K\}} \alpha^k(x,y) + \frac{\lambda^{F,k}(y)}{\xi^{W,k}(x)}([-\ln \mu(x,y) + \ln n(y)]\sigma^F + \gamma^k(x,y) - v(y)) - u(x)$$

Solving for $\mu(x,y)$ we get:

$$\ln \mu(x,y)(\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F) = \min_{k\in\{1,\dots,K\}} \xi^{W,k}(x)(\alpha^k(x,y) - u(x)) + \lambda^{F,k}(y)(\gamma^k(x,y) - v(y)) + \lambda^{F,k}(y)\sigma^F \ln n(y) + \xi^{W,k}(x)\sigma^W \ln m(x)$$

and then

$$\mu(x,y) = \min_{k\in\{1,\dots,K\}} m(x)^{\frac{\xi^{W,k}(x)\sigma^W}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W}} n(y)^{\frac{\lambda^{F,k}(y)\sigma^F}{\lambda^{F,k}(y)\sigma^F + \xi^{W,k}(x)\sigma^W}} \exp\frac{\xi^{W,k}(x)(\alpha^k(x,y)-u(x)) + \lambda^{F,k}(y)(\gamma^k(x,y)-v(y))}{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F}$$

The transfer may be obtained by equalizing equation (8) and equation (9) after summing and substracting $(1+\tau(y))t^k$ in equation (9)

$$n(y)\frac{\exp\left(\frac{\gamma^k(x,y)-(1+\tau(y))(w(x,y)-t^k)}{\sigma^F}\right)}{\exp\left(\frac{v(y)}{\sigma^F}\right)} = m(x)\frac{\exp\left(\frac{\alpha^k(x,y)+(1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))(w(x,y)-t^k)}{\sigma^W}\right)}{\exp\left(\frac{u(x)}{\sigma^W}\right)}$$

Solving for $w(x,y)$ we get

$$\sigma^W\sigma^F\ln\frac{n(y)}{m(x)} + \sigma^W(\gamma(x,y)^k + (1+\tau(y))t^k - v(y)) - \sigma^F(\alpha^k(x,y) - (1-\phi^k_{fd})(1-\phi^k_{sy})t^k - u(x)) = w(x,y)[(1+\tau(y))\sigma^W + (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))\sigma^F]$$

then manipulating the last expression we get

$$\sigma^W\sigma^F\ln\frac{n(y)}{m(x)} + \sigma^W(\gamma^k(x,y) + (1+\tau(y))t^k - v(y)) - \sigma^F(\alpha^k(x,y) - (1-\phi^k_{fd}(x))(1-\phi^k_{st}(y))t^k - u(x)) = \widetilde{w}^k(x,y)\left[\frac{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F}{\xi^{W,k}(x)\lambda^{F,k}(y)}\right]$$

Finally,

$$\widetilde{w}^k(x,y) = \frac{\xi^{W,k}(x)\lambda^{F,k}(y)\sigma^F\sigma^W}{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F}\ln\frac{n(y)}{m(x)} + \frac{\xi^{W,k}(x)\lambda^{F,k}(y)\sigma^W}{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F}(\gamma^k(x,y) - v(y)) - \frac{\xi^{W,k}(x)\lambda^{F,k}(y)\sigma^F}{\xi^{W,k}(x)\sigma^W + \lambda^{F,k}(y)\sigma^F}(\alpha^k(x,y) - u(x)) + (1-\phi^k_{st}(y))t^k$$

This is the same expression appearing in the square bracket of equation (11) which needs to multiplied by $\frac{1}{(1-\phi^k_{st}(y))}$ and minimized to get the equilibrium transfer $w(x,y)$.

B. Details on the harmonisation procedure

B.1. Silhouette method

The silhouette method evaluates the intra-cluster and inter-cluster distances both converging in the definition of the silhouette index, which it is used to compare different cluster configurations. For each group of states, we use ten different cluster configurations, starting from one cluster in which all the thresholds belong to the same cluster until reaching ten clusters where the thresholds split among ten clusters. The following formulas represent the core of the method:

$$a(i) = \frac{1}{|N_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

$$b(i) = \min_{k \neq i} \frac{1}{|N_k|} \sum_{j \in C_k} d(i, j)$$

$a(i)$ represents the average distance between each threshold i and all other thresholds j in the same cluster and $b(i)$ represents the minimum average distance of each threshold i to each cluster j. Then for each $a(i)$ and $b(i)$ (then for each threshold i), we define a silhouette value (for construction the $-1 \leqslant s(i) \leqslant 1$):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{21}$$

For each cluster in a specific cluster configuration, we compute the average silhouette value:

$$\overline{s}(k) = \frac{1}{N_k} \sum_{i=1}^{N_k} s(i) \tag{22}$$

where N_k is the number of thresholds in the cluster k (for a cluster configuration C_k with k clusters). Finally, we can compute the average silhouette value for the specific cluster configuration:

$$\overline{s}(C_k) = \frac{1}{C_k} \sum_{C_k} \overline{s}(k) \tag{23}$$

After computing the $\overline{s}(C_k)$ for each cluster configuration, we apply the Kaufman's rule (Leonard and Peter (1990)) delivering the optimal number of clusters such that the maximum value among $\overline{s}(C_k)$ is identified by the silhouette index (SI) with $0 \leqslant SI \leqslant 1$:

$$SI = \max_{C_k} \overline{s}(C_k) \tag{24}$$

The intuition behind the last formula is based on the requirement that we need to maximize the between cluster distance $b(i)$ while try to minimize the within cluster distance $a(i)$. Then the cluster configuration providing the largest value of $\overline{s}(C_k)$ is chosen. In this case, the total number of thresholds obtained when combining all the centroids (the average value of the thresholds in a cluster) is eight. In other words eight represents the overall number of thresholds obtained when combining all the optimal cluster solutions delivered by the silhouette analysis in each group of states.

Specifically, we got the following silhouette index for the three group of states:

- First group of US states: SI is equal to 0.934, this value corresponds to the optimal cluster configuration attributing two clusters.
- Second group of US states: SI is equal to 0.938, this value corresponds to the optimal cluster configuration attributing two clusters.
- Third group of US states: SI is equal to 0.8732, this value corresponds to the optimal cluster configuration attributing four clusters

However, for the second group of US states we adopt a slightly different evaluation of the optimal cluster based on Rousseeuw (1987), who provides the rank concerning the goodness of the silhouette index:

- SI $\leqslant 0.25$ indicates no cluster structure has been found
- $0.26 \leqslant$ SI $\leqslant 0.5$ indicates that the cluster structure is weak
- $0.51 \leqslant$ SI $\leqslant 0.7$ indicates a reasonable cluster structure has been found
- $0.71 \leqslant$ SI $\leqslant 1$ indicates a strong structure has been found

We adopt this rank to drive the choice of the optimal cluster solutions of the second group of states. The justification hinges on the fact that the number of thresholds in that case is equal to nine, this implies that the optimal solution would lead to pick the SI associated with the cluster configuration involving nine clusters (one threshold for each cluster provides the largest silhouette value for each cluster, and then the largest $\overline{s}(C_k)$. This choice would be meaningless. Moreover, in the attempt to select a cluster configuration with $SI \geq 0.71$ while trying to minimize the number of clusters we identify in two clusters our optimal solutions for the second group of states.

B.2. Gaussian Mixture Model

We proceed with a robustness check using the Gaussian Mixture Model (GMM) as conclusion of the cluster analysis. The intent is to use this probabilistic method based on Bayesian statistics to check the validity of the optimal cluster solutions achieved with k-means. The main differences with k-means are essentially two:

- It rests on soft clustering meaning the assignment of each threshold to a component (cluster) is occurring in a probabilistic fashion
- For each cluster, it updates not only the mean (centroids) but also the variance in the univariate case and the covariance matrix in a multivariate case.

The GMM is a convex combination of Gaussian component densities:

$$f(\boldsymbol{y}, \boldsymbol{\psi}) = \sum_{i=1}^{g} \pi_i \phi_p(\boldsymbol{y}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

$$\phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^T (\boldsymbol{y}-\boldsymbol{\mu})}$$

where $\boldsymbol{y}$ denotes a p-dimensional random vector, $\boldsymbol{\mu}$ is a p-dimensional mean vector, $\boldsymbol{\Sigma}$ is a p x p positive definite covariance matrix ($|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$), $\boldsymbol{\pi_i}$ are the mixing proportions satisfying $\pi_i \geq 0$ and $\sum_{i=1}^{g} \pi_i = 1$, the vector $\boldsymbol{\psi}$ is containing all the unknown parameters of the mixture model. In particular $\boldsymbol{\psi} = (\pi_1, \ldots, \pi_g; \theta_1, \ldots, \theta_g)$ and θ_i denotes the vector of unknown parameters of the i-th component density, which contains the elements of $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$.

The GMM employs an expectation-maximization algorithm:

- In the expectation phase, we apply the maximum a posteriori (MAP) rule to each threshold, meaning that we assign each threshold to the component density for which the posterior probability computed using the Bayesian theorem has the largest value.
- In the maximization phase, the posterior probability is used to update the vector of the mixing proportions, the mean vector and the covariance matrix.

The algorithm stops when the likelihood function improves less than a fixed tolerance.

In our case, the idea is to use the division of the US states achieved with the hierarchical clustering, and apply to each of this group of states the GMM algorithm. Before applying the GMM algorithm, we definitely need to fix one aspect of this computation. The K-means algorithm updates the centroids while GMM updates the means and the variances. Moreover, we need to

reduce the number of parameters updated. This is achieved using the pooled variance option in the GMM algorithm:

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_j - \overline{y}_i)$$

where s_p^2 is the pooled variance, k is the number of components for each group of states (in this case, the number of components is equal to the optimal number of clusters obtained from k-means), n_i is the number of thresholds for each component, $\overline{y}_i$ is the average threshold value for each component (centroids). With this specification, the variance would be updated at each iteration through the updated mean and not through the posterior probability. Moreover, the implicit robustness check obtained by applying the GMM rests on imposing to each group of states the optimal number of clusters obtained from k-means. Therefore, when we impose the optimal number of clusters derived from k-means to the GMM algorithm for each group of states and this returns similar mean value in each component that the one got from k-means, the logic implication is that the assignment of the thresholds to the appropriate cluster for GMM and k-means coincides. The mean value of each component computed by the GMM is the arithmetic average of the thresholds in that component (we provided the same definition for centroids concerning k-means).

In our case, the GMM method is rightly providing similar if not equal centroids compared to k-means.

B.3. Choice of the tolerance error and marginal tax rate assignment

The lack of theoretical clue concerning the assignment of the tolerance error drives our decision to adopt a pragmatic procedure. Moreover the tolerance error chosen is the one delivering marginal tax rates which better approximate the marginal tax rates either of the 34 US state taxations and of the federal taxation.

For each state and federal taxation, we explicitly compare the original marginal tax rates with the ones got with the final taxation in three different cases corresponding to three tolerance errors expressed in percent (3; 0.75; 0.2). We choose these tolerance errors because they reflect three different scenarios concerning the size of the final taxations. The first scenario (0.2 percent) generates a final taxation with almost the same thresholds obtained by merging the federal taxation and the synthetic state taxation, the second scenario (0.75 percent) generates a final taxation with half of the thresholds obtained by merging the federal taxation and the synthetic

state taxation and the third scenario (3 percent) generates a final taxation with one third of the thresholds obtained by merging the federal taxation and the synthetic state taxation.

Each tolerance error generates a distinctive final taxation. Practically, for each state and federal taxation within the tolerance error we compute the marginal tax rates associated with the final taxation and the percent errors committed in approximating the marginal tax rates of the original taxation. Moreover the sum of these percent errors represents the cumulative percent error of each state and of federal taxation. Finally, the tolerance error reporting the smallest cumulative percent errors has been picked.

An intermediate step entails within each tolerance error the imputation of the marginal tax rates associated with the final taxation for each state and for federal taxation. To be concrete, let's pick the state s with generic intervals $[t_s^k, t_s^{k+1}]$ and $[t_s^{k+1}, t_s^{k+2}]$ and marginal tax rates ϕ_s^k and ϕ_s^{k+1} and the final taxation with the generic interval $[t_{fin}^k, t_{fin}^{k+1}]$ (where the subscripts fin stands for the final taxation). Assuming for simplicity that $t_s^k < t_{fin}^k < t_s^{k+1}$ and $t_s^{k+1} < t_{fin}^{k+1} < t_s^{k+2}$ (this represents one of the possible interval combinations between the original taxation of state s and the final taxation), we compute the marginal tax rate $\phi_{fin,s}^k$ of the state s in the interval $[t_{fin}^k, t_{fin}^{k+1}]$ using:

- a weighted average using as weight w_s^k in the interval $[t_s^k, t_s^{k+1}]$ and w_s^{k+1} in the interval $[t_s^{k+1}, t_s^{k+2}]$:

$$\phi_{fin,s}^k = \frac{\phi_s^k w_s^k + \phi_s^{k+1} w_s^{k+1}}{w_s^k + w_s^{k+1}}$$

  where w_s^k is expressed through the equation 14. Similarly we can express w_s^{k+1}

In our case, the cumulative percent errors generated by the tolerance error of 3 percent are systematically higher than the cumulative percent errors associated with the tolerance error of 0.75 percent, and this appears consistent for each US state and for federal taxation. Instead the comparison of the cumulative percent errors associated with the tolerance error of 0.75 percent with the ones associated with the tolerance error 0.2 percent reveals similar values with only few US states exhibiting a marginal improvement in the approximation of the original state taxation. However the adoption of the final taxation derived assuming a tolerance error of 0.2 percent would produce a taxation with higher number of thresholds adding a further computational cost when running the model against a negligible gain in precision. Moreover in order to minimize the trade-off between precision and computational cost we identify in the final taxation generated by the tolerance error of 0.75 percent our best choice.

C. ASEC dataset

C.1. The evolution of the topcode in the Current population survey and its implication

Historically, the US Census topcoded procedure has been extensively used for matter of confidentiality. The procedure has evolved and dramatically changed over time.

From 1962 to 1995, the values exceeding the topcode were simply recorded with the threshold values. During 1996-2010, the Census introduced a new replacement procedure assigning to each individual disclosing an earning above the threshold for the specific year the mean income within a specific group (cell mean). The groups were formed based on 12 allocation groups depending on characteristics such as gender, race and full-time status. From 2011 onward the procedure dictates that the values exceeding the topcoded earning are rounded to two significant digits and then swapped among individuals within bounded intervals. This last method is called rank proximity swap.

Most of the economists using CPS and evaluating long-term earning used the correction proposed by Katz and Murphy (1992). This correction implies the multiplication of the topcoded earnings by a factor of 1.4 or 1.5 (Autor et al. (2008); Juhn et al. (1993); Lemieux (2006)). This last coming from the assumption that the distribution of the higher tail of the earnings is Pareto. Along this literature, Armour et al. (2016) support the idea that using a fixed factor across years and across changes in the threshold levels could lead to misleading estimation of the long-term trends in earnings. They proposed to use an ad hoc procedure to deal with topcoded earnings enabling the shape parameter of the Pareto distribution to adjust dynamically over time. This procedure is beneficial when using internal CPS data providing none topcode on earnings but higher censoring point. From 2011 the introduction of the rank proximity swap although masking the true correspondence of individual earnings by swapping this information among individuals within bounded intervals when earnings are above the public topcode, it allows to exploit the earning information between the public topcode and the internal censoring point. Then for this reason the formula of the shape parameter developed by Armour et al. (2016) is suitable for our ASEC supplement of 2015.

The formulation of the shape parameter proposed by Armour et al. (2016) can read as:

$$\alpha = \frac{M}{T\log(x_t) + \sum_{x_c \leqslant x_i < x_t} \log(x_i) - (M+T)\log(x_c)} \tag{25}$$

where M is the number of individuals with earnings between lower cutoff (x_c is the public threshold for wage and salary, $\$280,000$ in ASEC 2015) and the internal censoring point (x_t equal

to $\$1,099,999$), T is the number of individuals with earnings at or above the internal censoring point, and x_i is the earning of an individual i.

Hence the mean earning above threshold x_t

$$M(x_t) = \frac{\alpha}{(\alpha - 1)} x_t, \tag{26}$$

As reported by Armour et al. (2016), the above expression allows individuals between the cutoff and censoring point to contribute to the cumulative density function (CDF) with their actual earnings, while those at the cutoff or above to contribute to the CDF with the information that they have earnings at least as high as the censoring point.

Bibliography

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, volume 4, pages 1043–1171. Elsevier.

Armour, P., Burkhauser, R. V., and Larrimore, J. (2016). Using the pareto distribution to improve estimates of topcoded earnings. *Economic Inquiry*, 54(2):1263–1273.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in us wage inequality: Revising the revisionists. *The Review of economics and statistics*, 90(2):300–323.

Becker, G. (1964). Human capital. a theoretical and empirical analysis with special reference to education. new york: Columbia university press for the nber.

Berlingieri, F. and Erdsiek, D. (2012). How relevant is job mismatch for german graduates? *ZEW-Centre for European Economic Research Discussion Paper*, (12-075).

Boadway, R., Marchand, M., and Pestieau, P. (1991). Optimal linear income taxation in models with occupational choice. *Journal of Public Economics*, 46(2):133–162.

Brown, C. (1980). Equalizing differences in the labor market. *The Quarterly Journal of Economics*, 94(1):113–134.

Chan, M., Kroft, K., and Mourifie, I. (2019). An empirical framework for matching with imperfect competition. *Unpublished manuscript*.

Choo, E. and Siow, A. (2006). Who marries whom and why. *Journal of political Economy*, 114(1):175–201.

Converso, D., Sottimano, I., Guidetti, G., Loera, B., Cortini, M., and Viotti, S. (2018). Aging and work ability: the moderating role of job and personal resources. *Frontiers in psychology*, 8:2262.

David, H., Katz, L. F., and Kearney, M. S. (2005). Rising wage inequality: the role of composition and prices. Technical report, National Bureau of Economic Research.

De Hek, P. and van Vuuren, D. (2011). Are older workers overpaid? a literature review. *International Tax and Public Finance*, 18(4):436–460.

Duncan, G. J. and Hoffman, S. D. (1981). The incidence and wage effects of overeducation. *Economics of education review*, 1(1):75–86.

Dupuy, A. and Galichon, A. (2015). A note on the estimation of job amenities and labor productivity.

Dupuy, A., Galichon, A., Jaffe, S., and Kominers, S. D. (2017). Taxation in matching markets.

Galichon, A., Kominers, S., and Weber, S. (2017). Costly concessions: An empirical framework for matching with imperfectly transferable utility.

Galichon, A., Kominers, S. D., and Weber, S. (2015). The nonlinear bernstein-schrodinger equation in economics. In *International Conference on Networked Geometric Science of Information*, pages 51–59. Springer.

Galichon, A. and Salanié, B. (2015). Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*.

Hartog, J. (1980). Earnings and capability requirements. *The Review of Economics and Statistics*, pages 230–240.

Hartog, J. (2000). Over-education and earnings: where are we, where should we go? *Economics of education review*, 19(2):131–147.

Hwang, H.-s., Reed, W. R., and Hubbard, C. (1992). Compensating wage differentials and unobserved productivity. *Journal of Political Economy*, 100(4):835–858.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of political Economy*, 101(3):410–442.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics*, 107(1):35–78.

Lemieux, T. (2006). Postsecondary education and increasing wage inequality. *American Economic Review*, 96(2):195–199.

Lemieux, T. (2010). What do we really know about changes in wage inequality? In *Labor in the new economy*, pages 17–59. University of Chicago Press.

Leonard, K. and Peter, J. R. (1990). Finding groups in data: an introduction to cluster analysis. In *Probability and Mathematical Statistics. Applied Probability and Statistics*. Wiley Series.

Lockwood, B. B., Nathanson, C. G., and Weyl, E. G. (2017). Taxation and the allocation of talent. *Journal of Political Economy*, 125(5):1635–1682.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

Meyer, S.-C. and Hünefeld, L. (2018). Challenging cognitive demands at work, related working conditions, and employee well-being. *International journal of environmental research and public health*, 15(12):2911.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the united states. *Journal of Economic perspectives*, 25(3):173–96.

Park, J. H. et al. (1994). Estimation of sheepskin effects and returns to schooling using he old and the new cps measures of educational attainment. Technical report.

Parker, S. C. (2003). Does tax evasion affect occupational choice? *Oxford Bulletin of Economics and Statistics*, 65(3):379–394.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., Campion, M. A., Mayfield, M. S., Morgeson, F. P., Pearlman, K., et al. (2001). Understanding work using the occupational information network (o* net): Implications for practice and research. *Personnel Psychology*, 54(2):451–492.

Powell, D. and Shan, H. (2012). Income taxes, compensating differentials, and occupational choice: How taxes distort the wage-amenity decision. *American Economic Journal: Economic Policy*, 4(1):224–47.

Quintini, G. (2011). Right for the job: Over-qualified or under-skilled?

Rothschild, C. and Scheuer, F. (2013). Redistributive taxation in the roy model. *The Quarterly Journal of Economics*, 128(2):623–668.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sanders, C. (2011). Skill uncertainty, skill accumulation and occupational choice. washington university at st. Technical report, Louis, Mimeo.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of economic literature*, 31(2):831–880.

Sheshinski, E. (2003). Note on income taxation and occupational choice.

Skirbekk, V. (2004). Age and individual productivity: A literature survey. *Vienna yearbook of population research*, pages 133–153.

Stoevska, V. (2017). Qualification and skill mismatch: Concepts and measurement. *ILO*.

Thurow, L. C. and Lucas, R. E. (1972). *The American distribution of income: a structural problem*, volume 7. US Government Printing Office.

Uragun, B. and Rajan, R. (2013). The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modelling. *BMC neuroscience*, 14(1):114.

Wilemme, G. (2017). Optimal taxation to correct job mismatching.

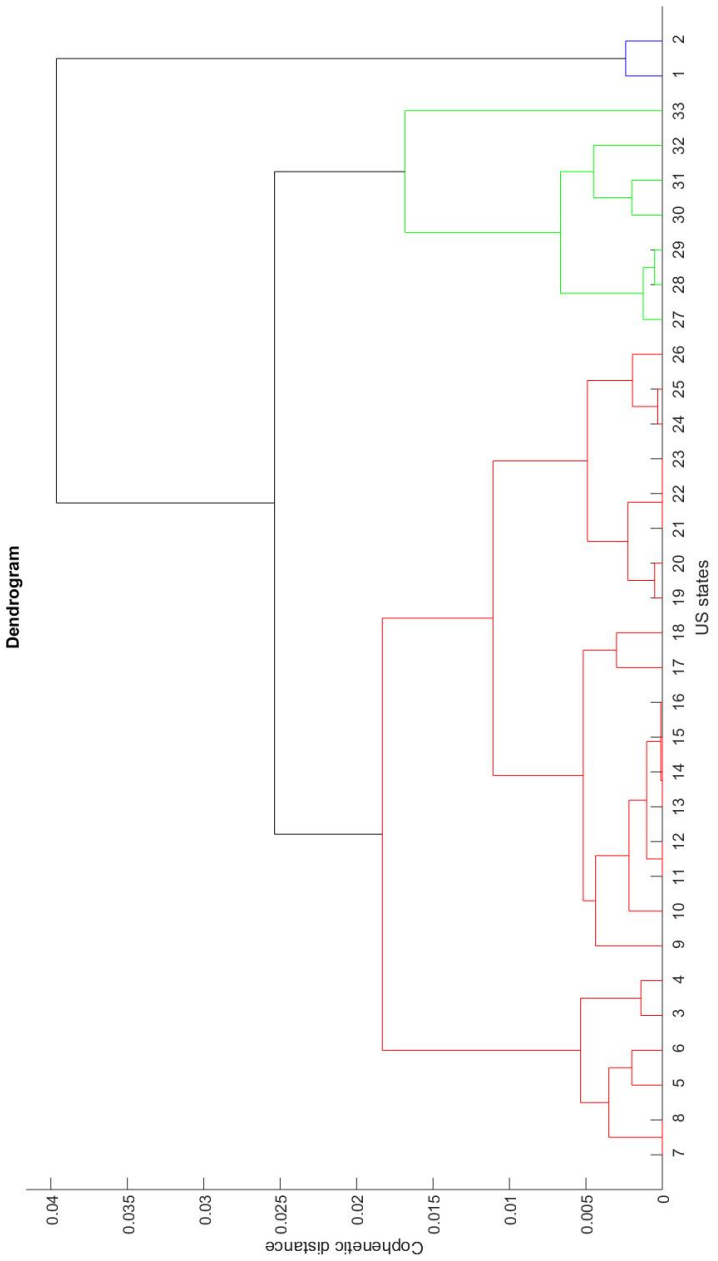


Figure 1: is providing the visual representation (dendrogram) of the three group of US states adopted in the k-means clustering: on the x-axis we have the 33 US states and on the y-axis we have the cophenetic distances. In particular in green we have the group 1 representing US states displaying high marginal tax rates by median wage in 2014, in blue we have group 2 representing US states displaying low marginal tax rates by median wage in 2014, in red we have group 3 representing US states displaying medium marginal tax rates by median wage in 2014.

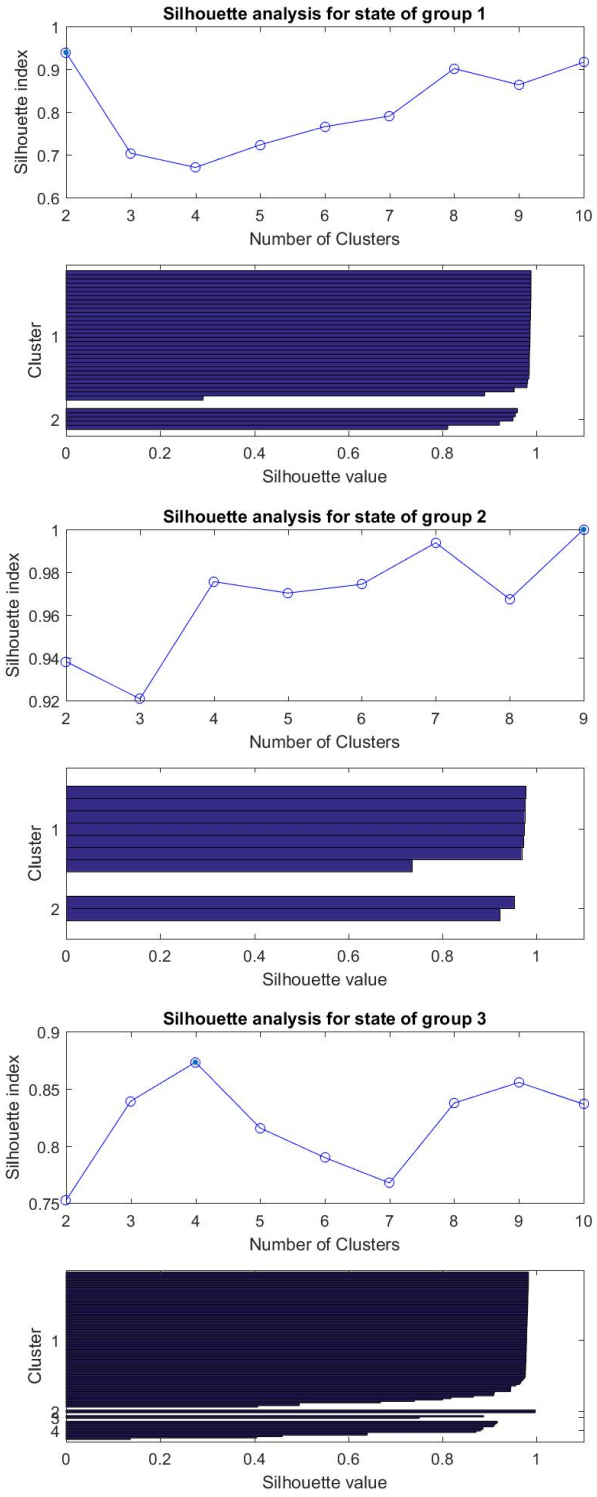
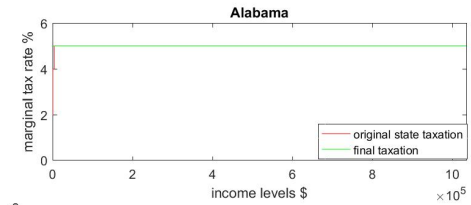
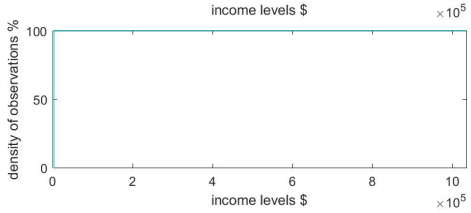
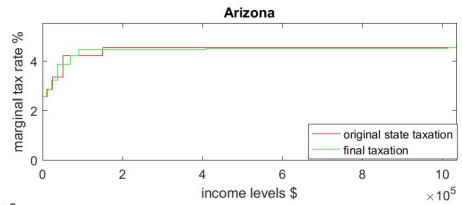
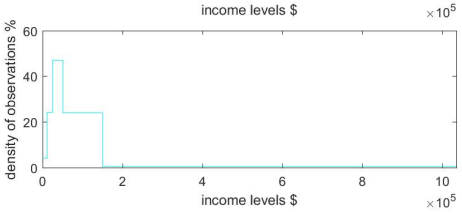


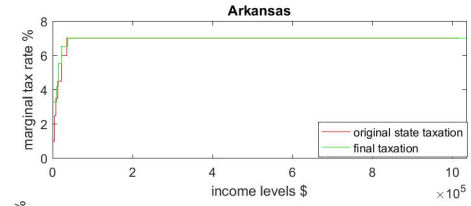
Figure 2: shows the optimal cluster solutions for each group of US states. The construction of the distinctive group of US states has been provided by the hierarchical clustering. For each group of US states: on the upper graph we have the silhouette index by cluster configurations, on the lower graph we have the silhouette values within each cluster forming the optimal cluster configuration. For the group 1 and group 3 we follow the Kaufman's rule to select the optimal cluster configuration (more detailed in appendix B.1). Instead for the group 2 we have two US states with cumulatively 9 thresholds, therefore using the silhouette index would suggest nine clusters as the best cluster configuration (upper graph). This choice appears meaningless with one threshold for each cluster, so we select the optimal cluster configuration based on Rousseeuw (1987) (lower graph) (more details in appendix B.1).

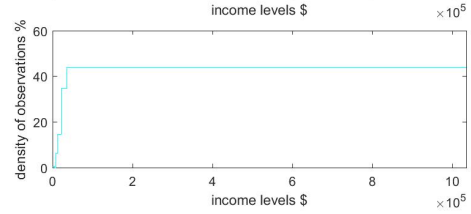

Alabama
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

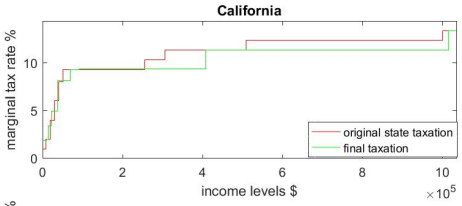

density of observations %
income levels $
×10^5

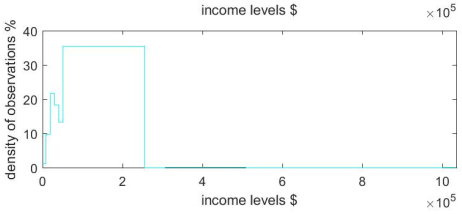

Arizona
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

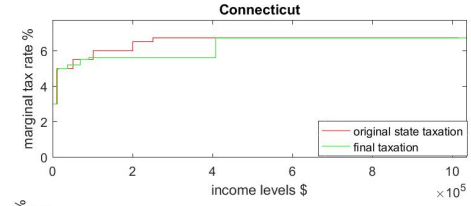

density of observations %
income levels $
×10^5

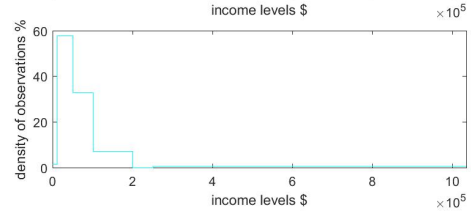

Arkansas
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

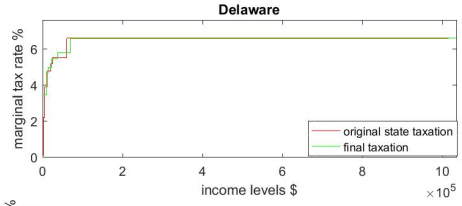

density of observations %
income levels $
×10^5

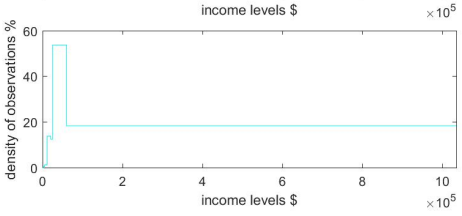

California
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

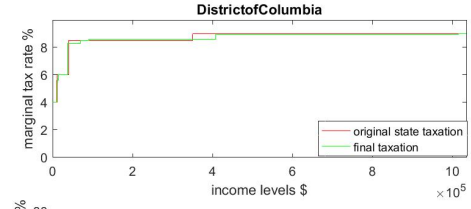

density of observations %
income levels $
×10^5

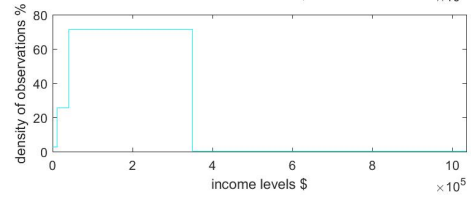

Connecticut
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

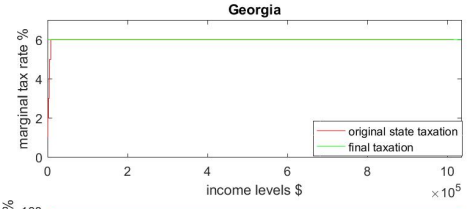

density of observations %
income levels $
×10^5

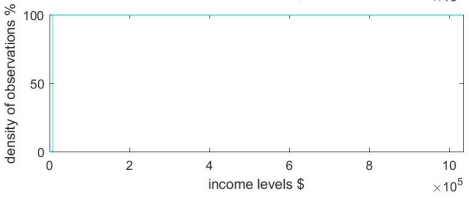

Delaware
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

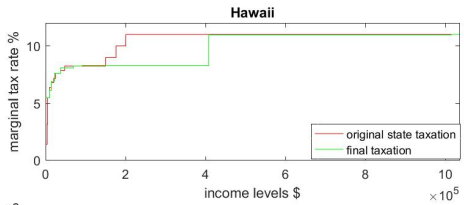

density of observations %
income levels $
×10^5

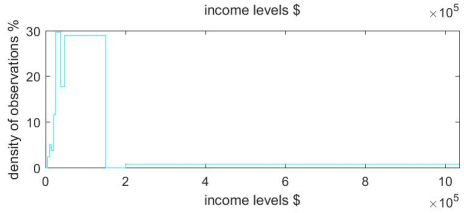

DistrictofColumbia
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

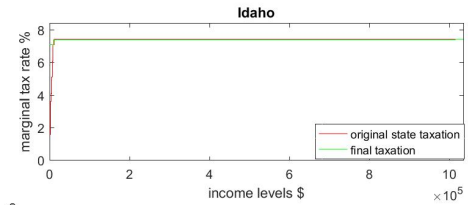

density of observations %
income levels $
×10^5

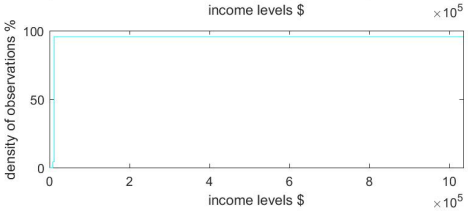

Georgia
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

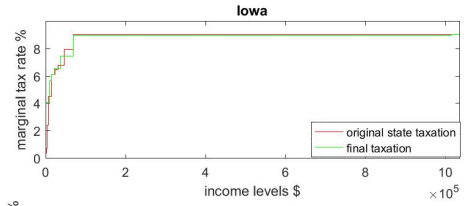

density of observations %
income levels $
×10^5

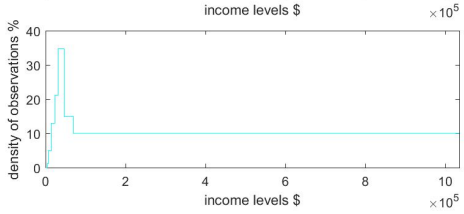

Hawaii
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

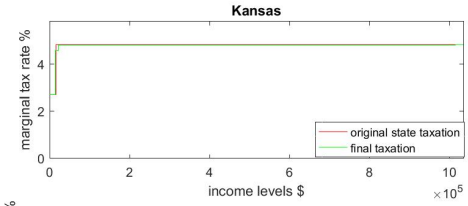

density of observations %
income levels $
×10^5

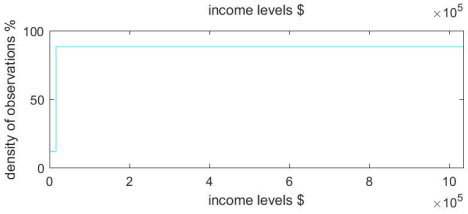

Idaho
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

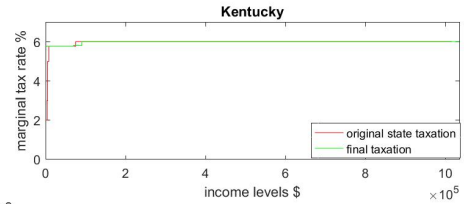

density of observations %
income levels $
×10^5

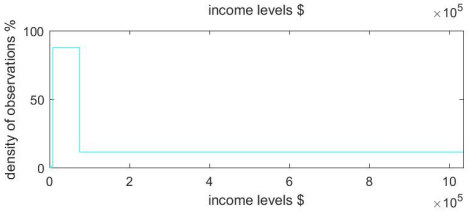

Iowa
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

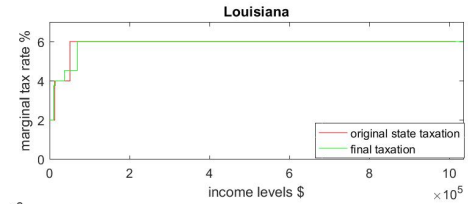

density of observations %
income levels $
×10^5

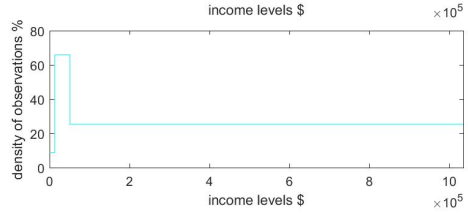

Kansas
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

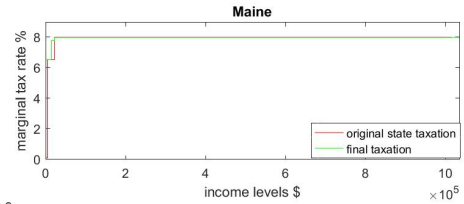

density of observations %
income levels $
×10^5

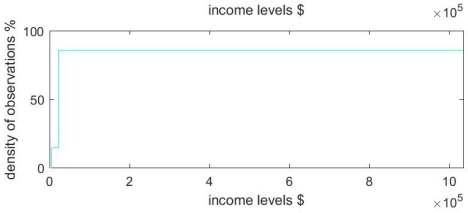

Kentucky
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

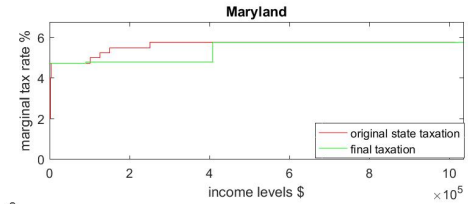

density of observations %
income levels $
×10^5

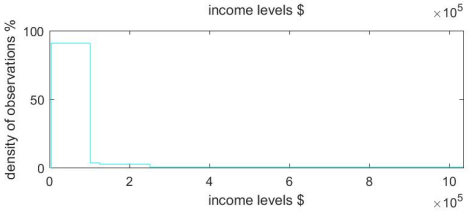

Louisiana
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

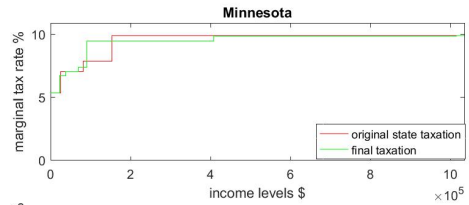

density of observations %
income levels $
×10^5

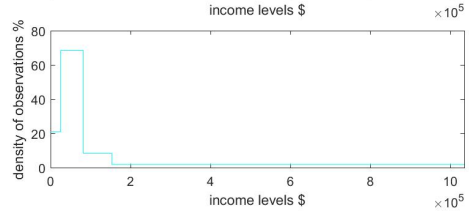

Maine
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

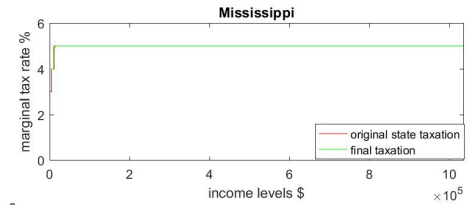

density of observations %
income levels $
×10^5

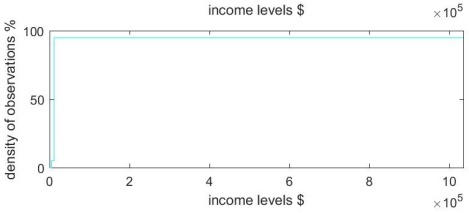

Maryland
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

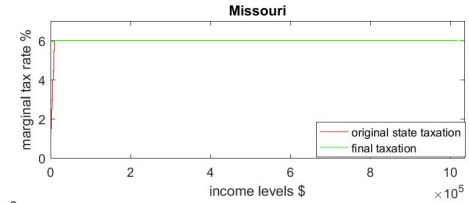

density of observations %
income levels $
×10^5

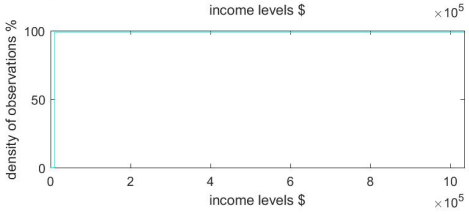

Minnesota
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

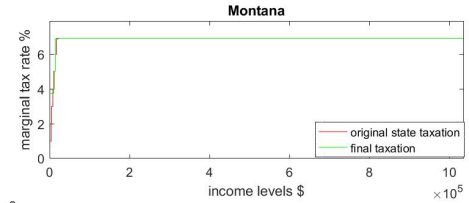

density of observations %
income levels $
×10^5

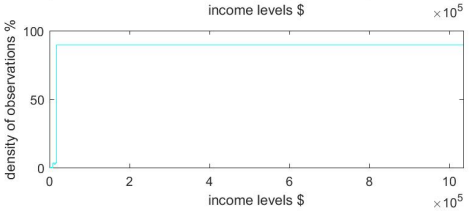

Mississippi
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

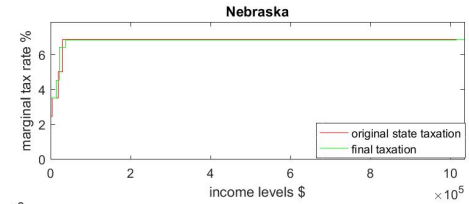

density of observations %
income levels $
×10^5

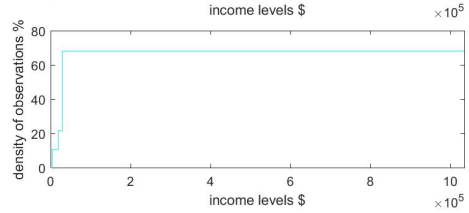

Missouri
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

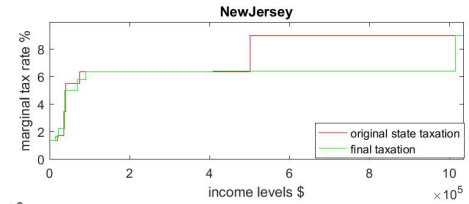

density of observations %
income levels $
×10^5

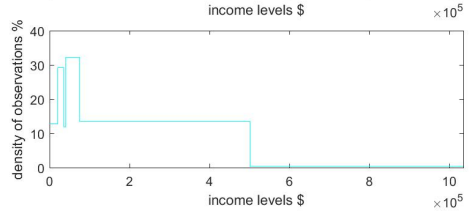

Montana
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

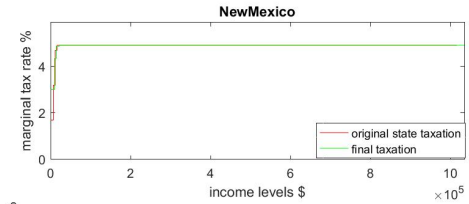

density of observations %
income levels $
×10^5

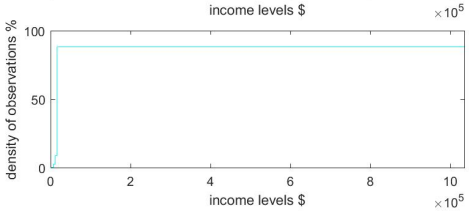

Nebraska
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

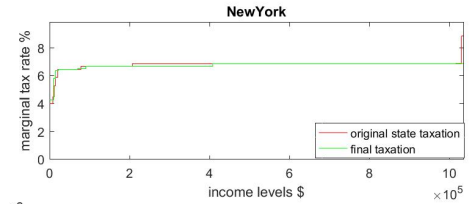

density of observations %
income levels $
×10^5

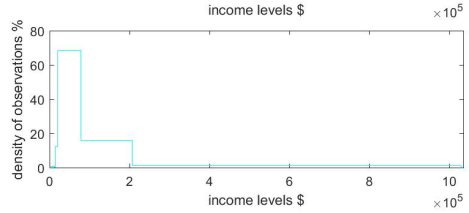

NewJersey
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

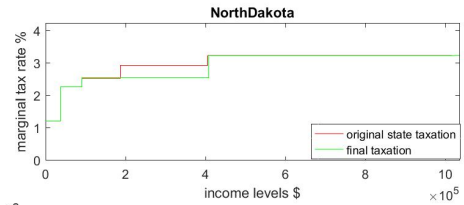

density of observations %
income levels $
×10^5

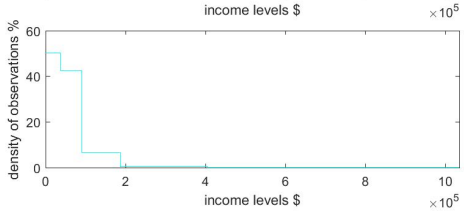

NewMexico
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

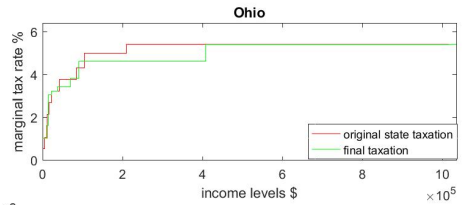

density of observations %
income levels $
×10^5

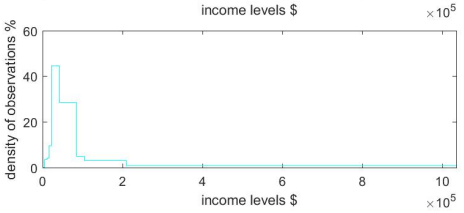

NewYork
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

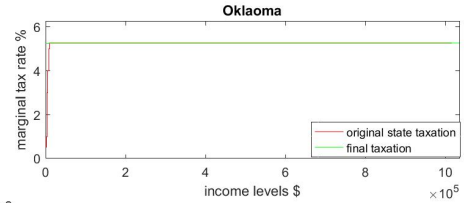

density of observations %
income levels $
×10^5

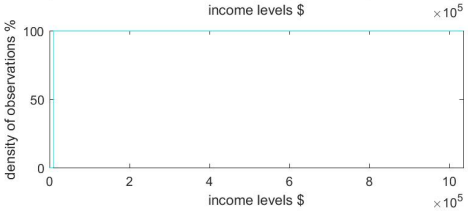

NorthDakota
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

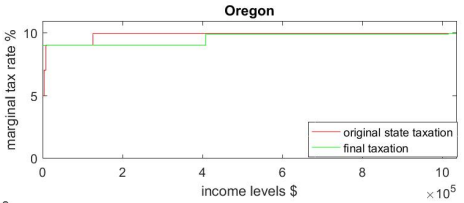

density of observations %
income levels $
×10^5

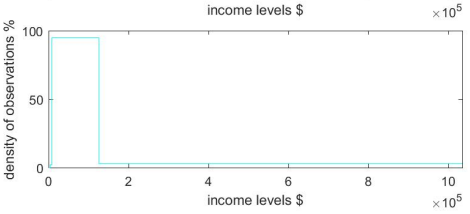

Ohio
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

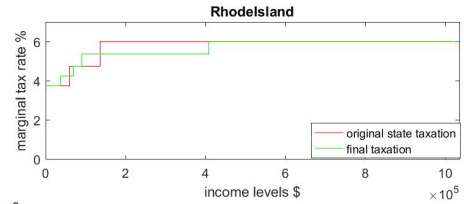

density of observations %
income levels $
×10^5

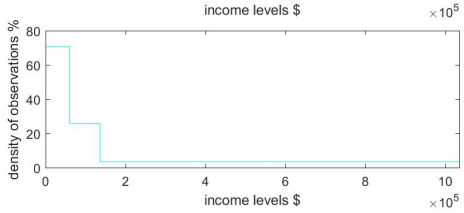

Oklaoma
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

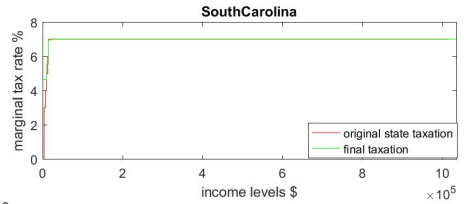

density of observations %
income levels $
×10^5

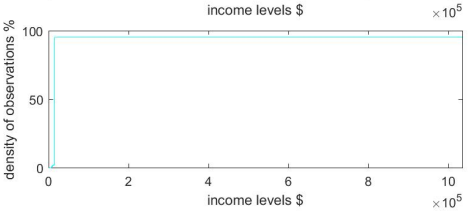

Oregon
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

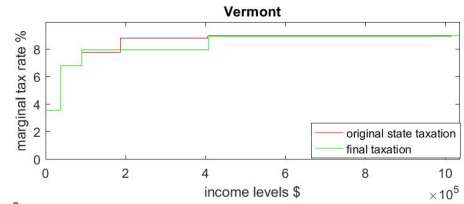

density of observations %
income levels $
×10^5

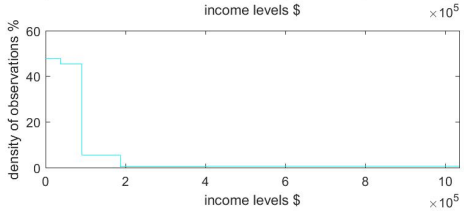

Rhodelsland
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5

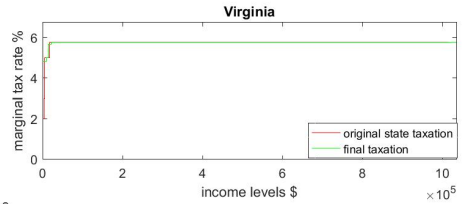

density of observations %
income levels $
×10^5

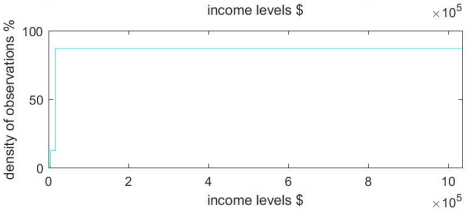

SouthCarolina
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5


density of observations %
income levels $
×10^5


Vermont
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5


density of observations %
income levels $
×10^5


Virginia
marginal tax rate %
original state taxation
final taxation
income levels $
×10^5


density of observations %
income levels $
×10^5

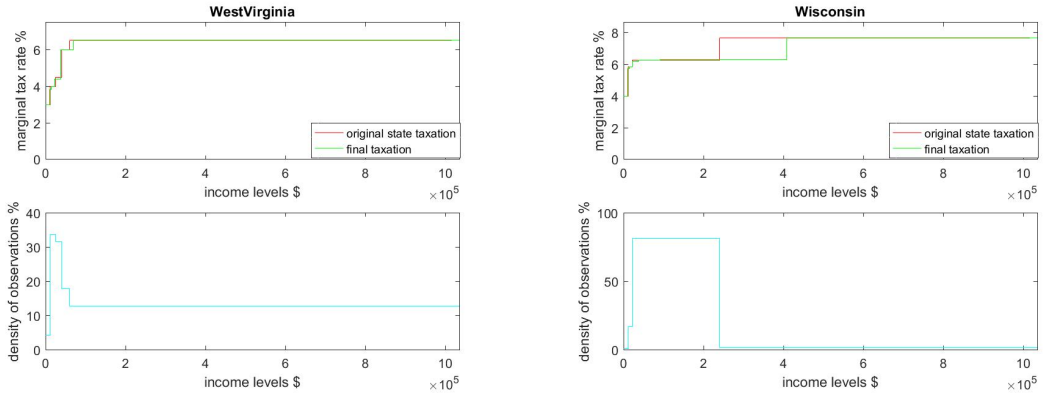


Figure 3: presents for each of the 34 US states the accuracy reached by the final taxation in approximating the state taxations. For each state: we have on the upper graph the marginal tax rates associated with the original state taxation (red) and the marginal tax rates associated with the final taxation (green). In the lower graph, we represent the mass of observations by income levels. It is clear by the graph that our procedure privileges the brackets where the mass of observations is relevant.

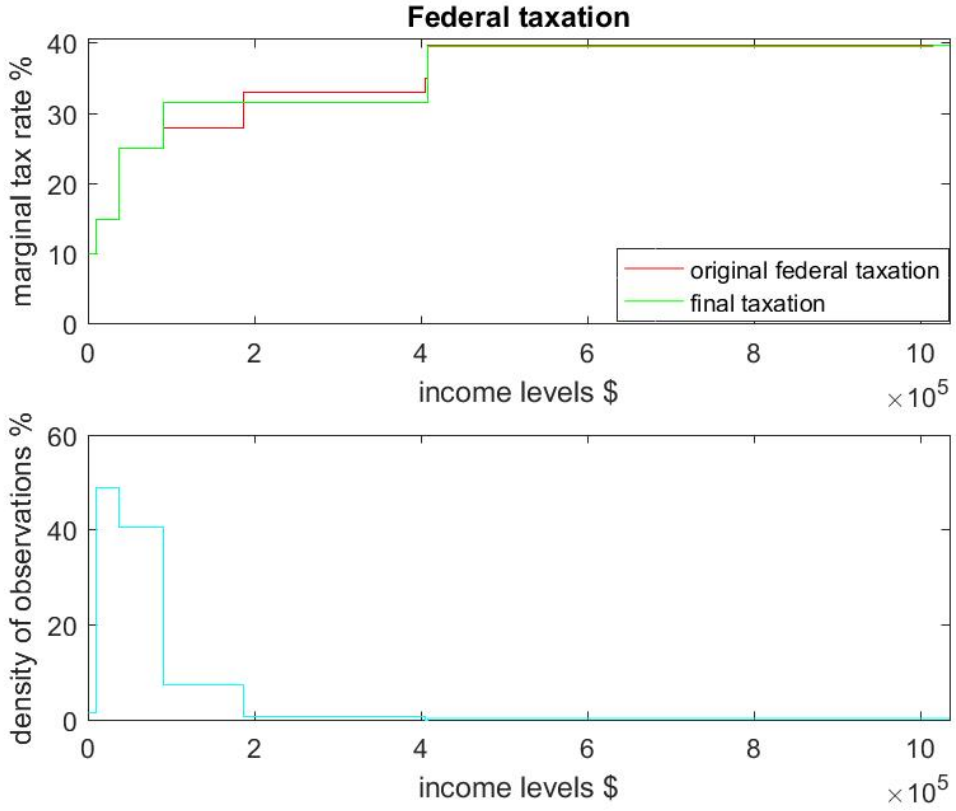


Figure 4: presents the approximation of the federal taxation reached by the final taxation. On the upper graph: in red we have the marginal tax rates associated with the federal taxation while in green we represent the marginal tax rates associated with the final taxation. On the lower graph: we represent the mass of observations by income levels. Our procedure privileges the brackets where the mass of observations is relevant. Indeed in that brackets the marginal tax rates associated with the federal taxation and final taxation are practically indistinguishable

Table 1: presents the summary statistics

	Mean	S.d.	Min	Max
Worker				
age (years)	38.04	13.47	16	64
years of education (years)	13.83	2.07	7	18
Firm				
non-routine cognitive: analytical	0	1	-2.58	2.43
non-routine cognitive: interpersonal	0	1	-2.04	2.91
non-routine manual: physical	0	1	-1.60	2.83
routine cognitive	0	1	-3.22	3.72
routine manual	0	1	-1.69	2.61
offshorability	0	1	-3.07	2.81
yearly wage (in M$)	0.046	0.034	0.003	0.3

Table 2: reports the main and interaction effects of worker and firm features on the formation of their preferences (expressed in M$). The covariates are standardized. In parentheses the standard errors are presented as obtained by bootstrapping our subsample 500 times.

	jobs productivities (main effect)	non-routine cognitive analytic	routine cognitive	routine manual
jobs amenities (main effect)		-0.0079	-0.0010	-0.0025
		(0.001)	(0.001)	(0.0000)
age		-0.0029	0.0005	-0.0004
		(0.0008)	(0.001)	(0.001)
years of education		-0.0003	0.0001	-0.0039
		(0.0008)	(0.001)	(0.001)
age	0.0080	0.011	-0.0006	0.0008
	(0.001)	(0.003)	(0.003)	(0.002)
years of education	0.010	0.023	-0.0024	-0.0197
	(0.001)	(0.002)	(0.002)	(0.003)
age squared	-0.0047			
	(0.001)			
constant	0.1772			
	(0.0042)			

Table 3: presents the 5 edzones identifying the educational attainment and 22 job categories associated to the two digits of the SOC code

Education categories	
edzone 1	less than high school diploma
edzone 2	high school diploma
edzone 3	some college but no degree, associate degree
edzone 4	bachelor degree
edzone 5	master degree and over
Job categories	
1	Management occupations
2	Business and financial operations occupations
3	Computer and mathematical science occupations
4	Architecture and engineering occupations
5	Life, physical, and social science occupations
6	Community and social service occupation
7	Legal occupations
8	Education, training, and library occupations
9	Arts, design, entertainment, sports, and media occupations
10	Healthcare practitioner and technical occupations
11	Healthcare support occupations
12	Protective service occupations
13	Food preparation and serving related occupations
14	Building and grounds cleaning and maintenance occupations
15	Personal care and service occupations
16	Sales and related occupations
17	Office and administrative support occupations
18	Farming, fishing, and forestry occupations
19	Construction and extraction occupations
20	Installation, maintenance, and repair occupations
21	Production occupations
22	Transportation and material moving occupations

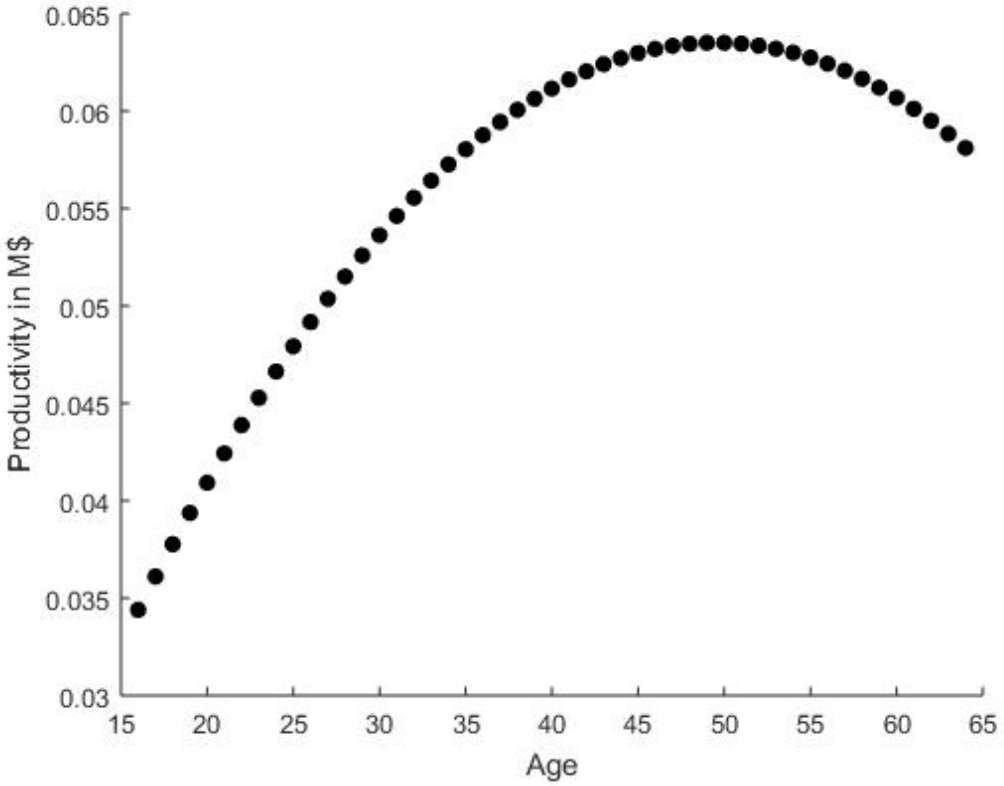
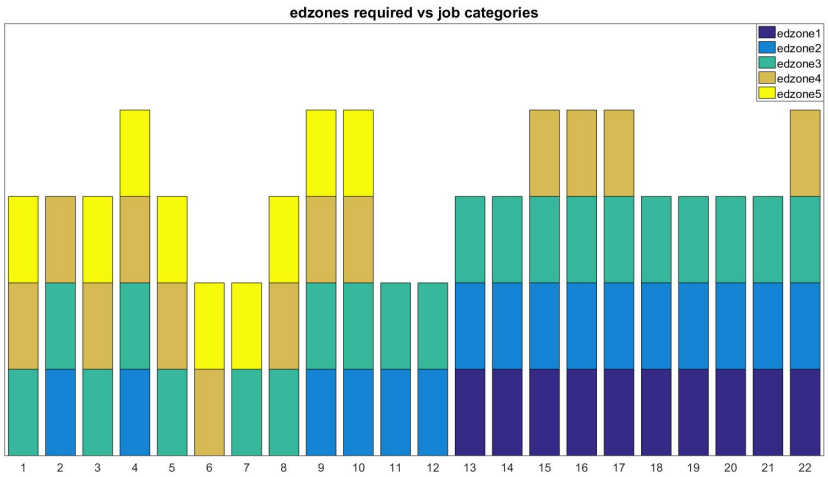
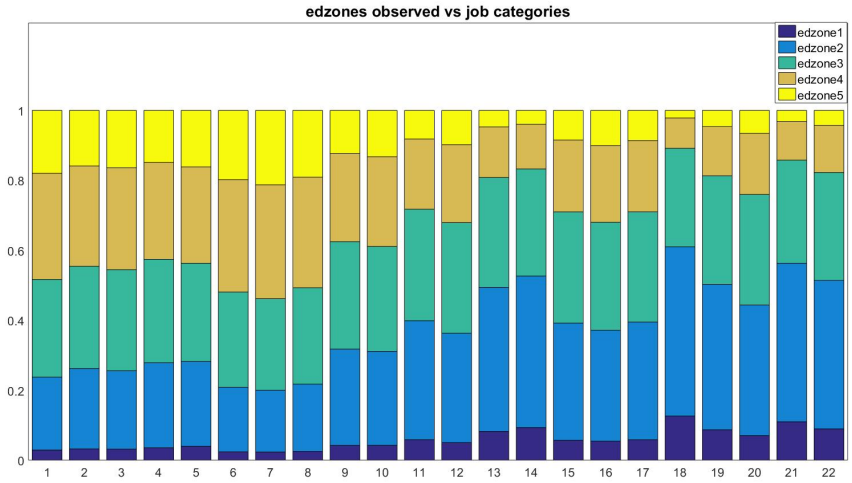


Figure 5: depicts the inverted U-shape between productivity and age

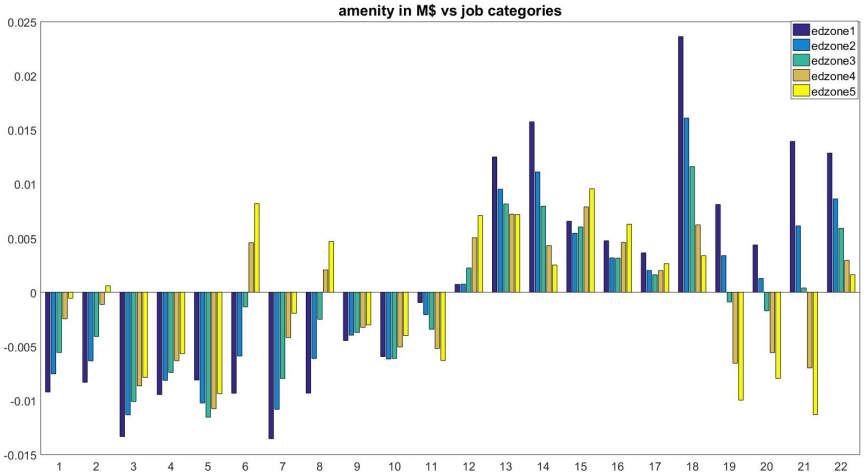


(a)

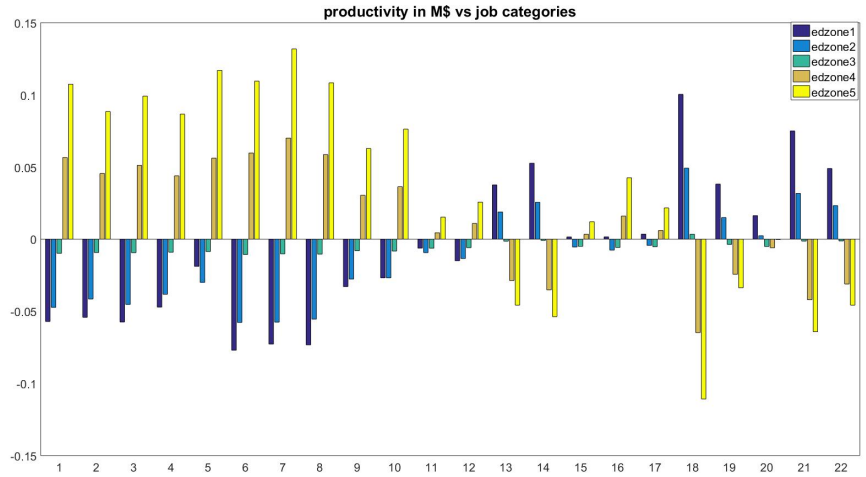


(b)

Figure 6: depicts the required level of education for job categories as derived by O*NET (a) and the observed percentage level of education for job categories (b). The percentage in (b) indicates the severity of the jobs mismatch by identifying the missing levels of education in (a).

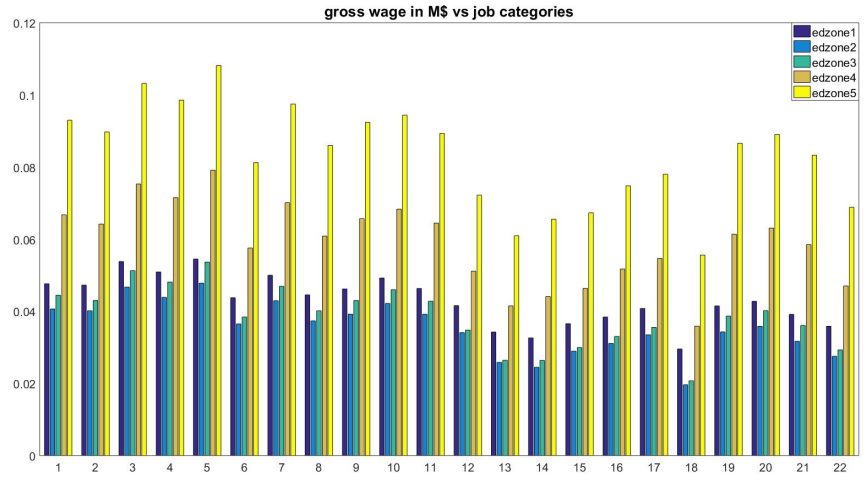


(a)



(b)

Figure 7: depicts the level of amenity by edzone and job category pairs (a) and the level of productivity by edzone and job category pairs (b). On the y-axis the level of amenity expressed in M$ (a) and the level of productivity expressed in M$ (b). On the x-axis the 22 job categories.



(a)

(b)

Figure 8: depicts the gross wage by edzone and job category pairs (a) and the difference between gross wage and after-tax wage by edzone and job categories (b). On the y-axis the gross wage expressed in M$ (a) and the difference expressed in M$ (b). On the x-axis the 22 job categories.

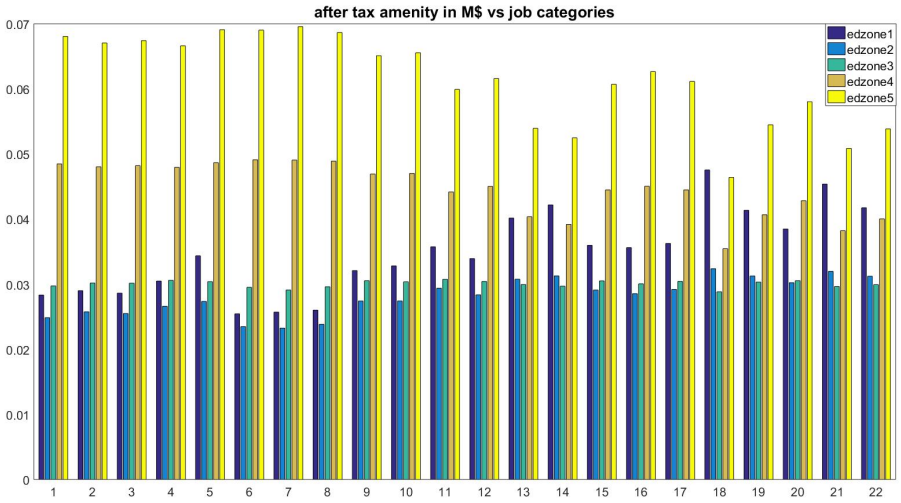


Figure 9: depicts the after-tax level of amenity (it is the combined effect of amenity and after-tax wage) by edzone and job category pairs. On the y-axis the after-tax level of amenity in M$. On the x-axis the 22 job categories.

Chapter 3

US firms innovation: the role of proximities in promoting sponsored research at university

1 Introduction

The relationship between university and firm has sharply changed over time. Historically, most of the research conducted at university was mainly focused on basic research with negligible interest regarding the possibility to commercialize new inventions. Firms from their side were implementing and expanding their internal R&D in order to innovate. Few occasions of knowledge spillover were possible between these two organizations (Santoro and Bierly (2006)). In recent decades, universities have become more aware about the commercial value of their work and the appealing possibility to commercialize their inventions fascinated many researchers. Moreover, universities switched from the *first revolution* where research was added to teaching activities to the *second revolution* where the interaction with firms is recognized to be a crucial step to capitalize the knowledge (Etzkowitz (1998)).

Several reasons enabled university and firm to establish a more close connection (Santoro and Bierly (2006); Rynes et al. (2001)).

From firm perspective, the globalization of the world economy has intensified the competition among companies which in turn has produced an apprehension concerning the financial performances. Firms looked at universities as sources of new valuable ideas with the aim to acquire competitive advantage over rival companies (Abrahamson (1996); Micklethwait and Wooldridge (1996); Pfeffer et al. (2000)). Further, several companies have drastically cut down the size of their R&D departments, creating the necessity to exploit the research resources of the universities (Cohen et al. (1998); Powell and Owen-Smith (1998)). Lastly, economic incentives were provided through public policy to encourage the university-firm collaborations either by providing tax breaks for corporate sponsoring research at university or by creating programs requiring collaboration between industry and firm to get access to the funding (Cohen et al. (1998)).

From university perspective, the need for additional funding created the conditions to search for new partnerships in the private sector. Indeed public funding for higher education as fraction of total revenues as well as federal funding consistently declined over the last decades (Cohen et al. (1998); National Center for Education Statistics (Ed)). Along this funding need, public policy massively promoted with some initiatives the collaboration between university and firm providing incentives for universities generating valuable knowledge for commercial use (Powell and Owen-Smith (1998); Press and Washburn (2017)). The best example supporting this statement is the Bayh-Dole-Act of 1980, enabling university to maintain the intellectual property of the inventions while licensing the inventions to private firms or granting exclusive license for their products. Indeed, several studies confirmed the viability of this act in boosting the number

of formal university-industry linkages (Burack (1999); Lawler et al. (1999)) and the number of patenting and licensing activities of universities (Colyvas et al. (2002)).

Even with these encouraging premises universities and firms belong to two different worlds with different scopes, motivations, values and organizational structures (Bruneel et al. (2010)), so their effective collaboration may be hard to realize despite the economic incentives. According to Alpaydın and Fitjar (2021) spatial proximity as well as non-spatial proximity may play a central role in enhancing the collaboration university-firm. Boschma (2005) identifies diverse types of proximities: geographic, cognitive, organizational, institutional and social. Geographic proximity indicates the spatial closeness of the actors, cognitive proximity hinges on the similarity of the knowledge base of the agents, organizational proximity defines the similarity of norms, practices and incentives of the actors, institutional proximity refers to politically and culturally embedded relationships, social proximity refers to the social relations due to repeated social interactions, friendship and kinship between individuals engaging in the collaborations.

Of the literature assessing the effects and the mechanisms of knowledge transmission occurring between university and firm (Jaffe et al. (1989); Nelson (1986)), the number of studies attempting to infer the role of the different types of proximities in shaping that collaboration is quite limited (D'Este et al. (2012); Garcia et al. (2015); Molina-Morales et al. (2014)). Furthermore the economic literature mainly proposed the use of empirical models to disentangle the effect of the different types of proximities in the formation of university-industry research collaborations.

D'Este et al. (2012) studied the formation of the university-firm collaboration by exploring the role of geographic and organizational proximity using a dataset from the UK Engineering and Physical Sciences Research Council (EPSRC) over the period 1999-2003. Geographic proximity is measured as the inverse of the square root of the distance between university and firm while the organizational proximity is measured by the number of prior partnerships university-firm observed in the 1999-2002 period. They employed a logit estimation to infer the likelihood of research partnerships, finding out that geographic and organizational proximity make the partnerships more likely.

Garcia et al. (2015) looked at the collaborations occurring between universities and firms in Brazil by assessing the link between geographic and cognitive proximity. Geographic proximity is measured as straight-line distance between research group and firm while the cognitive proximity is identified as the angular separation between scientific field and industry sector computed through the cosine index. They found out by applying an ordinary least square estimation that cognitive proximity may substitute geographic proximity in driving the partnership university-firm.

Further studies have also proposed alternative measures of cognitive proximity. In the context

of knowledge spillover between firms, Jaffe (1989) introduced a measure of cognitive proximity defined by the technological distance between industries. The technological closeness of two firms is retrieved from their patent activities. Therefore firms presenting similar patent portfolios would receive a value for their technological relatedness equal to 1, otherwise they would receive 0. Molina-Morales et al. (2014) studying the impact of geographic and cognitive proximity in explaining innovation performance within the footwear industry of Spain suggested a measure of cognitive proximity reflecting common goals and culture. They presented a two-item scale definition based on Simonin (1999).

Along this literature, our paper proposes to assess the interplay between geographic and cognitive proximity in shaping the formation of university department-firm collaboration by introducing a structural model, and to the best of our knowledge this represents the first attempt. Furthermore we measure geographic proximity as in D'Este et al. (2012) while we adopt the approach proposed in Garcia et al. (2015) concerning the measurement of cognitive proximity. The latter choice hinges on the assumption that the pattern of collaborations university department-firm is not randomly distributed, namely firms facing specific innovation issues would naturally tend to collaborate with research groups of certain scientific fields exhibiting a specific set of capabilities and knowledge. Therefore the measurement of cognitive proximity proposed in Garcia et al. (2015) seems appropriate since it is based on the application of the correspondence analysis to the contingency table, which it is determined by counting the joint occurrence of the collaborations between scientific fields and industrial sectors.

Our model builds on Dupuy et al. (2017) by proposing a matching framework with imperfect transferable utility and logit heterogeneity in preferences. This type of model enables to understand the decision process of both the agents within the studied market. Indeed, from one side the university department defined by a set of observable attributes as well as unobserved preferences decide with which firm wants to collaborate. We assume that the location of the university department is exogenous given their long-standing tradition. On the other side the firm characterized by a set of observables attributes as well as unobserved preferences decides where to locate and with which university department wants to collaborate. The location of the firm is determined endogenously.

Furthermore, the flexibility of the model results in a two-fold advantage: when the matching is observed it would allow to measure the cognitive proximity *ex-ante* and to estimate the optimal level of preference associated to it, and even more relevant when lacking the observed matching it would allow to measure the cognitive proximity *ex-post* by simulating a counterfactual matching. The observed or simulated matching (counterfactual) represents nothing but the joint occurrence

of the university department-firm pairs. Therefore it readily identifies the contingency table employed as starting point to measure the cognitive proximity.

We test the model for the US by creating a dataset *ad hoc*, given that we do not have access to a random sample of matching of firms and university departments. Therefore we first proceed by deriving a sample of firms from Compustat and a sample of university departments from the National Science Foundation (NSF), and finally we merge the two datasets. The last step represents a convenient but not necessary way to structure the working dataset and certainly it is not intended as one-to-one observed matching of university departments and firms. Indeed we simulate the counterfactual matchings by proposing two scenarios: in the first we vary the level of preference attached to the geographic proximity (for which it is possible to compute the measure *ex-ante*) and in the second we indirectly impact the cognitive proximity by varying the level of preference associated to the interaction of R&D intensity with the quality of university department, assuming that the complementarity of these two variables represents a good indicator of the cognitive similarity. In each scenario, the simulated matchings are then deployed to measure the variation of cognitive and geographic proximity by university department-industrial sector pairs.

The remainder of the chapter is organized as follows: Section 2 provides an overview of the types of collaboration; Section 3 introduces the types of proximity and their measurements; Section 4 presents the sponsored research agreement as crucial mechanism for firm and university department to collaborate; Section 5 develops the economic model; Section 6 discloses the estimation strategy; Section 7 describes the observables of firm and university department; Section 8 presents the dataset; Section 9 describes the simulations and their relative results; Section 10 discusses the findings; Section 11 concludes.

2 Types of collaboration

There are several ways the collaboration between university and firm can manifest.

Generally, social interactions and network effect constitute the primary and easier way for firm to establish a connection with researchers and this may eventually lead to a formal knowledge exchange (Lee (2019)). According to Bozeman et al. (2015) *much of the scientific and technical human capital is embedded in social and professional networks of technological communities.* Then, it is not a surprise that researchers having a remarkable social capital are more involved with industry (Perkmann et al. (2013)).

Networks can be highly beneficial in supporting the interaction university-firm and provide

introductions that can potentially evolve in consulting engagements (Lee (2019)). The latter represents the most adopted way of transferring the knowledge from university to a licensing firm (Argyres and Liebeskind (1998)). Empirical analyses support the evidence that the non-direct engagement of the faculty inventors is the cause of the 18% of commercialization failure (Thursby and Thursby (2004)). This is confirmed by Agrawal (2001) analyzing the licenses of three departments of the MIT (Mechanical Engineering, Electrical Engineering and computer Science) where he found out that the involvement of the inventors increases the likelihood of the commercialization and the amount of royalties generated.

Sponsored research represents another mechanism adopted when firm and university decide to collaborate. This kind of interaction has been favored through the decline of the federal funding and most of the activity is focused in improving the transmission of the knowledge that provides value for the sponsoring firm (Santoro and Bierly (2006)). Markman et al. (2005), using a sample made by 128 research universities in US, provided evidence that 11% of the university license strategy is associated with sponsored research. Further, about one third of the university licenses require compensation by the licensee in the form of sponsored research (Jensen and Thursby (2001)).

Lastly, the university spinoff is an ulterior mechanism widely adopted to transfer knowledge. The faculty inventor assigns her patents to the university which in turn licenses to a startup created by faculty inventor. Due to the number of failures occurring when commercializing the product without the direct participation of the inventors, the university spinoff is the most suitable way to realize new products requiring cutting edge technology (Di Gregorio and Shane (2003);Lowe (2006)). Empirical analysis has confirmed that the direct involvement of the faculty inventor is likely to increase the success of developing the new technology (Knockaert et al. (2011)).

Although recognizing the multiple strategies and opportunities available when firm and university decide to collaborate, many attempts to transmit the knowledge from university to firm in order to commercialize new products seem to be unsuccessful (Santoro and Bierly (2006)).

3 Types of proximity

With these premises, the role of the universities and researchers appears crucial in order to bridge the knowledge gap between university and firm and fosters innovation.

Robert (1988) pointed out that better human capital can raise the ability to develop new technologies and the benefit for the whole economy is undeniable. Skilled workers are essential

for promoting growth and innovations (Bartel and Lichtenberg (1987); Wozniak (1987)).

The better use of skilled workers resides in the research activity, capable of generating valuable ideas and in better implementing the cooperation with firms, capable of concretely realize these ideas. Nelson (1986) found out that research at university is the main source of technological innovation in private industry. This result is reached by direct cooperation (Cox (1985)) or by spillover effect (Bernstein and Nadiri (1988)).

Indeed Jaffe et al. (1989), adopting the production function model developed by Griliches et al. (1979), found out that innovation identified by the number of patents in private industry is positively correlated with the amount of research conducted at local universities. The number of patents is used as proxy for innovation while private corporate expenditures and university research expenditures are used as explanatory variables. Jaffe concluded that the knowledge spillover is highly localized at the state level, namely the mechanism of transmission of the knowledge is effective when firms and universities are co-located. Subsequently, the same finding is confirmed at regional level (Jaffe et al. (1993)). Similarly, Mansfield (Mansfield (1991, 1995); Mansfield and Lee (1996)) found out that the knowledge transmission is highly dependent from the geographical proximity between universities and firms. The findings show that the firms exhibit a preference to collaborate with universities within a distance of 100 miles while show a non-negligible reluctance to collaborate with universities at distance larger than 1000 miles. Recently Adams (2002) using the result of a 1997 survey of 208 private R&D laboratories in US studied the role of geographical proximity in acquiring the knowledge from universities and from other firms. The research provides evidence that the closeness between universities and firms matter far more than the adjacency of firms. Additionally, despite the increment of communication technologies, such as fiber optics, social networks, and satellite and in parallel the decrease in communications and transport costs that should have facilitated communications, geographical distance seems still matter in creating collaborations and spreading the knowledge. In this respect, Keller and Yeaple (2009) have provided evidence why face-to-face interaction is superior over communication like telephone calls and e-mails.

The reason why the geographical proximity matters in the knowledge spillover even if research results are publicly available ('public good') resides in the tacit knowledge. It is the part of the knowledge not codified which represents a crucial step to understand the research results. Indeed this aspect explains the important role of the geographical proximity in promoting the collaboration between university and industry (Gertler (2007)). Interactive learning processes including personal interactions and face-to-face contacts underline the importance of the proximity in bringing the benefits of collaboration to the firms (Abramovsky et al. (2007); Arundel and

Geuna (2004); D'Este et al. (2012); Muscio (2013)).

Although recognizing the relevance of the geographical proximity, other type of proximity may affect the decision of the university and firm to collaborate. Indeed, among them the cognitive proximity play an important role (Boschma (2005)). Within an organization, cognitive proximity can manifest through employee using the same language, scientific standards and technological formal codes (Wink (2008)). When organizations decide to interact with each others, cognitive proximity is perceived as similarity of knowledge, competencies and technological aspects and it is generally associated with technological proximity (Knoben and Oerlemans (2006)). In a broader sense, cognitive proximity is expressed as the homogeneity of competencies, capabilities, skills and knowledge bases possessed by independent organizations (Hautala (2011)). It is clear that cognitive proximity plays an important role to increase the efficiency of the knowledge transfer between academic research and firm. The tacit nature of the new knowledge can be better absorb by firm if university and firm share a common cognitive base that can ease the communication process (Boschma and Lambooy (1999)). In order for the cooperation to be fruitful it is essential that the cognitive distance between firm and university should be either not too close which would result in a lack of novelty nor too far which would prevent an efficient communication and an effective transmission of the knowledge (Nooteboom (2000)).

The cognitive proximity is highly related to the absorptive capacity of the firm, namely the ability to identify, digest and exploit the new knowledge (Cohen and Levinthal (1990)). Indeed firms presenting high absorptive capacity tend to better use the collaboration with universities (Boschma and Ter Wal (2007)) and in order to remain at the forefront of research they might need to cooperate with non-local universities, the latter providing a broad set of academic capabilities crucial to figure out innovative solutions (Bishop et al. (2011)). By contrast, firms with low absorptive capacity tend to rely more on geographical proximity and less on the quality of the universities. Clearly, the absorptive capacity of the firm becomes more important when the quality of the local universities is unable to meet the needs of the firm (Bishop et al. (2011)). The superior capacity of the firm to incorporate knowledge and to coordinate activities when interacting with non-local universities is crucial (Laursen et al. (2011)). Additionally, the cognitive proximity appears related to firm's size and industry sector. In principle, large firms possess the suitable capacities to undertake several academic collaborations (Fritsch and Lukas (2001)) and to successfully manage long-term distant interactions (Levy et al. (2009)) while small and medium firms tend to rely heavily in collaborations with local universities (Muscio (2013)). According to Wink (2008), cognitive proximity is more important than geographical proximity regarding knowledge integration for organizations operating in science-driven sectors.

3.1 Measurement of cognitive proximity

The construction of the cognitive proximity has been drawn from Garcia et al. (2015). We therefore start by creating the contingency table made by two categorical variables: one containing the university departments and the other containing the industrial sectors. The contingency table identifies the distribution of collaborations between each university department and industrial sector pair, determined by counting the joint occurrence of all possible combinations of pairs. Afterwards a correspondence analysis is conducted evidencing the association between the two categorical variables using a geometric approach. Indeed the methodology enables to associate the university department-firm pairs by the mean of coordinates in a low dimensional space. The dimension of this space is retrieved by the variance explained through the singular value decomposition. Finally, we got a cosine index (or cosine similarity) by computing the inner product of the coordinates of each university department and industrial sector pair. In formula we have

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \tag{1}$$

where A_i and B_i represent the coordinates for a university department-industrial sector pair, and n the dimensionality of the space.

The cosine index owns values between -1 and 1. Values approaching -1 represent strong negative association between the university department and the industrial sector. Values approaching 0 indicate independence between the university department and the industrial sector. Values close to 1 identify strong positive association between the university department and the industrial sector.

3.2 Measurement of geographic proximity

The geographical proximity is the inverse of the square root of the distance expressed in km between university department and firm as originally proposed in D'Este et al. (2012). The distance is provided by the straight-line joining the geographical coordinates expressed as latitude and longitude of the university department and firm. The distances (between the university departments and each firm) have been measured using the ArcGIS software.

4 Sponsored research and its taxation

4.1 Sponsored research agreement

In the context of the partnerships between university and firm, the sponsored research financed by corporate in the form of the sponsored research agreement (SRA) plays a fundamental role in performing advanced research in key technological areas.

The SRA mainly exists in two versions:

- Fixed price contract requires that the underlying payment intended in the contract for the costs incurred by the university in order to perform the research project is fixed (lump sum payment) and it cannot be subject to adjustments. This type of contract is carefully used only if the costs of the project are readily and easily definable.

- Cost reimbursement contract stipulates that the costs potentially incurred by the university during the research project are roughly estimated, establishing a ceiling that the university should not exceed. However an upsurge of the costs is tolerated whether it is consistent with research goals.

The type of costs covered in the research agreement are respectively the direct costs and indirect costs. The direct costs are determined with great accuracy within the project, i.e. the salaries of the faculty members involved in the projects, the travel expenses, the scientific equipments. Instead the indirect costs cannot be allocable to one project directly, i.e the salaries of the administrative staff, the office equipments, the maintenance of the general purpose equipment and facilities. They are computed as percent of the direct costs, generally around 40 to 50 percent of direct costs.

In these types of agreement two relevant aspects needs to be mentioned.

Firstly, the university engaged to conduct the research in reasonable best effort basis to provide successful results. Therefore due to the unpredictable nature of any experimental research the company implicitly accepts the uncertainty associated with the project. Furthermore the company cannot advance any claim under the form of guarantee or warranty regarding the results of the project. Hence, the firm is not entitled to impose penalties concerning the university failure to make progress as well as it cannot withhold the payment to the university whether it is not satisfied by the results.

Secondly, university and firm bargain over the intellectual property (IP) of the research results and their applicability through the license policy for profit purpose. Generally, the university is keeping the IP of the results with the right to publish them while allowing the company to

review the future publication during a minimum of three months to a maximum of six months to ensure that the confidential information and IP is sufficiently protected. In return the university is offering to the company a non-exclusive royalty-free license for commercial purpose meaning that the firm does not need to pay any royalties for the use of the research output for creating the new product. The non-exclusivity character of the license policy means that the the licensor (university) grants the licensee (company) the right to use the intellectual property of the research results, but the licensor it is free to exploit the same intellectual property. The rationale behind this agreement is to prevent the scenario in which the university may risk future infringement by allowing the university researchers to continue the study on the subject matter of the invention.

4.2 Qualified Research Activity

The advantage of the company engaged in research projects with the university is twofold: the company benefits from the expertise of university in a specific domain with the intent to remain competitive while it can acquire a tax credit if the partnership fulfills the appropriate requirements. The R&D tax credit is a mean used to stimulate private companies to produce more ideas and technologies by substantially decreasing the cost of these activities (Fichtner and Michel (2015)). Indeed several studies showed the positive effect of the tax incentive in increasing the R&D spending of private firms. According to Bloom et al. (2002) the credit stimulates $1.10 of research for every dollar of tax revenue. Another study found out that the R&D tax credit induces $2.96 of additional R&D investment for every tax dollar spent (Klassen et al. (2004)).

The R&D tax credit is a general corporate tax credit under the Internal Revenue Code (IRC) Section 41 granted by companies that are subject to research and development costs. It has been introduced in 1981 within the Economic Recovery Tax Act with the intent to encourage research investment in United States. The purpose was to stimulate the economic growth of the country and keep pace of the global competitiveness. Since the first expiration in 1985, the R&D tax credit has been extended fifteen times until the Obama's administration which made the tax credit as a permanent expenditure of the government bill.

In our case, the company may claim R&D tax credit if the partnership with university qualifies as qualified research activity (QRA) as expressed in the IRC section 41(d). The conditions that the SRA needs to satisfy to be identified as QRA are the followings:

- the research activity hinges on hard science, such as engineering, computer science, biological science, or physical science principles
- the research activity leads to the development of a new or improved business component,

defined as new or improved products, processes, internal use computer software, techniques, formulas, or inventions to be sold or used in the taxpayer's trade or business

- the research activity relies on a process of experimentation with the finality to test and assess alternative solutions to eliminate technological uncertainty

Based on the SRA, it is clear that the partnerships between university and firm may fall in the QRA. Within the QRA, the tax credit may be granted to the company for the following qualified research expenses (QRE) as expressed by the IRC section 41(b):

- wages paid to employees for qualified services (including amounts considered to be wages for federal income tax withholding purposes)
- supplies (defined as any tangible property other than land or improvements to land, and property subject to depreciation) used and consumed in the R&D process
- contract research expenses paid to a third party for performing QRAs on behalf of the taxpayer, regardless of the success of the research, allowed at 65 percent of the actual cost incurred
- basic research payments made to qualified educational institutions and various scientific research organizations, allowed at 75 percent of the actual cost incurred.

The next step is to understand if within the SRA the company can claim the tax credit.

We exclude the first two options because these are related to in-house research. Furthermore we exclude the basic research option because we are interested in the market of innovation where the company engages in research partnerships for a commercial purpose. Then, according to the definition provided by the National Science Fundation (NSF) *basic research is defined as experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts.* This would not fit the objective of the SRA.

We focus our attention on the third condition to be QRE. We found out that the SRA needs to fulfill the following three conditions to qualify as contract research expenses:

- the agreement is formulated before the performance of the qualified research
- the agreement must guarantee that research is performed on behalf of the company. This essentially means that the company has substantial rights over research results, even if it's not an exclusive right.
- the agreement must require that the company bears the expenses of the research even if it is not successful

The intent of the R&D tax credit is clear: the company is rewarded if only if it accepts to bear the overall risk involved in the research project. Each agreement that it is contingent to the success of the research cannot qualify as QRE, and none claim concerning tax credit can be granted to the company.

It appears clear by the description provided in the previous section that the SRA may be identified as contract research expenses. In that case the company may benefit of the R&D tax credit within the limit of 65 percent of the overall costs incurred during the research partnership.

4.3 Firm taxation under QRA

According to the IRS the computation of the tax credit granted by the company is based on the choice of two main methodologies:

- Regular tax credit : the credit rate is applied to the amount of the QREs exceeding the base amount. The latter is calculated at first by defining the fixed base percentage as the ratio of the QREs and the gross receipts for the period 1984-1988, and it is capped at 16 percent. Afterwards, the base amount is derived by multiplying the base percentage to the average annual gross receipts for the four year preceding the credit year. As reported by the IRS the base amount cannot be less than 50 percent of the current QREs. Hence, the greater between the two serves as the base amount. Lastly, the amount of the current QREs exceeding the base amount is multiplied by 13 percent before 2017 and 15.8 percent after 2017.

- Alternative simplified credit (ASC): the credit rate is applied to the amount of the QREs exceeding the base amount. The latter is the 50 percent average QREs of the previous three years relative to the current credit year. The amount of the current QREs exceeding the base amount is multiplied by 9.1 percent before 2017 and 11.06 percent after 2017. In the case there is no previous QREs in any one of the previous three years, the credit equals 6 percent of the QREs of the current year.

According to Fichtner and Michel (2015), the regular tax credit was introduced with fixed base with the aim to produce a marginal tax incentive. Nonetheless the Government of Accountability Office (GAO) realized that *over time, the historically fixed base of the regular credit becomes a very poor measure of the research spending that taxpayers would have done anyway. As a result, the benefits and incentives provided by the credit become allocated arbitrarily and inequitably across taxpayers, likely causing inefficiencies in resource allocation.* This represents the main reason why the ASC has been introduced in 2006 through the Tax Relief and Health Care Act.

Furthermore, the company preferred the ASC to the traditional research credit computation for three main reasons beside its evident simplicity: incomplete or complicated records for computing their base amount, relatively large base amount, or recent large growth in total receipts.

The information concerning the R&D spending, including the annual QREs of the firm, the base amount, and the reduced credit are available in the restricted-access IRS Statistics of Income (SOI) dataset and they are retrieved from the Form 1120 of the IRS concerning the firms' basic tax return and Form 6765 of the IRS concerning firms' R&D spending.

Hence, the accuracy of the SOI dataset enables to compute the actual credit rates that the companies face on their marginal dollar of R&D spending (Rao (2016)), independently of the methodologies used to compute the credit rates. The tables 1 and 2 illustrate the computations of the reduced credit using the two methodologies introduced above before and after the 2017 Tax Cuts and Jobs Act (TCJA) which reduced the corporate tax rate from 35 percent to 21 percent. In particular, the 65 percent appearing in the tables then changed in 79 percent after the TCJA originates from the intent of the US government to avoid that corporate may receive a double benefit from a tax credit and a tax deduction (the taxable corporate profits are equal to a corporation's receipts less allowable deductions) for the same research expenditures. In particular under section 174 of the IRC, the tax law gives the company the right to deduct research and experimental expenses (R&E). Nonetheless, the claim should be reduced by the amount of claimed R&D credit. However under section 280C(c) of the IRC, the company may reduce directly the R&D credit without adjusting the deductions claimed under section 174. Generally, the company cuts the claimed R&D credit by a proportion equal to the maximum statutory corporate tax rate (35 percent before 2017 or 21 percent after 2017)(Fichtner and Michel (2015)).

Unfortunately, the access to confidential information provided by the SOI is limited. However we can obtain an average credit rate (as percent per QREs $) by using the statistics of the SOI publicly available concerning corporations credit for increasing research activities retrieved from the Form 6765 of the IRS. Indeed SOI statistics report the QREs (including the wages for qualified services, the cost of supplies, the rental or lease costs of computers, the applicable percentage of contract research expenses), the base amount and the credit under regular and ASC computations with two main restrictions: the information is available at the aggregate level of the business sector (the first two digits of the NAICS code or for sub-sectors of the manufacturing sector) and the most recent year is the 2013. For each business sector we compute the credit rate as the ratio of the credit to the QREs and we average it by using the weights provided by the amount of QREs pertaining respectively to the regular and ASC computations. Afterwards the average credit

rate by business sector is multiplied by the applicable percentage of contract research expenses (65 percent as mentioned in the definition of QREs in section 4.2). This last step would supply a flat credit rate readily applicable to the individual firm (given the business sector) in order to quantify the credit the company is entitled to receive in each $ spent in sponsored research. Table 3 computes the flat credit rate for illustrative purpose of the food manufacturing sector. The same procedure is adopted to compute the flat credit rates for the other business sectors by employing the SOI public statistics concerning corporations credit for increasing research activities. Table 4 presents the flat credit rates by sub-sectors of the manufacturing sector.

4.4 University taxation under QRA

Under section 501(c)(3) of the IRS code, the university is non-profit organization whose principal aim is to educate the future generations, promote health and conduct scientific research. Moreover universities and colleges are tax-exempt entities under federal and state law. Nonetheless in recent times the university has shown a growing tendency to embark on sponsored research for commercial ends, likely generating an income stream.

Within that context it is crucial to assess whether the project sponsored by a business entity may qualify under the university's tax-exempt scheme or is subject to the Unrelated Business Income Tax (UBIT). For that reason the IRS has developed specific guidelines suggesting that commercial and industrial operations can no longer be qualified as research.

However the university does not need to pay taxes if the project is conducted in such a way that the university enforces its original mission. Hence, the project:

- is developed and supervised by professionals in order to solve a problem applying a scientific method through hypothesis, design, testing and data analysis
- provides additional knowledge
- involves the development of new ideas, skills, methods
- is conducted in the public interest and the results are made available to the public
- involves students and trainees actively participating to the project with specific tasks and duties

It seems clear that each project undertaken by the university with corporate labeled as QRA exempts university from paying taxes on the amount received.

5 Economic model

5.1 Firm location

The existence and diffusion of the knowledge have been widely studied by Jaffe et al. (1989) and Jaffe et al. (1993) whose reached the conclusion that the knowledge flow tend to be highly localized in space. A key limitation of that approach hinges on the assumption that the location of the firm is exogenous (Alcácer and Chung (2007)). In our paper, we overcome this limitation allowing the location of the firm to adjust endogenously in the model. We based our decision on two main reasons. Firstly, firms tend to maximize the knowledge spillover and the absorptive capacity of the firm would differentiate the companies based on their abilities to identify and digest new knowledge. Secondly, firms seek competitive advantage over rivals reducing whenever possible the knowledge leakage (Myles Shaver and Flyer (2000)). Moreover the firm would act strategically choosing locations which maximize the occurrence of knowledge spillover while minimizing its role as source of knowledge to competitors. In that context Alcácer and Chung (2007) using data regarding new entrants in US from 1985-1994 examined the location choices of the firms. They found out that *laggard firms* (defined as less technologically advanced) tend to locate in areas with any level of academic activity and high industrial activity which produces easier-to-digest commercial knowledge while *leader firms* (defined as more technologically advanced) would privilege locations with high levels of academic activity and discard economic area independently of the level of the industrial activity. The main idea conveyed in that paper is the following: co-location may not be beneficial for all firms equally, and due to the heterogeneity in technological abilities some firms would be more equipped to receive knowledge, deciding to locate far away from competitors in order to exploit their competitive advantage.

In line with this literature we introduce the location of the firm as an endogenous choice.

5.2 Set up

We develop a matching model with imperfect transferable utility. In that context the friction when firm is making a transfer to the university department is due to the research tax credit that the company may retain by engaging in QRA. We explicitly attribute the costs supported by the company within the SRA underwritten with the university department as contract research expenses. Moreover, the firm is entitled to claim credit on the 65 percent of the overall transfer. For each university department-firm pairs we identify the amount of friction by multiplying a flat credit rate identified by business sector to the transfer. The flat credit rate (introduced as $\tau(y)$ in section 5.4) has been derived by retrieving the average credit rate by business sector from the

SOI public statistics concerning corporation research credit and then by multiplying it to the applicable percentage of contract research expenses claimable.

Further we assume as given the location of the university departments while we assume as endogenous to the model the location of the firms. We introduce $\boldsymbol{Z}$ a finite set of locations, where each location identifies a market of innovation.

Finally, we assume that the unobserved heterogeneity in the random utility of both the agents of the market is logit. The following assumption allows to retrieve a closed form solution for the matching function as shown in section 5.5.

5.3 University department problem

Let's assume that university department characteristics be identified by a vector of observable attributes $x \in \boldsymbol{X} = \mathbb{R}^{d_x}$, including its location. More we acknowledge the existence of an heterogeneous part of the utility function (unknown to econometrician) expressing the unobservable preferences of the university department.

The university department maximizes its utility deciding the type of firm y located at z they want to accept. We define $\alpha(x, z, y)$ the utility derived by university department i of observable attributes $x_i = x$ choosing firm of type $y \in \boldsymbol{Y}$ located at $z \in \boldsymbol{Z}$. Furthermore we model the unobserved heterogeneity by introducing the term $\epsilon(i, z, y)$ capturing the idiosyncratic part of the university department utility when choosing firm of type y located at z.

Hence, the university department i of type x chooses firm of type y enjoys utility

$$\alpha(x, z, y) + t(x, z, y) + \sigma^U \epsilon(i, z, y)$$

where σ^U is the scaling factor associated with the unobserved taste of university department i (the superscript U stands for university).

The systematic utility of the university department is given by

$$U(x, z, y) = \alpha(x, z, y) + t(x, z, y)$$

Hence university department i of type x maximizes the utility function

$$u(i) = \max_{zy} \{U(x, z, y) + \sigma^U \epsilon(i, z, y)\} \tag{2}$$

5.4 Firm problem

Let's assume that firm characteristics be identified by a vector of observable attributes $y \in \boldsymbol{Y} = \mathbb{R}^{d_y}$. More we acknowledge the existence of an heterogeneous part of the utility function (unknown to econometrician) expressing the unobservable preferences of the firm.

The firm maximizes its utility deciding the location z and the type of university department x they want to invest. We define $\gamma(x, z, y)$ the utility derived by a firm j of observable attributes $y_j = y$ at location $z \in \boldsymbol{Z}$ choosing university department of type $x \in \boldsymbol{X}$. Furthermore we model the unobserved heterogeneity by introducing the term $\eta(x, z, j)$ capturing the idiosyncratic part of the firms' utility when choosing university department of type x at location z.

Hence, the firm j of type y chooses university department of type x enjoys utility

$$\gamma(x, z, y) - (1 - \tau(y))t(x, z, y) + \sigma^F \eta(x, z, j)$$

where σ^F is the scaling factor associated with the unobserved taste of firm j (the superscript F stands for firm) and $\tau(y)$ is the flat credit rate.

The systematic utility of the firm is given by

$$V(x, z, y) = \gamma(x, z, y) - (1 - \tau(y))t(x, z, y)$$

Hence firm j of type y maximizes the utility function

$$v(j) = \max_{xz} \{V(x, z, y) + \sigma^F \eta(x, z, j)\} \tag{3}$$

5.5 Matching function

In our economy the university department maximizes the utility given in (2) and the firm maximizes the utility given in (3) so the expected utility of the university department, assuming a large market, is

$$G_x(U(x, z, y)) = E_{S_x}\left[\max_{zy} \{U(x, z, y) + \sigma^U \epsilon(i, z, y)\}\right] \tag{4}$$

and the expected utility of the firm

$$H_y(V(x, z, y)) = E_{T_y}\left[\max_{xz} \{V(x, z, y)) + \sigma^F \eta(x, z, j)\}\right] \tag{5}$$

S_x and T_y are the known distributions of the unobserved heterogeneity of university departments and firms respectively.

From Daly-Zachary-Williams theorem[1] we know that for each couple (x, z, y) we have

$$\frac{\partial G_x(U(x,z,y))}{\partial U(x,z,y)} = \mu(zy|x) = \frac{\mu(x,z,y)}{n(x)}$$
$$\frac{\partial H_y(V(x,z,y))}{\partial V(x,z,y)} = \mu(xz|y) = \frac{\mu(x,z,y)}{m(y)}$$

where $\mu(zy|x)$ represents the mass of university departments of type x demanding firm of type y located at z while the $\mu(xz|y)$ represents the mass of firms y demanding at location z university departments of type x.

Definition 1. *μ is an equilibrium matching if and only if μ is*

- *feasible meaning that it satisfies the accounting constraints*

$$\begin{aligned} \sum_{zy} \mu(x,z,y) &= n(x) \\ \sum_{xz} \mu(x,z,y) &= m(y), \end{aligned}$$

  the market clearing condition

$$m(x)\frac{\partial G_x(U(x,z,y))}{\partial U(x,z,y)} = n(y)\frac{\partial H_y(V(x,z,y))}{\partial V(x,z,y)}$$

- *and it is solution to both problems* (4) *and* (5)

As underlined by Galichon and Salanié (2015) the large market assumption mitigates the concerns about the misrepresentation of the agent characteristics. The feasibility constraints are needed to ensure from the university department perspective that the mass of university department-firm pairs with university departments of type x coincides with the mass of university departments of type x and from the firm perspective that the mass of firm-university department pairs with firms of type y coincides with the mass of firms of type y. Finally, the market clearing condition satisfies the condition that at the equilibrium the mass of university departments of type x demanding firms of type y located at z coincides with the mass of firms of type y located at z demanding university departments of type x.

Assuming a Gumbel type I distribution for the idiosyncratic shocks, we know from McFadden

[1]The theorem states that the partial derivatives of the expected achieved utility is equal to the choice probabilities in an additive random utility model.

et al. (1973) that the probability choice is logit

$$G_x(U(x,z,y)) = \sigma^W \log\left(\sum_y \exp \frac{U(x,z,y)}{\sigma^W}\right)$$

and,

$$\frac{\partial G_x(U(x,z,y))}{\partial U(x,z,y)} = \frac{\exp \frac{(U(x,z,y))}{\sigma^W}}{\sum_{zy} \exp \frac{U(x,z,y)}{\sigma^W}}$$

According to Dupuy and Galichon (2015) we can then express the conditional probability of choosing firm y located at z given a university department of type x as

$$\mu(zy|x) = \exp \frac{U(x,z,y) - u(x)}{\sigma^W} \tag{6}$$

Similarly we can express the conditional probability of choosing university department x by a firm of type y deciding to locate at z as

$$\mu(xz|y) = \exp \frac{V(x,z,y) - v(y)}{\sigma^F} \tag{7}$$

we can now write the joint probabilities of (6) and (7) at the equilibrium respectively as

$$\mu(x,z,y) = m(x)\exp\left(\frac{\alpha(x,z,y) + t(x,z,y) - u(x)}{\sigma^U}\right) \tag{8}$$

$$\mu(x,z,y) = n(y)\exp\left(\frac{\gamma(x,z,y) - (1-\tau(y))t(x,z,y) - v(y)}{\sigma^F}\right) \tag{9}$$

Now manipulating the last expressions, we obtain the *aggregate matching function* (Galichon et al. (2017)) which can read as

$$\mu(x,z,y) = \left[m(x)^{\frac{\sigma^U}{\lambda^F(y)\sigma^F + \sigma^U}} n(y)^{\frac{\lambda^F(y)\sigma^F}{\lambda^F(y)\sigma^F + \sigma^U}} \exp \frac{(\alpha(x,z,y) - u(x)) + \lambda^F(y)(\gamma(x,z,y) - v(y))}{\sigma^U + \lambda^F(y)\sigma^F}\right] \tag{10}$$

where $\lambda^F(y) = \frac{1}{1-\tau(y)}$

The factor used to rescale the unobserved heterogeneity in the model express the heteroskedastic behavior of the distributions of the unobserved heterogeneity of the firm.

For sake of simplicity we can recast the equation (10) as:

$$\mu(x,z,y) = \left[m(x) \frac{\sigma^U}{\lambda^F(y)\sigma^F + \sigma^U} \, n(y) \frac{\lambda^F(y)\sigma^F}{\lambda^F(y)\sigma^F + \sigma^U} \exp \frac{(\widetilde{\Phi}(x,z,y) - u(x) - \widetilde{v}(y))}{\sigma^U + \lambda^F(y)\sigma^F} \right] \tag{11}$$

where $\widetilde{v}(y) = \lambda^F(y)v(y)$ and $\widetilde{\Phi}(x,z,y) = \alpha(x,z,y) + \lambda^F(y)\gamma(x,z,y)$ is the *systematic surplus*.

6 Estimation strategy

In our case, we have access to a representative sample of firms and university departments. For each university department we observe the characteristics X_i, including the location. For each firm we observe the characteristics Y_i and the location Z_i.

Therefore we proceed to the parametrization of $\alpha(x,z,y)$ and $\gamma(x,z,y)$, both contributing to the *systematic surplus*. In that regards we follow the specification proposed by Dupuy et al. (2017):

$$\alpha(x,z,y;A) = x^T A_0 y + A_1^T y + A_2 \, GP$$

$$\gamma(x,z,y;\Gamma) = x^T \Gamma_0 y + \Gamma_1^T x + \Gamma_2 \, GP$$

$$\widetilde{\Phi}(x,z,y;A,\Gamma) = \alpha(x,z,y;A) + \lambda^F(y)\gamma(x,z,y;\Gamma)$$

where the matrices of parameters A_0 and Γ_0 (*affinity matrix*) indicate the level (intensity) of complementarity or substitutability between the observables, A_1 and Γ_1 indicate the vectors of parameters assessing the direct effect of the observables, A_2 and Γ_2 are the parameters identifying the effect of the geographic proximity obtained by constructing the matrix of the inverse of the square root of the distances between university departments and firms.

We can now express the equilibrium matching using the sample counterpart of equation (11) and fulfill its *feasibility condition* by solving the scarcity constraint system introduced in the definition 1.

This allows to retrieve the functions $u(x_i;A,\Gamma)$ and $\widetilde{v}(y_j;A,\Gamma)$

$$\sum_{j=1}^{N} \exp \frac{(\widetilde{\Phi}(x_i,z_j,y_j;A,\Gamma) - u(x_i;A,\Gamma) - \widetilde{v}(y_j;A,\Gamma))}{\sigma^U + \lambda^F(y_j)\sigma^F} = \frac{1}{N}$$

$$\sum_{i=1}^{N} \exp \frac{(\widetilde{\Phi}(x_i,z_j,y_j;A,\Gamma) - u(x_i;A,\Gamma) - \widetilde{v}(y_j;A,\Gamma))}{\sigma^U + \lambda^F(y_j)\sigma^F} = \frac{1}{N}$$

with the normalization $u(x_1;A,\Gamma) = 0$

7 Variables

7.1 Firm variables

According to Cohen and Levinthal (1990), we define the absorptive capacity as the R&D intensity defined by the ratio of firms' R&D expenditures to firms' sales.

Then, we specify the size of the firm measured as the total number of employees. Generally, the range of the number of employees is quite ample: from micro size presenting few employees reaching large size with thousand employees.

Lastly we include the age of the firm as the number of years from its incorporation. This would provide a measure of the firms' ability to remain competitive, providing an indirect assessment of its innovation performances over long term (Zhang et al. (2019)).

7.2 University department variables

We define the size of the university department using the research expenditures. The intent is that this observable would supply implicitly the information contained in the traditional measure of the size attributed to the number of researchers (Garcia et al. (2015)) and further the degree of commitment university department shows for research. Indeed, the total research expenditures would capture two fundamental dimensions in promoting the interaction with companies:

- department spending more in research would naturally be more appealing for outstanding academics who in turn may potentially promote and attract valuable partnerships with firms
- it would also embed the costs concerning the maintenance and renovation of the physical assets that appears crucial in order to conduct innovative research.

We get the information of the research expenditures by US university departments from the NSF. According to the NSF research expenditures contains the following informations:

- salaries and wages of the R&D members whether full or part time, temporary or permanent, including fringe benefits.
- software purchased or license fees.
- movable equipments (computer, vehicles, furnitures), including ancillary cost related to delivery and setup
- other direct costs (travel, tuition waivers, services such as consulting, computer usage fees, and supplies)

Thereafter, we identify the quality of the research through the score provided by the University Ranking by Academic Performance dataset (URAP). We use the URAPs' score for precise reasons. Firstly, it is solely focused on bibliometric measurements such as publications, citations and the impact on their scientific fields, discarding teaching indicators such as student quality and teaching improvement. Indeed, the URAPs' score appears to be mainly interested in evaluating research-oriented institutions, and the quality of their research (Kivinen et al. (2017)). Secondly, according to *EUA report on Ranking for 2013* published by the European University Association, the URAPs' indicators tend to privilege research in science fields. In our case this makes the URAPs' score ideal to measure the quality of the university, mostly focused on research and innovation. The URAPs' score is determined by 6 indicators evaluating the research performance of the university for the current year.

- the number of articles published by the university in the current year.
- the number of citations measuring the impact of the research. The citations counted would cover a 4 years period of time previous the current year.
- the overall scientific productivity. This includes conference papers, reviews, letters, discussions, scripts in addition to journal articles published during the 4 years period of time previous the current year.
- the impact total factor (AIT) which evaluates the scientific production as the ratio between the citation per publication (CPP) of the university and the world average CPP, and then adjusted by the number of articles published during the 4 years period of time previous the current year.
- The citation impact total factor (CIT) which evaluates the impact of the research as the ratio between the CPP of the university and the world average CPP, and then adjusted by the number of citations received during the 4 years period of time previous the current year.
- the number of articles published with foreign universities.

We then construct a heuristic quality score for department by adjusting the university qualitys' score by the department research expenditures.

8 Dataset

Initially, we have derived a sample of firms using Compustat database produced by S&P and of US university departments using the HERD survey administered by the NSF. Then, we have

constructed our final dataset by randomly selecting university department and firm observations. For each university department we know:

- heuristic quality score
- research expenditures
- location

For each firm we know:

- business sector
- size
- age
- R&D intensity
- location

For each university department-firm pair we know

- geographic proximity

8.1 Firm sample

The observables of the firm has been retrieved using the Compustat database produced by S&P since 1962. This database offers a precious amount of accounting informations on annual or quarter basis through the financial report filled by public company. Indeed, it contains 98 percent of the total market capitalization with data backed to 1950.

We start by selecting the annual reports of public companies for the year 2013 from Compustat. Our year is the fiscal year as defined from June 2013 until May 2014 (not the calendar year), in line with standard rule established by the IRS for filling the taxation form 6765. Unfortunately we did not grant the access to the restricted area of IRS and therefore we use the 2013 as the most recent year for computing the credit rates associated with the R&D expenditures at the aggregate level of business sector (to be precise we compute the credit rates at the sub-sector level of the manufacturing sector as it will be clear later).

Our 2013 Compustat dataset is made of 11,871 companies. We select firms providing nonmissing, nonnegative and nonzero sales (Bayar et al. (2014)). We drop observations not reporting the number of employees, not located in US and we keep firms providing nonzero

or missing values concerning the R&D expenditures. This would reduce the dataset to 3,127 companies.

According to Hall (1988) we define the R&D intensity as the ratio of the item XRD (the R&D expenditures expressed in M$) to the item SALE (gross sales expressed in M$).

Afterwards we define the firm's size through the item EMP reporting the number of employees of the company and its consolidated subsidiaries expressed in thousands.

Thereafter we define the firms' age as incorporation age and merge our dataset with the hand-collected and updated version of the dataset of Prof. Ritter containing the incorporation dates. The Ritters' dataset even if it includes more than 11,000 firms, it is far from presenting a complete list of public companies reporting only those firms for which reliable informations regarding the incorporation dates are available. Indeed we obtained 1,564 residual companies (we invite the reader to explore appendix A to get a complete overview of the procedure).

Then we inspect the industry code used by the Compustat dataset. We opt for the NAICS code (SIC code was also available). The reason for that choice hinges on the fact that this would provide a direct link with the U.S Patenting and Trademark Office (USPTO) patents developed by NAICS industry category. We do this additional merging for the simple reason that our aim is to explore the collaborations university-firm intended as QRA. Indeed the USPTO dataset supplying the utility patent (known as *patents for inventions*) granted by industry sector between 2008-2012 would automatically pick those industry sectors mostly focused on the creation or improvement of products, processes or machines. The USPTO dataset reveals that only the manufacturing sector is concerned by the utility patents. This is not coming as surprise. Indeed the remarkable R&D activity conducted by the manufacturing sector is confirmed by two different sources:

- According to the Business Research and Innovation Survey (BRDIS) conducted by the NSF the R&D of the manufacturing sector accounts year after year for 60 to 70 percent of the domestic R&D activity in US.

- According to the SOI statistics for corporation research credit most of the US companies claiming tax credit for QREs belong to the manufacturing sector (above 60 percent). Indeed the SOI statistics concerning corporations credit for increasing research activities have been developed for business sector and also for sub-sectors of the manufacturing sector.

Within 1,564 companies, 809 companies belong to the manufacturing sector. We merge 793 companies using the USPTO dataset. Unfortunately 16 observations reporting the NAICS code 3241 at 4 digit level or 324 at 3 digit level were not present in the USPTO utility patent dataset

and then they have been dropped from the dataset.

Within the remaining 793 observations we have 108 firms exhibiting missing values of R&D intensity, these missings are associated to the blank R&D expenditures. We decide to replace the missing R&D intensity with the average R&D intensity by industry sector (we defer the curious reader to appendix B for further explanation). Indeed we rely in the BRDIS survey for 2013 conducted by the NSF which provides the R&D expenditures as percentage of the sales at 3 digit level of the NAICS code. According to the NSF report of 2013 the R&D intensities are computed by sampling 45,089 among private and public firms with more than 5 employees.

Thereafter we associate to the 793 companies their flat credit rates retrieved from the SOI tax statistics regarding the corporation research credit. As mentioned before the residual companies belong only to the manufacturing sector and the SOI tax statistics are available at 3 digit level of the manufacturing sector. Then we adopt the procedure introduced at the end of the section 4.3 to compute the flat credit rates by sub-sector of the manufacturing sector.

Lastly, as suggested by Griliches (2007) we plot the logarithm of the R&D expenditures on the logarithm of sales which can be assumed as the basic relationship between R&D activity and firm size (here roughly approximated by the sales). Similarly to Griliches (2007) we found out a positive correlation between size and R&D expenditures, although some small companies exhibit remarkable high R&D intensity and some medium companies exhibit remarkable low R&D intensity. These outliers can strongly affect the slope and the degree of curvature of the relationship making the R&D intensity distribution extremely skewed. In Griliches (2007) they trimmed the dataset excluding some firms exhibiting the lowest and highest R&D intensity. Unfortunately the same procedure adopted in our case would lead to the overall exclusion of 158 companies reducing the dataset to 635 companies. In order to avoid the deletion of precious observations while reducing the skewness of the R&D intensity distribution we winsorized the variable at 10th percentile and 90th percentile.

8.2 University department sample

The observables of the university department has been retrieved from the HERD survey conducted by the NSF. This represents the primary and most accurate source of accounting information for research expenditures by postsecondary institutions in the United States. It contains private and public universities reporting as total research expenditures at least 150,000$. In particular after 2010 each university campus headed by a campus-level president, chancellor, or equivalent is treated as separate institution rather than aggregating the campuses belonging to the same university, i.e. we have the University of California with its 9 campuses responding separately

to the HERD survey. In 2013, 917 academic institutions completed the survey but 26 of them reported Research expenditures less than 150,000$ and were excluded from the sample. NSF presents two version of the dataset, a short and long version. In the short version NSF includes 246 academic institutions reporting total research expenditures between 150,000 and 1M$, and designing that version with only few core questions compared to the long version. In the long version NSF includes 645 academic institutions with total research expenditures at least of 1M$, designing for that version an extended and detailed questionnaire. Indeed we use the latter version for which precise informations concerning the research expenditures by university department are available. Finally, the reliability of the HERD survey hinges on a response rate nearly close to 99 percent, ensuring that the imputation exercise is minimal.

Unluckily we do not have the precise information concerning type and amount of external collaboration between university and business. In principle, that information could be identified in question 3 of the HERD survey asking the part of university R&D financed externally by type of agreement. NSF mentions two types of agreement: contracts and grants, reimbursements, and all other agreements. The contract is defined as *legal commitments in which a good or service is provided by the institution that benefits the sponsor (firm). The sponsor specifies the deliverables and gains the rights to results.* This definition could fit the requirements to qualify as SRA, nonetheless we cannot disentangle the part of the contract received by federal sponsor and the part received by nonfederal sponsor (and among nonfederal sponsor we cannot make a distinction between noprofit and profit organizations). Furthermore even if this distinction was available, NSF would not supply that information detailed by university department. Moreover we decide to use the business fundings by university department as proxy for the overall SRA fundings received by the university for that department. This amount would constitute jointly with the other non federal sources and the federal sources the research expenditures by university department.

Moreover, we select universities receiving nonzero fundings from business at least in one of department (mentioned in the definition of QRA in section 4.2). This selection would reduce to 450 the number of institutions in our sample.

Then, we merge our remaining dataset with the USPTO dataset reporting the cumulative number of utility patents by university between 2008-2012. We use this information to shortlist universities most oriented to the commercialization of new products. We successfully merge 233 institutions.

Lastly, the remaining 233 academic institutions are merged with the URAPs' score dataset defining the quality of the university, reducing the dataset to 191 academic institutions corresponding to 1,317 university departments.

8.3 Final sample

The final dataset is derived by randomly merging (part of) 1,317 university departments with 793 firms. Indeed we implicitly fix as the greatest constraint for the overall number of potential partnerships university department-firm the number of observations in the firms' dataset (it therefore constitutes the size of the final sample).

The final dataset is then derived by constructing the key variable for appropriately merging the two datasets and by ensuring the randomness of the merging. We figured out the randomization by generating random numbers for the university department dataset. By sorting these numbers we ensure that the university departments are randomly picked. We secure the randomness and perform the merging by generating a variable keeping track of the order of the observations.

This last step is essential to structure our dataset in a convenient matrix where rows and columns represent university departments and firms respectively, and it is not intended as a procedure to create a one-to-one matching. Therefore, it does not represent an observed matching from which we can infer the optimal level of preferences but rather a pool of representative university departments and firms employed to simulate various counterfactual matchings.

Table 5 presents the summary statistics. The average department expenditure is 20.69M$ with a consistent dispersion indicated by the large standard deviation (52M$). The average firms R&D intensity is 0.45 with 3,800 employees and 33 years of life. The average geographic proximity is 0.0319.

9 Simulation

Using the final dataset we construct two different scenarios. Each scenario implies the computation of two distinct matchings obtained by varying the preferences, differentiated by changing the magnitude of the concerned parameter. The matching frequencies are then aggregated by 4 university departments and 18 industrial sectors (see table 6). The university departments refer to computer science, engineering, physics and life science. The industrial sectors refer to the sub-sectors of the manufacturing sector (3 digits in the NAICS classification).

Afterwards we compute for each university department-industrial sector pair the variation of the cognitive proximity and the geographic proximity (hereafter to ease the interpretation we visualize the variation of the geographic proximity as the variation of the geographic distance expressed in km).

In the first scenario:

- we compute the simulated matching by requiring low values of the parameter associated

to the interaction between the quality score of the university department and the R&D intensity of the firm. Thereafter we compute the cognitive proximity and the distance of collaboration for each university department-industrial sector pair given the matching values.

- we compute the simulated matching by requiring high value of the parameter associated to the interaction between the quality score of the university department and the R&D intensity of the firm, keeping ceteris paribus the parameter associated to the geographic proximity. Thereafter we compute the cognitive proximity and the distance of collaboration for each university department-industrial sector pair given the matching values.

In the second scenario:

- we compute the simulated matching by requiring low value of the parameter associated to the geographic proximity. Thereafter we compute the cognitive proximity and the distance of collaboration for each university department-industrial sector pair given the matching values.

- we compute the simulated matching by requiring high value of the parameter associated to the geographic proximity, keeping ceteris paribus the parameter associated to the interaction between the quality score of the university department and the R&D intensity of the firm. Thereafter we compute the cognitive proximity and the distance of collaboration for each university department-industrial sector pair given the matching values.

Henceforth we refer to the scenario involving the variation of the parameter related to the interaction of the university department quality score with the R&D intensity with the abbreviation QRD and to the scenario involving the variation of the parameter related to the geographic proximity with the abbreviation GP.

9.1 Scenario from low to high QRD

In this scenario we compute two distinct matchings by assuming low and high value of the preference associated with the QRD, keeping unchanged the parameter related to the geographic proximity.

Figure 1 represents the core of our findings obtained by determining the variation of the cognitive and the geographic proximity when we adjust from low to high value the preference associated with the QRD. The three tables display 72 elements where each element represents the variation occurring within a specific university department-industrial sector pair. Most of the

variation is occurring in life science and engineering concerning the university departments and chemical manufacturing and computers and electronic products manufacturing concerning the industrial sectors. Figure 1a quantifies the variation in the occurrence of the couples indicating with positive sign an increment in the number of couples and with negative sign a reduction in the number of couples when the parameter related to QRD is increased. Unsurprisingly, the increment in the number of couples is followed by an increase of the cosine index, pointing out that the cognitive proximity raises for these pairs (see figure 1b). Apparently, the elements concerning life science-chemical manufacturing and engineering-computer and electronic products manufacturing seem to have incremented only slightly their cognitive proximity in response to a consistent gain in the number of couples. However both have reached the greatest (positive) level of cosine index by imposing the high value to the parameter associated with the QRD. Generally, the variation of the geographic proximity results in a rise of the distance of collaboration in correspondence to a rise of the cognitive proximity, albeit for the pairs exhibiting large variations in the number of couples we have similar responses concerning the cognitive proximity and geographic proximity clearly indicating that a surge (reduction) of the cognitive proximity is followed by a decline (rise) of the distance of collaboration (see figure 1c).

The overall pattern of the cognitive proximity is further highlighted by the histogram displayed in figure 2a, reporting the distribution of cosine index values concerning the 72 elements representing the university department-industrial sector pairs. In particular the cosine index values are presented in blue in the case of low value of the parameter associated with the QRD and in orange in the case of high value of the parameter associated with the QRD. There is clear gain of the cognitive proximity driven by the upsurge of this preference. Instead the histogram representing the distances of collaboration (see figure 2b) unambiguously indicates that the increase of this preference implies a loss of geographic proximity (expressed by a growth of the distances of collaboration).

In the scatter plot of figure 3 we present the variation of the cosine index values on the variation of the distances of collaboration by university department (as obtained in figure 1). We drawn three main remarks:

- There is clear pattern indicating an improvement of the cognitive proximity independently of the industrial sector, with only life science showing a general worsening of the cognitive proximity for most of the collaborations with industrial sectors (figure 3a).
- There is an evident loss of geographic proximity with an increase of the distances of collaboration, and this is particularly true for life science (figure 3a).

- In general the university department-industrial sector pairs contribute equally to the increase of the surplus, with a unique exception regarding life science-chemical manufacturing pair showing a remarkable rise of the surplus (figure 3b).

The general improvement of cognitive proximity is confirmed by figure 4a and figure 5a. The figures present the scatter plot of the mean values of cosine index attributed to high value of the parameter associated with the QRD on the mean values of the cosine index attributed to low value of the parameter associated with the QRD respectively by university department and industrial sector. Life science confirms the loss of cognitive proximity as well as chemical manufacturing. Their mean values are precisely reflecting the fact that they are heavily close to each other (indeed that pair reaches the cosine index value of 1 with huge increase in the number of couples) while they are loosing cognitive proximity in the remaining matings. Furthermore it may appear unexpected the decrease of the distance of collaboration concerning life science (see figure 4b) whether we return to the above remarks. By the way, the contradiction is only apparent since in figure 3 we display the variation of the distance of collaboration for each university department-industrial sector pair while in figure 4b we present the average values concerning the distance of collaboration by university department, clearly indicating that the decrement of the distance of collaboration occurs for life science-industrial sector pairs where the number of couples is superior.

Finally, figure 5b indicates a comparable increase of the mean values of the distance of collaboration among the industrial sectors.

9.2 Scenario from low to high GP

This scenario requires the computation of two distinct matchings by imposing low value of the parameter attached to the GP in one case and high value of the parameter attached to the GP in the other case, while keeping unaltered the ones associated with QRD.

Figure 6 provides an overview of our findings. The three tables contain 72 elements, and each element expresses the variation occurring within a university department-industrial sector pair. In particular figure 6a and figure 6b present the variation respectively of the number of couples and cosine index. It is evident that no significant change in the number of couples and consequently in the cognitive proximity occurs in this setting. Instead remarkable changes take place for what concerns the distances of collaboration. There is consistent decrement of the distance for all pairs (see figure 6c). These findings are reinforced by the histogram in figure 7a and figure 7b. In blue it is depicted the histogram assigning low value to the parameter associated with the GP and in orange the one assigning high value to the parameter associated with the GP.

The histogram related to the distribution of the cosine index values shows a slightly decrease of the cognitive proximity while the histogram related to the distances of collaboration reveals a massive increment of the geographic proximity.

These patterns are overly confirmed by the scatter plots showing the variation of the cosine index values on the variation of the distances of collaboration by university department as illustrated in figure 8. We drawn three main remarks:

- There is apparently a tenuous variation of the cognitive proximity with one unique clear pattern related to life science department exhibiting a subtle improvement of the cognitive proximity (figure 8a)
- There is a significant increment of the geographic proximity, independently of the university department-industrial sector pair (figure 8a).
- There is a moderate increase of the surplus with equal contribution among the university department-industrial sector pair (figure 8b)

Figure 9a and figure 10a illustrate the scatter plots of the mean values of the cosine index related to high value of the parameter of GP on the mean values of the cosine index related to low value of the parameter of GP respectively by university department and industrial sector. In that case it appears evident that the slightly loss of cognitive proximity is prevailing. Although there is a moderate increase of the mean values of the cosine index for some industrial sectors and university departments these appear to be still negative.

Figure 9b and figure 10b illustrate the scatter plots of the mean values of the distance of collaboration related to high value of the parameter of GP on the mean values of the distance of collaboration related to low value of the parameter of GP respectively by university department and industrial sector. They both show a consistent decreasing of the distances of collaboration.

10 Discussion

The analysis of the first scenario shifting the preference from low to high value of the parameter associated with QRD shows a solid improvement of the cognitive proximity. The enhancement of the cognitive proximity is consistent for both the university departments and the industrial sectors. Unique exception is represented by the life science concerning the university departments and chemical manufacturing concerning the industrial sectors whose exhibit an overall loss of cognitive proximity. The explanation is simple: the life science-chemical manufacturing pair has experienced a remarkable increase in the number of couples leading to the sharp increment

in cognitive proximity but for both this arrangement has also produced the impairment of the cognitive proximity in the other matings. In this scenario the geographic proximity (expressed by the distance of collaboration in km) has deteriorated for almost all the university department and industrial sector pairs. Once more unique exception of this pattern is provided by the life science department and chemical manufacturing pair which displays against the general trend an improvement of the geographic proximity evidenced by the decrease of the average value of the distance of collaboration.

The second scenario examines the variation of the cosine index and the distance of collaboration generated by the shift from low to high value of the parameter associated with the GP. The geographic proximity shows a consistent improvement testified by a strong decrease of the distance of collaboration, independently of the university department and industrial sector. Unsurprisingly, the variation of the cosine index is rather weak given the subtle change in the number of couples by university department-industrial sector pair. However the mean values of cosine index computed for university departments and industrial sectors reveal a feeble impairment of the cognitive proximity.

The implementation of these scenarios definitely lead to the following remarks:

- The characteristics of the firm such as R&D intensity and the characteristics of the university department such as the quality of academic research are important factors whose may induce firms to collaborate with distant university departments. Indeed, the cognitive proximity generated by the higher value of the parameter associated with QRD may often substitute for geographic proximity in promoting university department-firm collaboration. This is in line with the empirical results obtained by Garcia et al. (2015).

- The theoretic model proved to be accurate enough to disclose the following findings: the general pattern showing a trade-off between cognitive and geographic proximity and the individual response of cognitive and geographic proximity for university department-industrial sector pairs (some of them showing a simultaneous increment of the geographic and cognitive proximity).

11 Conclusion

We contribute to that literature by proposing a new methodology to disentangle the effect of geographic and cognitive proximity on the willingness of university department and firm to collaborate. Indeed we propose a matching model with imperfect transferable utility à la Dupuy et al. (2017), allowing to separately quantify the effect of the different types of proximity. The

imperfection arises from the credit tax rate (here computed as flat credit rate by industrial sector) attributed to the firm within the QREs assumption.

The proposed scenarios generated by shifting from low to high value the parameter associated with QRD in one case and shifting from low to high value the parameter associated with GP in the other case shed a light on the potential substitutability between the cognitive and geographic proximity in shaping the university department-firm collaborations. This tendency already underlined by the past empirical literature have been successfully captured by the theoretical framework developed.

Moreover the encouraging results open avenues for future research. Indeed the forthcoming step would be to apply the model using a sample of real US partnerships university department-firm. The observed collaborations would allow to construct a measure of the cognitive proximity *ex ante* by applying the correspondence analysis to the observed matchings and by introducing it in the specification of the model as we did for the geographic proximity. This would enable to supply the optimal level of preference attached to the cognitive and geographic proximity, and figure out their relative importance when university department and firm decide to collaborate.

Finally we could also enrich the model specification by incorporating other types of proximity such as the organizational proximity. As underlined by D'Este et al. (2012) this type of proximity is able to soften the effect of spatial proximity by assessing the impact of past partnerships university department-firm on the propensity to collaborate.

12 Appendix

A. Firms'age

We review the definition of firms' age choosing the most suitable for our needs. Compustat dataset is supplying the information concerning the Initial Public Offering date (IPO), the year the company becomes public. This information may be used to define the *listing age*, namely the number of years plus one elapsed since the year of the companys' IPO. Shumway (2001) argues that the *listing age* is a crucial moment in a companys' life since listing may affect the capital structure, the ownership, and increase business opportunities. The *listing age* which assumes the year when it appears for the first time in the Center for Research in Security Prices database (CRSP) has been widely used by Fama and French (2001) and Pástor and Pietro (2003). However the inclusion of the firms' age in our model has the intent to provide an indirect measure of the firms' ability to innovate. Indeed we opt for the definition of the firms' age since the year of incorporation (*incorporation age*), namely the number of years plus one since the year of incorporation. For the scope we merge the Compustat dataset with the hand-collected and updated version of the dataset of Prof. Ritter containing the incorporation dates, the CRSP permanent IDs, the cusip identifier and the company names for firms that went public in the US during 1975-2019 (Loughran and Ritter (2004); Field and Karpoff (2002)). The Ritters' dataset even if it includes more than 11,000 firms, it is far from presenting a complete list of public companies reporting only those firms for which reliable informations regarding the incorporation dates are available (Loderer and Waelchli (2011)).

For that purpose we merge the 3,127 companies with the Ritters's dataset. The Ritters' dataset present two key variables to identify the company, namely LPERMNO and NCUSIP. Hence, we accomplish a double merging using LPERMNO as key variable for the first merge and NCUSIP as key variable for the second merge in order to maximize the companies matched. LPERMNO is a unique stock level identifier used in the CRSP database associated with the unique company identifier LPERMCO (GVKEY is the unique company identifier in the Compustat dataset). NCUSIP is an 8 digits code available in Compustat, the first 6 digits identify the company and the last 2 digits the type of security issued. In order to get the LPERMNOs (we can have multiple LPERMNO associated with one LPERMCO) we need to use the CRSP/Compustat merge tool. The latter enables to create a direct link between the identifier of the company in our main dataset (GVKEY) with the firms' identifier (LPERMCO) contained in the CRSP dataset. We retrieved 3,090 companies in the CRSP database, unfortunately for 37 firms we could not establish any link between Compustat and CRSP. Furthermore we delete the companies reporting

missing LPERMNO, this additionally reduces the companies to 2,557 observations. The first merging using LPERMNO results 1,505 companies matched while the second merging using NCUSIP accounts for 59 additional companies. This provides 1,564 residual companies.

B. R&D expenditures in Compustat

According to Koh and Reeb (2015) we treat the encoded zero R&D expenditures as the sign that the firm does not undertake any R&D activity during that year while we treat blank R&D expenditures as the voluntarily choice of the firm to not display their R&D expenditures by creating a glow of uncertainty about their potential level of innovation and growth vis-à-vis their competitors. This choice is supported by the empirical results obtained by Koh and Reeb (2015), explaining that missing R&D expenses may be a conscious decision of the firm's management to not separate the R&D expenses from other reported expenses. Indeed they found out that firms non-reporting R&D file 14 times more patent applications than firms reporting zero R&D. In the economic literature the patent activity is widely accepted as an indicator of innovation, providing a noisy measure of the intensity of R&D activity of the firm (Hall et al. (2005)). In that context it is emblematic the cases of Coca-Cola and MCI that they did not display R&D expenditures for more than 20 years (1980-2006) but during that period they filled over 500 patent applications. Moreover unsurprisingly Koh and Reeb (2015) found out that treating missing R&D as zero R&D may trigger substantial bias into the analysis while a better results is obtained by replacing missing R&D with the average R&D by industry sector. This approach is also in line with GAAP accountant requirements which dictates to the company to fulfill R&D expenditures section when it is larger than 1 percent of sales (the materiality threshold according to Accounting Series Released 125 (1972)). Therefore Compustat reports zero R&D expenditures when firms report zero R&D expenditures (defined as immateriality expenditures) and missing R&D when firms provides no information concerning R&D expenses.

Bibliography

Abrahamson, E. (1996). Management fashion. *Academy of management review*, 21(1):254–285.

Abramovsky, L., Harrison, R., and Simpson, H. (2007). University research and the location of business r&d. *The Economic Journal*, 117(519):C114–C141.

Adams, J. D. (2002). Comparative localization of academic and industrial spillovers. *Journal of Economic geography*, 2(3):253–278.

Agrawal, A. K. (2001). University-to-industry knowledge transfer: Literature review and unanswered questions. *International Journal of management reviews*, 3(4):285–302.

Alcácer, J. and Chung, W. (2007). Location strategies and knowledge spillovers. *Management science*, 53(5):760–776.

Alpaydın, U. A. R. and Fitjar, R. D. (2021). Proximity across the distant worlds of university–industry collaborations. *Papers in Regional Science*, 100(3):689–711.

Argyres, N. S. and Liebeskind, J. P. (1998). Privatizing the intellectual commons: Universities and the commercialization of biotechnology. *Journal of Economic Behavior & Organization*, 35(4):427–454.

Arundel, A. and Geuna, A. (2004). Proximity and the use of public science by innovative european firms. *Economics of Innovation and new Technology*, 13(6):559–580.

Bartel, A. P. and Lichtenberg, F. R. (1987). The comparative advantage of educated workers in implementing new technology. *The Review of Economics and statistics*, pages 1–11.

Bayar, T., Cornett, M. M., Erhemjamts, O., Leverty, T., and Tehranian, H. (2014). Product market competition and the efficient use of firm resources.

Bernstein, J. I. and Nadiri, M. I. (1988). Interindustry r&d spillovers, rates of return, and production in high-tech industries.

Bishop, K., D'Este, P., and Neely, A. (2011). Gaining from interactions with universities: Multiple methods for nurturing absorptive capacity. *Research Policy*, 40(1):30–40.

Bloom, N., Griffith, R., and Van Reenen, J. (2002). Do r&d tax credits work? evidence from a panel of countries 1979–1997. *Journal of Public Economics*, 85(1):1–31.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74.

Boschma, R. A. and Lambooy, J. G. (1999). Evolutionary economics and economic geography. *Journal of evolutionary economics*, 9(4):411–429.

Boschma, R. A. and Ter Wal, A. L. (2007). Knowledge networks and innovative performance in an industrial district: the case of a footwear district in the south of italy. *Industry and Innovation*, 14(2):177–199.

Bozeman, B., Rimes, H., and Youtie, J. (2015). The evolving state-of-the-art in technology transfer research: Revisiting the contingent effectiveness model. *Research Policy*, 44(1):34–49.

Bruneel, J., d'Este, P., and Salter, A. (2010). Investigating the factors that diminish the barriers to university–industry collaboration. *Research policy*, 39(7):858–868.

Burack, E. H. (1999). Bridging research to corporate application. In *Impact analysis*, pages 27–60. Psychology Press.

Cohen, W. M., Florida, R., Randazzese, L., and Walsh, J. (1998). Industry and the academy: uneasy partners in the cause of technological advance. *Challenges to research universities*, 171(200):59.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, 35(1):128–152.

Colyvas, J., Crow, M., Gelijns, A., Mazzoleni, R., Nelson, R. R., Rosenberg, N., and Sampat, B. N. (2002). How do university inventions get into practice? *Management science*, 48(1):61–72.

Cox, R. (1985). Lessons from 30 years of science parks in the usa. *Science Parks and Innovation Centres: Their Economic and Social Impact*, pages 18–24.

D'Este, P., Guy, F., and Iammarino, S. (2012). Shaping the formation of university–industry research collaborations: what type of proximity does really matter? *Journal of economic geography*, 13(4):537–558.

Di Gregorio, D. and Shane, S. (2003). Why do some universities generate more start-ups than others? *Research policy*, 32(2):209–227.

Dupuy, A. and Galichon, A. (2015). A note on the estimation of job amenities and labor productivity.

Dupuy, A., Galichon, A., Jaffe, S., and Kominers, S. D. (2017). Taxation in matching markets.

Etzkowitz, H. (1998). The norms of entrepreneurial science: cognitive effects of the new university–industry linkages. *Research policy*, 27(8):823–833.

Fama, E. F. and French, K. R. (2001). Disappearing dividends: changing firm characteristics or lower propensity to pay? *Journal of Financial economics*, 60(1):3–43.

Fichtner, J. J. and Michel, A. N. (2015). Can a research and development tax credit be properly designed for economic efficiency. *Washington: Mercatus Center-George Mason University*.

Field, L. C. and Karpoff, J. M. (2002). Takeover defenses of ipo firms. *The Journal of Finance*, 57(5):1857–1889.

Fritsch, M. and Lukas, R. (2001). Who cooperates on r&d? *Research policy*, 30(2):297–312.

Galichon, A., Kominers, S. D., and Weber, S. (2017). Costly concessions: An empirical framework for matching with imperfectly transferable utility.

Galichon, A. and Salanié, B. (2015). Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*.

Garcia, R., Araujo, V., Mascarini, S., Gomes Santos, E., and Costa, A. (2015). Looking at both sides: how specific characteristics of academic research groups and firms affect the geographical distance of university–industry linkages. *Regional Studies, Regional Science*, 2(1):518–534.

Gertler, M. S. (2007). Tacit knowledge in production systems: how important is geography? In *The economic geography of innovation*.

Griliches, Z. (2007). *R&D, patents and productivity*. University of Chicago Press.

Griliches, Z. et al. (1979). Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of economics*, 10(1):92–116.

Hall, B. H. (1988). The effect of takeover activity on corporate research and development. In *Corporate takeovers: Causes and consequences*, pages 69–100. University of Chicago Press.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, pages 16–38.

Hautala, J. (2011). Cognitive proximity in international research groups. *Journal of Knowledge Management*, 15(4):601–624.

Jaffe, A. B. (1989). Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2):87–97.

Jaffe, A. B. et al. (1989). Real effects of academic research. *American economic review*, 79(5):957–970.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Jensen, R. and Thursby, M. (2001). Proofs and prototypes for sale: The licensing of university inventions. *American Economic Review*, 91(1):240–259.

Keller, W. and Yeaple, S. R. (2009). Multinational enterprises, international trade, and productivity growth: firm-level evidence from the united states. *The Review of Economics and Statistics*, 91(4):821–831.

Kivinen, O., Hedman, J., and Artukka, K. (2017). Scientific publishing and global university rankings. how well are top publishing universities recognized? *Scientometrics*, 112(1):679–695.

Klassen, K. J., Pittman, J. A., Reed, M. P., and Fortin, S. (2004). A cross-national comparison of r&d expenditure decisions: tax incentives and financial constraints. *Contemporary Accounting Research*, 21(3):639–680.

Knoben, J. and Oerlemans, L. A. (2006). Proximity and inter-organizational collaboration: A literature review. *international Journal of management reviews*, 8(2):71–89.

Knockaert, M., Ucbasaran, D., Wright, M., and Clarysse, B. (2011). The relationship between knowledge transfer, top management team composition, and performance: the case of science–based entrepreneurial firms. *Entrepreneurship Theory and Practice*, 35(4):777–803.

Koh, P.-S. and Reeb, D. M. (2015). Missing r&d. *Journal of Accounting and Economics*, 60(1):73–94.

Laursen, K., Reichstein, T., and Salter, A. (2011). Exploring the effect of geographical proximity and university quality on university–industry collaboration in the united kingdom. *Regional studies*, 45(4):507–523.

Lawler, E. E., Mohrman, A. M., Mohrman, S. A., Ledford, G., and Cummings, T. G. (1999). *Doing research that is useful for theory and practice*. Lexington Books.

Lee, P. (2019). Tacit knowledge and university-industry technology transfer. *Research Handbook on Intellectual Property and Technology Transfer (2019, Forthcoming)*.

Levy, R., Roux, P., and Wolff, S. (2009). An analysis of science–industry collaborative patterns in a large european university. *The Journal of Technology Transfer*, 34(1):1–23.

Loderer, C. and Waelchli, U. (2011). Firm age and governance. *Çalışma metni*.

Loughran, T. and Ritter, J. (2004). Why has ipo underpricing changed over time? *Financial management*, pages 5–37.

Lowe, R. A. (2006). Who develops a university invention? the impact of tacit knowledge and licensing policies. *The Journal of Technology Transfer*, 31(4):415–429.

Mansfield, E. (1991). Academic research and industrial innovation. *Research policy*, 20(1):1–12.

Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *Review of Economics and Statistics*, 77(1):55–65.

Mansfield, E. and Lee, J.-Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial r&d support. *Research policy*, 25(7):1047–1058.

Markman, G. D., Phan, P. H., Balkin, D. B., and Gianiodis, P. T. (2005). Entrepreneurship and university-based technology transfer. *Journal of business venturing*, 20(2):241–263.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

Micklethwait, J. and Wooldridge, A. (1996). *The witch doctors: Making sense of the management gurus*. Crown Business.

Molina-Morales, F. X., García-Villaverde, P. M., and Parra-Requena, G. (2014). Geographical and cognitive proximity effects on innovation performance in smes: a way through knowledge acquisition. *International Entrepreneurship and Management Journal*, 10(2):231–251.

Muscio, A. (2013). University-industry linkages: What are the determinants of distance in collaborations? *Papers in Regional Science*, 92(4):715–739.

Myles Shaver, J. and Flyer, F. (2000). Agglomeration economies, firm heterogeneity, and foreign direct investment in the united states. *Strategic management journal*, 21(12):1175–1193.

National Center for Education Statistics (Ed), Washington, D. (1997). *Integrated Postsecondary Education Data System, 1994.[Cd-Rom].* National Center for Education Statistics.

Nelson, R. R. (1986). Institutions supporting technical advance in industry. *The American Economic Review*, 76(2):186–189.

Nooteboom, B. (2000). *Learning and innovation in organizations and economies*. OUP Oxford.

Pástor, L. and Pietro, V. (2003). Stock valuation and learning about profitability. *The Journal of Finance*, 58(5):1749–1789.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., et al. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research policy*, 42(2):423–442.

Pfeffer, J., Sutton, R. I., et al. (2000). *The knowing-doing gap: How smart companies turn knowledge into action*. Harvard business press.

Powell, W. W. and Owen-Smith, J. (1998). Universities and the market for intellectual property in the life sciences. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 17(2):253–277.

Press, E. and Washburn, J. (2017). The kept university. In *Academic Ethics*, pages 191–204. Routledge.

Rao, N. (2016). Do tax credits stimulate r&d spending? the effect of the r&d tax credit in its first decade. *Journal of Public Economics*, 140:1–12.

Robert, L. (1988). On the mechanics of economic development. *Journal of monetary economics*.

Rynes, S. L., Bartunek, J. M., and Daft, R. L. (2001). Across the great divide: Knowledge creation and transfer between practitioners and academics. *Academy of management Journal*, 44(2):340–355.

Santoro, M. D. and Bierly, P. E. (2006). Facilitators of knowledge transfer in university-industry collaborations: A knowledge-based perspective. *IEEE Transactions on engineering management*, 53(4):495–507.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1):101–124.

Simonin, B. L. (1999). Ambiguity and the process of knowledge transfer in strategic alliances. *Strategic management journal*, 20(7):595–623.

Thursby, J. G. and Thursby, M. C. (2004). Are faculty critical? their role in university–industry licensing. *Contemporary Economic Policy*, 22(2):162–178.

Wink, R. (2008). Gatekeepers and proximity in science-driven sectors in europe and asia: the case of human embryonic stem cell research. *Regional Studies*, 42(6):777–791.

Wozniak, G. D. (1987). Human capital, information, and the early adoption of new technology. *Journal of Human Resources*, pages 101–112.

Zhang, S., Yuan, C., and Wang, Y. (2019). The impact of industry–university–research alliance portfolio diversity on firm innovation: Evidence from chinese manufacturing firms. *Sustainability*, 11(8):2321.

Table 1: computes the credit amount expressed in $ using the regular method under section 280C(c) before and after TCJA

	before TCJA	**after TCJA**
Current Year QREs	2,500,000	2,500,000
Fixed Base Percentage	16%	16%
Average Gross Receipts prior 4 years	10,000,000	10,000,000
Base Amount	1,600,000	1,600,000
Base Amount or 50% of the Current Year QREs	1,600,000	1,600,000
Excess of QREs over Base Amount	900,000	900,000
Reduced Credit	**117,000**	**142,200**

Note: the reduced credit is obtained by multiplying the Excess of QREs over the Base Amount by 13 percent (obtained by multiplying the 65 percent to the 20 percent credit rate) before TCJA and by 15.8 (obtained by multiplying the 79 percent to the 20 percent credit rate) percent after TCJA
Source: INSIGHT: 2017 Tax Law Changes Increases Value of R&D Tax Credit

Table 2: computes the credit amount expressed in $ using the ASC method under section 280C(c) before and after TCJA

	before TCJA	**after TCJA**
Current Year QREs	2,500,000	2,500,000
Sum Prior 3 years of QREs	7,500,000	7,500,000
Average of the Prior 3 years of QREs	2,500,000	2,500,000
Base Amount	1,250,000	1,250,000
Excess of QREs over Base Amount	1,250,000	1,250,000
Reduced Credit	**113,750**	**138,250**

Note: the reduced credit is obtained by multiplying the Excess of QREs over the Base Amount by 9.1 percent (obtained by multiplying the 65 percent to the 14 percent credit rate) before TCJA and by 11.06 percent (obtained by multiplying the 79 percent to the 14 percent credit rate) after TCJA
Source: INSIGHT: 2017 Tax Law Changes Increases Value of R&D Tax Credit

Table 3: computes the flat credit rate for illustrative purpose of the food manufacturing sector from SOI relative to tax year 2013 using regular and ASC computation, all the amounts are in thousands of dollars.

	Regular
Wages For Qualified Services	452,629
Cost of supplies	258,090
Rental or lease costs of computers	5
Applicable percentage of contract research expense	79,362
Total QREs	763,086
Average Annual Gross Receipts	N/A
Base Amount	180,827
Regular Credit	55,736
Percent per QREs $	**7.3%**
Weight Regular Credit	**36.5%**

	ASC
Wages For Qualified Services	837,553
Cost of supplies	236,552
Rental or lease costs of computers	0
Applicable percentage of contract research expense	221,979
Total QREs	1,324,803
QREs for prior 3 years	3,625,874
Alternative Simplified Credit	47,556
Percent per QREs $	**3.6%**
Weight ASC credit	**63.4%**

Note: the flat credit rate is obtained by computing the average credit rate and multiplying it to the applicable percentage (65 percent) of contract research expenses. In this case the average credit rate is **4.94%** obtained by computing the following expression 0.365***7.3**+0.634***3.6** where 0.365 is the weight provided by the following ratio $\frac{763{,}086}{763{,}086+1{,}324{,}803}$ and 0.634 is the weight provided by the following ratio $\frac{1{,}324{,}803}{763{,}086+1{,}324{,}803}$. Then the flat credit rate is then **3.21%** (4.94*0.65)

Table 4: computes the flat credit rates for sub-sectors of the manufacturing sector, computing from SOI 2013

Industrial sectors	**Flat credit rates**
Beverage and tobacco products manufacturing	3.92%
Chemicals manufacturing	3.29%
Computers and electronic products manufacturing	3.84%
Electrical equipment and appliances manufacturing	3.80%
Food manufacturing	**3.21**%
Furniture and related product manufacturing	3.86%
Machinery manufacturing	2.43%
Nonmetallic mineral products manufacturing	3.11%
Plastics and rubber products manufacturing	2.21%
Primary metals and fabricated metal products manufacturing	2.97%
Transportation equipment manufacturing	3.63%
Wood products manufacturing	3.47%
Miscellaneous manufacturing	2.95%
Paper manufacturing	3.57%
Printing and related support activities	3.74%
Textile mills and textile products mils	4.66%
Apparel manufacturing	4.15%
Leather and allied products manufacturing	3.77%

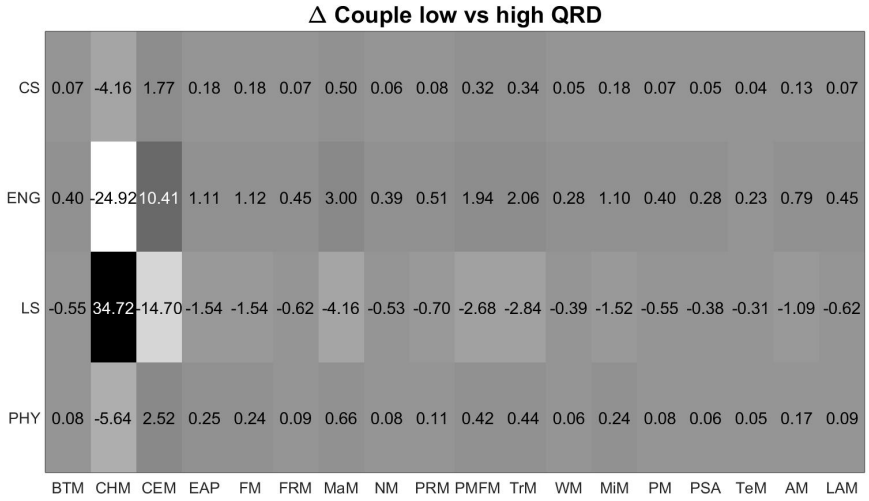
Table 5: presents the summary statistics

Variable	Description	Mean	Standard deviation
UDEs	University department expenditures in M$	20.69	52.05
HQS	Heuristic quality score computed for department; the university URAP score is weighted by the amount of department expenditures	28.24	44.51
Emp	Number of firms' employees in thousands	3.80	13.06
R&D intensity	Ratio of firms' R&D expenditures to firms' sales	0.45	0.77
Age	Firms' age in years	33.84	27.94
GP	Geographic proximity expressed as the inverse of the squared root of university-firm distance	0.0319	0.0374

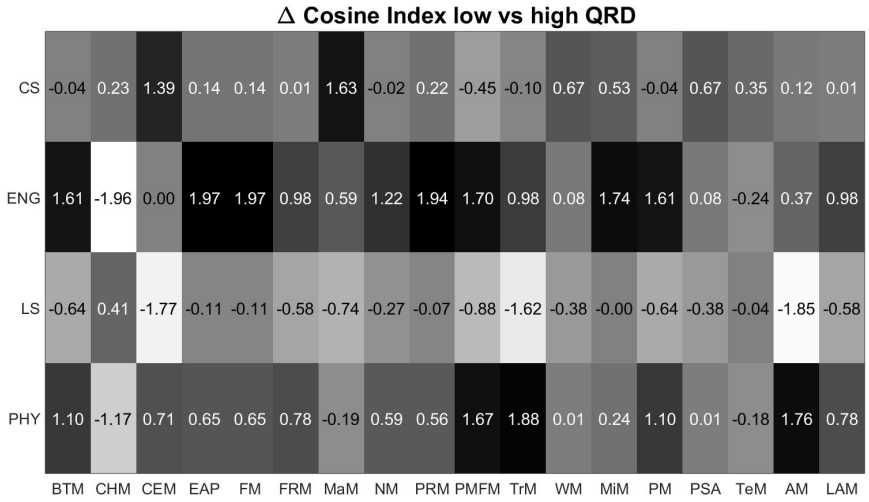
Table 6: summarizes the 4 university departments and the 18 industrial sectors used for the analysis, and their abbreviations

Industrial sectors	**Industrial sectors abb.**
Beverage and tobacco products manufacturing	BTM
Chemicals manufacturing	CHM
Computers and electronic products manufacturing	CEM
Electrical equipment and appliances manufacturing	EAP
Food manufacturing	FM
Furniture and related product manufacturing	FRM
Machinery manufacturing	MaM
Nonmetallic mineral products manufacturing	NM
Plastics and rubber products manufacturing	PRM
Primary metals and fabricated metal products manufacturing	PMFM
Transportation equipment manufacturing	TrM
Wood products manufacturing	WM
Miscellaneous manufacturing	MiM
Paper manufacturing	PM
Printing and related support activities	PSA
Textile mills and textile products mils	TeM
Apparel manufacturing	AM
Leather and allied products manufacturing	LAM

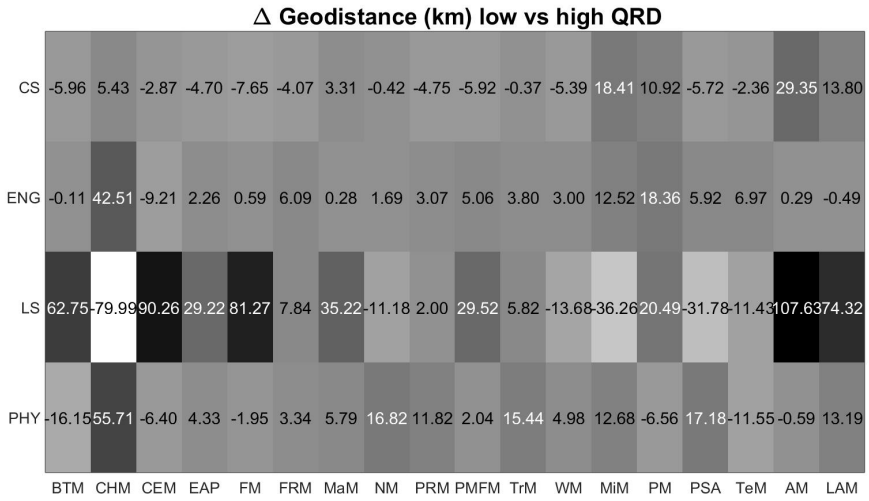
University departments	**University departments abb.**
Computer science	CS
Engineering	ENG
Life science	LS
Physics	PHY


Δ Couple low vs high QRD
CS 0.07 -4.16 1.77 0.18 0.18 0.07 0.50 0.06 0.08 0.32 0.34 0.05 0.18 0.07 0.05 0.04 0.13 0.07
ENG 0.40 -24.92 10.41 1.11 1.12 0.45 3.00 0.39 0.51 1.94 2.06 0.28 1.10 0.40 0.28 0.23 0.79 0.45
LS -0.55 34.72 -14.70 -1.54 -1.54 -0.62 -4.16 -0.53 -0.70 -2.68 -2.84 -0.39 -1.52 -0.55 -0.38 -0.31 -1.09 -0.62
PHY 0.08 -5.64 2.52 0.25 0.24 0.09 0.66 0.08 0.11 0.42 0.44 0.06 0.24 0.08 0.06 0.05 0.17 0.09
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TrM WM MiM PM PSA TeM AM LAM

(a)

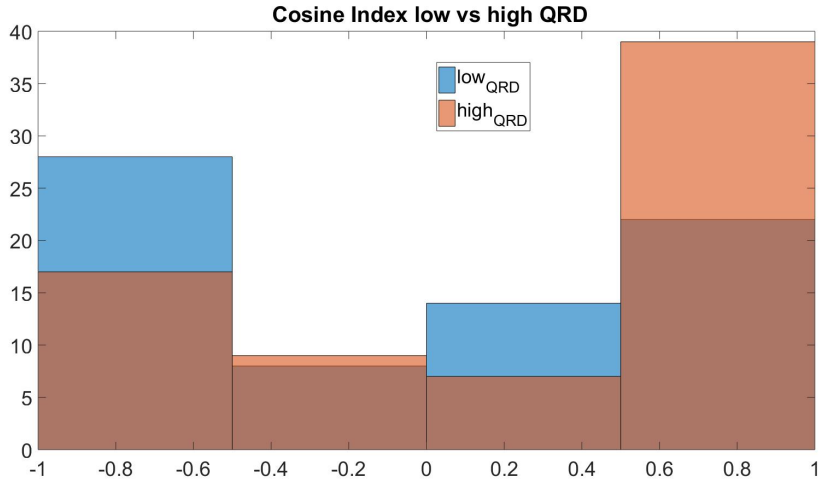

Δ Cosine Index low vs high QRD
CS -0.04 0.23 1.39 0.14 0.14 0.01 1.63 -0.02 0.22 -0.45 -0.10 0.67 0.53 -0.04 0.67 0.35 0.12 0.01
ENG 1.61 -1.96 0.00 1.97 1.97 0.98 0.59 1.22 1.94 1.70 0.98 0.08 1.74 1.61 0.08 -0.24 0.37 0.98
LS -0.64 0.41 -1.77 -0.11 -0.11 -0.58 -0.74 -0.27 -0.07 -0.88 -1.62 -0.38 -0.00 -0.64 -0.38 -0.04 -1.85 -0.58
PHY 1.10 -1.17 0.71 0.65 0.65 0.78 -0.19 0.59 0.56 1.67 1.88 0.01 0.24 1.10 0.01 -0.18 1.76 0.78
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TrM WM MiM PM PSA TeM AM LAM

(b)

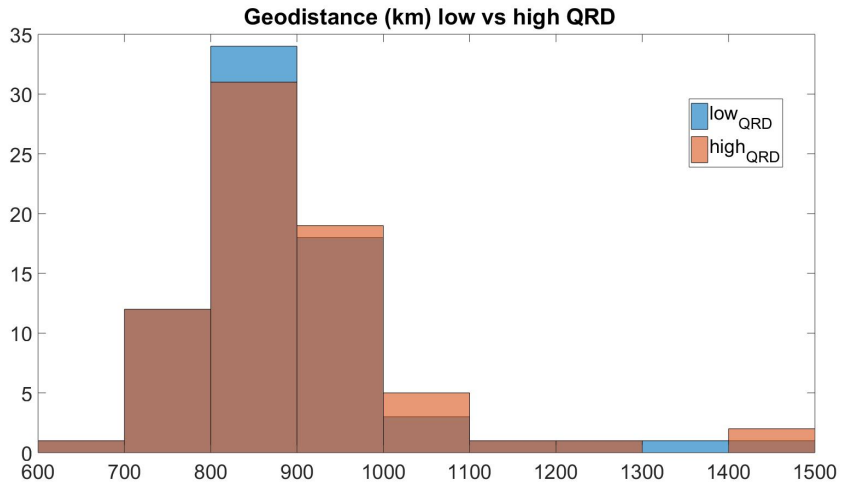

Δ Geodistance (km) low vs high QRD
CS -5.96 5.43 -2.87 -4.70 -7.65 -4.07 3.31 -0.42 -4.75 -5.92 -0.37 -5.39 18.41 10.92 -5.72 -2.36 29.35 13.80
ENG -0.11 42.51 -9.21 2.26 0.59 6.09 0.28 1.69 3.07 5.06 3.80 3.00 12.52 18.36 5.92 6.97 0.29 -0.49
LS 62.75 -79.99 90.26 29.22 81.27 7.84 35.22 -11.18 2.00 29.52 5.82 -13.68 -36.26 20.49 -31.78 -11.43 107.63 74.32
PHY -16.15 55.71 -6.40 4.33 -1.95 3.34 5.79 16.82 11.82 2.04 15.44 4.98 12.68 -6.56 17.18 -11.55 -0.59 13.19
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TrM WM MiM PM PSA TeM AM LAM

(c)

Figure 1: shows the variation of three quantities computed when the level of preference concerning QRD is switched from low to high. The three tables display 72 elements where each element represents the variation occurring within a specific university department-industrial sector pair. The variation of the number of couples is reported in (a); the variation of the cosine index is presented in (b); the variation of distance of collaboration is displayed in (c), the distances of collaboration have been previously normalized by the amount of couples in the university department-industrial sector pair.

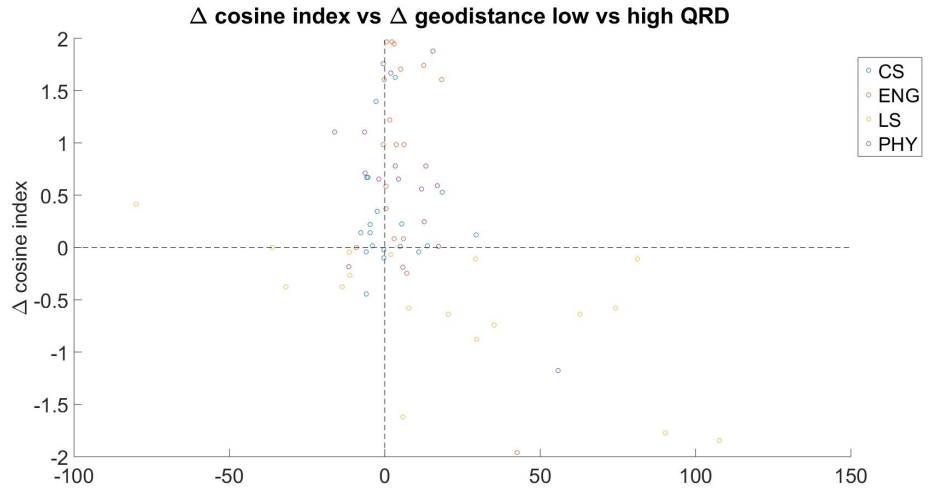


(a)

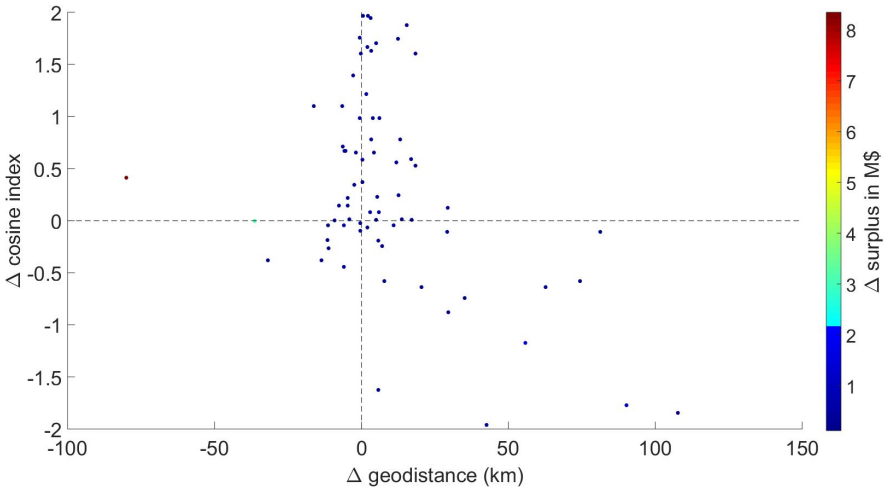


(b)

Figure 2: shows through the histogram representation the distribution of values regarding the cosine index (a) and the distances of collaboration (normalized by the amount of couples in the university department-industrial sector pair) (b) respectively in case of low value of the parameter associated with QRD (blue) and high value of the parameter associated with QRD (orange). Common values of the cosine index and distance of collaboration are colored in brown as resulting from the intersection of blue and orange.

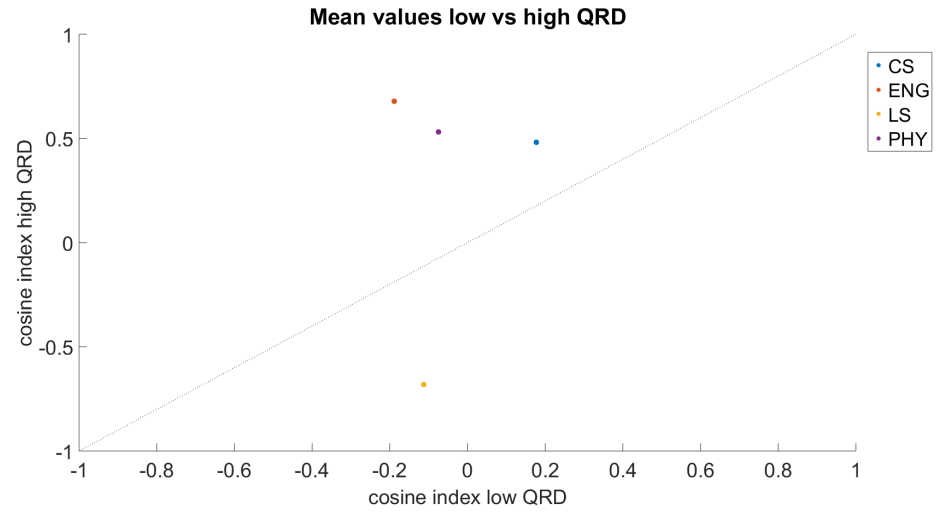


(a)

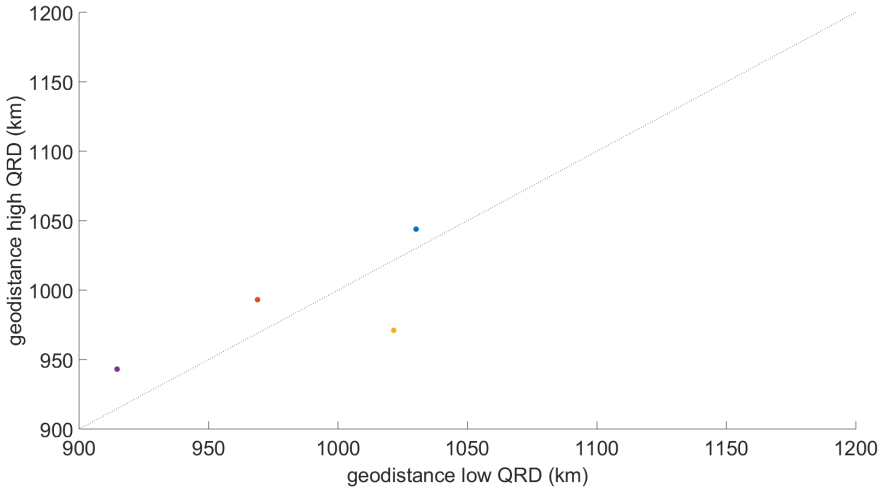


(b)

Figure 3: shows the scatter plot regarding the variations of the cosine index on the variations of the distance of collaboration for each of university department-industrial sector pair in (a) and the identical scatter plot with the additional information concerning the variations of the surplus in (b) when the level of preference concerning QRD is switched from low to high. The distances of collaboration and the surplus values have been previously normalized by the amount of couples in the university department-industrial sector pair.

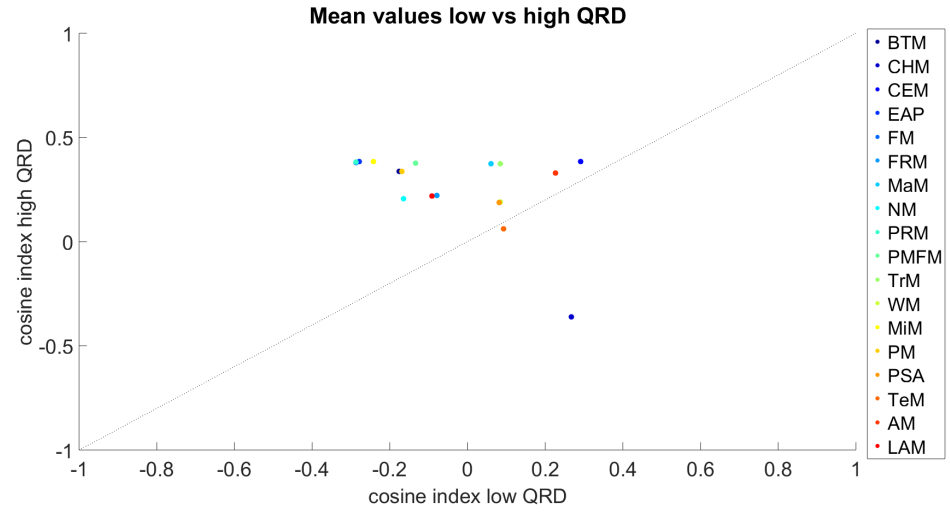


(a)

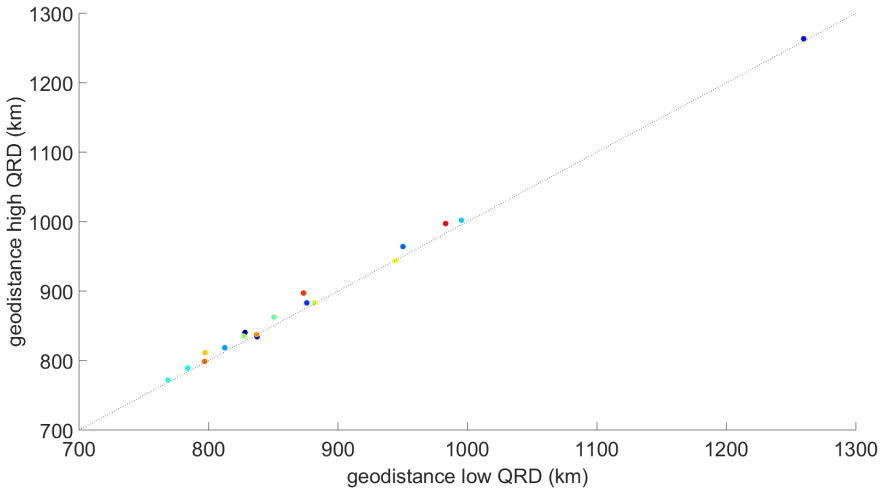


(b)

Figure 4: displays the scatter plot of the mean values of the cosine index generated by the high value of the parameter associated with QRD on the mean values of the cosine index generated by the low value of the parameter associated with QRD by university department (a) and the scatter plot of the mean values of the distance of collaboration generated by the high value of the parameter associated with QRD on the mean values of the distance of collaboration generated by the low value of the parameter associated with QRD by university department (b).

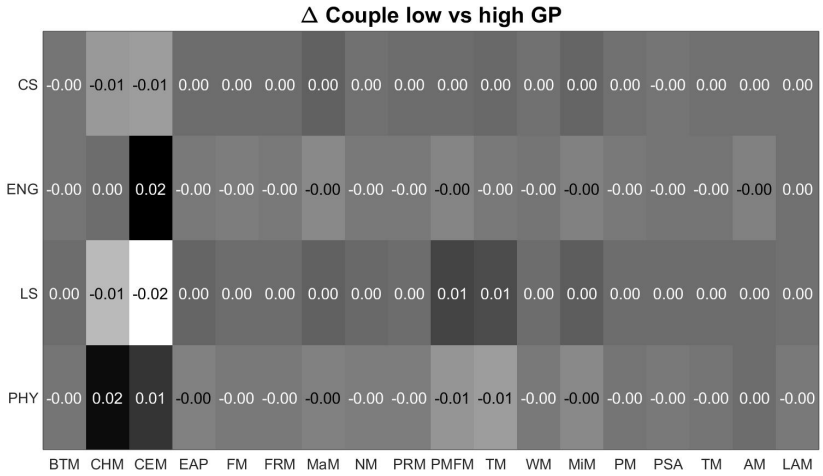


(a)

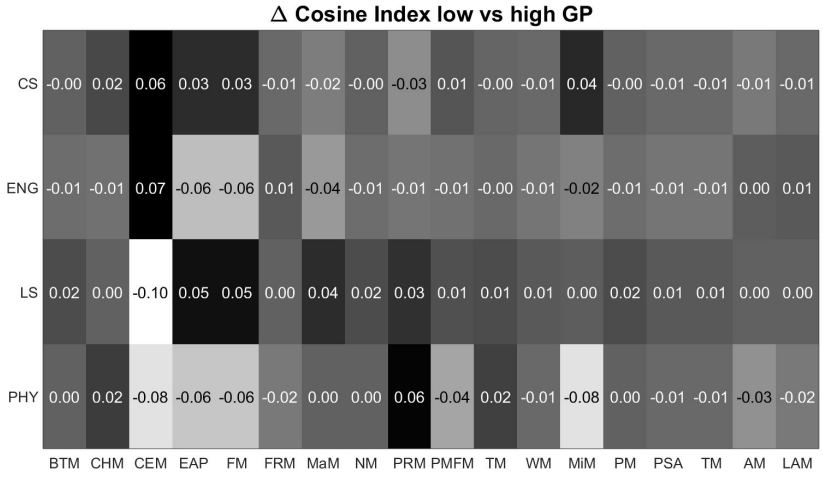


(b)

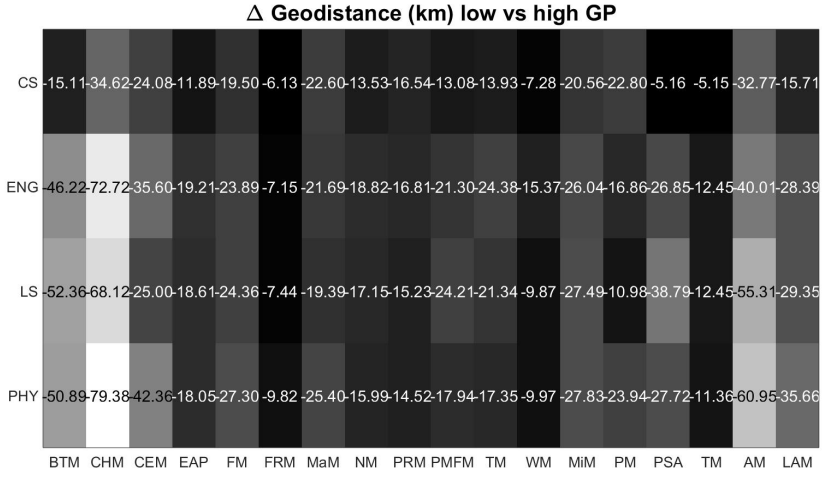
Figure 5: displays the scatter plot of the mean values of the cosine index generated by the high value of the parameter associated with QRD on the mean values of the cosine index generated by the low value of the parameter associated with QRD by industrial sector (a) and the scatter plot of the mean values of the distance of collaboration generated by the high value of the parameter associated with QRD on the mean values of the distance of collaboration generated by the low value of the parameter associated with QRD by industrial sector (b).


Δ Couple low vs high GP
CS -0.00 -0.01 -0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 -0.00 0.00 0.00 0.00
ENG -0.00 0.00 0.02 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.00
LS 0.00 -0.01 -0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00
PHY -0.00 0.02 0.01 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.01 -0.01 -0.00 -0.00 -0.00 -0.00 -0.00 0.00 -0.00
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TM WM MiM PM PSA TM AM LAM

(a)

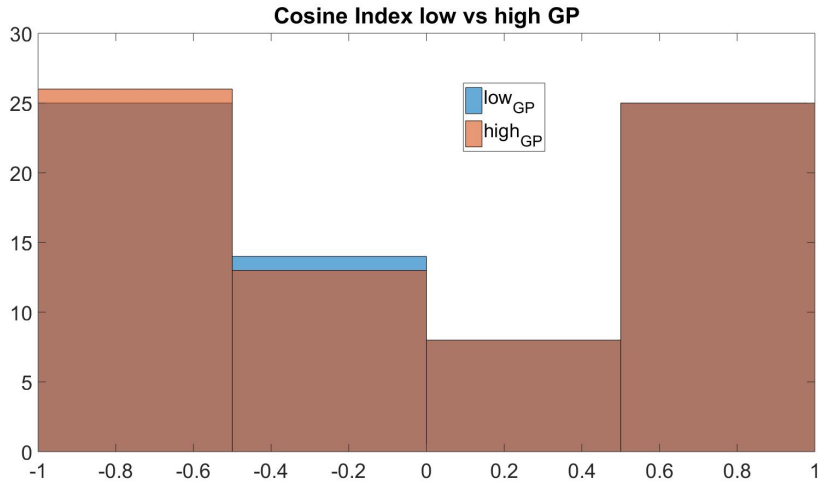

Δ Cosine Index low vs high GP
CS -0.00 0.02 0.06 0.03 0.03 -0.01 -0.02 -0.00 -0.03 0.01 -0.00 -0.01 0.04 -0.00 -0.01 -0.01 -0.01 -0.01
ENG -0.01 -0.01 0.07 -0.06 -0.06 0.01 -0.04 -0.01 -0.01 -0.01 -0.00 -0.01 -0.02 -0.01 -0.01 -0.01 0.00 0.01
LS 0.02 0.00 -0.10 0.05 0.05 0.00 0.04 0.02 0.03 0.01 0.01 0.01 0.00 0.02 0.01 0.01 0.00 0.00
PHY 0.00 0.02 -0.08 -0.06 -0.06 -0.02 0.00 0.00 0.06 -0.04 0.02 -0.01 -0.08 0.00 -0.01 -0.01 -0.03 -0.02
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TM WM MiM PM PSA TM AM LAM

(b)

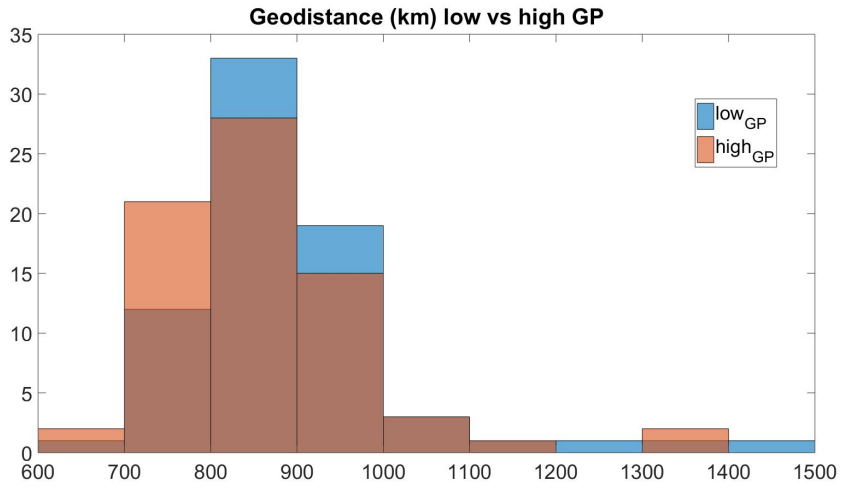

Δ Geodistance (km) low vs high GP
CS -15.11 -34.62 -24.08 -11.89 -19.50 -6.13 -22.60 -13.53 -16.54 -13.08 -13.93 -7.28 -20.56 -22.80 -5.16 -5.15 -32.77 -15.71
ENG -46.22 -72.72 -35.60 -19.21 -23.89 -7.15 -21.69 -18.82 -16.81 -21.30 -24.38 -15.37 -26.04 -16.86 -26.85 -12.45 -40.01 -28.39
LS -52.36 -68.12 -25.00 -18.61 -24.36 -7.44 -19.39 -17.15 -15.23 -24.21 -21.34 -9.87 -27.49 -10.98 -38.79 -12.45 -55.31 -29.35
PHY -50.89 -79.38 -42.36 -18.05 -27.30 -9.82 -25.40 -15.99 -14.52 -17.94 -17.35 -9.97 -27.83 -23.94 -27.72 -11.36 -60.95 -35.66
BTM CHM CEM EAP FM FRM MaM NM PRM PMFM TM WM MiM PM PSA TM AM LAM

(c)

Figure 6: shows the variation of three quantities computed when the level of preference concerning GP is switched from low to high. The three tables display 72 elements where each element represents the variation occurring within a specific university department-industrial sector pair. The variation of the number of couples is reported in (a); the variation of the cosine index is presented in (b); the variation of distance of collaboration is displayed in (c), the distances have been previously normalized by the amount of couples in the university department-industrial sector pair.

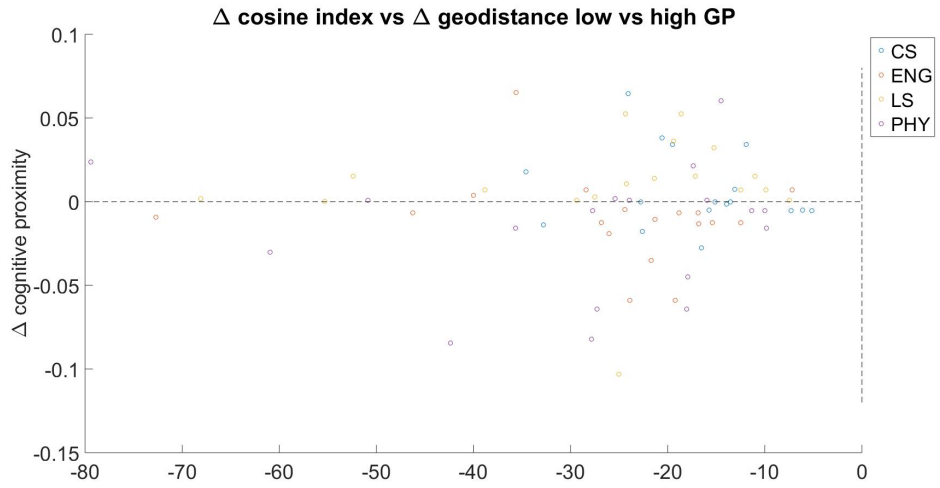


(a)

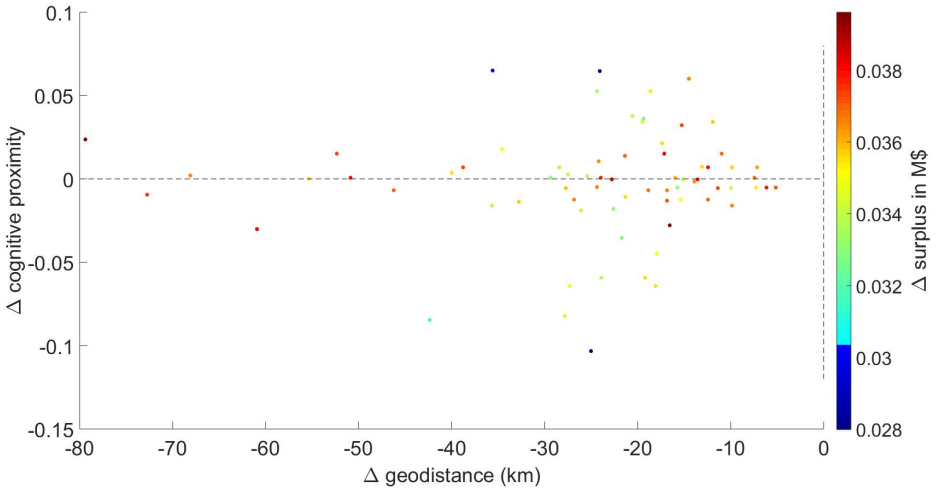


(b)

Figure 7: shows through the histogram representation the distribution of values regarding the cosine index (a) and the distances of collaboration (normalized by the amount of couples in the university department-industrial sector pair) (b) respectively in case of low value of the parameter associated with GP (blue) and high value of the parameter associated with GP (orange). Common values of the cosine index and distance of collaboration are colored in brown as resulting from the intersection of blue and orange.

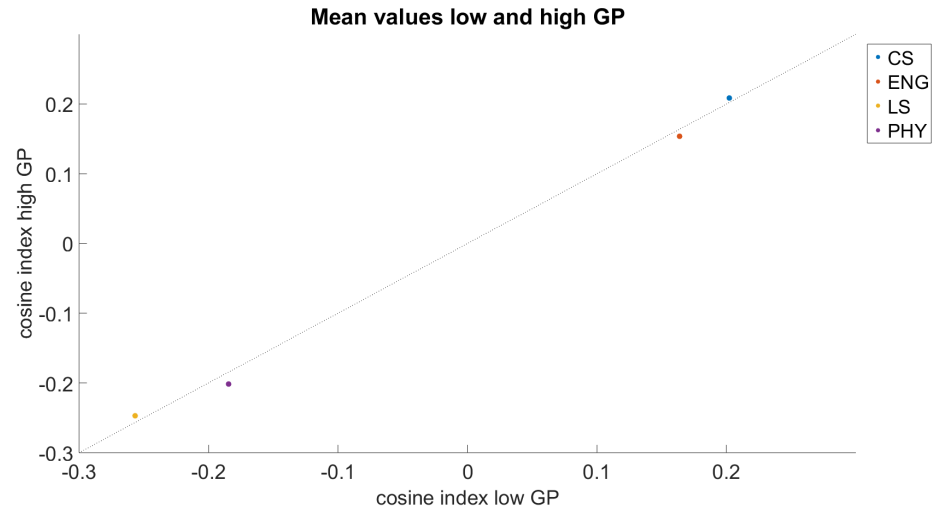


(a)

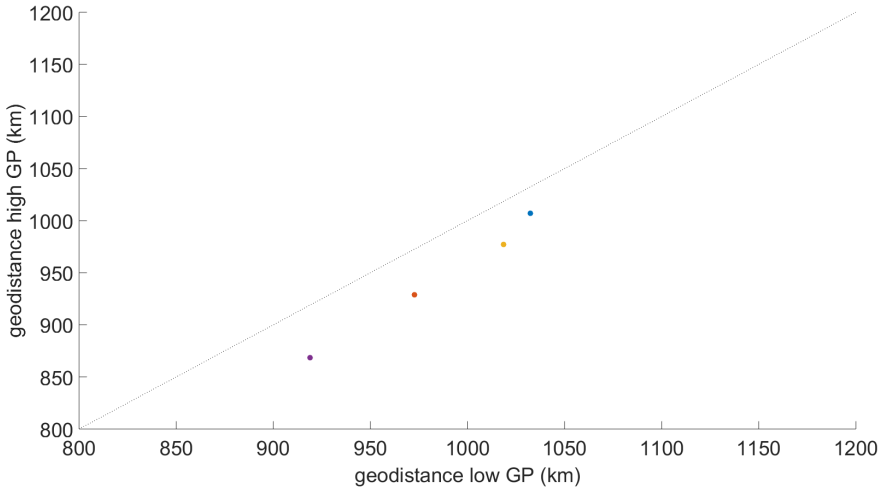


(b)

Figure 8: shows the scatter plot regrading the variations of the cosine index on the variations of the distance of collaboration for each of university department-industrial sector pair in (a) and the identical scatter plot with the additional information concerning the variations of the surplus in (b) when the level of preference concerning GP is switched from low to high. The distances of collaboration and the surplus values have been previously normalized by the amount of couples in the university department-industrial sector pair.

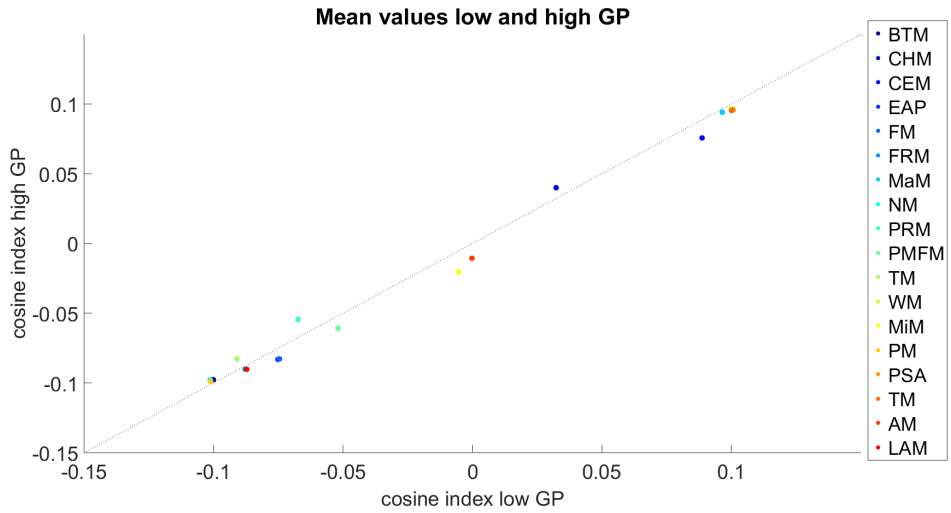


(a)

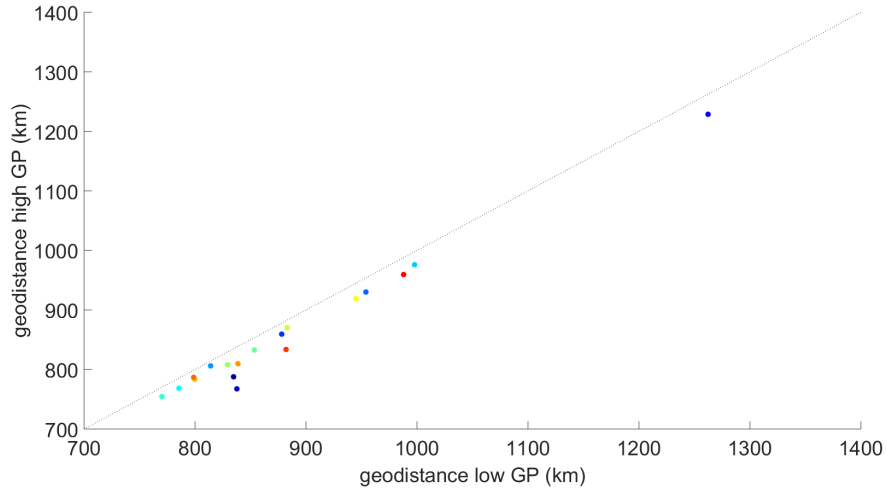


(b)

Figure 9: displays the scatter plot of the mean values of the cosine index generated by the high value of the parameter associated with GP on the mean values of the cosine index generated by the low value of the parameter associated with GP by university department (a) and the scatter plot of the mean values of the distance of collaboration generated by the high value of the parameter associated with GP on the mean values of the distance of collaboration generated by the low value of the parameter associated with GP by university department (b).



(a)



(b)

Figure 10: displays the scatter plot of the mean values of the cosine index generated by the high value of the parameter associated with GP on the mean values of the cosine index generated by the low value of the parameter associated with GP by industrial sector (a) and the scatter plot of the mean values of the distance of collaboration generated by the high value of the parameter associated with GP on the mean values of the distance of collaboration generated by the low value of the parameter associated with GP by industrial sector (b).

Conclusion

The dissertation has presented several applications of the matching model with imperfect transferable utility. The versatility of the theoretical framework proposed resides in its flexibility which allows to accommodate the friction in transfer independently of its nature. Indeed in chapter 1 the friction is originated from government price regulations in the Luxembourg childcare market. In chapter 2 and chapter 3 the frictions are respectively due to the income taxation of the US labor market and to the R&D tax credit of the US market of innovation.

The aim of the type of model proposed is to retrieve the preferences of the agents forming those markets attempting to understand the mechanisms driving the decision of the agents to pair. Indeed in chapter 1 we discuss the impact of the Luxembourg childcare policy on the welfare of the households; in chapter 2 we study the US labor market by measuring the job amenities and labor productivities and by highlighting the effect of the taxation on the jobs mismatch; in chapter 3 we shed light on the interplay between cognitive and geographic proximity when university department and firm decide to collaborate.

Finally, a crucial and interesting feature of that type of model is that it enables to naturally produce counterfactual scenarios, which are relevant in the context of policy recommendations. Indeed in chapter 1 given the preferences of the agents we simulate an alternative equilibrium matching by raising the quality standards of Luxembourg providers and by measuring its effect on the welfare of the households. In chapter 3 lacking the sample of observed matchings we produce simulated scenarios by varying respectively the level of preferences attached to the geographic and cognitive proximity and disclose its impact on the distribution of collaborations university department-firm.

Limitations

As highlighted in Dupuy and Galichon (2015) the matching pairs are typically observed in the data, which implies that only one part of the market is retained (singles are generally excluded). The drawback produced by this involuntary selection hinges on the fact that the identification of the systematic values of the matching is viable up to fixed effects. Specifically, unemployed individuals and inactive firms are not observed in the labor market case, therefore $\alpha(x, y)$ which is the deterministic value attributed to the worker can only be constructed by assuming the direct effect of the characteristics of the firm (y) and the interaction of that characteristics (y) with the features of the worker (x). Conversely, $\gamma(x, y)$ which is the deterministic value of the firm can only be identified by assuming the direct effect of the characteristics of the worker (x) and the interaction of that characteristics (x) with the features of the firm (y). With the appropriate dataset this limitation could be overcome by extending the theoretical framework to model the

whole market.

A further limitation resides in the computation of the sigmas, whose are introduced in the utility of the agents to retain the appropriate level of randomness needed to rationalize the data. Larger values indicate that the matching appears to the econometrician as driven by causality, conversely lower values imply that the matching is merely produced by the observable characteristics (known to the econometrician). The sigma values are generally retrieved by implementing a grid search procedure, which consists in running the optimization process by imposing different sigma values each time. The procedure ends when the sigma values yield the largest maximum value of the likelihood function. This process is clearly time-consuming, especially when the dataset is large and the specification of the benchmark model may include several observables. Indeed this is the main motivation driven the decision to use a subsample of the dataset available in the second chapter.

The time consuming practice associated with the computation of the sigmas deeply affects the dynamic nature of the model. Indeed as noticed along the whole thesis the preferences of the agents have been inferred from cross-sectional data. In principle the dynamicity of the model is ensured by the fact that those preferences may be derived from different cross-sectional data and readily compared since the pattern of matchings as well as the transfers are observed in the markets studied (Dupuy and Galichon (2015)). Indeed, each cross-sectional would then represent a pool of matched agents at different point in time, and it would therefore be possible to trace the change in time of their preferences. In practice this would entail to implement the grid search procedure for each cross-sectional, and the computing cost required to complete this task would be huge (depending on the size of the dataset and specification of the benchmark model).

Future research

In the second chapter we propose a matching model with ITU applied to the US labor market. The model enables to understand the preferences of the agents forming that market and to determine the effect of taxation on the jobs mismatch. As mentioned in the second chapter, the general features of the model facilitate its application to alternative labor markets, *in primis* the European labor market which represents a well-suited candidate.

The implementation of the model to the European labor market would allow to derive the preferences of the agents forming that market as done in the US case, and it would offer the opportunity to address a crucial (labor) migration question (for US and European labor market) which has been unexplored in this thesis: the appraisal of the mobility cost which may greatly influence the decision to move. Indeed the model adopted in chapter 2 may potentially

accommodate this measure by adjusting its specification to keep track of the origin and destination state of the individuals, and therefore price the attitude of the individuals to migrate. As outlined by Borjas et al. (1992) the mobility cost can be split in three types of cost affecting the decision to migrate: transportation, entry barrier and psychological cost. The latter identifies the value individuals attached to *the social, cultural, and physical amenities associated with remaining where he or she was born, including family, friends, and familiarity with old surroundings* (Borjas (2016)). Although the transportation and entry barrier cost in the European and US case may be denied, the effect of the psychological cost may be sizable (Borjas (2016)). The research question is definitely attractive and we expect different outcomes since the European and US labor market present diverse features. Indeed the European labor market appears to be less integrated and homogeneous than the US labor market (Dorn and Zweimüller (2021)). Even if the European as well as the US state borders do not represent legal barriers to the entry of labor migrants, the European migration is far less intense that the interstate US migration (Molloy et al. (2011)). In the European case a crucial obstacle to migrate hinges on the cultural and language barriers as well as on the heterogeneity in education, training and social security systems (Chiswick (2014); Adsera and Pytlikova (2015)). Moreover the European national borders even if they may no longer be considered as legal entity preventing the free mobility of the individuals they still matter for national identity (Dorn and Zweimüller (2021)). According to the survey of the European commission conducted in 2018 most of the Europeans exhibit strong attachment to their own country (57 percent) while few Europeans show similar feelings for the European Union (14 percent) (Eurobarometer and Wave (2018)).

Bibliography

Adsera, A. and Pytlikova, M. (2015). The role of language in shaping international migration. *The Economic Journal*, 125(586):F49–F81.

Borjas, G. J. (2016). *We wanted workers: Unraveling the immigration narrative*. WW Norton & Company.

Borjas, G. J., Bronars, S. G., and Trejo, S. J. (1992). Self-selection and internal migration in the united states. *Journal of urban Economics*, 32(2):159–185.

Chiswick, B. (2014). International migration and the economics of language in chiswick, barry and paul w. miller (eds.), handbook of the economics of immigration.

Dorn, D. and Zweimüller, J. (2021). Migration and labor market integration in europe. *Journal of Economic Perspectives*, 35(2):49–76.

Dupuy, A. and Galichon, A. (2015). A note on the estimation of job amenities and labor productivity.

Eurobarometer, S. and Wave, E. (2018). Standard eurobarometer 89 spring 2018. *European Commission*.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the united states. *Journal of Economic perspectives*, 25(3):173–96.