IMT School for Advanced Studies Lucca

Lucca, Italy

Essays on the Use of Big Data in Developmental Economics

PhD Program in Economics

XXVII Cycle

By

Omar Alonso Doria Arrieta

2016

The dissertation of Omar Alonso Doria Arrieta is approved.

Program Coordinator: Prof. Fabio Pammolli, IMT School for Advanced Studies Lucca

Supervisor: Prof. Fabio Pammolli, IMT School for Advanced Studies Lucca

Supervisor: ,

Tutor: Prof. Alex M. Petersen, Prof. Cristina Tealdi and Prof. Greg Morrison., IMT School for Advanced Studies Lucca

The dissertation of Omar Alonso Doria Arrieta has been reviewed by:

,

,

IMT Institute for Advanced Studies, Lucca

2016

A Carmen mi Madre, Paola mi Hermana y Leidy mi Esposa:
Las mujeres de mi vida.

Contents

List of Figures

List of Tables

Acknowledgements

Acknowledgments may be optional except in either of the following circumstances.

1. I thank to Prof. Fabio Pammolli and Prof. Alexander M. Petersen as coauthors of the first chapter. I also thank to Cristina Tealdi for helpful discussions. I acknowledge funding from the National Research Program of Italy (PNR) project "CRISIS Lab".
2. For Chapter 2, I wish to thank Prof. Daniel Toro, Prof. Jorge Alvis and Prof. Aaron Espinosa from the Faculty of Economics and Business of the Universidad Tecnológica de Bolivar (UTB) and the Ibero-American Laboratory of Research and Innovation in Development and Culture ("L+iD", by its name in Spanish), for receiving and sharing data and important comments. I am also grateful to Juan Mauricio Ramirez from Fedesarrollo, Colombia for sharing data. I would also like to acknowledge to Prof. Cristina Tealdi, Prof. Alex Petersen, and Prof. Rodolfo Metulini from the IMT School for Advanced Studies Lucca, for their assistance and important comments that made this work possible.
3. For Chapter 3, I wish to thank Prof. Greg Morrison as coauthor.
4. The student reproduces/reprints copyrighted material requiring permission to be reprinted/reproduced in which case the student is responsible for acquiring and acknowledging each permission to reprint/reproduce.
5. The student uses as text in a chapter either material based on co-authored published or about to be published arti-

cles or material based on co-authored papers in progress. The student should identify all co-authors, the journal where the article can be found and the journal publisher.

Vita

1982 Born, Cartagena de Indias, Colombia.

2009 Degree in Physics
Thesis Score: 4.5/5.0.
Universidad Nacional de Colombia.

2011 Master of Science in COMPUTATIONAL PHYSICS
Thesis Score: 9.5/10.0 cum laude
University of Barcelona & Barcelona Tech. Spain.

2015 Consultant Data Scientist
The Cambridge Management Consulting Labs.
Milan, Italy.

2015 Data Scientist
Sadako.
Barcelona, Spain.

2014 Consultant.
United Nations Development Programme (UNDP).
Cartagena de Indias, Colombia.

2011 Researcher and Data Modeling.
Hospital Clinic de Barcelona.
Barcelona, Spain.

Publications

1. O. Doria, "Ronchi Test in a Concave Mirror," in *Revista Colombiana de Física*. Vol. 38. N. 2.

Presentations

1. O. Doria, "Deprivation and Trade. Evidence from Colombia.," at *Universidad Tecnologíca de Bolivar.*, Cartagena de Indias, Colombia, 2014.
2. O. Doria, "Numeric Model of a Flow trough an Airfoil using Conformal Transformations in complex space," at *Universidad Nacional de Colombia.*, Bogotá, Colombia, 2008.
3. O. Doria, "Ronchi Test in a concave mirror," at *Physics National Meeting.*, Barranquilla, Colombia, 2006.
4. O. Doria, "Numeric approximation of Zernike's Polynomial for a concave Topology," at *Optical National Meeting. IX ENO*, Medellín, Colombia, 2005.

Abstract

In this thesis I approach problems within the literature of Development Economics. Using tools from policy evaluation, different quantitative methods and big data sources, I study the common problems that affect the development of the nations. I separate my thesis into three chapters. In Chapter 1, using policy evaluation techniques together with other quantitative methods, I study the effects of the policy integration for the academic sector within the European Union. In the second chapter, I study one of the most important subjects presented in this thesis: inequality. Using the case of Colombian municipalities, I examine how international trade affects social conditions measured by the Multidimensional Poverty Index. Finally, in the third chapter, I study the effect of the patent innovation using the complexity algorithm developed by Hidalgo and Haussmann. Here I do a comparative of the patent innovation using two aggregations: countries and cities. Next, I will explain in more detail the findings of each chapter.

Chapter 1: It is generally accepted that the frequency of cross-border collaborations has been increasing in recent decades, which is principally regarded as a symptom of globalization. While this is true on average, we uncover a more nuanced story by analyzing publications, and by disaggregating these R&D outputs by country, across 14 well-defined research subject areas. In this way, we are able to interpret trends in cross-border activities according to more domain-specific trends. We focus our analysis on the impact of entry into the EU by new member states by quantifying the rate of cross-border collaboration before and after the 2004 enlargement of the European Union. In this sense, we build upon recent studies aimed at quantifying the impact of European Research Area integration policies on the activity of the European innovation system. We combine descriptive complex networks techniques with panel regression (Difference in Difference) methods to reveal, counterintuitively, a decrease in cross-border activity by the new EU member states following their entrance. The results show that while the number of cross-border collaborations in academia is increasing in the old member countries, and despite that the number of cross-border publications in the new members is higher compared with past years, they would actually collaborate more being outside the European Union. We use data for the inventor mobility network to show that these counterintuitive trends are none other than the negative externalities of unification associated with brain-drain. Chapter 2: We empirically measure the effects of international trade on inequality. By studying the Colombian case, we found that the municipalities with exporter firms have an 11% greater probability for increasing their inequality through social deprivations compared to municipalities, where any of the firms are exporters. Furthermore, we aggregate firms' exports at the municipality level by using the minimum economical unit affected by the incoming wealth from foreign markets, considering spatial relations to account for direct, indirect and total effects. We define social inequality as the average

shortfall of social conditions by municipality. Specifically we use the Colombian Multidimensional Poverty Index. As a result, we found empirical evidence for a strong neighborhood effect, which helps make the decision that would be used to improve social conditions in those municipalities without exporter firms. Chapter 3: One of the most important questions in Developmental Economics is how technological innovation is able to shift development. Here, we use the Hidalgo and Hausmann complexity algorithm to estimate how the selection of the innovation field affects the leadership in innovation among countries by using the first patent of triadic families of the European, Japanese and United States patent office. In this analysis we rank countries, regions and patents using the Economic Complexity Index (ECI) and Product Complexity Index (PCI). Our findings highlight the United States as a leading country in patent innovation during most of the years. In contrast, using the region aggregation level, we find that the Japanese regions are the leaders in patent innovation during every year in our data. However, the most complex regions in the Unites States. On the patent side, we note that the fields related to chemistry, biotechnology and pharmaceuticals play a very important role in patent innovation. Finally, we compare our findings with similar works of other researchers, finding a strong relationship between academic research and patent innovation.

In this thesis, I explore quantitative methods and Data Science techniques applied to social studies for different aggregations. In the first chapter, I account for the collaboration in R&D between groups of countries. In the second chapter, I explore how one single country and its municipalities relate to the world through trade, and how its relationship could modify the condition, while its minimal political level affects the social conditions through a mechanism accepted and studied by classical economics. In Chapter 3, we study the behavior of nations, where every nation with innovative production is included. By alternating the level of aggregations from a nation to a city, we are able to determine the role that political regions play within the whole country. Recognizing the importance of the results and the methods that I use to explore these important development issues, such as international integration, inequality, international trade, innovation and relation country-cities, this thesis questions the classical economical methods and the relevance in accepting new methodologies based on Data Science to answer the question of classical problems that were answered by using theoretical methods in the past. These new methodologies based on Big Data are evolving every day with importance, and, moreover, with information collected by governments, social networks, international organizations and private institutions. Therefore, the methods exposed in this dissertation play a fundamental role in the reshaping of economics and the study and interpretation of the relation between governments, societies and citizens.

Chapter 1

Measuring the impact of European integration on the rate of cross-border collaboration and high-skilled labor mobility

1.1 Introduction

We use the 2004/2007 European Union (EU) enlargement by 12 member states to quantify the impact of EU efforts to expand and integrate the scientific competitiveness of the European Research Area (ERA). Using the synthetic control method applied to cross-border collaboration data extracted from millions of academic publications disaggregated across 14 subject areas and 32 European countries from 1997–2012, we show that levels of European cross-border collaboration would have been higher without EU enlargement, despite the new 2004/2007 EU entrants gaining access to EU resources incentivizing cross-border integration. To further illustrate the unintended consequence of the EU expansion, we use of-

ficial high-skilled mobility statistics to identify brain drain – principally east-to-west – as a major factor underlying the divergence in cross-border integration between western and eastern Europe. These results challenge central tenets underlying ERA integration policies, namely that unifying labor markets will increase the international competitiveness of the ERA.

Despite positive trends in the globalization of R&D, recent studies of international collaboration show that national borders are still a formidable hindrance to scientific integration, notwithstanding directed policies aimed at reducing barriers – as specifically is the case for the European Research Area (ERA) (Boyle, 2013; Chessa et al., 2013; Hoekman et al., 2009; Morescalchi et al., 2015). As a result, the globalization of science via international collaboration has not evolved uniformly across all countries and regions. For example, comparing the decade before and after 2004, while Western Europe and North America experienced a 36-42% increase in the rate of cross-border collaboration (per publication), Eastern Europe and Asia have experienced much slower 9% growth (see Fig. 1). These diverging trends point to the importance of historical, socio-technological, and geographic factors affecting the globalization of science (Delanghe et al., 2009; Geuna, 2015; Lepori et al., 2015; Scherngell, 2013).

So why have Western and Eastern Europe followed different cross-border collaboration paths? To provide insight into this phenomena, we constructed a longitudinal dataset for 32 European countries over the 16-year period 1997–2012 by aggregating annual (i) publication count, citation count, and international collaboration rate data (disaggregated across 14 research subject areas indexed here by s, e.g. s = 1300: "Biochemistry, Genetics, and Molecular Biology") from SCImago Journal & Country rank; (ii) government investment in R&D data from the World Bank; (iii) official country-country pairwise counts of incoming/outgoing EU high-skilled labor mobility from the EU Single Market Regulated professionals database (*European Commission: The EU Single Market Regulated professionals database (professionals moving abroad)*., n.d.); and (iv) global migration data from Abel & Sander (Abel and Sander, 2014) (see the Supplementary Appendix for further description of these datasets). In what follows, we show that high-skilled mobility – 'brain drain' (Ack-

ers and Gill, 2008; Ackers, 2005; Beine et al., 2001; Gibson and McKenzie, 2011; Grossmann and Stadelmann, 2011; Weinberg, 2011) – explains a significant portion of the East-West EU divergence following the integration of European labor markets. The mechanism underlying this effect is, rather, intuitive: Europe experienced a significant loss of cross-border integration because as mobile academics pursued career paths by following their previous collaboration channels, they subsequently brought – thereby eliminating – their cross-border links with them.

The outline of this chapter is as follows. We first outline the data used in our analysis, which were aggregated across several open data portals. In particular, we explain in detail the high-skilled mobility data which is central to our analysis, providing a graphical and numerical description of the data and developing quantitative country-level measures to be used later in our regression analysis. In the Results section we start with a demonstration of the Synthetic Control Method which identifies the disparity between the cross-border collaboration rates between the two groups of countries identified in our analysis – the 2004/2007 entrants and the complementary set of European countries. We then discuss the high-skilled mobility data in more detail and carry out a country fixed-effects panel data regression model which incorporates two important features:

1. the EU enlargement which is captured in a difference-in-difference dummy variable which captures the before-after shift of entrant countries entering into the "EU member" group status relative to the countries which do not change their "EU member" group status through the entire analysis, and

2. the relative amount of brain drain, as proxied by high-skilled labor mobility in or out of a country, in a given year.

In the Discussion section we consider in depth the implications of the regression estimates for features (1) and (2), in particular. And in the Conclusion section we summarize the policy implications of our analysis and results.

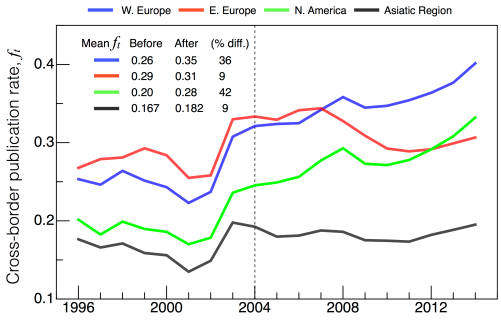


Figure 1: Eastern - Western European divergence. Global trends in cross-border collaboration by international region: 1996–2014. Source: SCImago Journal & Country Rank based on Scopus (*SCImago: SJR SCImago Journal and Country Rank.*, n.d.). Notably, the curves for W. Europe and E. Europe are, prior to 2004, characterized by a roughly constant offset, thereby satisfying the prior equal slopes condition of the difference-in-difference framework.

1.2 Materials and Methods

1.2.1 Countries analyzed

We analyzed 32 European countries over the 17-year period 1996-2012: Austria (AT), Belgium (BE), Bulgaria (BG), Croatia (HR), Cyprus (CY), Czech Republic (CZ), Denmark (DK), Estonia (EE), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Iceland (IS), Ireland (IE), Italy (IT), Latvia (LV), Liechtenstein (LI), Lithuania (LT), Luxembourg (LU), Malta (MT), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Romania (RO), Slovakia (SK), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), United Kingdom (GB). These countries can be grouped according to EU entry year: $g_{EU,i} = 1$ if existing EU member in 2004, $g_{EU,i} = 2$ if part of the 2004 EU enlargement (CY, CZ, EE, HU, LT, LV, MT, PL, SK, SI), $g_{EU,i} = 3$ if part of the 2007 EU enlargement (BG, RO), and $g_{EU,i} = 4$ (CH, HR, IS, LI, NO) if not part of the EU as of the end of 2012, corresponding to the final year of our analysis.

1.2.2 Publication data

We downloaded comprehensive publication data from SCImago Journal & Country rank, which is calculated using comprehensive *Scopus* data (*SCImago: SJR SCImago Journal and Country Rank.*, n.d.). From this data repository, we gathered four time series for each country i and each subject area s: (i) the total number of publications, $D^s_{i,t}$, (ii) the total number of citations received in year t, $C^s_{i,t}$, (iii) the fraction of publications involving international collaboration, $f^s_{i,t}$, and (iv) the total number of publications involving international collaboration $Y^s_{i,t} = f^s_{i,t} D^s_{i,t}$.

We analyzed 14 subject areas (indexed by s): "Agricultural and Biological, Sciences" (1100), "Biochemistry, Genetics, and Molecular Biology" (1300), "Business Management and Accounting" (1400), "Chemical Engineering" (1500), "Chemistry" (1600), "Computer Science" (1700), "Decision Sciences"(1800), "Energy" (2100), "Engineering" (2200), "Environmental Science" (2300), "Materials Science" (2500), "Medicine" (2700), "Pharmacology,Toxicology, and Pharmaceutics" (3000), "Physics and Astronomy" (3100).

In order to account for the censoring bias associated with the measurement of citations (publications from recent years have had less time to accrue citations than older publications), we normalized $C^s_{i,t}$ within s and t according to the logarithmic transform, $R^s_{i,t} \equiv (\ln C^s_{i,t} - \langle \ln C^s_{i,t} \rangle)/\sigma[\ln C^s_{i,t}]$, where $\langle ... \rangle$ and $\sigma[...]$ are the mean and standard deviation calculated within each s and t group, respectively. $R^s_{i,t}$ measures the scientific reputation of country i in subject area s in year t. Moreover, we confirmed that $R^s_{i,t}$ is approximatley distributed according to the $Normal(0,1)$ baseline distribution, independent of t. Thus, $R^s_{i,t}$ is comparable across both s and t, being independent of the disciplinary and censoring bias that are problefmatic in the comparison of raw citation counts.

1.2.3 World Bank country-level R&D data

As controls for country investment in R&D, which are for example related to the level of internationalization of higher educational institutions (Lepori et al., 2015), we used researcher population, government

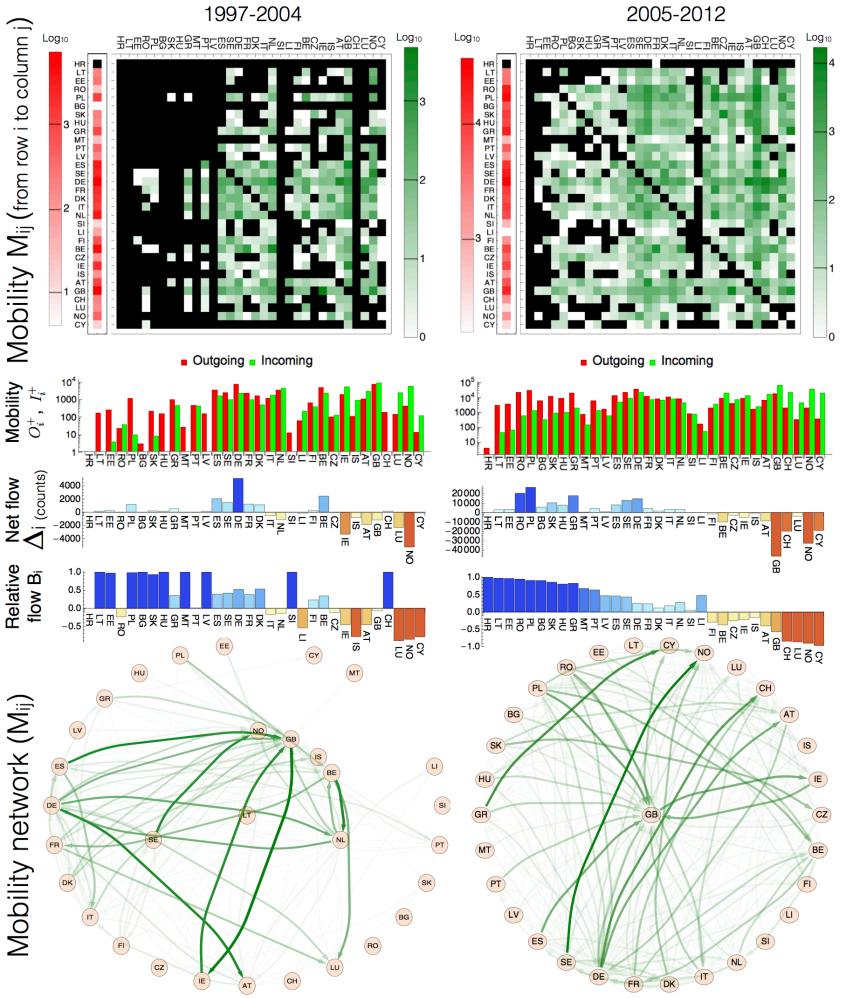


Figure 2: High-skilled mobility before and after the 2004 enlargement. Total mobility counts at the dyadic country-country level, M_{ij}, and aggregated at the country level: total outgoing O_i^+, incoming I_i^+, net flow out $\Delta_i = O_i^+ - I_i^+$, and relative brain-drain $B_i = (O_i^+ - I_i^+)/(O_i^+ + I_i^+)$. The red color scale to the left of each M_{ij} matrix visualization represents $\log_{10} O_i^+$, the total mobility out of country i (black cells indicates $\Delta_i < 4$ for 1997-2004 and $\Delta_i < 155$ for 2005-2012). The green color scale to the right of each M_{ij} visualization represents $\log_{10} M_{ij}$. The network links also have thickness/opacity nonlinearly related to $\log_{10} M_{ij}$ so that only the most prominent links are visible. Color values are not comparable across time periods. We use a circular layout which explicitly puts GB in the center in order to emphasize its central role.

spending, and gross domestic product (GDP) data from the World Bank data repository (*World Bank data sources.*, n.d.):

1. "Researchers in R&D (per million people)", given by $Spc_{i,t}$, with mean $\pm$ standard deviation $= 2,900 \pm 1,700$;

2. The total number of researchers in R&D, given by $S_{i,t}$ (calculated using Population data in combination with $Spc_{i,t}$), with mean $\pm$ std. dev. $= 47,000 \pm 72,000$;

3. "Research and development expenditure (% of GDP)", given by $e_{i,t}$, with mean $\pm$ std. dev. $= 1.47 \pm 0.89$; We then use GDP data to convert $e_{i,t}$ to the total R&D expenditure, $E_{i,t}$;

4. "GDP (current US$)", given by $GDP_{i,t}$, with mean $\pm$ std. dev. of $\log_{10} GDP_{i,t} = 11 \pm 0.75$; and

5. "GDP per capita (current US$)", given by $GDPpc_{i,t}$, with mean $\pm$ std. dev. of the log value $(\log_{10} GDPpc_{i,t}) = 4.4 \pm 0.36$.

We deflated all dollar amounts to 2010 USD$. Averaging across 32 EU and 57 large non-EU countries, the average annual growth rate of $S_{i,t}$ is 4-6%, and the average annual growth rate of the total R&D expenditure is between 8-9%; over this period, there is little difference between the EU and non-EU growth rates of total R&D expenditure (Pan et al., 2015).

1.2.4 Mobility data (EU High-skilled)

Competitiveness in the global economy is increasingly becoming linked to the high-skilled "knowledge" economy (Brown et al., 2001). And while Europe is certainly producing a large number of high-skilled laborers, it is also home to large stocks of high-skilled emigrants (Docquier and Rapoport, 2012), in particular scientists (Ackers, 2005; Geuna, 2015). The study of scientific researcher mobility has been aided by large publication datasets (Deville et al., 2014; Moed et al., 2013; Noorden, 2012), facilitating new studies into the supply-demand for researchers, which

can oftentimes be linked to specific policies and programmes. However, the availability of comprehensive researcher career data, as well technical (name disambiguation) problems that exist when attempting to extract researcher trajectories from raw publication metadata, mean that researcher mobility data is difficult to acquire and certainly not comprehensive in its coverage of all scientists.

As a proxy for researcher mobility trends, we used official EU Commission "Professionals moving abroad (Establishment)" data from the The EU Single Market Regulated professionals database. This database tracks the number of (high-skilled) professionals who obtained official certification in a given country of qualification (source country), and then applied for official recognition of their professional certification in a particular host country (destination country) (*European Commission: The EU Single Market Regulated professionals database (professionals moving abroad).*, n.d.). Lacking the mobility outcome data, we assume that the actual number of migrating professionals is highly correlated with the number of positive decisions to recognize the professional certification in a given destination country – i.e. we assume that if an individual has their application approved then they move with high probability. As such, we also assume that the information captured by the high-skilled mobility data is highly correlated to scientific mobility trends over the same period. The database covers a variety of certification "Recognition Regime" categories (e.g. "Pharmacist", "Doctor in basic and specialized medicine both listed in Annex V", etc.). We aggregated the data for all professions using the option "Recognition Regime =All". For a more specific description of their counting methods and the outcome statistics, see the data description page.

The data are grouped into 13 periods indexed here by $t = 1...13$ corresponding to 1997/1998, 1999/2000, 2001/2002, 2003/2004, 2005/2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014. We did not include the final 2 years of data in our analysis because the mobility data was either incomplete or still being updated and because the World Bank R&D data is incomplete for many countries after 2012. It is also worth explicitly stating that we divided the mobility headcount variables for periods in

$t \leq 2006$ by a factor of two so that these count values refer to mean annual rates. As such, in order to combine observations across these three datasets, it was also neccesary to aggregate the count data for publications and country-level economic indicators across the specified 2-year periods and then divide by a factor of 2, resulting in 2-year annual averages.

Thus, for each year period t we recorded $M_{ij,t}$, the total number of high-skilled migrations ("Total positive decisions") from country i ("Country of qualification") to country j ("Host country"). In all, the total mobility (headcounts) for a given time period x, $M_x = \sum_{ij} M_{ij,x}/2$, are 315,888 (1997–2012), 43,075 (1997–2004), and 272,813 (2005–2012). We also recorded the number of "Total negative decisions", $N_{ij,t}$, corresponding to those applications which were denied (for a variety of reasons). The total number of negative decisions by period are 24,046 (1997–2012), 4,734 (1997–2004), and 19,312 (2005–2012), representing roughly 7% of the total (positive and negative) decisions made.

We used this data to analyze the intra-EU mobility rates before ($<$) and after ($>$) the 2004 EU enlargement. The total incoming mobility before and after are given by $I^+_{i,<} = \sum_{j,t\leq 2004} M_{ji,t}$ and $I^+_{i,>} = \sum_{j,t\geq 2005} M_{ji,t}$, respectively; the total outgoing mobility before and after are then given by $O^+_{i,<} = \sum_{j,t\leq 2004} M_{ij,t}$ and $O^+_{i,>} = \sum_{j,t\geq 2005} M_{ij,t}$, respectively. Furthermore, the negative decisions can also be aggregated by country: $I^-_{i,<} = \sum_{j,t\leq 2004} N_{ji,t}$ and $I^-_{i,>} = \sum_{j,t\geq 2005} N_{ji,t}$ and $O^-_{i,<} = \sum_{j,t\leq 2004} N_{ij,t}$ and $O^-_{i,>} = \sum_{j,t\geq 2005} N_{ij,t}$. At the annual level, $I^y_{i,t}$ and $O^y_{i,t}$ refer to total incoming and outgoing counts within period t and decision type $y = \pm$.

The "success rate" of outgoing (incoming) applications contains information about the competitiveness (selectivity) of the source (host) country. We define the incoming and outgoing success rates using the relative frequency of positive (+) and negative (-) decisions, $\mathcal{P}^{in}_{i,t} = I^+_{i,t}/(I^+_{i,t} + I^-_{i,t})$ and $\mathcal{P}^{out}_{i,t} = O^+_{i,t}/(O^+_{i,t} + O^-_{i,t})$, respectively. As above, we use the notation $\mathcal{P}^{in}_{i,<}$ referring to the net success rates calculated by aggregating periods $t \leq 2004$, and $\mathcal{P}^{in}_{i,>}$ referring to the net success rates calculated by aggregating periods $t \geq 2005$. These success rates can also be generalized to country-country pairs at variable time resolutions ($x = \{t, <, >\}$) accord-

ing to the definitions $\mathcal{P}^{in}_{ij,x} = I^{+}_{ij,x}/(I^{+}_{ij,x} + I^{-}_{ij,x})$ and $\mathcal{P}^{out}_{ij,x} = O^{+}_{ij,x}/(O^{+}_{ij,x} + O^{-}_{ij,x})$.

We use the Gini index $G^{in}_{i,t}$ ($G^{out}_{i,t}$) to measure the concentration of the incoming (outgoing) mobility across the other EU member states. For example, $G^{out}_{i,t}$ is calculated using the 31 possible destination countries (j) of country i in the mobility network as $G^{out}_{i,t} = (\sum_{j=1}^{31}\sum_{k!=j}^{30}|M_{ij,t} - M_{ik,t}|)/(2(31-1)^2\langle M^{out}_{i}(t)\rangle)$ where $\langle M^{out}_{i}(t)\rangle$ is the average outgoing mobility of i in t; $G^{in}_{i,t}$ is calculated by swiching the order of i and j, k in the matrices to represent incoming counts. $G_{i,t}$ is particularly useful in our case because it is standardized over the fixed unit interval $[0, 1]$, thus it is less sensitive to the large variations in $M_{ij,t}$: the minimum value 0 represents the case in which the mobility is dispersed evenly across all the other countries, and the maximum value 1 represents the case in which the mobility is entirely concentrated on one country with no mobility to any other countries. Thus, this quantity controls for the strong variation in the incoming and outgoing links from any given i in the mobility network (see Fig. 2).

We define the relative brain drain $B_{i,x} = (O^{+}_{i,x} - I^{+}_{i,x})/(O^{+}_{i,x} + I^{+}_{i,x}) \in [-1, 1]$. This quantity measures the mobility polarization, with extreme values $B_{i,x} = -1$ corresponding to $O^{+}_{i,x} = 0$ and $I^{+}_{i,x} > 0$ (entirely incoming mobility) and $B_{i,x} = 1$ corresponding to $O^{+}_{i,x} > 0$ and $I^{+}_{i,x} = 0$ (entirely outgoing mobility). As illustrated in Fig. 2, this measure takes on positive values when there is more mobility out of a country i ('brain drain') than mobility into a country i ('brain gain'), and is useful as a relative measure to compare countries with total mobility rates that differ across several orders of magnitude.

1.2.5 Total migration data

In order to account for underlying global migration trends, we used data from Abel & Sander (Abel and Sander, 2014), who provide estimates of the bilateral migration (high-skilled + low-skilled) between countries i and j, given by the matrix $\tilde{M}_{ij,\tau}$, which they calculated aggregating official country statistics over three 5-year periods, $\tau = 1$ (1995–2000),

$\tau = 2$ (2000–2005), and $\tau = 3$ (2005–2010). This novel dataset uses sequential population stock tables, including census data about birthplace and refugee and population statistics, to reconstruct and estimate the aggregate $\tilde{M}_{ij,\tau}$ headcount data.

Here we use this data to calculate the analogs of the total mobility (I/O) and diversity (G) measures described above: the total migration from (to) country i given by $\tilde{O}_{i,\tau}$ ($\tilde{I}_{i,\tau}$) and the Gini index of the migration from (to) country i given by $\tilde{G}^{out}_{i,\tau}$ ($\tilde{G}^{in}_{i,\tau}$). We approximate the global migration data for $t = 2011, 2012$ using the $\tilde{M}_{ij,\tau}$ values for 2005-2010. As above for the high-skilled mobility, we also define the relative brain drain $\tilde{B}_{i,\tau} = (\tilde{O}_{i,\tau} - \tilde{I}_{i,\tau})/(\tilde{O}_{i,\tau} + \tilde{I}_{i,\tau}) \in [-1, 1]$, which measures the net migration flow as a percentage of the total migration in and out of the country i, which we use as a control in our regression model for total mobility rates.

1.3 Results

1.3.1 Synthetic Control Method

In order to measure the impact of the 2004 and 2007 EU enlargement on the rate of international collaboration in Europe, we used a combination of causal inference methods – the Synthetic Control Method (SCM) (Abadie et al., 2010) and a difference-in-difference (DiD) panel regression model. In each method we use the EU enlargement – 10 entrants in 2004 (CY, CZ, EE, HU, LT, LV, MT, PL, SK, SI) and two entrants in 2007 (BG, RO) – as a multi-country 2-stage policy intervention corresponding to the (treatment) years $t^* = 2004$ and 2007, respectively. As such, we separated the European countries into two groups, with the first comprising the 17 incumbent 2004 members, and the second comprising the 12 entrant countries. Then, for each country i, we analyzed the fraction $f_{i,t}$ of the total publications ($D_{i,t}$) that involved cross-border collaboration in year t, and the total number $Y_{i,t}$ of cross-border publications – before and after t^*. Implicit in our statistical methods are controls for global trends in cross-border collaboration, e.g. a sharp increase in $f_{i,t}$ –

within the EU and abroad around 2002 – possibly resulting from the 6th EU framework programme (FP6) which was the first to broadly include specific international collaboration criteria in its funding schemes.

In order to estimate the causal impact of the EU enlargement on the $f_{i,t}$ and $Y_{i,t}$, we used the SCM to estimate the counterfactual scenario – no EU enlargement – by extrapolating synthetic $\hat{f}_{i,t}$ and $\hat{Y}_{i,t}$ for $t > 2004$ based upon a control set of 26 non-European countries. More specifically, we estimated $\hat{f}_{i,t}$ and $\hat{Y}_{i,t}$ using a panel dataset comprised of 4 covariate time series: the total number of publications ($\log_{10} D^s_{i,t}$), the normalized citations ($R^s_{i,t}$), the per-capita GDP ($\log_{10} GDPpc_{i,t}$), and government expenditure on R&D as a % of GDP, $e_{i,t}$. In short, the SCM estimates an optimal set of weights allowing for the extrapolation of $\hat{Y}_{>}(\hat{f}_{>})$ for the EU entrant countries based upon their projection onto a subspace of external data representing 26 non-European control countries (see the Supplementary Appendix for further SCM details).

Fig. 3 shows the empirical curves (f_t and Y_t) measuring the cross-border activity for the average country within each group of EU states (incumbent and entrant), as well as the SCM estimates ($\hat{f}_t$ and $\hat{Y}_t$). The difference between $\hat{f}_{i,t}$ ($\hat{Y}_{i,t}$) and $f_{i,t}$ ($Y_{i,t}$) is thus a measure of the impact of EU enlargement, and its associated policies, on the scientific integration and the international competitiveness of the EU. For f_t, the counterfactual difference δ is the mean difference between $\hat{f}_t$ and f_t for $t \geq 2005$; for Y_t, we calculate the counterfactual difference as the percentage difference between $\hat{Y}^{>} = \sum_{t \geq 2005} \hat{Y}_t$ and $Y^{>} = \sum_{t \geq 2005} Y_t$, with $\delta(\%) = 100 \times (\hat{Y}^{>} - Y^{>})/Y^{>}$.

By and large, the difference in the real and synthetic curves indicate that the 2004 entrants lost out on collaborative integration within the global science system by entering the EU – suffering a 0.062 decrease in f_t and a 9% decrease in Y_t. Interestingly, the incumbent pre-2004 EU countries also suffered a 15% decrease in Y_t, however, the per-publication rate f_t was actually better with the EU enlargement as compared to the counterfactual, indicating a disparity between the incumbent and new EU member countries.

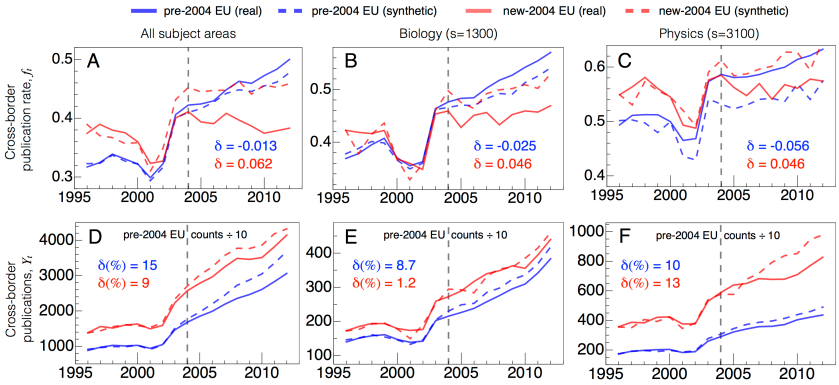


Figure 3: Comparing synthetic (counterfactual) and actual cross-border collaboration after the 2004 EU enlargement. The fraction f_t of cross-border publications (A-C) and the total number Y_t of cross-border publications (D-F), by subject area. The dashed curves represent the estimates, $\hat{Y}_t$ and $\hat{f}_t$, measuring the counter-factual cross-border activity – had the new 2004 EU members not joined the EU. Estimates are made using the Synthetic Control Method (Abadie et al., 2010), implemented using a control group of 26 non-EU countries to best-fit Y_t (f_t) for $t < 2004$ and then to extrapolate $\hat{Y}_t$ ($\hat{f}_t$) for $t \geq 2004$. Note that the Y_t representing the incumbent pre-2004 EU countries are divided by 10 in order to facilitate visualizing all the curves on the same scale. δ and $\delta(\%)$ represent the difference between the real and synthetic curves after 2004, providing estimates of the "2004 EU enlargement" effect on cross-border European integration.

1.3.2 High-skilled mobility in Europe: 1997-2012

To identify the source of this disparity in cross-border activity, we analyzed the net flow of high-skilled labor across Europe over the period 1997–2004. We denote the high-skilled mobility (head counts) from country i to country j in year t as $M_{ij,t}$, and then aggregate these counts over the two time periods 1997–2004 ($<$) and 2005–2012 ($>$). Totaling $M_{ij,t}$ across all EU countries gives the total outgoing $O^+_{i,t} = \sum_{j,t} M_{ij,t}$ and incoming $I^+_{i,t} = \sum_{j,t} M_{ji,t}$ mobility, from and to country i. Together, these mobility statistics indicate a 7-fold increase in intra-EU high-skilled mobility after the 2004 enlargement. However, this increased labor flow was not distributed evenly across the member states, which we captured by measuring incoming (outgoing) mobility Gini-index $G^{in}_{i,t}$ ($G^{out}_{i,t}$) for each country based on the incoming and outgoing dyadic flows. Indeed, after 2004, 29% of the mobility was from Eastern Europe (defined as 2004/2007 EU entrants) to Western Europe (defined as incumbent and non-EU countries CH, HR, IS, LI, NO), a significant increase over the 5% value observed for the pre-2004 period.

By analyzing the net flow between two countries $\Delta_{ij,t} \equiv |M_{ij,t} - M_{ji,t}|$, we also captured the information contained in the brain-drain network. Fig. 4 compares the brain-drain networks $\Delta_{ij,<}$ and $\Delta_{ij,>}$, illustrating the notable shifts in high-skilled labor concentration. At a country level, we measure the net brain drain $\Delta_{i,t} = O^+_{i,t} - I^+_{i,t}$ and the relative brain drain $B_{i,t} = \Delta_{i,t}/(O^+_{i,t} + I^+_{i,t})$.For example, the mean brain-drain values after 2004 for the incumbent and entrant countries are $\langle B^{\text{old EU}}_{>} \rangle = 0.06$ and $\langle B^{\text{new EU}}_{>} \rangle = 0.53$, respectively; the countries with significant brain-gain, i.e. $B_{i,>} \leq -0.5$, were CY, LU, and GB, whereas PT, GR, MT, HU, SK, BG, PL, RO, EE, and LT were the countries with the largest brain drain, i.e. $B_{i,>} \geq 0.5$ (see Fig. 2).

Figure 2 shows the high-skilled mobility matrix, before $(M_{ij,<})$ and after $(M_{ij,>})$ the 2004 enlargement. The countries are ordered according to decreasing B_i calculated across the entire period 1997–2012 (the country with largest relative brain drain was HR, and the country with largest relative brain gain was CY). In order to visualize the pairwise

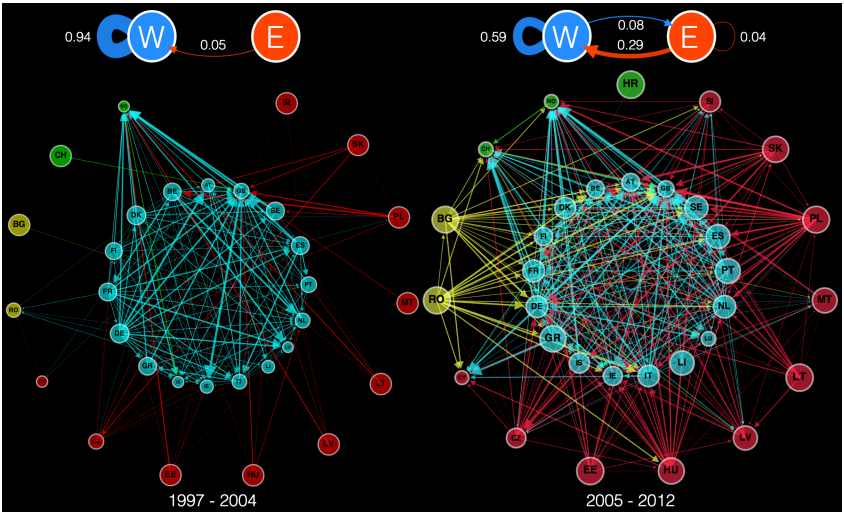


Figure 4: High-skilled brain-drain networks (Δ_{ij}), before and after the 2004 EU enlargement. (top) Mobility between the 2004/2007 entrant countries ("E") and the rest of the incumbent European countries ("W"). The networks in each period are calculated from a total of 43,075 head counts (1997–2004) and 272,813 head counts (2005–2012), respectively. Link thickness (shown) represents the fraction of the total mobility, with link direction the same as the source node. (bottom) The node color represents the EU entry year group ($g_{EU,i}$); the node size is proportional to the relative brain drain, $1 + B_i$ (larger values indicating larger mobility out of country i); link thickness is proportional to $\log(|\Delta_{ij}|)^2$ between countries i and j, with the arrow pointing in the direction of the net flow and link color corresponding to the source node. The size/thickness scales used for both networks are the same.

mobility counts, which can range across several orders of magnitude, we show $\log_{10} M_{ij}$. Comparing the periods 1997-2004 to 2005-2012, the total mobility across all countries increased roughly 7-fold, from $M_<$ =43,075 (1997–2004) to $M_> = 272{,}813$ (2005–2012). The significant increase in high-skilled labor mobility was distributed across all the European countries, thereby resulting in a reorganization of the entire mobility network, as some countries transitioned from being major sources to major recipients of brain gain (e.g. CH). One constant across the two time periods is the role played by Great Britain as the major mobility hub, which largely gained from the EU enlargement, going from a relatively small sink before the enlargment ($B_{GB,<} \approx -0.1$), to a relatively large sink afterwards ($B_{GB,>} \approx -0.6$).

The regulated professionals mobility data also contains the application success rates for professional license transfer. Because application approval is a precondition for migration, it serves as an additional quantitative indicator of each country's competitiveness ($\mathcal{P}_i^{out}$, as in the case of outgoing mobility) and selectivity ($\mathcal{P}_i^{in}$, as in the case of incoming mobility). Interestingly, Cyprus and the Czech Republic, two of the wealthiest countries over the entire study period in terms of per capita $GDP(PPP)$ (constant 2005 international dollars), are the only two enlargement countries with $B_i < 0$ before and after 2004. In the case of CZ, this is largely owing to its relatively high incoming success rate, $\mathcal{P}_{>,<}^{in}$. Countries with a notable decrease in their "labor import" selectivity, corresponding to a significant increase in $\mathcal{P}_i^{in}$, are GR, DE, and PL. Countries with a notable increase in their "labor export" competitiveness, corresponding to a significant increase in $\mathcal{P}_i^{out}$, are AT, CH, FR, IT, LT, LV, PT and SI; CY, BG and RO are two countries with a notable decrease in their "labor export" competitiveness. The difference in outgoing and incoming success rates, $\overline{\mathcal{P}}_{ij}^{out} - \overline{\mathcal{P}}_{ij}^{in}$, may be useful for identifying countries with mismatches in their competitiveness and selectivity, and indicative of labor market inefficiencies within Europe's "single market" (Boyle, 2013).

1.3.3 Measuring the effect of EU enlargement and subsequent Brain drain on EU science integration

This difference-in-difference (DiD) approach has also been used to study the impact of economic or political regime change (e.g. liberalization in the former, or democratization in the latter case) on a country's economic growth (Giavazzi and Tabellini, 2005; Persson and Tabellini, 2006). Here we implement a DiD panel regression to estimate the impact of brain drain and EU enlargement on the per-publication rate of cross-border activity, $f^s_{i,t}$, a proxy for European science integration.

The DiD interaction represents the cross-term between EU membership and the country's entry year, thereby measuring the impact of a change in EU membership status on a new member state's rate of international collaboration. In this way, the control group consists of the countries that did not change their EU membership status over 1997–2012 (members of country groups $g_{EU,i} = 1$ and 4), and the treated group are those that did change their EU membership status over the course of 1997–2012 (members of country groups $g_{EU,i} = 2$ and 3). We estimated the parameters of the following linear panel data model with country fixed-effects, which controls for scientific productivity and impact, R&D investment, high-skilled mobility, total migration in particular, and research subject area:

$$\begin{aligned} f^s_{i,t} &= \beta_t t + \beta_T T_{EU,i,t} + \{\beta_D \log_{10} D^s_{i,t} + \beta_R R^s_{i,t} \\ &+ \beta_E \log_{10} E_{i,t} + \beta_{GDPpc} \log_{10} GDPpc_{i,t} + \beta_{Spc} \log_{10} Spc_{i,t} + \\ &+ \beta_B B_{i,t} + \beta_I \log_{10} I^+_{i,t} + \beta_{P(in)} \mathcal{P}^{in}_{i,t} + \beta_O \log_{10} O^+_{i,t} + \beta_{P(out)} \mathcal{P}^{out}_{i,t} \\ &+ \beta_{G(in)} G^{in}_{i,t} + \beta_{G(out)} G^{out}_{i,t} + \beta_{\tilde{B}} \tilde{B}_{i,\tau} + \beta_{\tilde{O}} \log_{10} \tilde{O}_{i,\tau} \\ &+ \beta_{\tilde{I}} \log_{10} \tilde{I}_{i,\tau} + \beta_{\tilde{G}(in)} \tilde{G}^{in}_{i,\tau} + \beta_{\tilde{G}(out)} \tilde{G}^{out}_{i,\tau}\} \\ &+ \beta_s \cdot \mathbf{SA}(s) + \beta_{i,0} + \epsilon_{i,t} \\ &= \beta_t t + \beta_T T_{EU,i,t} + \{\beta \cdot \mathbf{x}_{s,i,t}\} + \beta_s \cdot \mathbf{SA}(s) + \beta_{i,0} + \epsilon_{i,t} \end{aligned} \tag{1.1}$$

The "EU Enlargement" treatment effect is estimated using the indicator value $T_{EU,i,t}$ capturing the EU-vs-non EU and before-vs-after cross-term: it is 1 for countries belonging to the EU in year t and 0 otherwise. Thus,

there are three groups of countries: (i) the incumbent EU countries with $T_{EU,i,t} = 1$ for all t, (ii) the group of new entrants with a transition from $T_{EU,i,t} = 0$ to $T_{EU,i,t} = 1$ in $t = 5$ for the ten 2004 entrants (CY, CZ, EE, HU, LT, LV, MT, PL, SK ,SI), and $t = 8$ for the two 2007 entrants (BG, RO), and (iii) the three Eurozone countries (CH, HR, and NO) that were not part of the EU as of the end of 2012 with $T_{EU,i,t} = 0$ for all t.

1.3.4 Demonstration of model robustness with partial models.

Table 1 shows the parameter estimates for partial models (A-E) that do not include one or more of the data types (Scientific productivity and impact, R&D investment, High-skilled mobility, Total migration). We also ran the same regression as the Full model, but this time restricting the data to the two periods before and the two periods after 2004 (4-period model F). This 4-period model better satisfies the difference-in-difference model condition that the number of countries (units) be much larger than the number of time periods analyzed, $31 = |i| \gg |t| = 4$. In all, the coefficients estimated across all model estimates shown in Table 1 are consistent in magnitude, sign, and significance, demonstrating the full model's robustness.

In addition to T_{EU} and $B_{i,t}$, there are several other parameters which are of particular interest. First, an increasing total incoming mobility I_t^+ is related to smaller f_t ($\beta_I < 0$, $p \leq 0.026$ in all regressions), consistent with the mobility mechanism whereby countries receiving foreign high-skill labor are at the same time losing the cross-border activity that was previously being channeled across the same foreign collaborator. This effect was also observed for the total mobility data ($\beta_{\tilde{I}} < 0$, $p \leq 0.005$ in all regressions). If, however, the foreign collaboration channel is maintained, then the cross-border activity is sustained. Thus, we observe that countries with higher outgoing mobility O_t^+ have higher f_t ($\beta_O > 0$, $p \leq 0.034$ in regressions A-E). Thus, an important caveat is whether or not the cross-border mobility results in the termination of cross-border activities, a causal effect that we are not able to estimate given the limitations of our data.

Second, the model indicates that more concentrated (nonuniform) distribution of outgoing mobility (larger G_t^{out}) is related to larger f_t ($\beta_{G(out)} > 0$, $p \leq 0.050$ in all regressions). This effect is consistent with the maintenance and investment in the cross-border activities among the core of more selective countries, principally the old EU members, which are characterized by a less-dispersed outward mobility (see Fig. 2).

Third, among the two scientific productivity and impact covariates we included, we observe a negative relation between the quantity of scientific output ($\beta_D < 0$, $p \leq 0.027$ in all regressions) implying a saturation effect in the capability to collaborate internationally. More importantly, we confirm the prestige effect represented by the citation impact R_t of each country ($\beta_R > 0$, $p \leq 0.000$ in all regressions) capturing the positive feedback between reputation and future collaborative opportunities at the aggregate level of countries which has also been observed for individuals (Petersen, 2015).

And finally, the subject area controls indicate that "Biochemistry, Genetics, and Molecular Biology" (1300) and "Physics and Astronomy" (3100) are the most collaborative domains, with "Agricultural and Biological, Sciences" (1100), "Chemistry" (1600), "Materials Science" (2500), and "Medicine" (2700) forming a middle group, and the rest of the subject areas comprising a third, relatively "low-collaboration" subset. The high-f_t group of biology and physics is largely due to the emergence of large team science stemming from globalizing endeavors (e.g. European Organization for Nuclear Research – CERN) and initiatives (e.g. the Human Genome Project, ENCODE) (Petersen et al., 2014).

Table 1: Parameter estimates for the panel data model for the collaboration rate $f^s_{i,t}$ (see Eq. 1.1), implemented with country fixed-effects and robust standard error estimates. Red and blue highlights indicate parameters significant at the $p \leq 0.05$ level. Beta coefficient are estimated using standardized variables for the non-categorical variables ($\log_{10} D^s_{i,t}$ thru $\tilde{G}^{out}_{i,\tau}$). For the full model (first column), $N_{obs.} = 4494$, Adj. $R^2 = 0.66$, and $N_c = 31$ countries. As a visual aid, we colored the coefficient estimates – red = significantly negative and blue = significantly positive – at the $p \leq 0.05$ level.

	Full model parameter estimates			Partial model parameter estimates (A-E)										4-period model (F)	
$f^s_{i,t}$ (fraction)	Eq. [4] Coeff.	Stand. var. (beta)	p-value	Model A	p-value	Model B	p-value	Model C	p-value	Model D	p-value	Model E	p-value	Model F	p-value
$Year, t$	0.015 ± 0.001	0.095 ± 0.008	0.000	0.019 ± 0.001	0.000	0.009 ± 0.001	0.000	0.011 ± 0.002	0.000	0.015 ± 0.001	0.000	0.018 ± 0.002	0.000	0.016 ± 0.005	0.006
T_{EU} (EU entry – treatment effect)	-0.058 ± 0.019	-0.376 ± 0.122	0.004	-0.043 ± 0.018	0.027	-0.070 ± 0.020	0.001	-0.076 ± 0.021	0.001	-0.055 ± 0.021	0.012	-0.053 ± 0.019	0.011	-0.044 ± 0.013	0.003
Scientific productivity and impact															
$\log_{10} D^s_{i,t}$ (publications)	-0.223 ± 0.037	-1.253 ± 0.208	0.000	-0.216 ± 0.041	0.000					-0.224 ± 0.039	0.000	-0.214 ± 0.037	0.000	-0.292 ± 0.041	0.000
$R^s_{i,t}$ (normalized citations)	0.164 ± 0.023	0.935 ± 0.132	0.000	0.159 ± 0.024	0.000					0.164 ± 0.024	0.000	0.159 ± 0.023	0.000	0.196 ± 0.035	0.000
R&D investment															
$\log_{10} E_{i,t}$ (Gvt. expenditure on R&D)	-0.080 ± 0.047	-0.467 ± 0.275	0.100			-0.096 ± 0.042	0.031			-0.065 ± 0.043	0.138			0.014 ± 0.058	0.811
$\log_{10} GDPpc_{i,t}$ (per capita GDP)	0.217 ± 0.058	0.505 ± 0.135	0.001			0.129 ± 0.060	0.040			0.151 ± 0.062	0.021			0.456 ± 0.110	0.000
$\log_{10} Spc_{i,t}$ (per capata researchers)	0.164 ± 0.063	0.292 ± 0.113	0.015			0.171 ± 0.062	0.010			0.182 ± 0.062	0.006			0.201 ± 0.060	0.002
High-skilled mobility															
$B_{i,t}$ (high-skilled brain-drain polarization)	-0.043 ± 0.013	-0.169 ± 0.049	0.002					-0.056 ± 0.012	0.000			-0.046 ± 0.011	0.000	-0.095 ± 0.024	0.000
$\log_{10} I^{hs}_{i,t}$ (total incoming mobility)	-0.024 ± 0.010	-0.187 ± 0.080	0.026					-0.028 ± 0.012	0.023			-0.023 ± 0.011	0.042	-0.056 ± 0.015	0.001
$\log_{10} O^{hs}_{i,t}$ (total outgoing mobility)	0.019 ± 0.008	0.130 ± 0.059	0.034					0.030 ± 0.009	0.003			0.027 ± 0.009	0.004	0.011 ± 0.011	0.307
$\mathcal{P}^{in}_{i,t}$ (incoming success rate)	-0.015 ± 0.040	-0.036 ± 0.101	0.722					-0.002 ± 0.043	0.967			-0.019 ± 0.042	0.651	-0.004 ± 0.049	0.940
$\mathcal{P}^{out}_{i,t}$ (outgoing success rate)	-0.110 ± 0.045	-0.206 ± 0.084	0.020					-0.068 ± 0.047	0.159			-0.067 ± 0.050	0.189	-0.201 ± 0.056	0.001
$G^{in}_{i,t}$ (incoming mobility Gini index)	0.011 ± 0.035	0.026 ± 0.079	0.746					0.003 ± 0.040	0.937			0.017 ± 0.038	0.660	0.009 ± 0.043	0.828
$G^{out}_{i,t}$ (outgoing mobility Gini index)	0.135 ± 0.044	0.237 ± 0.077	0.004					0.101 ± 0.048	0.044			0.103 ± 0.051	0.050	0.272 ± 0.056	0.000
Total migration															
$\tilde{B}_{i,\tau}$ (migration polarization)	-0.040 ± 0.020	-0.169 ± 0.084	0.052					-0.027 ± 0.018	0.145			-0.036 ± 0.020	0.087	-0.060 ± 0.033	0.078
$\log_{10} \tilde{I}_{i,\tau}$ (total incoming mobility)	-0.044 ± 0.011	-0.186 ± 0.048	0.000					-0.050 ± 0.011	0.000			-0.048 ± 0.011	0.000	-0.154 ± 0.051	0.005
$\log_{10} \tilde{O}_{i,\tau}$ (total outgoing mobility)	0.024 ± 0.013	0.172 ± 0.092	0.071					0.016 ± 0.011	0.165			0.019 ± 0.012	0.138	0.018 ± 0.014	0.199
$\tilde{G}^{in}_{i,\tau}$ (incoming migration Gini index)	0.212 ± 0.104	0.086 ± 0.042	0.051					0.242 ± 0.119	0.051			0.264 ± 0.102	0.015	0.204 ± 0.162	0.217
$\tilde{G}^{out}_{i,\tau}$ (outgoing migration Gini index)	-0.030 ± 0.044	-0.023 ± 0.034	0.502					-0.025 ± 0.039	0.528			-0.017 ± 0.043	0.687	-0.137 ± 0.052	0.014
Subject Area (s) (publication-level)															
"Agricultural and Biological, Sciences" (1100)	-0.031 ± 0.032	-0.202 ± 0.208	0.339	-0.028 ± 0.034	0.419	0.043 ± 0.009	0.000	0.062 ± 0.015	0.000	-0.032 ± 0.033	0.342	-0.027 ± 0.033	0.415	-0.067 ± 0.031	0.036
"Biochemistry, Genetics, and Molecular Biology" (1300)	0.044 ± 0.017	0.284 ± 0.112	0.016	0.046 ± 0.018	0.016	0.077 ± 0.013	0.000	0.097 ± 0.008	0.000	0.044 ± 0.018	0.019	0.047 ± 0.017	0.012	0.026 ± 0.017	0.127
"Business Management and Accounting" (1400)	-0.359 ± 0.056	-2.324 ± 0.361	0.000	-0.350 ± 0.060	0.000	-0.126 ± 0.010	0.000	-0.106 ± 0.017	0.000	-0.360 ± 0.058	0.000	-0.348 ± 0.056	0.000	-0.456 ± 0.062	0.000
"Chemical Engineering" (1500)	-0.177 ± 0.040	-1.145 ± 0.259	0.000	-0.171 ± 0.043	0.000	-0.020 ± 0.010	0.045	-0.001 ± 0.011	0.924	-0.178 ± 0.041	0.000	-0.169 ± 0.041	0.000	-0.236 ± 0.045	0.000
"Chemistry" (1600)	-0.028 ± 0.025	-0.184 ± 0.164	0.269	-0.025 ± 0.027	0.355	0.037 ± 0.012	0.003	0.056 ± 0.013	0.000	-0.029 ± 0.026	0.272	-0.025 ± 0.026	0.349	-0.058 ± 0.023	0.017
"Computer Science" (1700)	-0.135 ± 0.027	-0.874 ± 0.175	0.000	-0.132 ± 0.029	0.000	-0.075 ± 0.011	0.000	-0.056 ± 0.016	0.002	-0.135 ± 0.028	0.000	-0.131 ± 0.028	0.000	-0.156 ± 0.029	0.000
"Decision Sciences" (1800)	-0.315 ± 0.062	-2.043 ± 0.404	0.000	-0.305 ± 0.068	0.000	-0.023 ± 0.014	0.125	-0.003 ± 0.017	0.848	-0.317 ± 0.065	0.000	-0.303 ± 0.063	0.000	-0.393 ± 0.066	0.000
"Energy" (2100)	-0.248 ± 0.050	-1.604 ± 0.325	0.000	-0.240 ± 0.054	0.000	-0.034 ± 0.016	0.039	-0.015 ± 0.019	0.450	-0.249 ± 0.052	0.000	-0.238 ± 0.051	0.000	-0.321 ± 0.060	0.000
"Engineering" (2200)	-0.068 ± 0.020	-0.439 ± 0.129	0.002	-0.066 ± 0.021	0.003	-0.060 ± 0.009	0.000	-0.040 ± 0.015	0.010	-0.068 ± 0.020	0.002	-0.066 ± 0.020	0.003	-0.060 ± 0.020	0.006
"Environmental Science" (2300)	-0.118 ± 0.038	-0.762 ± 0.243	0.004	-0.113 ± 0.040	0.008	(omitted)		0.019 ± 0.014	0.181	-0.118 ± 0.039	0.005	-0.112 ± 0.038	0.006	-0.164 ± 0.038	0.000
"Materials Science" (2500)	0.003 ± 0.022	0.017 ± 0.145	0.906	0.006 ± 0.024	0.818	0.056 ± 0.011	0.000	0.075 ± 0.014	0.000	0.002 ± 0.023	0.922	0.006 ± 0.023	0.791	-0.004 ± 0.023	0.871
"Medicine" (2700)	(omitted)	baseline Subj. Area		(omitted)		-0.035 ± 0.021	0.097	-0.016 ± 0.011	0.158	(omitted)		(omitted)		(omitted)	
"Pharmacology,Toxicology, and Pharmaceutics" (3000)	-0.191 ± 0.038	-1.240 ± 0.246	0.000	-0.185 ± 0.041	0.000	-0.019 ± 0.014	0.180	(omitted)		-0.192 ± 0.040	0.000	-0.183 ± 0.038	0.000	-0.254 ± 0.040	0.000
"Physics and Astronomy" (3100)	0.151 ± 0.019	0.976 ± 0.126	0.000	0.152 ± 0.020	0.000	0.164 ± 0.013	0.000	0.183 ± 0.017	0.000	0.150 ± 0.020	0.000	0.153 ± 0.020	0.000	0.147 ± 0.019	0.000
constant	-29.1 ± 2.5	-190 ± 17	0.000	-36.2 ± 2.6	0.000	-16.9 ± 2.6	0.000	-20.8 ± 3.3	0.000	-29.1 ± 2.6	0.000	-34.5 ± 3.5	0.000	-32.2 ± 10.5	0.004
Adjusted R^2	0.66	0.66		0.65		0.61		0.61		0.65		0.66		0.67	
Number of observations	4494	4494		4494		4494		4494		4494		4494		1680	
Number of countries	31	31		31		31		31		31		31		31	

1.4 Discussion

The rate of international collaboration has been increasing as a result of globalization, with a large contributor to this trend being the countries with smaller science programs which integrate with large R&D hubs (Chessa et al., 2013; Morescalchi et al., 2015; Petersen et al., 2014). As such, over the 1997-2012 period of analysis, we also observe an increase in the per-publication cross-border collaboration rate $f_{i,t}$, especially during the early 2000s. For the incumbent EU countries, the mean (averaged over countries and 14 subject areas) cross-border collaboration rate before and after 2004 were $\langle f_{<} \rangle = 0.41$ and $\langle f_{>} \rangle = 0.53$ (significantly different mean values, with difference-in-means Student T-test p-value = 10^{-14}); for the 2004 non-EU countries, the mean cross-border collaboration rates before and after 2004 were $\langle f_{<} \rangle = 0.42$ and $\langle f_{>} \rangle = 0.46$ (significantly different mean values, with difference-in-means Student T-test p-value = 0.001). The increase in $f_{i,t}^{All}$ is stronger for the incumbent EU countries (Fig. 1). Interestingly, the notable increase between 2002 and 2003 may be attributable to EU Framework Programme (FP6) funding initiatives introducing explicit cross-border collaboration requirements.

Meanwhile, the rate of high-skilled mobility between EU members also increased over the same period (see Ackers and Gill, 2008; Kahanec, 2013, for an in-depth review of the impact of EU enlargement on labor mobility). However, the incoming and outgoing rates for each country are typically not equal, representing a large-scale reorganization of high-skilled labor in Europe. While the in-to-out mobility ratio $I_{i,t}/O_{i,t}$ for the incumbent EU countries was distributed more evenly above and below unity, $I_{i,t}/O_{i,t}$ is mostly less than unity for the 2004 non-EU countries, representing the major brain-drain imbalance between eastern and western Europe.

In order to quantify the impact of brain drain on $f_{i,t}$, we implemented a panel data model with country fixed-effects, including controls for scientific productivity and impact, R&D investment, high-skilled mobility, total migration (high-skilled + low skilled), and controls for publication subject area. The results of our main model parameter estimates

are shown in the first column of Table 1; The remaining columns show partial model parameter estimates demonstrating the robustness of our model and its results.

Most importantly, the difference-in-difference term T_{EU}, which measures the relative change in $f_{i,t}$ – between the entrants and incumbent groups, before versus after 2004 – shows that the entrant countries suffered a -0.058 decrease in $f_{i,t}$ due to the EU enlargement ($p = 0.004$). The divergence in $f_{i,t}$ between the entrant and incumbent EU countries was further exacerbated by the net polarization in high-skilled brain drain ($B_{i,t}$), which we find to be roughly half as strong as the T_{EU} effect. The net difference in $f_{i,t}$ explained by the combination of the EU enlargement effect and the brain-drain effect is $-0.058 + (-0.043) \times (\langle B_{>}^{\text{new EU}} \rangle - \langle B_{>}^{\text{old EU}} \rangle) = -0.078$. The actual difference-in-difference in the mean collaboration rates before and after is $(\langle f_{>}^{\text{new EU}} \rangle - \langle f_{>}^{\text{old EU}} \rangle) - (\langle f_{<}^{\text{new EU}} \rangle - \langle f_{<}^{\text{old EU}} \rangle) = -0.085$ (calculated from the mean f_t curves in Fig. 2(A,B) for the 14 s). Thus, we estimate that T_{EU} and $B_{i,t}$ explain roughly 92% of the European east-west divergence in f_t over the period of analysis. Other factors certainly contribute to the divergence between eastern and western Europe cross-border collaboration rates, such as inequality in R&D funding within the EU framework programmes, institutions, and the location of central scientific facilities (Abbott and Schiermeier, 2014; Hoekman et al., 2013).

1.5 Conclusion

In summary, our analysis reveals the counterintuitive *decrease in cross-border activity* by the new member states following their entry into the EU. Despite gaining access to EU resources incentivizing cross-border integration, we find that both the number and rate of cross-border collaboration would have increased *had they not joined the EU*. It is important for EU policy makers to consider the possible unintended consequences of EU labor market integration, especially considering the EU goals for a unified industrial and academic R&D innovation system (Boyle, 2013; European Commission and Directorate-General for Research and Inno-

vation, 2013; Nedeva and Stampfer, 2012). When a researcher not only moves abroad, but also brings their international links along with them, this represents a loss of social capital – in addition to human capital and tacit knowledge (Agrawal et al., 2011, 2006) – that may further reduce the potential for knowledge spillovers across countries. As such, this net flow of high-skilled labor to the large GDPpc countries (GB, CH, NO) may negatively impact the convergence of human and technological capital within Europe (Grossmann and Stadelmann, 2011), especially when considering the long-term impacts of losing elite scientists (Weinberg, 2011).

The EU should, however, be commended for implementing "twinning" and "teaming" policies within the Horizon2020 framework to counter the divergence in scientific competitiveness and specialization between regions (*European Commission Horizon 2020: Spreading excellence and widening Participation*, n.d.) – hopefully it's not "too little too late". Notwithstanding the negative effects of brain drain, it is important to mention some of the positive impacts of brain drain that have been proposed in the literature, such as increased educational incentives within the emigrant country and positive network externalities on trade and technological adoption (Beine et al., 2001; Docquier and Rapoport, 2012; Gibson and McKenzie, 2011). Specific to European science, we emphasize that opportunities for talented researchers to study abroad are extremely important, both from an incentive and a social-capital perspective – afterall, this is also a key component of an "open" and "competitive" globalized science system. However, there must be a more concerted effort to contain brain drain and to foster the right conditions for home-return knowledge transfer (Wang, 2015), so that a 'brain regain' follows organically from global 'brain circulation' (Ackers, 2005; Dustmann et al., 2011; Wiesel, 2014). The starting point is within the existing framework of mobility fellowships (e.g. Marie Curie and other cross-border fellowships), possibly via implementing stronger incentives and criteria for home-country return, and encouraging less-attractive countries to implement local strategies for maintaining ties with their high-skilled expatriates (Lepori et al., 2015).

Chapter 2

Inequality and International Trade. Evidence from Colombia.

2.1 Introduction

In recent years Social Inequality has grown as a field of academic enquiry. It has gained, at least, the same perceived level of importance as other macroeconomics topics, such as Economic Growth. Most macroeconomists have focused their research on studying the mechanisms that affect growth. The current attention on inequality has made it possible to reformulate the growth question: "how to increase growth equally, or how to redirect growth to improve equality levels?". This question has relevance not only for development economics, but also for highly developed economies. To answer both questions, we started searching for the mechanisms through which inequality increases. Moreover, seeking these mechanisms is one of the most important challenges that researchers and policy makers have faced in recent years. Specifically, we face the problem of inequality by studying the local effects of international trade. From an economic view, international trade is the mechanism through which markets expand and through which production

is optimized. It is implemented, typically, by reducing taxes on cross-border trade, and by generating an open market across economies. The relevance of international trade in terms of inequality consists in that during the optimization process the market structure changes, having an impact on labor mobility and firms' dynamics, among others economic factors. Likewise, since economic internationalization leads to the reallocation of wealth within the exporting economy, it also impacts levels of poverty and inequality, providing a link with socioeconomic conditions (Nissanke et al., 2008; Ravallion, 2001, 2006). Social problems caused by economic internationalization have been studied extensively. For instance, recent gender studies have shown that increasing labor mobility, increases in gender inequality (De Hoyos et al., 2012). Also, further evidence (Buera and Shin, 2013) links trade liberalization and financial frictions on the emergence of the "miracle economies". Models of competitive firm growth indicate that better positioned firms, in terms of capital, services and geographical location, are those that are more capable of exploiting international commerce. However, the literature is inconsistent in determining whether exploitative firms, those with higher survival probability, tend to alter the levels of inequality. Within this active topic of research, there are two controversial positions: the first supposes that economic internationalization policies increase growth, bringing benefits to society; the second position supposes that open trade increases inequality eroding social standards (Lundberg and Squire, 2003). Aligned with the second position, Helpman et al. (2010) develops a theoretical model to explain inequality which assumes an open trade economy, defining inequality through the differences between the highest and the lowest wages, which intrinsically assumes inequality over the poverty line, by assuming labor formality. In favor with the last position, we use inequality as shortfalls on social standards, which we measure within individuals below the poverty line. Therefore, our empirical work extends Helpman et al. (2010), using a different Poverty Gap framework studying the effect of economic internationalization on the population below the poverty line.

Here, we study the Colombian case. To justify our case study, we

have taken into account that Colombia has been referred to as one of the most unequal societies in the world by Piketty (2014). Specifically on Colombia, Ramirez et al. (2016) found that tax centralization and multidimensional poverty specific to geographic regions contribute to a lack of efficiency. In contrast, Eslava et al. (2013) measures the manufacturing plant efficiency in Colombia, finding that "Trade liberalization also increases the productivity of incumbent plants and improves the allocation of activity". Seminal papers in this direction by Dijkstra (2000) and Wood (1997) measured the inefficiency of the manufacturing, finding remarkably high levels of plant inefficiency, even in the most industrialized Latin American countries, including Colombia. These contradictory results open an interesting niche for research.

Instead of using the view of productivity to reckon the effect of international trade, we use inequality. To measure it, we use the Multidimensional Poverty Index (MPI) at municipality level (Elhorst, 2010; LeSage and Pace, 2008), for 1086 counties, including all main cities, taken from the Colombian 2005 Census provided by the Colombian National Statistical Agency (DANE). Within this context, it is important to note that inequality in Colombia, a good example of a developing economy, continued to steadily (Fig. 20) grow throughout the 1990s due to the open trade policy implemented in the same decade (Attanasio et al., 2004; Goldberg and Pavcnik, 2004). We use export data for Colombia over the period 2002-2004 and import data from its principal importers in order to provide new quantitative evidence for the impact of economic internationalization on inequality. To study the relation between inequality and international trade at a municipality level, we assume that the effect of geographical properties cannot be overlooked in the analysis of social systems. This is especially true in regional economics where regions are highly influenced by their neighbors (Anselin, 1988). To go deeper in our case study, observing evidence of high degrees of homogeneity across Colombia's municipalities, we use spatial econometric estimations to control for spatial correlations. In our effort to disentangle geographical effects, political factors, and idiosyncratic efficiency, we also use an instrumental variable to account for possible sources of en-

dogeneity (Drukker et al., 2013).

Our research contributes to the literature on dynamic industrialization, the open economy in Latin America, inequality sources and policy making. The main finding of this work is that international trade is a mechanism through which inequality increases within municipalities. Inequality has increased at a higher percentage in our comparison of municipalities with and without exporting firms. Our results differ from previous similar works, such as the theoretical model by Helpman et al. (2010) and the empirical work of Goldberg and Pavcnik (2005), because we approximate the Poverty Gap by the the average shortfall of deprivations, instead of accounting for the difference in wages, which intrinsically assumes labor formality. Moreover, using this approach, we find empirical evidence that inequality increases through an increase of exports, where with our inequality definition, we are able to capture the increase in inequality as the average of deprivations of the poor. Therefore, we account for a decrease in the social standard for people below the poverty line. We use this definition for the purpose of measuring whether international trade improves or worsens the life quality for the poor.

We define the present chapter of the thesis as follows: section 2.2 explains in details the data used in this chapter, including the Multidimensional Poverty Index, the international trade data and different variables to control for idiosyncratic effects on municipalities, together with the explanation of the spatial econometrics estimators. Section 2.3 explains the model specification, as well as assumptions and limitations of the model. In section 2.4, we show the results of the estimations and in section 2.5 we conclude the chapter.

2.2 Data

2.2.1 Poverty and Inequality

The poverty measure can be divided into two simpler measures: identification and aggregation. Identification explains who the poor are. It is

also called the *Headcount Ratio* or *Incidence* and counts how many people are below the poverty line. Aggregation joins in one single poverty measure all of the population under the poverty line, and it is is also known as *Intensity* (Sen, 1976). Additionally, the method suggested by Foster, Greer and Thorbecke (*FGT Measure*) introduces a multiplier effect for the poor, transforming the poverty measure using a non negative exponent α. When this coefficient takes the value of 0, the measure remains the Headcount Ratio, for $\alpha = 1$ the measure turns into the Poverty Gap, and when $\alpha = 2$ we have the *FTG index*, which is defined as the average of the square normalized shortfalls within the population increasing the effect of the poorest. This is also known as *Severity*.

We use the Multidimensional Poverty Index (MPI) developed by Alkine and Foster (Alkire and Foster, 2011a). The MPI is a poverty measure composed by different dimensions in order to monitor how public policies affect individuals. It has two versions: global and regional. The global version compares poverty across countries with an uniform methodology. It is composed by three principal dimensions: Health, Education and Living Standards. The regional version of MPI compares poverty across regions, with a difference that it is focusing on specific indicators of development for the region. Therefore, it cannot be used to make comparisons across countries. In the context of the MPI, deprivations have a lack of accessibility, under a certain determined level, to the dimensions defined in the MPI. In this work, we use the regional version of the MPI and municipalities as its regional units. We justify the municipality choice as a regional unit because it represents an aggregation level small enough to be sensitive to local polices, while they are large enough to capture changes in the distribution of inequality.

Specifically, we use the Colombian Multidimensional Public Index (CMPI), which is a regional Multidimensional Public Index for Colombia aggregated by municipalities based in the 2005 Colombian Census [1], it is defined by five dimensions: Education, Childhood and Youth conditions, Employment, Health and Access to public utilities and housing

[1]Source: Departamento Nacional de Estadística (DANE, National Statistical Department), it is based in a survey to 1.3 million households.

conditions, using 15 indicators aggregated in five dimensions, Ramirez et al. (2016)[2], see Table 2 and Table 3 for details. The CMPI uses household as analysis unit. This implies that deprivations are experienced simultaneously among every member of the household instead on individuals. For instance, if one child within the household in the age of attending school (6 to 16 years old) do not attend school, it is assumed that the whole household is deprived. This implies that all individuals living in this household are consider deprived with respect to the attending school indicator. We denote y_{ij} as the achievement of a household i of indicator j.

For category of each indicator a threshold z_i is introduced such that the condition $y_{ij} < z_j$, denotes deprivation with respect to the indicator j. This threshold z_j is the firs cutoff condition and it is applied to indicators. There is another cutoff condition k, which determines how many indicators a household must have to be defined as poor. For instance, if $k = 2$, it means that a household deprived in two or more dimension is consider poor. At this point we introduce the deprivation matrix $g_{ij} = 1$ for $y_{ij} < z_j$ and 0 otherwise, where rows are individuals i and columns are indicators j. After the second cutoff condition k, we get the censored deprivation matrix $g_{ij}(k)$, technically, it is the deprivation matrix after changing the all the indicators of non deprived households into non deprived ($g_{ij} = 0$). For instance,

$$g^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{matrix} 1 \\ 2 \\ 4 \\ 0 \end{matrix} \Longrightarrow g_{ij}^{(0)}(k) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{matrix} 0 \\ 2 \\ 4 \\ 0 \end{matrix}$$

For the case of the CMPI the second cutoff condition is $k = 5$. To be able to understand the estimation of the deprivation matrix of the CMPI with each threshold z_j, we explain each indicator literally taken from the technical report (see Angulo et al., 2013, p.14-21), where quotes are neglected due to the long extension of the direct reference:

[2]I am grateful to Juan Ramirez from Fedesarrollo, Colombia for sharing the CMPI data.

Dimension of household education condition

Educational achievement. This indicator is measured by the average level of education for individuals 15 years old and over within the household. However, it is worth noting that if a household member selects preschool as the highest level of education approved, zero years of schooling is assigned to such a member. In terms of the cutoff point used by this indicator, a household is considered deprived when the average years of schooling of its members aged 15 and over is below nine years of schooling. But, when there are no household members aged 15 years old and over within the household, the household is automatically considered as deprived in terms of educational achievement.

Literacy. This indicator is defined as the percentage of people aged 15 or above in the household that know how to read and write. A household is considered deprived if at least one of the household members aged 15 or older does not know how to read or write (i.e. less than 100% of its members 15 years old and over are able to read and write). When there are no household members 15 years old or over, the household is considered deprived.

Dimension of childhood and youth conditions

School attendance. This indicator is calculated as the proportion of school-age children (6 to 16 years old) in a household who attend an educational institution. According to this indicator, a household is considered deprived if at least one of the children between 6 and 16 years old do not attend school (i.e. less than 100% of children 6 to 16 years old are attending school). Households with no children between 6 and 16 years old are not considered deprived in this indicator.

No school lag. School lag is calculated for the households with children between the ages of 7 and 17. The school lag of each child is defined as the difference between the number of legally expected years of schooling by age and the number of school years completed in fact. The legally

expected years of schooling by age are defined by the Sector Plan for Education 20062010 presented by the National Ministry of Education. To the age of 7 is expected to have 1 year of school completed, successively until to the age of 17 is expected to have 11 years of school completed. A household is considered as deprived in this variable if any of the children between 7 and 17 years are lagging in school. In other words, the desired result is 100% of children in a household without school lag. Households with no children between 7 and 17 years old are not deprived in this indicator.

Access to childcare services. This indicator provides the percentage of children 0 to 5 years old in each household who have access to childcare services (health, proper nutrition, and adult supervision or education) simultaneously. A household is considered to be deprived in access to childcare services if there is at least one child between 0 and 5 years old with no simultaneous access to all childcare services. Thus, a household is not deprived if its children under the age of 5: i) spend most of the week at a community home, nursery or preschool, or are under the care of a responsible adult; ii) are covered by health insurance; and iii) receive lunch in the care facility where they spend most of time (the latter in the case of children going to a community home, nursery or preschool).

Children not working. According to the International Labour Organization (ILO) [3] and the Colombian National statistical Department (DANE), child labour refers to children under 18 years old that carry out household chores for more than 15 hours per week, children under 14 years old classified as employed, and children under 18 years old involved in hazardous work. In the case of the CMPI and given the data constraints of the LSMS, the CMPI only includes the percentage of children in the household between 12 and 17 who are employed. The indicator of children not working is defined as the percentage of children who are out of the labor market. A household is deprived in this variable if at least

[3] See ILO convention No 138 on the minimum age for admission to employments and work and ILO convention No 182 on the worst forms of child labour, 1999.

one child between 12 and 17 years old is employed. A household with no children between 12 and 17 years old is considered not deprived.

Dimension of employment.

Absence of long-term unemployment. This indicator measures the percentage of the Economically Active Population (EAP) in the household that has been unemployed for more than 12 months. The indicator is calculated as $(1 - (\text{ long term unemployment}/EAP))$. A household where there is at least one person in long-term unemployment is considered to be in deprivation. Households with no economically active population are considered deprived in this variable, with the exception of households made up of people living on a pension.

Formal employment. This indicator takes the proportion of the economically active population within the household that is employed and actively affiliated to a pension fund (affiliation to a pension fund is taken as a proxy of formality). A household is considered deprived when less than 100% of the EAP has formal employment (employees affiliated to a pension fund / EAP). This indicator also captures unemployment. For this reason, the long-term unemployed are removed from the denominator in order to avoid counting them in deprivation twice. Children under the age of 18 who hold a job are also eliminated in order to be congruent with the non-child employment policy. Households with no EAP are considered deprived.

Dimension of health

Health insurance coverage. Health insurance coverage is defined as the proportion of household members covered by the Social Security Health System. A household is deprived if any of its members is not affiliated with a health insurance regime. Given that the access-to-childcare-services variable takes into account the health insurance status of children between 0 and 5 years old, this indicator is measured only for the population older than five.

Access to health services in case of need. This indicator measures the proportion of people in a household who have access to health services in case of need. A household is not deprived in access to healthcare services if all of its members who in the last 30 days have suffered an illness, an accident, dental problems or any other health issues that have not required hospitalization, have been attended by a doctor, specialist, dentist, therapist or health institution. Households where no one has had a need for healthcare services are not considered to be deprived.

Dimension of access to public utilities and living conditions

Access to improved drinking water. This indicator was defined using WHO-UNICEF guidelines, where urban households are considered deprived when they have no access to public water services. In rural areas, households are considered deprived when they have no access to public water services and the water used to prepare food is obtained from a well, rainwater, a river, spring water source, public tap or standpipe, water truck, water carrier or any other source other than piped water.

Adequate elimination of sewer waste. In this case urban households without access to a public sewer system are considered deprived. Rural households are considered deprived if they have a toilet without a sewer connection, a latrine or if they simply do not have a toilet.

Adequate floors. Households with dirt floors are considered deprived.

Adequate exterior walls. An urban household is considered deprived when the exterior walls are built of untreated wood, boards, planks, guadua (a type of bamboo) or other vegetation, zinc, cloth, cardboard, waste material or when no exterior walls exist. A rural household is considered deprived when exterior walls are built of guadua or other vegetation, zinc, cloth, cardboard, waste materials or if no exterior walls exist.

No critical overcrowding. An urban household is considered critically overcrowded, and therefore deprived, when the number of people sleeping per room (excluding kitchen, bathroom and garage) is greater than or equal to three; a rural household is considered deprived when the number is more than three people per room.

Here ends up the quotation about of the technical report (see Angulo et al., 2013, p.14-21).

Table 2 shows the definition of each indicator specifying their weights into the CMPI. Table 3 shows how each indicator is calculated giving their cutoff point.

Table 2: Dimensions, variables, weights and poverty lines of the implemented CMPI

Dimension	Variable	Indicator	Cutoff
Household education conditions (0.2)	Educational achievement (0.1)	Percentage of people living 15 and older who holds at least 9 years of education.	9 years old
	Illiteracy (0.1)	Percentage of people living in a household 15 and older who know how to read and write.	100%
Childhood and youth conditions (0.2)	School attendance (0.05)	Percentage of children between the ages of 6 and 16 in the household that attend schooling	100%
	No school lag (0.05)	Percentage of children and youths (717 years old) within the household that are not suffering from school lag (according to the national norm).	100%
	Access to childcare services (0.05)	Percentage of children between the ages of 0 and 5 in the household who simultaneously have access to health, nutrition and education.	100%
	Children not working (0.05)	Percentage of children between 12 and 17 years old in the household that are not working.	100%
Employment (0.2)	No one in long-term unemployment (0.1)	Percentage of household members from the economic active population that are not facing longterm unemployment (more than 12 months).	100%
	Formal employment (0.1)	Percentage of employed household members that are affiliated to a pension fund (formality proxy).	100%
Health (0.2)	Health insurance (0.1)	Percentage of household members over the age of 5 that are insured by the Social Security Health System.	100%
	Access to health services (0.1)	Percentage of household members that had access to a health institution in case of need.	100%
Access to public utilities and housing conditions (0.2)	Access to dwelling services (0.16)	Percentage of dwelling services that the household has access to; this out of (i) water source, (ii) adequate elimination of sewer waste, (iii) adequate floors (iv) adequate external walls.	1
	No critical overcrowding (0.04)	Number of people sleeping per room, excluding the kitchen, bathroom and garage.	Urban: 3 or more people per room, Rural: more than 3 people per room.

Source: Angulo et al. (2013), National Planning Department (NPD), Social Development Unit (SDU), Social Promotion and Quality of life Division (SPQLD). 2011. **Notes:** The weight assigned to each dimension and variable is shown in parenthesis.

Table 3: Estimation of the indicator at household level, where $g_{ij} = (1 - y_{ij}/z_j)$ for $y_{ij} < z_j$ and $g_{ij} = 0$ otherwise.

Indicator	Cutoff	Poverty Gap Calculation (Every indicator is multiplied by 100)
Education (9+ years of schooling)	Household ave. 9 years †	$1 - \frac{\text{People 15 y. old and over with 9 or more schooling y.}}{\text{People 15 years old and over in the household}}$
Literacy	100%	$1 - \frac{\text{People 15 y. old and over that know how to read ad write.}}{\text{People 15 years old and over in the household}}$
School attendance	100%	$1 - \frac{\text{Children between 6-16 attending school.}}{\text{Children between 6-16 years old}}$
No school lag	100%	$1 - \frac{\text{Children between 7-17 with no school lag.}}{\text{Children between 7-17 years old}}$
Access to childcare services	100%	$1 - \frac{\text{Children between 0-5 in the household with simultaneous access to health, nutrition and education.}}{\text{Children between 0-5 years old}}$
Children not working	100%	$1 - \frac{\text{Children between 12-17 that are not working.}}{\text{Children between 12-17 years old}}$
No one in long-term unemployment	100%	$\frac{\text{Long term unemployed.}}{\text{Economical Active Population}}$
Formal employment	100%	$1 - \frac{\text{Employed and affiliated to a pension fund.}}{\text{Economical Active Population}}$
Health insurance	100%	$1 - \frac{\text{People over years old with insurance.}}{\text{People over years old}}$
Access to health services	100%	$1 - \frac{\text{People with access to a health institution.}}{\text{People in need of health care}}$
Critical overcrowding	Urban: 3 or more people for room	$1 - \left(\frac{\text{(Number of rooms * 3) - 1}}{\text{Total household members}}\right)$
	Rural: more than 3 people for room	$1 - \left(\frac{\text{Number of rooms * 3}}{\text{Total household members}}\right)$

Source: Angulo et al. (2013), National Planning Department (NPD), Social Development Unit (SDU), Social Promotion and Quality of life Division (SPQLD). 2011. **Notes:** † The cutoff point for the calculation of the Incidence is a household average of 9 years of education, while the poverty gap is calculated as the percentage of adults who have fewer than 9 years of education. This means that some households which are not classified as deprived on this indicator for the purposes of the incidence, will have one or more adult members with fewer than 9 years of schooling, and thus would be indicated as having a poverty gap on this indicator. However, the gap for these households is not included in the calculations of M1 and M2, because gaps are defined only for households deprived on each dimension.

A formal definition of the deprived household follows by considering the household-dimension matrix g_{ij}. The net poverty state of the household i is column sum of the matrix $c_i = |g_{ij}|$. The total number of deprived households in dimension j is the row sum of the matrix $|g_{ij}|$. The higher-order FTG measures ($\alpha = 0, 1, 2$) are defined by the unidimensional method, such that:

$$g_{ij}^1 = g_{ij}^0 \left(\frac{z_j - y_{ij}}{z_j} \right), \tag{2.1}$$

where z_j is the minimum level for deprivation in dimension j; and y_{ij} is the individual achievement of i in dimension j. More generally

$$g_{ij}^\alpha = (g_{ij}^1)^\alpha, \tag{2.2}$$

with $\alpha \geq 0$, (see Alkire and Foster, 2011b; Foster et al., 1984, for details). Following the example of the censored deprivation matrix $g_{ij}^{(0)}(k)$ defined before, we have:

$$g_{ij}^{(0)}(k) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \begin{matrix} 0 \\ 1 \\ 1, \\ 0 \end{matrix}$$

$$g_{ij}^{(1)}(k) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0.4 & 0 \\ 0.3 & 0.6 & 0.1 & 0.2 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad g_{ij}^{(2)}(k) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.2^2 & 0 & 0.4^2 & 0 \\ 0.3^2 & 0.6^2 & 0.1^2 & 0.2^2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

According to Foster et al. (1984), the Adjusted Headcount Ratio is defined as the average of the censored deprivation matrix $M_0 = \langle g_{ij}^{(0)}(k) \rangle$, $\langle \cdot \rangle$ for average, which in the dummy example $M_0 = \langle g_{ij}^{(0)}(k) \rangle = 6/16 = 0.375$. In general for the FTG Measure are defined as $M_\alpha = \langle g_{ij}^{(\alpha)}(k) \rangle$, which in our dummy example $M_1 = (0.2+0.4+0.3+0.6+0.1+0.2)/16 = 0.11$, and $M_2 = (0.2^2+0.4^2+0.3^2+0.6^2+0.1^2+0.2^2)/16 = 0.04$. From the censored deprivation matrix, we can also estimate the Headcount Ratio (H), it is defined as the fraction of rows of matrix $g_{ij}(k)$ with at least one non zero element, it gives the percentage of the population that is poor,

in our example $H = 2/4 = 0.5$. The Average Intensity (A) is defined as the average value of deprived dimension per poor, it can be expressed as $A = M_0/H$, (Alkire and Foster, 2011a,b).

The DNP reports in the CMPI the Headcount Adjusted Ratio (M_0), the Adjusted Poverty Gap (M_1), the Adjusted Severity and the Headcount Ratio (H).

Also, using the definition in Foster et al. (1984), we have that M_1 is the *Adjusted Poverty Gap*, it is also interpreted as the result if you multiply the Adjusted Headcount Ratio (M_0) times the "*Average Poverty Gap*" (G). It represents the sum of the normalized gaps of the poor, it is used to estimate the average poverty gap by $G = M_1/M_0 = \langle(1 - y_{ij}/z_j)\rangle$. M_1 is the average normalized shortfall of deprived dimensions from minimum levels z_i until reaching the household achievements y_{ij}. It measures how far is an average household will move out of the poverty line with values between 0 and 1, with 1 as the most unequal state. M_1 increases more than (M_0) when a household decreases its personal achievement: $y_{ij}^{t+1} < y_{ij}^{t} \iff {M_1}^{t+1} \ll {M_1}^{t}$, giving a more sensitive measure for inequality than the former. Furthermore, the Adjusted Poverty Gap is a measure of social convergence, e.g., if a society decreases its *inequality* means that it also decreases its M_1 value.

Again, using the definition in Foster et al. (1984) for $\alpha = 2$, we have that M_2 is the *Adjusted FGT index or Severity*. It is a square shortfall between personal achievement and minimum standard levels. The Adjusted FGT index represents the average of severity measure of a deprived households. This version of the MPI is useful in comparing the spread of extreme poverty over different regions, having a bigger value for a bigger Severity, for instance, having two different regions A and B with similar Adjusted Poverty Gap $(M_1^A \approx M_1^B)$, but with an FTG index significantly different $(M_2^A \gg M_2^B)$ means that region A has a stronger move toward extreme poverty than region B. M_2 is used to determine the Severity by $S = M_2/M_0 = \langle(1 - y_{ij}/z_j)^2\rangle$.

For doing public policy is interesting to aggregate in one measure the number people and the quality of the poor. Although, for our case, we are interesting in avoiding the size effect of the inequality measure.

Therefore, we use the average poverty gap G over of the adjusted poverty gap M_i, because we are interested in intensive measures, eliminating the size effect in the extensive measure M_i. G does not account for the number of people within municipalities, but only for the average deprivations in that specific area.

For a wider definition see Foster et al. (1984), and Alkire and Foster (2011b).

In order to explore the CMPI data, we present Figure 5, Figure 6 and Table 4. Fig. 5 shows the relation between *Incidence* (H) vs. *Intensity* (A) for municipalities and Fig.5 shows the same relation for Departments (states). By visual inspection, Fig. 5 shows a higher non-linear relation between Incidence and Intensity compared with Fig. 6. This difference shows that, at municipality level, we are able to study non-linear effects, which are not obvious at higher aggregation levels. Table 4 shows a summary of the statistics of the variables included in the CMPI.

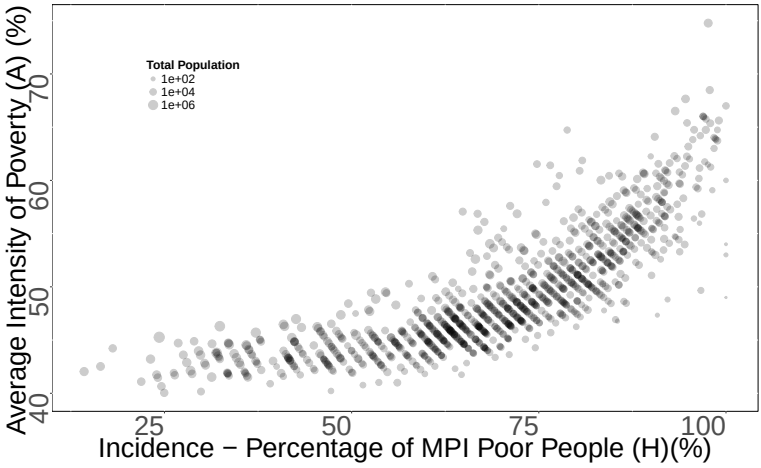


Figure 5: Colombian Multidimensional Poverty Index by Municipality (counties).

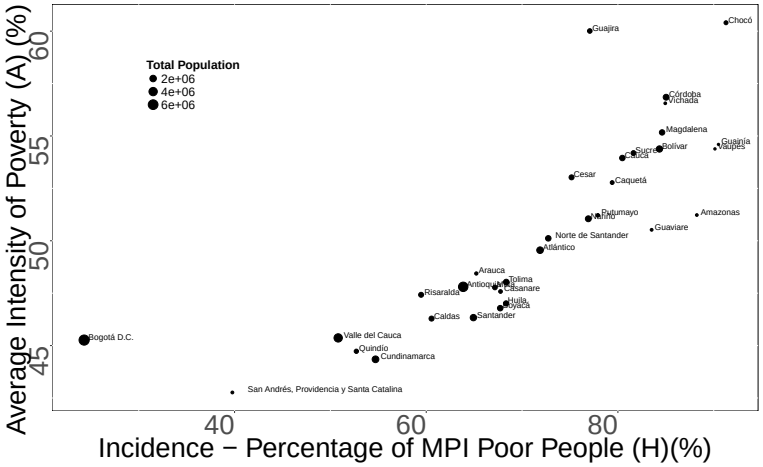


Figure 6: Colombian Multidimensional Poverty Index by Departments (states).

Table 4: CMPI statistics from 2005, summarizing the magnitude of poverty in Colombia.

CMPI	Mean	(Std. Dev.)	Min.	Max.
Headcount (individuals)	18849.56	(65685.75)	290	1638155.5
Log of Headcount (individuals)	9.078	(1.047)	5.670	14.310
Poverty (%) Total	68.09	(15.86)	14.27	100
Poverty (%) Urban	52.04	(17.48)	13.23	100
Poverty (%) Rural	79.05	(13.42)	22.99	100
M_0	0.34	(0.11)	0.06	0.73
M_1	0.24	(0.09)	0.04	0.67
M_2	0.21	(0.08)	0.04	0.65
N		1086		

Table 5 shows a correlation analysis of the variables. Here, we evidence that the intensive variables A, G and S are not correlated with the extensive variables of the CMPI exposed in Table 4. The small correlation among intensive and extensive is due to that we exclude the size effect of the Adjusted FTG measures. Although, among the average poverty gap G and the severity S, we find a significant high correlation of 0.97, which says that most of poverty gap is severe.

Table 5: Cross-correlation table of the Colombian Multidimensional Poverty Index (CMPI)

CMPI	H	H(urb)	H(rur)	M_0	M_1	M_2	A	G	S
H	1.0000								
H(urb)	0.7763	1.00							
	(0.00)								
H(rur)	0.8802	0.7089	1.00						
	(0.00)	(0.00)							
M_0	0.9650	0.8066	0.8667	1.00					
	(0.00)	(0.00)	(0.00)						
M_1	0.8927	0.8200	0.8264	0.9548	1.00				
	(0.00)	(0.00)	(0.00)	(0.00)					
M_2	0.8847	0.8050	0.8186	0.9481	0.9973	1.00			
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)				
A	0.81	0.75	0.78	0.93	0.93	0.93	1.00		
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)			
G	0.07	0.32	0.15	0.16	0.43	0.44	0.28	1.00	
	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
S	0.13	0.33	0.21	0.22	0.48	0.50	0.32	0.97	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	

P-value inside parenthesis. H for Incidence (Headcount Ratio), (urb) for Urban Area, (rur) for Rural Area, M_α with $\alpha = 0, 1, 2$ represents the FTG Measure of the Colombian Multidimensional Poverty Index (CMPI). A is the Intensity ($A = M_0/H$), G is the Poverty Gap ($G = M_1/M_0$) and S is the Severity ($S = M_2/M_0$)

2.2.2 Economic Internationalization and Trade

We model the effect of international trade using Free On Board (FOB), it represents the amount of U.S. dollars (in 1994) paid by the exporter to ship the commodity. This amount includes the price of the commodity and the transportation fee until the ship. The data covers one export value per year of 357,812 Colombian firms during the period 2002-2004. We aggregate FOB by municipality, using the municipality where the firm has its principal legal association. We use the resulting aggregated FOB value as a proxy for the export capacity of a given municipality, which we associate with municipality descriptors, such as education level, income per capita, directed national funds, rurality, population demographics, etc.

The motivation of this work is to study the relation between international trade and inequality. It is likely that there exists a double causation

between the descriptors of individual municipalities and the resources, wealth and taxes generated by exports. This endogeneity likely occurs in the following way: the earnings obtained via exports have a large impact locally through commerce and taxes that impact all levels of economic agents (employees, firms and institutions) within the local economy. Of great importance is the fact that the income wealth due to exports comes from international sources serving as a financial injection into the local economy. Neighbor municipalities may also be affected via "spatial spillovers", capturing the important endogenous relation between international trade and the individual characteristics of municipalities. In an attempt to externalize the system of analysis, we include measures for international commodity demand on municipal exports.

Table 7: Summary Statistics for FOBs and MI by municipality.

Variable	Mean	(Std. Dev.)	Min.	Max.
log of Exports (2002-2004)	2.1	5.5	0	24.2
N		1086		

2.2.3 More on Explanatory Variables

We begin the explanation of the municipality descriptor variables that account for the high level of variability in living standards across Colombian municipalities. The principal issue is to determine which descriptive variables to include in our model for establishing the relation between international trade and CMPI living conditions within municipalities (see Table 2), accounting for the ceteris paribus condition (see Wooldridge, 2009, Section 6.3.).

The subsections that follow outline the controls which have been divided into categorical groups: political, socioeconomic and demographic. This classification does not imply within group effects, but rather, is mainly to differentiate the context of the regional explanatory factors.

Political Controls

We include the "Sistema General de Participaciones" SGP (General System of Participation) as political controls to represent features of the Colombian, which outlines the "transfers from the central government to counties, districts and municipalities to finance the service under their charge, in health, education and others defined in Article 76 of the Law 715 from 2001"[4]. Under these regulations, 58.5% of the funds are earmarked for education, 24.5% for health services and the remaining 17% for general purposes including water, and other public services. On average, the SGP income accounts for 40% of net municipality income, compared to 20% which typically comes from internal taxes (BM, 2012). In our analysis we include SGP data from 2005

We also include the "System of cities" classification of Colombian municipalities defined by the "Departamento Nacional de Desarrollo" DNP (National Department of Development). The DNP determines whether or not to include a given municipality in the system of cities, depending on the economic and administrative importance of the municipality with respect to its regional neighbors. We incorporate this classification using a dummy variable which takes a value of 1 if the municipality belongs to the system of cities (DNP, 2012) and 0 otherwise.

We use the Rurality Index to capture the characteristic level of urban planning within each region. While every municipality may have both urban and rural areas, the density of organized land and designated urban centers may vary widely, reflecting geographic, historic, economic, and political factors which have led to the current state of urbanization. For an extended explanation of the rurality index, see PNUD (2011) and Ramirez et al. (2016).

We also use the measure of institutional vulnerability, defined by the United Nations Development Programme (UNDP), to control for the administrative capacity and the fiscal performance of a municipality.

We control a wide set of dummy variables for each Colombian department (the U.S. analog of state). The municipalities are inside depart-

[4]Definition from Colombia's Economics Department.

ments, where each department has a low but existing level of autonomy.

Table 8: Political Controls

Variable	Mean	Std. Dev.	Min.	Max.
SGP 2005 (USD).	8073658263.425	48669830677.802	1102888.61	1360772982089
Log of SGP 2005 (USD).	21.32	2.23	13.91	27.94
System of cities	0.145	0.352	0	1
Rurality Index	45.251	10.659	0	86.8
Institutional Vulnerability	47.755	19.913	0	100
N		1086		

Socioeconomic Controls

The aim of the set of Socioeconomic controls is to account for socioeconomic activities that may have a direct impact on the population living standards. This set of controls includes records of agricultural and commercial units and income per capita. These measures are complemented by a set of vulnerabilities indices pertaining to environmental, human capital, violence and economic factors. This data is taken from the PNUD (2010, 2011). The environmental vulnerability index data is taken from (PNUD, 2010) and captures biological, social and physical vulnerability to climate change. The human capital vulnerability index measures the Illiteracy Rate and the number of people of working age. The violence vulnerability index records the total number of homicides, massacres, displaced people, political victims and area of coca crops. The economical vulnerability index incorporates the Gini coefficient of territory and income. These data sources are principally from the 2005 census and are listed in the 2011 report "Colombia Rural, Razones para la esperanza", (Rural Colombia, the reason for hope) (PNUD, 2011) which quotes 2005 data sources. We also include attacks from guerrilla groups such as ELN and FARC, and incursions by Colombian military, recorded in 2008 data of the "Policía Nacional de Colombia."

Demographic Controls

In order to finish the list of controls, we describe in this subsection additional variables measuring the indigenous and afro-descended per-

Table 9: Socioeconomic Controls

Variable	Mean	Std. Dev.	Min.	Max.
Agricultural Units	1613.403	1474.845	11	18166
Comercial Units	717.586	5157.508	1	151975
Income per capita	532.208	425.811	124	6485
FARC Att.	0.178	0.809	0	11
ELN Att.	0.01	0.116	0	2
Military Operations.	0.574	2.139	0	42
Environmental Vulnerability	49.502	16.74	0	100
Human Capital Vulnerability	50.257	20.586	0	100
Violence Vulnerability	49.657	19.824	0	100
Economic Vulnerability	50.098	20.057	0	100
N		1086		

centages of the population, the illiteracy rate, the total population, and the number and density of people per area living in urban and rural areas. Additionally, we also include an index quantifying Demographic risk-of-violence defined by the United Nations and the Demographic Vulnerability Index, which includes the average number of people per family, percentage of families with a female figurehead, the average number of adults older than 64 years per family and the average number of children younger than 4 years per family.

Table 10: Demographic Controls

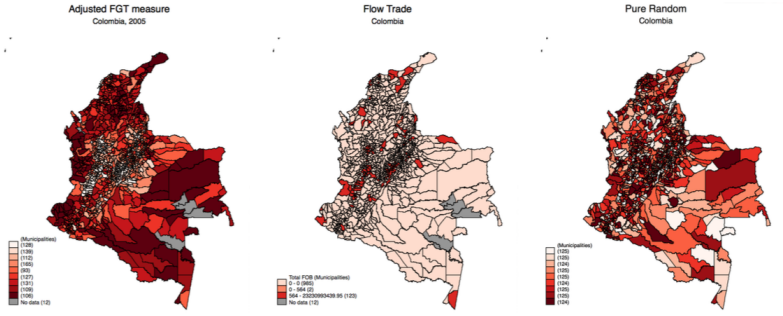
Variable	Mean	Std. Dev.	Min.	Max.
Afro-descendant Pop.	3870.886	23572.053	0	542039
Indigenous Pop.	1245.289	5411.892	0	106366
Rural Pop.	9624.627	10020.006	162	111180
Urban Pop.	30026.882	237682.999	64	6763325
Total Pop.	38912.628	238012.16	290	6740859
Density. (Pop / km^2)	147.939	636.272	0.47	13687.06
Illiteracy Rate	14.191	7.229	1.6	67.8
Demographic Vulnerability	52.033	20.337	0	100
N		1086		

2.3 Methodology

Poverty and Inequality in Colombia has a very strong spatial component in which the poorest municipalities are grouped in specific geographic zones. In order to consider this effect, it is necessary to use empirical techniques that include spatial correlations. In contrast, non spatial econometrics models assume random variability among individuals, which contradicts the empirical evidence.

Empirical evidence shows a clustered distribution within social dimensions such as poverty, income per capita and others, which is a consequence of similar endowments, weather, soil and resources shared among them, which justify the spatial econometrics approach as is shown in Figure 7a.

In order to see the spatial effect, Figure 7c shows how there would be a random pattern within the Colombian municipalities. The important issue is to account for how much is predictable within the variables of Colombian territory.



(a) $CMPI : M_2$ **(b)** Exports (2002-2004). **(c)** Random Pattern.

Figure 7: Measures over Colombian territory.

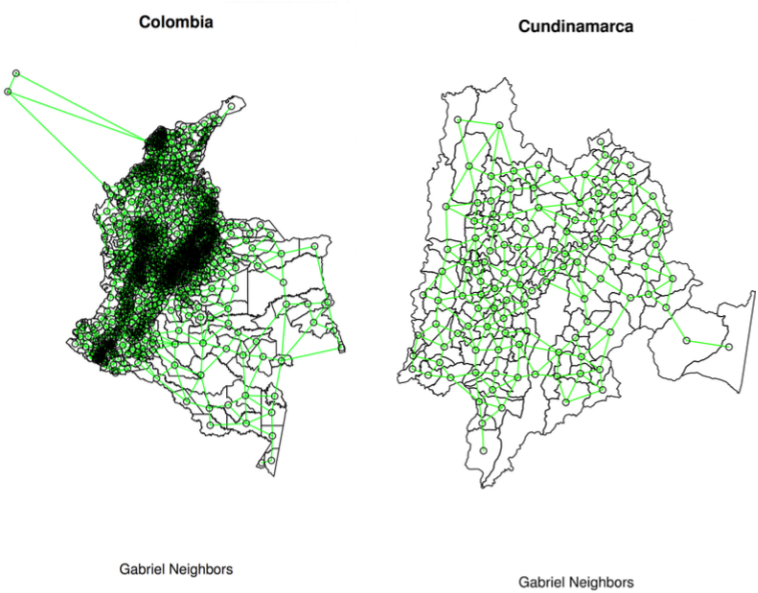
2.3.1 Spatial Facts

We use the Gabriel Neighborhood method for defining neighbors since it combines features of two other basic neighbor definitions that we tested: the Queen contiguity and the K-neighbor. The first order Queen contiguity network defines a neighbor when at least one border is shared with another region, and the K-Neighbor gives a fixed number of neighbors to every region, choosing as neighbors those regions that are K-ranked using the closest distance between centers. As the definition of Queen contiguity, the Gabriel Neighborhood incorporates geographical information related to regional borders in order to connect regions within the neighborhood matrix. Additionally a neighbor must be included in the area defined by the circle having a diameter connecting the two municipalities, if and only if, there is not any other municipality within the circle (Gabriel and Sokal, 1969; Matula and Sokal, 1980).

In the Supplementary Materials section we discuss our econometric results using both the Queen neighborhood matrix and the K-Neighbor neighborhood matrix. Together, our results using three contiguity matrix definitions serves as a robustness check for the estimation of our econometric model parameters.

Figure 8 shows the Gabriel neighborhood matrix for all municipalities with CMPI data and highlights in 8 the territory of Cundinamarca, which is the department of highest municipality density and which includes the capital.

This method is advantageous because the total number of neighbors a municipality receives is proportional to the density of surrounding municipalities. Hence, central municipalities, such as Bogota in Cundinamarca, are assigned many neighbors, whereas municipalities on the frontier are assigned few. Here, we limit our problem to the Colombian territory, accounting as neighbors only those municipalities within the frontier.



(a) Gabriel Neighbors over Colombian territory.

(b) Gabriel Neighbors over Cundinamarca.

Figure 8: Spatial Correlation: Gabriel Neighbor Definition.

2.3.2 Testing for spatial correlations using Moran's I coefficient

We use the Moran autocorrelation coefficient (Moran's I coefficient.), which is an extension of the Person product-moment correlation coefficient to a univariate series (Getis, 1995; Moran, 1950b), to test spatial correlations in regional variables. For each variable we test the significance of the Moran's I coefficient. As a result, we relax the assumption of random distributed variables to use spatial correlated variables within regions (e.g. Fig7).

We estimate the Moran's I coefficient to justify the use of spatial econometrics. Table 22 shows the Moran's I coefficient for each variable within

the econometrics exercise. Variables such as Afro-descendent population, Military operations and the number of attacks from guerrilla groups have bigger value for the contiguity spatial correlation matrix. To interpret these results we must refer back to the colonial times, when the slaves were transported from the cost to the gold mines and working zones (Acemoglu et al., 2012). They moved across contiguous municipalities where they settled; military attacks and guerrilla operations move across the territory, giving a similar result for the Moran's I coefficient. The three spatial correlation matrices give different perspectives of the spatial lag. Such differences reveal information about the way social, demographic and political information are spread among the territory.

We use municipalities trade (FOB) for each year since 2002 until 2004. They do not have a significant Moran's I coefficient. Although the Moran's I coefficient is not statistically significant, the export municipalities are few compared to the total number of municipalities, with only 12% of the total. Those municipalities are principal cities or they are around one, see Figure 7. Regardless of the lack of spatial correlation, there is a social dependence on principal cities. Therefore, a non spatial correlated variable, such as international trade by a municipality is able to spread poverty through neighbors, generating a spatial dependent problem. The effect over a region through the variables of its neighbors is known as Undirected Spatial Effect. Here we may also find Direct Spatial Effects, which are those affected by the value of the variable of the specific region and the combination of these two effects, Direct and Undirected, which is called the Total Spatial Effects.

2.3.3 Spatial Model

Spatial models are developed to include the spatial correlation effects, where the assumption of random chosen individuals does not hold. The spatial correlation could be presented in any regional measures. There are three principal models where the spatial correlation is assumed to exists. For instance, the Spatial Lag Model (SL) includes the effect of correlation in the dependent variable; the Spatial Error Model (SAR) includes the spatial correlation into the error term; and the Spatial Durbin Model

(SDM) includes the spatial correlation in the explanatory variables and in the dependent variable. For any case, moreover doing regional analysis, it is necessary to include these spatial correlation considerations in order to hold the assumption of the gaussian distribution in the error term. Combinations of these approaches confirm the family of the spatial dependence models (see Elhorst, 2010).

We use the spatial evidence on the independent variables, the CMPI measures and the explanatory variables to choose the spatial model. For this reason we use the Spatial Durbin Model (SDM), which includes the spatially lagged dependent variable and the spatially lagged explanatory variables,

$$\mathbf{Y_i} = \rho \mathbf{W Y_i} + \beta \mathbf{X_i} + \Theta \mathbf{W X_i} + \varepsilon_i, \tag{2.3}$$

where the dependent variable is $\mathbf{Y}$, the spatial correlation lag on the dependent variable is given by $\mathbf{WY_i}$, the independent variables are $\mathbf{X}$ and its spatially lagged version $\mathbf{WX_i}$, ρ and Θ are the spatial coefficients, which captures the spatial correlation. Here, we refer to the spatially lagged variable $\mathbf{WX_i}$ of a variable $\mathbf{X_i}$ as:

$$\mathbf{WX_i} = \sum_{j \in i-\text{neighbors}} X_j, \tag{2.4}$$

where in our analysis the summation $(j \in i - \text{neighbors})$ strictly refers to the neighbors defined in the Gabriel neighborhood matrix definition, as in Fig. 8, (see Gabriel and Sokal, 1969; Matula and Sokal, 1980, for details). An alternative view for a spatial lagged variable $\mathbf{WX_i}$ is based in the Adjacency matrix $\mathbf{W}$, which is a square matrix of dimension equal to the number of municipalities with binary elements, where an element $w_{ij} = 1$ means that municipalities i and j are neighbors, and $w_{ij} = 1$ means that i and j are not neighbors, with $w_{ii} = 0 \ \forall i$. Abusing of the notation we write:

$$\mathbf{WX_i} = \sum_j w_{ij} X_j = \mathbf{W} \cdot X_i. \tag{2.5}$$

We prefer to use the Spatial Durbin Model over the Spatial Lag Autocorrelation model, because of the evidence of the high spatial correlations measured with the Moran's I coefficient in the independent variables. In case these pair correlations would have a similar value the Spatial Error Model (SAR) would be the best choice, but the intensity of the spatial correlation changes among pair variables, which explains why the SDM is a better choice over the SAR model. The Moran's I coefficient shows a high variability among explanatory variables, this is an effect of their different nature.

2.3.4 The Model

Variable Selection

We use the correlation matrix represented in Figure 9 to choose the variables used in the model. The variables must have orthogonal information, therefore we must choose those that are not correlated. Additionally, we apply a significance test using a 95% confidence level and the Wards hierarchical agglomerative clustering method (within the black boxes) to choose the variable which best represents the clustered group to order to selected it.

We choose the variable number of Agricultural Units among the variables Agricultural and Commercial Units per capita, Rural, Indigenous, Afrocolombian and Literacy population per capita, which compose the biggest cluster. The number of Agricultural Units is the variable with the least number of correlations and represents demographics as Rural, Ethnic minorities and Socio-economic conditions such as the Literacy Population. Additionally, we account for the Total number of export firms as the principal dependent variable, therefore we must avoid double counting the total units by not adding the variable Commercial Units, which belongs to the same cluster. From the second biggest cluster, we neglect Total Vulnerability which is the variable which is most correlated with the others. Despite the fact that almost every additional vulnerability belongs to the same cluster, we decided to take them all because they are not correlated with the other Vulnerabilities. To finish the variable selec-

tion, we neglect the Environmental Vulnerability and the system of cities, because they are too correlated as negative and positive respectively with the FOB through the production, and by its definition.

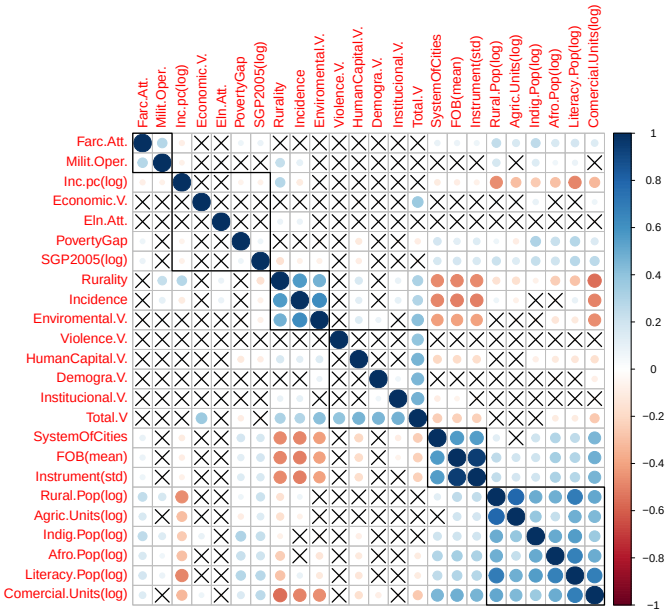


Figure 9: Correlation Matrix. The variables are ordered and using the Wards hierarchical agglomerative clustering method (within black boxes). The X represents those pair-correlation not correlated within a 95% confidence level. The bar to the right has the scale that goes from -1 (dark red) for perfect anti correlation and 1 (dark blue) for perfect correlation.

As an interesting fact, we highlight the small correlation between the Income per capita in 2005 and the average FOB from 2002 until 2004. We explain this result, by arguing that there are inefficient institutions, which lead to a scarce redistribution effect in municipalities with the higher exports. The money obtained by exports represented by the FOB does not stay within the municipality, which is typical case for a centralized economy.

To resume, using the significance and cluster analysis, we choose the following dependent variables: the logarithm average value in dollars of the exports between 2002-2004 (FOB) per capita, Incidence, number of Agricultural Units per capita (Agric.Units p.c.), Income per capita (Inc.pc in log base 10), number of Attacks for the ELN guerrillas (Eln.Att.), number of Attacks for the FARC guerrillas (Farc.Att.), number of Military Operations (Milit.Oper.), the logarithm of the General System of Participation (SGP2005, which is a measure per capita), Demographic Vulnerability (Demogra.V.), Human Capital Vulnerability (HumanCapital.V.), Violence Vulnerability (Violence.V.), Institucional Vulnerability (Institucional.V.) and Economic Vulnerability (Economic.V.) from 2005.

Identification Strategy

Our econometrics model is a version of the Spatial Durbin Model,

$$G_i = \rho \mathbf{W} G_i + \alpha FOB_i + \alpha^w \mathbf{W} FOB_i + \beta X_i + \beta^w \mathbf{W} X_i + u_i, \qquad (2.6)$$

where the subindex i refers to each municipality, the dependent variable G represents the Poverty Gap; $\mathbf{W}G_i$ represents the poverty of the neighbors of municipality i, here we use the abuse of notation of Eq. 2.5; The logarithm of average Free On Board quantities exported by each municipality within the years 2002-2004 is given by FOB_i; $\mathbf{W}FOB_i$ represents the exports of the neighbors of municipality i, with the abuse of the notation explained in Eq. 2.5; X which are the explanatory variables; $\mathbf{W}X_i$ represents the values of the explanatory variables of the neighbors of municipality i, also abusing of the notation (Eq. 2.5); u_i is the error term; ρ is the spatial correlation coefficient, it measures the relation of the dependent variable for a municipality i with its neighbors.

We explain the identification technique by measuring how international trade trough the FOB_i variable affects the Poverty GAP (G) using intensive variables. Using the theoretical model in Helpman et al. (2010), we know that under international trade a high quality of workers tend to make more than a low quality of workers, generating an increment

on wealth distribution for the highest worker, which generates an increase in inequality. Instead, we face the question of how international trade affects people below the poverty line. This question is relevant because it can lead to opposite international trade effects over inequality. A positive effect would be that despite the increase in inequality through the wage difference of high and low quality workers, international trade could improve the quality of the poor through an efficient re-distribution system. In contrast, we may find a negative effect due to an inefficient re-distribution system, because the poor are excluded from the production chain. In this last view, poor people are worse off because of the the price increase of goods and services, generated by a high inflation due to a greater income of high quality workers. Specifically in the Colombian case, most of exports ($\sim 90\%$) are commodities, generating two principal negative effects: the detriment of natural resources and a low labor inclusion for the poor, among other negative effects that are frequent in unequal societies.

2.4 Results

We use two different estimations techniques: Ordinary Least Square (OLS) and a Durbin Spatial Model (SDM). For each technique we consider two Export dependent variables: the Export dummy, which gives the value of 1 for export municipalities and 0 otherwise and the logarithm of the average export since 2002 until 2004. We use the same set of control variables in both cases.

We present the results of the two estimation techniques for every type of Export (dummy and FOB) in Table 11. Table 12 shows the impacts estimates for the Spatial model estimation, which includes: the Direct Spatial effect measures how the regional area is affected by its own measure, the Indirect spatial effects measures how the region is affected by the variables of their neighbors; and the Total spatial effect is the sum of the Direct and Indirect, which accounts for the total impact of the specific control variable over the Poverty Gap.

The Table 11 shows four high significant variables in each model.

Two of them indicate a reduction in the poverty gap within municipalities: the number of Agricultural Units (in logarithms) and the average Income per capita (in logarithms). Variables such as the System of General Participation (in logarithms) and Military operations as a presence of the power of violence from the state also show a reduction of the power of inequality, although they are not significant. The positive effect of poverty Incidence shows that a municipality with a high number of poor individuals also has a high shortfall of deprivations. Additionally, we find non-significant but positive variables, such as the number of attacks by the non-state armed groups guerrillas (F.A.R.C. and E.L.N.). Both Export variables (dummy and FOB) are positive and significant, and this result leads to an increment of the poverty gap within export municipalities. With the use of variables in percentage and logarithms, we are able to read the coefficients directly, for instance one percentage of increment produces the value of the coefficient increment (depending whether is positive or negative). Therefore, using the SDM we interpret the effect as an increase in inequality within the group of export municipalities. Additionaly, it has the statistically contrary effect of increasing the Income per capita or adding new Agricultural Units. Furthermore, using the average of exports we find that the increment quantity exported in dollars increases inequality. To interpret the results of the spatial model, we use the Impacts Estimates with its different spatial effects.

Table 11: Regressions results of the OLS and Spatial Durbin Model using the Poverty Gap as dependent variable.

	Dependent variable: Poverty Gap (G)			
	OLS		*SDM*	
	(1)	(2)	(3)	(4)
Export Dummy	0.067^{***}		0.070^{***}	
lg Mean $Exports_t$		0.004^{***}		0.005^{***}
Incidende	0.190^{**}	0.184^{**}	0.203^{***}	0.197^{***}
lg Agric. U.	-0.015^{***}	-0.015^{***}	-0.015^{***}	-0.015^{***}
lg Income p.c.	-0.016^{**}	-0.017^{**}	-0.014^{*}	-0.015^{**}
ELN Atta.	0.013	0.013	0.010	0.011
FARC Atta.	0.002	0.003	0.001	0.002
Military Inc.	-0.001	-0.001	-0.001	-0.001
lg SGP (2005)	-0.0004	-0.0004	-0.0002	-0.0002
Demographic V.	0.00005	0.00004	0.00002	0.00000
Violence V.	-0.0001	-0.0001	-0.0001	-0.0001
Institutional V.	-0.0002	-0.0001	-0.0001	-0.0001
Economic V.	0.0001	0.0001	0.0001	0.0001
Lagg. Export Dummy			0.046^{**}	
Lagg. lg Mean $Exports_t$				0.004^{**}
Lagg. Incidende			0.181	0.167
Lagg. lg Agric. U.			-0.002	-0.002
Lagg. lg Income p.c.			0.017	0.017
Lagg. ELN Atta.			-0.055	-0.054
Lagg. FARC Atta.			0.002	0.002
Lagg. Military Inc.			0.002	0.002
Lagg. lg SGP (2005)			0.0002	0.0001
Lagg. Demographic V.			-0.0003	-0.0003
Lagg. Violence V.			0.0004	0.0004
Lagg. Institutional V.			0.0002	0.0003
Lagg. Economic V.			0.0003	0.0003
Constant	0.888^{***}	0.894^{***}	0.654^{***}	0.666^{***}
Dummies for Colombian Departments (States)				
Observations	1,086	1,086	1,086	1,086
Adjusted R^2	0.345	0.340		
Log Likelihood			1,433.925	1,430.964
Note:				*p<0.05; **p<0.01; ***p<0.001

The Table 12 shows the two different spatial effects using the Spatial

Durbin Model. Neighbors increase the effect of the Export dummy up to 3%. The effect of the average of Exports is 8% accounting for spatial effects. Both cases show a greater direct effect compared with the indirect effects. by including the neighbors variables into the model, we find that most of the negative effect of trade spreads a rate of 3% more inequality through the network. The Total effect of the income per capita is not significant. Also by increasing 1% the number of Agricultural Units and the variables clustered with it, for instance the number of Commercial Units, municipalities decrease inequality by 1.5%.

Table 12: Impacts Estimate of the 2SLS Spatial Durbin Model using the Logarithm of the average of deprivations as dependent variable.

$y = G$						
(Poverty Gap)	Direct	Indirect	Total	Direct	Indirect	Total
Export Dummy	0.069***	0.038*	0.107***			
	(0.007)	(0.015)	(0.016)			
lg Mean $Exports_t$				0.005***	0.003**	0.008***
				(0.001)	(0.001)	(0.001)
Incidende	0.200***	0.154	0.354**	0.194**	0.141	0.335**
	(0.056)	(0.118)	(0.130)	(0.058)	(0.113)	(0.127)
lg Agric. U.	−0.015***	−0.001	−0.016*	−0.015***	−0.0003	−0.015*
	(0.003)	(0.006)	(0.006)	(0.003)	(0.006)	(0.006)
lg Income p.c.	−0.014**	0.017	0.003	−0.015**	0.017	0.002
	(0.005)	(0.011)	(0.012)	(0.005)	(0.011)	(0.012)
ELN Atta.	0.011	−0.053	−0.041	0.012	−0.051	−0.040
	(0.018)	(0.040)	(0.043)	(0.018)	(0.040)	(0.044)
FARC Atta.	0.001	0.001	0.002	0.002	0.002	0.004
	(0.003)	(0.006)	(0.007)	(0.003)	(0.006)	(0.007)
Military Inc.	−0.001	0.002	0.001	−0.001	0.002	0.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
lg SGP (2005)	−0.0002	0.0002	0.00002	−0.0002	0.0001	−0.0001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Demographic V.	0.00002	−0.0002	−0.0002	0.00001	−0.0003	−0.0003
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
Violence V.	−0.0001	0.0004	0.0003	−0.0001	0.0004	0.0003
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
Institutional V.	−0.0001	0.0002	0.0001	−0.0001	0.0003	0.0002
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
Economic V.	0.0001	0.0002	0.0003	0.0001	0.0002	0.0003
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
			Dummies for Colombian Departments (States)			

Note: Standard deviation between parenthesis.
"L." for logarithm and V. for Vulnerability. *p<0.05; **p<0.01; ***p<0.001

2.5 Conclusions

By related inequality measures with aggregated exports at municipality level, we studied how the export distribution across municipalities affects the Poverty Gap. We chose municipalities at a regional level to capture the minimum political area where exports have economic consequences. We also account for spatial effects to include how neighbor

municipalities are affected through international trade.

We use the Poverty Gap as a dependent variable. We take the definition of the Poverty Gap, included in the FTG measures of the Colombian Multidimensional Poverty Index (CMPI), where $G = M_1/M_0$, to have the intensive version of the average Poverty Gap. We choose the Poverty Gap (G) as it represents the average shortfall of the socio-economic dimensions measured for every person within the Colombian Census of 2005. As dependent variables, we use a set of control variables selected from a bigger group, using a cluster analysis and a correlation matrix. We also chose two independent variables for Exports: a Dummy export, which gives a 1 to those municipalities that have exporter firms and the logarithm of the average Exports since 2002 and 2004. We use two estimation techniques: Ordinary Least Squares (OLS) and the Spatial Durbin Model (SDM).

In Figure 10, we draw the empirical distribution function of the poverty gap G for export municipalities and no export municipalities. It shows a slight overall increase of the poverty gap in export municipalities compared with non export municipalities, and a remarkable additional peak corresponding to the extreme poverty in exporting municipalities. Table 13 shows those exports municipalities with a poverty gap greater of 0.85, which we believe are part of the right peak in Fig. 10. We compare the departments of Table 13 with the department aggregated commodity exports in Table 14[5], it shows that most of the departments in Table 13 are within the 15 highest commodity exports, Caldas is the only one excluded. We also find a strong neighborhood effect, the most frequent department in Table 13 is Cundinamarca, which geographically includes Bogota (Fig. 8b). Bogota as the Colombian capital seems to spread poverty on its neighbors. However, there are other affected departments: Cesar, Guajira, Norte de Santander, Tolima and Casanare, which are well known as commodity exporters, each one in different commodities (coal, mineral, petroleum and precious stones). This evidence lead us to relate commodities and the poverty gap, although this relation should be tested

[5]Here we would prefer to report municipalities but they do not exist, instead we include statistics on departments (states).

with a more detailed evidence and futures measures of the CMPI by municipalities, which today does not exist. Table 15 shows the descriptive statistics of the poverty measures define in the CMPI, section 2.2. Exporter municipalities show higher values in poverty gap (G) and severity (S) compared with Non exporter municipalities. G and S are poverty measures taken only over the poor, their greater values in municipalities with exports, together with the lower values of the Headcount Ratio, show that in average, the exporter municipalities have a smaller fraction of population below the poverty line, however the poor in these municipalities are more deprived compared with the non export municipalities.

We find evidence that shows the relation between the international trade and the higher values of the export municipalities compared with non export municipalities, as is shown in the Fig. 10. As a fact, the difference between export and no export municipalities is larger for high values of inequality. This fact gives relevance to the present analysis. Using empirical econometric analysis, we attempt to disentangle other possible factors, which could generate the poverty gap difference between these two types of municipalities.

Thus, the empirical evidence opens an interpretation of how international trade affects inequality based in the nature of the exports and the efficiency of the firm optimization problem. For instance, we can address the following question: "Is this specific selection of municipalities a natural selection chosen by export firms in function on their productivity accounting for high inequality condition for their benefit?". A positive answer to that question may lead to a belief that the productive chain is based in poor social standards. For the Colombian case, we can not affirm that firms choose where to locate, as a fact, in 2005 the internal economy was based in the export of commodities. Firms cannot choose the location when they have to extract minerals from the land. This fact implies that the inequality, produced by export firms, is not an effect of the efficient firm maximization problem, instead we interpret that the decreasing in social standards is a cost that has to be paid in order to export commodities. This view resolve a possible source of endogeneity in our empirical analysis respect to the firm allocation.

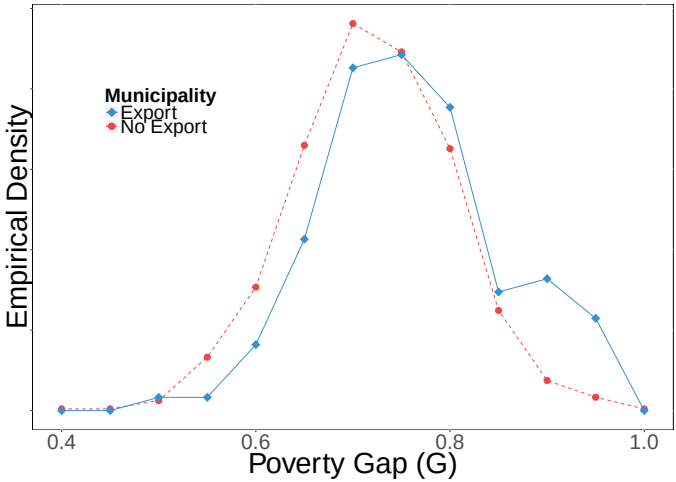


Figure 10: Empirical Distribution function of Poverty Gap (G) for municipalities with exports (continuous line with blue diamonds points) and for municipalities without exports (dashed line with red circles points).

In regards of the results show in Table 11 and Table 12, we analyze the relation between the Number of Agricultural Units, as a proxy for the development of rural areas and inequality in municipalities, moreover using the fact of its significant anti-correlation with inequality, we find an asymmetry in the effect of international trade over urban and rural development. We can affirm that the more developed is the rural area, less negative influence it will have from international trade. Under this view, poor households in exports municipalities are more likely to worsen their life condition in urban centers than in rural areas, moreover if there is a high number agricultural firms. This asymmetry between urban and rural areas leads to a disconnection of the production chain involved in exports and rural development. Furthermore, we argue that this asymmetry is due to an efficient firm optimization effect of agricultural units over an inefficient optimization process of commodity extraction firms. For agricultural firms, the optimization process is en-

Table 13: Descriptive statistics of right peak in Export municipalities for $G > 0.85$.

Municipality	Department	Poverty Gap
Bogota D.C.	Bogot D.C.	0.91
La Dorada	Caldas	0.91
Aguachica	Cesar	0.86
Cajica	Cundinamarca	0.90
Chia	Cundinamarca	0.86
Choconta	Cundinamarca	0.86
El Rosal	Cundinamarca	0.93
Funza	Cundinamarca	0.90
Girardot	Cundinamarca	0.93
Madrid	Cundinamarca	0.89
Sopo	Cundinamarca	0.91
Tocancipa	Cundinamarca	0.86
Riohacha	Guajira	0.86
Maicao	Guajira	0.88
Cucuta	Norte de Santander	0.87
Ibague	Tolima	0.87
Yopal	Casanare	0.89

Additional peak municipalities corresponding to the extreme poverty, with a poverty gap grater than 0.85. We also report the department (state) of each municipality.

dogenous leading to a improvement of life condition for households in that area. This goes in opposite direction, and sign, of the exports, which are a proxy for commodities extraction.

The main contribution of this work is in the analysis of the way international trade increases inequality through deprivations within municipalities. This result gives a new perspective to the discussion of how international trade affects inequality. We found a negative effect of international trade over inequality. This result is aligned with the theoretical evidence in Helpman et al. (2010). In addition, we contribute with a new perspective in defining inequality, which instead of using differences in wages as a proxy for inequality, we use the derived variable from FTG measure for $\alpha = 1, 2$, the average Poverty Gap $G = M_1/M_0$, given contained in the Colombian Multidimensional Poverty Index (CMPI). The

Table 14: Extractive Exports (F.O.B.) in 2004 dollars by Colombian departments (states).

Department	Coal	Mineral	Petroleum	Precious Stones	Total	Share
Cesar	13101307714.97		364.13		13101308079.10	0.204
Guajira	11465612934.15	3148.76	1195835855.97	0.46	12661451938.88	0.198
Antioquia	10705932.4	137076061.5	22551374.27	9182423282	9352756650	0.146
Meta	0	172.24	6919255164	0	6919255337	0.108
Bolívar	104283132.8	2095905.97	5451538518	3536448.52	5561454005	0.087
Casanare	0	8	0	4969438342.95	4969438350.95	0.078
Arauca	0	1025.57	2498692962.77	0	2498693988.34	0.039
Santander	2626920.53	26975007.85	1359673164.62	79846153.23	1469121246.23	0.023
Huila	0	958994.28	1285010817	0	1285969811	0.020
Boyaca	812543805.9	2415888.78	200	447915053.9	1267043977	0.020
Cundinamarca	892297714.5	892297714.5	211590026.6	625925.15	1106929555	0.017
Bogota	263778081	55903538.49	334131574.5	444150439.3	1097963633	0.017
Valle del cauca	179310185.6	11442983.25	9145841.63	820986768.7	1020885779	0.016
Norte de Santander	681167036.6	74273800.88	13686502.23	518780.8	769646120.5	0.012
Tolima	0	92516.2	596181751	32522.77	596306790	0.009

Data from the statistical national department DANE. We report the 15 highest commodities departments. We account Coal, Mineral, Petroleum and Precious stones as principal commodities. We also report Total and Share values including the rest of departments.

Table 15: Descriptive statistics of Export and Non Exports municipalities.

Mun.	Obs.	H	H_u	H_r	A	G	S	M_0	M_1	M_2
NExp.	976	0.72	0.55	0.81	0.50	0.70	0.62	0.36	0.26	0.23
Exp.	122	0.46	0.38	0.64	0.46	0.74	0.64	0.22	0.16	0.14

Mun. for municipality; NExp. for municipalities without export; Exp. for municipalities with exports; Obs. for number of observations; H for Headcount Ratio; H_u for urban headcount ratio; H_r for rural headcount ratio; A for Intensity; G for Poverty Gap: S for Severity; M_0 for the Adjusted Headcount Ratio; M_1 for the Adjusted Poverty Gap; and M_2 for the Adjusted Severity.

CMPI gives a more detailed information of the social standards, connecting them to local policy. Therefore, we add the deprivation state of the poor to the definition of the inequality. This work shows a specific case where international trade reduces the life conditions of the poor. The result shows evidence that poor people have worse conditions in a high export municipality and even worse living in a municipality which is a neighbor of one of the export municipalities. We attribute the negative effect of international trade on the poverty gap in the Colombian case to the very high percentage of commodities in the exports. This generates an inefficient optimization process on the chose of the geographical ar-

eas, which affect local market and development. Consequently, this generates price increment of goods and services. Additionally, commodities generate negative effects such as the depletion of natural resources and the exclusion of labor from the the poor, among other negative effects that are common in highly unequal societies.

Colombian development policy should be focused in generating conditions to firm creation in those places affected by the commodities economy. We show in this chapter that these efforts could have positive effects in the life condition of the poor. We recommend that the Colombian efforts should be directed in providing the sufficient conditions to the creation of agricultural and commercial firms, we believe these are efficient alternatives to improve life conditions affected by the commodities extraction. Additionally, we highlight that it must be considered spacial spillovers for policy design, as result we find that the commodities economy has negative effects for the poor, not only in the municipalities where are located the extraction firms, but also in their neighborhood.

Chapter 3

Innovation competitiveness of Nations and Regions: A view from Patent Innovation

3.1 Introduction

Network techniques have been growing in modern economics for generations. Nonetheless, agents and their relations have been within the fundamentals of the economics literature only since the paradigm of network science, which is also composed by agents and their relations. The thanks to the inclusion of network science in the economic literature, it has been possible to study how the dynamic of those micro-iterations among agents generate macro-outputs. Network science started in physics as a way of linking micro particles with the macro measures such as temperature. In economics, network science links people, countries or other economic agents with poverty, inequality, inflation and growth among other aggregated measures. The aim of using in economics techniques from other sciences, such as physics in the case of network science, is to complement the understanding of the classic measures and to open

new perspectives for creating measures with relevant meaning, instead of replacing the classic approach. Despite the fact that most of the techniques migrate from physics to economics, sometimes the migration row reverses its direction. For instance, using criminal networks Ballester et al. (2006) finds a relation between the classic Nash Equilibrium and the network science algorithm Bonacich Centrality, which is a well-known centrality measure. Using these fundamentally different techniques, it reaches the same answer, finding *the key player: "the player who, once removed, leads to the optimal change in aggregate activity"*.

Using network sciences, we link inventors with production in countries and cities. In particular, we use a network science measure (Hausmann, ed, 2013) called *The HH complexity algorithm*, (HH by its authors: Hausmann and Hidalgo). The HH algorithm uses a specific type of network with two types of agents *(bipartite networks)*: Countries and Goods. The HH algorithm, in its original version, uses the export performance of countries. Using a specific metric, it measures *Diversity*, which is defined as the *capabilities* of a country to export, and *Ubiquity*, which is the required *capability* that a country need to export certain goods. Based on the definitions of Diversity and Ubiquity, the HH algorithm ranks countries and goods.

To rank countries and goods, it is necessary to create comparable measures. Therefore, using a normalized version of the Ubiquity, Hausmann, ed (2013) defined by Economic Complexity Index (ECI), it measures the capability of countries to export complex goods. Likewise, to be able to compare goods, using a normalized version of the Diversity, the authors define the Product Complexity Index (PCI): it measures the capability needed by countries to export a certain good. Countries with relatively high ECI values are those able to produce *complex* goods, while goods with relatively high PCI values are produced by most *complex* countries. Here, complexity is a measure of capabilities: *produce capability* for countries and *capability needs* for goods. In order to avoid the nonsense of a cyclic definition among Diversity and Ubiquity, i.e., countries with a high produce capability produce goods that need a high capability to be produced (and vice-versa), it is necessary to study in

depth a global view of why ECI and PCI, as is defined by Hausmann and Hidalgo, are important in understanding the development and their relation to economic growth.

The idea behind the HH algorithm is to define a measure that explains the effect of how the portfolio of exported goods, by a specific country, incentivises the creation of other products. The capability of a country to export goods, which at the same time are able to produce other goods, is what the authors call country complexity, as represented in the ECI. Meanwhile, the capability needed for a certain good to be produced, because its creation also depends on the creation of other goods, is what the authors call product complexity, represented in the PCI. Here there is a hidden implication linking complexity with development. It is assumed that a highly-complex economy, which is able to export complex products, should have efficient institutions such as courts, regulators and educational systems among others, which implies high social investment and development. Although high complexity goods are only produced for a few high complex countries, the existence of goods, which are produced in a few countries, this does not imply that these countries have a high level of complexity. For instance, we find that some countries have natural resources, e.g. a small group of countries are able to export diamonds, but the knowledge to obtain a diamond is lower compared with the knowledge to obtain a microchip, and the capability to use diamonds to build other products is lower compared to the capability of microchips to create other products. If we compare two countries with one single good, with first one producing diamonds, while the second producing microchips, the latter country is expected to have a higher complexity in the long run, because the capability to produce microchips is more *applicable* to produce other types of goods, even though both countries produce only one good. Using this framework, the authors explain economic growth though a function of complexity.

Among other approaches of complexity measures Cristelli et al. (2013) defines the PC algorithm (called the *PC algorithm* by two of its authors: Pietronero and Cristelli). Basically, it differs from the HH algorithm as it uses the harmonic function (Rao et al., 2014) as metric. Cimini et al.

(2014) uses the PC algorithm over Publication data to reckon the complexity of academic subjects.

We use the HH algorithm to find a complexity measure using Patents databases. Specifically we use the OECD REGPAT DATABASE (Dernis and Khan, 2004). This database also has information about the triadic family of Patents. The triadic family is a set of patents that have a series of corresponding patents filed at the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO) and the Japan Patent Office (JPO). For the same invention, by the same applicant or inventor, they are defined to protect a single invention (Dernis and Khan, 2004). Likewise the original version of the HH algorithm in Hausmann, ed (2013), we define a bipartite network between locations (Country and Regions) and the number of Patent classes. We take the countries from the Inventors information within the database to extract the regions: we map NUTS3 regions (Dernis and Khan, 2004), reported by Inventors; and we aggregate the number of Patents classes at city level to identify a well-defined geographical locations. Overall, we choose NUTS3 regions, or simply Regions as a proxy for Cities. We use "proxy," because the metropolitan area of Paris, for instance, has several NUTS3 regions. Additionally, we take the number of Patent classes defining two principal assumptions: 1) triadic families are approximations for technological innovation, neglecting all the technological developments that are not intellectually protected through patent regulation and 2) the first patent of the triadic family is the most representative patent of the triadic family, where we choose its classes as to represent the classes for the triadic family, (see Dernis and Khan, 2004; Martinez, 2010; Zuniga and Guellec, 2009). We do these assumptions to include only the technologies that have had an important impact and to avoid the double counting of the same technology by including Patents of the same triadic family.

This results in finding a set of complex Patent classes, countries and cities. This work attempts to find a specific group of patents classes, as a proxy of innovative technology, that a certain country would like to invest in in order to have the best impact in its economy. These results are between the findings in Hausmann, ed (2013) and Cimini et al. (2014)

where the patent classes are a bridge between education subjects and goods.

In the following section we explain the two methods that we use in this work: the HH algorithm and the Hierarchical Cluster Analysis. We present the results in Section 3.3. To finalize we discuss this in Section 3.4.

3.2 Data and Methods

3.2.1 Patent Data

We analyzed the last Edition of the OECD REGPAT DATABASE (July 2014). This dataset of Patents has been regionalized across OECD countries (28 EU countries, Brazil, China, India, the Russian Federation and South Africa). Furthermore, it includes detailed information of the Patents, such as the application date, Patent Classes and Inventors. The last are regionalized at the Country and NUTS3 level.

By definition, according to the New Oxford American Dictionary, a Patent is *a license conferring a right or title for a set period, especially the sole right to exclude others from making, using, or selling an invention.* Patents have two main points: it is an invention, and it gives to the holder the right to use, sell, offer for sale and/or imports any product or technology protected by the patents claims. It is a document that means of the Patent owner has intellectual property over an invention. Under this aim, *The World Intellectual Property Organization (WIPO) was created in 1967 "to encourage creative activity, to promote the protection of intellectual property throughout the world."* (Dernis and Khan, 2004). The World Intellectual Property Organization defines types of Patent Classifications through the International Patent Classification (IPC) system. It *is used to classify patents and utility models according to the different areas of technology to which they pertain* (Organization, 2016). We use the Patent database from The World Intellectual Property Organization. It is composed by three principal tables: Patent Details, Patent Classes and Patent Inventors. Within each table we choose to focus on the following fields:

- Patent Details
  - Publication Number: it changes depending of the Classification system. For instance, in the USPTO it is a number composed for a country code (only two letters: US) and a serial number from 1-12 digits.
  - Triadic Family Number: it is a serial number to identify Patents within the same triadic family, i.e., two Patents belong to the same triadic family if the have the same Triadic Family Number.
  - Application Date: it is the actual filing date of the patent application. It is also known as the filing date, as it sets a cutoff date after which any public disclosures will not form prior art. After a Patent is filed, it receives a Publication Number.
- Patent Classes
  - Publication Number
  - Class: it is a system for examiners of patent offices to classify the Patent according to the technical features of their content. It is established by the International Patent Classification (IPC). There are other classifications, such as the United States Patent Classification (USPC) by the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO) and the Japan Patent Office (JPO) among others.
  - Class Type: it is a letter that represents the type of classification. Specifically we are using the IPC, USPC, EPO and JPC.
- Patent Inventor
  - Publication Number
  - Country: it is the Country reported by the Inventor of the Patent.
  - NUTS3 Region: it is the NUTS3 Region associated with the address reported by the Inventor of the Patent. The NUTS3 is

the third level of Nomenclature of Territorial Units for Statistics (NUTS).

We use 7-digits (IPC7) of the International Patent Classification system to determine the class of the Patent. The 7-digit code is composed by:

- Section: it represents the whole body of knowledge which may be properly regarded to the field of invention patents. Sections are the highest level of the classification and have one of the capital letters (A-H).
  - A: Human Necessities.
  - B: Performing Operations and Transporting.
  - C: Chemistry and Metallurgy.
  - D: Textiles and Paper.
  - E: Fixed Constructions.
  - F: Mechanical Engineering, Lighting, Heating, Weapons and Blasting.
  - G: Physics.
  - H: Electricity.
- Class: it has two digital numbers and represents the second hierarchical level of the classification:
  - Example: H01 Basic Electric Elements.
- Subclass: It has one capital letter and is the third hierarchical level of the classification.
  - Example: H01S Devices Using Stimulated Emission.
- Group/Subgroup: it has one- to three-digit numbers, the oblique stroke and the number 00. Subgroups are subdivisions under the main groups. It differs from the group because at least two digits are different from 00.

- Group Example: H01S 3/00 Lasers.
- Subgroup Example: H01S 3/14 Lasers are characterized by the material used as the active medium.

Preprocessing the Patent Data

We use the three tables of the Patent database:

- Patent Details: it has the fields of the Patent Number, filing year and triadic number.

  1. We ascendantly order by Triadic Number, Year and Patent Number.

  2. We use window operations to choose only the first row within each window defined by the ordered Triadic Number, Year and Patent Number.

- Patent Class: it has the fields of the Patent Number, Classification Number and Classification type.

  1. We eliminate the the oblique stroke, take the first seven numbers of each Patent and allow only the Classification type that we are interested in: IPC, USPT, JPO and EPO.

  2. As this table has duplicate rows for Patent Numbers, because one Patent could have multiple classification codes, we aggregate into one row the object of every Patent Number.

- Patent Inventor: it has the fields of the Patent Number, country and NUTS3 Region.

Finally, we join the three tables using the Patent Number as the merge key and unwrap it by the Patent Classification Code. This result in having two new tables: the Country table, which contains the filing Year, Country and 7-digits classification number and the Region table with the filing Year, Region and 7-digits classification number.

Descriptive Statistics of Country table

We present some descriptive statistics. In Table 18, we show the Top-10 IPC7 classes in countries/regions, and how many Triadic Families have the Top-10 countries/regions. Likewise in Table 16 and 17, we show the Top-10 countries and regions respectively.

Table 16: The Top-11 countries, in order of their frequency in the specified years Country Table. The first column lists Countries, and the second shows the number of Triadic Families in the countries

Country	Number of Triadic Families
United States	635916
Germany	455229
Japan	372764
France	134042
United Kingdom	129259
Switzerland	61063
The Netherlands	48359
Italy	42395
Sweden	40868
Belgium	33171
Canada	32362

Table 17: The Top-10 Regions (+1 Not Classified), in order of their frequency in the Country Table by specified years. The first column lists the Regions, and the second shows the number of Triadic Families in the Regions.

Region	Number of Triadic Families
Tokyo	89453
Osaka	44378
Kanagawa	43982
Not classified	36030
Middlesex County, MA	24316
Santa Clara County, CA	22042
Hyogo	20665
Aichi	20625
San Diego County, CA	19066
Saitama	17672
San Mateo County, CA	17516

Table 18: The Top-10 categories (as defined by the WIPO industrial aggregation), in order of their frequency in the specified years Country/Region Table. The first column lists the 7-digits IPC (IPC7) code of the first triadic patent class, the second shows the frequency of the classes and the third column is a brief description of the class

IPC7	Freq.	Description	
A61K031	59535	Pharma	Medicinal preparations containing organic active ingredients.
C12N015	24691	Pharma	Mutation or genetic engineering; DNA or RNA concerning genetic engineering, vectors, e.g. plasmids, or their isolation, preparation or purification.
A61K038	20402	Pharma	Medicinal preparations containing peptides.
A01N043	19694	Pesticides agro chemical products	Biocides, pest repellants or attractants, or plant growth regulators containing heterocyclic compounds.
G01N033	19645	Measuring instruments	Investigating or analyzing materials by specific methods not covered by the preceding groups.
C07K014	18776	Pharma	Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives.
A61P035	18616	Pharma	Antineoplastic agents.
A61P025	18203	Pharma	Drugs for disorders of the nervous system.
H01L021	15870	Electronic components	Processes or apparatus adapted for the manufacture or treatment of semiconductor or solid state devices or of parts.
A61K009	15557	Pharma	Medicinal preparations characterized by special physical form.

3.2.2 The HH Algorithm

The Hausmann and Hidalgo (HH) algorithm (Hausmann, ed, 2013) uses a bipartite network with two types of agents: countries and goods. For each type of agent, it defines a metric: *Diversity* for countries, which is define as the capabilities of a country to export, and *Ubiquity* for goods, which is the required capability that a country need to export certain good. In its original version, uses the export performance of countries. Based on the definitions of Diversity and Ubiquity, the HH algorithm ranks countries and goods using the following algorithm[1]:

$$\mathbf{K} \equiv k_{c,N} = \frac{1}{k_{c,0}} \sum_{p} M_{cp} \cdot k_{p,N-1} \tag{3.1}$$

$$\mathbf{Q} \equiv k_{p,N} = \frac{1}{k_{p,0}} \sum_{c} M_{cp} \cdot k_{c,N-1} \tag{3.2}$$

The $k_{c,N}$ and $k_{p,N}$ represent the *Diversity* for countries and *Ubiquity* for goods respectively. These two fundamental measures are normalized in each step. Their final values are given by an iterative process and their initial values are:

$$k_{c,0} = \sum_{p} M_{cp} \tag{3.3}$$

$$k_{p,0} = \sum_{c} M_{cp} \tag{3.4}$$

The converge condition is given by:

$$|\max(\mathbf{k_N} - \mathbf{k_{N-1}})| < \epsilon \tag{3.5}$$

where $\max()$ is the maximum value of the i_{th} component of the difference and $\epsilon = 10^{-7}$

To define the matrix M_{cp}, we must define a measure that allows us to

[1] For a deeper discussion oh the Hausmann and Hidalgo (HH) algorithm see the section 3.1.

compare the country and products. For this, we use the Revealed Comparative Advantage (RCA). The Revealed Comparative Advantage represents the "relevance" for a country c in a product p. A country has a Revealed Comparative Advantage ($RCA_{cp} = 1$) if its share of that product is bigger than the share of the total world trade of the same product. For instance, *"in 2008, with exports of $42 billion, soybeans represented 0.35% of world trade. Of this total, Brazil exported nearly $11 billion, and since Brazils total exports for that year were $140 billion, soybeans accounted for 7.8% of Brazils exports. This represents around 21 times Brazils "fair share" of soybean exports (7.8% divided by 0.35%), so we can say that Brazil has Revealed Comparative Advantage in soybeans"* (see Hausmann, ed, 2013, p25).

Formally, having X_{cp} represents the exports of a country c of a good p in dollars, we write the Revealed Comparative Advantage (RCA) as:

$$RCA_{cp} = \frac{X_{cp}}{\sum_{p'} X_{cp'}} \Bigg/ \frac{\sum_{c'} X_{c'p}}{\sum_{c'p'} X_{c'p'}}. \tag{3.6}$$

With this measure we are able to include only those exports which are representative for each country. We then can define a binary version of the RCA matrix:

$$M_{cp} = 1, \quad \text{if } RCA_{cp} \geq 1, \text{ otherwise}$$
$$M_{cp} = 0.$$

Using the analogy for exporting goods, we can translate each concept in the Patent production. X_{cp} represents the number of Patent classes products p produced by a country c. M_{cp} is binary version of a matrix RCA, which represents whether a country c produce a bigger share of Patent class products p rather than the share of total world on that Patent class. The Diversity and Ubiquity vectors represent the ability of a country has to develop technologies, and how much ability is required to create a certain technology.

We also estimate a normalized version of the Diversity and Ubiquity through the Economical Complexity Index (ECI) and Product Complexity Index (PCI):

$$ECI = \frac{\mathbf{K} - < \mathbf{K} >}{\text{stdev}(\mathbf{K})}, \text{and} \tag{3.7}$$

$$PCI = \frac{\mathbf{Q} - < \mathbf{Q} >}{\text{stdev}(\mathbf{Q})}, \tag{3.8}$$

where $\mathbf{K} = k_{c,N}$ is the country vector after N iterations, $\mathbf{Q} = k_{p,N}$ is the product vector after N iterations, $<>$ is the average and stdev() is the standard deviation.

3.2.3 Hierarchical Clustering and Dendrograms

Hierarchical Clustering Analysis (HCA) is a statistical method that defines a hierarchical structure of clusters. It ranks cluster pairs based on a dissimilarity measure. The dissimilarity is a measure which is opposite to correlation. Lower values of dissimilarities imply high values of correlation. Among the different methods of HCA, we use the *complete linkage method*. It links mostly similar clusters among them (Garber et al., 2001).

Formally, for the data matrix $X = \{x_{ij}\}$ where columns represents countries and rows year, we choose the dissimilarity known as the *correlation* method, which is defined by one minus the Person correlation coefficient, which is correlated among two variables with the aim to convert correlations in distances.

$$1 - \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}. \tag{3.9}$$

This is the dissimilarity between j_{th} and k_{th} country or region, where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ and x_{ij} are the elements of a matrix X, which in our dataset columns represent countries and regions, and rows represent the years.

The complete linkage method, the furthest neighbor or the maximum method determine the hierarchical process between clusters. After estimating the correlation matrix among variables, we link the least dissimilar variables. This new link defines a cluster of those two variables linked. Here we must choose which is the new dissimilar measure for the recent defined cluster. The maximum method chooses the maximum value of the dissimilarity measure among the two variables clustered and the others by this its name. There are other similar methods: the minimum linkage and the average linkage. As their names suggest, they take the minimum and average value among the dissimilarity measure of the two variables clustered and others. We repeat this process until it links every variable in our dataset, reporting which dissimilarity value they are clustered with and building links among clusters. Specifically, we use the GNU R-package `pvclust` to estimate and plot the Hierarchical cluster, where the representation is known as *dendrogram*. The R-package estimates two types of p-values: Approximately Unbiased (AU) p-value and Bootstrap Probability (BP) value. AU p-value is calculated using multi-scale bootstrap resampling, while BP value is calculated by the ordinary bootstrap resampling (Suzuki and Shimodaira, 2006).

Our aim is to understand the dynamics of complexity among regions and countries. Using the correlation dissimilarity, we are able to capture synchronous changes among regions and countries. Additionally, through the complete linkage method, we choose the most homogeneous regions. With these choices, we attempt to find those regions and countries which developed technologies at the same time, perhaps by collaboration or because of competition.

The complexity approach, together with the hierarchical cluster analysis over it, is relevant in understanding how the most important centers of development evolved and which are the most *lucrative*, in terms of continuation, Patent classes to research. This methodology may be useful for new centers. For instance, our results answer the question of what to research in order to have a better future, and how I should choose my partners in order to increase production in R&D.

3.3 Results

We analyzed the last Edition of the OECD REGPAT DATABASE (Dernis and Khan, 2004). This database of Patents has been regionalized across OECD countries, (28 EU countries, Brazil, China, India, the Russian Federation and South Africa). Furthermore, it includes detailed information of the Patents, such as application date, Patent Classes and Inventors. These last ones are regionalized at country and regional level. Using Big Data techniques and a Hadoop server[2], we manipulate the three Patent databases (detailed, classes and inventors) in order to relate the application date, country/region and classes for the first Patents of the triadic families.

We use the number of the first patent of every triadic family since 1980 until 2011 to build the country and region diversity $k_{c,0}$ (Equation 3.3) and the product ubiquity $k_{p,0}$ (Equation 3.4). After that, we use the HH algorithm (Equations 3.1 and 3.2) to estimate the Fitness for countries and regions and Complexity for Classification Codes of Patents. Intrinsically, we use the network approach by joining countries (and regions) with Patent classes, where the link is weighted using the M_{cp} matrix. Furthermore, we interpret the results of the HH algorithm using the number of Patent classes instead of the number of products exported, as in the original HH algorithm application.

To explain the results, we divide this section using the two type of tables: Countries and Regions.

[2] "The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures." taking from their webpage `https://hadoop.apache.org`.

3.3.1 Country Table

We rank IPC7 Patent classes by their Product Complexity Index (PCI) (Equation 3.8) for every year since 1980-2011. We report the IPC3 classification in Figure 11 instead of the IPC7 classification for making more legible. The most common 2-level class in the IPC7 and IPC3 of the Top-PCI Patent Classes is the Pharmaceutical, which is also the most common among countries, see Table 18. Also, the IPC7 class that has increased the most among the period of study is the H04L209 (signaling and real-time protocols for multimedia conference).

We also rank countries by their Economic Complexity Index (ECI) (Equation 3.7), for every year from 1980 to 2011. The results are reported in Figure 12. The Republic of Korea is not in Table 16 but it is in Figure 12, because it has been the country that has the most increment of ECI during the period of study. Figure 13 includes snapshots of the complexity rank in 2005 for the world. Additionally, we use the complete linkage method, presented in Figure 14, to estimate the cluster of countries by their ECI. We find two country clusters at a 95% Approximately Unbiased (AU) p-value: cluster 1 is composed by France and Germany, and the cluster 2 is composed by Sweden, Italy and Netherlands.

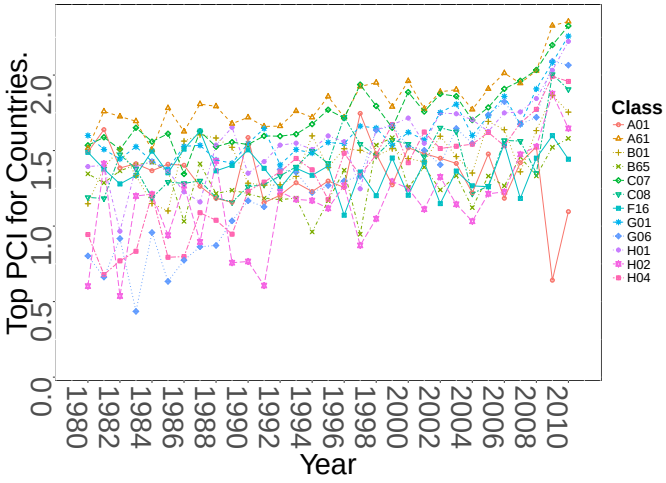


Figure 11: National Product Complexity Index (PCI) for IPC3. Evolution of the Product Complexity Index (PCI) by year of the Top-PCI for countries that produce Triadic Families for every year from 1980 until 2010 using the HH algorithm. A01: agriculture, forestry, animal husbandry, hunting, trapping, fishing; A61: medical or veterinary science, hygiene; B01: physical or chemical processes or apparatus in general; B65: conveying, packing, storing, handling thin or filamentary material; C07:organic chemistry; C08: organic macromolecular compounds; F16: engineering elements and units; G01: measuring and testing; G16: computing, calculating and counting; H01: basic electric elements; H02: generation, conversion or distribution of electric power; H04: electric communication technique.

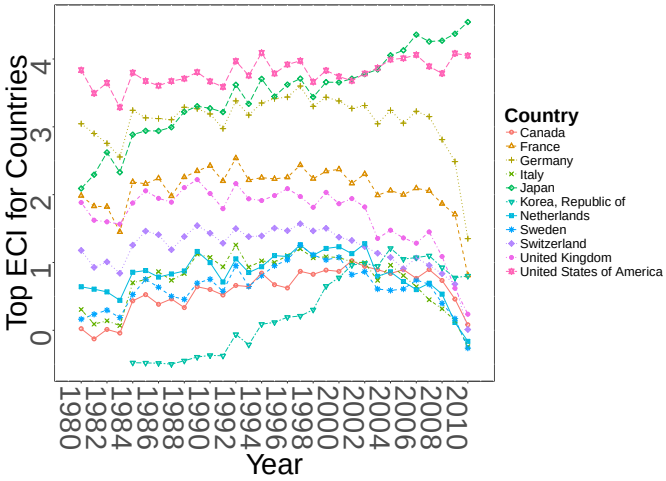


Figure 12: Economic Complexity Index (ECI) for Countries Evolution of the Economic Complexity Index (ECI) by year of the Top-ECI for countries that produces Triadic Families for every year from 1980 until 2010 using the HH algorithm.

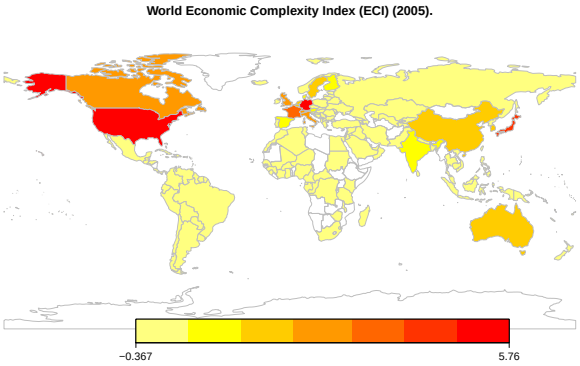


Figure 13: Map of the Economic Complexity Index (ECI) Rank in the World (2005).

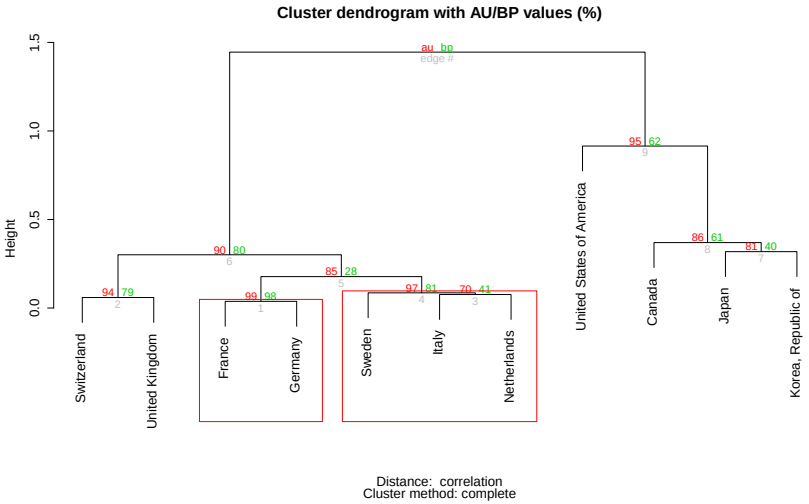


Figure 14: Cluster of Country Fitness during the period 1980 - 2010 using the Complete Linkage Method. Approximately Unbiased (AU) p-value in red and Bootstrap Probability (BP) value in green. Red boxes show 0.95 level of significance of the AU p-value using Bootstrapping with 1000 iterations.

3.3.2 Region Aggregation

We use the same methodology for Regions as for country aggregation. Figure 15 shows the Top-PCI IPC3 Patent classes. Again we present IPC3 instead of IPC7 to make it more legible, also the most common IPC3 Patent 2-level class is Pharmaceutical. Compared with Table 18, we find the same IPC7 Patent classes plus others. At the regional level, the IPC7 Patent classes, that has increased the most during the period 1980-2011, is A61K047 (Medicinal preparations containing active ingredients), and the one that has decreased the most is C07C067 (Preparation of carboxylic acid esters).

Figure 16 shows that the most complex regions, giving the HH algorithm, are the Japanese cities of Tokyo, Osaka and Kanagawa. We also study, within the Top-PCI IPC7 Patent classes, regions from United States, Germany, Switzerland, France and The Republic of Korea. Again, the region that has increased its ECI the most during the period 1980-2011 is in The Republic of Korea (Gyeonggi) which agrees with the findings in the country aggregation. Compared with Table 17, we do not find regions in every Top-ECI country, such as Italy, Netherlands, Sweden, Canada and the United Kingdom. Figure 17 shows the dendrogram of the Hierarchical Cluster Analysis (HCA) using the ECI for Regions. We find that most of the significant clusters are composed for regions of the same country, with the exception of Basel-Landschaft (Switzerland) which shares a cluster with German regions, and Paris, which shares cluster with Japanese regions. Within the United States, the most complex regions are in California with the exception of Middlesex which is in Massachusetts.

Using the highest ECI for the period of 1980-2011, we find that the Top-ECI region is a Japanese region (Tokyo), where the Top-ECI country is the United States. We explain this effect arguing that the innovation in the United States is more diversified among regions compared to Japan. A similar argument is used to explain why Italy, Netherlands, Sweden, Canada and the United Kingdom do not have regions within the Top-ECI Regions. Using this methodology we can build a diversification of inno-

vation index including the relation between country/region ECI. This aims to study how region diversification affects country output. Our results show that the least centralized countries, in terms of region diversification of innovation, are better ranked compared to non-diversified countries. The HCA results suggest that the growth of Region-ECI occurs in regions within the same country. This result agrees with the positive effect on region diversification of innovation. We neglect the size-effect within the ECI for countries. Instead we uses the NUTS3 regions, which are chosen by the thresholds of the average population size (150,000 - 800,000 Millions of people).

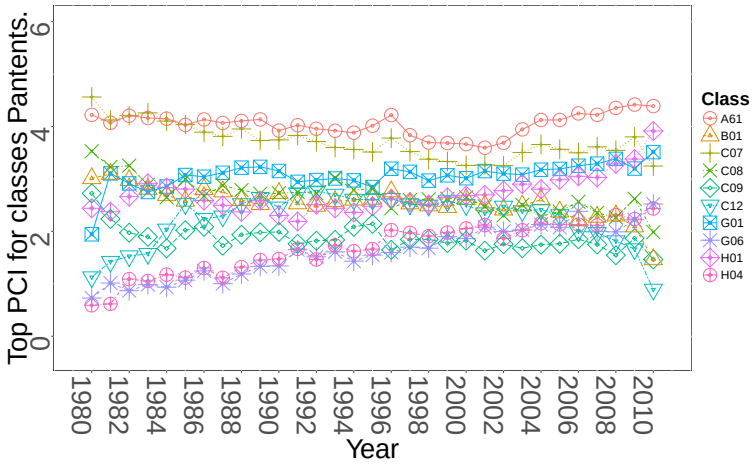


Figure 15: Regional Product Complexity Index (PCI) for IPC3. Evolution of the Product Complexity Index (PCI) by year of the Top-ECI on regions that produces Triadic Families for every year from 1980 until 2010 using the HH algorithm. A61: medical or veterinary science, hygiene; B01: physical or chemical processes or apparatus in general; C07:organic chemistry; C08: organic macromolecular compounds; C09: dyes, paints, polishes, natural resins, adhesives; C12: biochemistry; beer; spirits; wine; vinegar; microbiology; enzymology; mutation or genetic engineering; G01: measuring and testing; G06: computing, calculating and counting; H01: basic electric elements; H04: electric communication technique.

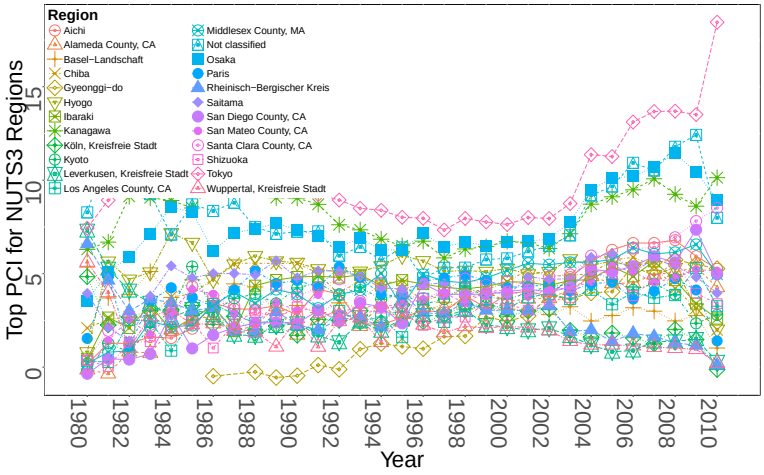


Figure 16: Economic Complexity Index (ECI) for Regions. Evolution of the Economic Complexity Index (ECI) by year of the Top-ECI on regions that produces Triadic Families for every year from 1980 until 2010 using the HH algorithm.

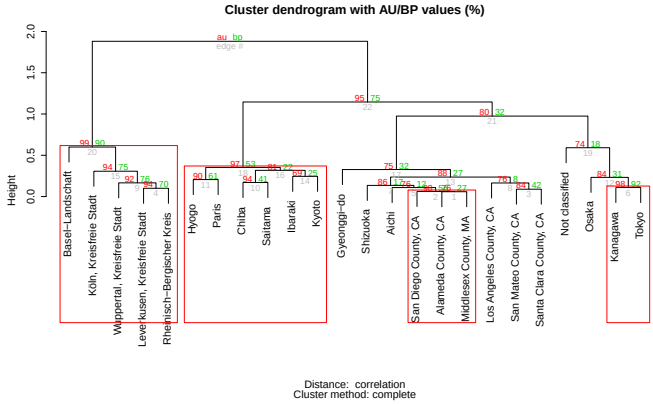


Figure 17: Fitness cluster of the top Regions (cities approx) during the period of 1980 - 2010 using the Complete Linkage Method. Approximately Unbiased (AU) p-value is in red and Bootstrap Probability (BP) value is in green. Red boxes are for the 0.95 level of significance of the AU p-value, using Bootstrapping with 1000 iterations.

3.3.3 National ECI & GDP, PCI & Patents counts: a comparison between aggregation levels.

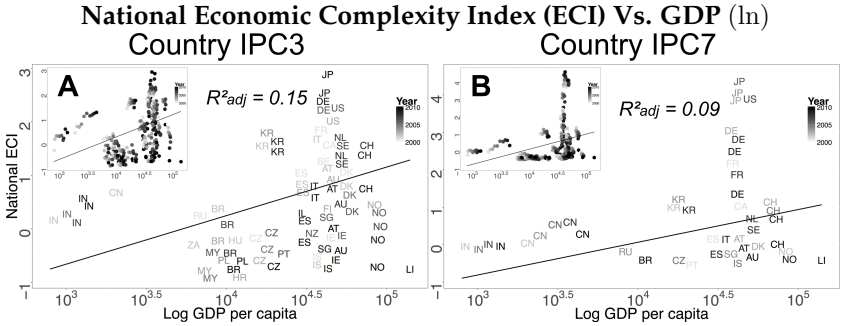


Figure 18: National ECI Vs. (ln) of Gross Domestic Product (GDP) per capita (constant 2010 USD$, source from the World Bank indicators). The plot includes data from 2000 until 2010 aggregated by different Patent Classification. A: Aggregated at Country and IPC3; B: Aggregated at Country and IPC7; IPC refers to the International Class Classification. We plot only the non-overleaped points on the external Figure and every point in the internal. We regress $ECI \sim \ln(gdp) + \epsilon$, (black line) where ECI is the National Economic Complexity Index, gdp is Gross Domestic Product per capita. Adjusted R^2 are reported.

Figure 18 shows the relation between the National Economic Complexity Index aggregated at IPC3 (Fig. 18A) and IPC7 (Fig. 18B), and the Gross Domestic Product per capita. The National ECI calculated with the IPC3 level of aggregation shows an $R^2_{adj} = 0.15$, higher with the National ECI calculated with the IPC7 level of aggregation $R^2_{adj} = 0.09$. This aggregation level (Fig. 18B) shows that the United States, Japan, Germany and France are countries that have a higher National ECI compared with the regression. Here, we observe that the National ECI have an additional information compared with the GDP per capita.

Figure 19 shows the relation between PCI and the number of Patent classes in logs. We aggregate these quantities using different aggregation levels to study how the Product Complexity Index is related with the Patent counts. Here we test how much additional information we can learn with the complexity algorithm, compared with an obvious one,

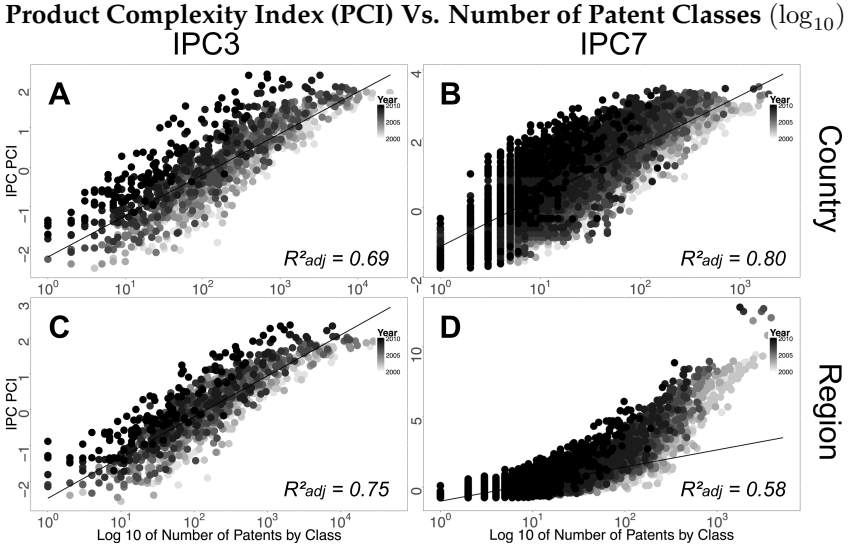


Figure 19: PCI Vs. $(\log_{10})$ of Number of Patent Classes. The plotted data includes different aggregation by geographical zones (countries and regions) and Patent classification (IPC3 and IPC7) from 2000 until 2010. A: Aggregated at Country and IPC3; B: Aggregated at Country and IPC7; C: Aggregated at NUTS3 Region and IPC3; D: Aggregated at NUTS3 Region and IPC7. We regress $PCI \sim \log_{10}(cnts) + \epsilon$, where PCI is the Product Complexity Index, $cnts$ are the counts of the Patents by IPC classification (black line). Adjusted R^2 are reported in the same color of the year.

such as patent counts. To account for the differences given by the aggregation levels, we use the HH algorithm over different aggregated data varying over regions types and classification depth. To quantify these differences, we use a linear regression between the PCI and the number of Patents by class and Year in logarithms ($PCI = \log_{10}(cnts) + \epsilon$). We observe that for a coarser aggregation level (country or IPC3) the complexity measure has almost the same information the logarithm of the counts. Instead for a lower aggregation level, such as region and the number of IPC7, we find a different relation beyond the exponential. Having the fact that the lowest value of the adjusted R^2 is the one of the lowest aggregation level (Fig. 19D), we evidence that most of the infor-

mation obtained from the HH algorithm is higher at lower aggregation levels, which differs the most from an obvious exponential relation with the patent counts.

3.4 Discusion

Despite the fact that the United States and Japan are the leaders in technology innovation measured by Triadic Families, together the members of the European Union also represent, as an economic union, an important player in the development of technology. The presence of the European Union members, and their partnership shown in the Hierarchical Cluster Analysis (HCA), shows the importance of the cross-border collaboration among countries. Bearing in mind the fact that the Republic of Korea is the country that has grown the most in the National ECI during the data period (1980-2011), it is the one with the lowest National ECI within the Top-ECI group of countries.

Using the HCA results, our findings show that the most independent country, in term of the evolution of the National ECI, in respect to the others is the United States; it has ~ 1.0 of the dissimilarity measure, which is equivalent to the ~ 0.0 correlation. However, for the time period, the regions of the United States are significantly clustered with a low dissimilarity distance. The result supposes a close relationship between the regions of the United States (~ 0.1 dissimilarity) and their distant relationship with other countries. We find evidence that regions within countries are more likely to be clustered, and for this, have similar dynamics with regions across borders. Most of those regions belong to Japan, the United States and Germany, which are the Top-ECI countries. In this sense, Regional ECI clusters, evidenced by the HCA, has a positive impact on National ECI.

In the phrase *"economic complexity reflects the amount of knowledge that is embedded in the productive structure of an economy"*, Hausmann, ed (2013) relates the economic complexity with the knowledge embedded in the productive structure. Inspired in this statement, we explore the relation of the economic complexity with the productive structure varying the ag-

gregation level of the product and geographical zones. The result of the comparison between the National ECI with the GDP per capita suggests that the amount of knowledge embedded in a lower aggregation level is higher. The more specific the information about the product is, for instance IPC7 compared with IPC3, the higher the amount of knowledge is in National ECI for countries. Furthermore, this effect is confirmed by the result of the comparison between PCI and patent counts, and that shows at the lowest level the higher difference with the amount of knowledge of the patent counts.

The interpretation of the economic complexity presented in the original work (see Hausmann, ed, 2013, p27) cannot be used directly in this work. Here, we use knowledge production with the nationality of the authors, which are not necessary the nationality of the technology, therefore saying that we can use these nationalities as a measure for knowledge production for countries would be an assumption. Instead, the nationality taken from the patent database, as we use it, the country of the address reported by the inventor, is a proxy for the development of the country institutions, which we believe that are chosen generally in function of their facilities to develop the inventors activities and the capacity to generate the conditions favorables for inventors and their activities. Through institutions and social conditions, the higher complexity countries and regions attract more inventors, which generates more knowledge.

3.5 Conclusions

We use the first Patent of a triadic family and the Hausmann and Hidalgo (HH) complexity algorithm to explore knowledge production in the competitiveness of nations and regions. We also use a Hierarchical Cluster Algorithm (HCA) to study the similarities of the economic complexity dynamics on regions and countries. Additionally, we explore different aggregation levels, and their role on the interpretation of the knowledge production in countries and regions.

Our findings show certain clusters of regions and countries which

are clusterised through their similarities in Regional and National ECI. Our results suggest that regional clusters are related to the output at a national level. For a subseqquent work, we suggest studying the effect of the regional decentralization into National ECI, and how the cross-border collaboration impacts National ECI.

Using the view of economic complexity, based on knowledge creation as a measure of economic development through favorables institutions for inventors, our findings agree with Hausmann's view of economic complexity, which suggests that the greater to economic complexity, the faster the growth. The knowledge embedded in the economic complexity calculated with patent data is proportional to the institutional development and consequently for growth (Acemoglu and Robinson, 2012). Here, the National and Regional ECI are related to growth through the institutional design. Countries and regions with a higher National and Regional ECI attract inventors of the highest complex patent classes through their institutions. Furthermore, analysing the economic complexity for different aggregation levels, we find that most of the addtional information, compared with GDP per capita and patent counts, is presented at lower aggregation levels of the National ECI and National and Regional PCI.

To conclude, we compared our results with the findings of previous works. Specifically, Cimini et al. (2014) ranks academic topics as products (1996 - 2012) according to its PCI: 1) Biochemistry, Genetics, Molecular Biology; 2) Neurosciences; 3) Pharmacology, Toxicology, Pharmaceutics Earth & Planetary Sciences 4) Agricultural & Biological Sciences; and 5) Environmental Sciences. We also compared the results with Hausmann, ed (2013) in exported goods (2005): 1) Machines and mechanical appliances having individual functions; 2) Equipment for photographic laboratories; 3)Acrylic polymers in primary forms; 4) Chemical preparations for photographic uses; and 5) Tool plates/tips/etc, sintered metal carbide & cermet. The HH for 2010 included: 1) Glass, drawn or blown; 2) Photographic plates and film, exposed and developed, not motion-picture film; 3) Nickel tubes, pipes and tube or pipe fittings; 4) Apparatus based on the use of X-rays or of alpha, beta or gamma radiations; 5)

Fork-lift trucks. Although we used the HH algorithm as in Hausmann, ed (2013), our results were closer to the Cimini et al. (2014), where the authors uses the FC algorithm. Because of this, and assuming that the intentions of the two measure are similar, they differed in the metric. Our results were closer to the academic complexity than to the industrial development. This was confirmed in the case of the United States, where the two most innovative cities were those where the most important universities were located: Los Angeles and Massachusetts.

Appendix A

Measuring the impact of European integration on the rate of cross-border collaboration and high-skilled labor mobility

A.0.1 Estimating the negative impact of joining the EU using the Synthetic Control Method

In order to explain the divergence in cross-border collaboration between Western and Eastern Europe, we use the 2004 EU enlargement as a policy shift experiment characterized by a large subset of 10 countries with coinciding "policy intervention" (treatment) year $t^* = 2004$.[b] A naive assumption might be that the 2004 entrants would produce more cross-border publications ($Y^s_{i,t}$) after entry into the EU because of increased access to EU framework programme funding and collaborative opportunities facilitated by the "integrated" EU R&D system. However,

[2]Bulgaria and Romania also serve as a second policy shift experiment with coinciding "treatment" year $t^* = 2007$.

we find the contrary to be true, that new entrants *would have produced more publications* – both in frequency $f^s_{i,t}$ per publication and number $Y^s_{i,t}$ – had they not entered the EU. This result provides a partial explanation for why the EU cross-border collaboration rate grew no faster than international rates during this period, representing a "stagnation" of the EU integration process Chessa et al. (2013).

We demonstrate this counterintuitive outcome on cross-border collaboration within the EU using the Synthetic Control Method (SCM) Abadie et al. (2010); Abadie and Gardeazabal (2003). This method estimates the effect of the counterfactual outcome – that each EU entrant country *had not participated in the EU enlargement* – on our two measures of cross-border integration: the fraction $\hat{f}^s_{i,t}$ and total number $\hat{Y}^s_{i,t}$ of cross-border publications. We used a control group of $N_c = 26$ non-EU countries $\{j\}$ ={AR, AM, AZ, BY, CA, CN, CO, CU, IN, IL, JP, KZ, KW, KG, MG, MX, MN, PA, RU, RS, SG, KR, TT, TR, UA, US} to estimate the counterfactual cross-border trends $\hat{f}^s_{i,t}$ and $\hat{Y}^s_{i,t}$ for $t \geq 2004$. Thus, the difference δ between the synthetic outcome and the real outcome corresponding to the "EU Enlargement Effect". Because none of the control group countries belong to the EU, the implicit assumption of no interference between units is satisfied –i.e. enlargement of the EU should not be signifcantly correlated to international collaboration rates in Japan, for example.

The SCM produces an optimal representation of the actual time series of interest, $Z_{i,t}$ ($= \log_{10} Y^s_{i,t}$ or $f^s_{i,t}$)[c], based upon best-fit weights calculated using the control country data for the time period before the "EU treatment" ($t < 2004$). The covariate data ($X_{i,t}$) we used to model $Z_{i,t}$ are the total number of publications ($\log_{10} D^s_{i,t}$), the normalized citations ($R^s_{i,t}$), the per-capita GDP ($\log_{10} GDPpc_{i,t}$), and government expenditure on R&D as % of GDP, $e_{i,t}$.[d] The factor model representation of the de-

[3]For the total number of cross-border documents, we estimated the model using $\log_{10} Y^s_{i,t}$ which is less sensitive to large deviations in scale across the control countries as well as the EU countries. We then exponentiated the SCM results in order to estimate the difference $\delta(\%)$ and plot the results in Figs. 3

[4]Because the World Bank data for researcher population data is incomplete for many of the control countries, we were unable to include it without severely reducing the number

pendent variable is given by

$$Z_{i,t} = \gamma_t + \theta_t X_i + \lambda_t \mu_i + \epsilon_{it} \ , \tag{A.1}$$

where γ_t represents global factors affecting all countries equally, θ_t is a vector representing the covariate effects associated with the vector of observed covariates X_i, λ_t generalizes the model to include a vector of unobserved common factors and their loadings μ_i, and ϵ_{it} is the country-specific error term. We shorthand the SCM algorithmic procedure using the representation of a multi-dimensional projection of $Z_{i,t}$ onto the complementary vector space of control time series given by $Z_{j,t}$. In this way, the normalized weights can be conceptualized as

$$w_j = \frac{\langle Z_{i,t}, X_{i,t} | Z_{j,t}, X_{j,t} \rangle}{\sum_j \langle Z_{i,t}, X_{i,t} | Z_{j,t}, X_{j,t} \rangle} \in [0, 1] \ , \tag{A.2}$$

which satisfy $\sum_{j=1}^{N_c} w_j = 1$. The SCM algorithm then finds the optimal weight vector $\mathbf{w}^*$ that sufficiently satisfies the following equalities

$$\begin{aligned} \sum_{j=1}^{N_c} w_j^* Z_{j,t} &= Z_{i,t} \ \text{, for } \ t \in [1996, 2003] \ , \\ \sum_{j=1}^{N_c} w_j^* X_j &= X_i \ . \end{aligned} \tag{A.3}$$

This method is reliable as long as the number of number of periods prior to 2004 (i.e. 7 years in our case) is large with respect to the timescale of ϵ_{it}. For the longhand description and derivation of the SCM, with application to the 1988 California tobacco control program (Proposition 99) in the USA, we refer the interested reader to Abadie, Diamond, and Hainmueller Abadie et al. (2010).

Using the optimal weighted coefficients w^* which best reproduce the actual $Z_{i,t}$ for $t < 2004$, the weighted linear combination is extrapolated for $t \geq 2004$, thereby producing the counterfactual time series $\hat{Z}_{i,t}$. This method is well-suited for this policy intervention scenario because it ac-

of control countries (N_c).

counts for the global trends in cross-border collaboration already existing before and persisting after 2004, as captured by γ_t (implicit in the non-EU global control set).

We now return to the two scenarios of interest, first where the outcome variable is the fraction of publications that involved cross-border collaboration, $Z_{i,t} \equiv f^s_{i,t}$, and in the second case where the outcome variable is total number of cross-border publications $Z_{i,t} \equiv Y^s_{i,t}$. In both cases we measure the "EU enlargement effect" by computing the difference in the post-2004 totals, $Z^>_i = \sum_{t=2005}^{2012} Z_{i,t}$ and $\hat{Z}^>_i = \sum_{t=2005}^{2012} \hat{Z}_{i,t}$. In the case of $f^s_{i,t}$ we define the post-treatment difference as a difference in means, $\delta = (\hat{f}^> - f^>)/(2012 - 2005 + 1)$, and in the case of $Y^s_{i,t}$ we define the post-treatment difference as a percent difference, $\delta(\%) = 100 \times (\hat{Y}^> - Y^>)/Y^>$.

For the case of $f^s_{i,t}$, we observe opposite effects for the new and incumbent EU countries. Figure 3 shows $\delta > 0$ values for the 2004 entrant EU countries and $\delta < 0$ values for the incumbent EU countries. This pattern is robust for three different estimations: for all subject areas aggregated (s =All), as well as for the individual subject areas $s = 1300$ representing "Biochemistry, Genetics, and Molecular Biology" (Biology), and $s = 3100$ representing "Physics and Astronomy" (Physics), the two most collaborative subject areas. The diverging trends provide a key insight into the substitution effect due to high-skilled mobility: had there been no enlargement, the counterfactual number of intra-border publications ($D^s_{i,t} - Y^s_{i,t}$) would have decreased relative to $Y^s_{i,t}$ for the incumbent EU countries because there would have been more researchers to potentially collaborate with abroad. However, since the net flow of high-skilled mobility was towards the pre-2004 EU countries – contributing to their stock of internationally reputable scientists along with their international connections – this left the new 2004 EU entrant countries at a loss of international collaboration opportunities.

The case of $Y^s_{i,t}$ further demonstrates negative effect on the intensity of Europe's science integration, as measured by cross-border collaboration. For both incumbent and new EU countries, there would have been more cross-border publications had there been no EU enlargement.

For example, for all subject areas aggregated (s =All), we calculated a $\delta(\%) = 15$ counterfactual effect for the average incumbent EU country, and a $\delta(\%) = 9$ percent effect for the average entrant country (see Fig. 3). These results were also consistent when applying the method to just the Biology and Physics subject area data.

Appendix B

Deprivation and Trade. Evidence from Colombia.

B.1 Supplementary Materials

B.1.1 Robustness check

Table 19: Instrumental Variables Tests. F-Statistics (P-values) Reported.

Test	$y = \log(\sum_{ij} CMPI_H)$		$y = \log(\sum_{ij} g^1_{ij})$		$y = \log(\sum_{ij} g^2_{ij})$	
	Exp. (D)	$\log \sum$ Exp.	Exp. (D)	$\log \sum$ Exp.	Exp. (D)	$\log \sum$ Exp.
Weak Insttruments	<2e-16***	<2e-16***	<2e-16***	<2e-16***	<2e-16***	<2e-16***

Note: *p<0.05; **p<0.01; ***p<0.001

Exp. (D) for Export Dummy and $\log \sum$ *Exp. for the logarithm of the total exports (2002-2004)*

Table 20: Lagrange Multiplier diagnostics for spatial dependence in linear models. F-Statistics (P-values) Reported.

Test	$y = \log(\sum_{ij} CMPI_H)$		$y = \log(\sum_{ij} g^1_{ij})$		$y = \log(\sum_{ij} g^2_{ij})$	
	Exp. (D)	$\log \sum$ Exp.	Exp. (D)	$\log \sum$ Exp.	Exp. (D)	$\log \sum$ Exp.
LMerr	7.883e-15***	1.121e-14***	6.071e-12***	6.130e-13***	1.228e-10***	1.263e-11***
LMlag	<2.2e-16***	<2.2e-16***	1.010e-14***	6.661e-16***	2.128e-13***	2.298e-14***
RLMerr	0.05357	0.04537*	0.1061	0.0003741***	0.05796	0.0003323***
RLMlag	9.229e-06***	7.373e-06***	9.724e-05***	3.268e-07***	6.146e-05***	4.995e-07***
SARMA	<2.2e-16 ***	<2.2e-16***	2.687e-14***	<2.2e-16***	3.303e-13***	3.331e-16***

Note: *p<0.05; **p<0.01; ***p<0.001

Exp. (D) for Export Dummy and $\log \sum$ *Exp. for the logarithm of the total exports (2002-2004)*

To test whether the SDM is the specification that fits better the data,

we must run the Likelihood Ratio Test. The results of the Table 20 suggests that for each model the error term is spatially correlated with the dependent variable. As result, the SDM has a bigger likelihood compared with the Spatial Error Model.

Table 21: Likelihood Ratio Test. Likelihood Ratio (LR) and P-values (p-val) reported.

Test	$y = \log(\sum_{ij} CMPI_H)$		$y = \log(\sum_{ij} g^1_{ij})$		$y = \log(\sum_{ij} g^2_{ij})$	
	LR	p-val	LR	p-val	LR	p-val
Durbin/Error	129.2911	2.366e-07***	117.8482	5.786e-06***	117.7423	5.953e-06

Note: (df=58). *p<0.05; **p<0.01; ***p<0.001

Exp. (D) for Export Dummy and $\log \sum$ *Exp. for the logarithm of the total exports (2002-2004).*

B.1.2 Plots

B.1.3 Spatial Facts

The poverty gap within Colombia has a very strong spatial component in which the poorest municipalities are grouped in specific geographic zones. In order to consider this effect, it is necessary to use empirical techniques that includes spatial correlations. In contrast, non spatial econometrics models assume random variability among individuals, which contradicts the empirical founds.

There is empirical evidence that shows a clustered distribution within social dimensions as poverty, income per capita and others, which is a consequence of similar endowments, weather, soil and resources shared among them, which justify the spatial econometrics approach as is shown in Figure 7.

In order to see the spatial effect Figure 7 shows how would be a random pattern within the colombian municipalities, the important issue here is to account how much is the un-randomness within the variables of colombian territory. This measure is given by Moran's Index.

The Moran's I coefficient (Moran, 1950b) gives a zero value for those

variables distribution which do not have an spatial correlation and one for those with high spatial correlation.

Measuring Moran's Index also implies a spatial relationship between individuals which in this case are municipalities. The relationship between municipalities is given by the spatial correlation matrix. Within the present work, three types of spatial relationship had been taken into consideration, the first is the contiguity matrix $\mathbf{C}$ with a *Queen* boundary which is a squared matrix $(N \times N)$, where N is the number of municipalities, each matrix element is equal to one $(C_{ij} = 1)$ for those municipalities which share a common edge, zero for the diagonal $(C_{ii} = 0 \; \forall i \in N)$ and for all those municipalities which do not share a common boundary. The *Queen* type of spatial matrix relates every other adjacent municipality in every direction, as the moves of a *queen* within a chess game. Figure 21 shows how the queen neighborhood are spread within the territory (Figure 21a), furthermore Figure 21b shows and amplification in how these relations behaves within the county *Cundinamarca*, which includes the largest and the country's capital city.

Queen's neighborhood relation includes every single adjacent municipality. As some municipalities are isolated, which leads to an isolation also in the dependent and explanatory variables. To measure the effect of further neighbors and to guaranty that every municipality has neighbors, a K-neighbor correlation matrix is included in the present empirical exercise. The k-neighbor matrix used in the present work, is a four neighbor matrix (K_4), this relation connects every municipality with its 4-nearest neighbors measured from center to center. Figure 22 shows how the K_4 neighbor matrix relates every municipality defining four edges for every municipality for the whole territory as in Figure 22a and for the *Cundinamarca* county (Figure 22b).

The K-neighbor spatial correlation matrix must connect every municipality k-times, this may be used as constraint where may connect municipalities without direct relation, overall in those located at the extreme of

the territory. To roughly combine the both mentioned spatial relationship it is used the Gabriel Neighbor spatial matrix correlation (see subsection 2.3.1).

The three kinds of spatial correlation matrix give, each one, different information about neighbor relation. The contiguity Queen neighbor spatial relation increases the effect of the variables spread by adjacent neighborhood and reduces the effect over the isolated municipalities. The K-neighbor spatial correlation homogenize the spatial correlation effect increasing the effect of those variables that equally spread among to the closest neighbors. The Gabriel spatial correlation instead, has both components, increase the effect of the variables spread among adjacent neighbors and connects every single municipality within the territory proportionally to its *neighbor capacity* giving more neighbors to central municipalities and less to those locates in the frontiers.

To test whether is valid or not to use the spatial econometrics approach, it is necessary to estimate the Moran's Index. Table 22 shows the Moran's Index for each variable within the econometrics exercise. Variables as Afro-descendent population, Military operations and the number of attacks from guerrilla groups have bigger value for the contiguity spatial correlation matrix. Slavery in colonial's times was spread it by foot, the slaves were transported from the cost to the gold mines and working zones (Acemoglu et al., 2012), moving easier across contiguos municipalities. Military attacks and guerrilla operations are also spread by troops which travels on foot giving the same result. In contrast, social variables as IMP Headcount, the number of Agricultural and Commercial Units, the rural, urban and total population have bigger K_4 Moran's Index coefficient. These variables spread equal and symmetrically in every direction to the closest municipalities. The Gabriel Moran's Index coefficients instead, are bigger for more complex variables are HDI, HDI Adjusted, Income per Capita and others. These three spatial correlation matrices give a different perspective of the spatial lag, such differences reveal information about the how social, demographic and political in-

formation are spread among the territory.

Table 22: Moran's I coefficient

Variable	Queen	K_4	Gabriel
IMP Headcount	0.042**	0.045**	0.037*
IMP M_0	0.639***	0.653***	0.65***
IMP M_1	0.678***	0.691***	0.692***
IMP M_2	0.665***	0.675***	0.677***
MI Norm.	0.029***	0.013*	0.013*
Rurality Index	0.633***	0.678***	0.685***
Indigenous Pop.	0.522***	0.554***	0.495***
Afro-descendant Pop.	0.441***	0.427***	0.109***
Illiteracy Rate	0.588***	0.592***	0.595***
Log of Urban Pop.	0.354***	0.376***	0.364***
Log of Rural Pop.	0.347***	0.391***	0.374***
Log of Total Pop.	0.339***	0.359***	0.347***
Log of Density.	0.571***	0.622***	0.628***
Log of Agricultural Units	0.378***	0.406***	0.386***
Log of Comercial Units	0.245***	0.259***	0.252***
Log of Income per capita	0.375***	0.393***	0.399***
ELN attc.	0.227***	0.138***	0.215***
Military Operations.	0.439***	0.294***	0.392***
FARC att.	0.242***	0.166***	0.196***
Log of SGP(2005)	0.018	0.023	0.036*
Log of SGP(2009)	0.032*	0.03	0.031
System of cities	0.343***	0.413***	0.383***
Environmental Vulnerability	0.331***	0.326***	0.335***
Demographic Vulnerability	-0.005	0.03**	0.009
Human Capital Vulnerability	0.004	0.023	0.041*
Violence Vulnerability	0.009	0.015	0.038*
Institutional Vulnerability	0.006	0.041*	0.014
Economic Vulnerability	0.016	0.013	0.015
HDI	0.48***	0.487***	0.492***
Adjusted HDI.	0.447***	0.445***	0.459***

Legend: $*p < 0.05; **p < 0.01; ***p < 0.001$

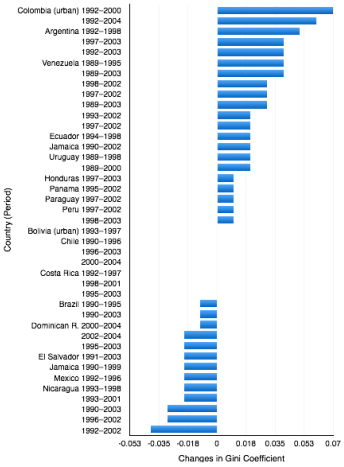
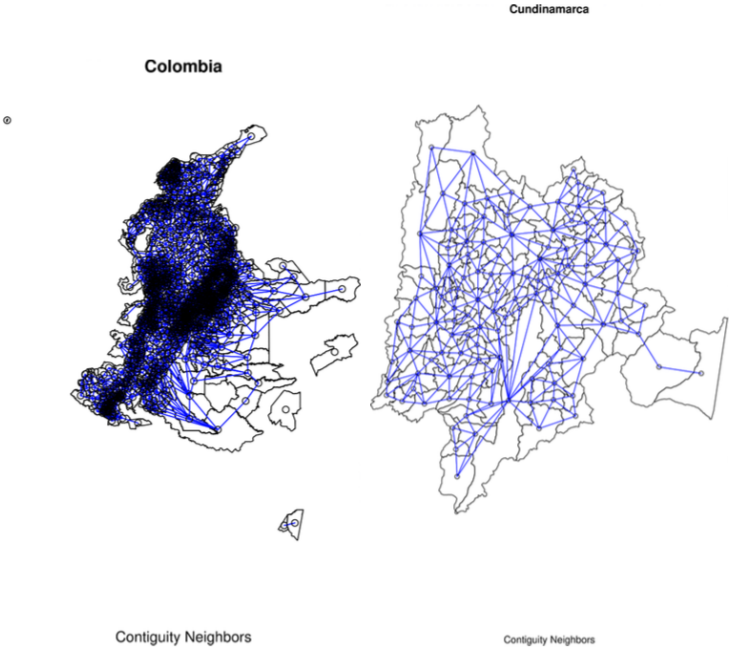
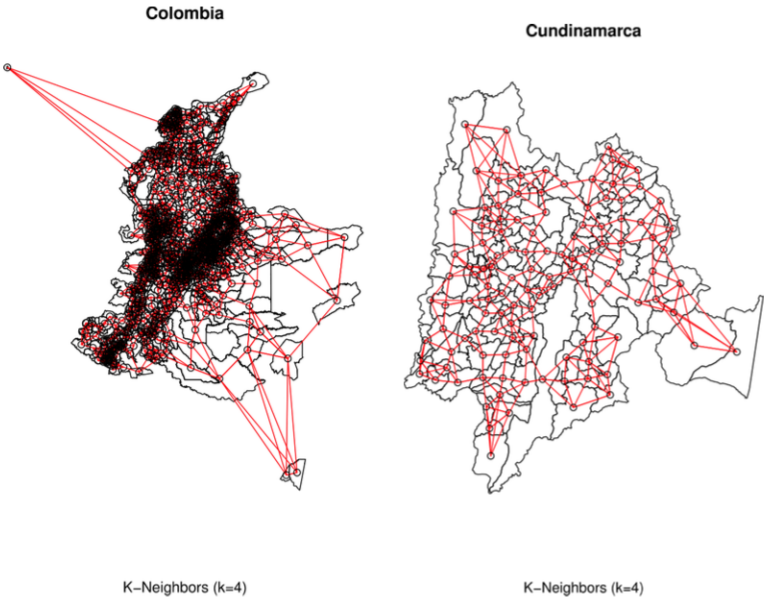


Figure 20: Latin America: Changes in Gini coefficients (%). Distribution of household per capita income. Source: Gasparini et al. (2007).



(a) Contiguity Queen Matrix over colombian territory. **(b)** Contiguity Queen Matrix over Cundinamarca.

Figure 21: Queen Correlation Matrix. Contiguity Queen Matrix Definition.



(a) Contiguity Queen Matrix over colombian territory.

(b) Contiguity Queen Matrix over Cundinamarca.

Figure 22: K-Neighbor Correlation Matrix Contiguity Queen Matrix Definition.

Table 23: Impacts Estimate of the 2SLS Spatial Durbin Model using the Logarithm of the average of deprivations as dependent variable.

$y = G$						
(Poverty Gap)	Direct	Indirect	Total	Direct	Indirect	Total
Incidence	−0.062**	0.081	0.019	−0.061**	0.083	0.022
	(0.020)	(0.054)	(0.060)	(0.020)	(0.054)	(0.059)
Exports (dummy)	4.268***	6.817**	11.085***			
	(0.810)	(2.633)	(3.008)			
Exports (average)				0.919***	1.468*	2.387***
				(0.176)	(0.575)	(0.648)
L.Agric.Units	−3.271***	−1.799	−5.070*	−3.295***	−1.837	−5.133**
	(0.654)	(1.857)	(2.063)	(0.651)	(1.819)	(2.029)
L.Inc.pc	−4.257***	1.996	−2.261	−4.390***	1.784	−2.606
	(1.204)	(4.116)	(4.553)	(1.226)	(3.943)	(4.374)
Eln.Att.	1.189	6.607	7.796	1.151	6.547	7.698
	(1.662)	(5.746)	(6.263)	(1.682)	(5.731)	(6.280)
Farc.Att.	0.273	1.479	1.752	0.256	1.452	1.708
	(0.274)	(0.848)	(0.969)	(0.277)	(0.894)	(1.018)
Milit.Oper.	−0.033	−0.414	−0.447	−0.038	−0.421	−0.459
	(0.109)	(0.361)	(0.412)	(0.109)	(0.355)	(0.396)
L.SGP	−0.016	−1.083	−1.100	−0.034	−1.112	−1.146
	(0.214)	(0.735)	(0.860)	(0.213)	(0.781)	(0.907)
Demogra.V.	0.001	0.053	0.054	0.001	0.053	0.054
	(0.010)	(0.035)	(0.041)	(0.010)	(0.036)	(0.042)
HumanCapital.V.	−0.016	−0.031	−0.047	−0.016	−0.030	−0.046
	(0.010)	(0.036)	(0.041)	(0.010)	(0.036)	(0.041)
Violence.V.	−0.011	−0.013	−0.024	−0.013	−0.015	−0.027
	(0.010)	(0.036)	(0.041)	(0.010)	(0.033)	(0.038)
Institucional.V.	−0.010	0.006	−0.004	−0.010	0.006	−0.004
	(0.011)	(0.038)	(0.044)	(0.011)	(0.038)	(0.045)
Economic.V.	0.001	0.072*	0.073	0.002	0.072*	0.073
	(0.010)	(0.036)	(0.041)	(0.011)	(0.037)	(0.043)
Antioquia	18.843	−36.260	−17.417	18.432	−36.916	−18.484
	(14.760)	(19.027)	(11.652)	(14.831)	(19.412)	(11.466)
Arauca	14.573	−9.035	5.538	14.594	−9.002	5.592
	(15.034)	(22.047)	(15.267)	(15.031)	(22.046)	(14.958)
Atlantico	18.373	−36.377	−18.004	17.864	−37.190	−19.326
	(15.465)	(19.522)	(11.939)	(15.390)	(19.789)	(11.639)
Bogota D.C.	37.491*	5.443	42.934	36.473*	3.817	40.290
	(16.187)	(31.991)	(31.888)	(15.942)	(33.472)	(32.721)
Caldas	22.627	−42.328*	−19.701	22.250	−42.931*	−20.681
	(14.738)	(19.249)	(12.067)	(14.651)	(19.370)	(11.871)
Cauca	26.222	−42.382*	−16.160	25.741	−43.151*	−17.410
	(14.848)	(19.331)	(11.868)	(14.602)	(19.343)	(11.749)
Choco	24.311	−36.417	−12.106	23.877	−37.110	−13.233
	(14.801)	(19.148)	(11.818)	(14.859)	(19.571)	(11.615)
Cordoba	21.943	−32.733	−10.790	21.399	−33.602	−12.203
	(15.210)	(19.480)	(11.947)	(14.863)	(19.526)	(11.625)
Guajira	33.466*	−29.891	3.575	33.024*	−30.596	2.427
	(15.766)	(19.501)	(11.845)	(15.748)	(20.092)	(11.696)
Huila	20.733	−37.134*	−16.401	20.324	−37.788*	−17.464
	(14.585)	(18.980)	(11.836)	(14.634)	(19.311)	(11.695)
Nariño	33.293*	−38.407*	−5.114	32.904*	−39.027*	−6.123
	(15.200)	(19.368)	(11.689)	(15.045)	(19.553)	(11.530)
Norte de Santander	22.173	−33.740	−11.568	21.783	−34.363	−12.580
	(14.896)	(18.888)	(11.778)	(14.946)	(19.472)	(11.662)
Putumayo	33.221*	−42.938*	−9.717	32.844*	−43.540*	−10.696
	(15.039)	(19.877)	(12.503)	(14.985)	(20.036)	(12.078)
Quindio	17.836	−39.473*	−21.637	17.525	−39.970*	−22.445
	(14.800)	(19.373)	(12.407)	(14.846)	(19.399)	(11.768)
Risaralda	21.589	−39.596*	−18.007	21.076	−40.416*	−19.340
	(14.843)	(19.731)	(12.651)	(14.772)	(19.942)	(12.392)
San Andres	7.162	−54.003*	−46.842**	6.810	−54.566*	−47.756**
	(15.590)	(23.761)	(18.083)	(15.857)	(25.017)	(18.126)
Observations	1,086	1,086	1,086	1,086	1,086	1,086
R^2	0.368	0.368	0.368	0.368	0.368	0.368
Adjusted R^2	0.341	0.341	0.341	0.341	0.341	0.341
Residual Std. Error (df = 1040)	6.795	6.795	6.795	6.795	6.795	6.795
F Statistic (df = 45; 1040)	13.460***	13.460***	13.460***	13.460***	13.460***	13.460***

Note: "L." for logarithm and V. for Vulnerability. *p<0.05; **p<0.01; ***p<0.001

Appendix C

Innovation competitiveness of Nations and Regions: A view from Patent Innovation

C.1 Data

C.1.1 Patent data

We analyzed the last Edition of the OECD REGPAT DATABASE (July 2014). This dataset of Patents has been regionalized across OECD countries, EU 28 countries, Brazil, China, India, the Russian Federation and South Africa. Furthermore, it includes detailed information of the Patents as application date, Patent Classes and Inventors, these last regionalized at NUTS3 level.

Countries used as controls: Argentina, Australia, Bulgaria, Belarus, Brazil, Canada, Switzerland, Chile, China, Colombia, Egypt Arab Rep., Hong Kong SAR, China, Croatia, Indonesia, India, Iran Islamic Rep., Israel, Jordan, Japan, Korea, Rep., Kuwait, Lebanon, Morocco, Mexico,

Malaysia, Nigeria, New Zealand, Pakistan, Romania, Russian Federation, Saudi Arabia, Singapore, Thailand, Turkey, Ukraine, United States, Venezuela, RB, South Africa.

References

a. P. Moran, P., "Notes on continuous stochastic phenomena," *Biometrika*, June 1950, *37* (1-2), 17–23.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program," *J. Amer. Stat. Assoc.*, 2010, *105* (490), 493–505. xii, 11, 13, 94, 95

__ **and Javier Gardeazabal**, "The Economic Costs of Conflict: A Case Study of the Basque Country," *Amer. Econ. Rev.*, 2003, *93* (1), 113–132. 94

ABBott, Alison, "Science fortunes of Balkan neighbours diverge," *Nature*, 2011, *469*, 142–143.

Abbott, Alison and Quirin Schiermeier, "After the Berlin Wall: Central Europe up close," *Nature*, November 2014, *515* (7525), 22–25. 22

Abel, Guy J. and Nikola Sander, "Quantifying Global International Migration Flows," *Science*, 2014, *343* (6178), 1520–1522. 2, 10

Acemoglu, Daron and James A. Robinson, *Why Nations Fail: The Origins of Power, Prosperity and Poverty*, 1st ed., New York: Crown, 2012. 91

__ **, Camilo Garca-Jimeno, and James A. Robinson**, "Finding Eldorado: Slavery and long-run development in Colombia," *Journal of Comparative Economics*, 2012, *40* (4), 534 – 564. Slavery, Colonialism and Institutions Around the World. 49, 101

__ **, Mara Anglica Bautista, Pablo Querubin, and James A Robinson**, "Economic and political inequality in development: the case of Cundinamarca, Colombia," Technical Report, National Bureau of Economic Research 2007.

Ackers, L. and B. Gill, *Moving People and Knowledge Scientific Mobility in an Enlarging European Union.*, Cheltenham, UK.: Edwar Elgar, 2008. 2, 21

Ackers, Louise, "Moving People and Knowledge: Scientific Mobility in the European Union," *Int. Migr.*, 2005, *43* (5), 99–131. 3, 7, 23

Adams, Jonathan, "Collaborations: The fourth age of research," *Nature*, May 2013, *497* (7451), 557–560.

Agrawal, A., D. Kapur, J. McHale, and A. Oettl, "Brain drain or brain bank? The impact of skilled emigration on poor-country innovation," *J. Urban Econ.*, 2011, *69*, 43–55. 23

_ , I. Cockburn, and J. McHale, "Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships," *J. Econ. Geog.*, 2006, *6*, 571–591. 23

Alkire, Sabina and James Foster, "Counting and multidimensional poverty measurement," *Journal of Public Economics*, 2011, *95* (78), 476 – 487. 28, 38

_ and _ , "Understandings and misunderstandings of multidimensional poverty measurement," *The Journal of Economic Inequality*, June 2011, *9* (2), 289–314. 37, 38, 39

Almeida, Hector and Daniel Ferreira, "Democracy and the Variability of Economic Performance," *Economics & Politics*, 2002, *14* (3), 225–257.

Angulo, Roberto, Yadira Diaz, and Renata Pardo, "Multidimensional poverty in Colombia, 1997-2010," ISER Working Paper Series 2013-03, Institute for Social and Economic Research January 2013. 29, 34, 35, 36

Anselin, Luc, *Spatial Econometrics: Methods and Models*, Vol. 4 of *Studies in Operational Regional Science*, Dordrecht: Springer Netherlands, 1988. 26

_ , Anil K. Bera, Raymond Florax, and Mann J. Yoon, "Simple diagnostic tests for spatial dependence," *Regional Science and Urban Economics*, February 1996, *26* (1), 77–104.

Attanasio, Orazio, Pinelopi K Goldberg, and Nina Pavcnik, "Trade reforms and wage inequality in Colombia," *Journal of development Economics*, 2004, *74* (2), 331–366. 26

Bailey, Trevor C and Anthony C Gatrell, *Interactive spatial data analysis*, Vol. 413, Longman Scientific & Technical Essex, 1995.

Ballester, Coralio, Antoni Calv-Armengol, and Yves Zenou, "Who's Who in Networks. Wanted: The Key Player," *Econometrica*, September 2006, *74* (5), 1403–1417. 65

Beine, Michel A. R., Frdric Docquier, and Caglar Ozden, "Dissecting Network Externalities in International Migration," January 2011.

Beine, Michel, Frderic Docquier, and Hillel Rapoport, "Brain Drain and Human Capital Formation in Developing Countries: Winners and Losers*," *Econ. Journal*, 2008, *118* (528), 631–652.

_ , Frdric Docquier, and Cecily Oden-Defoort, "A Panel Data Analysis of the Brain Gain," *World Development*, 2011, *39* (4), 523–532.

_ , _ , and Hillel Rapoport, "Brain drain and economic growth: theory and evidence," *J. Dev. Econ.*, 2001, *64* (1), 275–289. 3, 23

Bernstein, Alan, "Collaboration: Link the world's best investigators," *Nature*, April 2013, *496* (7443), 27–27.

BM, DNP, *Sistema de Ciudades. Una aproximacin visual al caso colombiano*, PLANEACION, D. N. D. (ed.). Bogota: Mimeo, 2012. 43

Boyle, Paul, "Policy: A single market for European research," *Nature*, September 2013, *501* (7466), 157–158. 2, 16, 22

Breschi, S. and F. Lissoni, "Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows," *J. Econ. Geog.*, 2009, *9*, 439–468.

Brown, P., A. Green, and H. Lauder, *High Skills: Globalization, Competitiveness, and Skill Formation*, Oxford, UK: Oxford Univ. Press, 2001. 7

Buera, Francisco J and Yongseok Shin, "Financial Frictions and the Persistence of History: A Quantitative Exploration," *Journal of Political Economy*, 2013, *121* (2), 221–272. 25

Cameron, A Colin and Trivedi, Pravin K, *Microeconometrics: methods and applications*, Cambridge University Press, 2005.
Capacity building: Architects of South American science

***Capacity building: Architects of South American science**, Nature, June 2014, 510 (7504), 209–212.*

Chessa, A., A. Morescalchi, F. Pammolli, O. Penner, A. M. Petersen, and M. Riccaboni, *"Is Europe Evolving Toward an Integrated Research Area?," Science, February 2013, 339 (6120), 650–651. 2, 21, 94*

Cimini, Giulio, Andrea Gabrielli, and Francesco Sylos Labini, *"The Scientific Competitiveness of Nations," PLoS ONE, December 2014, 9 (12), e113470. 66, 67, 91, 92*

Commentary, *"Architects of South American Science," Nature, 2014, 510, 209–212.*
Comparative-Benchmarking-of-European-and-US-Research-Collaboration-and-Researcher-Mobility_sept2013.pdf

Comparative-Benchmarking-of-European-and-US-Research-Collaboration-and-Researcher-Mobility_sept2013.pdf.

Cristelli, Matthieu, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero, "Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products," *PLoS ONE*, August 2013, *8* (8), e70726. 66

Dasgupta, Kunal, "Learning and knowledge diffusion in a global economy," *Journal of International Economics*, 2012, *87* (2), 323 – 336.

Delanghe, H., U. Muldur, and L. Soete, *European science and technology policy: Towards integration or fragmentation?*, Cheltenham, UK.: Edwar Elgar, 2009. 2

Dernis, Hlne and Mosahid Khan, "Triadic Patent Families Methodology," OECD Science, Technology and Industry Working Papers, Organisation for Economic Co-operation and Development, Paris March 2004. 67, 68, 79

Deville, P., D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabasi, "Career on the Move: Geography, Stratification, and Scientific Impact.," *Sci. Rep.*, 2014, *4*, 4770. 7

Dijkstra, A. Geske, "Trade Liberalization and Industrial Development in Latin America," *World Development*, 2000, *28* (9), 1567 – 1582. 26

DNP, *Algunos Aspectos Del Anlisis Del Sistema De Ciudades Colombiano*, PLANEACION, D. N. D. (ed.). Bogota: Mimeo, 2012. 43

Docquier, Frdric and Hillel Rapoport, "Globalization, Brain Drain, and Development," *J. Econ. Lit.*, 2012, *50* (3), 681–730. 7, 23

__ **, alar Ozden, and Giovanni Peri**, "The Labour Market Effects of Immigration and Emigration in OECD Countries," *Econ. Journal*, 2014, *124* (579), 1106–1145.

Drukker, David M, Peter Egger, and Ingmar R Prucha, "On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors," *Econometric Reviews*, 2013, *32* (5-6), 686–733. 27

Dustmann, Christian, Itzhak Fadlon, and Yoram Weiss, "Return migration, human capital accumulation and the brain drain," *J. Dev. Econ.*, 2011, *95* (1), 58 – 67. 23

Editorial, "Eastern promise," *Nature*, 2003, *426*, 369.

__ , "Can Europe build framework for success?," *Nature*, 2011, *473*, 421.

__ , "Science without borders," *Nature*, October 2013, *502* (7469), 5–5.

_ , "Capacity building: Architects of South American science," *Nature*, June 2014, *510* (7504), 209–212.

_ , "Global collaboration," *Science*, October 2014, *346* (6205), 47–49.

Elhorst, J Paul, "Applied spatial econometrics: raising the bar," *Spatial Economic Analysis*, 2010, *5* (1), 9–28. 26, 50
■Érdi et al.

Érdi, Pter, Kinga Makovi, Zoltn Somogyvri, Katherine Strandburg, Jan Tobochnik, Pter Volf, and Lszl Zalnyi, "Prediction of emerging technologies based on analysis of the US patent citation network," *Scientometrics*, June 2012, *95* (1), 225–242.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler, "Trade and market selection: Evidence from manufacturing plants in Colombia," *Review of Economic Dynamics*, 2013, *16* (1), 135 – 158. 26

European Commission, "http://ec.europa.eu/internal_market/qualifications/regprof/," 2015.

_ **and Directorate-General for Research and Innovation**, *European Research Area facts and figures 2013.*, Luxembourg: Publications Office, 2013. 22

European Commission Horizon 2020: Spreading excellence and Widening Participation., "http://ec.europa.eu/programmes/horizon2020/en/h2020-section/spreading-excellence-and-widening-participation," 2015.

European Research Area, "http://ec.europa.eu/research/era," 2015.

European Research Council, "http://erc.europa.eu," 2015.
European Commission Horizon 2020: Spreading excellence and widening Participation

***European Commission Horizon 2020: Spreading excellence and widening Participation**, http://ec.europa.eu/programmes/horizon2020/en/h2020-section/spreading-excellence-and-widening-participation. 23*
European Commission: The EU Single Market Regulated professionals database (professionals moving abroad).

European Commission: The EU Single Market Regulated professionals database (professionals moving abroad)., `http://ec.europa.eu/growth/tools-databases/regprof/`. Retrieved August, 2015. 2, 8
European Research Area

***European Research Area**, http://ec.europa.eu/research/era. 2015.*
European Research Council

European Research Council, http://erc.europa.eu. 2015.

Foray, D. and B. Van Ark, "Smart specialisation in a truly integrated research area is the key to attracting more R&D to Europe," Knowledge Economists Policy Brief 1, Knowledge for Growth Expert Group, European Commission 2007.

Foster, James, Joel Greer, and Erik Thorbecke, "A class of decomposable poverty measures," *Econometrica: Journal of the Econometric Society*, 1984, pp. 761–766. 37, 38, 39

Fraser, Barbara, "Research training: Homeward bound," *Nature*, June 2014, *510* (7504), 207–207.

Freeman, Richard B. and Wei Huang, "Collaboration: Strength in diversity," *Nature*, September 2014, *513* (7518), 305–305.

Frenken, K., "A new indicator of European integration and an application to collaboration in scientific research," *Economic Systems Research*, 2002, *14* (4), 345–361.

Gabriel, K. Ruben and Robert R. Sokal, "A New Statistical Approach to Geographic Variation Analysis," *Systematic Biology*, 1969, *18* (3), 259–278. 47, 50

Galvis, Luis and Adolfo Meisel, "Fondo de Compensacin Regional: Igualdad de oportunidades para la periferia colombiana," DOCUMENTOS DE TRABAJO SOBRE ECONOMA REGIONAL 006634, BANCO DE LA REPBLICA - ECONOMA REGIONAL January 2010.

__ **and** __ , "Persistencia de las desigualdades regionales en Colombia: Un anlisis espacial," DOCUMENTOS DE TRABAJO SOBRE ECONOMA REGIONAL 006631, BANCO DE LA REPBLICA - ECONOMA REGIONAL January 2010.

Garber, Mitchell E., Olga G. Troyanskaya, Karsten Schluens, Simone Petersen, Zsuzsanna Thaesler, Manuela Pacyna-Gengelbach, Matt van de Rijn, Glenn D. Rosen, Charles M. Perou, Richard I. Whyte, Russ B. Altman, Patrick O. Brown, David Botstein, and Iver Petersen, "Diversity of gene expression in adenocarcinoma of the lung," *Proceedings of the National Academy of Sciences*, November 2001, *98* (24), 13784–13789. 77

Gasparini, Leonardo, Federico Gutirrez, and Leopoldo Tornarolli, "Growth and Income Poverty in Latin America and the Caribbean: Evidence from Household Surveys," *Review of Income and Wealth*, June 2007, *53* (2), 209–245. xv, 103

Gehlke, CE and Katherine Biehl, "Certain effects of grouping upon the size of the correlation coefficient in census tract material," *Journal of the American Statistical Association*, 1934, *29* (185A), 169–170.

Getis, Arthur, "Cliff, A.D. and Ord, J.K. 1973: Spatial autocorrelation. London: Pion," *Progress in Human Geography*, June 1995, *19* (2), 245–249. 48

Geuna, Aldo (Ed.), *Global mobility of research scientists*, Academic Press, 2015. 2, 7

Giavazzi, Francesco and Guido Tabellini, "Economic and political liberalizations," *Journal of Monetary Economics*, 2005, *52* (7), 1297 – 1330. 17

Gibson, John and David McKenzie, "Eight Questions about Brain Drain," *J. Econ. Perspectives*, 2011, *25* (3), 107–28. 3, 23
Global collaboration

***Global collaboration**, Science, October 2014, 346 (6205), 47–49.*

Goldberg, Pinelopi K. and Nina Pavcnik, *"Trade, Inequality, and Poverty: What Do We Know? Evidence from Recent Trade Liberalization Episodes in Developing Countries," Working Paper 10593, National Bureau of Economic Research June 2004. 26*

Goldberg, Pinelopi Koujianou and Nina Pavcnik, *"Trade, wages, and the political economy of trade protection: evidence from the Colombian trade reforms," Journal of international Economics, 2005, 66 (1), 75–105. 27*

Grossmann, Volker and David Stadelmann, *"Does international mobility of high-skilled workers aggravate between-country inequality?," J. Dev. Econ., 2011, 95 (1), 88 – 94. 3, 23*

Hausmann, Ricardo, ed., *The atlas of economic complexity: mapping paths to prosperity, updated edition ed., Cambridge, MA: The MIT Press, 2013. 65, 67, 75, 76, 89, 90, 91, 92*

Helpman, Elhanan, Oleg Itskhoki, and Stephen Redding, *"Inequality and unemployment in a global economy," Econometrica, 2010, 78 (4), 1239–1283. 25, 27, 53, 61*

_ , _ , Marc-Andreas Muendler, and Stephen J. Redding, *"Trade and Inequality: From Theory to Estimation," Working Paper 17991, National Bureau of Economic Research April 2012.*

Hoekman, J., K. Frenken, and F. Van Oort, *"The geography of collaborative knowledge production in Europe," The Annals of Regional Science, 2009, 43, 721–738. 2*

__ , **T. Scherngell, K. Frenken, and R. Tijssen**, *"Acquisition of European research funds and its effect on international scientific collaboration," J. Econ. Geog., 2013, 13 (1). 22*

Hoekman, Jarno, Koen Frenken, and Frank van Oort, *"The geography of collaborative knowledge production in Europe," The Annals of Regional Science, July 2008, 43 (3), 721–738.*

Hoyos, Rafael E De, Maurizio Bussolo, and Oscar Nez, *"Exports, Gender Wage Gaps, and Poverty in Honduras," Oxford Development Studies, 2012, 40 (4), 533–551. 25*

Jaffe, A. B., M. Trajtenberg, and R. Henderson, *"Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," The Quarterly Journal of Economics, 1993, 434, 578–598.*

Juhn, Chinhui, Gergely Ujhelyi, and Carolina Villegas-Sanchez, *"Men, women, and machines: How trade impacts gender inequality," Journal of Development Economics, 2014, 106 (0), 179 – 193.*

Kahanec, Martin, *"Labor mobility in an enlarged European Union," in A. F. Constant and K. F. Zimmermann, eds., International Handbook on the Economics of Migration, Cheltenham, UK.: Edwar Elgar, 2013, chapter 7, pp. 137–152. 21*

Kelejian, Harry H. and Ingmar R. Prucha, *"A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model," International Economic Review, May 1999, 40 (2), 509–533.*

__ **and** __ , *"Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances," Journal of Econometrics, 2010, 157 (1), 53 – 67. Nonlinear and Nonparametric Methods in Econometrics.*

Laursen, Lucas, *"Europe waters down transnational 'research buddy'plan," Nature, 2013, p. doi:10.1038/nature.2013.14406.*

Lepori, B., M. Seeber, and A. Bonaccorsi, *"Competition for talent. Country and organizational-level effects in the internationalization of European higher education institutions," Research Policy, 2015, 44 (3), 789 – 802. 2, 5, 23*

LeSage, James P and R Kelley Pace, *"Spatial Econometric modeling of origin-destination flows," Journal of Regional Science, 2008, 48 (5), 941–967. 26*

Lopez, J Humberto and Guillermo Perry, *"Inequality in Latin America: determinants and consequences," World Bank 2008.*

Lundberg, Mattias and Lyn Squire, *"The simultaneous evolution of growth and inequality*," The Economic Journal, 2003, 113 (487), 326–344. 25*

Maraut, S, H Dernis, C Webb, V Spiezia, and D Guellec, *"The OECD REGPAT Database: a Presentation," Working Paper 2, OECD 2008.*

Marchiori, Luca, I.-Ling Shen, and Frdric Docquier, *"BRAIN DRAIN IN GLOBALIZATION: A GENERAL EQUILIBRIUM ANALYSIS FROM THE SENDING COUNTRIES' PERSPECTIVE," Economic Inquiry, 2013, 51 (2), 1582–1602.*

Markl, Hubert S., *"Battle for the brains," Science, 2005, 310, 1585.*

Martinez, Catalina, *"Insight into Different Types of Patent Families," OECD Science, Technology and Industry Working Papers, Organisation for Economic Co-operation and Development, Paris February 2010. 67*

Matula, David W. and Robert R. Sokal, *"Properties of Gabriel Graphs Relevant to Geographic Variation Research and the Clustering of Points in the Plane," Geographical Analysis, 1980, 12 (3), 205–222. 47, 50*

Meja, Daniel and Marc St-Pierre, *"Unequal opportunities and human capital formation," Journal of Development Economics, 2008, 86 (2), 395–413.*

Moed, Henk F., M Aisati, and A Plume, *"Studying scientific migration in Scopus," Scientometrics, 2013, 94, 929–942. 7*

Moran, Patrick AP, *"Notes on continuous stochastic phenomena," Biometrika, 1950, 37 (1-2), 17–23. 48, 99*

Morescalchi, Andrea, Fabio Pammolli, Orion Penner, Alexander M. Petersen, and Massimo Riccaboni, *"The evolution of networks of innovators within and across borders: Evidence from patent data," Research Policy, April 2015, 44 (3), 651–668. 2, 21*

Moro-Martin, Amaya, *"A call to those who care about Europes science," Nature, 2014, 514, 141.*

Morrison, Greg, Eleftherios Giovanis, Fabio Pammolli, and Massimo Riccaboni, *"Border sensitive centrality in global patent citation networks," Journal of Complex Networks, December 2014, 2 (4), 518–536.*

Murtagh, Fionn and Pierre Legendre, *"Wards Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Wards Criterion?," Journal of Classification, October 2014, 31 (3), 274–295.*

Nedeva, M. and M. Stampfer, *"From Science in Europe to European Science," Science, 2012, 336 (982). 23*

Newman, M. E. J., *"Finding community structure in networks using the eigenvectors of matrices," Phys. Rev. E, Sep 2006, 74, 036104.*

_ , *"Modularity and community structure in networks," Proc. Nat. Acad. Sci. USA, 2006, 103 (23), 8577–8582.*

_ , *"Spectral methods for community detection and graph partitioning," Phys. Rev. E, Oct 2013, 88, 042822.*

Nissanke, Machiko, Erik Thorbecke, and Guido Porto, *"The Impact of Globalization on the World's Poor: Transmission Mechanisms," Journal of Economic Literature, 2008, 46 (1), 179–182. 25*

Noorden, Richard Van, *"Science on the move," Nature, 2012, 490, 326. 7*

of Medicine (US) Committee on Accelerating Progress in Obesity Prevention, Institute and Dan Glickman, *Accelerating progress in obesity prevention: solving the weight of the nation, National Academies Press Washington, DC:, 2012.*

_ **and** _ , *Accelerating progress in obesity prevention: solving the weight of the nation, National Academies Press Washington, DC:, 2012.*
On the mend

On the mend, *Nature*, November 2014, *515* (7525), 7–8.

Organization, The World Intellectual Property, "http://www.wipo.int/classifications/en/," 2016. 68

Paci, Raffaele and Stefano Usai, "Knowledge flows across European regions," *The Annals of Regional Science*, 2009, *43* (3), 669–690.

Pan, Raj K., Alexander M. Petersen, Fabio Pammolli, and Santo Fortunato, "The memory of Science: Inflation, Myopia, and the Knowledge Network," *Under Review*, 2015. 7

Persson, Torsten and Guido Tabellini, "Democracy and Development: The Devil in the Details," *Am. Econ. Rev.*, 2006, *96* (2), 319–324. 17

Petersen, Alexander M., Ioannis Pavlidis, and Ioanna Semendeferi, "A Quantitative Perspective on Ethics in Large Team Science," *Science and Engineering Ethics*, June 2014, *20* (4), 923–945. 19, 21

Petersen, Alexander Michael, "Quantifying the impact of weak, strong, and super ties in scientific careers," *Proc. Natl. Acad. of Sci.*, 2015, *112* (34), E4671–E4680. 19

Piketty, Thomas, *Capital in the Twenty-First Century*, Harvard University Press, March 2014. 26

_ **and Emmanuel Saez**, "Inequality in the long run," *Science*, 2014, *344* (6186), 838–843.

PNUD, "Pobreza y cambio climtico. Trabajo realizado para Naciones Unidas. Programa conjunto de cambio climtico.," Technical Report, Bogot: PNUD. 2010. 44

__ , *Colombia Rural: Razones para una esperanza. Informe de Desarrollo Humano*, Bogot: INDH PNUD, Septiembre., 2011. 43, 44

Ramirez, Juan Mauricio, Yadira Diaz, and Juan Guillermo Bedoya, "Decentralization in Colombia: A Search for Equity in a Bumpy Economic Geography," Technical Report 2016. IARIW-IBGE Conference on Income, Wealth and Well-Being in Latin America (Rio de Janeiro, Brazil, September 11-14, 2013). 26, 29, 43

Rao, Calyampudi Radhakrishna, Xiaoping Shi, and Yuehua Wu, "Approximation of the expected value of the harmonic mean and some applications," *Proceedings of the National Academy of Sciences of the United States of America*, November 2014, *111* (44), 15681–15686. 66

Ravallion, Martin, "Growth, inequality and poverty: looking beyond averages," *World development*, 2001, *29* (11), 1803–1815. 25

__ , "Looking beyond averages in the trade and poverty debate," *World Development*, 2006, *34* (8), 1374–1392. 25

Scherngell, T and M J Barber, "Distinct spatial characteristics of industrial and public research collaborations: Evidence from the 5th EU Framework Programme," *Ann. Reg. Sci.*, 2011, *46*, 247–266.

__ **and R Lata**, "Towards an integrated European Research Area? Findings from Eigenvector spatially filtered spatial interaction models using European Framework Programme data," *Papers in Regional Science*, 2013, *92*, 555–577.

Scherngell, Thomas (Ed.), *The Geography of Networks and R&D Collaborations*, Springer International Publishing, 2013. 2

Science Europe, "http://www.scienceeurope.org," 2015.
Science Europe

***Science Europe**, http://www.scienceeurope.org. 2015.*
Science without borders

Science without borders, *Nature*, October 2013, *502* (7469), 5–5.

SCImago Journal \& Country rank (Accessed 2014), "http://www.scimagojr.com/compare.php," 2014.
SCImago: SJR SCImago Journal and Country Rank.

SCImago: SJR SCImago Journal and Country Rank., *http://www.scimagojr.com/compare.php. Retrieved September, 2015, from http://www.scimagojr.com. xi, 4, 5*

Sen, Amartya, *"Poverty: An Ordinal Approach to Measurement," Econometrica, 1976, 44 (2), 219–231. 28*

Sloan, Susan Sauer and Joe Alper, *Culture matters 2014.*

Smaglik, Paul, *"Europe: Swedish success story," Nature, October 2013, 502 (7473), 711–712.*

Statistics - Professionals moving abroad (establishment) European Commission, *"http://ec.europa.eu/growth/tools-databases/regprof/index.cfm?action=statistics&b_services=false."*

Suttmeier, Richard P., Cong Cao, and Denis Fred Simon, *""Knowledge Innovation" and the Chinese Academy of Sciences," Science, 2006, 312, 58–59.*

Suzuki, Ryota and Hidetoshi Shimodaira, *"Pvclust: an R package for assessing the uncertainty in hierarchical clustering," Bioinformatics, June 2006, 22 (12), 1540–1542. 78*

Taylor, David A., *"Mali Researcher Shows How To Reverse Brain Drain," Science, 2011, 332, 1498–1499.*

The League of European Research Universities, *"Memorandum of Understanding on ERA (www.leru.org)," July 2012.*

Trachana, Varvara, *"Austerity-led brain drain is killing Greek science," Nature, 2013, 496, 271.*

Unwin, David J, *"GIS, spatial analysis and spatial statistics," Progress in Human Geography, 1996, 20 (4), 540–551.*

Verbeek, Marno, *A Guide to Modern Econometrics, 4th ed., John Wiley & Sons, 2008.*

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

Waller, Lance A and Carol A Gotway, *Applied spatial statistics for public health data, Vol. 368, John Wiley & Sons, 2004.*

Wang, Dan, *"Activating Cross-border Brokerage: Interorganizational Knowledge Transfer through Skilled Return Migration," Admin. Sci. Quart, 2015, 60 (1), 133–176. 23*

Weeks, John R, *"The role of spatial analysis in demographic research," Spatially integrated social science, 2004, pp. 381–399.*

Weinberg, Bruce A., *"Developing science: Scientific performance and brain drains in the developing world," J. Dev. Econ., 2011, 95 (1), 95 – 104. 3, 23*

Wiesel, Torsten, *"Fellowships: Turning brain drain into brain circulation," Nature, June 2014, 510 (7504), 213–214. 23*

Wood, Adrian, *"Openness and Wage Inequality in Developing Countries: The Latin American Challenge to East Asian Conventional Wisdom," The World Bank Economic Review, January 1997, 11 (1), 33–57. 26*

Wooldridge, Michael, *An introduction to multiagent systems, John Wiley & Sons, 2009. 42*

World Bank, *"http://data.worldbank.org/indicator/," 2015.*
World Bank data sources.

World Bank data sources., `http://data.worldbank.org/indicator`. Accessed: 2015-08. 7

Wrigley, Neil, "Revisiting the modifiable areal unit problem and the ecological fallacy," *Diffusing geography*, 1995, pp. 123–181.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi, "The Increasing Dominance of Teams in Production of Knowledge," *Science*, May 2007, *316* (5827), 1036–1039.

Xie, Yu, Chunni Zhang, and Qing Lai, "Chinas rise as a major contributor to science and technology," *Proceedings of the National Academy of Sciences*, July 2014, *111* (26), 9437–9442.

Zappe, Hans, "Innovation: Bridging the market gap," *Nature*, September 2013, *501* (7468), 483–485.

Zuniga, Pluvia and Dominique Guellec, "Who Licenses Out Patents and Why? Lessons from a Business Survey," SSRN Scholarly Paper ID 1399144, Social Science Research Network, Rochester, NY January 2009. 67