Universidad de San Andrés

Departamento de Economía

Doctorado en Economía

Sampled Networks, Peer Effects and Spatial VAR model

Alumno:
Alejandro Izaguirre
DNI: 29945269

Director:
Gabriel Montes Rojas

Resumen:

La presente tesis está constituida por tres artículos diferentes cuyo contenido, en forma muy general, pertenece al área de estudio de la estadística-econometría.

El primer trabajo, denominado, *“Horvitz-Thompson estimator under partial information with an application to network degree distribution”*, es un trabajo metodológico donde se presenta una extensión del estimador de Horvitz-Thompson (H-T) cuya función es estimar el total de conjuntos de un tamaño dado en una población pero utilizando una muestra, no de conjuntos, sino de los elementos que forman dichos conjuntos. Básicamente, lo que tenemos es un problema de información parcial dado que no observamos las unidades que queremos estimar (conjuntos), sino una parte de ellas (sus elementos).

Una vez adaptado el estimador de H-T al escenario de información incompleta lo que hacemos es utilizarlo para estimar la *network degree distribution* que es una descripción de la frecuencia relativa de los nodos de una red con diferentes números de links. El estimador propuesto se puede utilizar bajo un amplio rango de muestreos de tipo probabilísticos. En el trabajo lo adaptamos a un tipo particular de muestreo conocido como *“induced subgraph sampling”*.

Para finalizar realizamos una simulación de tipo Monte Carlo para evaluar la performance del estimador propuesto y compararlo con una estimación de tipo “naive”.

El segundo trabajo, denominado *“Regional and state heterogeneity of monetary shocks in Argentina”* analiza la heterogeneidad espacial del impacto de la política monetaria en Argentina. Se plantean dos niveles de desagregación espacial, uno por provincias y otro por regiones, y se evalúa cómo los cambios en la tasa de interés afectan la actividad económica (medida a través del empleo formal) en las distintas áreas espaciales consideradas. El modelo utilizado es un modelo de vectores autorregresivos espaciales (SpVAR) que tiene la particularidad de considerar la interacción de las unidades espaciales (provincias o regiones según el caso).

Por último, el tercer trabajo, denominado *“Exploring peer effects in education in Latin America and the Caribbean* es un trabajo aplicado cuyo objetivo es evaluar la existencia de peer effects en la educación, es decir la existencia de interacciones o influencias dentro el aula que tengan impacto en los resultados académicos. Para ello se utiliza la base de datos TERCE que reúne información sobre el desempeño académico en matemática, ciencia y lengua, de los alumnos de tercer y sexto grado

para los países de América Latina.

El modelo utilizado para el análisis es el propuesto en (Lee, 2007) que permite identificar dos de los tres efectos que constituyen los peer effects, a saber, efectos endógenos y exógenos, controlando por efectos correlacionados a nivel grupal.

A su vez se evalúa la existencia de posibles heterogeneidades en dichos efectos, específicamente, se divide la muestra en escuelas públicas, privadas y rurales y se estima el modelo para cada subconjunto de datos.

Capítulo I

Horvitz-Thompson estimator under partial information with an application to network degree distribution[1]

Abstract

We present an extension of the Horvitz-Thompson estimator for estimating the total number of sets of a given size within a population. We study the limitations of the Horvitz-Thompson under partial information where we have a sample of population elements rather than of sets. The developed estimator is the chained Horvitz-Thompson. We apply this estimator for estimating the network degree distribution under probability sampling designs, in particular, induced subgraph. Finally, we present Monte Carlo simulations to assess the accuracy of the proposed estimator.

Key words: *Horvitz-Thompson estimator, networks, network sampling designs, degree distribution, average degree.*

JEL classification: C13, C4.

[1] Una versión extendida de este artículo fue publicada en *Communications in Statistics-Simulation and Computation* (2019): 1-24, en coautoria con Gabriel Montes Rojas.

1 Introduction

A growing literature related to social networks and their implications in economic outcomes emerged during the last years (see Jackson et al., 2008). A network represents a set of connections (edges) among a collection of agents (nodes). Most networks investigated today are parts of much larger networks. Although many applied works speak of *the network* when presenting empirical results, frequently it is only a sampled version of some larger underlying network. Sampling is of particular interest in the context of online social networks because they are usually very large.

Although the first works on network sampling started in the late 1960s by the hand of Ove Frank and his colleagues, the subject showed a growing interest in the last years in a number of fields such as economics, epidemiology, statistics, sociology and computer science, among others.

There are many papers in which the focus is on understanding the extent to which characteristics of a sampled network correspond to the complete network (see Zhang et al., 2015, for a literature review). Focusing on the analysis of the internet topology, this issue is studied in Lakhina et al. (2003) and Achlioptas et al. (2009). Typical characteristics of interest include degree distribution, density, diameter, clustering coefficient, average path length, among others. In general, under many sampling schemes, these measures are biased when we use sampled networks. One of the first works in proposing an estimator for network characteristics (average degree and density) based on a sampled network was Granovetter (1976).

Network models are widely used to represent information among interacting units and the structural implication of these relations, and the estimation of such models based on sampled networks suffers bias problems too. See, for instance, Handcock & Gile (2010), Santos & Barrett (2008) and Chandrasekhar & Lewis (2011). In the latter, the authors not only show how the bias arise when we use sampled data in the estimation of econometrics models, but also propose procedures to correct such biases. One of them is based on a graphical reconstruction, that is, using sample information to predict the full network. This approach requires a series of assumptions about the network formation process.

In this paper we focus our attention on the network degree distribution. The degree distribution of a network is a description of the relative frequencies of nodes that have different degrees. This measure is one of the most fundamental characteristic associated with a network and it is affected by sampling, sometimes dramatically.

Network sampling is a highly relevant topic in the field of network science. Some references for the sampling literature are Rothenberg (1995), Handcock & Gile (2010), (Kolaczyk, 2009, ch.5), Ahmed et al. (2014), among many others. In this paper we restrict our attention to probability sampling designs.

The main problem we face in this article is given by the lack of information. Briefly, we are interested in sets but we do not observe sets, we observe elements of those sets instead. Given this issue, in the first part we extend the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952) to make it feasible under this setting. In the second part, we propose an estimator for the network degree distribution to be used under any probability sampling designs, and then we apply it for a widely used probability sampling design known as induced subgraph (i.e., nodes sampling). A related work is Frank (1977), which presents an adaptation of the HT estimator to be used in networks contexts, but under the same framework as the original HT estimator. Finally we do a Monte Carlo simulation to assess the analytical results.

The paper is organized as follows. Section 2 presents the limitations of the Horvitz-Thompson estimator under partial information and Section 3 the corrected estimators. Section 4 develops the chained Horvitz-Thompson estimator. Section 5 applies the proposed estimators to the network degree distribution. Section 6 presents Monte Carlo simulations. Section 7 concludes.

2 The Horvitz-Thompson estimator under partial information

The HT estimator is a well known estimator mostly used for estimating population totals under probability sampling designs (see Fuller, 2011). In this paper we are interested in estimating population totals, but the HT estimator is unfeasible under the framework we work with.

The elements of the main population are grouped into sets and we want to know how many sets of a given size there are into the population. The issue for using the HT estimator is because the sample we have has only partial information about the sets. In particular, we have a sample of elements of the population, which means that we do not have a sample of sets, rather a sample of elements of the sets.

Consider a population and a collection of sets given by the following definitions.

Definition 1. Let $U = \{u_1, u_2, \ldots, u_N\}$ be a population of size $N < \infty$ whose elements are grouped into $J < \infty$ sets (not necessarily exclusive).

Definition 2. Let $\Theta = \left\{\theta^1, \ldots, \theta^J\right\}$ be a collection of the J sets in which the elements of U are grouped, with generic element $\theta = \{u_j, u_h, \ldots, u_l\}$, $1 \leq j, h, \ldots, l \leq N$, θ_k is a generic element of size k, with $k = 1, 2, \ldots, p$ and T_k is the total number of sets of size k in Θ.

Θ contains all the sets in which the elements of U are grouped into, such elements are not necessarily exclusive.

The total number of sets of size k in Θ can be written as $T_k = \sum_{\theta \in \Theta} I\left[\#\theta = k\right]$, where $\#\theta$ denotes the cardinality of θ, that is, the number of elements in θ. We want to estimate how many sets of size k there are into the population, that is, we want to estimate T_k. If we had a sample of Θ we could use the HT estimator as follows,

$$\hat{T}_k^{HT} = \sum_{\theta \in \Theta^s} \pi_\theta^{-1} I\left[\#\theta = k\right], \tag{1}$$

where $\Theta^s \subset \Theta$ is a sample of Θ and π_θ is the probability of selection of θ. If we have a sample of U instead of Θ the HT estimator is unfeasible because we do not know θ.

In this paper we propose an unbiased estimator for T_k (and its variance) based on the HT estimator to be used when we have a sample of U instead of Θ. The only requirement is that the sample has certain information about θ (see Assumption 1).

As an example suppose we want to know the total number of blocks in a city with different numbers of houses, that is, how many blocks are with, for example, 20 houses, how many with 25, and so on. Under this setting our population U is given by the N houses of the city which are grouped into J blocks, being θ a generic block and θ_k a generic block with k houses. Θ is a set (or collection) that contains all θ, and T_k is the total number of blocks with k houses. If we have a probability sample of Θ (a sample of blocks) we could use (1) for estimating T_k, but suppose we have instead a probability sample of U (a sample of houses instead of a sample of blocks), if we do not observe θ the estimator (1) is unfeasible.

3 Extending the HT estimator for partial information

We have a population whose elements are grouped into sets, and our main goal here is to estimate how many sets of a given size there are in the population. The problem is that we partially observe those sets, that is, we only observe subsets.

For an intuitive introduction to the methodology proposed here suppose we divide each set of the population Θ into all possible combinations, that is, we generate all possible subsets from *each set* of Θ, then suppose we arrange these subsets into a new population named Γ. In other words, we create a new population given by all subsets we can get from *each set* of Θ.

The total number of subsets of a given size in Γ is a linear combination of the total number of sets in Θ. This relation is the key for the estimator we propose here. Our proposal consists in estimating the total number of subsets in Γ (since we observe them by assumptions), and then to estimate the total number of sets in Θ based on the implied linear relationship.

Let L_c be the total number of subsets of size c in Γ, we have that $L_c = \sum_{k=1}^{p} T_k \begin{pmatrix} k \\ c \end{pmatrix}$. As noted above, the total number of subsets of size c in Γ, that is L_c, is a linear combination of the total number of sets in Θ, T_k.

As an example, suppose Θ contains 4 sets of size 5 and 3 sets of size 4, that is, $T_5 = 4$ and $T_4 = 3$, so there are $L_4 = 4 \begin{pmatrix} 5 \\ 4 \end{pmatrix} + 3 \begin{pmatrix} 4 \\ 4 \end{pmatrix}$ subsets of size $c = 4$ in Γ.

Generalizing the previous statement we have the following equality

$$L = AT, \tag{2}$$

where $L = (L_1, L_2, \ldots, L_p)^{'}$ is a $(p \times 1)$ vector, $A = [a_{ij}]$, where $a_{ij} = \begin{pmatrix} j \\ i \end{pmatrix}$ with i, $j = 1, \ldots, p$ is a $(p \times p)$ upper unitriangular matrix (the matrix A is a submatrix of the Pascal Matrix, we present some properties about A in the Appendix), and $T = (T_1, T_2, \ldots, T_p)^{'}$ is a $(p \times 1)$ vector.

The element T_k of T is the total number of sets of size k in Θ, the element L_c of L is the total number of subsets of size c in Γ. The matrix A transforms the total number of sets in Θ into the total number of subsets in Γ. The previous equation summarizes the linear relation between the total number of sets (of different sizes) and the total number of subsets (of different sizes) we can get from them.

Our methodology consists in estimating L and then replace it in $T = A^{-1}L$ to get an estimator for T. Given that A is a non stochastic upper unitriangular matrix its inverse exists, furthermore, $A^{-1} = \left[\bar{a}_{ij}^{-1}\right]$ where $\bar{a}_{ij}^{-1} = (-1)^{j+i} \begin{pmatrix} j \\ i \end{pmatrix}$.

3.1 The Horvitz-Thompson estimator for L_c

Consider the following definitions and assumptions.

Definition 3. Let S be a probability sample of U. Let $\mathcal{S} = 2^U$ be the σ-field of all possible samples of U, and Π be a probability measure, $\Pi : \mathcal{S} \mapsto [0, 1]$. The triplet $(U, \mathcal{S}, \Pi)$ is a probability space.

Our raw material here are the subsets we can get from the sets of Θ, we construct an "artificial" population of such subsets as follows:

Definition 4. Let $\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$ be a collection of all possible subsets we can get from ***each element*** of Θ where $P(\theta)$ is the power set of θ, γ is a generic element of Γ, and γ_c a generic element of size c, with $c = 1, 2, \ldots, p$. The total number of elements of size c in Γ is given by $L_c = \sum_{\gamma \in \Gamma} I[\#\gamma = c]$.

Although the elements of Γ (and Γ^s) are sets we call them subsets to highlight that they are subsets of θ. To avoid confusions we refer to the elements of Θ as sets and to the elements of Γ (and Γ^s) as subsets.

Definition 5. Let $\Gamma^s = \{\gamma | \forall \gamma \in S\}$ be the collection of all subsets γ in sample S, $\Gamma^s \subseteq \Gamma$. Let $p^s \leq p$ be the maximum size of $\gamma \in S$.

To clarify, S is a sample of U, and $\gamma \subseteq \theta$ are subsets composed by elements of U. What we mean by $\gamma \in S$ is that all $u \subseteq \gamma$ are in S.

Definition 6. Let $\pi_\gamma = Prob(\gamma \in \Gamma^s)$ be the probability of selection of γ and $\pi_{\gamma\gamma'} = Prob(\gamma \cup \gamma' \in \Gamma^s)$ be the joint probability of selection of γ and γ'. Given definition 5, $\pi_\gamma = Prob(\gamma \in S)$ and $\pi_{\gamma\gamma'} = Prob(\gamma \cup \gamma' \in S)$.

Consider now the following assumptions.

Assumption 1. *For all $u \in S$ we can identify every θ in Θ such that $u \in \theta$.*

Assumption 2. $\pi_\gamma > 0 \; \forall \left[\#\gamma \leq p\right] \in \Gamma$. *The probability of selection of any element in Γ has to be positive.*

Assumption 3. $\pi_{\gamma\gamma'} > 0 \; \forall \left[\#\gamma \leq p\right] \vee \left[\#\gamma' \leq p\right] \in \Gamma$. *The joint probability of selection of any pair of elements in Γ has to be positive.*

Assumption 1 is the key assumption for the proposed estimator. It does not imply that we know all the elements in θ, it only assumes we know to which set (or sets) each element of the sample belongs. In other words, we can match every $u \in S$ with any $\theta \in \Theta$ such that $u \in \theta$. The idea behind this assumption is that we can group the elements in S as they are grouped in Θ, and thus we have a sample of subsets of θ. These subsets could indeed be θ, but we do not know it a priori.

Consider the following example. Let $U = \{u_1, u_2, u_3, u_4, u_5\}$ be a population of individual elements, let $\Theta = \left\{\theta^1, \theta^2, \theta^3\right\}$ be a population of sets, where $\theta^1 = (u_1, u_3)$, $\theta^2 = (u_2, u_3, u_5)$, $\theta^3 = (u_4, u_5)$, and let $S = \{u_1, u_2, u_5\}$ be a sample of U. Assumption 1 assumes we know that u_1 belongs to θ^1, u_2 belongs to θ^2, and u_5 belongs to θ^2 and θ^3.

Assumption 1 allows to create $\Theta' = \left\{\theta'^1, \theta'^2, \theta'^3\right\}$, where $\theta'^1 = (u_1)$, $\theta'^2 = (u_2, u_5)$ and $\theta'^3 = (u_5)$. We do not necessarily recover θ in S but we recover a subset θ' instead.

Assumption 1 guarantees that we can identify all γ in S, this is because γ is no more than a combination of elements of θ, and if we can identify the elements of θ into the sample then we can identify any combination of them. Following with the previous example, let

$$\Gamma = \left\{\gamma^{11}, \gamma^{12}, \gamma^{13}, \gamma^{21}, \gamma^{22}, \gamma^{23}, \gamma^{24}, \gamma^{25}, \gamma^{26}, \gamma^{27}, \gamma^{31}, \gamma^{32}, \gamma^{33}\right\}$$

be the population of all possible subsets from each element of Θ, where $\gamma^{11} = (u_1)$, $\gamma^{12} = (u_3)$, $\gamma^{13} = (u_1, u_3)$, $\gamma^{21} = (u_2)$, $\gamma^{22} = (u_3)$, $\gamma^{23} = (u_5)$, $\gamma^{24} = (u_2, u_3)$, $\gamma^{25} = (u_2, u_5)$, $\gamma^{26} = (u_3, u_5)$, $\gamma^{27} = (u_2, u_3, u_5)$, $\gamma^{31} = (u_4)$, $\gamma^{32} = (u_5)$ and $\gamma^{33} = (u_4, u_5)$. Let $\Gamma^s = \left\{\gamma^{11}, \gamma^{21}, \gamma^{23}, \gamma^{25}, \gamma^{32}\right\}$ be the collection of all subsets γ in S.

The elements of Γ^s are identifiable into S because they are subsets of the elements of Θ', that is, $\Gamma^s = \cup_{\theta' \in \Theta'} \left\{P\left(\theta'\right) - \emptyset\right\}$. The identifiability of Γ^s by S is the reason why we create the artificial population Γ.

Assumptions 2 and 3 impose standard restrictions on the sampling designs.

Table 1 summarizes the concepts and assumptions presented above.

Table 1: Populations and samples, summary

Population	Definition	Example	Parameter
$U = \{u_1, u_2, \ldots, u_N\}$	Population of individual elements.	Population of houses.	-
$\Theta = \left\{\theta^1, \ldots, \theta^J\right\}$	Population of sets in which the elements of U are grouped.	Population of blocks.	T_k =Total number of sets of size k in Θ.
$\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$	Population of all possible subsets we can get from each element of Θ. It is an "artificial" population.	All possible combinations of houses we can get from each block.	L_c = Total number of subsets of size c in Γ.
$S = \{u_1, u_2, \ldots, u_n\}$	Sample of U.	Sample of houses.	-
$\Gamma^s = \{\gamma \| \forall\, \gamma \in S\}$	Sample of Γ built from S.	Sample of houses grouped as they are in Θ and all possible combinations of such groups.	-

Briefly, we do not have a sample of the population whose parameters we are interested in, that is, we do not have a sample of Θ, but under Assumption 1 we have a sample of an "alternative" population Γ and we can estimate certain parameters of Γ which are linearly related with the parameters we are interested in.

We can summarize the proposed methodology in the following steps.

1. Transform the population of sets Θ into an artificial population of subsets, Γ. We know the total number of elements of a given size in both populations are linearly related.

2. Use S (a sample of U) to construct Γ^s (a sample of Γ). Assumption 1 allows to transform the sample S into the sample Γ^s.

3. Based on Γ^s estimate L estimating L_c for all c (see section below).

4. Once we estimate L, estimate T_k based on $T = A^{-1}L$.

3.2 Estimator for L_c

We can estimate L_c using the HT estimator as follows,

$$\hat{L}_c^{HT} \;=\; \sum_{\gamma \in \Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right], \tag{3}$$

where$I[\cdot]$ is the indicator function.

Lemma 1. *Under Assumptions 1 and 2, we have that* $E\left(\hat{L}_c^{HT}\right) = L_c$.

Proof. Let I_γ be an indicator function which takes the value one if $\gamma \in \Gamma^s$ and zero otherwise, then we have that $E\left(I_\gamma\right) = \pi_\gamma$. The estimator (3) can be expressed as $\hat{L}_c^{HT} = \sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] I_\gamma$, thus we have $E\left(\hat{L}_c^{HT}\right) = \sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] E\left(I_\gamma\right)$, and given that $E\left(I_\gamma\right) = \pi_\gamma$, we have that $\hat{L}_c^{HT}$ is an unbiased estimator of L_c. □

It is worth to note here that Assumption 2 only requires that any subset in Γ has positive probability to be sampled, it does not mean that the sample Γ^s must have elements γ of all different sizes. For $\hat{L}_c^{HT}$ be unbiased, it does not matter the size of the subsets γ in Γ^s, we only have to be sure that any subset could be sampled.

Let $p^s \leq p$ be the maximum size of γ in Γ^s, we have that $\hat{L}_c^{HT} = 0$ for all $c > p^s$, and also that $\hat{L}_c^{HT} > 0$ for all $c \leq p^s$. This statement is a consequence of the definition of Γ: if we have γ in the sample we also have all possible combinations of its elements, for instance, if we observe $\gamma^1 = (u_2, u_5)$ in Γ^s we also observe $\gamma^2 = (u_2)$ and $\gamma^3 = (u_5)$. In other words, if we observe γ we also observe every $\gamma' \subset \gamma$.

3.3 Estimator for L

We propose to estimate $L = \left(L_1, L_2, \ldots, L_p\right)'$ replacing each element by $\hat{L}_c^{HT}$, thus we have,

$$\hat{L}^{HT} = \left(\hat{L}_1^{HT}, \hat{L}_2^{HT}, \ldots, \hat{L}_p^{HT}\right)'. \tag{4}$$

Lemma 2. *Under Assumptions 1 and 2 we have that* $E\left(\hat{L}^{HT}\right) = L$.

Proof. Under Assumptions 1 and 2 $\hat{L}_c^{HT}$ is unbiased for all c, thus $\hat{L}^{HT}$ is unbiased. □

3.4 Variances and covariances for $\hat{L}_c^{HT}$

Let γ_c be an element of Γ of size c, and let $\gamma'_{c'}$ be an element of Γ of size c'. Then, $Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$ can be expressed as follows,

$$\sigma_{cc'} = Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right]\left(\pi_{\gamma\gamma'} - \pi_\gamma\pi_{\gamma'}\right).$$

It is worth to note that, for $c = c'$ the $Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = Var\left(\hat{L}_c^{HT}\right)$.

The proposed estimator for the covariance is given by

$$\hat{\sigma}_{cc'} = \widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) = \sum_{\gamma\in\Gamma^s}\sum_{\gamma'\in\Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right]\left(\pi_{\gamma\gamma'} - \pi_\gamma\pi_{\gamma'}\right)\pi_{\gamma\gamma'}^{-1}.$$

Lemma 3. *Under Assumptions 1-3, we have that* $E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] = Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$.

Proof. See Appendix 8.2.1. □

Provided that the sample will not necessarily include subsets γ of all sizes, then $\hat{\sigma}_{cc'} = 0$ for $c \wedge c' > p^s$.

3.5 Variance-Covariance matrix for $\hat{L}^{HT}$

Let $\Omega_L = [\sigma_{cc'}]$ be the $(p \times p)$ variance-covariance matrix of $\hat{L}^{HT}$, replacing each element of Ω_L by its unbiased estimator we have the estimator for Ω_L, $\hat{\Omega}_L = [\hat{\sigma}_{cc'}]$, where $\hat{\sigma}_{cc'} = \widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)$.

Lemma 4. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_L\right) = \Omega_L$

Proof. Under Assumptions 1-3, $\hat{\sigma}_{cc'}$ is unbiased for all c and c', thus $\hat{\Omega}_L$ is unbiased. □

4 The chained Horvitz-Thompson estimator for T

Based on equation (2) the estimator we propose for T is given by,

$$\hat{T}^{CHT} = A^{-1}\hat{L}^{HT}. \tag{5}$$

Basically, it consists in replacing L by $\hat{L}^{HT}$ in $T = A^{-1}L$.

Lemma 5. *Under Assumptions 1 and 2, we have that* $E\left(\hat{T}^{CHT}\right) = T$.

Proof. Under Assumptions 1 and 2 $\hat{L}^{HT}$ is unbiased , and given that A^{-1} is a non-stochastic matrix, $\hat{T}^{CHT}$ is also unbiased. □

The variance-covariance matrix of $\hat{T}^{CHT}$ is given by $\Omega_T = A^{-1}\Omega_L A'^{-1}$, our proposed estimator for Ω_T is given by

$$\hat{\Omega}_T = A^{-1}\hat{\Omega}_L A'^{-1}. \tag{6}$$

Lemma 6. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_T\right) = \Omega_T$.

Proof. Under Assumptions 1-3 $\hat{\Omega}_L$ is unbiased, and given that A^{-1} is a non-stochastic matrix, $\hat{\Omega}_T$ is also unbiased. □

4.1 Feasible estimator

We assume that we do not know p so the previous estimators are unfeasible because they depend on p (L and A^{-1} are of order p). As a special case, if we know p we use (5) and (6).

We assume that all we know about the size of the subsets γ is based on the sample, thus the largest size we observe is $p^s \leq p$. This fact could lead us to think that we can only estimate $L_{c\leq p^s}$, but this is not true because under Assumption 2 if we do not observe subsets of sizes $c > p^s$ is due to randomness, therefore $\hat{L}^{HT}_{c>p^s} = 0$ because $I\left[\#\gamma = c\right] = 0$ for $c > p^s$. The problem caused by not knowing p is that we do not know if there exist subsets of size $c > p^s$, and their estimates are zero.

Since the largest size of subsets we observe in the sample is p^s, then the estimate of the total number of sets of size larger than p^s is zero. Then $\hat{L}^{HT}_{c>p^s} = 0$ makes that $\hat{T}^{CHT}_{k>p^s} = 0$.

We write L, T and A as follows, $L = \left(L_A^{'}, L_B^{'}\right)^{'}$ where $L_A = (L_1, \ldots, L_{p^s})^{'}$ and $L_B = (L_{p^s+1}, \ldots, L_p)^{'}$, $T = \left(T_A^{'}, T_B^{'}\right)^{'}$ where $T_A = (T_1, \ldots, T_{p^s})^{'}$ and $T_B = (T_{p^s+1}, \ldots, T_p)^{'}$, and $A^{-1} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}$ where $\bar{A}_{11}$ is a $(p^s \times p^s)$ upper-left submatrix of A^{-1}, $\bar{A}_{12}$ is a $(p^s \times (p - p^s))$ upper-right submatrix of A^{-1}, $\bar{A}_{21} = [0]$ is a $((p - p^s) \times p^s)$ lower-left submatrix of A^{-1} and $\bar{A}_{22}$ is a $((p - p^s) \times (p - p^s))$ lower-right submatrix of A^{-1}.

Given that $T = A^{-1}L$ and the previous definitions, we have that $T_A = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} L$ and $T_B = \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} L$, given this we divide the estimator of T into two parts, on the one hand we estimate T_A and on the other hand we estimate T_B.

Using block matrix multiplication we have that,

$$\begin{aligned} \hat{T}_A^{CHT} &= \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} \hat{L}^{HT} \\ &= \begin{bmatrix} \bar{A}_{11}\hat{L}_A^{HT} & \bar{A}_{12}\ \hat{L}_B^{HT} \end{bmatrix} \\ &= \bar{A}_{11}\hat{L}_A^{HT}. \end{aligned} \tag{7}$$

The previous result is given by the fact that $\hat{L}_B^{HT} = 0$.

It is worth to note here that $\bar{A}_{11}$ is a submatrix of A^{-1}, if we need to know A^{-1} in order to get $\bar{A}_{11}$, the estimator $\hat{T}_A^{CHT}$ would be unfeasible because it would still depend on p. In appendix (10) we prove that $\bar{A}_{11} = A_{11}^{-1}$ where $A_{11} = A\left[1 : p^s,\, 1 : p^s\right]$, that is, $\bar{A}_{11}$ is the inverse of a submatrix of A given by its first p^s rows and columns, therefore, it is enough to know p^s to get $\bar{A}_{11}$.

In the same line

$$\begin{aligned} \hat{T}_B^{CHT} &= \begin{bmatrix} \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \hat{L}^{HT} \\ &= \begin{bmatrix} 0\hat{L}_A^{HT} & \bar{A}_{22}\ \hat{L}_B^{HT} \end{bmatrix} \\ &= 0. \end{aligned} \tag{8}$$

The previous result is given by $\bar{A}_{21} = 0$ and $\hat{L}_B^{HT} = 0$.

The key issue here is that, regardless of whether we know p, the estimate for T_B is equal to zero because its estimator depends entirely on $\hat{L}_B^{HT}$. In other words, the estimate of the total number of sets larger than p^s is zero, and the estimate of the total number of sets of sizes less than or equal to p^s is given by $\hat{T}_A^{CHT}$.

Summarizing, if we do not know p the only feasible estimator is $\hat{T}_A^{CHT}$. It does not mean we only have an estimator for T_A, we know that, if T_B exist, its (unbiased) estimate is zero.

Lemma 7. *Under Assumptions 1-2, we have that* $E\left(\hat{T}_A^{CHT}\right) = T_A$ *and* $E\left(\hat{T}_B^{CHT}\right) = T_B$.

Proof. Under Assumptions 1 and 2 $\hat{L}^{HT}$ is unbiased, and given that $\bar{A}_{11}$, $\bar{A}_{12}$, $\bar{A}_{21}$ and $\bar{A}_{22}$ are a non-stochastic matrices, $\hat{T}_A^{CHT}$ and $\hat{T}_B^{CHT}$ are also unbiased. □

Let Ω_{TA} be the variance-covariance matrix of $\hat{T}_A^{CHT}$, the estimator of Ω_{TA} is given by,

$$\hat{\Omega}_{TA} \quad = \quad A_{11}^{-1}\hat{\Omega}_{L,11}A_{11}^{-1\prime}, \tag{9}$$

where $\hat{\Omega}_{L,11} = \hat{\Omega}_L\left[1 : p^s,\ 1 : p^s\right]$ and $A_{11}^{-1} = \bar{A}_{11}$.

Let Ω_{TB} be the variance-covariance matrix of $\hat{T}_B^{CHT}$, the estimator of Ω_{TB} is given by $\hat{\Omega}_{TB} = [0]$ (we show this equality in Appendix 8.2.2).

Lemma 8. *Under Assumptions 1-3, we have that* $E\left(\hat{\Omega}_{TA}\right) = \Omega_{TA}$ *and* $E\left(\hat{\Omega}_{TB}\right) = \Omega_{TB}$.

Proof. See Appendix 8.2.2. □

4.2 Numerical Example

We present in this section a simple numerical example. Suppose we have a population of 10 houses grouped into 6 blocks in the following way: 3 blocks have 1 house ($T_1 = 3$), 2 blocks have 2 houses ($T_2 = 2$) and 1 block has 3 houses ($T_3 = 1$).

$U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$ is a population of $N = 10$ houses, $\Theta = \left\{\theta^1, \theta^2, \theta^3, \theta^4, \theta^5, \theta^6\right\}$ is a population of $J = 6$ blocks, where $\theta^1 = (u_1)$, $\theta^2 = (u_2)$, $\theta^3 = (u_3)$, $\theta^4 = (u_4, u_5)$, $\theta^5 = (u_6, u_7)$, $\theta^6 = (u_8, u_9, u_{10})$, the maximum size of sets (blocks) is $p = 3$.

The population of subsets is given by

$\Gamma = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, (u_4, u_5), (u_6, u_7), (u_8, u_9), (u_9, u_{10}), (u_8, u_{10}), (u_8, u_9, u_{10})\}$, the sizes of the elements of Γ are given by $c = 1, 2, 3$. Each element in Γ is a possible combination of elements of each set in Θ.

We draw a random sample S of size $n = 4$ from U. The probability of selection of γ_c is given by

$$Prob\left(\gamma_c \in S\right) = \begin{pmatrix} N \\ n \end{pmatrix}^{-1} \begin{pmatrix} N-c \\ n-c \end{pmatrix}.$$

Given that $c = 1, 2, 3$ we have that $Prob\left(\gamma_1 \in S\right) = \frac{n}{N}$, $Prob\left(\gamma_2 \in S\right) = \frac{n(n-1)}{N(N-1)}$ and $Prob\left(\gamma_3 \in S\right) = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}$.

Let $S = \{u_1, u_3, u_6, u_7\}$ be a sample of U, and let $\Gamma^s = \{u_1, u_3, u_6, u_7, (u_6, u_7)\}$be the sample of Γ that we get from S. The size of largest subset in Γ^s is $p^s = 2 < p$.

We first estimate $L = (L_A,\, L_B)$ by (4), where $L_A = (L_1,\, L_2)$ and $L_B = (L_3)$, we have that

$$\hat{L}_1^{HT} = 4\pi_{\gamma 1}^{-1} = 4\left(\frac{4}{10}\right)^{-1} = 10$$

and

$$\hat{L}_2^{HT} = 1\pi_{\gamma 2}^{-1} = 1\left(\frac{12}{90}\right)^{-1} = 7.5$$

with the previous we have that $\hat{L}_A^{HT} = (10,\, 7.5)^{'}$.

We are assuming we do not know p, so we neither know that L_B exist, that is, we do not know there are subsets of size $c = 3$ in Γ but, under assumption 2, we know that if there are subsets of size larger than $p^s = 2$ the estimate of their total number is zero, that is, $\hat{L}_3^{HT} = 0$. Assumption 2 establishes that any subset in Γ should have positive probability to be sampled, given the size of the largest subset in Γ is 3, as long as $n \geq 3$ Assumption 2 holds.

On the other hand we have $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$ so $A_{11}^{-1} = (A\,[1:2,\,1:2])^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}$. Given the previous we have that

$$\hat{T}_A^{CHT} \quad = \quad \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \hat{L}_A^{HT},$$

thus we have that $\hat{T}_A^{CHT} = (-5,\,7.5)^{'}$.

As we see, in theory some T_k could be negative, but in practice they should be positive. We can impute the value zero for every $\hat{T}_k^{CHT} < 0$, this generate some bias but reduce the Mean Square Error.

We do not have an estimate of T_3 because we do not know there are subsets of size 3, but we know that the estimate of every $T_{k>2}$ is zero (if it exist!).

5 Estimating the network degree distribution

In this section we use the methodology proposed above to estimate the network degree distribution, which is one of the most fundamental features of a network. Some of the papers related with this topic are the following.

5.1 Literature review

Frank (1980, 1981) shows that, under certain networks sampling designs, the expectation of the observed degree relative frequencies (i.e., the degree distribution) is a linear combination of the true degree relative frequencies, and he proposes an estimator which depend on an inverse matrix that in some cases is not invertible and, even when it is, the result may not be non-negative.

Zhang et al. (2015) proposes a method to overcome these problems, and apply it to a few common networks sampling designs where inclusion probabilities are known. They do not make assumptions about the structure of the network. The methodology is assessed by a simulation study that considers the effects of several factors on the accuracy of the estimators. The results show, among others aspects, that sampling schemes have a considerable impact on the performance of the estimators.

Thompson (2006) proposes a sampling design and discusses inference procedures on the average degree and the degree distribution under such sampling designs. In the same line, Ribeiro & Towsley (2012) studies the mean squared error associated with different sampling methods for the degree distribution.

Stumpf & Wiuf (2005) discusses two sampling schemes for selecting random subnets from a network and investigate how the degree distribution is affected for this two types of sampling. The central question addressed by the authors is whether the degree distribution of randomly sampled subnets has the same properties as the degree distribution of the overall network, they derive a necessary and sufficient condition that guarantees this equality and describe some situations under which this condition is satisfied, however, for the majority of the networks this condition is no be met.

Internet topology is one of the areas where networks are widely used. Faloutsos et al. (1999) shows that the internet degree distribution has a power-law form, Lakhina et al. (2003) show that when graphs are sampled using traceroute-like methods, the resulting degree distribution (based on sampled network) can differs sharply from the true, they explore the reason of such bias and propose a test for determining when sampling bias is present. Achlioptas et al. (2009) study the traceroute sampling systematically and extend the results in Lakhina et al. (2003).

5.2 Some network definitions

A common way to represent a network is listing all their nodes and links among them. Let $\mathcal{N} = \{1, 2, ..., N\}$ be a collection of N nodes and let g be a collection of links. Then a network is defined by $\mathcal{G} = (\mathcal{N}, g)$. If $ij \in g$, then node i is linked to node j, alternatively, we can use the notation $g_{ij} = 1$ if nodes i and j are linked, $g_{ij} = 0$ otherwise.

The relation among nodes can be directed or undirected. In directed networks if i is linked to j, j is not necessarily linked to i, that is, $g_{ij} \neq g_{ji}$ is possible. In undirected networks there exists reciprocity, and then, if i is linked to j, j is linked to i as well, $g_{ij} = g_{ji}$.

The degree of a node i, $d_i(\mathcal{G})$, is the number of links i has. The degree distribution of a network, $P_{\mathcal{G}}$, is a description of the relative frequencies of nodes that have different degrees. Let

$T_k = \sum_{i=1}^{N} I\left[d_i\left(\mathcal{G}\right) = k\right]$ the total number of nodes with degree k in $\mathcal{G}$, we have that T_k/N is the relative frequency of nodes with degree k in $\mathcal{G}$. Knowing T_k for all k, we can get the degree distribution of $\mathcal{G}$.

Let $S \subset N$ be a subset of nodes, we name $\mathcal{G}_S = (S,\, g|_S)$ to the network $\mathcal{G}$ restricted only to the set S. That is,

$$g|_{Sij} = \begin{cases} 1 & \text{if } i\,j \in S,\, g_{ij} = 1 \\ 0 & \quad\text{otherwise} \end{cases} .$$

In the network literature there exist multiple networks sampling designs. We only work with two of them known as *induced subgraph* and *incident subgraph*. In induced subgraph sampling a set of nodes is selected and then all edges among selected nodes are observed. In incident subgraph links are selected and then all nodes that correspond to selected links are observed. The selection can be done under simple random sampling or under Bernoulli sampling.

5.3 Estimating network degree distribution by CHT estimator

In this subsection we apply the CHT estimator for estimating the network degree distribution. The proposed estimator can be used under fairly general contexts, the main requirement being to have a probability sample of links.

As noted above the CHT estimator works in a particular framework. This framework is given by the existence of a population U whose elements are grouped into sets in Θ, the main goal of the CHT estimator is to estimate the total number of sets of a given size in Θ based on a sample of U. In order to apply the CHT estimator we need to adapt some networks concepts to the framework established in Sections 2 and 3.

5.3.1 Definitions

We begin by defining a network, a population of links and a sample of links.

Definition 7. Let $\mathcal{G} = (\mathcal{N},\, g)$ be a fixed finite network with $\mathcal{N} = \{1, 2, ..., N\}$ nodes. Let $U = g$ be a population given by the links of $\mathcal{G}$, and let $g|_S$ be a probability sample of U .

Definition 8. Let $\theta^i = \{g_{ij}|\, g_{ij} = 1,\, for\, j \in \mathcal{N}\}$ be the set of all links that i has, with $i = 1, \ldots, J$

and $j = 1, \dots, N$, where J is the total number of nodes in $\mathcal{G}$ with at least one link.

The *neighborhood* of a node i is the set of all nodes that i is linked to, $N_i(g) = \{j : g_{ij} = 1\}$. The set θ^i is conceptually close to $N_i(g)$, the latter contains ***the nodes*** linked to i while the former contains ***the links*** i has. It is important to note here that, for $d_i(\mathcal{G}) \neq 0$, $d_i(\mathcal{G}) = \#\theta^i$, the degree of the node i is equal to the total number of elements in θ^i.

Definition 9. Let $\Theta = \cup_{i=1}^{J} \theta^i$, be the collection of $\theta's$.[2]

Given the previous definitions, the total number of sets of size k in Θ is equal to the total number of nodes with degree k in $\mathcal{G}$, therefore $T_k = \sum_{i=1}^{J} I\left[\#\theta^i = k\right]$. Our main goal here is to estimate T_k.

Up to here we defined a population U whose elements (links) are grouped into sets which belong to Θ. At this point it is worth to remember that the methodology proposed in Section 3 allows to estimate the total number of sets of a given size based on a sample of elements from the sets instead of a sample of sets, so based on a sample of U (links) we can estimate the total number of sets of size k in Θ, what is no more than the total number of nodes with degree k in $\mathcal{G}$. In other words, based on a sample of links we can estimate the total number of nodes with degree k in $\mathcal{G}$.

Next we define an "artificial" population constructed from Θ.

Definition 10. Let $\Gamma = \cup_{\theta \in \Theta} \{P(\theta) - \emptyset\}$ be a collection of all possible subsets we can get from ***each element*** of Θ, where $P(\theta)$ is the power set of θ, γ is a generic element of Γ and γ_c a generic element of size c, with $c = \{1, 2, \dots, p\}$. The total number of elements of size c in Γ is given by $L_c = \sum_{\gamma \in \Gamma} I[\#\gamma = c]$.

The elements of θ are links then, given a node i, the collection Γ contains all combinations of one link of i, all combinations of two links of i, all combinations of three links, etc.[3], and this is so for all i.

Definition 11. Let $\Gamma^s = \{\gamma | \forall \gamma \in g|_S\}$ be the collection of all subsets γ in $g|_S$. Let $p^s \leq p$ be the maximum size of $\gamma \in g|_S$.

$\gamma \subseteq \theta$ are subsets of links, what we mean by $\gamma \in g|_S$ is that all links in γ are in $g|_S$. By now we do not say anything about $g|_S$, it is only a probability sample of links.

[2] We define Θ as a collection because their elements are not necessarily exclusive.

[3] Assuming that $d_i(G) \geq 3$.

Definition 12. Let $\pi_\gamma = Prob\left(\gamma \in \Gamma^s\right)$ be the probability of selection of γ and $\pi_{\gamma\gamma'} = Prob\left(\gamma \cup \gamma' \in \Gamma^s\right)$ be the joint probability of selection of γ and γ'. Given definition 11, $\pi_\gamma = Prob\left(\gamma \in g|_S\right)$ and $\pi_{\gamma\gamma'} = Prob\left(\gamma \cup \gamma' \in g|_S\right)$.

Given that γ is a subset of links, the probability of selection of γ is the joint probability of selection of such links.

5.3.2 Assumptions

We are going to use the CHT estimator for estimating the network degree distribution. The unbiasedness and feasibility of the estimator require assumptions 1 and 2 hold, and the unbiasedness of the variance estimator requires assumptions 1-3 hold. We present below these assumptions adapted to the network context.

First we assume some mild conditions under which Assumption 1 holds.

Assumption 4. *Let $g|_S$ be a probability sample of links from $\mathcal{G}$. For all $[g|_S]_{ij} = 1$ we observe i.*

Let $[g|_S]_{ij}$ be a sampled link between i and j. Assumption (4) establishes we know such link belongs to node i. This requirement makes feasible to group the sampled links as they are grouped in θ so we have a sample of subsets of θ. The "artificial" population Γ contains all possible subsets of every θ, therefore we have a sample of Γ, that is Γ^s.

The next two assumption refer to the individual and joint probability of selection of γ.

Assumption 5. $\pi_\gamma > 0 \; \forall \gamma \in \Gamma$. *The probability of selection of any element in Γ has to be positive.*

Assumption 6. $\pi_{\gamma\gamma'} > 0 \; \forall \gamma \vee \gamma' \in \Gamma$. *The joint probability of selection of any pair of elements in Γ has to be positive.*

Assumptions 5 and 6 are equivalent to assumptions 2 and 3. The elements in Γ are subsets of links that belong to the same node, therefore we need that any set of links that belongs to the same node has positive probability to be sampled (individual probability), and not only that but also we need that any combination of two sets of links has positive probability to be sampled (joint probability).

5.3.3 Estimator

Let $\tilde{T} = \left(T_0, T^{'}\right)^{'}$ be a $((p+1)\times 1)$ vector with $T = (T_1, T_2, \ldots, T_p)^{'}$, where T_0 and T_k are the total number of nodes with degree zero and k in $\mathcal{G}$ respectively[4], therefore we have that $P_{\mathcal{G}} = N^{-1}\tilde{T}$. Following we present an estimator for $P_{\mathcal{G}}$ based on the methodology proposed in the previous section.

Based on definitions (7), (8) and (9) we have a population of links named U whose elements are grouped into sets which belongs to Θ. We know that the total number of sets of size $0 < k \leq p$ in Θ is equal to the total number of links with degree k in $\mathcal{G}$, that is, T_k. We propose to use the methodology presented above for estimating T_k.

An important point to note here is we assume p is unknown, that is, we do not know the maximum degree in $\mathcal{G}$, therefore, following the exposed in section (4.1), we only going to have estimates for $T_{k\leq p^s}$, but it does not mean we can only estimate T partially, because we know that the estimates for $T_{k>p^s}$ are zero.

Being L_c the total number of subsets of size c in Γ, the HT estimator for L_c is given by

$$\hat{L}_c^{HT} \quad = \quad \sum_{\gamma\in\Gamma^s} \pi_\gamma^{-1} I\left[\#\gamma = c\right]. \tag{10}$$

Given we are assuming p is unknown we propose to use the estimator (7) for estimating T as follows

$$\hat{T}_A^{CHT} \quad = \quad \bar{A}_{11}\hat{L}_A^{HT}, \tag{11}$$

where $\hat{L}_A^{HT} = \left(\hat{L}_1^{HT}, \ldots, \hat{L}_{p^s}^{HT}\right)^{'}$ and $\hat{T}_A^{CHT} = \left(\hat{T}_1^{CHT}, \ldots, \hat{T}_{p^s}^{CHT}\right)^{'}$.

Despite the previous estimator only estimates $T_{k\leq p^s}$, we know that, if there exist nodes with degree larger than p^s the estimate of their total number is zero.

Up to here we estimate $T_{k\geq 1}$, we do not have an estimator for T_0, that is, we do not have an estimator for the total number of nodes without connections, but given that $T_0 = N - \sum_{k=1}^{p} T_k$, assuming we know N we can estimate T_0 by $\hat{T}_0^{CHT} = N - 1_{p^s}\hat{T}_A^{CHT}$ where 1_{p^s} is a $(1\times p^s)$ vector

[4] We divide the vector $\tilde{T}$ into T_0 and T because our methodology allows to estimate T, that is, the total number of nodes with degree $k > 0$. For estimating the total number of nodes with degree zero, T_0, we present another estimator.

of ones. Here we are assuming we know N, but if we do not, N should be replaced by any unbiased estimator.

Our proposed estimator for the network degree distribution is given by

$$\hat{P}_{\mathcal{G}}^{CHT} = N^{-1}\hat{\tilde{T}}_{A}^{CHT}, \tag{12}$$

where $\hat{\tilde{T}}_{A}^{CHT} = \left(\hat{T}_{0}^{CHT},\, \hat{T}_{A}^{CHT\prime}\right)^{\prime}$.

Given we are assuming p is unknown, we do not know if there are nodes with degree greater than p^{s}, but we know if they exist the estimate of their total number is zero, so the estimate of the relative frequencies of nodes with degree greater than p^{s} is zero.

The $Var\left(\hat{T}_{0}^{CHT}\right) = Var\left(N\right) + Var\left(1_{p^{s}}\hat{T}_{A}^{CHT}\right) - 2Cov\left(N,\, 1_{p^{s}}\hat{T}_{A}^{CHT}\right)$, and given that N is not random, we have that $Var\left(\hat{T}_{0}\right) = Var\left(1_{p^{s}}\hat{T}_{A}^{CHT}\right)$. The estimator of $Var\left(\hat{T}_{0}^{CHT}\right)$ is given by $\widehat{Var}\left(\hat{T}_{0}^{CHT}\right) = 1_{p^{s}}\hat{\Omega}_{TA}1_{p^{s}}^{\prime}$.

The variance estimator of $\hat{P}_{\mathcal{G}}^{CHT}$ is given by

$$\widehat{Var}\left(\hat{P}_{\mathcal{G}}^{CHT}\right) = N^{-2}\left(\widehat{Var}\left(\hat{T}_{0}^{CHT}\right),\, Diag\left(\hat{\Omega}_{TA}\right)\right).$$

5.4 Estimating degree distribution under induced and incident subgraph sampling designs. Probabilities

All of the previous developments are valid for any probability sampling design. Here, we derive the probabilities of selection under induced subgraph sampling for directed networks. The extension to undirected networks is straightforward.

5.4.1 Probability of selection under induced subgraph sampling

Under induced subgraph sampling we draw a sample of nodes and we know their relation (we know the links among sampled nodes), therefore we have to construct the probability of selection of links based on the probability of selection of nodes.

The probabilities of selection of a link is given by the joint probability of selection of the two nodes involved in such link. It is important to remember that γ is a subset of links that belong to the same node, therefore all links in γ have a common node, thus the total number of nodes involved in γ is equal to the total number of links in γ plus one. Thus, the probability of selection of γ_c is given by the joint probability of selection of the $(c+1)$ nodes involved in γ_c.

$$Prob\left(\gamma_c \in g|_S\right) \; = \; \begin{pmatrix} N \\ n \end{pmatrix}^{-1} \begin{pmatrix} N-(c+1) \\ n-(c+1) \end{pmatrix},$$

where n is the number of sampled nodes, $\begin{pmatrix} N \\ n \end{pmatrix}$ are all possible samples of size n, and $\begin{pmatrix} N-(c+1) \\ n-(c+1) \end{pmatrix}$ are all possible samples of size n which contain all the nodes involved in the c links of γ.

On the other hand, $\pi_{\gamma\gamma'}$ is the joint probability of selection of γ and γ', this is equal to the joint probability of selection of all nodes involved in the links of γ and γ'. The nodes involved in the links of γ and γ' could be repeated, for instance, being g_{ij} a link in γ, the link g_{ji} could be in γ', thus both links are formed by the same two nodes.

Let $\omega_{\gamma\gamma'} = \{i, j|\, g_{ij} \in \gamma \cup \gamma'\}$ be a set of all nodes (without repetitions)[5] involved in the links of γ and γ', we have

$$Prob\left(\left(\gamma_c \cup \gamma'_{c'}\right) \in g|_S\right) \; = \; \begin{pmatrix} N \\ n \end{pmatrix}^{-1} \begin{pmatrix} N-\omega_{\gamma\gamma'} \\ n-\omega_{\gamma\gamma'} \end{pmatrix}.$$

In case the links involved in γ_c and $\gamma'_{c'}$ do not have nodes in common, $\#\omega_{\gamma\gamma'} = (2c+2)$, and in case the links involved in γ and γ' have all nodes in common, $\#\omega_{\gamma\gamma'} = (c+1)$.

In order to the estimator (12) be unbiased it is necessary that $\pi_\gamma > 0$ for all γ and this happens as long as $n \geq (c+1)$ for all c, and given $c = \{1, 2, \ldots, p\}$ this implies that the sample size should be, at least, equal to the number of nodes involved in the links of γ_p, that is, $p+1$.

Similarly, in order to the estimator of the variance be unbiased, in addition to the previous assumption on π_γ, it is necessary to assume that $\pi_{\gamma\gamma'} > 0$ for allγ and γ', and this happens as long

[5] Since $\omega_{\gamma\gamma'}$ is an union of nodes, if there are some common nodes among the links in γ and γ' they will appear only once in $\omega_{\gamma\gamma'}$.

as $n \geq \#\omega_{\gamma\gamma'}$ for all γ and γ'. This implies that the sample size should be, at least, equal to the number of nodes involved in the links of all pairs of subsets γ and γ', therefore, in order to $\pi_{\gamma\gamma'} > 0$ we need that $n \geq max\left(\#\omega_{\gamma\gamma'}\right)$.

6 Monte Carlo simulations

In order to evaluate the performance of the CHT estimator for degree distribution we present Monte Carlo simulations.

We simulate the coordinates of N points (nodes) by two $U(0,1)$ independent random variables, and then we establish links among points following the p-nearest neighbors criteria, and construct the corresponding adjacency matrix. The networks created in this way ensure that each node has degree p (where p is the number of neighbors), i.e., the so called “regular networks” (networks whose nodes have all the same degree).

Second, in order to construct networks whose nodes have different degrees we randomly delete some links by taking rows and columns from the adjacency matrix and filling them with zeros. We randomly select the id’s of the rows and columns to fill with zeros drawing two random samples from $\{1,\ldots,N\}$. We draw one sample of size $\lceil\alpha_1 N\rceil$ for selecting rows and other of size $\lceil\alpha_2 N\rceil$ for selecting columns, with $0 < \{\alpha_1, \alpha_2\} < 1$. In this way we have networks whose nodes have different degrees but none of them are greater than p.

We consider networks of two different sizes, $N = 300, 600$, whereas the sampling rates are $\{0.9, 0.7, 0.5\}$. The maximum degree is fixed in $p = 4$, so the degree distribution has domain in $k = \{0, 1, \ldots, 4\}$.

We want to ensure that the relative frequencies of nodes with degree k are not smaller than 0.1, that is, we want to ensure that $P_{\mathcal{G}}\left(d = k\right) > 0.1$ for all k. For doing that we fix $\alpha_1 = 0.1$ and $\alpha_2 = 0.4$.

Given a sample size, we generate one network and draw $R = 500$ samples for each sampling rate. The networks are the same at each repetition for all sampling rates.

For induced subgraph sampling we randomly choose n numbers from $\{1,\ldots,N\}$ and select the nodes that match with such numbers. Under this sampling design we know all links among sampled nodes, so we can construct an adjacency matrix for the sampled nodes. Once we have the links and the probability of selection (previously presented) we construct the CHT estimator and its variance estimator.

We group all the links (within the sample) incident to a given node and then we construct all the different subsets we can make based on it. We do this for each sampled node, these subsets of links are the subsets γ in $g|_S$.

We assume that we do not know p, so the estimator for $T_{k>p^s}$ is zero. We take into account such zeros when we construct the estimators.

We report the sampling rates (n/N). For each Monte Carlo exercise we report the true degree distribution, $P_{\mathcal{G}}\left(d=k\right)$ and the true average degree , $ave.P_{\mathcal{G}}$. We also report the estimates for the average degree. Then we report the bias and root-mean squared error (RMSE), that is,

$$Bias\left(d=k\right)=\frac{1}{R}\sum_{r=1}^{R}\left(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-P_{\mathcal{G}}\left(d=k\right)\right),$$

$$Bias\left(ave\right)=\frac{1}{R}\sum_{r=1}^{R}\left(ave.\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-ave.P_{\mathcal{G}}\left(d=k\right)\right),$$

$$RMSE\left(d=k\right)=\left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right)-P_{\mathcal{G}}\left(d=k\right)\right)^{2}\right]^{1/2},$$

$$RMSE\left(ave\right)=\left[\frac{1}{R}\sum_{r=1}^{R}\left(ave.\hat{P}_{\mathcal{G}}^{r}-ave.P_{\mathcal{G}}\right)^{2}\right]^{1/2},$$

We also report the average of the estimated standard deviations, $sd\left(d=k\right)=\frac{1}{R}\sum_{r=1}^{R}\hat{sd}(\hat{P}_{\mathcal{G}}^{r}\left(d=k\right))$ and $sd(ave)=\frac{1}{R}\sum_{r=1}^{R}\hat{sd}(ave.\hat{P}_{\mathcal{G}}^{r})$ in order to assess the proposed estimator of the variance.

We compare our proposed estimator with the "naive" one, where the in-sample degree is calculated. We also report the bias and RMSE for this case.

Tables 2 and 3 report the simulation results for $N=300$ and $N=600$ network sizes, respectively.

Table 2: Induced subgraph sampling $N = 300$.

		CHT estimator			Naïve estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_G(0) = 0.120$	0.000	0.011	0.011	0.005	0.011
	$P_G(1) = 0.163$	0.000	0.021	0.022	0.023	0.027
	$P_G(2) = 0.293$	0.000	0.030	0.031	-0.010	0.021
	$P_G(3) = 0.283$	-0.001	0.036	0.035	-0.061	0.064
	$P_G(4) = 0.140$	0.001	0.025	0.026	-0.056	0.058
	Average degree 2.16	0.000	0.047	0.049	-0.213	0.219
0.7	$P_G(0) = 0.120$	0.000	0.030	0.029	0.022	0.027
	$P_G(1) = 0.163$	0.000	0.074	0.071	0.047	0.052
	$P_G(2) = 0.293$	-0.007	0.118	0.111	-0.079	0.082
	$P_G(3) = 0.283$	0.006	0.125	0.112	-0.173	0.174
	$P_G(4) = 0.140$	0.000	0.068	0.062	-0.116	0.117
	Average degree 2.16	0.008	0.104	0.103	-0.643	0.647
0.5	$P_G(0) = 0.120$	0.005	0.080	0.079	0.040	0.044
	$P_G(1) = 0.163$	-0.017	0.260	0.246	0.021	0.029
	$P_G(2) = 0.293$	0.032	0.440	0.388	-0.176	0.178
	$P_G(3) = 0.283$	-0.033	0.417	0.331	-0.248	0.249
	$P_G(4) = 0.140$	0.013	0.118	0.123	-0.135	0.135
	Average degree 2.16	0.000	0.183	0.175	-1.082	1.087

Table 3: Induced subgraph sampling $N = 600$.

		CHT estimator			Naïve estimator	
n/N	Degree distribution	Bias	RMSE	std.dev.	Bias	RMSE
0.9	$P_G(0) = 0.131$	0.000	0.007	0.007	0.001	0.007
	$P_G(1) = 0.136$	0.000	0.015	0.015	0.029	0.031
	$P_G(2) = 0.286$	-0.002	0.024	0.025	0.000	0.015
	$P_G(3) = 0.336$	0.001	0.021	0.023	-0.086	0.087
	$P_G(4) = 0.108$	0.000	0.015	0.015	-0.044	0.045
	Average degree 2.15	0.000	0.034	0.035	-0.212	0.217
0.7	$P_G(0) = 0.131$	0.000	0.020	0.020	0.013	0.017
	$P_G(1) = 0.136$	0.000	0.052	0.052	0.065	0.067
	$P_G(2) = 0.286$	0.002	0.085	0.083	-0.063	0.066
	$P_G(3) = 0.336$	-0.002	0.073	0.073	-0.224	0.225
	$P_G(4) = 0.108$	0.000	0.039	0.040	-0.089	0.090
	Average degree 2.15	0.000	0.069	0.073	-0.645	0.649
0.5	$P_G(0) = 0.131$	0.000	0.058	0.056	0.028	0.031
	$P_G(1) = 0.136$	0.000	0.179	0.174	0.046	0.049
	$P_G(2) = 0.286$	0.010	0.272	0.260	-0.167	0.168
	$P_G(3) = 0.336$	-0.011	0.221	0.203	-0.301	0.302
	$P_G(4) = 0.108$	0.003	0.098	0.079	-0.104	0.104
	Average degree 2.15	0.000	0.128	0.124	-1.077	1.081

The simulation results highlight that the naive estimator of the degree distribution that uses the sampled network as if it were the true one produces considerable bias in the estimation of both, the degree distribution and the average degree. The bias increases when the sampling rate decreases, and the same problem arises for both types of sampling designs. Our proposed CHT estimator, has a small bias in all cases. The CHT estimator also has a good performance in terms of RMSE.

The average of the estimated standard deviation is close to the simulated RMSE, and this suggests the estimator of the variance performs well in small samples.

With respect to the sampling rates and the population sizes the results are expected. The greater the sampling rate is, the better the estimators behave, and in the same line, given a sampling rate, the greater the population is, the better the estimators behave.

7 Conclusion

This paper presents a general methodology for estimating the total number of sets of a given size based on a sample of elements of such sets. The resulting estimator is the chained Horvitz-Thompson (CHT) estimator.

We apply such methodology to derive an estimator for the degree distribution (and its variance) to be used under probability sampling designs, and we adapt it to a widely used sampling design known as induced subgraph. The results obtained from the Monte Carlo simulations show that the estimator perform well in terms of bias and RMSE, in line with the analytical results.

This is a basal paper, and there are many interesting lines in which to continue working. It would be worth to derive the asymptotic distribution of the CHT estimator. This should be relatively straightforward because the CHT estimator is no more than a linear combination of the HT estimator, and under some regularity conditions such estimator is asymptotically normal.

It would also be interesting to adapt the degree distribution estimator to other probability sampling designs and to evaluate them under other schemes. In the same line, more structure could be added to the degree distribution estimator. By assuming some particular degree distribution (scale-free, Poisson, etc.) the results may improve.

8 Appendix

8.1 Some properties of A and A^{-1}

The $((p+1)\times(p+1))$ Pascal Matrix $P=[p_{ij}]$ is defined by $p_{ij}=\begin{pmatrix} j \\ i \end{pmatrix}$ with $i,\, j=0,1,\ldots,p$.

The inverse $P^{-1}=\left[\bar{p}_{ij}^{-1}\right]$ is defined by $\bar{p}_{ij}^{-1}=(-1)^{j+i}\begin{pmatrix} j \\ i \end{pmatrix}$.

Lemma 9. *Let $A=P\,[1:p,\,1:p]$ be a $(p\times p)$ submatrix of P, then $A^{-1}=P^{-1}\,[1:p,\,1:p]$.*

Let $P_n=\begin{bmatrix} B & C \\ D & A \end{bmatrix}$, where $B_{(1\times1)}=[1]$, $C_{(1\times p)}=[1,\ldots,1]$ and $D_{(p\times1)}=[0,\ldots,0]^{'}$. By block matrix inversion we have that

$$\begin{bmatrix} B & C \\ D & A \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1}+B^{-1}C\left(A-DB^{-1}C\right)^{-1}DB^{-1} & -B^{-1}C\left(A-DB^{-1}C\right)^{-1} \\ -\left(A-DB^{-1}C\right)^{-1}DB^{-1} & \left(A-DB^{-1}C\right)^{-1} \end{bmatrix}$$

Given $D_{(p\times1)}=[0,\ldots,0]^{'}$ we have

$$\begin{bmatrix} B & C \\ D & A \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1} & -CA^{-1} \\ D & A^{-1} \end{bmatrix}$$

The inverse of B and A exists.

Given the previous we have that $P^{-1}\,[1:p,\,1:p]=A^{-1}$, and thus $\bar{a}_{ij}^{-1}=(-1)^{j+i}\begin{pmatrix} j \\ i \end{pmatrix}$ with $i,\, j=1,\ldots,p$.

The matrix P has a well known inverse, and given that A is a submatrix of P, we proved above that the inverse of$A$ is a submatrix of the inverse of P.

Lemma 10. *Let $A_{11}=A\,[1:p^s,\,1:p^s]$ be a $(p^s\times p^s)$ upper-left submatrix of A with $p^s\leq p$, then $A_{11}^{-1}=A^{-1}\,[1:p^s,\,1:p^s]$.*

$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ where A_{11} is a $(p^s\times p^s)$ upper-left submatrix of A, A_{12} is a $(p^s\times(p-p^s))$ upper-right submatrix of A, $A_{21}=[0]$ is a $((p-p^s)\times p^s)$ lower-left submatrix of A and A_{22} is a $((p-p^s)\times(p-p^s))$ lower-right submatrix of A.

By block matrix inversion we have that

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} \left(A_{22} - A_{21} A_{11}^{-1} A_{12}\right)^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} \left(A_{22} - A_{21} A_{11}^{-1} A_{12}\right)^{-1} \\ -\left(A_{22} - A_{21} A_{11}^{-1} A_{12}\right)^{-1} A_{21} A_{11}^{-1} & \left(A_{22} - A_{21} A_{11}^{-1} A_{12}\right)^{-1} \end{bmatrix},$$

given $A_{21} = [0]$ we have

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1} A_{12} A_{22}^{-1} \\ A_{21} & A_{22}^{-1} \end{bmatrix},$$

A_{11} and A_{22} are unitriangular matrices so their inverses exist.

Given the previous we have that $A^{-1}[1:p^s,\, 1:p^s] = A_{11}^{-1}$, that is, the inverse ofA_{11} is the $(p^s \times p^s)$ upper-left submatrix of A^{-1}. It is so relevant because this implies we do not need to know A to get A_{11}^{-1}

8.2 Variance and covariance

As we saw, the estimator $\hat{L}_c^{HT}$ can be expressed as $\hat{L}_c^{HT} = \sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] I_\gamma$, the first indicator select the elements γ of size c in Γ and the second indicator identifies which are in the sample. In the same way we have $\hat{L}_{c'}^{HT} = \sum_{\gamma' \in \Gamma} \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] I_{\gamma'}$.

$$\begin{aligned} Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) &= Cov\left(\sum_{\gamma \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] I_\gamma,\, \sum_{\gamma' \in \Gamma} \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] I_{\gamma'}\right) \\ &= \sum_{\gamma \in \Gamma} \sum_{\gamma' \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] Cov\left(I_\gamma,\, I_{\gamma'}\right) \\ &= \sum_{\gamma \in \Gamma} \sum_{\gamma' \in \Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left[E\left(I_\gamma I_{\gamma'}\right) - E\left(I_\gamma\right) E\left(I_{\gamma'}\right)\right]. \end{aligned}$$

The $E\left(I_\gamma\right) = \pi_\gamma$, the expectation of γ be sampled is equal to the probability of γ be sampled, and the same is for $E\left(I_{\gamma'}\right) = \pi_{\gamma'}$. On the other hand $E\left(I_\gamma I_{\gamma'}\right) = \pi_{\gamma\gamma'}$, where $\pi_{\gamma\gamma'}$ is the joint probability of selection of γ and γ'.

Then using the previous results we have

$$Cov\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_\gamma \pi_{\gamma'}\right).$$

8.2.1 Variance-covariance estimator for L_c

The covariance estimator proposed in (3.4) can be written as

$$\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right) \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_\gamma \pi_{\gamma'}\right) \pi_{\gamma\gamma'}^{-1} \left(I_\gamma I_{\gamma'}\right),$$

where $\left(I_\gamma I_{\gamma'}\right)$ is an indicator function that takes the value one if $\gamma \wedge \gamma' \in \Gamma^s$ and zero otherwise. With this we have

$$E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] \;=\; \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma} \pi_\gamma^{-1} I\left[\#\gamma = c\right] \pi_{\gamma'}^{-1} I\left[\#\gamma' = c'\right] \left(\pi_{\gamma\gamma'} - \pi_\gamma \pi_{\gamma'}\right) \pi_{\gamma\gamma'}^{-1} E\left(I_\gamma I_{\gamma'}\right),$$

with $E\left(I_\gamma I_{\gamma'}\right) = \pi_{\gamma\gamma'}$, and thus $E\left[\widehat{Cov}\left(\hat{L}_c^{HT},\, \hat{L}_{c'}^{HT}\right)\right] = Cov\left(\hat{T}_c^{HT},\, \hat{T}_{c'}^{HT}\right)$.

8.2.2 Variance-covariance matrix estimator for $\hat{T}_A^{CHT}$ and $\hat{T}_B^{CHT}$

By equation (7), the variance-covariance matrix of $\hat{T}_A^{CHT}$ is given by $\Omega_{TA} = \left[\begin{array}{cc} \bar{A}_{11} & \bar{A}_{12} \end{array}\right] \Omega_L \left[\begin{array}{cc} \bar{A}_{11} & \bar{A}_{12} \end{array}\right]'$. Replacing Ω_L by $\hat{\Omega}_L$ we have the estimator for Ω_{TA} . Under assumptions 1-3 $\hat{\Omega}_L$ is unbiased, therefore, $\hat{\Omega}_{TA}$ it is also.

By equation (8), the variance-covariance matrix of $\hat{T}_B^{CHT}$ is given by $\Omega_{TB} = \left[\begin{array}{cc} \bar{A}_{21} & \bar{A}_{22} \end{array}\right] \Omega_L \left[\begin{array}{cc} \bar{A}_{21} & \bar{A}_{22} \end{array}\right]'$. Replacing Ω_L by $\hat{\Omega}_L$ we have the estimator for Ω_{TB} . Under assumptions 1-3 $\hat{\Omega}_L$ is unbiased, therefore, $\hat{\Omega}_{TB}$ it is also.

To show that the previous expressions can be reduced we divide $\hat{\Omega}_L$ into four submatrix as follows, $\hat{\Omega}_{L,11} = \hat{\Omega}_L\left[1:p^s,\, 1:p^s\right]$, $\hat{\Omega}_{L,12} = \hat{\Omega}_L\left[1:p^s,\, (p^s+1):p\right]$, $\hat{\Omega}_{L,21} = \hat{\Omega}_L\left[(p^s+1):p,\, 1:p^s\right]$ and $\hat{\Omega}_{L,22} = \hat{\Omega}_L\left[(p^s+1):p,\, (p^s+1):p\right]$.

Given that $\hat{\sigma}_{cc'} = 0$ for $c \wedge c' > p^s$ we have that $\hat{\Omega}_{L,12} = [0]$, $\hat{\Omega}_{L,21} = [0]$ and $\hat{\Omega}_{L,22} = [0]$, therefore

$$\begin{aligned}\hat{\Omega}_{TA} &= \left[\begin{array}{cc}\bar{A}_{11} & \bar{A}_{12}\end{array}\right]\left[\begin{array}{cc}\hat{\Omega}_{L,11} & 0\\ 0 & 0\end{array}\right]\left[\begin{array}{cc}\bar{A}_{11} & \bar{A}_{12}\end{array}\right]'\\ &= A_{11}^{-1}\hat{\Omega}_{L,11}A_{11}^{-1\prime}.\end{aligned}$$

given that $\bar{A}_{11} = A_{11}^{-1}$.

$$\begin{aligned}\hat{\Omega}_{TB} &= \left[\begin{array}{cc}\bar{A}_{21} & \bar{A}_{22}\end{array}\right]\left[\begin{array}{cc}\hat{\Omega}_{L,11} & 0\\ 0 & 0\end{array}\right]\left[\begin{array}{cc}\bar{A}_{21} & \bar{A}_{22}\end{array}\right]'\\ &= \bar{A}_{21}\hat{\Omega}_{L,11}\bar{A}_{21}'\end{aligned}$$

And given $\bar{A}_{21} = [0]$ this implies $\hat{\Omega}_{TB} = [0]$.

where $A^{-1}\left[1 : p^s,\, 1 : p\right] = \left[A_R^{-1},\, D\right]$ with $D = A^{-1}\left[1 : p^s,\, (p^s + 1) : p\right]$, $\hat{\Omega}_L = \left[\begin{array}{cc}\hat{\Omega}_{LR} & 0\\ 0 & 0\end{array}\right]$ with $\hat{\Omega}_{LR} = \hat{\Omega}_L\left[1 : p^s,\, 1 : p^s\right]$.

The same as before, $\hat{\Omega}_{LR}$ disregards the terms equal to zero in $\hat{\Omega}_L$, that is $\hat{\sigma}_{cc'}$ for $c \wedge c' > p^s$. Providing that $\hat{\sigma}_{cc'}$ is unbiased for all c and c' (regardless some of them are equal to zero) $\hat{\Omega}_{TR}$ is also unbiased.

References

Achlioptas, D., Clauset, A., Kempe, D., & Moore, C. (2009). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)*, 56(4), 21.

Ahmed, N. K., Neville, J., & Kompella, R. (2014). Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2), 7.

Bernstein, D. S. (2005). *Matrix mathematics: Theory, facts, and formulas with application to linear systems theory*, volume 41. Princeton University Press Princeton.

Brawer, R. & Pirovino, M. (1992). The linear algebra of the pascal matrix. *Linear Algebra and Its Applications*, 174, 13–23.

Chandrasekhar, A. & Lewis, R. (2011). Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29 (pp. 251–262).: ACM.

Frank, O. (1977). Estimation of graph totals. *Scandinavian Journal of Statistics*, (pp. 81–89).

Frank, O. (1980). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1), 45–50.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological methodology*, 12, 110–155.

Fuller, W. A. (2011). *Sampling statistics*, volume 560. John Wiley & Sons.

Granovetter, M. (1976). Network sampling: Some first steps. *American Journal of Sociology*, (pp. 1287–1303).

Handcock, M. S. & Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1), 5.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260), 663–685.

Jackson, M. O. et al. (2008). *Social and economic networks*, volume 3. Princeton university press Princeton.

Jackson, M. O., Rodriguez-Barraquer, T., & Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *The American Economic Review*, (pp. 1857–1897).

Kolaczyk, E. D. (2009). *Statistical analysis of network data: methods and models*. Springer Science & Business Media.

Lakhina, A., Byers, J. W., Crovella, M., & Xie, P. (2003). Sampling biases in ip topology measurements. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1 (pp. 332–341).: IEEE.

Ribeiro, B. & Towsley, D. (2012). On the estimation accuracy of degree distributions from graph sampling. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* (pp. 5240–5247).: IEEE.

Rothenberg, R. (1995). Commentary: Sampling in social networks. *Connections*, 18, 104–110.

Santos, P. & Barrett, C. (2008). What do we learn about social networks when we only sample individuals. *Not Much. Department of Applied Economics and Management Working Paper Cornell University*.

Stumpf, M. P. & Wiuf, C. (2005). Sampling properties of random graphs: the degree distribution. *Physical Review E*, 72(3), 036118.

Thompson, S. K. (2006). Adaptive web sampling. *Biometrics*, 62(4), 1224–1234.

Yates, F. & Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 253–261).

Zhang, Y., Kolaczyk, E. D., Spencer, B. D., et al. (2015). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *The Annals of Applied Statistics*, 9(1), 166–199.

Capítulo II

Regional and state heterogeneity of monetary shocks in Argentina[1]

Abstract

This paper empirically investigates how economic activity in Argentina at regional and provincial (i.e., state) levels responds to central national monetary policy shocks, as given by a change in the interest rate. Regional heterogeneity of monetary shocks exists in Argentina. At the regional level the long-term effects of increasing the interest rate are negative and statistically significant. At the provincial level, 11 provinces show a negative and significant impact of a shock on the interest rate over employment. However, there are 13 provinces in which the effect is not statistically significant, including the City of Buenos Aires and Buenos Aires Province.

Keywords: *Monetary policy, Monetary transmission, Regional Effects*
JEL classification: E52, G21, R11, R12

[1] Una versión extendida de este artículo fue publicada en *The Journal of Economic Asymmetries* 20 (2019): e00129, en coautoria con Emilio Blanco, Pedro Elosegui y Gabriel Montes Rojas.

1 Introduction

As noted by Carlino and DeFina (1999) the idea that policy changes affect states differently is intuitive given the heterogeneity of state economies and their financial and trade networks. State heterogeneity in a state's response to U.S. Federal Reserve Board actions can be deduced from traditional and new credit-based theories (Bernanke and Blinder, 1988; Kashyap, Stein, and Wilcox, 1993; Kashyap and Stein, 1994) of the monetary policy transmission mechanism. Park and Hewings (2012) found that industry mix and even more critically the place in the value chain production contributed to the asymmetries. In the business literature, the notion of a whipsaw effect has been introduced to show how economies with production systems at the early stages of a value chain experience greater fluctuations that those whose production is close to final goods. Further, the latter economies' business cycles will be more highly correlated with the national ones than the former economies.

As a result, it is important to account for feedback effects among regions when modeling regional responses to aggregate shocks, and policymakers actions should take into account potential extreme or unexpected effects in some regions. The simple estimation of a standard vector autoregressive (VAR) for each region, as is being done in empirical macroeconomic and monetary studies may result in serious misspecification since indirect effects of policy actions (operating, for instance, through trade and financial linkages among regions) are neglected. See the literature review in Dominguez-Torres and Hierro (2018) for a recent discussion of different models implemented in this context, and empirical evidence for the U.S., Europe and few other countries.

This paper is the first to empirically investigate how economic activity in Argentina at regional and provincial (i.e., state) levels responds to central or national monetary policy shocks. To do this we implement different spatial macro-type structural vector autoregressive (SVAR) models where we study how a change in the interest rate (i.e., monetary policy shock) affect employment in regions or provinces within Argentina, taking into account the spatial correlations among them. We thus evaluate the short-, medium- and long-term effects of monetary shocks on Argentine regions by computing the impulse response functions.

We find that regional heterogeneity exists in Argentina, resulting in differential effects of monetary policy shocks. At the regional level it is interesting to note that the North-East (NEA) region is the only one that does not show a significant impact of the shock on the interest rate on employment. In all other cases, the results are statistically significant, showing that a tightening of the monetary policy results in a negative effect on employment. Ciudad Autónoma de Buenos Aires (CABA) and Great Buenos Aires (GBA) together with the Centro region show a similar behavior to that of the national aggregate. Meanwhile, the Sur, North-West (NOA) and Cuyo regions show the largest negative effect on regional employment. At the provincial level, 11 provinces show a negative and significant impact of the shock on the interest rate over employment, accumulated to 10 periods. However, there are 13 provinces in which the effect is not statistically significant. Among the latter, the two main jurisdictions (GBA-CABA and Buenos Aires) are noteworthy due to their non-significant impact, together with other relatively less developed provinces, such as Formosa and Patagonian provinces. On the other hand, the provinces that show significant impacts have diverse ranges of economic and financial development.

This paper is organized as follows. Section 2 develops the econometric model used to estimate and evaluate the shocks. Section 3 describes the Argentinean data and section 4 presents the econometric results for Argentina. Section 5 concludes and proposes further lines of research.

2 Econometric model

2.1 Maximum likelihood model

As mentioned in the Introduction, the aim of this work is to account for the spatial heterogeneity of macroeconomic shocks.

In Carlino and DeFina's (1998, 1999) approach interdependence across states is dealt with by allowing the lagged output of other regions to enter the equations of each specific region or state. However, no contemporaneous feedback is allowed (i.e., simultaneous propagation of economic disturbances among regions is excluded). This assumption is reflected in the identification scheme that is adopted, which rules out any contemporaneous interdependence among states by means of a set of overidentifying restrictions imposed on the contemporaneous VAR coefficients matrix. As a result, spatial propagation of monetary policy shocks is assumed to take place at least with a one-period time lag. De Lucio and Izquierdo (1999) contribution, while ruling out lagged feedback effects among regions, does allow for contemporaneous correlation among the VAR model residuals.

Their preferred specification consists of a set of regional macro-type SVARs, jointly estimated using seemingly unrelated regression (SUR) techniques.

Di Giacinto (2003) uses geographical proximity in the model specification assuming that information with respect to the nearest neighboring areas is relevant in predicting the process at a given location. He follows the standard approach in spatial econometrics (see, e.g., Anselin, 1988, chap. 3; Martin and Oeppen, 1975; Pfeifer and Deutsch, 1980; Pfeifer and Bodily, 1990) where a priori information on the spatial connectivity structure underlying the observed data is made operational within the VAR model through a sequence of spatial weights matrices, defined according to a proper spatial weighting scheme. Through the sequence of spatial weights matrices, a set of parameter restrictions is imposed on the VAR coefficients matrices. On one hand, these restrictions allow for the identification and estimation of a single monetary policy shock series for all regions by eliminating the degrees-of-freedom constraint incurred by VAR models as the cross-sectional dimension of the model increases. On the other hand, spatial constraints are useful in modeling contemporaneous interdependence among regions while preserving a sufficiently large number of restrictions for structural parameter identification.

Bertanha and Haddad (2008) apply Di Giacinto's model to Brazilian states and analyze the presence of regional asymmetries in the impact of monetary shocks for the 27 states of Brazil. The authors use a SVAR model with spatial weighted matrices. In fact, they can test the difference between the contiguity matrix and a trade-weighted matrix, as well as the importance of lagged and direct spatial effects. The direct effects predominate in the results, while the trade matrix enhances the impact of the shock in the state of São Paolo and Manaus (tax-free zone) where trade is a highly relevant sector. This is in fact the closest paper to our analysis.

We follow the model proposed in Di Giacinto (2003) that constructs a structural VAR (SVAR) model with temporal as well as spatial lags. The spatial SVAR model adds spatial information in the model making use of techniques commonly employed in spatial econometrics. Broadly speaking, the idea of spatial heterogeneity is given by the fact that the output of any spatial unit could be directly or indirectly affected by the output of any of the other units. Such idea can be covered by the traditional SVAR as in Carlino and DeFina (1998, 1999), Fraser et al. (2014) and Guo and Tajul (2017). The innovation of the spatial SVAR model is the introduction of the contiguity matrix in the context of SVAR.

The model considers three sets of variables. The first set, denoted as $\mathbf{x}_t = [x_{1t}, x_{2t}, \ldots, x_{Kt}]'$, represents K macroeconomic aggregate control variables. Under our specification, such variables are given by consumer's price index (CPI), the U.S. dollar / peso exchange rate and gross domestic product (GDP), i.e., $K = 3$. These variables correspond to the aggregate or national level. The second set of variables, denoted by $\mathbf{y}_t = [y_{1t}, y_{1t}, \ldots, y_{Nt}]'$, includes the stacked values of the output variable measured on the N spatial units. Our spatial variable is total formal employment in each regional/state unit. As discussed below this is the only variable for which we have spatial as well as temporal heterogeneity in Argentina. The third set is given by a single variable, the monetary policy instrument, the interest rate in our model, denoted by r_t. Following the macroeconomics literature, our interest is to estimate the effect of a shock in this variable on the output variables, which is measured by employment.

Setting $\mathbf{z}_t = [\mathbf{x}_t', \mathbf{y}_t', r_t]'$, the spatial SVAR model has the following expression

$$\mathbf{C}_0\mathbf{z}_t \quad = \quad \mathbf{C}_1\mathbf{z}_{t-1} + \ldots + \mathbf{C}_p\mathbf{z}_{t-p} + \mathbf{u}_t, \tag{1}$$

where $\mathbf{u}_t = [u^x_{1t}, \ldots, u^x_{Kt}, u^y_{1t}, \ldots, u^y_{Nt}, u^r_t]$ is an orthogonal multivariate white-noise series, i.e., $E(\mathbf{u}_t) = \mathbf{0}$, $E(\mathbf{u}_t\mathbf{u}'_{t-h}) = \Omega = diag([\sigma^2_{x1}, \ldots, \sigma^2_{xK}, \sigma^2_{y1}, \ldots, \sigma^2_{yN}, \sigma^2_r]')$ if $h = 0$ and $E(\mathbf{u}_t\mathbf{u}'_{t-h}) = \mathbf{0}$ elsewhere for $h \geq 0$.

The $\mathbf{C}_0$ matrix has the following block triangular structure

$$\mathbf{C}_0 \quad = \quad \begin{bmatrix} \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_0^{yy} & \mathbf{0} \\ -\mathbf{C}_0^{rx} & -\mathbf{C}_0^{ry} & 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{C}_0^{rx}$ is a $(1 \times K)$ vector of unrestricted coefficients relating the policy instrument to the contemporaneous values of the macro variables $\mathbf{x}$, and where

$$\mathbf{C}_0^{ry} \quad = \quad a_0^{ry}\omega'. \tag{3}$$

a_0^{ry} is a scalar parameter to be estimated and ω is a vector of N fixed coefficients representing the average weight of employment of each spatial unit with respect to the national aggregate. This determines that the interest rate is affected by employment only through a national weighted average. This restriction is motivated by the assumption that only aggregate output enters the Central Bank information set and, hence, the monetary instrument response function. The $\mathbf{C}_0^{yy}$ matrix models simultaneous spatial interdependence by the following structure

$$\mathbf{C}_0^{yy} \quad = \quad \mathbf{I}_N - \phi_0 W, \tag{4}$$

where $\phi_0 = diag([\phi_0^1, \phi_0^2, \ldots, \phi_0^N]')$ and W is the $N \times N$ spatial weights matrix with typical element $w(i,j) > 0$ if locations i and j are contiguous (in a broad sense) and $w(i,j) = 0$ elsewhere and if $i = j$.

Two types of restriction are imposed on the $\mathbf{C}_h$ matrices ($h = 1, \ldots, p$). First,

$$\mathbf{C}_h \quad = \quad \begin{bmatrix} \mathbf{C}_h^{xx} & \mathbf{C}_h^{xy} & \mathbf{C}_h^{xr} \\ \mathbf{C}_h^{yx} & \mathbf{C}_h^{yy} & \mathbf{C}_h^{yr} \\ \mathbf{C}_h^{rx} & \mathbf{C}_h^{ry} & \mathbf{C}_h^{rr} \end{bmatrix}. \tag{5}$$

Second, spatial restrictions are imposed on blocks $\mathbf{C}_h^{yy}$ that have structure

$$\mathbf{C}_h^{yy} \quad = \quad \sum_{k=1}^{\lambda_h} \phi_h W, \tag{6}$$

where $\phi_h = diag([\phi_h^1, \phi_h^2, \ldots, \phi_h^N]')$. Coefficients $\mathbf{C}_h^{xy}$ and $\mathbf{C}_h^{ry}$ relating the macro variables and the monetary instrument to past values of the spatial output series are constrained as follows

$$\mathbf{C}_h^{xy} \quad = \quad \mathbf{a}_h^{xy} \omega' \tag{7}$$

$$\mathbf{C}_h^{ry} \quad = \quad \mathbf{a}_h^{ry} \omega' \tag{8}$$

where $\mathbf{a}_h^{xy}$ and $\mathbf{a}_h^{ry}$ are, respectively, a k-dimensional vector and a scalar to be estimated. All remaining blocks are left unrestricted, as in the standard VAR specification. Di Giacinto (2003) derives consistent estimators of model parameters applying Full Information Maximum Likelihood method. Further details on the estimation procedure can be found in that paper.

Shock identification is embedded in the structural model described above. As noted by Dominguez-Torres and Hierro (2018) this is the most common structure for identification of monetary shocks in spatial models, where the policy instrument (i.e., r_t) is regressed on all other contemporaneous variables and temporal lags. Their meta-analysis suggests, however, "that the choice of the identification scheme appears to have no effect on the pattern of the responses yielded by these studies, since such responses broadly exhibit a hump-shaped trajectory (when a contractive shock is analysed) irrespective of the identification scheme implemented."(p.4) Our preliminary evidence also confirm that the results are robust to different identification procedures.

2.2 Models

We estimate three different models. One the one hand we estimate a SVAR model that ignores spatial heterogeneity, and use this models as a benchmark. This corresponds to a national-level model, a standard empirical macroeconomics SVAR in the line of Christiano et al. (1996).[2]

On the other hand, based on the general setting presented above, we estimate two spatial models, the *State Model* (SM) and the *Regional Model* (RM). The main difference between them is given by the level of spatial aggregation. The SM considers $N = 24$ spatial units given by the 23 states plus a conglomerate formed by the City of Buenos Aires and its contiguous neighborhood (known as *Gran*

[2] This model considers the macro variables, the interest rate and the aggregate employment (y_t), that is, it considers the same variables as the spatial models but employment is aggregated at the national level. We maintain the structural form of the non-spatial model as similar as possible to the spatial ones. In particular we consider a model of the form

$$\mathbf{B}_0\mathbf{z_t} \quad = \quad \mathbf{B}_1\mathbf{z_{t-1}} + \ldots + \mathbf{B}_p\mathbf{z_{t-p}} + \mathbf{u_t}, \tag{9}$$

where $\mathbf{z}_t = [x_t', e_t, r_t]$ and $\mathbf{u}_t = [u_{1t}^x, \ldots, u_{Kt}^x, u_t^y, u_t^r]$ is an orthogonal multivariate white-noise series. The $\mathbf{B}_0$ matrix has the following block triangular structure

$$\mathbf{B}_0 \quad = \quad \begin{bmatrix} \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ -\mathbf{B}_0^{rx} & -\mathbf{B}_0^{ry} & 1 \end{bmatrix}, \tag{10}$$

where $\mathbf{B}_0^{rx}$ is a $(1 \times K)$ vector of unrestricted coefficients relating the policy instrument to the contemporaneous values of the macro variables x, and $-\mathbf{B}_0^{ry}$ is a coefficient relating the policy instrument to the contemporaneous values of the aggregate employment. As in spatial models, the temporal lags were set to $p = 2$.

Buenos Aires, CABA-GBA), while the RM considers $N = 6$ spatial units given by 5 regions (groups of states) plus the previously defined conglomerate. See section 3 for a description of the Argentine regional structure.

Regarding the spatial structure of the models, for the SM we used a Queen type contiguity matrix, that is, two states are considered neighbors if they have a common border. [3] For the RM, however, we used a distance based contiguity matrix with the following structure. Let $W[w_{ij}]$ be the contiguity matrix,

$$w_{ij} \quad = \quad \frac{d(i,j)^{-1}}{\sum_{j=1}^{N} d(i,j)^{-1}}.$$

The strength of the relation between two spatial units, w_{ij}, is given by the inverse of the distance, as measured by centroids, among regions, $d(i,j)^{-1}$, considering the inverse of the distance to all the regions, $\sum_{j=1}^{N} d(i,j)^{-1}$. This configuration gives more weight to closer units. Unlike the Queen matrix, under this configuration all regions are considered neighbors. Both contiguity matrix are row-normalized.

For both models the spatial lag order was set to 1, and the temporal lags were set to $p = 2$, thus the matrix (5) has the following elements:

- $\mathbf{C}_h^{xx}$ is a (3×3) matrix relating the macro variables to their own values h periods ago;
- $\mathbf{C}_h^{xr}$ is a (3×1) matrix relating the macro variables to the interest rate $h = 1, 2,$ periods ago;
- $\mathbf{C}_h^{yx}$ is a $(N \times 3)$ matrix relating the total employment in each spatial unit to the macro variables $h = 1, 2,$ periods ago;
- $\mathbf{C}_h^{yr}$ is a $(N \times 1)$ matrix relating the total employment in each spatial unit to the interest rate $h = 1, 2,$ periods ago;
- $\mathbf{C}_h^{rx}$ is a (1×3) matrix relating the interest rate to the macro variables $h = 1, 2,$ periods ago;
- $\mathbf{C}_h^{rr}$ is a (1×1) matrix relating the interest rate to his own value $h = 1, 2,$ periods ago.

[3] Stakhovych and Bijmolt (2008, p.408) find that spatial models estimated using the first-order contiguity weights matrix perform better on average than those using the nearest neighbours or inverse distance weights matrices in terms of their higher probabilities of detecting the true model and they have lower MSE of the parameters.

Regarding the elements $\mathbf{C}_h^{xy}$ and $\mathbf{C}_h^{ry}$, they relate the macro variables and the interest rate to a weighting average of the total employment $h = 1, 2$, periods ago, respectively.

Following equations (7) and (8) we have that

- $\mathbf{a}_h^{xy}$ is a (3×1) matrix relating the macro variables to a weighting average of the total employment $h = 1, 2$, periods ago.

- $\mathbf{a}_h^{ry}$ is a (1×1) matrix relating the interest rate to a weighting average of the total employment $h = 1, 2$, periods ago.

We follow Bertanha and Haddad (2008) for defining the weights vector ω,

$$\omega' \quad = \quad (\omega_1, \omega_2, \ldots, \omega_N),$$

with $\omega_j = \dfrac{\sum_{t=1}^{T}(TotEmp_{jt}/NatEmp_t)}{T}$, where $TotEmp_{jt}$ is the total employment in spatial unit 1 at time t and $NatEmp_t = \sum_{n=1}^{N}(TotEmp_{nt})$ is the total employment at national level at time t. The weight of each spatial unit is thus given by its relative importance in terms of national employment along the analyzed period.

2.3 Impulse response functions

From the model estimates our main interest lies in constructing the impulse response functions (IRFs) from a unit shock (i.e., 1% increase in the interest rate) in u_t^r on the y_t regional variables. That is, in evaluating the effect at provincial/regional level of an aggregate monetary shock corresponding to a tightening of the monetary policy via an increase of the reference interest rate. We study the effect of this shock on the difference in logarithm of employment, and thus the effects are evaluated on employment growth. The shock is determined by the identification strategy given by the structural model.

Given the complex nature of the maximum likelihood model presented above and the fact that we are not necessarily confident in the Gaussian nature of the shocks, we compute bootstrap standard errors of all parameter estimates. In particular, we consider non-parametric bootstrap samples, with replacement, of quarters with the corresponding structure of lags (using 2 lags), maintaining the geographic structure intact throughout the analysis. IRFs analysis is evaluated using 20% confidence intervals where we generate a ranking order from the bootstrap samples for each of the 12 periods-ahead used in the IRFs.

3 Data description

3.1 Brief description of Argentina's regional structure

Argentina is a federal country, with 23 provinces and a semi-autonomous city, Ciudad Autónoma de Buenos Aires. As a federation, provinces reserve all powers not delegated to the federal government. They can dictate their own constitutions and manage autonomous budget and public policies (e.g. education, health) collecting local turnover, property and stamp taxes. Although the national constitution contemplates the possibility that provinces agree to be grouped into regions, the basic institutional jurisdiction is always the provincial level. Regions are nevertheless a common (and historical) way to group and analyze at a subnational level.[4] In general, they are selected by geographical contiguity, historical traditions and economic and financial similarities.[5] Figure 1 shows a traditional regional division that comprises six different regions: Centro, NOA (North-West), NEA (North-East), Cuyo, Sur, and the Buenos Aires metropolitan region, that comprises the city of Buenos Aires and its metropolitan area (Ciudad Autónoma de Buenos Aires plus Gran Buenos Aires, GBA-CABA). We consider GBA-CABA as a different spatial unit because they have considerable economic and structural differences in relation to the rest of the country, and in the Argentinean case, they concentrate a considerable portion of economic activity and population. Together the 6 regions are used in the regional model (RM) described above.

As can be seen in Table 1, provinces included in the Centro (34% GDP) and GBA-CABA (40% GDP) regions are the more economic developed provinces. The NEA (4.2% GDP) and to a lesser extent, the NOA (6.9% of GDP) provinces are the less developed ones. Whereas the Cuyo (6,1% of GDP) region includes provinces with an intermediate level of economic development. Finally, the Patagonian Sur region (9.2% GDP) includes intermediated developed provinces with large areas and low population density.

[4] See for instance the National Production Ministry, `http://mapaprod.produccion.gob.ar`

[5] See, for instance, Elosegui, Anastasi, Sangiácomo, and Blanco (2010) for an analysis of economic determinants of use and availability of banking services at a local level for the 1998-2009 period.

3.2 Variables used in the econometric models

Table 2 describes the data used and its sources. All variables have quarterly periodicity, and the time span considered for our exercises is 2003q1-2017q2. The series were seasonally adjusted (when needed) using X-13 ARIMA-SEATS, detrended or differentiated to make them stationary and finally log transformed. Population data between Census was interpolated using a linear polynomial.

One of the major issues working with Argentina is the lack of good data. At the regional and provincial level we can only rely on employment (total formal employment) to construct a panel data from which we can study the spatial interactions. The macro variables (consumer's price index CPI, US dollar/peso exchange rate and GDP) as well as the 'spatial variable' are in logarithm, the interest rate (30-59 days term deposits rate) is in percentage. After these transformations, based on augmented Dickey-Fuller tests, all variables are stationary. Figure 2 plots the main macroeconomic variables.[6]

For our subsequent analysis (see section **??** for the Bayesian model averaging analysis) we add to employment and macro data three sets of subnational indicators: one that captures the production mix of the province, and other that is specific to the stance of the provincial economy and the last that accounts for financial sector indicators (see Table 1).

4 Empirical results

4.1 Spatial correlation estimates

The spatial SVAR models proposed here depends on the existence of spatial effects. Such interaction is captured by the coefficients ϕ_{hk}, where h refers to the temporal lag and k refers to the lag order of the contiguity matrix, we use $h = 0, 1, 2$ and $k = 1$. If $\phi^{n}_{0k} \neq 0$, with $n = (1, \ldots, N)$, that means that a change in employment in the neighborhood of spatial unit n has a (direct) contemporaneous impact on employment of unit n. Furthermore, given that all spatial units are, directly or indirectly connected, a change in employment in any spatial unit has a (direct or indirect) contemporaneous

[6] Constructing appropriate data for Argentina for the 2007-2015 period is a controversial issue. First, during those years the official statistical office (INDEC, Instituto Nacional de Estadísticas y Censos) has been manipulated to report lower inflation. Nevertheless, we use the official CPI. Preliminary evidence using other alternative CPI provide the same results. Second, the 2011-2015 period is one of exchange rate controls. Thus, the official exchange rate (OER) differed from the unofficial exchange rate (UER) also known as "blue". The former applied to imports and exports, but most economic agents had quantity restrictions on buying US dollars and could only could buy dollars at the unofficial market. We use the OER for our analysis.

impact on employment of all other spatial units. This multiplicative impacts depend on the value (and significance) of ϕ^n_{hk}. This is also valid for ϕ^n_{1k} and ϕ^n_{2k}, but now the impact is with one and two time lags respectively.

Tables 3 and 4 present the point estimates and the bootstrap standard errors (with 200 bootstrap simulations) of the estimates of ϕ_{hk} for the Regional and State models, respectively. The analysis confirms that most spatial effects are statistically significant and positive with a few exceptions.

4.2 Regional and non-spatial aggregate models

Figures 3 and 4 summarize the IRFs from a monetary shock (a 1% increment in the interest rate) at under both the regional level (using the *RM* model) the non-spatial national aggregate models. A simple comparison shows that most regions show a negative effect when the interest rate increases, except for the NEA region (where the effect is not statistically significant). There is however a marked heterogeneity in the effects. The Sur and Cuyo regions are the most affected. Note that the IRF for the non-spatial model does not correspond to a simple average of the others, although it is close to the GBA-CABA region (the largest and most concentrated region).

As a comparison we also compute the effect of the same monetary shock on employment, for each region separately (see figure 5). The results confirm the negative impact of increasing the interest rate on employment. These results confirm that monetary shocks have a negative impact on regional employment. Note, however, that the *RM* spatial model produces larger effects than what is captured by each region separately. While the separate VAR has effects ranging between -0.003 and -0.008, the *RM* model ranges from -0.010 to -0.025 (ignoring the positive NEA effect). The ranking among regions also changes. While GBA-CABA has the largest effect when considering by a separate VAR, it the smallest effect (in absolute value) in the *RM* estimation.

4.3 State model

Consider now the IRFs from a monetary shock at the provincial level. These results are summarized in figures (6)-(8). This analysis shows greater heterogeneity among provinces. Many of them are not statistically significant although the point estimate is negative.

First, most short- and long-term effects are negative, except for Neuquén, Santiago del Estero and Tierra del Fuego. Thus, increasing the interest rate has a negative impact on employment growth at the Argentinean states.

Second, most of them have a short term negative and significant effect. Exceptions are CABA-GBA and Buenos Aires provinces. This could be due to the limited nature of our database because the national level estimate is indeed significant, and both units have a large share of the national aggregate total employment.

Third, for those provinces with a long-term statistically significant effect, employment growth decreases by between 1% and 2% after a 1 percentage point increment in the interest rate.

5 Discussion and conclusion

This paper empirically investigates how economic activity, as measured by total formal employment in Argentina at regional and provincial (i.e., state) levels respond to central or national monetary policy shocks, given by a change in the policy interest rate. The results confirm that there is considerable regional heterogeneity across regions and states within Argentina, resulting in differential effects of monetary policy shocks. At the regional level the long-term effects are negative and statistically significant. At the provincial level, 11 provinces show a negative and significant long-term impact of the shock on the interest rate over employment. However, there are 13 provinces in which the effect is not statistically significant, including GBA-CABA and Buenos Aires province.

Macroeconomic policies are generally "blind" at regional level (Hewings, 2014) and this ignores potentially large asymmetric effects across regions. The results in this paper indicate that further research should be applied benefiting from the large literature on spatial analysis of macroeconomic effects.

The non-homogeneous synchronization of regional business cycles may also be an important factor for the observed heterogeneity of regional impact of monetary shocks. In fact, local or regional cycles may not be in phase, with other regions or the national economy. In this sense, the time window and the territorial unit of analysis may be crucial to understand the potential heterogeneity.

Indeed, high frequency data may augment the heterogeneity by capturing better region to region interaction, as documented by Park and Hewings (2012). Given our short time span we focused on quarterly data that is at the same time more relevant to analyze the impact of monetary policy considering the traditional implementation lag of monetary policy shocks. Also, the authors showed that different industry mix and/or the place in the value chain production may contribute to the cyclical asymmetries, in line with the importance of such characteristics for the monetary policy transmission mechanism literature.

As Dominguez-Torres and Hierro (2018) emphasize, not only the time domain but also the space domain is important when considering cyclical heterogeneities. Asymmetries can be traced back to the interaction between regions, or more likely between provinces within the regions. The authors reviewed several empirical studies having comparable results with different aggregation levels both in terms of qualitative (trajectory) and quantitative (magnitude) results. In the case of US and China, authors find that the results are broadly maintained when using different levels of territorial aggregation. However, in the cases of Brazil and Canada, there are no such regularities and consistencies among the empirical analysis when using different levels of territorial aggregation. See also Mejía, P. y D. Lucatero (2011) for the case of Mexico.

It should be noted that as an initial methodological approach to the provincial business cycle in Argentina, we focus on the impact of a macroeconomic shock (monetary policy) to the regions or provinces considering the spatial interactions. However, as noted by an anonymous referee, our next research agenda should include at least two interesting issues. First, a better distinction between aggregated shocks and local to local shocks. Despite been an active research agenda for developed federal countries, there are not much research in developing federal countries. For instance, Bai and Wang (2012) was used by Chung and Hewings (2015) to capture regional asymmetries using a multi-level (in this case, two-level) approach. Along the same lines, Ramajo et al. (2017) developed a multiregional spatial vector autoregressive (MultiREG-SpVAR) model applied to study the spatiotemporal transmission of macroeconomic shocks across the regions in Spain. Second, our results indicate that the nation influences regions and there is some “contamination” from other regions but the strength and direction of these effects are not fully revealed as in the work by Hayashida and Hewings (2009) or through the Dendrinos-Sonis log-linear relative dynamic approach applied by Postiglione and Hewings (2008) for the case of Italy.

References

[1] Anastasi, A., Blanco, E., Elosegui, P. and M. Sangiácomo (2010) "Bankarization and determinants of availability of banking services in Argentina." *Ensayos Económicos*, 60.

[2] Anselin, L. (1998) *Spatial Econometrics: Models and Applications.* Kluwer, Dordrecht, The Netherlands.

[3] Arnold, I. (2001) "Regional effects of monetary policy in Europe." *Journal of Economic Integration*, 16(3), 399-420.

[4] Arnold, I. and E. Vrugt (2002) "Regional effects of monetary policy in the Netherlands." *International Journal of Business and Economics*, 1(2), 123-134.

[5] Bai, J. and P. Wang (2012) "Identification and estimation of dynamic factor models." Discussion Paper No. 1112-06, Department of Economics, Columbia University, New York.

[6] Bertanha, M. and M. A. Haddad (2008) "Efeitos regionais da política monetéria no Brasil: Impactos e transbordamentos espaciais." *Revista Brasileira de Economia* 62(1), 3-29.

[7] Bernanke, B. and A. Blinder (1988) "Credit, Money, and Aggregate Demand." NBER Working Paper No. 2534, March.

[8] Carlino, G. A. and R. DeFina (1998) "The differential regional effects of monetary policy." *Review of Economics and Statistics*, 80, 572-87.

[9] Carlino, G. A. and R. DeFina (1999) "The differential effects of monetary policy: Evidence from US States and Regions. Journal of Regional Science, 39, 339-358.

[10] Christiano, L., Eichenbaum, M. and C. Evans (1996) "The effects of monetary policy shocks: Evidence from the flow of funds." *Review of Economics and Statistics*, 78, 16-34.

[11] Chung, S. (2016) "Assessing the regional business cycle asymmetry in a multi-level structure framework: A study of the top 20 US MSAs." *Annals of Regional Science*, 56, 229-252.

[12] Chung, S. and G.J.D. Hewings (2015) "Competitive and complementary relationship between regional economies: A study of the Great Lake States." *Spatial Economic Analysis*, 10, 205-229.

[13] De Lucio, J. and M. Izquierdo (1999) “Local responses to a global monetary policy: The regional structure of financial systems.” Documento de Trabajo 99-14. Fundación de Estudios de Economia Aplicada FEDEA.

[14] Di Giacinto, V. (2003) “Differential regional effects of monetary policy: A geographical SVAR approach.” *International Regional Science Review*, 26(3), 313-341.

[15] Dominguez-Torres, H. and L. Hierro (2018) “The regional effects of monetary policy: A survey of the empirical literature.” Forthcoming *Journal of Economic Surveys.*

[16] Fraser, P., Macdonald, G. A., and A. Mullineux (2014) “Regional monetary policy: an Australian perspective.” *Regional Studies*, 48(8), 1419-1433.

[17] Guo, X. and A. M. Tajul (2017) “Regional effects of monetary policy in China: evidence from China’s provinces.” *Bulletin of Economic Research*, 69(2), 178-208.

[18] Hayashida, Motonari and G. Hewings (2009) “Regional business cycles in Japan.” *International Regional Science Review*, 32, 110-147.

[19] Hewings, G.J.D. (2014) “Spatially blind trade and fiscal impact policies and their impact on regional economies.” *Quarterly Review of Economics and Finance*, 54(4), 590-602.

[20] Kashyap, A., Stein, J. and D. Wilcox (1993). “Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance.” American Economic Review 83 (Mar.): 78-98.

[21] Kashyap, A. and J. Stein (1994). “The Impact of Monetary Policy on Bank Balance Sheets.” NBER Working Paper No. 4821, August

[22] Martin, R. L. and J. E.Oeppen (1975) “The identification of regional forecasting models using spacetime correlation functions.” Transactions of the Institute of British Geographers, 66, 95-118.

[23] Mejía, P. and D. Lucatero (2011) “Trends, structural breaks and economic growth regimes in the states of Mexico, 1940-2006.” *Paradigma Económico*, 3(1), 5-36.

[24] Park, Y. and G.J.D. Hewings (2012) “Does Industry Mix Matter in Regional Business Cycles?” *Studies in Regional Science*, 42, 39-60.

[25] Pfeifer, P. E. and S. E. Bodily (1990) "A test of space-time ARMA modelling and forecasting of hotel data." Journal of Forecasting, 9, 255-72.

[26] Pfeifer, P. E. and S. J. Deutsch (1980) "A three-stage iterative procedure for space-time modeling." Technometrics, 22, 35-47.

[27] Postiglione, P. and G.J.D. Hewings (2008) "Spatial hierarchical analysis of Italian regions." *Journal of Geographical Systems*, 10, 369-82.

[28] Ramajo, J., Márquez, M.A., and G.J.D. Hewings (2017) "Spatiotemporal analysis of regional systems: A multiregional spatial vector autoregressive model for Spain." *International Regional Science Review*, 40, 75-96

[29] Ridhwan M., Groot, H., Rietveld, P., and P. Nijkamp (2014) "The regional impact of monetary policy in Indonesia." *Growth and Change* , 45(2), 240-262.

[30] Runnemark, E. (2012) "Regional effects of monetary policy in Sweden." Working Paper 2012:9. Lund University.

[31] Serrano, F. and M. Nakane (2015) "Impacto regional da política monetaria no Brasil: Uma abordagem Bayesiana." Department of Economics FEA/USP, Working Paper Series 2015-44.

[32] Stakhovych, S. and T.H.A. Bijmolt (2008) "Specification of spatial models: A simulation study on weights matrices." *Papers in Regional Science*, 88(2), 389-408.

[33] Von Hagen, J. and C. J. Waller (2000) "Regional aspects of monetary policy in Europe." In *ZEI Studies in European Economics and Law*. Kluwer, Dordrecht, The Netherlands.

[34] Zeugner, S. (2011) "Bayesian Model Averaging with BMS. R package." `https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf`

Table 1: Selected economic and financial variables by provinces and regions

			Employment by sector share			Financial shares				Firms by size shares			
Province	Region	GDP share	Industry	Services	Public emp. pc	Loans	Deposits	Branches	Branch per 10K	Small	Large	Public Bank	Exports pc
GBA-CABA	GBA-CABA	39.7%	20%	57%	1%	54%	54%	18%	0.6789	89%	4%	1	0.00018
Buenos Aires	Centro	13.9%	21%	42%	9%	12%	12%	31%	2.2905	92%	4%	1	0.00340
Cordoba	Centro	7.8%	21%	45%	4%	7%	6%	10%	1.3811	91%	4%	1	0.00286
Entre Rios	Centro	2.4%	19%	37%	7%	2%	2%	3%	1.1489	89%	5%	0	0.00117
La Pampa	Centro	0.9%	12%	36%	8%	1%	1%	2%	3.4175	88%	7%	1	0.00118
Santa Fe	Centro	8.8%	25%	42%	5%	8%	6%	10%	1.4838	90%	4%	1	0.00469
Mendoza	Cuyo	3.9%	19%	41%	8%	2%	2%	4%	0.9374	88%	6%	0	0.00082
San Juan	Cuyo	1.1%	17%	36%	8%	1%	1%	1%	0.5726	84%	9%	0	0.00230
San Luis	Cuyo	1.1%	31%	36%	7%	0%	2%	1%	1.2260	82%	11%	0	0.00141
Chaco	NEA	1.3%	11%	40%	8%	1%	1%	1%	0.6160	85%	8%	1	0.00031
Corrientes	NEA	1.2%	15%	34%	6%	1%	1%	2%	0.9571	85%	8%	1	0.00022
Formosa	NEA	0.5%	7%	32%	9%	0%	1%	1%	0.4716	83%	11%	0	0.00006
Misiones	NEA	1.3%	19%	37%	6%	1%	1%	1%	0.5901	85%	8%	0	0.00039
Catamarca	NOA	0.9%	21%	35%	16%	0%	0%	1%	0.6797	82%	12%	0	0.00240
Jujuy	NOA	0.8%	21%	34%	13%	1%	1%	1%	0.4901	84%	10%	0	0.00062
La Rioja	NOA	0.6%	31%	25%	15%	0%	0%	1%	0.8692	80%	13%	1	0.00076
Salta	NOA	1.7%	13%	39%	8%	2%	1%	1%	0.5682	85%	8%	0	0.00081
Santiago del Estero	NOA	1.2%	11%	41%	7%	1%	1%	1%	0.6178	84%	9%	0	0.00081
Tucuman	NOA	1.7%	15%	41%	8%	2%	1%	1%	0.1933	85%	8%	0	0.00062
Chubut	Sur	2.2%	11%	33%	8%	1%	1%	2%	1.9839	87%	7%	1	0.00430
Neuquen	Sur	3.1%	7%	40%	12%	2%	1%	2%	1.8866	85%	8%	1	0.00032
Rio Negro	Sur	1.3%	9%	38%	10%	1%	1%	2%	1.1274	87%	7%	0	0.00083
Santa Cruz	Sur	1.7%	5%	36%	1%	1%	1%	1%	1.8616	83%	9%	0	0.00671
Tierra del Fuego	Sur	0.8%	34%	37%	14%	1%	1%	2%	6.6035	81%	12%	1	0.00130

Notes: Sources: Provincial GDP (2004) INDEC. Employment shares (Ministry of Labor statistical office). Export and population information INDEC. Financial information, Central Bank of Argentina. See Table 2 for sources.

Table 2: Variable description and sources

State/Regional Model	**Source**	**Seasonally adjusted**
Industrial employment data	Ministry of Labor	Yes
Population Data	INDEC	Yes
Macro Variables		
National GDP	INDEC	Yes
CPI Inflation	INDEC	No
30-59 days term deposits rate	BCRA	No
Bilateral Peso/USD Exchange rate	BCRA	No

Notes: INDEC: Instituto Nacional de Estadísticas y Censos (National Statistical Office), BCRA: Central Bank of Argentina.

Table 3: Estimates of ϕ_{01} and ϕ_{11} for RM.

	Centro	Cuyo	CABA-GBA	NEA	NOA	Sur
ϕ_0	0.619	0.749	0.464	-0.969	0.895	0.684
	(0.146)***	(0.296)***	(0.123)***	(0.506)**	(0.136)***	(0.240)***
ϕ_1	0.662	0.388	0.154	0.399	0.558	0.918
	(0.198)***	(0.290)*	(0.147)	(0.384)	(0.259)***	(0.254)***

Notes: Bootstrap standard errors in parenthesis. * Significant at 0.2 level. ** Significant at 0.1 level. *** Significant at 0.05 level. The estimates of ϕ_{21} are only significant for CABA-GBA, Sur and NEA at 0.2 level.

Table 4: Estimates of ϕ_{01} and ϕ_{11} for SM.

States	ϕ_{01}	ϕ_{11}	States	ϕ_{01}	ϕ_{11}
Buenos Aires	0.263	-0.039	Mendoza	0.451	0.699
	(0.833)	(1.316)		(0.118)***	(0.211)***
Córdoba	0.330	0.197	Misiones	0.171	0.276
	(0.162)***	(0.126)*		(0.078)***	(0.124)***
Catamarca	0.979	0.438	Neuquén	0.852	0.344
	(0.285)***	(0.316)*		(0.193)***	(0.300)
Chaco	0.566	0.467	Río Negro	0.237	0.424
	(0.468)	(0.472)		(0.138)**	(0.158)***
Chubut	0.420	0.263	Salta	-0.088	-0.097
	(0.114)***	(0.205)*		(0.153)	(0.226)
GBA-CABA	0.001	0.015	San Juan	0.172	-0.031
	(0.034)	(0.024)		(0.072)***	(0.125)
Corrientes	0.588	0.670	San Luis	0.272	0.225
	(0.129)***	(0.218)***		(0.140)**	(0.205)
Entre Ríos	0.301	0.171	Santa Cruz	0.982	-0.065
	(0.062)***	(0.097)**		(0.287)***	(0.325)
Formosa	0.513	0.072	Santa Fe	0.150	0.340
	(0.287)**	(0.402)		(0.110)*	(0.118)**
Jujuy	0.131	0.135	Santiago del Estero	-0.141	-0.198
	(0.074)**	(0.138)		(0.129)	(0.393)
La Pampa	0.247	0.351	Tierra del Fuego	-0.625	0.320
	(0.163)**	(0.265)*		(0.198)***	(0.158)***
La Rioja	-0.126	0.930	Tucumán	0.506	0.417
	(0.255)	(0.541)**		(0.160)***	(0.325)*

Notes: Bootstrap standard errors in parenthesis. * Significant at 0.2 level. ** Significant at 0.1 level. *** Significant at 0.05 level. The estimates of ϕ_{21} are only significant for Mendoza and San Juan at 0.05 level and for Tierra del Fuego and for Santa Fe at 0.1 and 0.2 level respectively.

Figure 1: Regions of Argentina

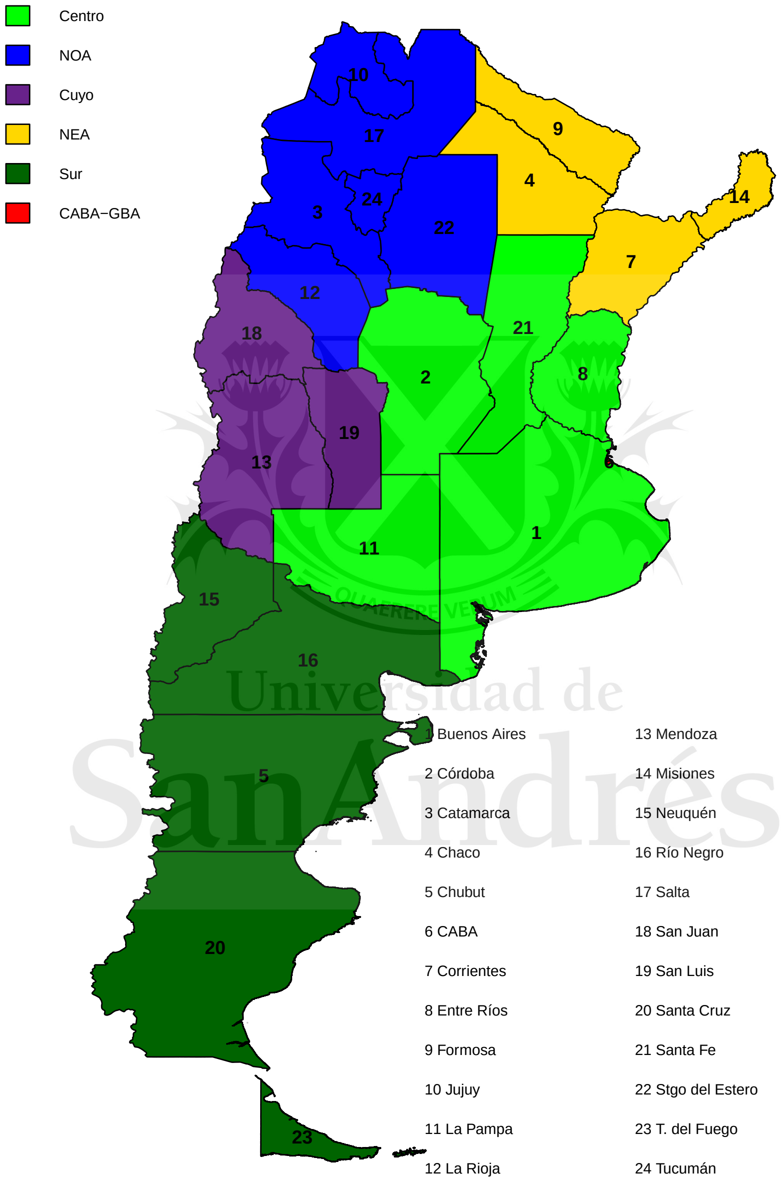


Figure 2: Macro variables

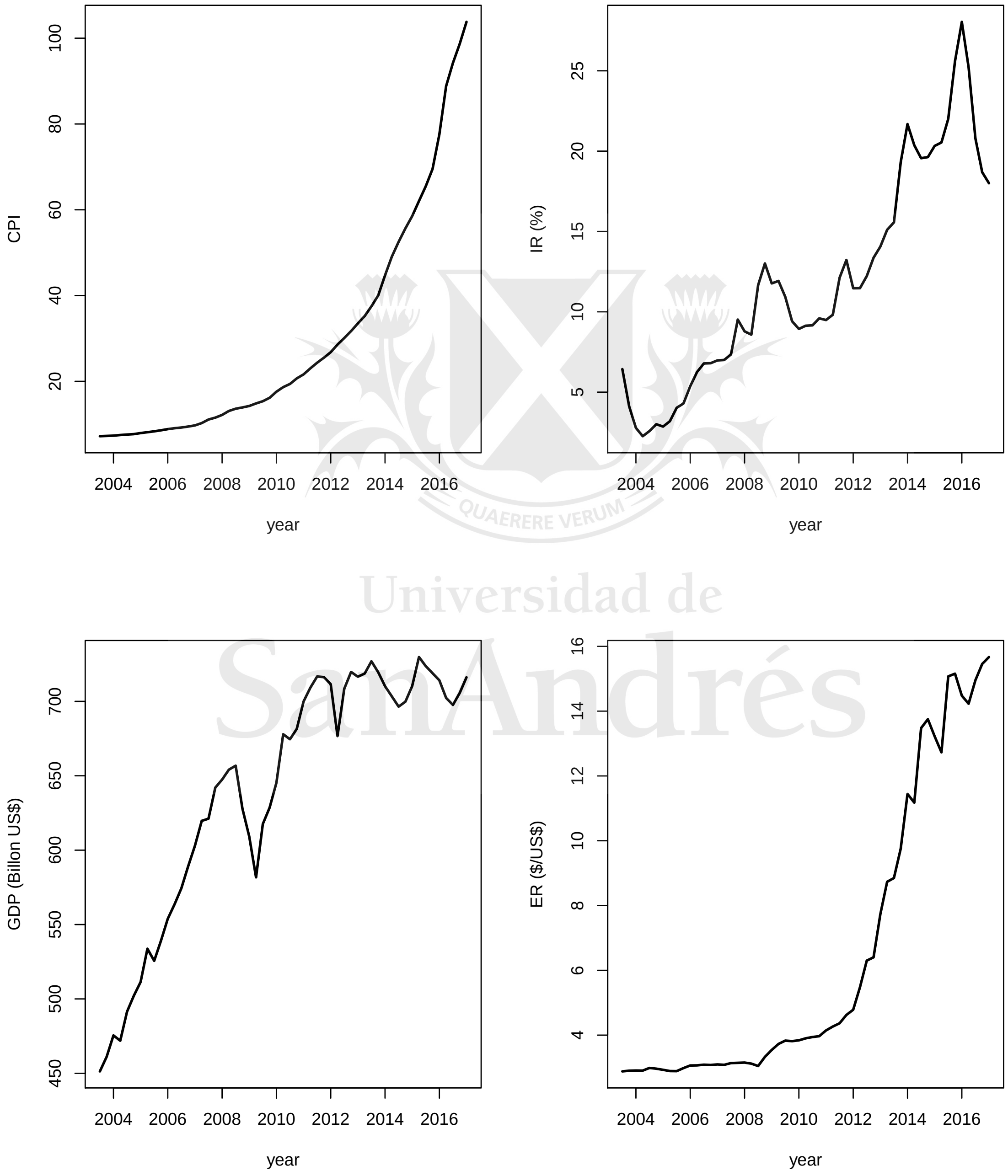
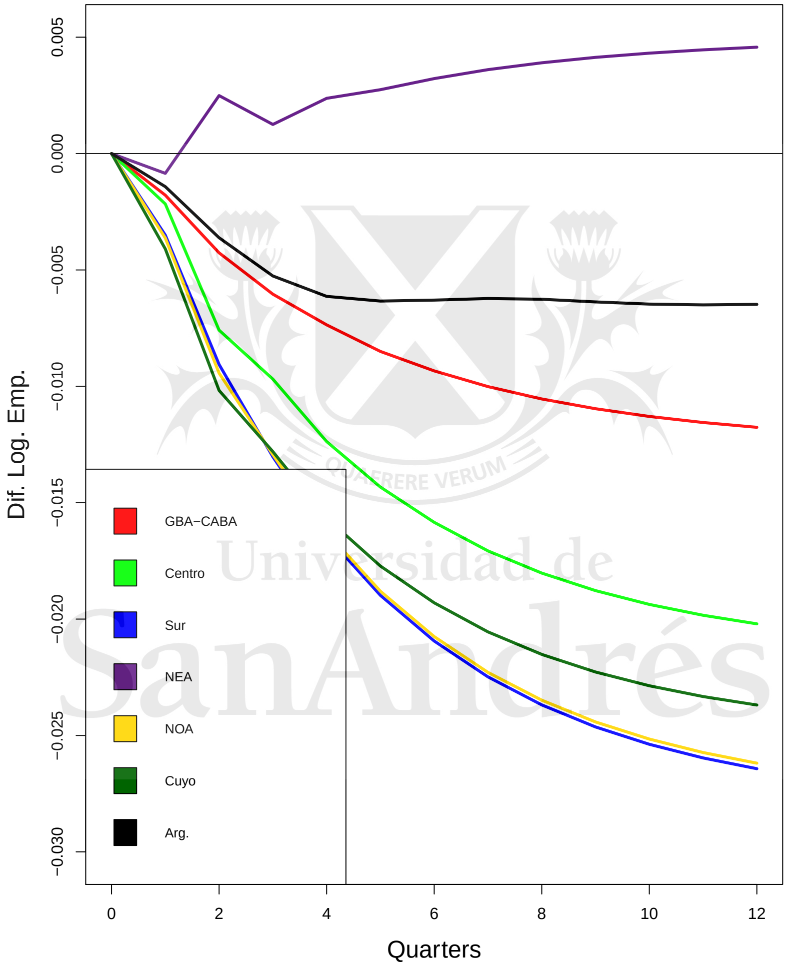
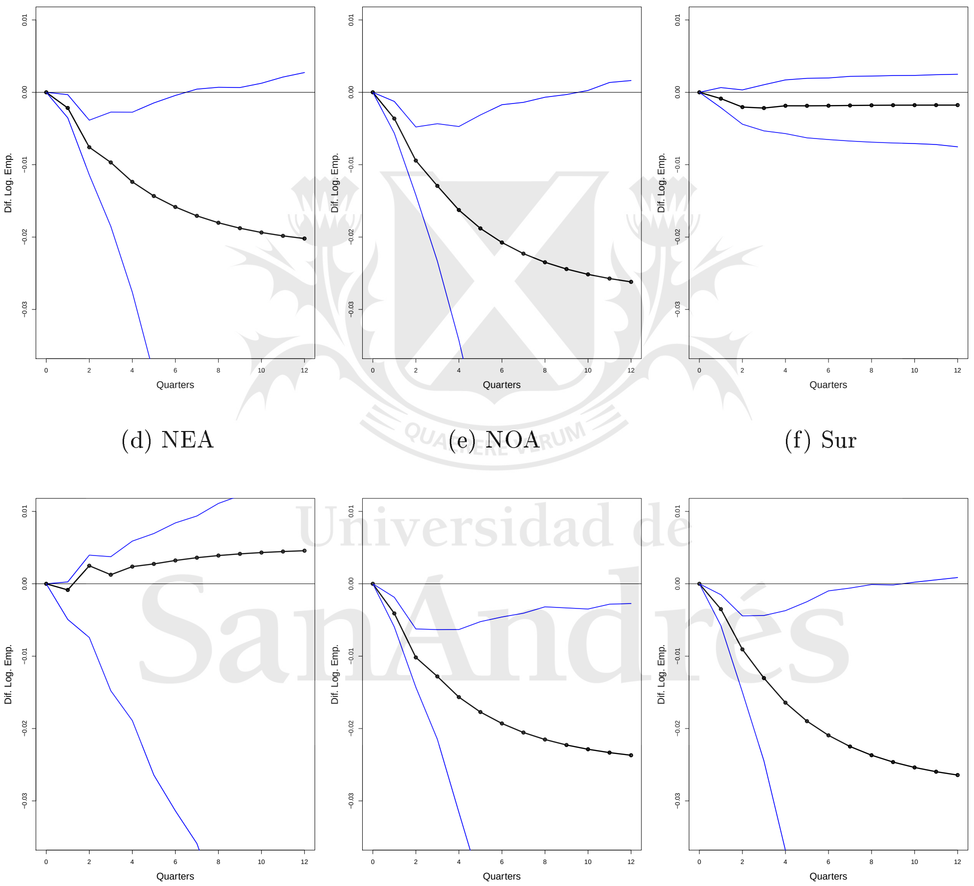


Figure 3: IRFs for national and regional model



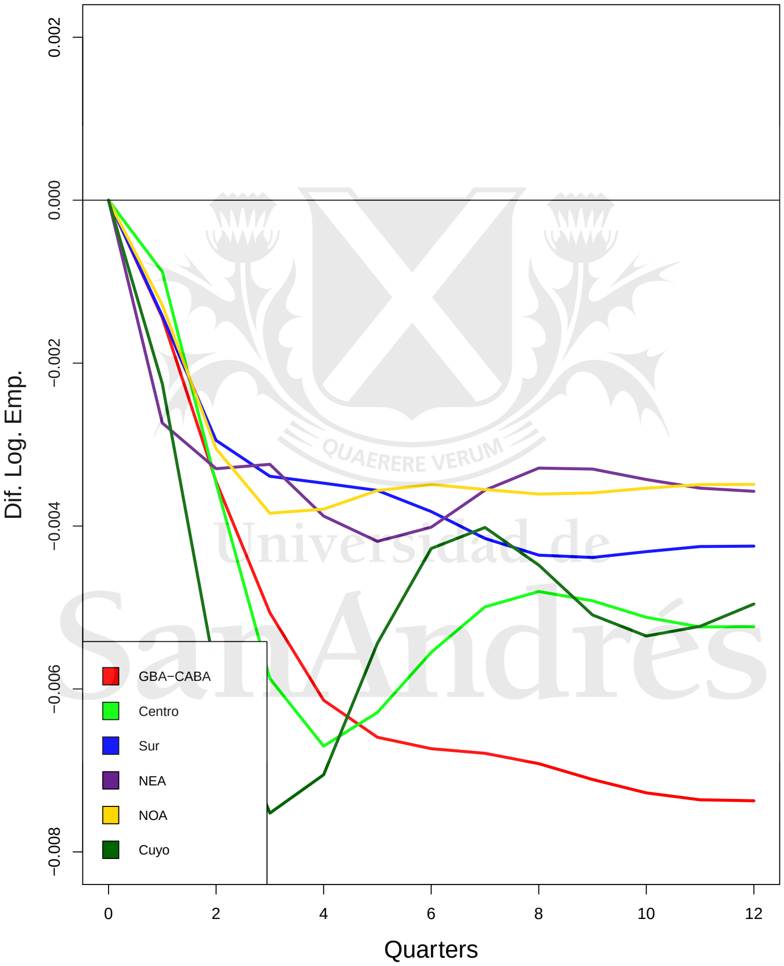
Note: IRFs of a 1% increment in the interest rate using the national aggregate VAR model and the RM spatial model.

Figure 4: IRFs by Regions



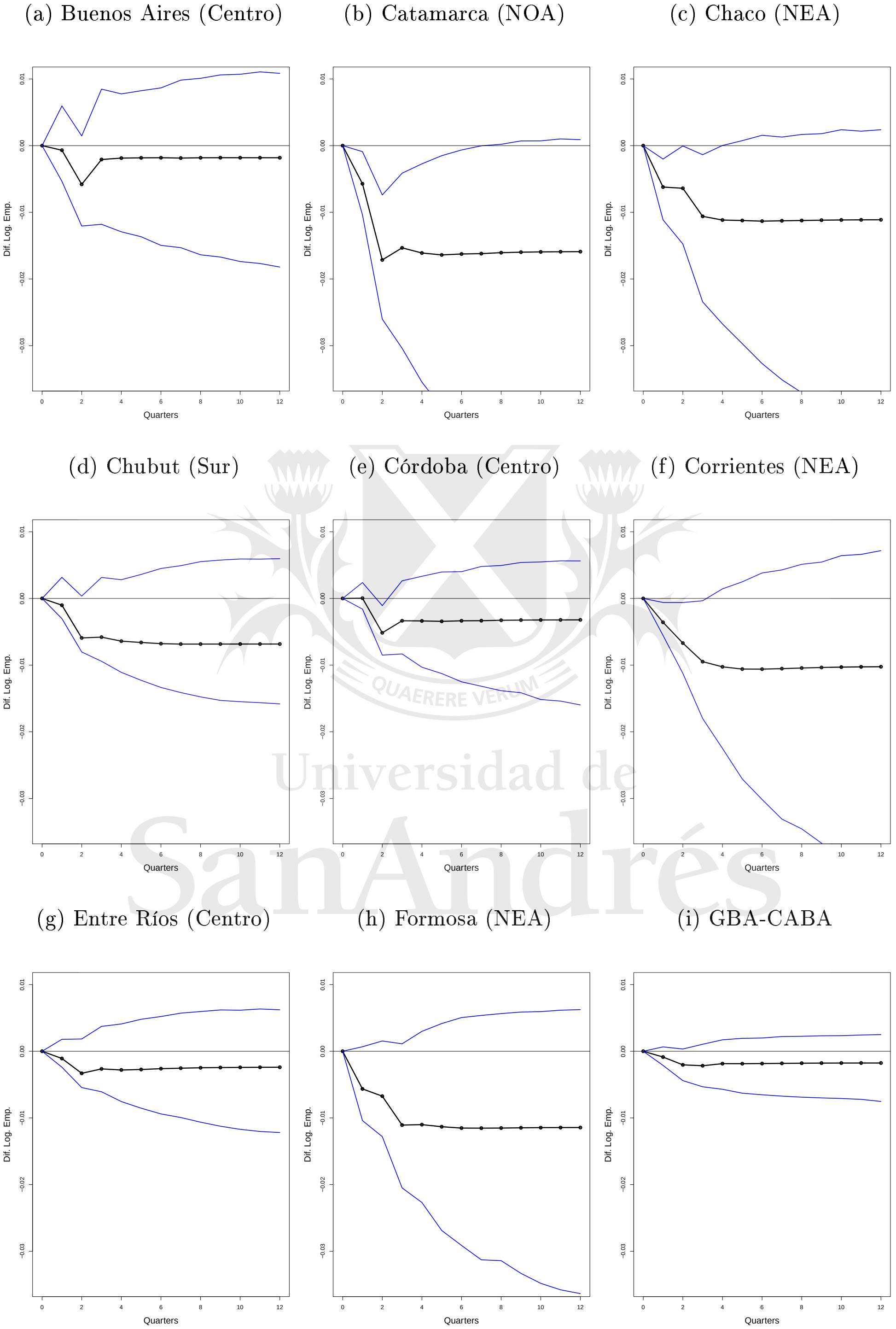
Note: IRFs of a 1% increment in the interest rate using the *RM* spatial model. 80% confidence interval are reported using bootstrap with 200 repetitions.

Figure 5: IRFs for each region separately VAR model



Note: IRFs of a 1% increment in the interest rate in a VAR model, separately for each region.

Figure 6: IRFs by states



(a) Buenos Aires (Centro)

(b) Catamarca (NOA)

(c) Chaco (NEA)

(d) Chubut (Sur)

(e) Córdoba (Centro)

(f) Corrientes (NEA)

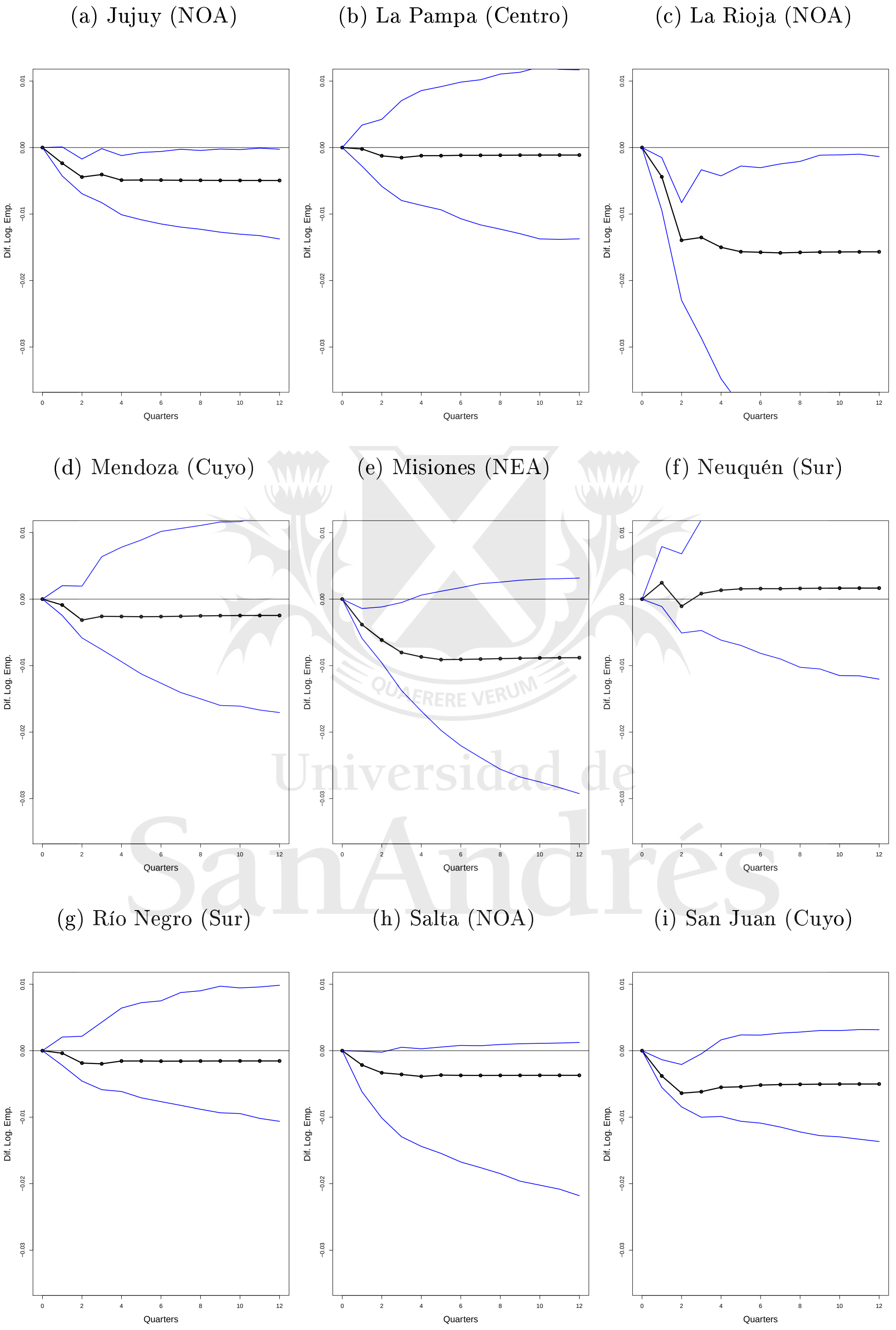
(g) Entre Ríos (Centro)

(h) Formosa (NEA)

(i) GBA-CABA

Note: IRFs of a 1% increment in the interest rate using the SM spatial model. 80% confidence interval are reported using bootstrap with 200 repetitions.

Figure 7: IRFs by state (cont.)



(a) Jujuy (NOA)

(b) La Pampa (Centro)

(c) La Rioja (NOA)

(d) Mendoza (Cuyo)

(e) Misiones (NEA)

(f) Neuquén (Sur)

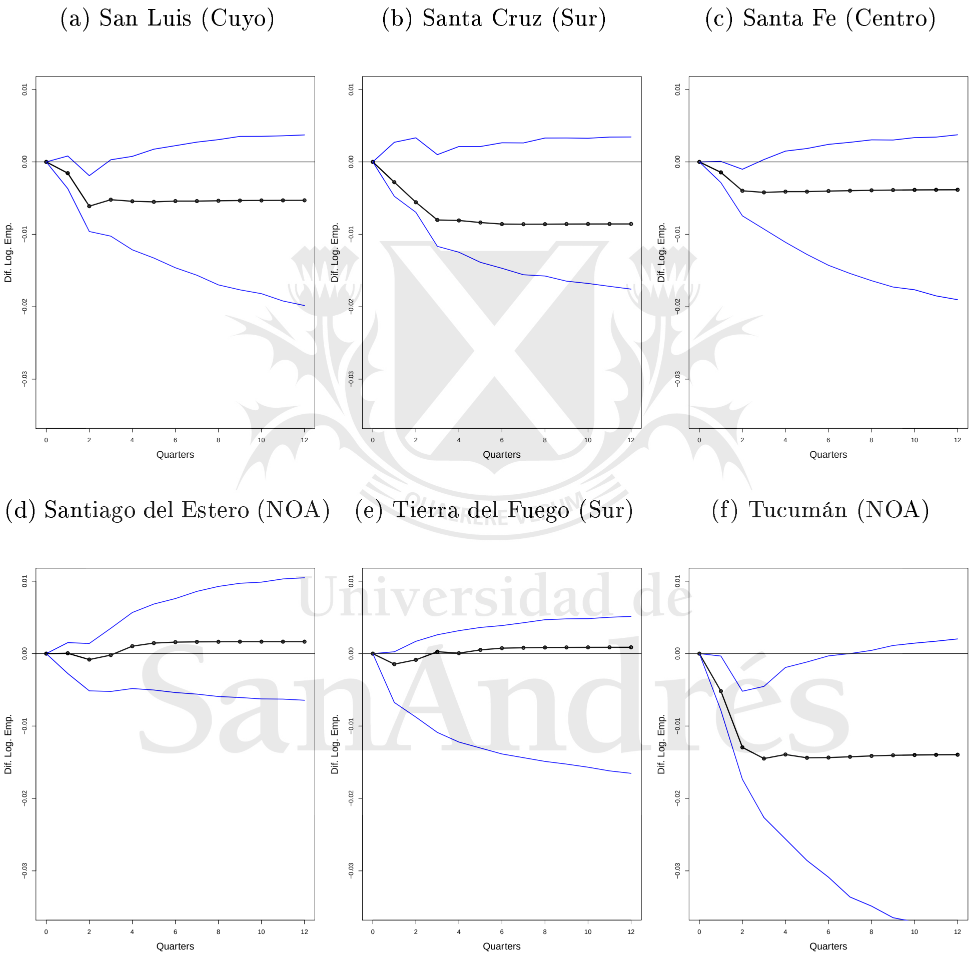
(g) Río Negro (Sur)

(h) Salta (NOA)

(i) San Juan (Cuyo)

Note: IRFs of a 1% increment in the interest rate using the SM spatial model. 80% confidence interval are reported using bootstrap with 200 repetitions.

Figure 8: IRFs by state (cont.)



(a) San Luis (Cuyo)

(b) Santa Cruz (Sur)

(c) Santa Fe (Centro)

(d) Santiago del Estero (NOA)

(e) Tierra del Fuego (Sur)

(f) Tucumán (NOA)

Note: IRFs of a 1% increment in the interest rate using the *SM* spatial model. 80% confidence interval are reported using bootstrap with 200 repetitions.

Capítulo III

Exploring peer effects in education in Latin America and the Caribbean

Abstract

This paper assesses peer group influence on academic performance of primary school students in Latin America and the Caribbean. Based on TERCE data set, we investigate peer effects in mathematics, language and sciences tests outcomes among sixth grade students. We apply a social interaction model which allows to identify endogenous and exogenous peer effects while controlling for group-level fixed effects. We explore some heterogeneities related to the school type (private, public or rural). The estimates suggest the existence of endogenous peer effects but their magnitude and significance depend on subject and school type.

Key words: *Peer effects, group interaction, academic performance.*
JEL classification: *C31, I21*

1 Introduction

Social scientists have long been interested in peer effects because of their far reaching implications at the individual and collective level. These non-market interactions represent how an individual's decision or outcome is directly influenced by his peer's outcome or characteristics. The sociological literature has placed great emphasis on the importance of social interactions arguing that they play an important role in determining behavioral and economic outcomes. In fact, a number of theoretical approaches such as collective socialization theories, contagion-based or epidemic theories, information asymmetries and network theories (Andrews et al., 2002) have been developed to account for contextual influence on individual's outcomes and behaviors regarding diverse aspects of life (such as criminal activity, use of public services, labor markets outcomes, etc.).

Among the various spheres in which peer effects may manifest themselves, the school context is especially important considering the vital role educational attainments have on future living conditions of individuals. Human capital accumulation has intertemporal repercussions given the proven relationship between years of schooling and labor incomes (Mincer, 1974; Becker, 1994). The analysis of peer effects in education has received considerable attention, notably since the publication of the Coleman report (Coleman et al., 1966). A common hypothesis is that student outcomes are higher in the presence of favorable peer groups, conditional on individual characteristics and family background (McEwan, 2003).

Evaluating peer effects in academic achievements is important for parents, teachers and schools; but crucially from a public policy perspective. A major question in the economic literature is whether or not interactions among students lead to large social multipliers (Epple & Romano, 1996). Depending on the nature of peer effects, there may be social gains from their existence (Hoxby, 2000). Furthermore, many researchers have studied the relative importance of peer effects in students academic performance versus the influence of other factors such as school infrastructure and teachers qualifications (Hanushek et al., 1998; Greene et al., 1999). As a matter of fact, peer effects have played a prominent role in educational policy debates concerning ability grouping, racial integration and school vouchers (Sacerdote, 2001; Gaviria & Raphael, 2001; Lin, 2005).

In this paper we analyse the possible existence of peer effects in educational achievements among sixth grade students participating in the Third Regional Comparative and Explanatory Study

(TERCE) conducted by United Nations Educational, Scientific and Cultural Organization (UNESCO). Since this survey focuses on primary school students, TERCE data provides a unique opportunity to explore peer effects in education in its early stages. Given the fact that primary education is a phase in which public policy can make a difference for students coming from vulnerable contexts, a better understanding of the educational production function shall improve equity in the education system. The latter has much relevance taking into consideration early education's welfare implications for future living standards of individuals and their families. Therefore, a deep understanding of the nature and characteristics of peer effects in education is not only central for educational policies but also for general policies targeting at social inequality.

One important difficulty in dealing with peer effects is that they are hard to identify with observational data since it is not easy to distinguish between the impacts that actually result from social interactions from the choices of with whom to interact with [1] and the existence of a common environment among group members (Manski, 1993). For this reason, disparities in educational attainments may actually reflect children and families with similar characteristics sorting together at the school level or facing similar exogenous factors. Consequently, divergence in academic performance of students could in fact reflect broader inequalities in the economy and thus policy implications differ greatly. To deal with these problems, recent developments in network literature allow to study outcomes of social interactions taking into consideration the problems caused by endogenous association of members within a group and cofounding factors (Moffitt et al., 2001; Bramoullé et al., 2009; Lee, 2007).

With this research we expect to contribute to the recent empirical literature on peer effects in education. Besides, this paper should specifically add to the scarce existing evidence on the magnitude and characteristics of peer effects in education in Latin America and the Caribbean. The article will explore personal, family and contextual factors associated with mathematics, language and sciences learning achievements for sixth grade students of those countries participating in TERCE. We also explore some heterogeneities in results depending on whether the school is urban public, urban private or rural. As this survey was applied in fifteen countries in the region, the data provides a general perspective of this subject in Latin American and Caribbean countries.

The paper is organized as follows. Section 2 reviews the existing literature on peer effects in educa-

[1] This refers to selection into peer groups based on common unobserved characteristics (homophily).

tion. Section 3 presents the methodological approach and econometric model used for estimations. Section 4 describes the data and variables used in the analysis, and explains how we deal with missing observations. Finally, section 5 shows estimation results while conclusions are provided in Section 6.

2 Literature review

The problem of heterogeneity of results in the education process, that manifests itself in significant differences in academic performance or achievements of students, has long attracted considerable attention in the economic literature (Hanushek, 1979; Burgess, 2016). In this line of research, the influence of peers on educational outcomes has been extensively studied. The milestone in this field is the 1966's Equality of Educational Opportunity Report (Coleman et al., 1966), known as Coleman report for its director. This report pushed peer effects into the limelight when concluding *'finally, it appears that a pupil's achievement is strongly related to the educational backgrounds and aspirations of the other students in the school'* (Coleman et al., 1966, pg. 22). Since this research, the empirical literature on peer effects has grown (Sacerdote, 2001; Hanushek et al., 2003; Angrist & Lang, 2004; Stinebrickner & Stinebrickner, 2006; Ammermueller & Pischke, 2009). However, the evidence regarding the magnitude of peer effects on student's achievement is far from conclusive.

The aforementioned lack of consensus partly reflects various econometric issues that any empirical study on peer effects must address. Trying to explain the common observation that people belonging to the same group tend to behave similarly, in a pioneer study Manski (1993) differentiates three kinds of social effects: *endogenous effects*, wherein the propensity of an individual to behave in some way varies with the behavior of the group; *exogenous (contextual) effects*, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of the group; *correlated effects*, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments.

Distinguishing between endogenous and exogenous effects is important because they have different implications for policy interventions. Endogenous effects may give rise to bidirectional influences and consequently to the possibility of social multipliers, while the repercussion of exogenous effects does not necessarily imply amplified responses to exogenous shocks (Gaviria & Raphael, 2001). As regards correlated effects, they arise when students in the same reference group achieve similar educational outcomes because they share a common set of characteristics. In this case, for example, it could imply that families send their children to the same schools according to their willingness and ability to pay for better peer influences (Gaviria & Raphael, 2001).

Researchers have used various approaches to solve these issues, but there is no simple methodological answer to face the existing challenges (Calvó-Armengol et al., 2009). Manski (1993) shows that endogenous and exogenous effects cannot be separately identify in a linear-in-means model[2] due to the *reflection problem*. Thus by using this kind of econometric models only aggregate parameters are estimated (Sacerdote, 2001; Ammermueller & Pischke, 2009). Many empirical studies have addressed this issue imposing alternative structures or excluding effects on the original model. As another strategy, some use instruments to obtain consistent estimates of the endogenous peer effect (Evans et al., 1992; Gaviria & Raphael, 2001; Atkinson et al., 2008). The key here is the suitable choice of those variables which are correlated with the endogenous peer effect but not correlated with the error terms in the model.

With respect to correlated effects, some studies explicitly account for this source of bias. Researchers have used three main strategies to handle this problem. They have either exploited data where group members are randomly or quasi-randomly assigned within their groups (Angrist & Lavy, 1999; Boozer & Cacciola, 2001; Sacerdote, 2001; Zimmerman, 2003; Kang et al., 2007), they have used an instrumental variable strategy (Evans et al., 1992; Rivkin, 2001), or a family fixed effect strategy (Aaronson, 1998; Plotnick & Hoffman, 1999).

Bramoullé et al. (2009) consider an extended version of the linear-in-mean model where interactions are structured through a social network allowing the existence of correlated effects. By doing so they provide necessary and sufficient conditions for identification; such conditions generalize a number of previous results due to Manski (1993), Moffitt et al. (2001) and Lee (2007).

In Lee et al. (2010) the model proposed in Lee (2007) is extended to consider network structures and correlated disturbances among connected individuals. The possible endogeneity of the network is a particular concern in settings where peer effects hypothetically raise from networks that are

[2] In the linear in means model, the outcome of each individual depends linearly on his own characteristics, on the mean outcome of his reference group and on its mean characteristics.

formed by individuals making choices to establish links, because such endogenity may bias estimates. Goldsmith-Pinkham & Imbens (2013) and Hsieh & Lee (2016) propose correcting this selection bias by modelling the endogenous network formation process.

Considering the fact that the model specified in Lee (2007) adequately deals with the above mentioned difficulties, it has been used as reference in various empirical researches (Lin, 2010; Lee et al., 2010; Boucher et al., 2014), especially when studying peer influences in the school context. Therefore, unlike various strategies proposed to address the basic issues affecting peer effects estimations, the one developed by Lee (2007) has the advantage of fully identifying peer effects not requiring panel data or strong assumptions that are difficult to motivate and may not hold in practice (Boucher et al., 2014).

Finally, another source of bias in empirical research comes from the determination of reference groups. The choice of reference groups is often severely constrained by the availability of data. Consequently many studies of peer effects in education focus either on the grade-within-school level (Hoxby, 2000; Hanushek et al., 2003; Angrist & Lang, 2004), or analyse peer effects at the classroom level (Kang et al., 2007; Burke & Sass, 2008; Atkinson et al., 2008; Ammermueller & Pischke, 2009). The data set used in this research does not provide information on students social networks, but allows estimations at the classroom level.

This paper advances the literature on peer effects in education in Latin America and the Caribbean, providing, to our knowledge, the first application based on Lee (2007). Although there are a few other works that analyses peer effects in the region (McEwan, 2003; Dieye et al., 2014; De Melo, 2014; Mariño Fages, 2015), they do not use the same methodological approach. This social interaction model proposed in Lee (2007) considers group interaction and the existence of the three effects mentioned above (e.i. endogenous, exogenous and correlated effects).

3 Methodological approach and Econometric model

As mentioned previously, the model considered in this paper is the one proposed in Lee (2007), this model relies in two key assumptions. First, individuals interact in groups that are known for the modeller. Under our setting these groups are formed by classmates, so students are affected by all others in their groups (classrooms) but by none outside it. Second, individual outcome is determined by a linear-in-means model with group fixed effects. Thus, the test score of a student is affected by his characteristics and by the average test score and characteristics in his group of peers. In addition, it may be affected by any kind of correlated group-level unobservables.

Suppose there are R groups and there are m_r units in the rth group. At group level, the structural model is given by

$$Y_r \;\; = \;\; \lambda_0 W_r Y_r + X_{r1}\beta_{r10} + W_r X_{r2}\beta_{r20} + I_{m_r}\alpha_r + e_r, \quad r = 1, ..., R,$$

with $W_r = \dfrac{1}{m_r - 1}(l_{m_r} l'_{m_r} - I_{m_r})$ where l_{m_r} is the m_r-dimensional vector of ones, and I_{m_r} is the m_r-dimensional identity matrix. Y_r, X_{r1}, X_{r2} are the vector and matrices of the m_r observations in the rth group. Equivalently in terms of each unit i in a group r,

$$y_{ri} \;\; = \;\; \lambda_0 \Big(\frac{1}{m_r - 1} \sum_{j=1, j \neq i}^{m_r} y_{rj} \Big) + x_{ri,1}\beta_{10} + \Big(\frac{1}{m_r - 1} \sum_{j=1, j \neq i}^{m_r} x_{rj,2} \Big) \beta_{20} + \alpha_r + e_{ri},$$

with $i = 1, ..., m_r$, and $r = 1, ..., R$, where y_{ri} is the ith individual in the rth group, $x_{ri,1}$ and $x_{rj,2}$ are, respectively, k_1 and k_2-dimensional row vectors of exogenous variables, and e_{ri} are i.i.d $N(0, \sigma_0)$[3]. Variables $\Big(\dfrac{1}{m_r - 1} \sum_{j=1, j \neq i}^{m_r} y_{rj} \Big)$ and $\Big(\dfrac{1}{m_r - 1} \sum_{j=1, j \neq i}^{m_r} x_{rj,2} \Big)$ are the peer group means of the outcome and the exogenous variables respectively[4].

In section (2), following (Manski, 1993), we defined endogenous effects as the propensity of an individual to behave in some way varies with the behavior of the group, such effects are captured by parameter λ_0 because it reflects peers outcome influence. Exogenous (contextual) effects were defined as the propensity of an individual to behave in some way varies with the exogenous characteristics of the group, such effects are captured by parameters β_{20} because it reflects peers exogenous characteristics influence. The individual effects are given by the influence of the own exogenous variables, such effects are captured by parameter β_{10}. Correlated effects arise because individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar

[3] For G2SLS estimation we do not need to assume normality.

[4] As we can see in the summations, a student is not assumed to be one of his own peer. This creates individual variations in average peer attributes. These variations survive the elimination of common unobservables.

institutional environments, such effects are captured by α_r, it represents the mean unobservables of the r-th group. As those unonbservables may correlate with exogenous variables, they are treated as fixed effects.

The vector of all exogenous variables x_{ri}'s must vary across individuals in a group, as all group invariant variables will be captured in α_r. In a general setting, $x_{ri,1}$ and $x_{rj,2}$ are subvectors of x_{ri}, which may or may not have common elements.

Lee (2007) proposes two ways to estimate the model, generalized two-stage least squares (G2SLS) and conditional maximum likelihood (CML), and shows that the identification of endogenous and exogenous effects is possible if there are sufficient group size variation in the sample. The identification, however, can be weak if of all groups are of large sizes.

The model assumes that W_r is exogenous conditional on the unobserved effect α_r[5], i.e. $E(e_{ri}|x_{ri}, W_r, \alpha_r) = 0$. This assumption can accommodate many situations where W_r is endogenous. Suppose, for instance, W_r depends on unobserved common characteristics of the student's group (i.e. their preferences for sports, for physical infrastructure, and so on), the model admits this kind of correlation. Nevertheless, this assumption fails to hold, for instance, if some unobserved characteristics affect both the likelihood to be in the group (classroom) and the outcome, and *differs* among individuals in the same group.

4 Data

4.1 Third Regional Comparative and Explanatory Study (TERCE)

In recent years, quantitative research on students outcomes in Latin America and the Caribbean has benefited a lot from the growing availability of international comparable data. The Third Regional Comparative and Explanatory Study (TERCE) is an example of this kind of data source. Implemented in 2013 by UNESCO, TERCE is a large scale study of learning achievements carried out in 15 countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru and Uruguay, as well as in

[5] Under the group interaction assumption all students in a classroom are peers, so the conditional exogeneity of W_r is equivalent to the conditional exogeneity of the group size, m_r.

the Mexican state of Nuevo León. Its main goals are to provide information for the discussion on educational quality in the region and to orientate decision making in public policies. TERCE is the third study of its kind in primary education conducted by UNESCO Regional Bureau of Education for Latin America and the Caribbean, preceded in 1997 by the First Regional Comparative and Explanatory Study (PERCE) and in 2006 by the Second Regional Comparative and Explanatory Study (SERCE).

TERCE assessed the performance of pupils in third and sixth grades primary school in Mathematics and Language (reading and writing skills); and students achievements in Natural Sciences, in the case of sixth grade. In order to measure learning achievements, the study applied tests regarding common elements of the school curricula in the region. To assure cultural adaptation to each country and to prevent from imposing foreign standards, the design and implementation of the study was done following a collaborative process with participating countries (Flotts et al., 2015).

In addition to students academic performance, context questionnaires aiming to collect information on associated factors that influence student's learning achievements were also implemented. Among the variables considered in these questionnaires, importance was given to socio-economic context, family life and personal issues, as well as educational policies and school processes. Therefore, the study also collected data on the characteristics of students and their families, teachers, the school and its principal.

The TERCE data base consists of $N_T = 67,582$ observations on students which are grouped in $R_T = 3,115$ classrooms along the 15 countries and the state of Nuevo León[6].

In table (1) we present the total number of classrooms and the quartiles of the classroom sizes distribution by country.

[6] For an in depth description of TERCE's sample design and survey's contents refer to (Flotts et al., 2015).

Table 1: Classrooms and sizes. Original data.

Country	Number of classrooms	Quartiles of classroom sizes		
		Quartile 1	Quartile 2	Quartile 3
Argentina	207	14	20	26
Brazil	126	21	29	34
Chile	197	20	28	35
Colombia	149	23	31	36
Costa Rica	197	12	19	24
Dominicana	170	13	22	30
Ecuador	210	16	26	35
Guatemala	232	14	22	31
Honduras	203	10	18	28
Mexico	168	14	23	30
Nicaragua	180	9	22	31
Panama	187	15	20	26
Paraguay	205	10	17	25
Peru	285	7	16	25
Uruguay	238	12	19	24
Nuevo León	161	21	27	35

4.2 Variables

To analyze student's learning achievements, the dependent variables used are individual results on students mathematics, language and sciences tests[7]:

Score_math: irt standardized mathematics score.

Score_lang: irt standardized language score.

Score_scien: irt standardized sciences score.

Regarding explanatory variables, individual characteristics, family background and peer's influence were taken into account. Following the literature (Sacerdote, 2001; Gaviria & Raphael, 2001; Lin, 2005, 2010; Lee et al., 2010; Boucher et al., 2014), we consider these variables:

Isecf: standardized index of the economic, material and sociocultural condition of the student's household. This index is directly estimated by UNESCO, and to construct it information on the mother's education level and occupation, as well as household income and goods and services available at the house is collected.

[7] Estimated as the standardized score following the Item Response Theory (see Flotts et al. (2015) for a thorough explanation on how this scores are calculated).

Mothereduc: highest education level of the mother. This is a categorical variable using UNESCO's International Standardized Education Classificator (CINE-P, for its acronym in Spanish), which takes the following values:

- 1 Without studies
- 2 Primary school/Low secondary school [cine-p 1-2]
- 3 High secondary school [cine-p 3]
- 4 Post secondary education/Tertiary education [cine-p 4-5]
- 5 University [cine-p 6]
- 6 Master degree/Ph.D. [cine-p 7-8]

Age: student's age measured in years.

Gender: dummy variable taking value one if the student is male and zero if female.

Indigenous: dummy variable taking value one if at least one of these conditions is met and zero otherwise:

- the mother or father self-define themselves as indigenous
- at least one of the parents speaks an indigenous language
- parents speak in an indigenous language to the student
- the student self-defines him or herself as indigenous
- the student speaks in an indigenous language

Contextual effects: average values of all the explanatory variables over the student's classmates.

Endogenous effects: average result in tests of the student's classmates.

The following Table shows basic statistical measures for all the variables considered above.

Table 2: Descriptive statistics.

Variable	Mean	S.D.
Score_math	712.3	105.3
Score_lang	711.3	103.0
Score_scien	709.4	103.1
Age	12.41	0.940
Gender	0.503	0.499
Indigenous	0.234	0.423
Mothereduc	2.898	1.223
Isecf	0.142	1.047

4.3 Missing data treatment

As it happens in most surveys, many observations present missing data in some variables. The percentage of missing values in the total sample of sixth grade students is 5% for language score, 4% for mathematics score, 3% for sciences score, 15% for isecf index and 23% for mothereduc indicator. There are several methods to deal with missing data in the literature (Little, 1992; Pigott, 2001; Enders, 2010). Their accuracy depends crucially on the assumptions about the missing data mechanisms generating it. For missing values in explained variables we apply complete cases method, which consists of only using observations for which we have the value for the explained variable (language, mathematics or sciences scores). For missing data in explanatory variables we use random imputation.[8]

With respect to only using observations for which the explained variable is defined, while this is an accurate method when the mechanism generating missing data is random, in models where the explained variable is also used as explanatory variable it is not advisable to use it[9]. Given the fact that in this case overall missing data in explained variables is relatively small (5, 4 and 3 % respectively), we expect that any bias that could be introduced in estimates by using complete observations (complete on explained variables) shall be negligible.

Consequently, to estimate the models we only use cases in which we observe the explained variables. Besides, we also dismiss all observations from classrooms where the percentage of missing values in any variable (explained or explanatory) exceeds 50%; and those cases where there is only one student in the classroom. Furthermore, as the neighborhood violence level is one of the variable we

[8] See appendix 8.2 for details

[9] The model proposed in this paper has an important link with spatial economnetric models, in particular with the Saptial Lag Model. The treatment of missing data under the latter model has some particular issues, see Wang & Lee (2013), LeSage & Pace (2004) and Kelejian & Prucha (2010) for details

use to impute the mothereduc indicator and as this variable has missing values for a few classrooms, we drop observations from those classrooms.

Finally, as observations with missing values differs among subjects the final data bases are not the same. The final data bases for the three subjects consist in nearly 90% of the students and classrooms from the original sample[10].

5 Empirical results

We estimate the model for the whole sample and for different subsamples in order to capture some heterogeneities, we named such models ***Whole sample Model*** and ***Subsamples Model*** respectively. We present the estimates for both models in the following subsections.

5.1 Whole sample Model

Tables (3) and (4) display estimates of the proposed model for student's mathematics, language and sciences academic outcomes based on both, CML and B2SLS estimation methods respectively.

[10] See 8.1 and 8.3 for more details on missing data.

Table 3: CML Estimation.

	Random Imputation		
	Mathematics	Language	Sciences
Endogenous Effects	0.326***	0.102*	0.091
	(0.046)	(0.054)	(0.063)
Individual Effects			
Isecf	13.28***	12.61***	11.88***
	(0.872)	(0.886)	(0.988)
Age	-7.91***	-9.36***	-7.23***
	(0.600)	(0.683)	(0.692)
Mothereduc	2.69***	5.69***	5.82***
	(0.720)	(0.621)	(0.658)
Gender	12.41***	-8.59***	2.80***
	(0.957)	(1.040)	(0.980)
Indigenous	-2.80**	-7.27***	-4.32***
	(1.171)	(1.211)	(1.188)
Contextual Effects			
Isecf	14.56	-14.14	-10.65
	(14.26)	(15.41)	(17.39)
Age	12.08	14.03	14.17
	(9.089)	(11.83)	(12.40)
Mothereduc	-20.38	-5.94	1.08
	(13.80)	(10.55)	(11.23)
Gender	24.22	20.67	8.92
	(15.73)	(18.36)	(15.30)
Indigenous	28.55	-25.27	15.94
	(18.63)	(21.77)	(19.10)
$Corr^2(\hat{y}, y)$	0.345	0.148	0.142

Notes: Standard Errors in parenthesis.(***) indicates 1% significance level.
(**) indicates 5% significance level.(*) indicates 10% significance level.

Table 4: B2SLS Estimation

	Random Imputation		
	Mathematics	Language	Sciences
Endogenous Effects	1.406	-0.040	0.222
	(1.166)	(0.053)	(2.926)
Individual Effects			
Isecf	13.42***	12.62***	11.86***
	(0.988)	(0.904)	(1.122)
Age	-7.64***	-9.40***	-7.19***
	(0.771)	(0.619)	(1.146)
Mothereduc	2.30***	5.72***	5.80***
	(0.802)	(0.630)	(0.718))
Gender	12.50***	-8.58***	2.76**
	(1.132)	(1.004)	(1.166)
Indigenous	-2.49	-7.07***	-4.40**
	(1.801)	(1.551)	(1.860)
Contextual Effects			
Isecf	4.67	-11.81	-12.72
	(19.09)	(14.23)	(53.62)
Age	26.73	11.75	16.11
	(21.34)	(8.447)	(46.42)
Mothereduc	-33.37*	-4.32	-0.05
	(18.54)	(10.64)	(25.97)
Gender	15.32	19.36	7.74
	(16.72)	(14.99)	(24.02)
Indigenous	40.63	-21.76	14.98
	(33.06)	(23.75)	(26.08)
$Corr^2(\hat{y}, y)$	0.383	0.037	0.221

Notes: Standard Errors in parenthesis.(***) indicates 1% significance level.
(**) indicates 5% significance level.(*) indicates 10% significance level.

Regarding the estimation methods, results are quite similar for significant estimates in both methods. The main difference arise in the estimate of the endogenous effect, which turns out to be significant only under CML method.

The differences among estimation methods are consistent with theory, CML are more efficient than B2SLS, furthermore, B2SLS suffers from weak instrument problem. The instrument quality is positively related with the relation between the explained and the explanatory variables, and we can see in the results that such relation is poor.

As explained above, the methodological approach used here allows to account for the incidence of endogenous effects (i.e., the influence of peer outcomes), student's individual characteristics and contextual or exogenous effects (i.e., the influence of exogenous peer characteristics); while filtering fixed effects at the group level. These fixed effects include not only observable characteristics of the group (such as country of residence, school infrastructure, teacher's qualifications, etc.) but also unobservables, as well as common shocks faced by the group. Endogenous peer effects estimates are listed at the top of tables (3) and (4) and contextual effects at the bottom.

Before analyzing the social-interaction effects, a brief discussion of the performance of the control variables is necessary. Concerning personal background controls, i.e. student's age, gender and ethnicity, they are all statistically significant in determining academic performance. Student's age is negatively related to academic achievements for all the three subjects. This variable may be an indirect indicator of late schooling or grade repetition among students and consequently could be reflecting individual difficulties in school performance.

Regarding gender, results are consistent with the empirical literature (Hyde et al., 1990). Female students tend to outperform males in language, while males students achieve better results than females in mathematics and sciences tests. Turning to ethnicity, results indicate that students with indigenous background achieve poorer academic results than the rest, which is in line with previous research (Verdisco et al., 2009). Furthermore, this disadvantage seems stronger when it comes to language outcomes possibly indicating that indigenous children suffer from idiomatic limitations that condition their academic achievements (Flotts et al., 2015). Finally, family sociocultural and economic condition as well as mother education, all have positive significant influence in student's academic achievements, reinforcing existing findings on these topics (Davis-Kean, 2005).

With respect to those effects that surge from social interaction, exogenous peer characteristics or contextual effects do not significantly influence student's academic outcomes. However, the fact that variables reflecting peer's characteristics turn out not to be statistically significant may be related to their high correlation with student's own ones.

On the other hand, endogenous peer effects, under CML estimation, prove to be statistically significant for both mathematics and language academic outcomes, though they are not in the case of sciences results. Endogenous peer effects in mathematics scores are highly significant and somehow stronger than in language. This may reflect the fact that mathematics provide more opportunities for interactions among students. Nevertheless, peer outcomes also impact language tests' results at 10% significance. These findings are in accordance with previous empirical studies (Boucher et al., 2014; Carrell et al., 2009; Hoxby, 2000; Hanushek et al., 2003; Vigdor & Nechyba, 2004; Zimmer &

Toma, 2000).

It is worth to note, however, that the model does not explain much of the variability of the data[11] suggesting the existence of other factors that may explain student's academic performance besides those explicitly considered here. Even so, it is clear that classmates academic outcomes do affect student's performance at school and therefore attention should be paid to these findings.

5.2 Subsamples Model

The model proposed up to here does not allow interactions between the endogenous peer effect and other factors, i.e. we have the same endogenous peer effect for the whole population. This assumption can be restrictive, specially under such a heterogeneous population.

One of the sources of heterogeneity in peer effects is the type of school. The sample considers three types of schools, namely, urban public, urban private, and rural. They differ in many aspects, such as infrastructure, socio-economic level, financial access and facilities among others. Rurality is associated with greater poverty and their access to resources and infrastructure is lower compared to schools in urban areas (UNESCO, 2016). To gain some insight into heterogeneities we estimate the proposed model for the three different type of schools. We do that in the same way as before, i.e. under two estimation methods and three missing data manage alternatives.

With respect to sample sizes, rural schools are about 42% of the schools in the final data base, whereas urban private and urban public schools are 25% and 32% respectively. Regarding to students, 34% attend rural schools, whereas 28% and 38% attend urban private and urban public schools respectively. The results are shown below.

[11] Based on $Corr^2(\hat{y}, y)$.

Table 5: CML Estimation.

	Mathematics			Language			Science		
	Urban Public	Urban Private	Rural	Urban Public	Urban Private	Rural	Urban Public	Urban Private	Rural
Endogenous Effects	0.363***	0.400***	0.008	-0.216***	0.505***	-0.060	0.249***	0.469***	-0.148
	(0.034)	(0.055)	(0.070)	(0.041)	(0.051)	(0.048)	(0.042)	(0.074)	(0.111)
Individual Effects									
Isecf	11.97***	11.73***	10.05***	11.56***	7.37***	12.53***	11.44***	10.31***	10.96***
	(2.028)	(1.923)	(1.326)	(2.232)	(2.341)	(1.389)	(2.145)	(2.036)	(1.450)
Age	-6.52***	-6.72***	-8.01***	-10.42***	-6.57***	-9.12***	-7.35***	-4.32***	-6.93***
	(1.573)	(1.559)	(0.820)	(1.177)	(1.655)	(0.848)	(1.454)	(1.672)	(0.932)
Mothereduc	5.13***	2.10	2.52**	9.50***	6.04***	3.72***	10.12***	5.38***	4.31***
	(1.429)	(1.336)	(1.218)	(1.671)	(1.173)	(1.133)	(1.708)	(1.344)	(1.137)
Gender	12.17***	15.93***	9.90***	-7.75***	-11.10***	-6.89***	3.05	2.94	1.66
	(2.255)	(2.072)	(1.398)	(2.212)	(1.908)	(1.560)	(2.324)	(2.146)	(1.485)
Indigenous	-4.30*	-2.01	-1.67	-5.81**	-2.34	-8.56***	-5.41**	-2.14	-4.36**
	(2.616)	(2.653)	(1.656)	(2.492)	(2.405)	(1.785)	(2.721)	(2.820)	(1.762)
Contextual Effects									
Isecf	-81.20*	34.65	16.07	-109.46**	-78.94	13.29	-74.95	6.26	-8.50
	(46.58)	(15.36)	(17.95)	(54.21)	(49.52)	(17.76)	(49.54)	(36.43)	(20.46)
Age	67.43*	25.35	-2.07	12.18	50.50	9.99	54.94*	28.77	8.37
	(37.25)	(8.672)	(9.331)	(23.63)	(30.82)	(10.40)	(33.55)	(28.91)	(13.53)
Mothereduc	23.59	26.24	-18.59	98.36**	-6.94	-30.22*	102.97**	-13.33	-7.02
	(32.39)	(11.68)	(18.58)	(41.63)	(20.99)	(16.80)	(42.93)	(25.42)	(15.76)
Gender	14.97	41.03	9.22	12.88	40.21	24.01	-31.81	50.83	1.15
	(49.00)	(15.43)	(16.31)	(47.79)	(33.33)	(20.17)	(50.51)	(42.17)	(17.18)
Indigenous	15.99	45.00	30.46	1.94	108.18**	-53.59*	6.81	39.95	11.44
	(61.37)	(24.49)	(24.15)	(57.16)	(48.67)	(27.63)	(63.16)	(47.85)	(24.46)
$Corr^2(\hat{y}, y)$	0.002	0.243	0.090	0.046	0.030	0.077	0.001	0.209	0.026

Notes: Standard Errors in parenthesis.(***) indicates 1% significance level.(**) indicates 5% significance level.(*) indicates 10% significance level.

Table 6: B2SLS Estimation.

	Mathematics			Language			Science		
	Urban Public	Urban Private	Rural	Urban Public	Urban Private	Rural	Urban Public	Urban Private	Rural
Endogenous Effects	2.064**	0.448	0.397	-1.399	1.005**	-0.500***	0.985	0.728	-9.616***
	(0.858)	(1.061)	(1.214)	(1.793)	(0.488)	(0.148)	(0.748)	(1.680)	(0.201)
Individual Effects									
Isecf	10.61***	11.73***	10.17***	13.70***	6.969***	12.33***	10.73***	10.31***	12.14***
	(2.295)	(2.056)	(1.470)	(3.252)	(2.198)	(1.356)	(2.257)	(2.153)	(1.192)
Age	-5.04***	-6.70***	-8.00***	-9.82***	-6.239***	-9.34***	-7.01***	-4.22**	-10.81***
	(1.854)	(1.854)	(0.852)	(1.178)	(1.803)	(0.829)	(1.653)	(2.069)	(0.714)
Mothereduc	4.61***	2.09*	2.28*	9.27***	6.107***	4.05***	10.70***	5.42***	8.17***
	(1.515)	(1.302)	(1.375)	(1.561)	(1.215)	(1.079)	(1.691)	(1.336)	(0.921)
Gender	12.22***	15.98***	9.75***	-6.25**	-10.865***	-6.91***	2.86	3.19	11.87***
	(2.541)	(2.681)	(1.553)	(2.436)	(2.221)	(1.530)	(2.509)	(2.996)	(1.307)
Indigenous	-4.59	-1.95	-1.56	-7.02**	-1.781	-7.33***	-4.39	-2.07	8.17***
	(4.312)	(3.636)	(2.532)	(3.904)	(3.624)	(2.214)	(4.337)	(3.807)	(1.928)
Contextual Effects									
Isecf	-141.57**	16.03	14.98	-36.69	-94.15**	14.84	-103.53*	3.94	127.36***
	(58.63)	(35.55)	(17.76)	(107.25)	(44.79)	(16.73)	(55.78)	(39.63)	(14.76)
Age	120.83**	27.17	1.08	13.04	62.69*	1.90	70.99*	32.67	-130.55***
	(48.77)	(37.91)	(15.56)	(23.42)	(33.04)	(9.92)	(41.97)	(44.60)	(8.01)
Mothereduc	3.52	-28.61	-24.68	98.60***	-8.67	-21.33	113.82***	-13.89	112.95***
	(34.06)	(26.15)	(25.42)	(32.80)	(21.82)	(14.31)	(37.34)	(28.77)	(11.88)
Gender	-2.54	67.65	3.40	39.02	52.21	19.27	-39.60	56.47	181.18***
	(54.31)	(44.37)	(23.03)	(48.16)	(44.07)	(18.38)	(54.05)	(60.05)	(14.87)
Indigenous	17.80	51.46	34.38	-35.63	125.82*	-33.69	37.160	42.96	174.72***
	(99.13)	68.39)	(37.79)	(94.15)	(67.66)	(26.05)	(101.21)	(71.34)	23.47)
$Corr^2(\hat{y}, y)$	0.127	0.257	0.261	0.104	0.000	0.024	0.032	0.229	0.173

Notes: Standard Errors in parenthesis.(***) indicates 1% significance level.(**) indicates 5% significance level.(*) indicates 10% significance level.

The conclusions about different estimation methods are the same as before. As regards estimates, the impact of controls variables are similar in both models (homogeneous and heterogeneous): age and ethnicity have a negative impact when significant; males have better results in mathematics, females in language, and gender is not significant in science results, while mother education and isecf show a positive impact in almost every estimation; and contextual effects are mostly not significant. Focusing on heterogeneity by school types, ethnicity has no impact in urban private schools, the effect of mother education seems to be stronger in urban public schools and isecf appears to be weaker in urban private schools but only for language. Endogenous peer effect are greater for urban private schools whereas for rural schools it is mostly not significant. Even though such differences depend on the subject, the are more noticeable in language followed by science.

6 Concluding remarks

Estimating peer effects is challenging because of the existence of the reflection problem, omitted variable bias problem as well as data limitation. In this paper we applied the social interaction model proposed in Lee (2007) to academic results of primary school students. This model allows separate estimations of endogenous and exogenous effects, while controlling for correlated effects. The results found in this research add empirical evidence that supports the hypothesis of peer effects in education, affecting in this particular application primary school attendants in Latin America and the Caribbean. We show that peer influence plays a significant role in early education academic achievements, mainly through endogenous effects. Considering the multiplying characteristics of these effects, results found here are important from a public policy perspective. On the other hand, we explore some heterogeneities showing that the impact of some control variables, as well as the endogenous effect, not only depend on the subject, but they are also of different magnitudes depending on the school types.

Hopefully, this paper has contributed to a better visualization of the impacts of social interactions in human capital accumulation. These results may add new inputs to be considered in the educational policy agenda of the region. Undoubtedly, the issues regarding the accumulation of human capital are sure to remain a fertile ground for future research. In fact, we expect to extend this research to third grade pupils also evaluated in TERCE as well as investigating possible non-linearities and other sources of heterogeneities of the effects so as to achieve a more precise picture of peer effects influence on students academic performance in the region.

7 References

References

Aaronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources*, (pp. 915–946).

Ammermueller, A. & Pischke, J.-S. (2009). Peer effects in european primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, 27(3), 315–348.

Andrews, D., Green, C., & Mangan, J. (2002). Neighbourhood effects and community spillovers in the australian youth labour market. *LSAY Research Reports*, (pp.2̃8).

Angrist, J. D. & Lang, K. (2004). Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, 94(5), 1613–1634.

Angrist, J. D. & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2), 533–575.

Atkinson, A., Burgess, S., Gregg, P., Propper, C., Proud, S., et al. (2008). The impact of classroom peer groups on pupil gcse results. *Centre for Market and Public Organiziation Working Paper*, 8, 187.

Becker, G. S. (1994). Human capital revisited. In *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition)* (pp. 15–28). The university of Chicago press.

Boozer, M. & Cacciola, S. E. (2001). Inside the'black box'of project star: Estimation of peer effects using experimental data.

Boucher, V., Bramoullé, Y., Djebbari, H., & Fortin, B. (2014). Do peers affect student achievement? evidence from canada using group size variation. *Journal of applied econometrics*, 29(1), 91–109.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1), 41–55.

Burgess, S. M. (2016). Human capital and education: The state of the art in the economics of education.

Burke, M. & Sass, T. (2008). Vclassroom peer effects and student achieve'mentv. *Federal Reserve Bank of Boston Working Paper*, (08), 5.

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4), 1239–1267.

Carrell, S. E., Fullerton, R. L., & West, J. E. (2009). Does your cohort matter? measuring peer effects in college achievement. *Journal of Labor Economics*, 27(3), 439–464.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & Robert, L. (1966). York. 1966. *Equality of educational opportunity*, 2.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2), 294.

De Melo, G. (2014). *Peer effects identified through social networks: Evidence from Uruguayan schools*. Technical report, Working Papers, Banco de México.

Dieye, R., Djebbari, H., & Barrera-Osorio, F. (2014). Accounting for peer effects in treatment response.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Epple, D. & Romano, R. E. (1996). Public provision of private goods. *Journal of political Economy*, 104(1), 57–84.

Evans, W. N., Oates, W. E., & Schwab, R. M. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy*, 100(5), 966–991.

Flotts, M. P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C., & García, M. J. (2015). *Informe de resultados TERCE: logros de aprendizaje*. UNESCO Publishing.

Gaviria, A. & Raphael, S. (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics*, 83(2), 257–268.

Goldsmith-Pinkham, P. & Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3), 253–264.

Greene, J. P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The milwaukee experiment. *Education and Urban Society*, 31(2), 190–213.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, (pp. 351–388).

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of applied econometrics*, 18(5), 527–544.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Does special education raise academic achievement for students with disabilities?* Technical report, National Bureau of Economic Research.

Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation.* Technical report, National Bureau of Economic Research.

Hsieh, C.-S. & Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2), 301–319.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, 107(2), 139.

Kang, C. et al. (2007). Does money matter? the effect of private educational expenditures on academic performance. *National University of Singapore. Department of Economics Working Paper*, 704.

Kelejian, H. H. & Prucha, I. R. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of geographical systems*, 12(3), 241–257.

Lee, L.-f. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2), 333–374.

Lee, L.-f., Liu, X., & Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2), 145–176.

LeSage, J. P. & Pace, R. K. (2004). Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, 29(2), 233–254.

Lin, X. (2005). Peer effects and student academic achievement: an application of spatial autoregressive model with group unobservables. *Unpublished manuscript, Ohio State University.*

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics*, 28(4), 825–860.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420), 1227–1237.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3), 531–542.

Mariño Fages, D. (2015). Efecto de pares en el desempeño académico de alumnos de primaria y secundaria. *L Reunión Anual Asociación Argentina de Economía Política.*

McEwan, P. J. (2003). Peer effects on student achievement: Evidence from chile. *Economics of education review*, 22(2), 131–141.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Moffitt, R. A. et al. (2001). Policy interventions, low-level equilibria, and social interactions. *Social dynamics*, 4(45-82), 6–17.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4), 353–383.

Plotnick, R. D. & Hoffman, S. D. (1999). The effect of neighborhood characteristics on young adult outcomes: Alternative estimates. *Social Science Quarterly*, (pp. 1–18).

Rivkin, S. G. (2001). Tiebout sorting, aggregation and the estimation of peer group effects. *Economics of Education Review*, 20(3), 201–209.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2), 681–704.

Stinebrickner, R. & Stinebrickner, T. R. (2006). What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics*, 90(8-9), 1435–1454.

UNESCO, O. (2016). Informe de resultados. terce: Tercer estudio regional comparativo y explicativo. factores asociados. resumen ejecutivo. *Perfiles Educativos*, 38(152).

Verdisco, A., Cueto, S., Thompson, J., Engle, P., Neuschmidt, O., Meyer, S., González, E., Oré, B., Hepworth, K., & Miranda, A. (2009). Urgency and possibility results of pridi a first initiative to create regionally comparative data on child development in four latin american countries technical annex.. *Technical Annex. Inter-American Development Bank, Washington DC.*

Vigdor, J. & Nechyba, T. (2004). Peer effects in elementary school: Learning from ?apparent?random assignment. *Unpublished manuscript*.

Wang, W. & Lee, L.-F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *The Econometrics Journal*, 16(1), 73–102.

Zimmer, R. W. & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 19(1), 75–92.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, 85(1), 9–23.

8 Appendix

8.1 Missing data descriptions

As we mentioned, the percentage of missing values in the total sample of sixth grade students is 5% for language score, 4% for mathematics score, 3% for sciences score, 15% for isecf index and 23% for mothereduc indicator. The overall percentage of *classrooms* with at least one missing value in language score, mathematics score, sciences score, isecf index and mothereduc indicator are 44%, 36%, 30%, 60% and 86% respectively. Despite the fact that the number of classrooms with missing data is high (specially for explanatory variables), the percentage of missing values within classrooms is considerable lower. In fact, the 80% of classrooms with missing data in language, mathematics and sciences scores do not have more than 8%, 6% and 5% of missing values respectively; whereas the 80% of classrooms with missing values of isecf index and mothereduc indicator do not exceed 20% and 33% of missing values respectively.

The aforementioned information on missing data concerns the sample as a whole, but the proportion of missing values varies considerably between countries, classrooms and variables. To get some insights in the distribution of missing values we calculate both, the percentage of missing values by country and the distribution of the percentage of missing values by classroom in each country. We report the 8th quantile of such distributions.

Table 7: Missing data by country and classrooms in explained and explanatory variables.

Country	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Argentina	64	17	66	16	62	15	89	38	95	46
Brazil	95	25	97	30	94	25	98	40	99	53
Chile	50	6	51	5	50	7	76	22	90	29
Colombia	46	5	34	3	42	4	47	8	91	16
Costa Rica	13	0	23	3	13	0	23	4	76	15
Dominicana	17	0	36	5	14	0	43	7	92	25
Ecuador	20	2	16	0	21	2	43	9	80	20
Guatemala	23	3	41	8	19	0	70	100	88	100
Honduras	16	0	38	6	15	0	44	11	80	28
Mexico	39	4	45	7	37	4	60	15	82	22
Nicaragua	36	5	68	15	30	3	68	20	86	33
Panama	63	16	62	12	52	10	76	26	92	37
Paraguay	37	7	46	10	35	6	65	22	91	40
Peru	14	0	24	4	11	0	38	10	73	22
Uruguay	32	5	36	6	32	5	70	100	86	100
Nuevo León	42	5	53	6	35	4	66	13	89	18

(1) % Classrooms with missing data in Mathematics score. (2) % Missing data in Mathematics score by classrooms, Q_8 . (3) % Classrooms with missing data in Language score. (4) % Missing data in Language score by classrooms, Q_8. (5) % Classrooms with missing data in Sciences score. (6) % Missing data in Sciences score by classrooms, Q_8 . (7) % Classrooms with missing data in isecf index. (8) % Missing data in isecf index by classrooms, Q_8. (9) % Classrooms with missing data in mothereduc indicator . (10)% Missing data in mothereduc indicator by classrooms, Q_8.

The percentage of classrooms with missing data in explained variables shows wide variability when measured by country. Regarding mathematics score it ranges from 13 to 95%, for language score it goes from 16 to 97 %, while in the case of sciences score it varies from 11 to 94%. In almost every country the percentage of classrooms with missing values in language score is slightly grater than the percentage of classrooms with missing data in mathematics and sciences scores. Something worth noting is that, regardless the number of classrooms with missing data, the percentage of missing values by classroom is relatively low. For instance, the 1st, 3rd and 5th columns of table (7) show that Brazil has missing values in almost every classroom in all the three subjects (mathematics, language and sciences scores), but 80% of such classrooms do not have more than 25 or 30% of missing data in those variables respectively.

As regards explanatory variables, they show more classrooms with missing data as well as a higher number of missing information by classroom.

8.2 Random imputation

Supposing that the missing problem is confined to a single variable, y, and that we observe a set of variables X for all units, then the method consists in estimating a regression model based on observed data. As we know all X, we impute the missing y using the estimated model.

Let y^o and y^u be the observed and unobserved y respectively, we estimate $y^o = \beta X^o + e^o$, where $e \sim N(0, \sigma_e)$, and then we impute the missing y by $\hat{y}^u = \hat{\beta} X^u + \hat{e}^u$ (consider we completely observe X). It is worth noting that we add an error term, $\hat{e}^u$, to the imputed values $\hat{y}^u$ (hence the name *random imputation*), which is generated by simulating their distribution, $\hat{e}^u \sim N(0, \hat{\sigma}_e)$.

The model we use to impute isecf index and mothereduc indicator when these variables show missing data has the following structure,

$$y_{ir} \quad = \quad \beta_1 \overline{y}_{-ir} + \beta_2 x_{2,ir} + \beta_3 x_{3,ir} + e_{ir}$$

where y_{ir} is the y value (isecf index or mothereduc indicator) for the $i-th$ student in $r-th$ classroom, $\overline{y}_{-ir}$ is the mean of y in classroom r without considering y_i, $x_{2,ir}$ is the mean of isecf index and it is present in both model, whereas $x_{3,ir}$ is the kind of school (public or private) in the model for isecf index, and the level of neighbor violence in the model for mothereduc indicator, $e_{ir} \sim N(0, \sigma_e)$ is an error term.[12] The intra-classroom autocorrelation of isecf index and mothereduc indicator is relatively high. That is why we use $\overline{y}_{-ir}$ as explanatory variable. However, this triggers another issue because the variable $\overline{y}_{-ir}$ is a classroom mean of the partially observed variable y, so it is also partially observed. We ignore this fact because the goal here is not causal inference but simply accurate prediction. Therefore it is acceptable to use any input in the imputation model to achieve this goal, and given $\overline{y}_{-ir}$ is helpful for explaining y, we consider it in the model.

8.3 Final data

As mentioned previously, to estimate the models we dismiss some observations due to missing data problems. The observations with missing values in mathematics score differ from those with missing

[12] We have selected the explanatory variables in order to maximize the R^2.

values in language score and sciences score, so the final data bases used for each subject differ.

The mathematics data base consists of $N_m = 58,817$ observations (students) which are grouped in $R = 2,736$ classrooms. That is, we work with the 87% of observations and with the 88% of classrooms from the original data. The overall percentage of missing values in isecf index and mothereduc indicator is 7 and 15% respectively. Whereas the overall percentage of *classrooms* with some missing value in isecf index and mothereduc indicator is 53 and 84% respectively. The 80% of classrooms with missing data of isecf index and mothereduc indicator do not have more than 12 and 24% of missing values respectively.

The language data base consists of $N_l = 58,224$ observations (students) which are grouped in $R = 2,730$ classrooms. Consequently, we work with the 86% of the observations and with the 88% of classrooms from the original data. The overall percentage of missing values in isecf index and mothereduc indicator is 5 and 13 % respectively. The total percentage of *classrooms* with some missing value in isecf index and mothereduc indicator is 44 and 83% respectively. The 80% of classrooms with missing data in isecf index and mothereduc indicator do not have more than 10 and 22% of missing values respectively.

The sciences data base consists of $N_c = 59,051$ observations (students) which are grouped in $R = 2,737$ classrooms. Consequently, we work with the 87% of the observations and with the 88% of classrooms from the original data. The overall percentage of missing values in isecf index and mothereduc indicator is 7 and 15 % respectively. The total percentage of *classrooms* with some missing value in isecf index and mothereduc indicator is 53 and 84% respectively. The 80% of classrooms with missing data in isecf index and mothereduc indicator do not have more than 12 and 25% of missing values respectively.

Given the fact that the percentage of missing data varies across countries, the *missing filtering process* impacts differently on each country data. In the following lines we present some measures on missing data distribution by country and by subject.

Table 8: Classrooms, sizes and missing data distribution. Reduced sample.

Country	Mathematics data						Language data					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Argentina	80.2	19.0	83.1	25.0	93.4	33.3	79.7	19.0	80.6	25.0	92.1	33.3
Brazil	71.4	27.0	80.0	20.0	96.7	33.3	71.4	25.5	71.1	16.7	95.6	29.5
Chile	91.4	28.0	72.8	17.3	90.0	24.4	91.4	28.0	72.2	17.4	90.0	23.9
Colombia	97.3	30.0	42.8	6.3	90.3	15.4	97.3	31.0	38.6	6.2	89.7	15.4
Costa Rica	98.5	19.0	21.6	3.5	74.7	15.0	98.5	19.0	19.6	0.0	74.7	15.0
Dominicana	94.7	22.0	41.6	6.2	92.5	24.3	94.7	22.0	23.6	3.0	91.9	23.5
Ecuador	88.6	22.5	59.8	10.9	83.3	21.1	88.6	26.0	31.7	4.0	79.0	15.4
Guatemala	75.0	22.5	59.8	10.9	83.3	21.1	75.0	22.0	32.8	4.6	76.4	16.7
Honduras	91.6	19.0	42.5	8.3	79.0	23.1	91.1	18.0	30.3	4.0	76.8	20.1
Mexico	91.7	23.0	55.2	13.6	82.5	20.3	92.3	23.0	45.8	8.7	76.8	18.2
Nicaragua	91.1	22.0	68.3	15.5	86.0	25.3	90.0	21.0	37.0	5.0	83.3	20.0
Panama	89.3	18.0	67.1	16.7	91.0	30.3	88.8	19.0	60.8	16.0	88.0	26.7
Paraguay	86.3	16.0	55.4	14.3	88.7	30.7	85.9	16.0	43.8	9.1	88.6	28.6
Peru	95.1	17.0	37.3	8.3	72.7	20.0	95.1	16.0	32.1	6.7	71.6	20.0
Uruguay	68.5	18.0	49.7	13.5	77.3	23.1	68.1	18.0	48.8	12.5	78.4	23.0
Nuevo León	98.1	27.0	64.6	12.0	88.0	17.5	98.1	27.0	55.1	10.6	86.1	15.9

(1) % of classrooms from the complete sample. (2) Median of classroom size. (3) % classrooms with missing values in isecf index. (4) % missing data in isecf index by classrooms, Q_8. (5) % classrooms with missing data in edumother indicator. (6) % missing data in edumother indicator by classrooms, Q_8.

Table 9: Classrooms, sizes and missing data distribution. Reduced sample.

Country	Sciences data					
	(1)	(2)	(3)	(4)	(5)	(6)
Argentina	80.2	19.0	83.1	25.8	94.0	33.3
Brazil	71.4	26.5	80.0	19.2	96.7	33.3
Chile	91.4	28.0	73.3	17.8	90.0	25.0
Colombia	97.3	31.0	42.8	6.3	89.7	15.5
Costa Rica	98.5	19.0	21.1	3.5	74.2	15.0
Dominicana	94.7	22.0	41.6	6.2	92.5	25.0
Ecuador	88.6	26.0	36.0	4.8	79.6	15.0
Guatemala	75.0	22.5	59.8	10.9	83.3	21.1
Honduras	91.6	19.0	42.5	8.3	79.0	23.1
Mexico	92.3	23.0	54.8	13.6	81.9	20.2
Nicaragua	91.1	23.0	68.9	16.2	86.0	27.8
Panama	89.3	19.0	68.9	17.2	91.0	31.2
Paraguay	86.3	16.0	55.9	14.3	89.3	30.0
Peru	95.1	17.0	37.3	8.3	72.7	20.0
Uruguay	68.5	18.0	49.7	14.1	76.7	22.7
Nuevo León	98.1	27.0	65.2	12.3	88.6	17.2

(1) % of classrooms from the complete sample. (2) Median of classroom size. (3) % classrooms with missing values of isecf index. (4) % missing data in isecf index by classrooms, Q_8. (5) % classrooms with missing data in edumother indicator. (6) % missing data in edumother indicator by classrooms, Q_8.