SAPIENZA
UNIVERSITÀ DI ROMA

The Location of Immigrants: Data-Driven Analysis of Network and Local Effects

Facoltà di Ingegneria dell'informazione, informatica e statistica

Corso di Laurea Magistrale in Data Science

Candidate

Sara Romiti
ID number 1550825

Thesis Advisors

Prof. Aris Anagnostopoulos
Prof. Ioannis Chatzigiannakis

Academic Year 2017/2018

Thesis defended on
in front of a Board of Examiners composed by:

(chairman)

The Location of Immigrants: Data-Driven Analysis of Network and Local Effects

Master thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: July 20, 2018

Author's email: s.romiti01@gmail.com

Contents

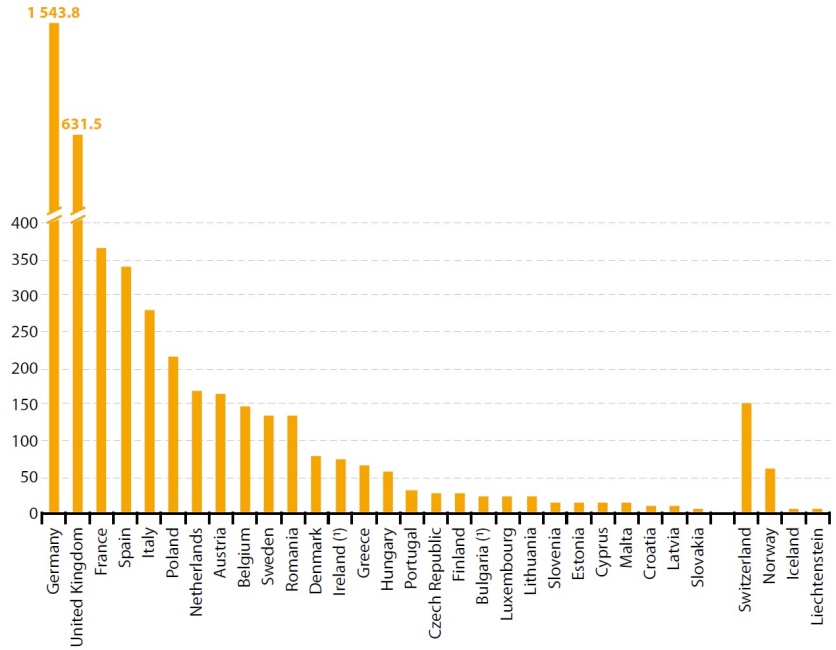
Chapter 1

Introduction

After a long tradition of emigration until the 1970s, Italy has suddenly turned into an immigration country initially in 1980s, but more prominently in 1990s.

Even if the phenomenon in Italy is relatively new with respect to other European countries, it reached the *"historical migrants destination"* (e.g. Germany, United Kingdom, France) in the 2015 EUROSTAT ranking.

Number of immigrants, 2015 (thousands)



The first massive migration to Italy goes back to 1991 from Albania, after the breakdown of the communist regime, when 27.000 migrants arrived in a single day. Italy had been the main immigration target for Albanians leaving their country, this may due to the fact that an enclave of Albanians population exists in the South of Italy since the XV century.

As of January 1st 2017, immigrants resident in Italy are about 5.144.440 and represent 8.5% of the total population.

As Italy, Spain has recently experienced large-scale immigration for the first time in modern history while nowadays, within Europe, it is one of the prefered destination country for immigrants. The Spanish economic development, particularly since 1986, when the country entered the European Union, led to a large transformation. From being a territory of emigrants it became a country of immigration.

The growth rate has intensified since 1996 and especially exponentially since 2000. There are 4.572.807 foreign residents in Spain today, it represents the 9.8% of the total population.

The aim of this thesis project is to predict the foreigners-born population legal flow and distribution across Italy and Spain territories from official data, i.e. how a given foreigner population distributes across Italian regions (or Spanish autonomous communities) and how additional variables may influence this choice. The reason for the study is that migration has become a major determinant of demographic change in the EU. It is one of the five demographic challenges that The Commission's Communication of the European Union on The demographic future of Europ)e — from challenge to opportunity identified. Also, EUROSTAT is focusing on migration projection between now and 2050.

The analysis focuses on Italy and Spain since it was not possible to find compatible data from any other EU country. Since it was not possible to retrieve data about irregular flows, all the data refer to regular immigrants.

The thesis is structured as follows. In the next section, we introduce the data and carry out some first level analysis based on data visualizations. In Section 3 we explain briefly the models used in the study then, in Section 4, we present the results of the models applied to the data. Section 5, concludes.

Chapter 2

Data

In this section, we will present the Italy and Spain dataset. Also, we will show a descriptive analysis of the two datasets based on the visualizations.

2.1 Italy

The only available data at a regional level we could retrieve are from ISTAT [15]. On ISTAT website it is possible to find some useful information about the yearly foreigners' population on January 1st, such as the Italian destination territory, the foreigners origin countries, the data collection year and the gender. Similar information is also available for the resident population, i.e. native plus foreigners population.

The data are collected from 2003 on, but until 2011 the values are not the results of continuous observation, they are statistical reconstruction between the two censures of 2001 and 2011.

2.1.1 Data Quality

The ISTAT dataset seems to be the more accurate for the data concerning Italy. Even though it is a good open data sources, it has some problems. The website is not user-friendly: the data labels are not always sufficiently explanatory, e.g. *resident population* includes the foreigners or is it just about the native? To answer the question, we needed to read an ISTAT yearly report. The flag legend is not clear, e.g. what does *data not reconstructed with respect to the census population 2011* mean? The tables structure is not always coherent, the resident population before 2011 is a separated table with respect to the one after 2011 while the data about foreigners population are together. Moreover it is not very clear what ISTAT considers as a country. It includes Palestine and Kosovo as separated countries because they both are partially recognized states. The two territories also have a cultural and religious independence with respect to Israel and Serbia. But it does not consider other partially recognized states like South Ossetia and Abkhazia or autonomous communities like Kurdistan and Basque Country. From a statistical point of view it may be more interesting to consider separately all population with their own culture, language and national identity even if they are not politically recognized.

ISTAT organizes the Italian territories into the following five zones:

- Center
- Islands
- North East
- North West
- South

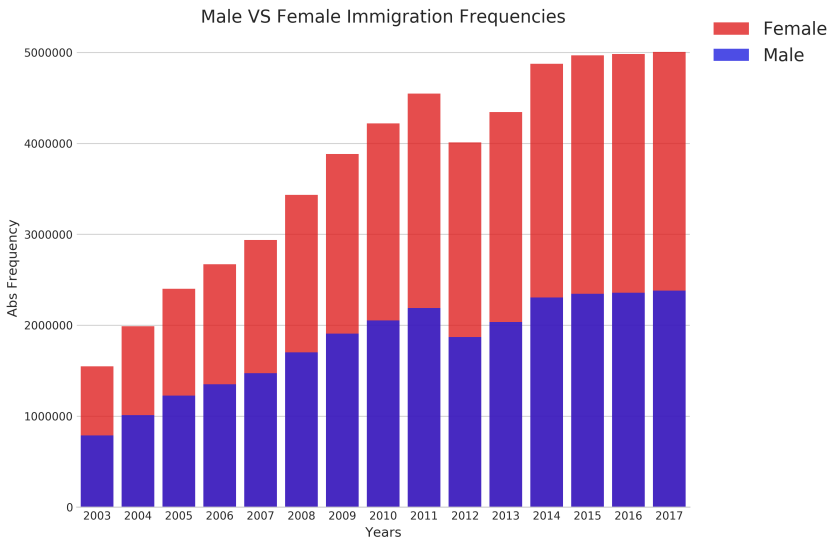
Moreover, the zones are divided into regions:

- Abruzzo
- Aosta Valley
- Apulia
- Basilicata
- Calabria
- Campania
- Emilia-Romagna
- Friuli-Venezia Giulia
- Lazio
- Liguria
- Lombardy
- Marches
- Molise
- Piedmont
- Sardinia
- Sicily
- South Tyrol
- Trentino
- Tuscany
- Umbria
- Veneto

Thus, the statistical initial analysis is referred to those territories.

2.1.2 Dataset Distributions

As first, we study the gender distribution over the years, aggregated both on the destination and the origin territories.



From this stacked barplot, the distribution across the gender seems to be quite balanced. Thus, the data will be aggregated for the study.

The aim is now to show the distributions of the foreign-born citizens across the five Italian zones. They are studied without distinction over the origin country.

Initially, the total foreign-born population is analyzed as follows:

- the absolute value distribution
- the growth distribution
- the growth rate distribution, i.e. $\frac{value_t - value_{t-1}}{value_{t-1}}$

Based on the above figures, even if the growth rate is significantly decreasing, the immigration is still an increasing phenomenon. Observe that in 2014, the growth rate is notably reduced. A possible explanation could be that in the same year (07/10/2014) there was the first agreement[12] between the Italian government, the regions and local authorities on a national level to face the flow of foreign-born population.

Furthermore, between 2010 and 2011 there is a significant decrease in the growth rate. This could be justified by the fact that the data before 2011 is a statistical

estimation, while from 2011 is an actual continuous observation. Thus, ISTAT may have overestimated the phenomenon.

While the growth trend is quite similar in the five zones, the absolute value is not. Three different trends could be distinguished from this analysis:

- North West - initially between 2003 and 2011 the absolute population is higher than any other region. After 2014 the growth rate seems to stabilize

- North East and Center - for these regions, the absolute population is at similar levels

- Islands and South - these two zones have attracted less foreign-born population than the other zones. Although the absolute numbers are different, both zones experience similar growth trends

2.1.3 Zones by Origin Distributions

In this section, we study the different origin territories at two aggregation levels:

- continent

- country

A plot for each Italian zone is here shown, we perform an aggregation over the origin continent.

It is interesting that, even if the continent distributions are similar across the Italian zones, there are some differences.

Europe is always the continent that contributes the most to the Italy foreigners flow and the growth rate decreases from 2014 (as aforementioned). The five trends are quite similar even if with different absolute values, especially in South and Islands. All the zones report an increase in the 2007-2008 period, mainly remarkable in South and Islands.

The others continent flow is less important, in terms of absolute value, but the growth rate is not the same for all the zones: it decreases in Center, North East and North West and increases in Islands and South (which are in the "third cluster").

This flow rate consideration holds also for Europe born foreigners.

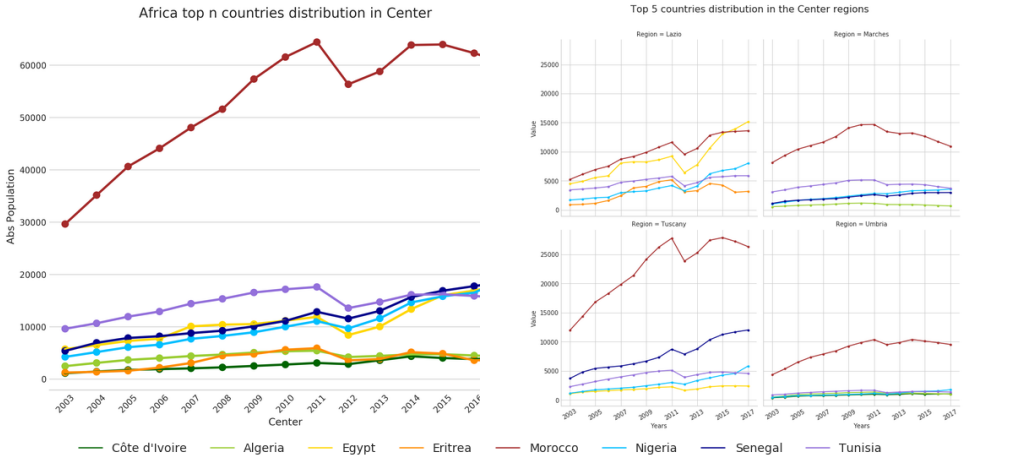
The flow from Pacific is very similar in all the zones and could be overlooked. American born population is significant only in North West and Center.

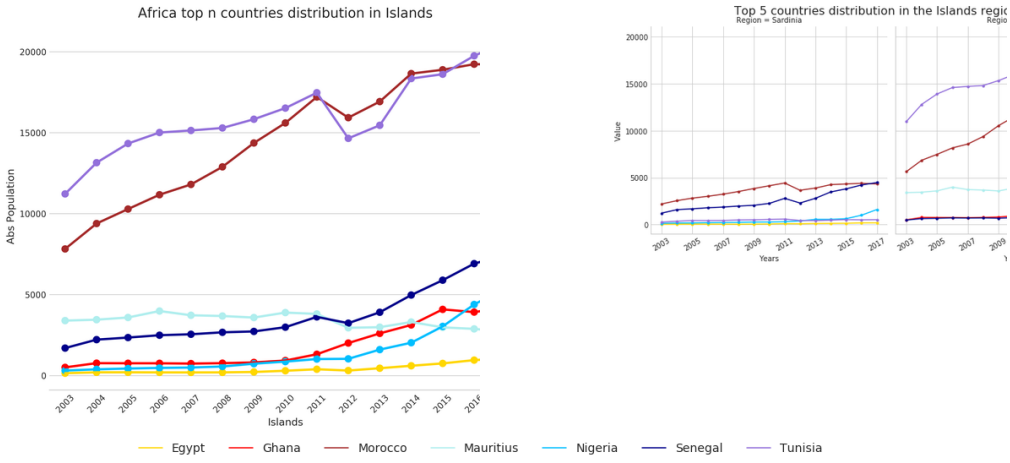
The goal now is to show the distribution of the origin countries that contribute the most to the Italian foreigners-born population flow.

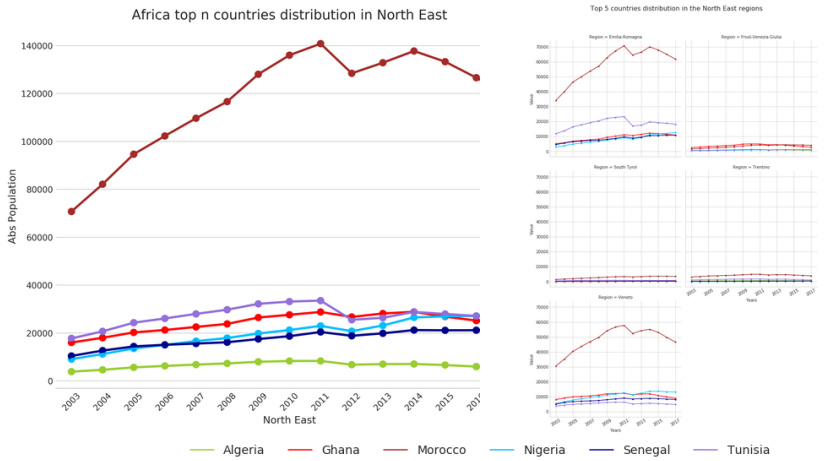
For each couple Continent-Zone and for all the regions in the particular zone, the distributions of the top 5 origin countries of the specific continent are shown.

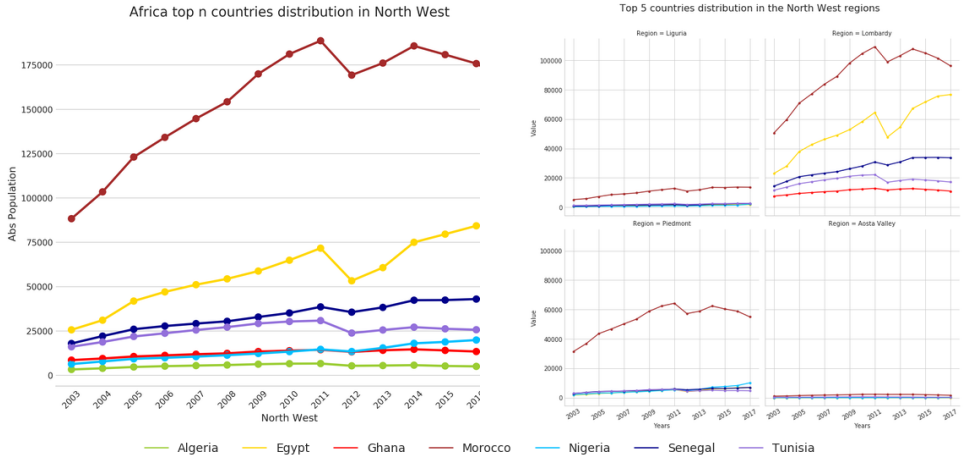
An aggregating plot is also included. It represents the distribution of the top n countries in the zone. The top n countries covered all the top 5 countries found in the zone regions.

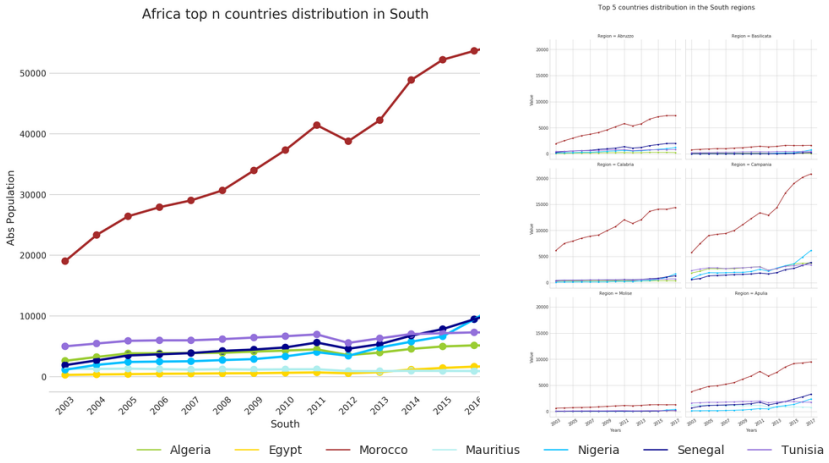
Africa











In all of the five zones, Morocco is the country that contributes the most to the total African-born population flow.

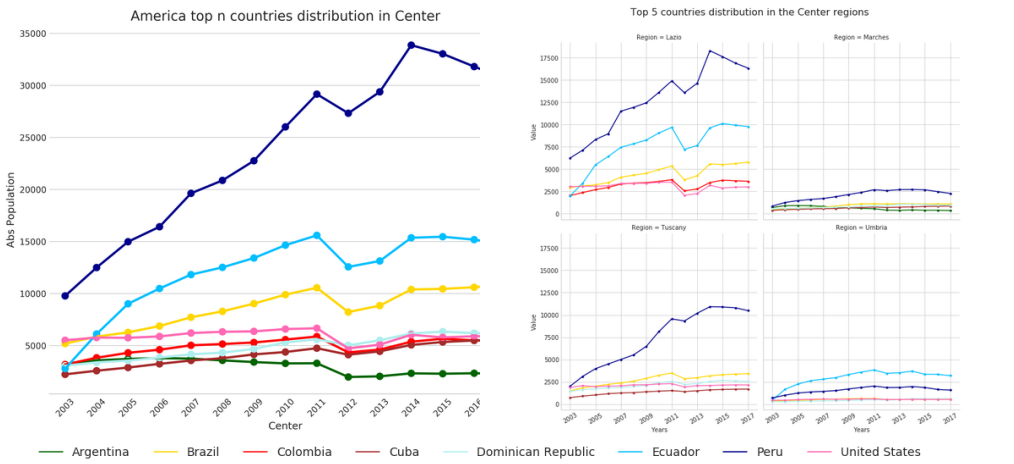
Until 2013, the Morocco's trend is quite similar, with different absolute values, in the Italian zones. From 2014 on, the destination territories of the Moroccan

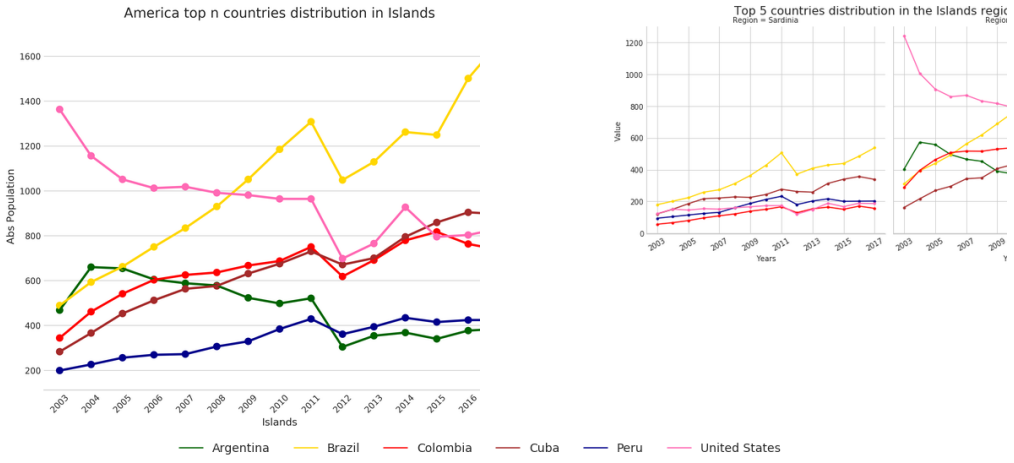
population seem to be changed. In particular, the growth rate of the zones with a higher absolute value (Center, North East, North West) is decreasing, while it is increasing in the other zones Islands and South. Here are the numerical values:

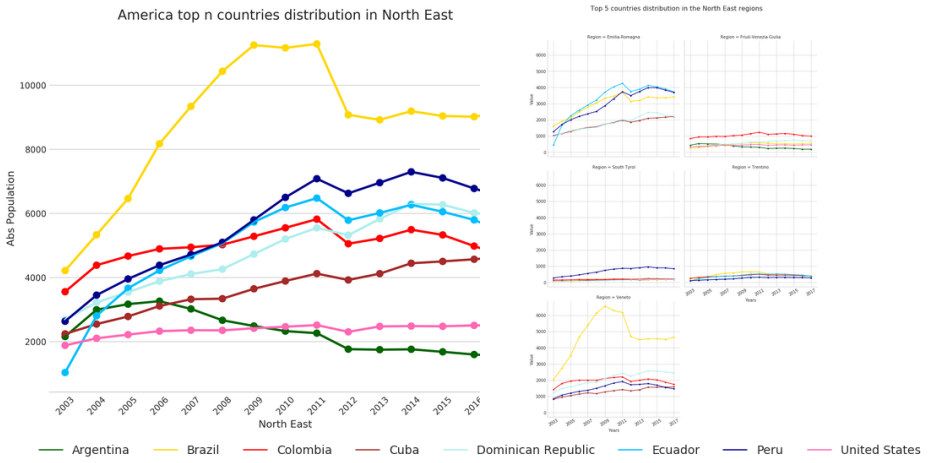
Zone	$flow_{2014}$	$flow_{2017}$	$flow_{2017} - flow_{2014}$	$\frac{flow_{2017}-flow_{2014}}{flow_{2014}}$
Center	63823	60365	-3458	-0.0542
Islands	18648	19202	554	0.0297
North East	137724	119727	-17997	-0.1307
North West	185718	166432	-19286	-0.1038
South	48860	54925	6065	0.1241

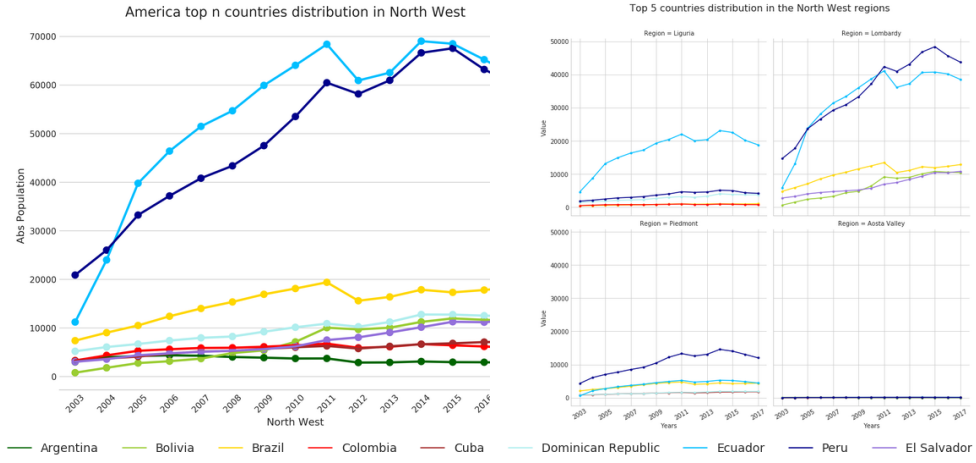
Others notable flows are the once from Tunisia to Sicily and from Egypt to Lombardy.

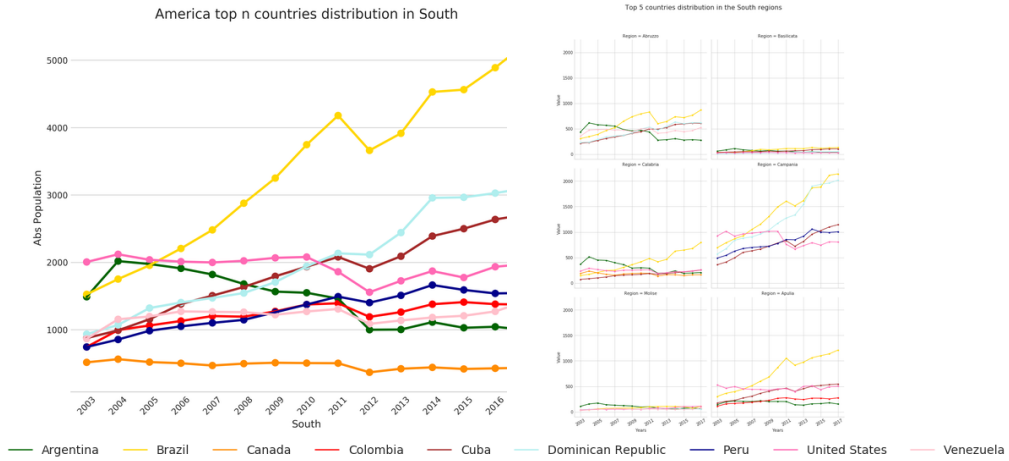
America











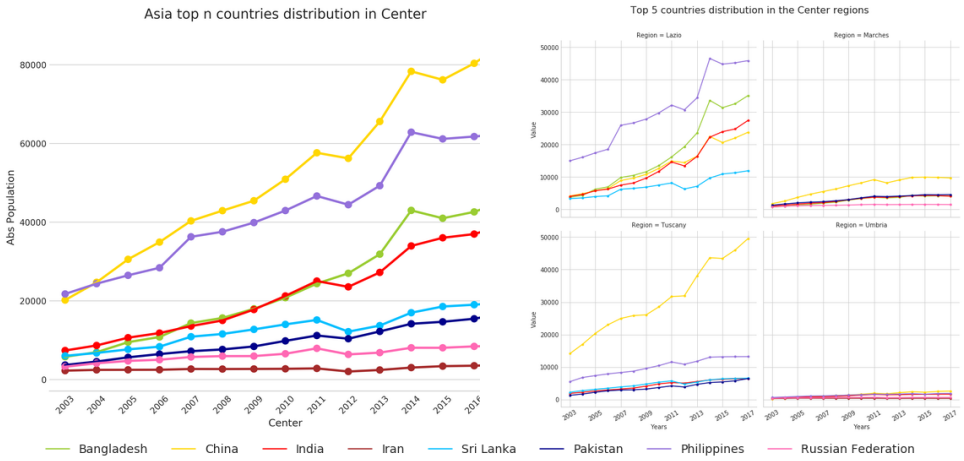
The American flow is pretty different in the observed zones.

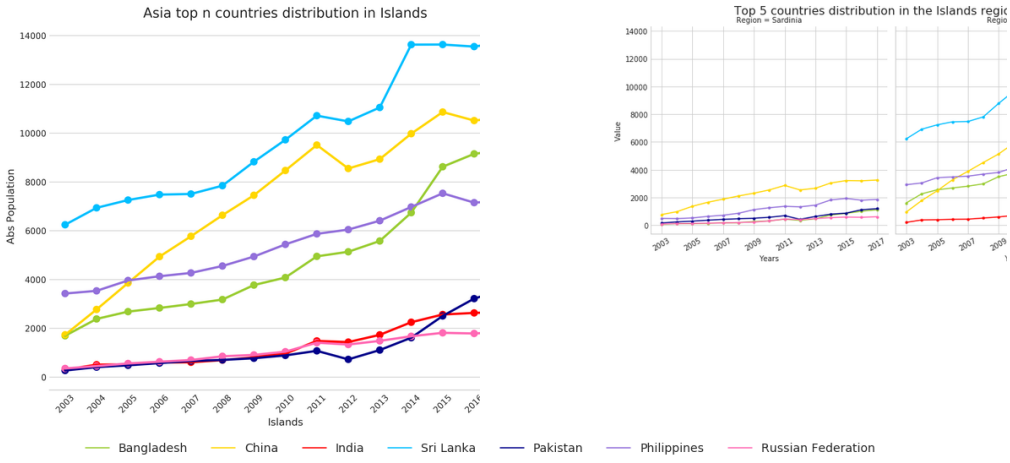
In Center and North West the countries giving the most to the aggregate flow are Peru and Ecuador. More in detail, the most common destination regions are Lazio and Lombardy for both Peruvians and Ecuadorians, then Tuscany and Liguria

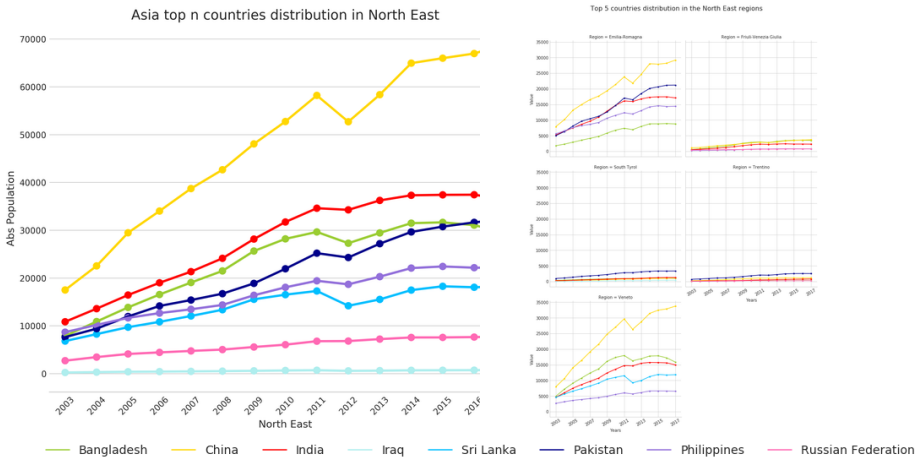
for Peruvians and Ecuadorians respectively. Once again, looking at the plots, it is possible to say that the absolute values are significantly different: the once from Peru and Ecuador are about double and quadruple respectively in the North West with respect to the Center. Both the growth rates start decreasing having negative values from 2014.

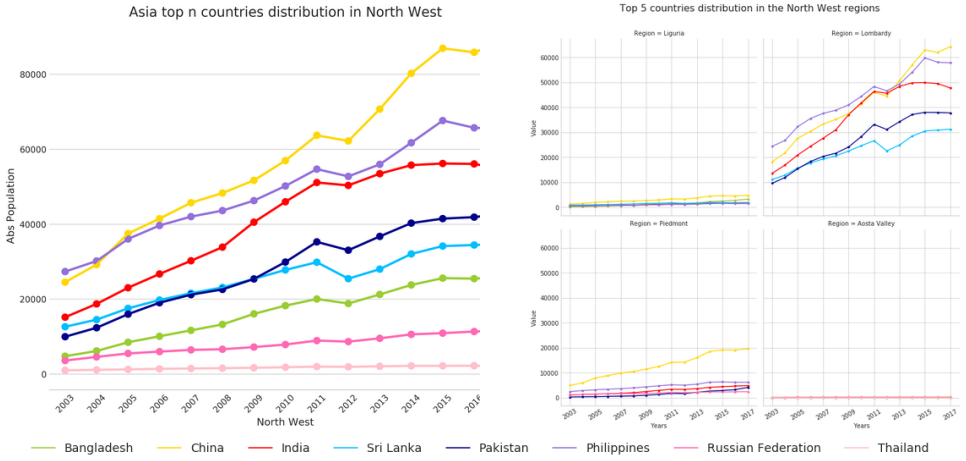
In North East, Islands, and South it is possible to notice that Brazil is the most common origin country. In North East the growth rate seems to be stable from 2014 on, while in Islands and South it is still increasing.

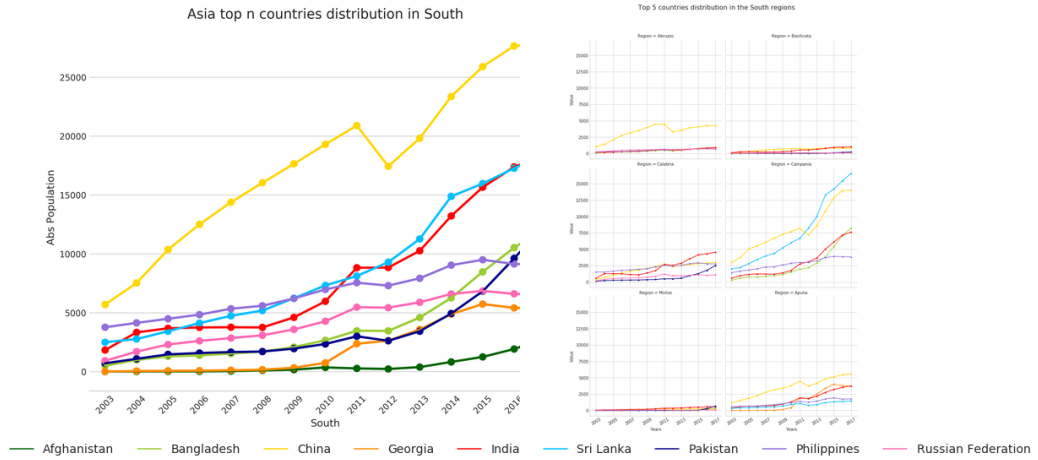
Asia









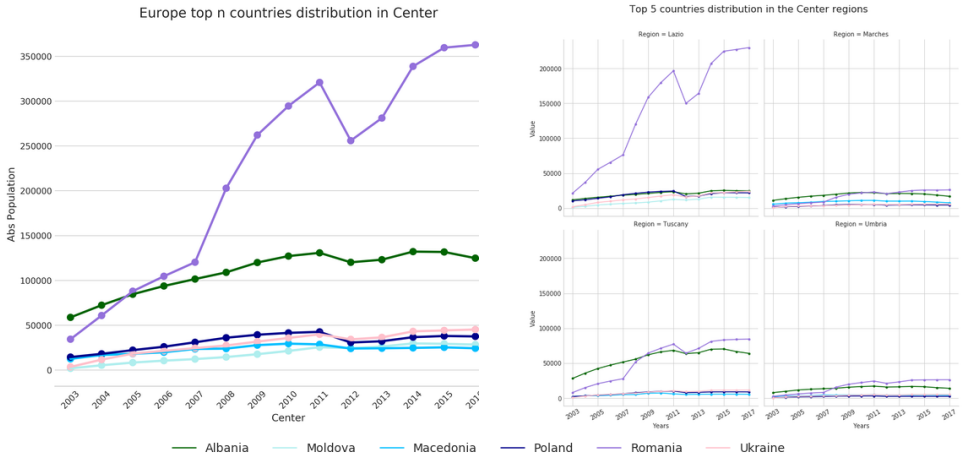


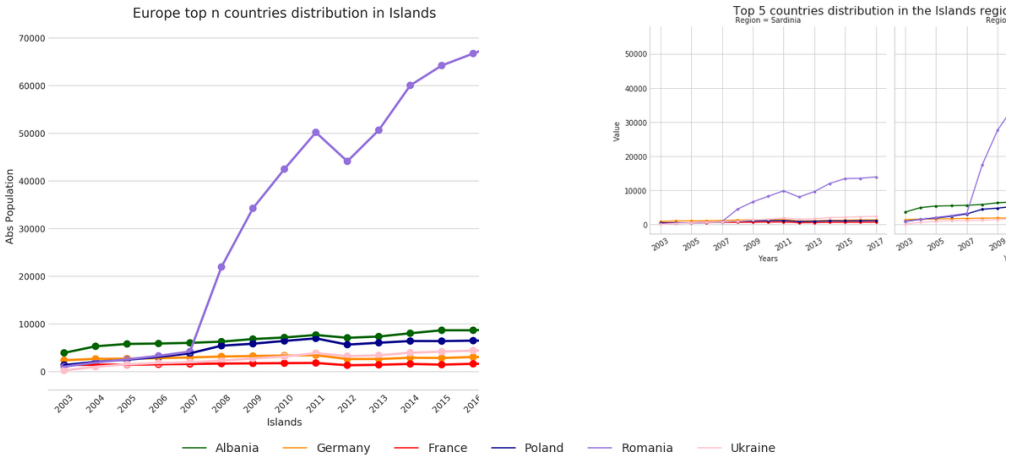
Concerning the Asian-born population flow distributions, the plots show a similar trend at the zone level. That is: all the countries have a comparable behavior even if with significantly different scale (in absolute values).

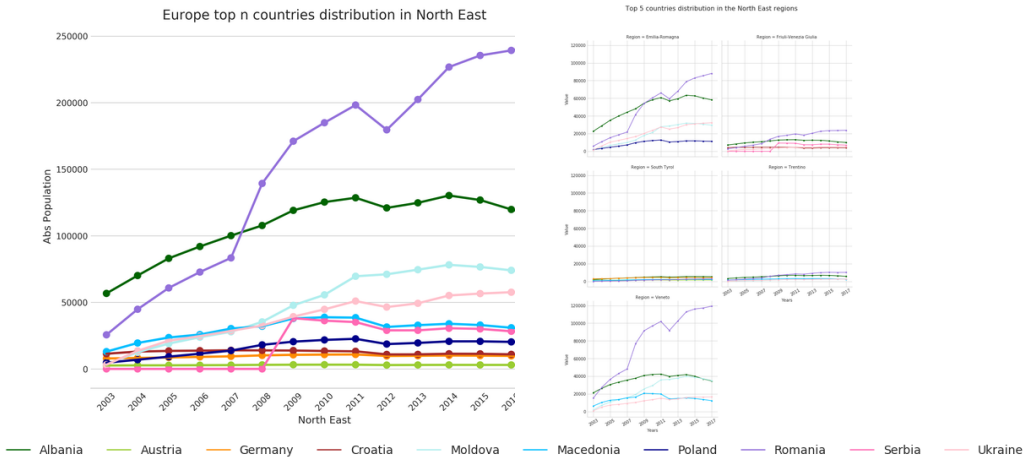
Here is evident, more than in the others continent flows, that just one or two (in Center and North East) regions contributes to the aggregate distribution.

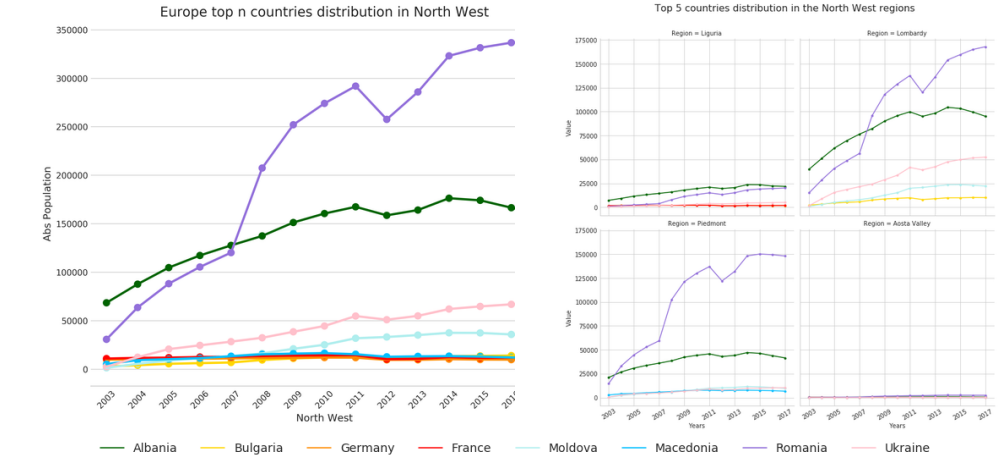
The Chinese-born flow is always the highest but in Lazio and Sicily, where the highest values correspond to the Philippines and Sri Lanka.

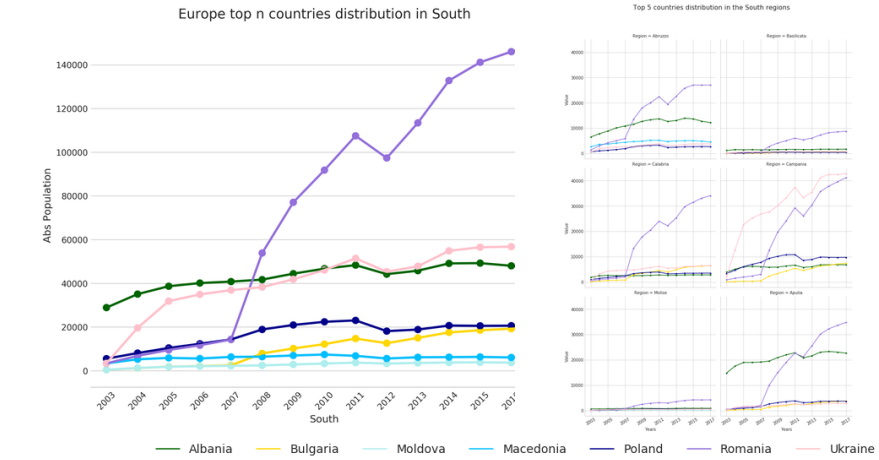
Europe









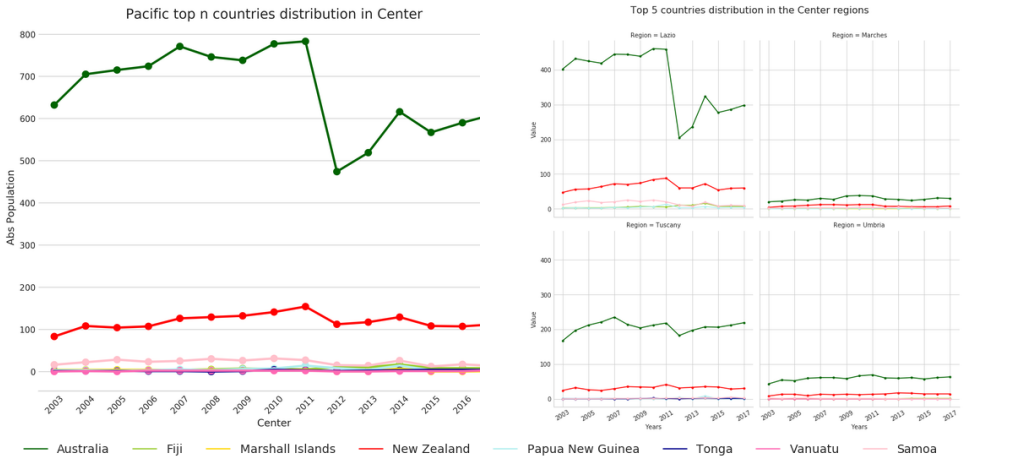


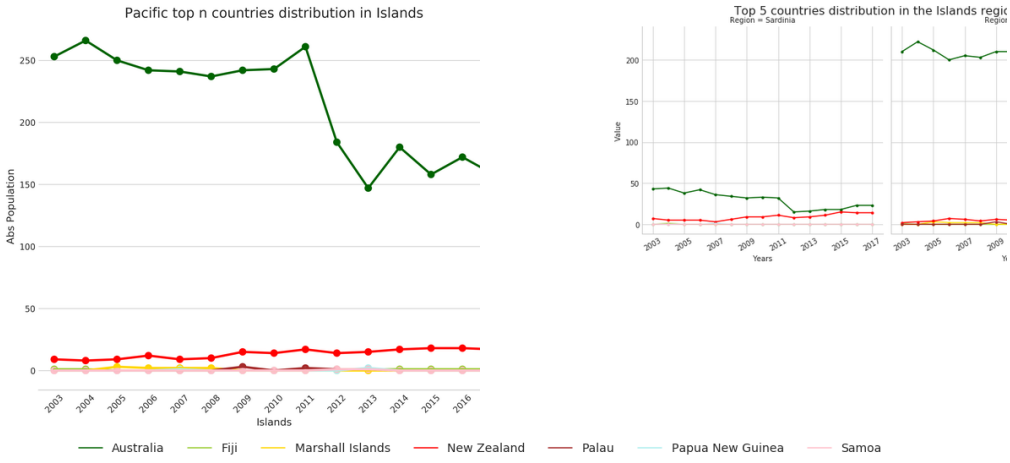
Romania is the country that contributes the most to the total European-born population flow. Even if the growth rate seems to decrease, it still has an increasing trend.

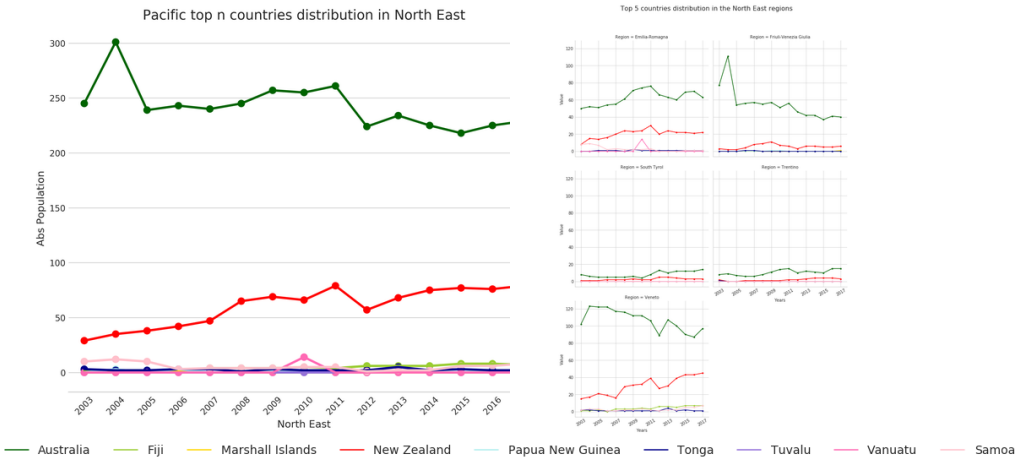
In regions like Tuscany, Lombardy, Apulia, and Abruzzo the Albanian flow is remarkable even if from 2014 it looks like to have a negative growth rate.

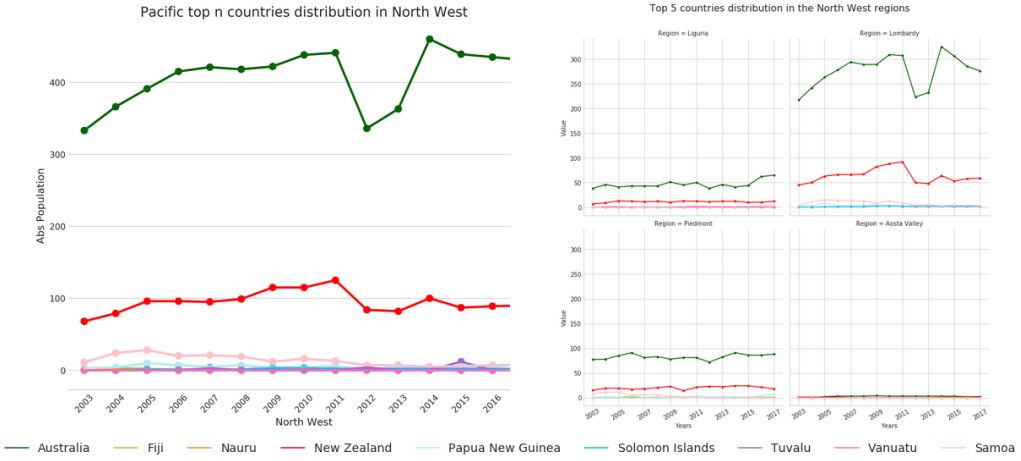
In Campania, there may be an important Ukrainian community.

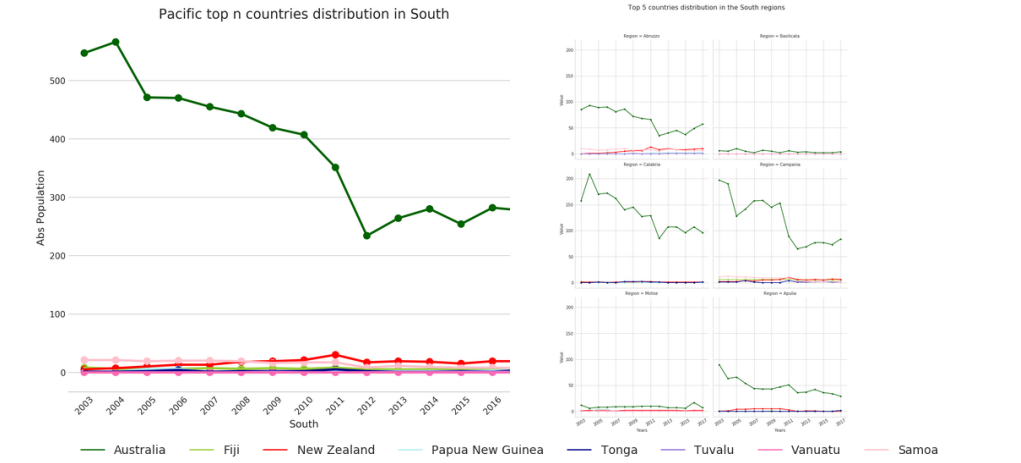
Pacific











It is possible to see that here the scales are very different from the once in the above plots. As aforementioned, the Pacific flow could be overlooked.

At the end of this section, it is also possible to come up with some comments about the zone-regions relation.

Lazio and partially Tuscany represent most of the Center total distribution.

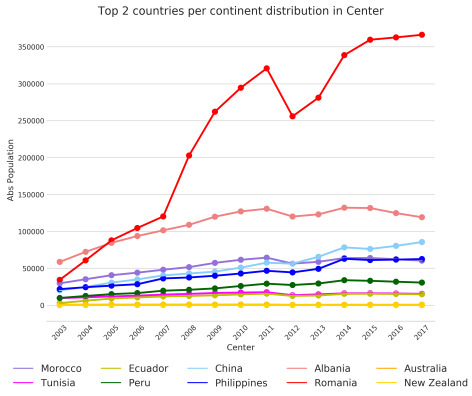
Sardinia's data could be overlooked with respect to the Sicily once.

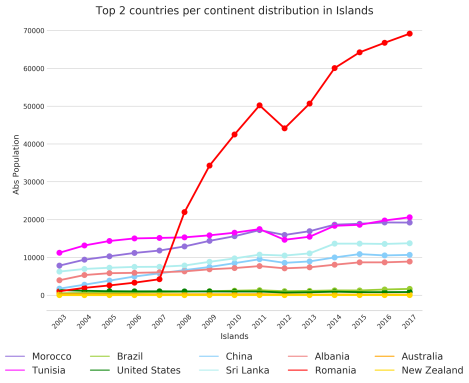
Lombardy constitutes almost the entire North West flow. In specific case also Piedmont and Liguria contribute to the aggregate values.

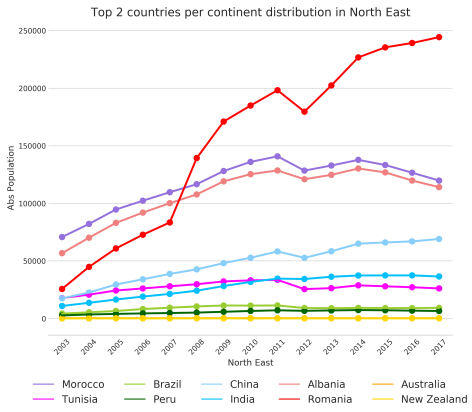
Emilia-Romagna and Veneto seem to split up all the North East flow.

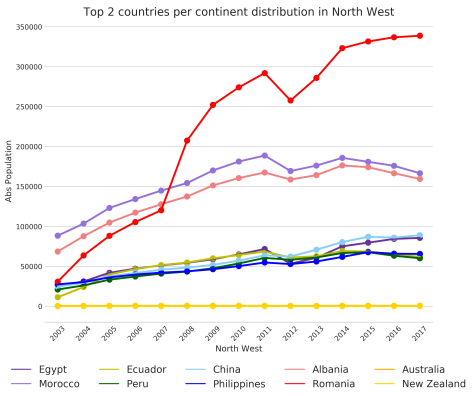
In South, it is not possible to point out a significant difference between the regional trends. However, it is shown that Molise and Basilicata may be ignored in the analysis.

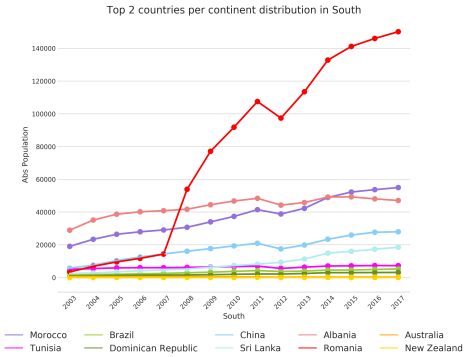
To conclude this analysis, in the following plots, we shown for each of the five Italian zones the top two countries per continent. Top two countries per continent represent the two countries, for each continent, contributing the most to the zone flow.











These plots seem to confirm what come out from the previous section.

Moreover, it is highlighted the difference between the Romanian flow and all the other flows.

To conclude that it would be possible to say that these two last sections confirm what assumed before about the three different behaviors of the five zones.

2.1.4 Additional Features

On the ISTAT website, it is possible to find some interesting features that could be included in the formulation of a model aim to explain the immigration flow to Italy.

To begin, we perform an initial data filtering. We consider only the economic, data that are supposed to have a relation with the variable of interest.

Some features can not be used, due to data availability:

- *Basic health care* only 2004-2013
- *Expenditure for interventions and social services* only 2013-2014
- *Expenditure for the house of families with foreign components* only 2009
- *Aspect of daily life - Interpersonal Trust* only 2010
- *Hospitalizations* missing from 2003 to 2012 plus 2016 and 2017
- *Aspects of daily life - general life degree of satisfaction* missing from 2003 to 2009

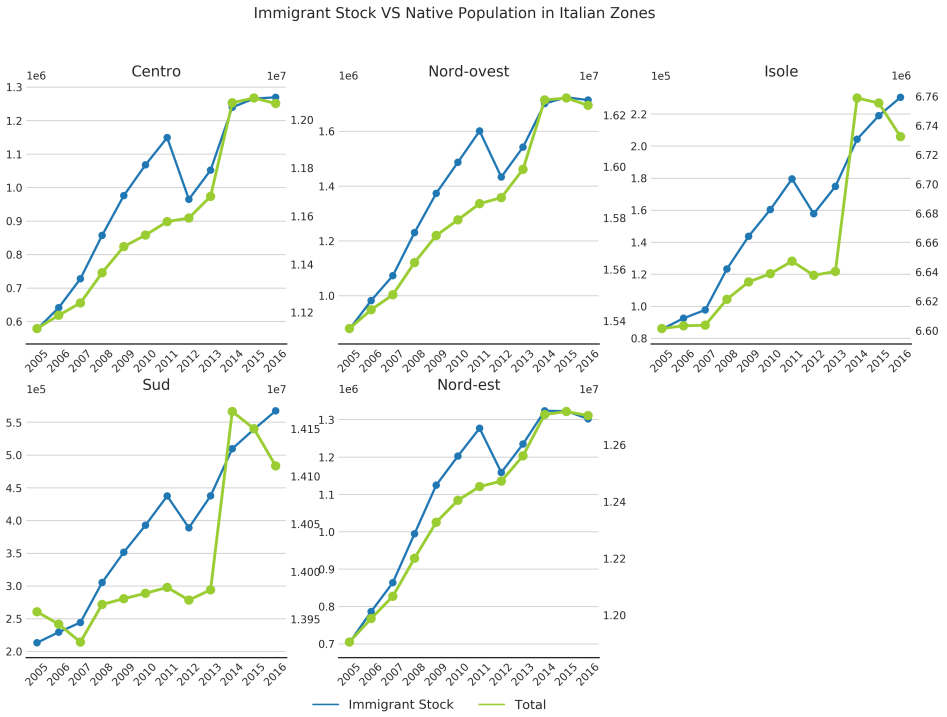
Some others, due to statistical problems:

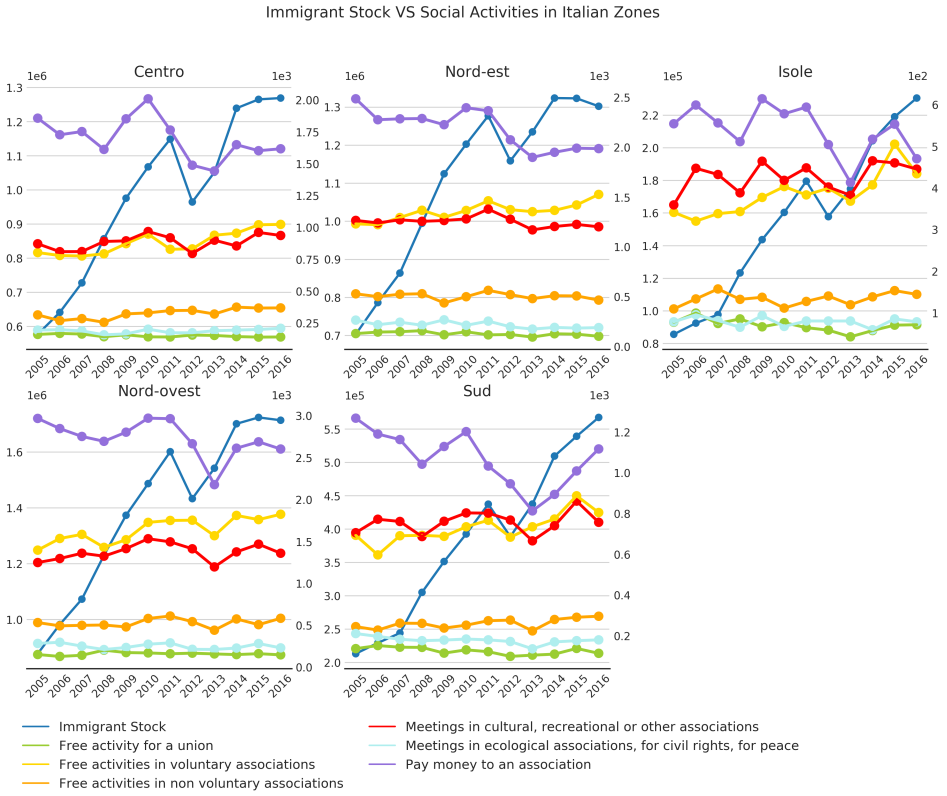
- *Economic situation opinions ("Famiglie per capacità di arrivare a fine mese"* [1]*)*: around 10.4% of data are not statistically significant and 4.6% do not reach the half of the minimum (ISTAT definition: "Il dato si definisce poco significativo nel caso in cui corrisponda ad una numerosità campionaria compresa tra 20 e 49 unità" [2].)

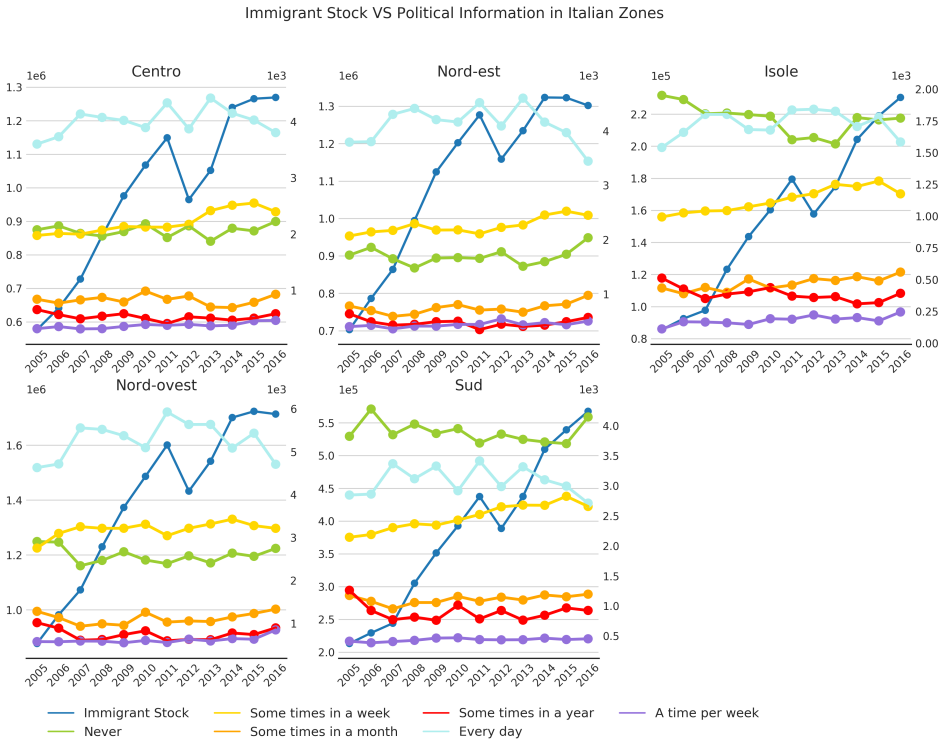
[1] Families by ability to get to the end of the month.

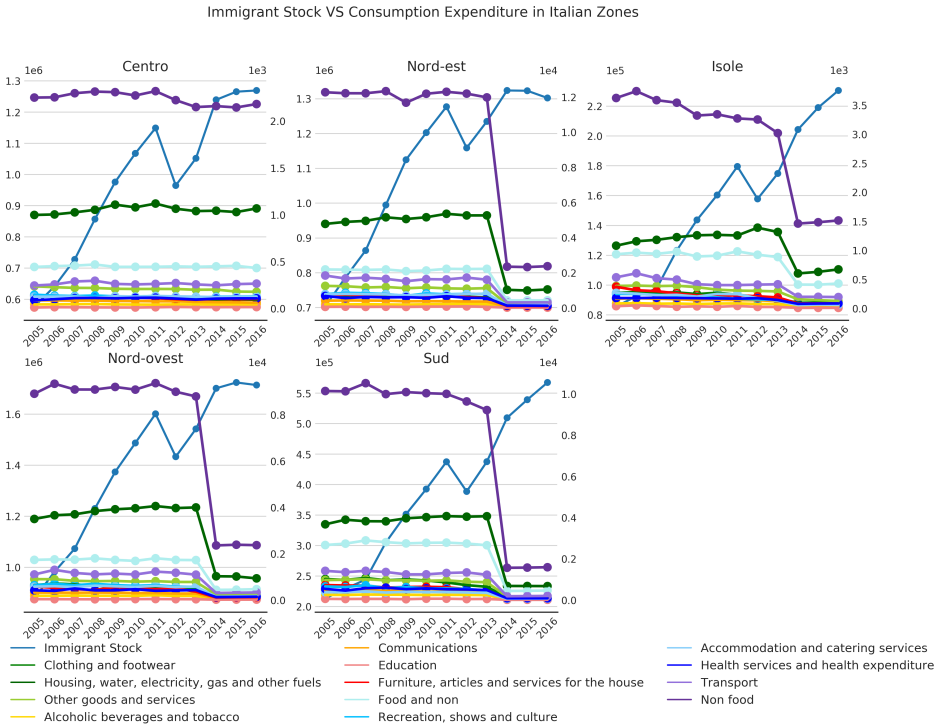
[2] Data is not very significant if it corresponds to a sample size between 20 and 49 units.

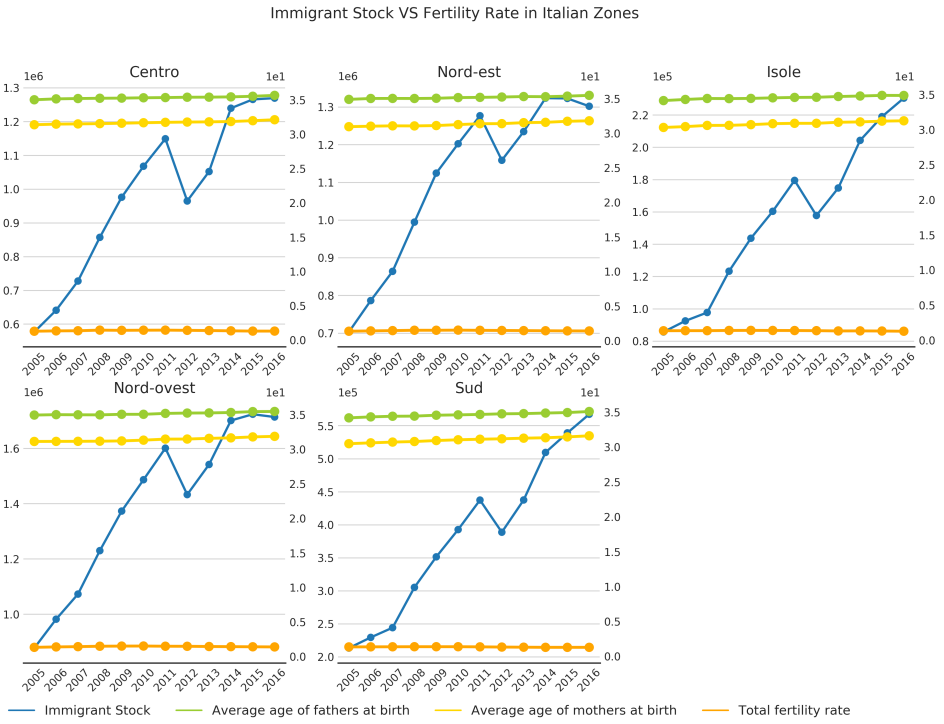
In order to see the relationship between the Immigrant flow and other additional features, let's plot them. The zone level is sufficient to understand the behaviour.

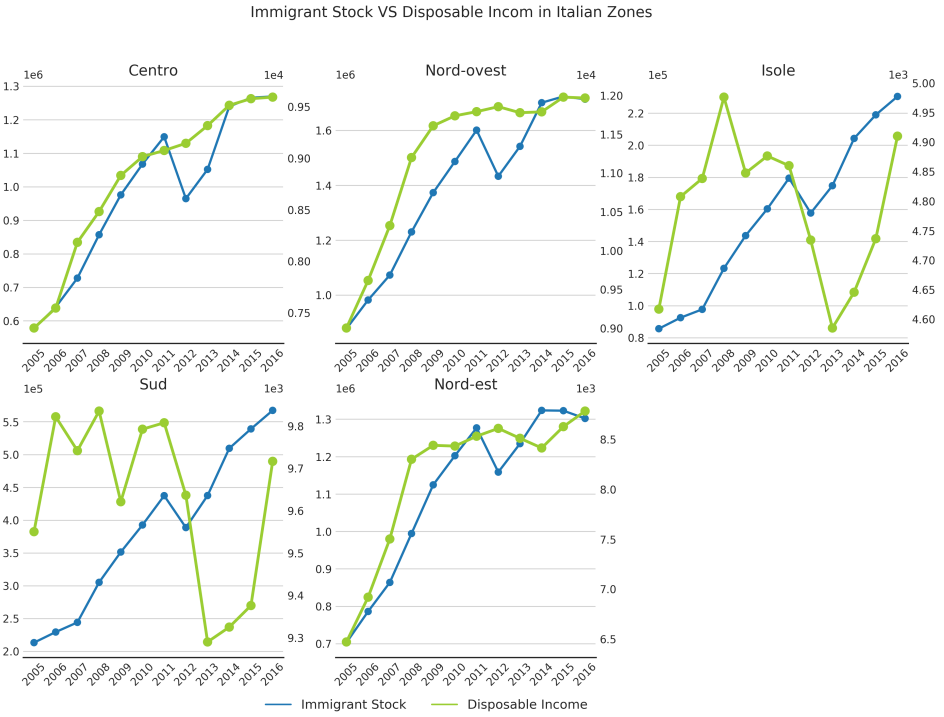


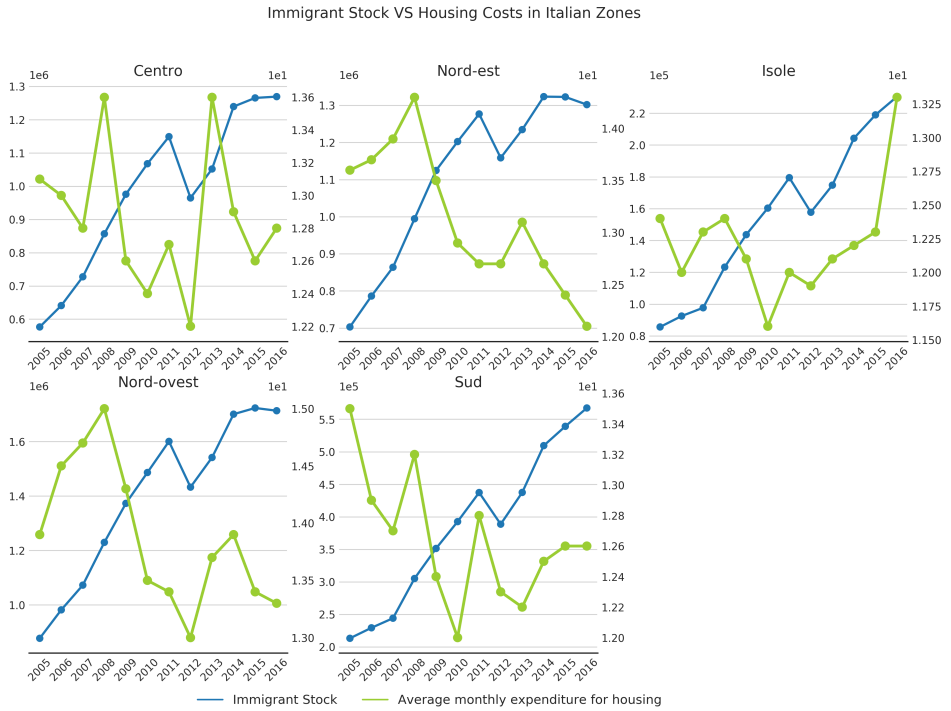


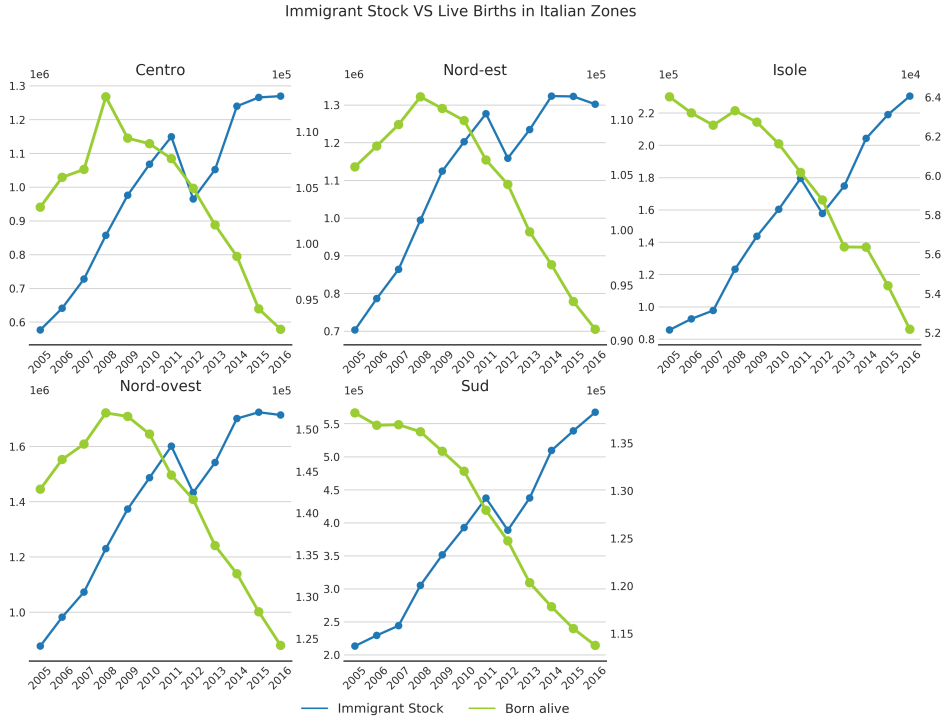

Immigrant Stock VS Political Information in Italian Zones
Centro
Nord-est
Isole
Nord-ovest
Sud
Immigrant Stock
Never
Some times in a week
Some times in a month
Some times in a year
Every day
A time per week

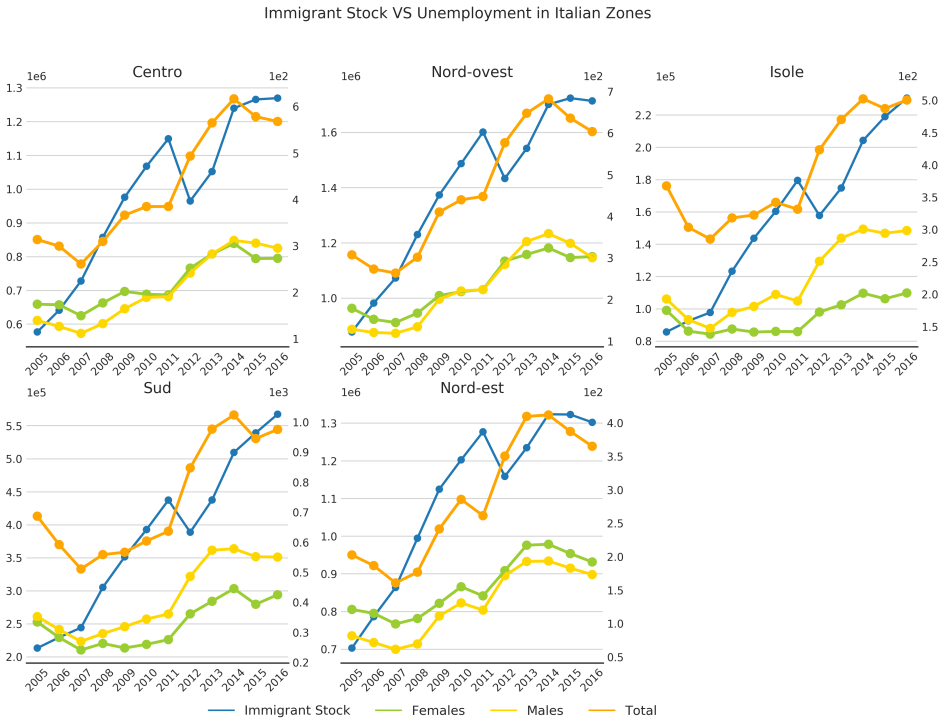

Immigrant Stock VS Consumption Expenditure in Italian Zones
Centro
Nord-est
Isole
Nord-ovest
Sud
Immigrant Stock
Clothing and footwear
Housing, water, electricity, gas and other fuels
Other goods and services
Alcoholic beverages and tobacco
Communications
Education
Furniture, articles and services for the house
Food and non
Recreation, shows and culture
Accommodation and catering services
Health services and health expenditure
Transport
Non food

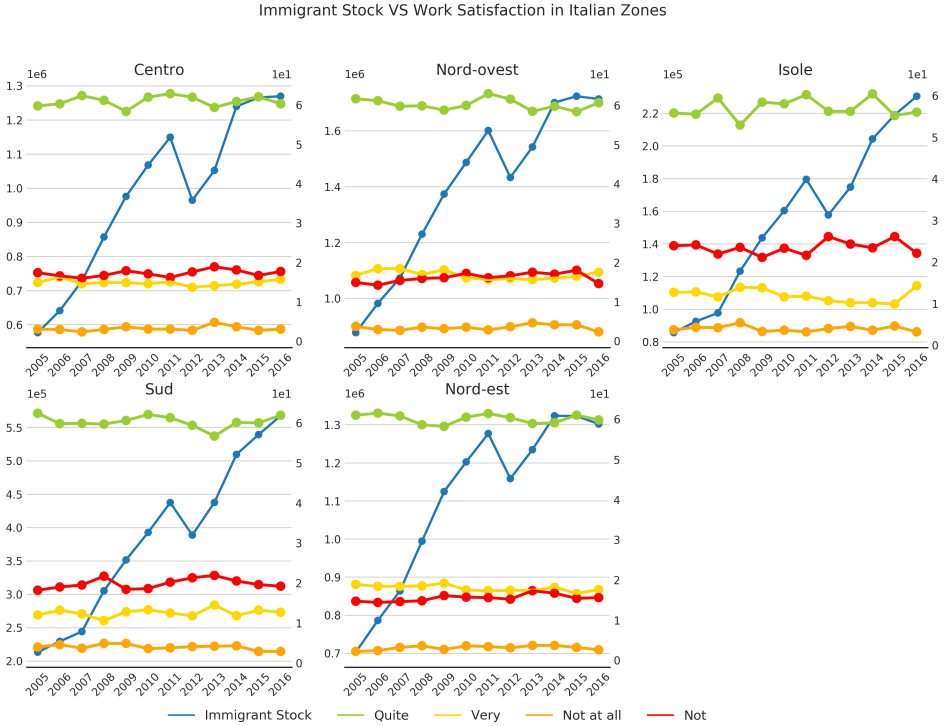

Immigrant Stock VS Fertility Rate in Italian Zones
Centro
Nord-est
Isole
Nord-ovest
Sud
Immigrant Stock
Average age of fathers at birth
Average age of mothers at birth
Total fertility rate

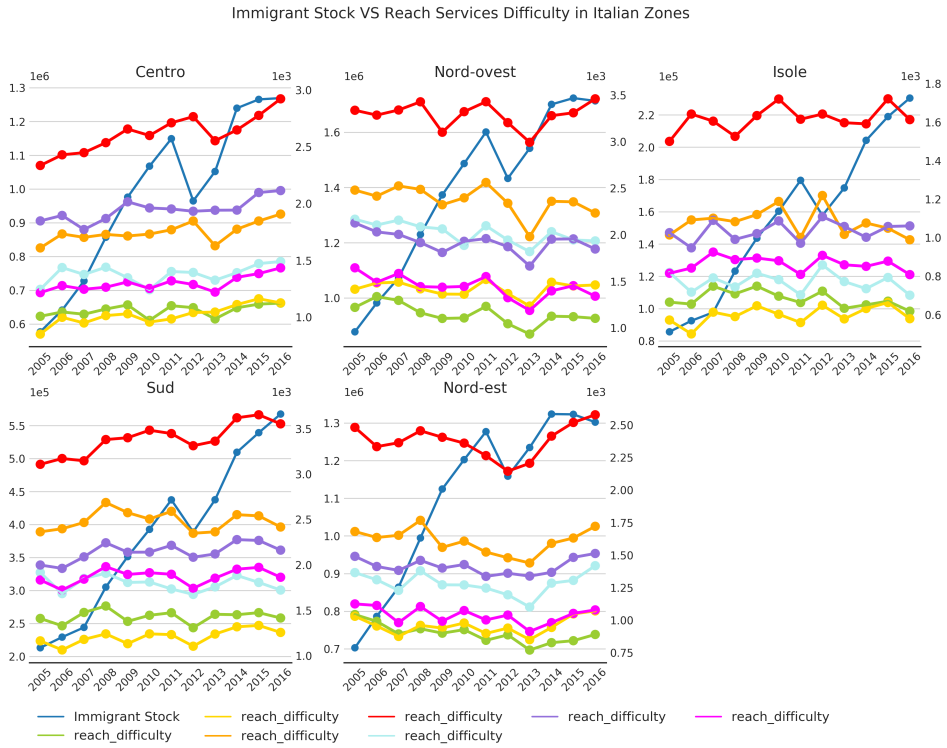

Immigrant Stock VS Disposable Incom in Italian Zones
Centro
Nord-ovest
Isole
Sud
Nord-est
Immigrant Stock
Disposable Income

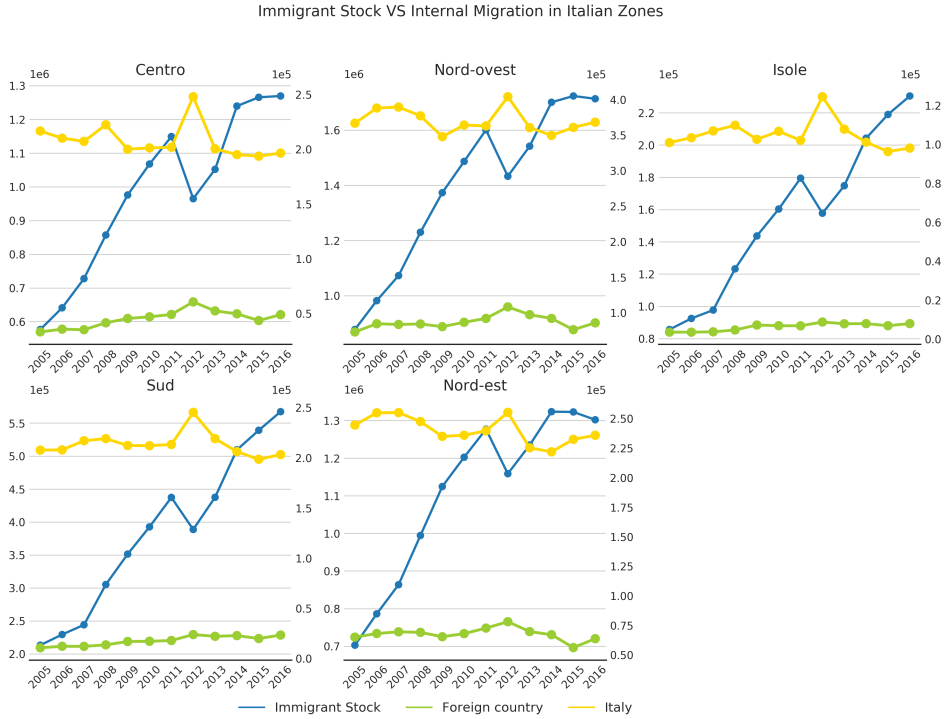

Immigrant Stock VS Housing Costs in Italian Zones
Centro
Nord-est
Isole
Nord-ovest
Sud
Immigrant Stock
Average monthly expenditure for housing


Immigrant Stock VS Live Births in Italian Zones
Centro
Nord-est
Isole
Nord-ovest
Sud
Immigrant Stock
Born alive

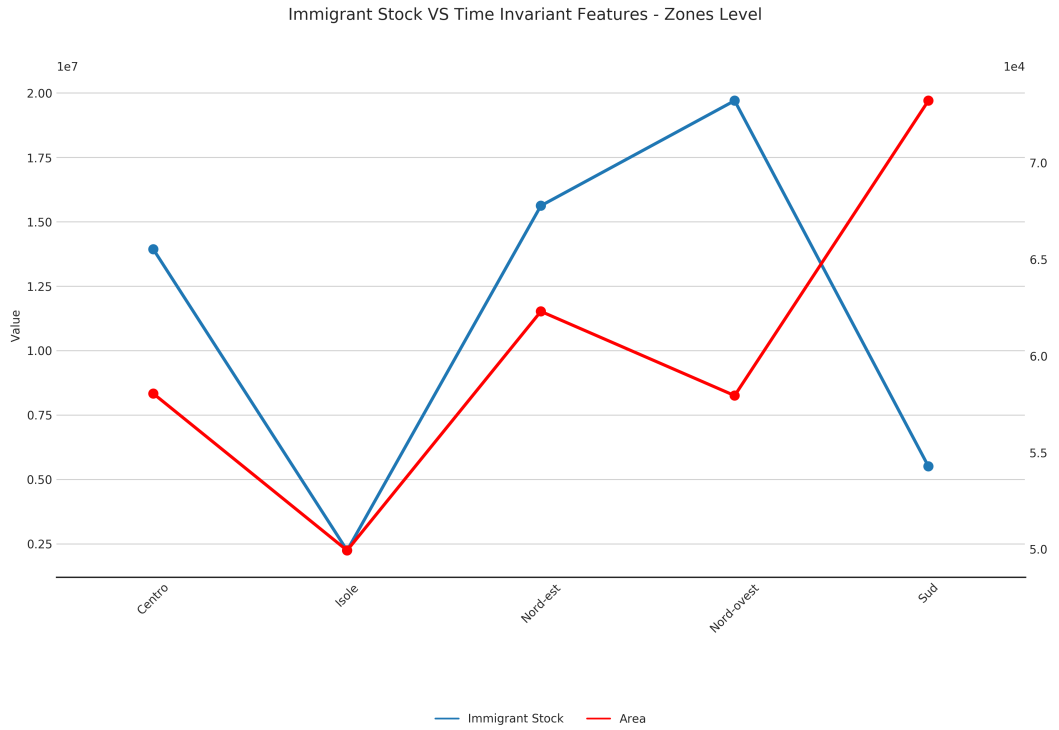

Immigrant Stock VS Unemployment in Italian Zones
Centro
Nord-ovest
Isole
Sud
Nord-est
Immigrant Stock
Females
Males
Total


Immigrant Stock VS Work Satisfaction in Italian Zones
Centro
Nord-ovest
Isole
Sud
Nord-est
Immigrant Stock
Quite
Very
Not at all
Not





Let's see also the "Area" variable which is time-invariant. Here as "Immigrant Stock" the mean over the years 2005-2015 is considered.



Some variables like:

- Native population
- Difficulty to reach services (e.g. pharmacy)
- Housing costs
- Net Income
- Social activities
- Unemployment

seem to effec one or more location-specific immigrant flow.

2.2 Spain

Detailed data about the immigrants stock in Spain can be found on the INE (Istituto Nacional de Estadística) website [14]. Information about the foreigners' population by origin country, Spanish province, gender and year are available. More in detail, data concerning 136 origin territories, 53 destination provinces and 20 years (from 1998 to 2017) are collected.

2.2.1 Data Quality

Similar to ISTAT website, also on the INE one it is possible to access an excellent and useful database. It is one of the few to provide such disaggregated data, but it

also has some problems. First of all, it was very hard to meet the dataset: the google query performed in English didn't show any interesting results, on the contrary it was possible to access the entire INE database by typing the same query in Spanish. Moreover, we met the needed dataset from the Spanish website version, not from the English one. Another important problem is about the data architecture: data are not provided in tables but in a text structure, so a challenging preprocessing was needed.

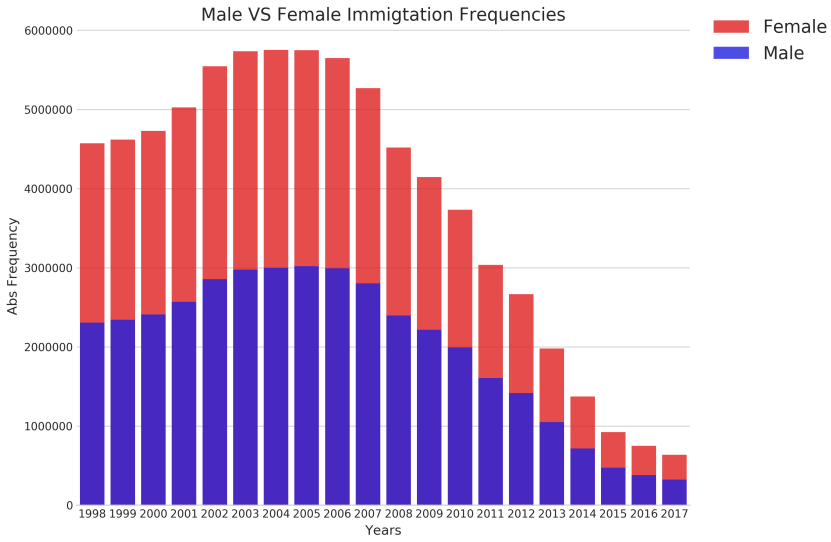
Based on the INE territories classification, we will consider the following autonomous communities for the project:

- Andalucía
- Aragón
- Canarias
- Cantabria
- Castilla y León
- Castilla-La Mancha
- Cataluña
- Comunidad Foral de Navarra
- Comunidad Valenciana
- Comunidad de Madrid
- Cueta
- Extremadura
- Galicia
- Illes Balears
- La Rioja
- Melilla
- País Vasco
- Principado de Asturias
- Región de Murcia

Follows a statistical analysis of the dataset.

2.2.2 Dataset Distributions

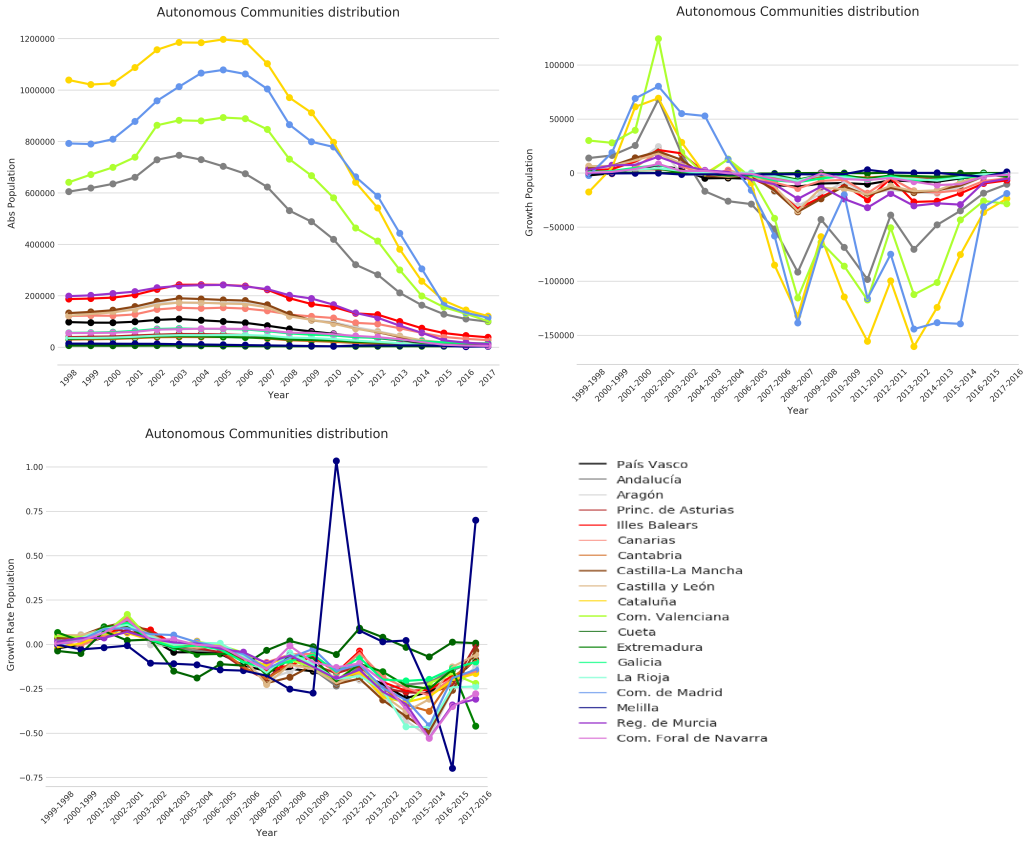
Again, the gender distribution over the years, aggregated both on the destination and the origin territories, is shown.



As for the Italy case, also for Spain there is no evident difference between the male and female immigrants distribution.

Let's now study the foreigners' flow over the years and across the 19 considered destination territories.

As for the Italian case, the absolute value distribution, the growth distribution and the growth rate distribution are studied.



It is possible to see that the immigrants flow into Spain could be divided in different periods:

- 1998-2005: there is a huge increase in foreigners
- 2006-2009: foreigners continue to arrive but in a smaller amount. The decreasing growth may due to the global financial crisis and the restrictive policies of the European Union
- 2010-2013: immigrants growth is significantly decreasing. A probable case is the economic crisis and the labor difficulties in Spain
- 2014-2017: the phenomenon is still decreasing, but the growth rate is notably increasing with respect to the previous 2010-2013 period

Moreover, it is possible to identify 4 "major" destination territories in terms of the number of immigrants:

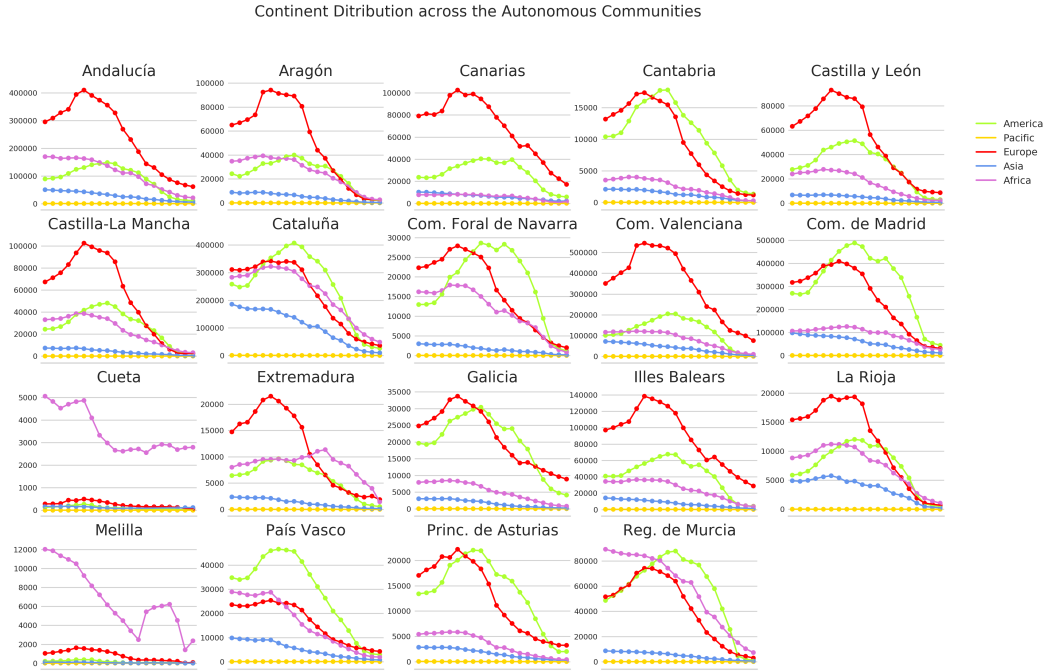
- Cataluña
- Comunidad de Madrid
- Comunidad Valenciana
- Andalucía

It is very interesting to notice that the difference, in terms of the number of immigrants, between the 4 aforementioned territories and the others is keeping decreasing over the time period considered.

2.2.3 Autonomous Communities by Origin Distributions

We will now study the flow by autonomous Communities and origin territories.

As first, let's see the distribution by origin country.

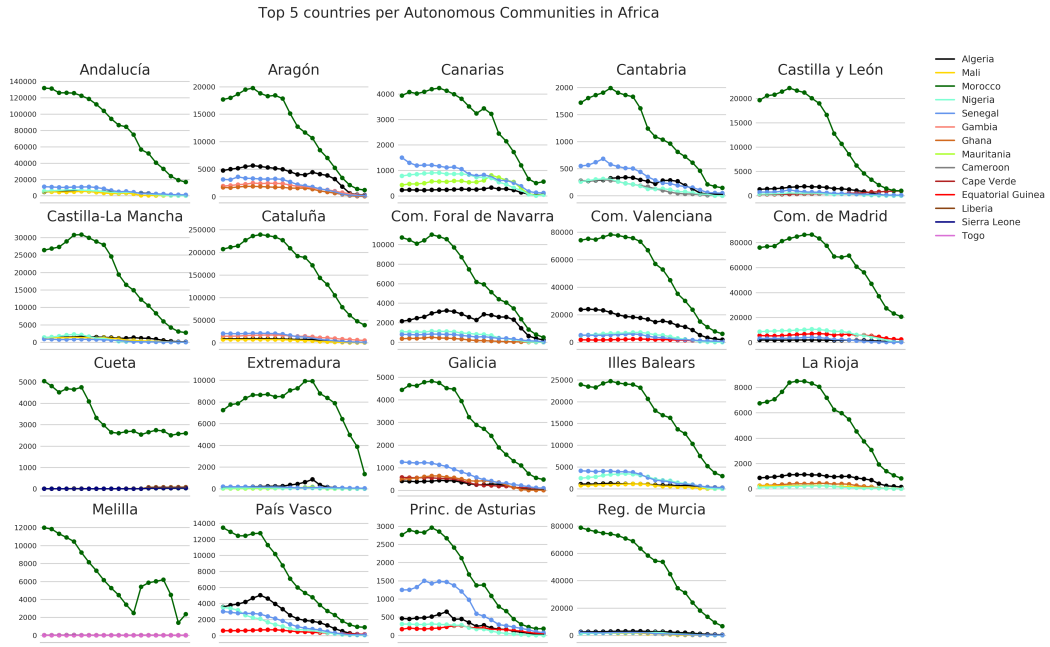


Europe is almost always the continent giving the most to the overall immigrants' flow.

The trends in the 19 autonomous communities are quite similar (but in Cueta and Melilla), and reflect what observed in the Spanish foreigners' flow in the previous subsection. While trends are comparable the absolute values are very different. In most of the cases also the flows coming from America and Africa contribute significantly to the total flow.

Pacific and Asia, but in a few cases, could be overlooked.

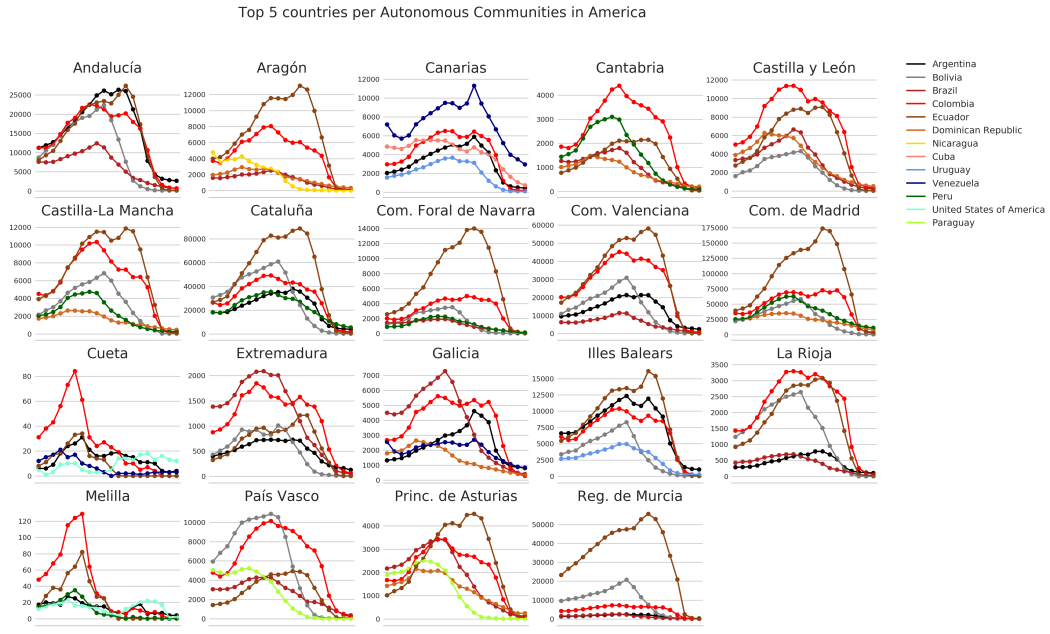
Let's now analyze the distribution by origin country. Because of showing all the 136 origin countries is not feasible nor interesting for the study, just the top 5 countries, in terms of total amount of immigrants, per continent are studied.



Morocco's flow basically explains all the African-born population flow in all the Spanish territories. The trends are quite similar in the autonomous communities, while the absolute values are not.

From 2006 a notable decreasing is observed in most of the destinations. In Extremadura, the decreasing starts in 2010.

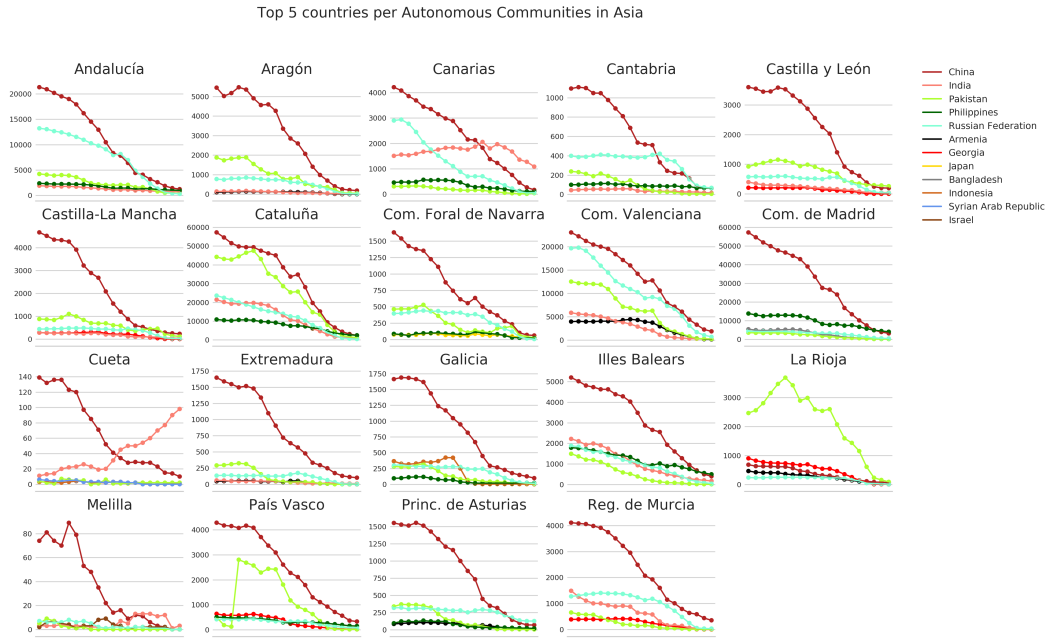
In some cases also the flows from Senegal and Algeria are noteworthy.



The American's countries flow cannot be generalized over the analyzed Spanish destinations.

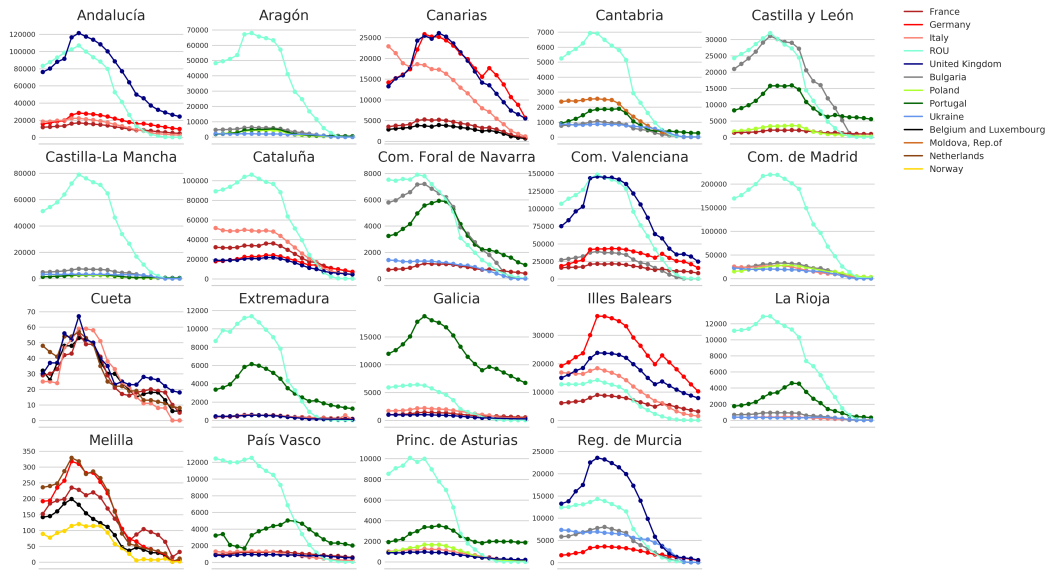
In most of the cases, Colombia or Brazil or both are the "main" flow. For some autonomous communities also Argentina and Bolivia flow should be considered.

As observed in all the previous analysis, also here it is possible to see an important decreasing starting from 2006 or 2010, it depends on the origin-destination pair.



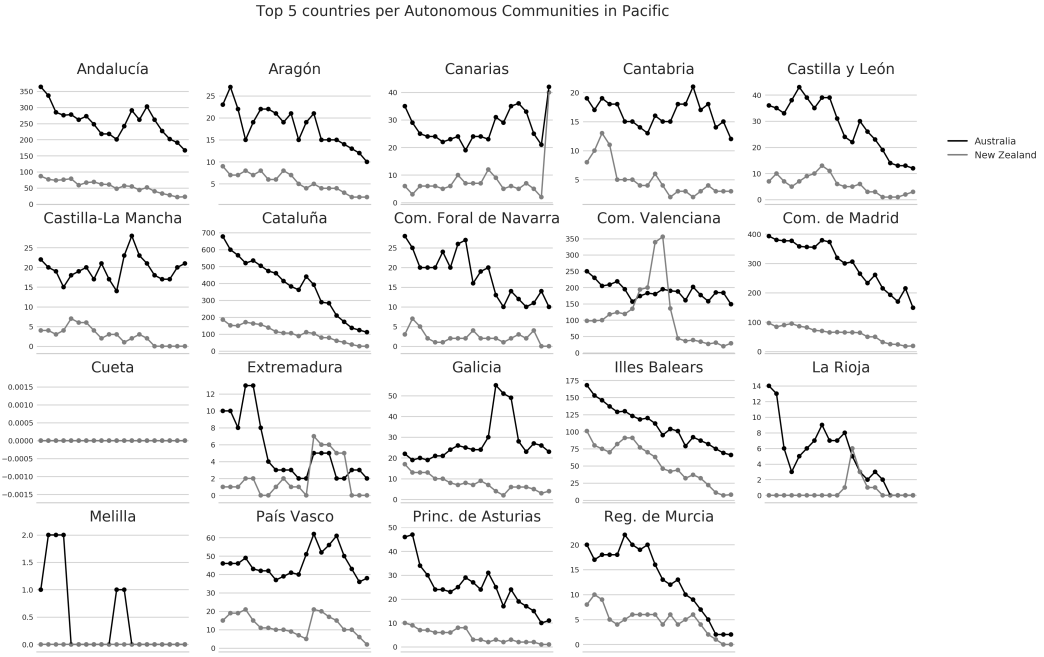
China is the country that contributes the most to the total Asian-born population flow for all the autonomous communities but La Rioja. The China flow has, almost, a linearly decreasing trend.

Other remarkable flows are the ones from India, Pakistan, and Russia.



The "main" flow of European immigrants is the one from Romania for most of the considered destination territories. As before, also this flow starts decreasing significantly from 2006 on.

Depending on the autonomous communities, there may be other denoting flows, but they cannot be generalized. In Andalucía, Canarias, Comunidad Valenciana, Ceuta, Illes Balears and Región de Murcia the United Kingdom flow is relevant. In other territories, also Germany and Republic of Moldova contribute to the total flow.



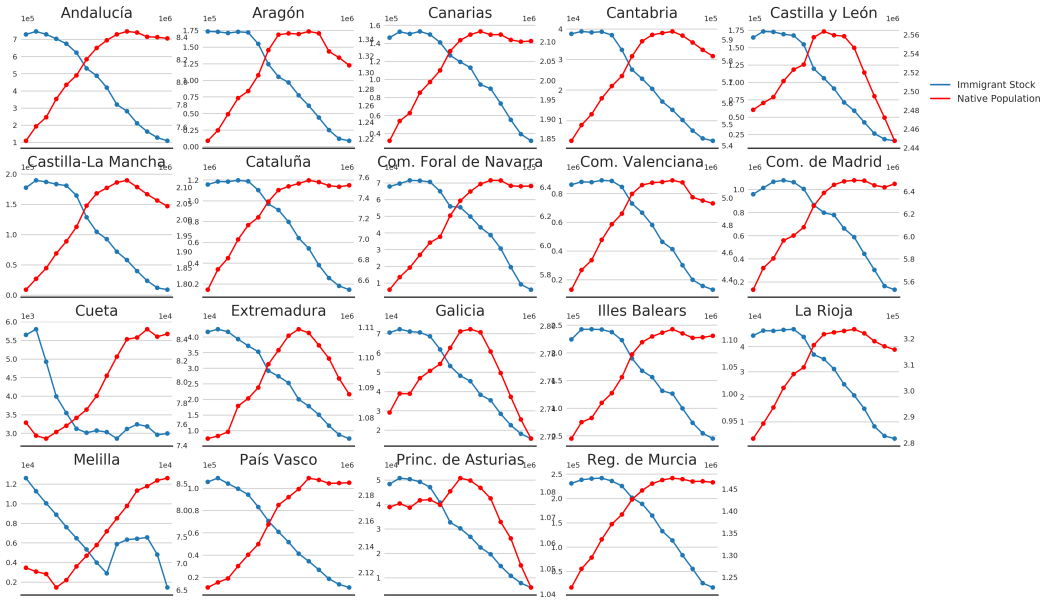
As beforementioned, the flow for the Pacific-born population can be left out.

2.2.4 Additional Features

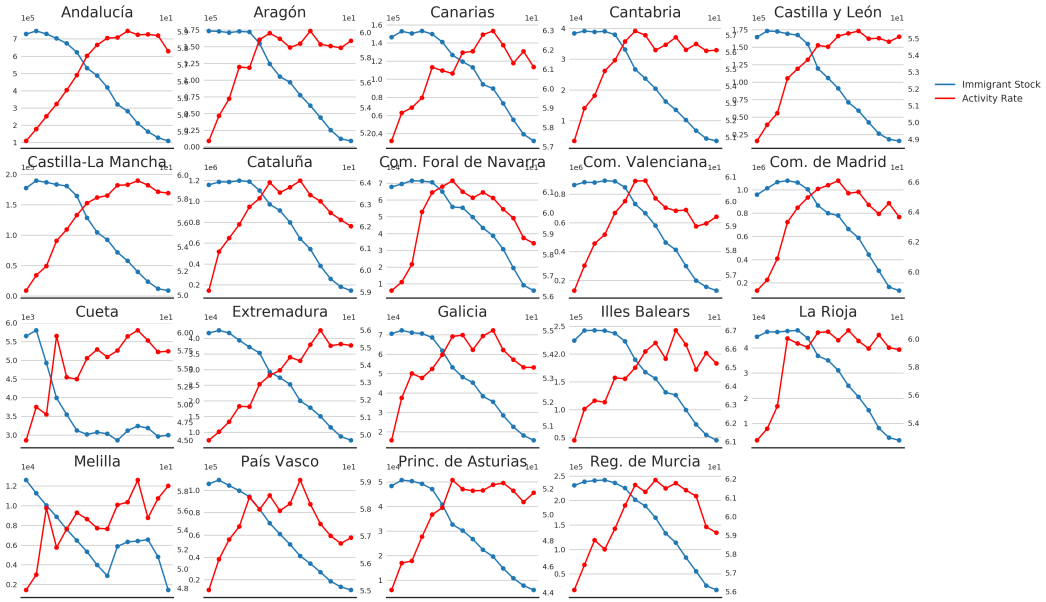
In order to build any model, we need additional features. Some interesting variables can be found on the INE database.

Follows the plots showing each feature time series compared with the immigrant flow series.

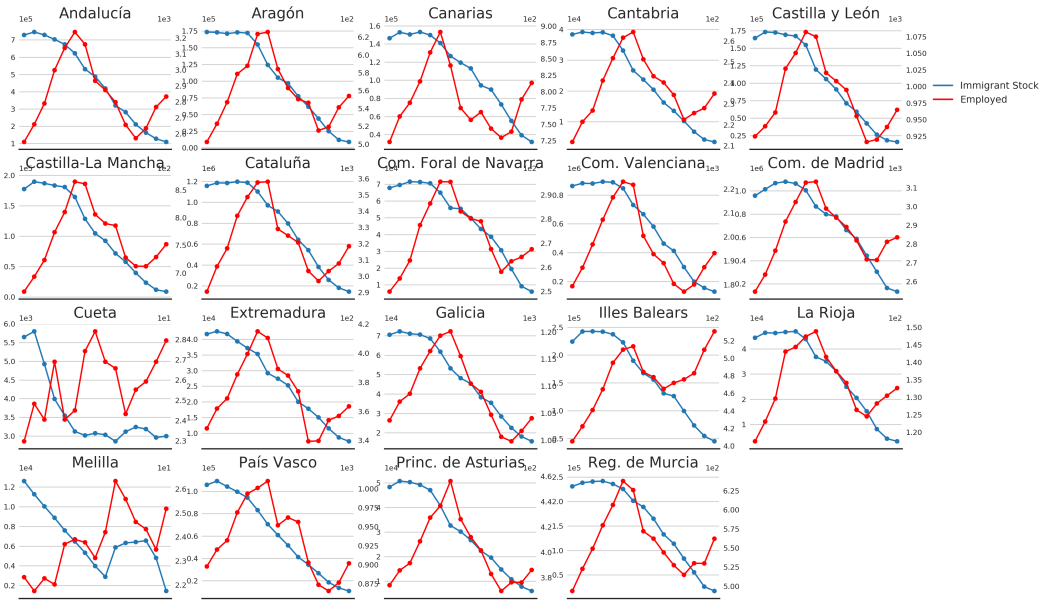
Immigrant Stock VS Native Population in Autonomous Communities



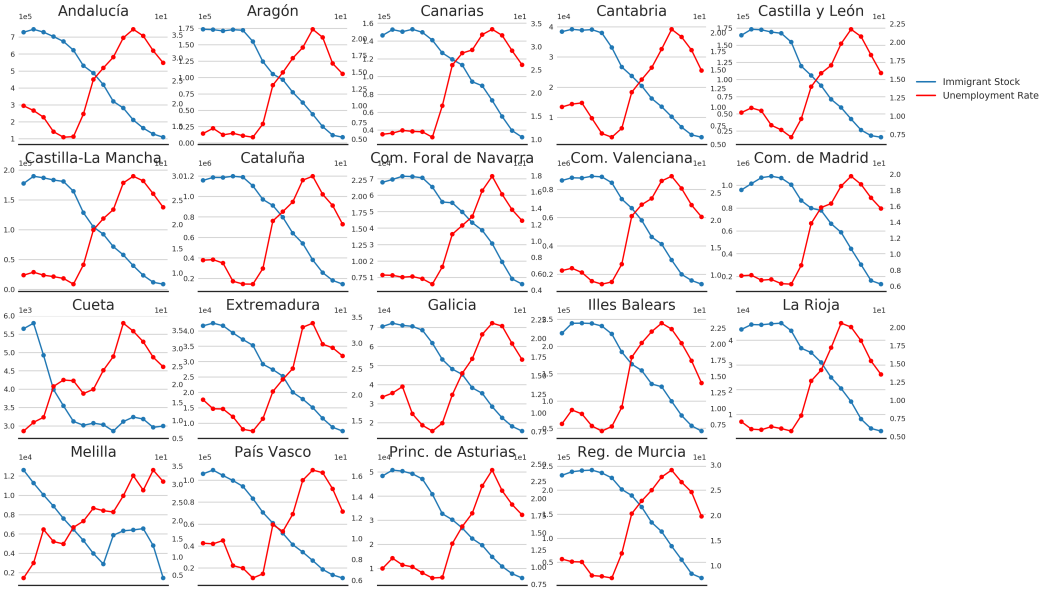
Immigrant Stock VS Activity Rate in Autonomous Communities



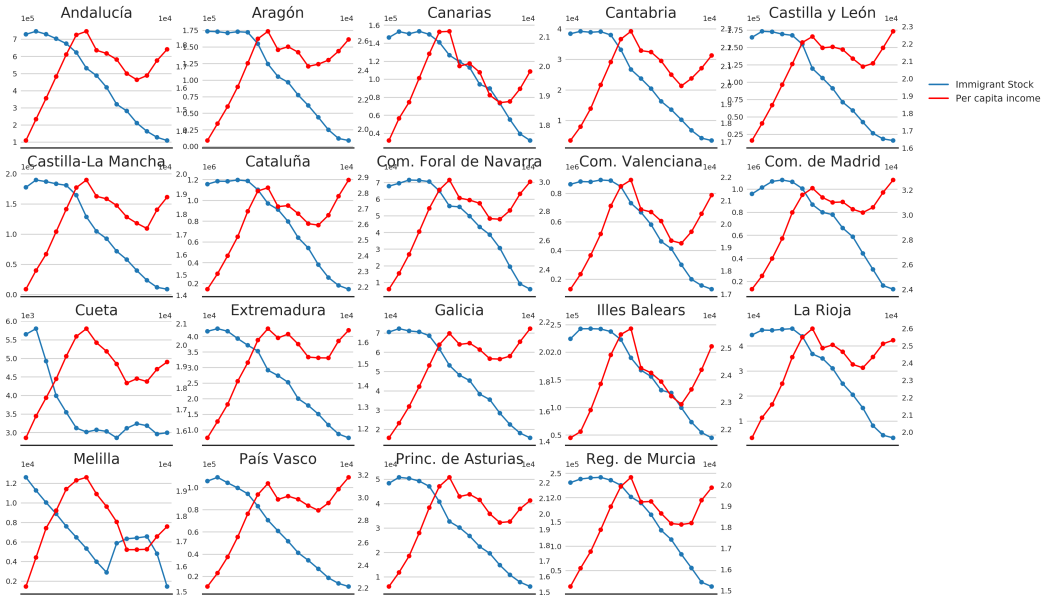
Immigrant Stock VS Employed in Autonomous Communities



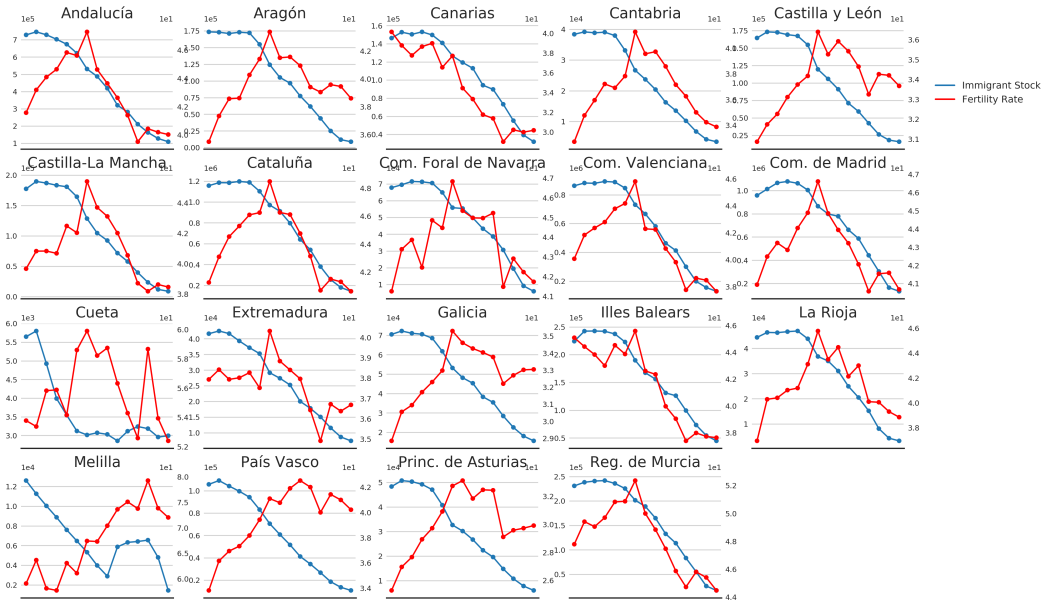
Immigrant Stock VS Unemployment Rate in Autonomous Communities

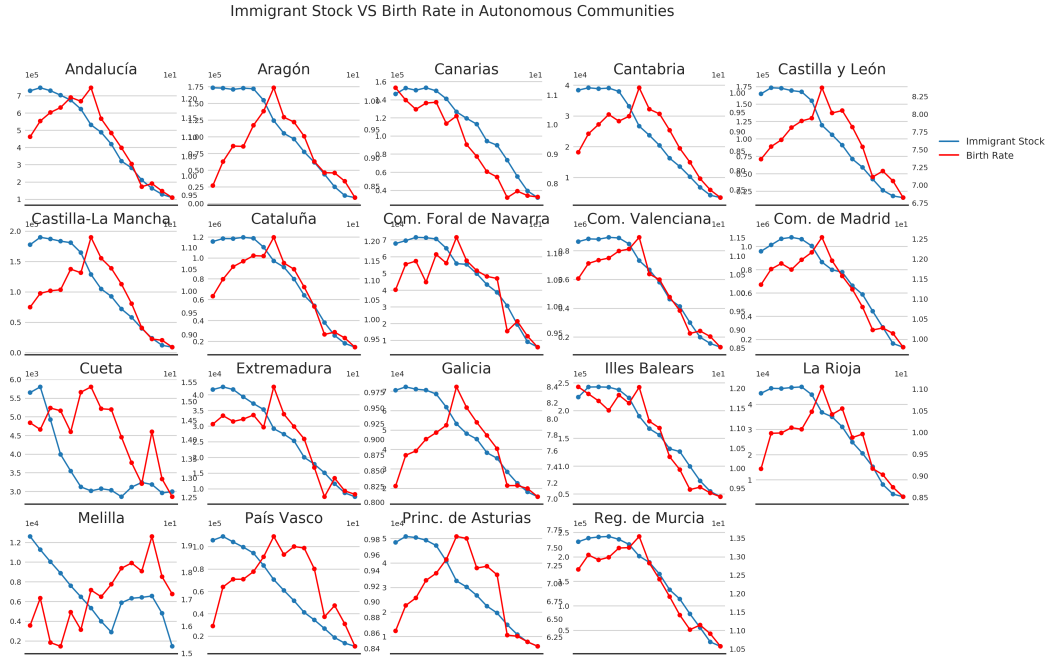


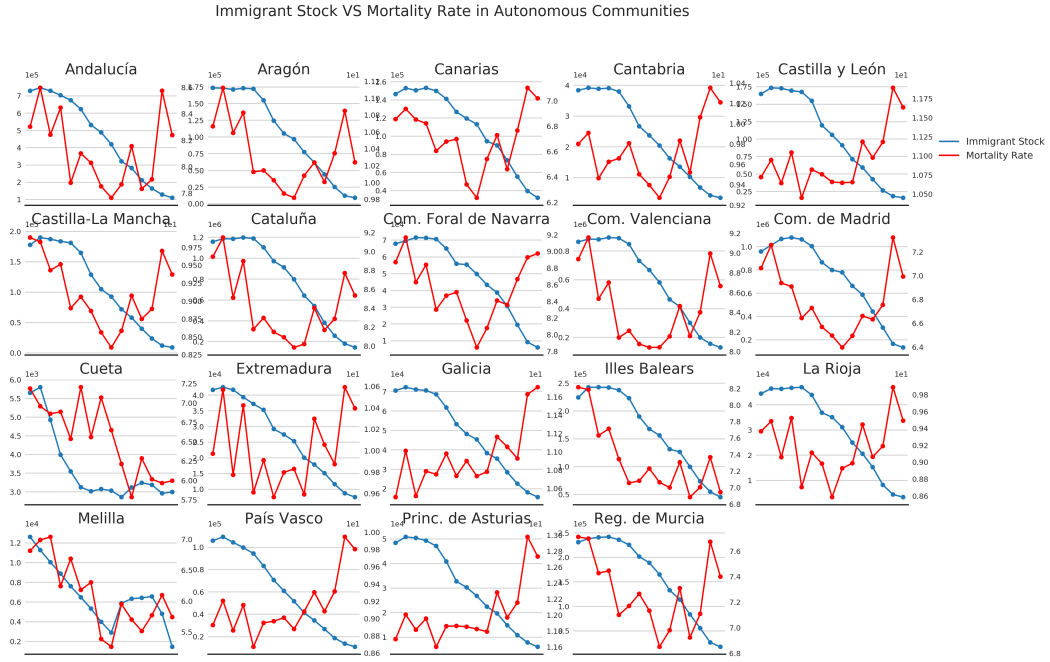
Immigrant Stock VS Per capita income in Autonomous Communities



Immigrant Stock VS Fertility Rate in Autonomous Communities







Unfortunately from the above plots is not possible to understand very well the relation between feature and immigrant flow. The only variables that seem to be interesting are Employed and Birth Rate.

Chapter 3

Models

In this Chapter, we will introduce the theoretical models used for our prediction aim: two very naive forecasting methods, a basic linear regression and an econometric model previously proposed in H. Jayet et al. paper [5].

3.1 Naive Forecasting Methods

As first, two of the most used time-series methods for forecasting are considered: Simple Moving Average (SMA)[16] and Exponential Smoothing (ES)[17].

3.1.1 Simple Moving Average

SMA is a technique to get an overall idea of the trends in a data set. It is an average of any subset of numbers. MSA is extremely useful for forecasting long-term trends.

SMAs are calculated by adding values over a given number of periods, then dividing the sum by the number of periods.

So a m-years SMA is defined by:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-m-1}}{m} \tag{3.1}$$

Year	Actual	Forecast	Calculation
2003	4		
2004	3		
2005	2		
2006	1.5	3	(4+3+2)/3
2007	1	2.67	(3+2+3)/3
2007		2.55	(2+3+2.67)/3

Table 3.1. 3-years SMA - Example

3.1.2 Exponential Smoothing

Exponential Smoothing is another simple forecasting method. It assigns exponentially decreasing weights as the observation get older. In other words, recent observations

are given relatively more weight in forecasting than the older ones. ES is usually a way of smoothing out the data by removing much of the noise (random effect) from the data by giving a better forecast.

ES is defined as:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t = \hat{Y}_t + \alpha(Y_t + \hat{Y}_t), \tag{3.2}$$

with $\hat{Y}_0 = Y_0$ and $\alpha \in [0,1]$ is the smoothing parameter.

It means that the next forecast is computed by interpolating between the last value and the forecast that had been made for it.

3.2 Regression Model for Panel Data

A simple regression model is also performed. The general expression is given by:

$$y_{i,t} = \beta^T x_{i,t-1} + \epsilon_{i,t}, \quad i = 1, \ldots, n \text{ and } t = 1, \ldots, T \tag{3.3}$$

where:

- $y_{i,t}$ is the dependent variable, here is the stock of foreign-born in territory i at time t
- β is the parameters vector
- $x_{i,t-1}$ is the independent variables vector, here considered at time $t-1$ not t

No unit-specific effect is used in the model, since it is assumed the randomness depends both on the time and the unit.

3.3 Spatial Error Model

In this section, some attempts to replicate the model specified in the H. Jayet et al. paper[5] are performed.

The initial model is defined as:

$$ln(n_{i,t}) = \beta ln(n_{i,t-1}) + \alpha_i + \gamma_t + u_{i,t} \quad i = 1, \ldots, I \text{ and } t = 1, \ldots, T, \tag{3.4}$$

where:

- $n_{i,t}$ is the stock of foreign-born in territory i at time t
- β is the parameter describing the network effect. The network effect is a local phenomenon that influences the immigrant choice, that is: foreign-born population tend to migrate to territories where a community of the same ethnic already exists
- $\alpha_i = x_i'\theta + \eta_i$, x_i' is the vector of all the time invariant observable location factors (feature vector), θ is the vector of coefficients and η_i is a random error term. The general form of a weights matrix is then: $w_{ij} \geq 0, w_{ij} = 0$ if $i = j, \sum_{i=1}^{n} w_{ij} = 1$

- γ_t measures the fixed time effect

$u_{i,t}$ is the *error term*. It can be modeled using a *Spatial Autoregressive Model*, so that $U_t = (u_{i,t}, \ldots, u_{I,t})$ follows:

$$u_t = \rho W u_t + w_t,$$

where:

- ρ is the autoregressive parameter
- W is a symmetric spatial weights matrix. For ease of interpretation, the weights matrix is often normalized so that the elements of each row sum up to one. This ensure all weights are between 0 and 1, moreover each row-normalized weight can be interpreted as the fraction of all spatial influence on unit i attributable to unit j
- $w_t \sim N(0, \sigma^2 I)$

The model just defined cannot be estimated using both the time and the location fixed effects α_i and γ_t. With an increasing sample size, maximum likelihood (ML) methods are asymptotically consistent, efficient and normally distributed. The ML estimates' consistency depends on the assumption that the number of parameters remains constant as the sample size increases.

Since the number of locations cannot increase, a larger sample in this model means a longer period. Thus, if the sample size increases, the number of fixed time effects increases. It is necessary to suppress them to consistently estimate the model.

In a general fixed model

$$y_{i,t} = \beta x_{i,t} + \alpha_t + u_{i,t} \quad i = 1, \ldots, I \text{ and } t = 1, \ldots, T,$$

it is possible to eliminate the fixed effect γ_t by:

- *fixed random transformation*: $y_{i,t} - \bar{y}_t = \beta(x_{i,t} - \bar{x}_t) + (\alpha_t - \bar{\alpha}) + (u_{i,t} - \bar{u}_{i,t})$
- *differencing with respect to a fixed unit*: $y_{i,t} - y_{I,t} = \beta(x_{i,t} - x_{I,t}) + (\alpha_t - \alpha_t) + (u_{i,t} - u_{I,t})$

Here, the differentiating method with respect to a reference location, I, is used:

$$\begin{aligned} ln(n_{i,t}) - ln(n_{I,t}) &= \beta\left(ln(n_{i,t-1}) - ln(n_{I,t-1})\right) + \alpha_i - \alpha_I + \gamma_t - \gamma_t + u_{i,t} - u_{I,t} \\ ln\left(\frac{n_{i,t}}{n_{I,t}}\right) &= \beta ln\left(\frac{n_{i,t-1}}{n_{I,t-1}}\right) + a_i + v_{i,t}, \end{aligned} \tag{3.5}$$

with:

- $i = 1, \ldots, I-1$ and $t = 1, \ldots, T$
- $a_i = \alpha_i - \alpha_I$
- $v_{i,t} = u_{i,t} - u_{I,t}$

Thus, $v_t = Qu_t$ with $v_t = (v_{1,t}, \ldots, v_{I-1,t})$, $u_t = (u_{1,t}, \ldots, u_{I,t})$, and $Q = [I_{I-1} \quad -1_{I-1}]$, being I_{I-1} an identity matrix of dimension $(I-1 \text{x} I-1)$ and -1_{I-1} the column vector of dimension $(I-1$ with all its element equal to -1.

Since the unobserved effects are correlated with the observed explanatory variable (i.e. $Cov\left(ln\left(\frac{n_{i,t-1}}{n_{I,t-1}}\right)\right) \neq 0$), *ordinary least squares estimates* are not consistent. To estimate the model is necessary to use the ML method.

In order to write the ML model, some computations are needed.

$$\begin{aligned} &u_t = \rho W u_t + w_t \Rightarrow u_t = (I-\rho W)^{-1}, \text{ where } w_t \sim N(0, \sigma^2 I), \text{ so} \\ &u_t \sim N\left(0, (I-\rho W)^{-1} \sigma^2 \left(I-\rho W^T\right)^{-1}\right) \\ &u_t = Qu_t \Rightarrow v_t \sim N\left(0, Q(I-\rho W)^{-1} \sigma^2 \left(I-\rho W^T\right)^{-1} Q^T\right) \end{aligned} \tag{3.6}$$

Let's call:

- $n_t = (n_{1,t}, \ldots, n_{I-1,t})$
- $n_{t-1} = (n_{1,t-1}, \ldots, n_{I-1,t-1})$
- $a = (a_1, \ldots, a_{I-1})$

then:

$$\begin{aligned} &ln\left(\frac{n_t}{n_{I,t}}\right) = \beta ln\left(\frac{n_{t-1}}{n_{I,t-1}}\right) + a + v_t, \quad t = 1, \ldots, T \\ &\text{so: } ln\left(\frac{n_t}{n_t}\right) \sim N\left(\beta ln\left(\frac{n_{t-1}}{n_{t-1}}\right) + a, Q(I-\rho W)^{-1} \sigma^2 \left(I-\rho W^T\right)^{-1} Q^T\right) \end{aligned} \tag{3.7}$$

Let's denote:

- $\mu = \beta ln\left(\frac{n_{t-1}}{n_{I,t-1}}\right) + a$
- $\Sigma = Q(I-\rho W)^{-1} \sigma^2 \left(I-\rho W^T\right)^{-1} Q^T$

then, it is possible to write the distribution of each vector $ln\left(\frac{n_t}{n_{I,t}}\right)$:

$$\begin{aligned} &f\left(ln\left(\frac{n_t}{n_{I,t}}\right)\right) = |2\pi\Sigma|^{-1/2} exp\left\{-\frac{1}{2}\left(ln\left(\frac{n_t}{n_{I,t}}\right) - \mu\right)^T \Sigma^{-1} \left(ln\left(\frac{n_t}{n_{I,t}}\right) - \mu\right)\right\} \\ &\Rightarrow |2\pi\Sigma|^{-1/2} = 2\pi^{-\frac{I-1}{2}} |\Sigma|^{-1/2} \\ &\Rightarrow |\Sigma|^{-1/2} = |Q(I-\rho W)^{-1} \sigma^2 \left(I-\rho W^T\right)^{-1} Q^T|^{-1/2} \\ &\qquad = |\sigma^2 Q(I-\rho W)^{-1} \left(I-\rho W^T\right)^{-1} Q^T|^{-1/2} \\ &\qquad = (\sigma^2)^{-\frac{I-1}{2}} |Q(I-\rho W)^{-1} \left(I-\rho W^T\right)^{-1} Q^T|^{-1/2} \\ &\qquad = (\sigma^2)^{-\frac{I-1}{2}} |L|^{-1/2} \end{aligned} \tag{3.8}$$

and $\Sigma^{-1} = \frac{1}{\sigma^2}\left[Q\left(I-\rho W\right)^{-1}\sigma^2\left(I-\rho W^T\right)^{-1}Q^T\right]^{-1} = \frac{1}{\sigma^2}L^{-1}$

Thus the distribution can be written as:

$$f\left(ln\left(\frac{n_t}{n_{I,t}}\right)\right) = (2\pi\sigma^2)^{\frac{I-1}{2}}|L|^{-1/2}exp\left\{-\frac{1}{2\sigma^2}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)\right\} \tag{3.9}$$

Let $ln\left(\frac{n}{n_I}\right)$ be a $(T\text{x}I-1)$ matrix whose T rows are $\left(ln\left(\frac{n_1}{n_{I,1}}\right),\ldots,ln\left(\frac{n_T}{n_{I,T}}\right)\right)$.

The matrix normal distribution of $ln\left(\frac{n}{n_I}\right)$ is defined as follows (it can be computed since the row component are conditionally- to the previous one - independent):

$$f\left(ln\left(\frac{n}{n_I}\right)\right) = \prod_{t=1}^{T}(2\pi\sigma^2)^{\frac{I-1}{2}}|L|^{-1/2}exp\left\{-\frac{1}{2\sigma^2}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)\right\} \tag{3.10}$$

The log-likelihood is then:

$$\begin{aligned} L\left(ln\left(\frac{n}{n_I}\right),\beta,a,\rho\right) &= -\frac{I-1}{2}\cdot T\cdot ln(2\pi\sigma^2) - \frac{T}{2}ln|L| - \\ &\quad - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right) \\ &\sim -T\cdot ln|L| - \frac{1}{\sigma^2}\sum_{t=1}^{T}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right), \end{aligned} \tag{3.11}$$

assuming a unitary variance, it can be written as:

$$L\left(ln\left(\frac{n}{n_I}\right),\beta,a,\rho\right) = T\cdot ln|L| - \sum_{t=1}^{T}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right), \tag{3.12}$$

maximize $L\left(ln\left(\frac{n}{n_I}\right),\beta,a,\rho\right)$ is equivalent to:

$$\min\ T\cdot ln|T| + \sum_{t=1}^{T}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right)^T L^{-1}\left(ln\left(\frac{n_t}{n_{I,t}}\right)-\mu\right) \tag{3.13}$$

Once $\hat{\beta}$ and $\hat{a}$ have been found, it is possible to estimate the coefficients of the time-invariant features through the formula:

$\hat{a}_i = a_i + \xi_i = \alpha_i - \alpha_I + \xi_i = (x_i^T - x_I^T)\theta + \epsilon_i + \xi_i$,

where $\epsilon_i = \eta_i - \eta_I$, θ is then estimated by ordinary least squares.

Chapter 4

Estimation Results

In this Section, we will report the estimation results obtained using the models introduced in the previous section on the Italy and Spain data. For each method, we will show the parameter estimates (if any), three different error metrics and a plot representing the real VS the predicted immigrant flow.

The model are trained using the Python Programming Language. All the code can be found on the github repositories [18][19]. The selection of the foreigners' nationalities included in the study is based on the distribution analysis performed in the chapter **??**. The selected origin countries are:

- Germany
- Morocco
- Peru
- Poland
- Romania

The four countries Morocco, Peru, Poland and Romania are chosen because they have a relevant migrant flow both to Italy and Spain. Germany is considered as a "control country", in the sense that the location motivation will be different for obvious reasons.

4.1 Italy

In order to compare the four models, the time period 2005-2016 is considered.

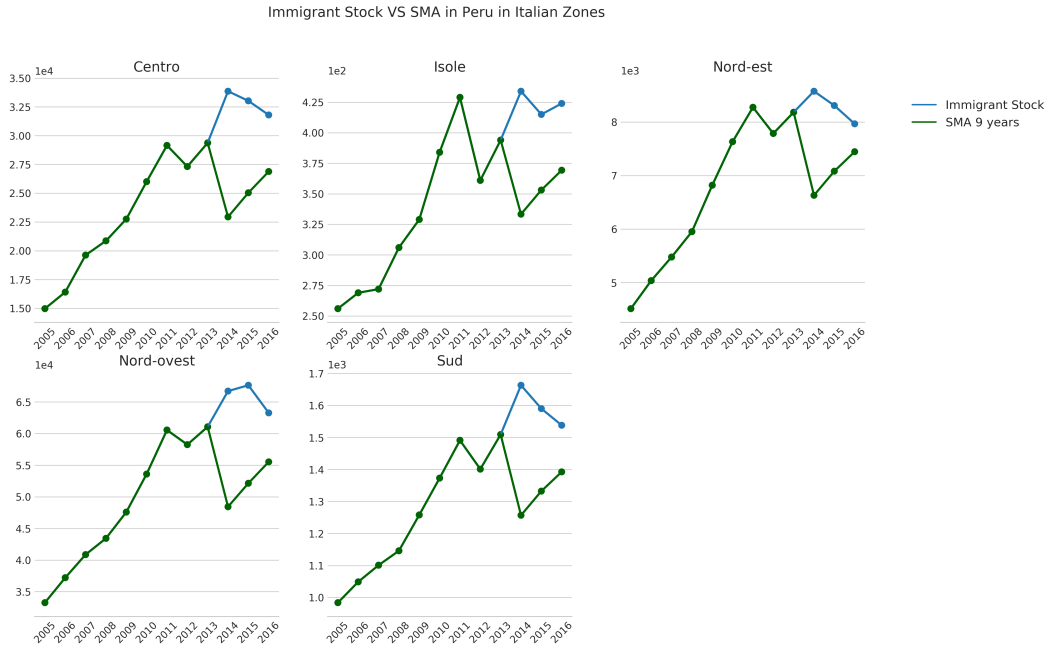
4.1.1 Simple Moving Average

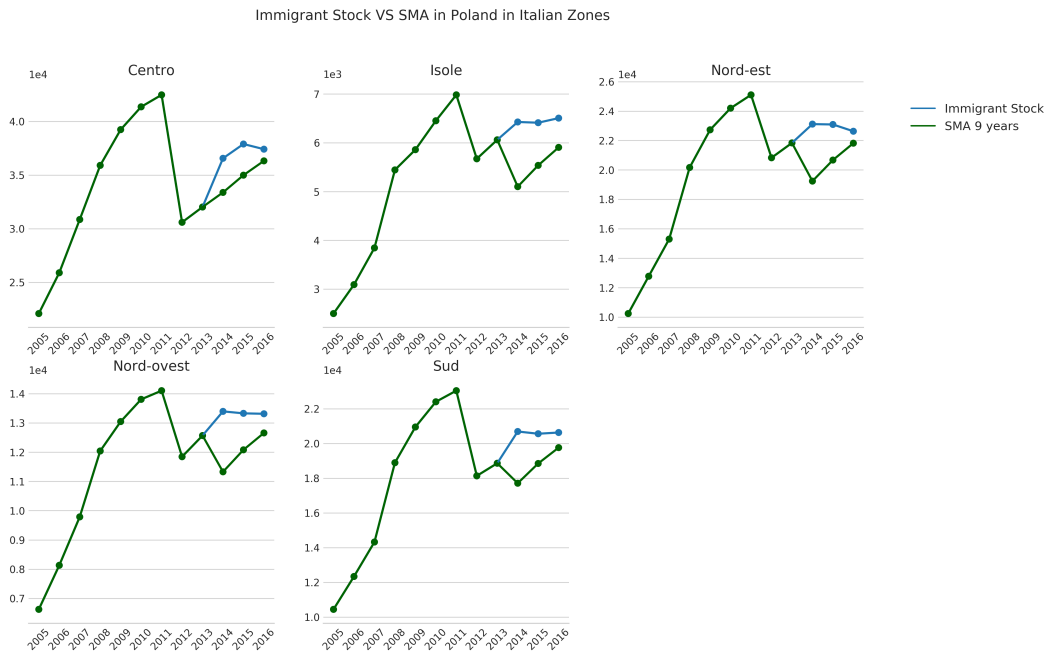
Here, for each of the five territories, a three-years SMA is used to forecast the last three years.

In order to understand the goodness of the method:

- a plot for each zone representing the real VS the forecasted value is included

- three different metrics are computed: Mean Absolute Error, Mean Squared Error, Root Mean Squares Error. The values shown in the table are the mean over the considered territories


Immigrant Stock VS SMA in Peru in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
SMA 9 years

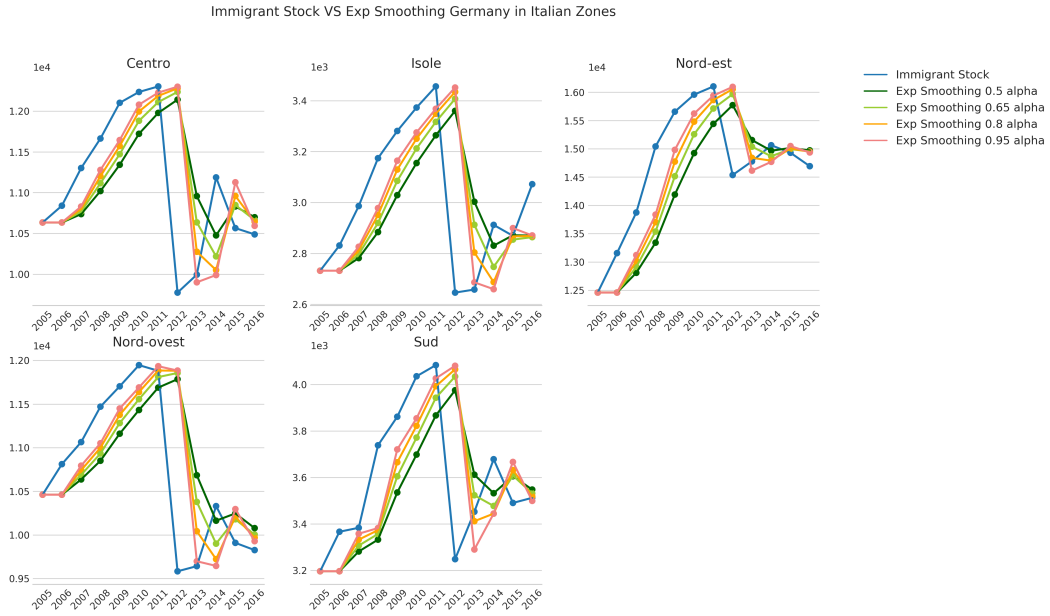

Immigrant Stock VS SMA in Poland in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
SMA 9 years

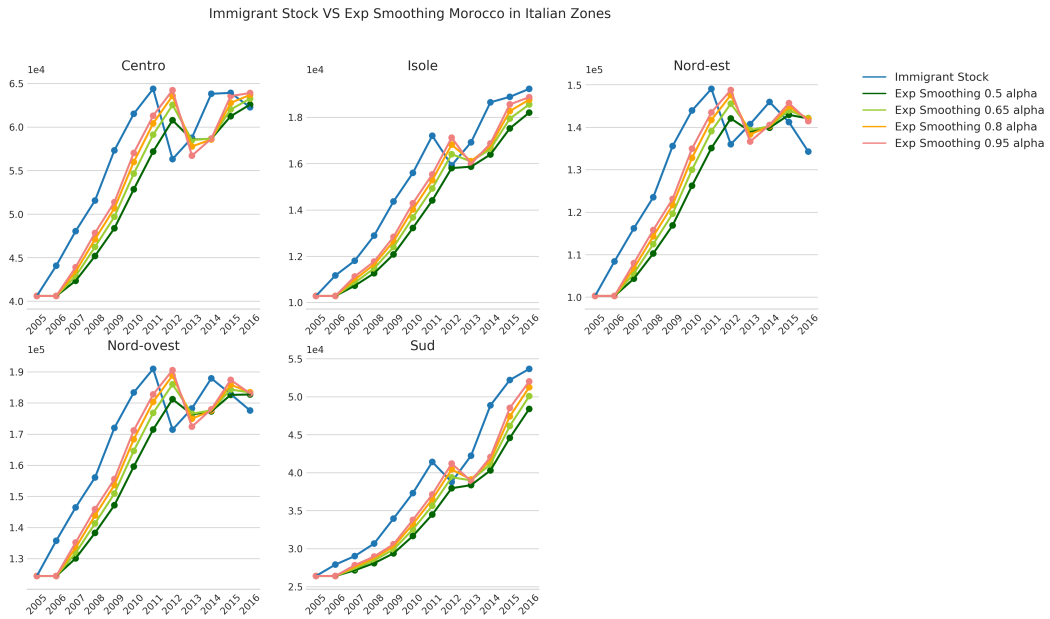
Country	MAE	MSE	RMSE
Germany	380.1	2.652e+05	515
Morocco	9959	1.426e+08	1.194e+04
Peru	4673	5.655e+07	7520
Poland	1774	4.199e+06	2049
Romania	7.494e+04	6.547e+09	8.092e+04

As it is possible to see from both the errors and the plots, this method seems to be too naïve to be able to predict the immigrant stock value.

4.1.2 Exponential Smoothing

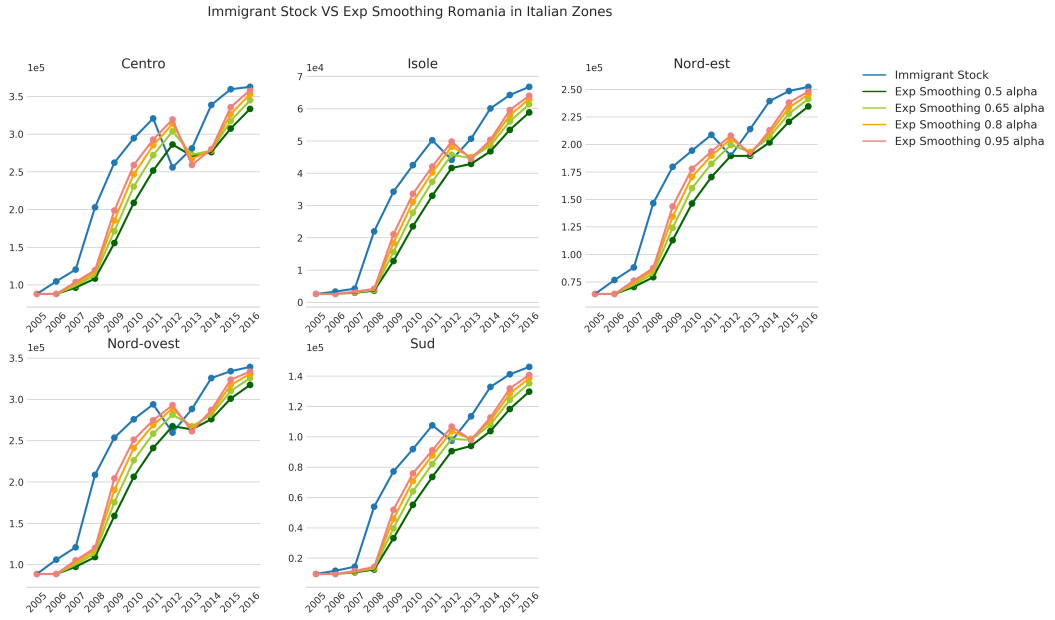
The Exponential Smoothing is then used on multiple smoothing parameters: 0.5, 0.65, 0.8, and 0.95. Again the three metrics are performed and the real VS forecasted plots shown.


Immigrant Stock VS Exp Smoothing Germany in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha


Immigrant Stock VS Exp Smoothing Morocco in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha


Immigrant Stock VS Exp Smoothing Peru in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha


Immigrant Stock VS Exp Smoothing Poland in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha



Metric	α	Germany	Morocco	Peru	Poland	Romania
MAE	0.5	513	6.68e+03	1.99e+03	2.6e+03	3.3e+04
	0.65	465	5.99e+03	1.65e+03	2.29e+03	2.8e+04
	0.8	424	5.55e+03	1.47e+03	2.02e+03	2.47e+04
	0.95	400	5.21e+03	1.39e+03	1.85e+03	2.22e+04
MSE	0.5	5.14e+05	8.29e+07	1.13e+07	1.22e+07	1.83e+09
	0.65	4.61e+05	6.28e+07	7.95e+06	9.89e+06	1.34e+09
	0.8	4.3e+05	5.17e+07	6.21e+06	8.45e+06	1.07e+09
	0.95	4.15e+05	4.51e+07	5.22e+06	7.53e+06	8.95e+08
RMSE	0.5	717	9.11e+03	3.37e+03	3.49e+03	4.27e+04
	0.65	679	7.93e+03	2.82e+03	3.14e+03	3.66e+04
	0.8	656	7.19e+03	2.49e+03	2.91e+03	3.26e+04
	0.95	644	6.72e+03	2.28e+03	2.74e+03	2.99e+04

The Exponential Smoothig leads to better results than the SMA. As it is possible to see, the choice of the smoothing parameter influences the performance of the method: the greater the weight to recent values, the best the ES.

4.1.3 Regression Model for Panel Data

For each origin country, multiples regression models are trained using different independent variables:

- only the previous time stock
- the previous two times stock
- the previous three times stock

- a set of 3, 5, 7, 10, 15 variables selected through the Mutual information criterion. MI between two random variables is a non-negative value, which measures the dependency between the variables. It quantifies the *amount of information* obtained about one random variable, through the other random variable. In the case of continuos random variables, it is defined as: $MI(X,Y) = \int_Y \int_X p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dx\, dy$. Where $p(x,y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability density function of X and Y respectively.

- the previous two times stock plus a set of seven features selected manually by looking at the plots

Each model is then trained on the period 2005-2013 and tested on the period 2014-2016.

Following the real VS predicted plots and tables including both the estimated parameters and the test errors.

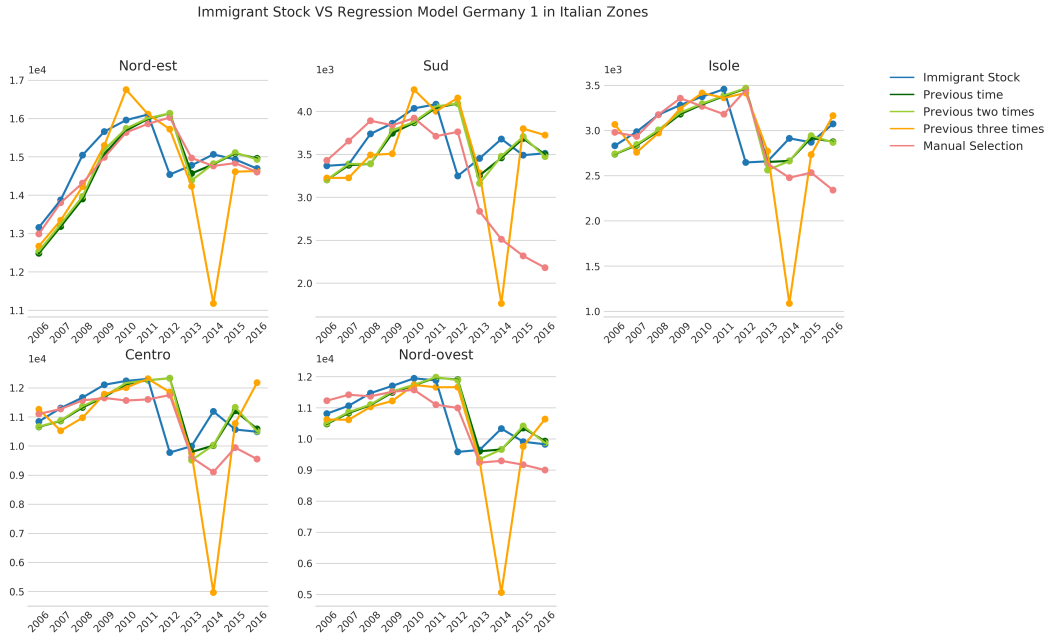
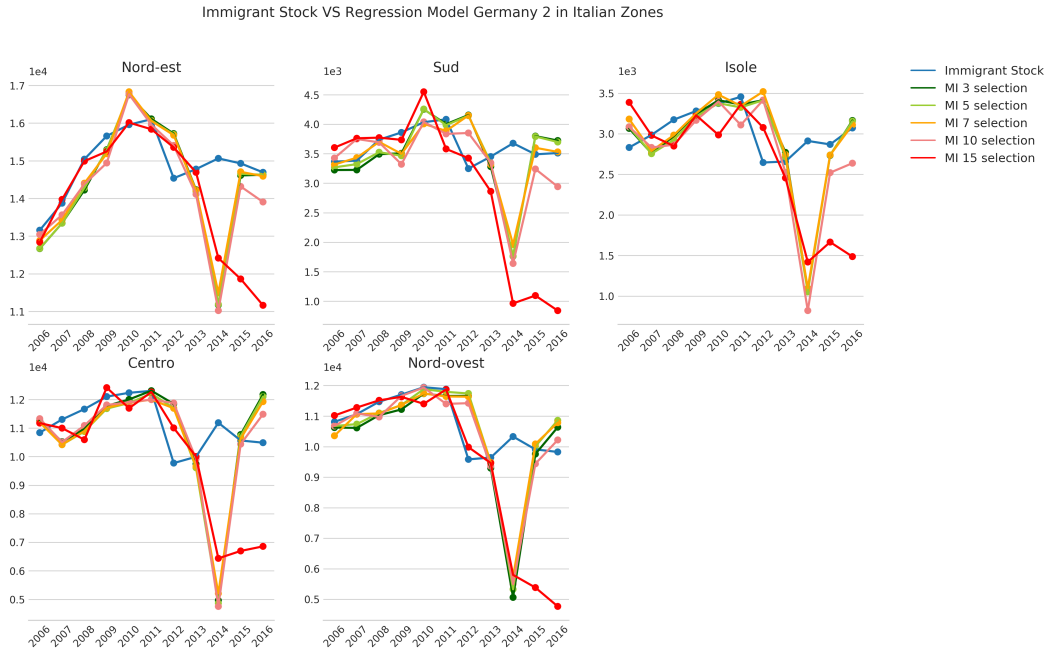


Table 4.1. Regression Results Germany

Indep. Var	I	II	III	IV	V	VI	VII	VIII	IX
y_{t-1}	1	1.11	0.79	0.728	0.79	0.791	0.767	0.72	0.373
y_{t-2}		-0.108	1.91	0.241	1.91	1.91	1.78	1.88	1.31
y_{t-3}			-1.76		-1.76	-1.79	-1.65	-1.71	-0.643
Native population - Total				0.0007		0.0001	0.0002	0.0003	-0.0011
Free activities in voluntary associations				-0.962				-2.66	2.62
Meetings in cultural, recreational or other associations				-0.921					
Disposable Income				-0.324					-1.2
Average monthly expenditure for housing				68.1					
Unemployment - Total				-2.89					
Reach Difficulty - Emergency room				-1.17					
Reach Difficulty - Post offices						-0.519	-2.93	-2.64	1.57
Internal Migration - Foreign country							-0.0041	0.0088	0.011
Reach Difficulty - Pharmacy							2.54	2.48	-3.77
Reach Difficulty - Municipal offices								-0.723	1.9
Pay money to an association								0.714	-1.91
Reach Difficulty - Supermarket									-3.27
Born alive									0.235
Accommodation and catering services									1.86
Non food									-0.0787
MAE	316	321	1.54e+03	794	1.54e+03	1.53e+03	1.4e+03	1.63e+03	3.18e+03
MSE	1.87e+05	1.99e+05	6.16e+06	8.92e+05	6.16e+06	6.07e+06	5.27e+06	6.18e+06	1.15e+07
RMSE	433	446	2.48e+03	944	2.48e+03	2.46e+03	2.3e+03	2.49e+03	3.39e+03



From both the figures and the errors it is possible to say that the variables influencing the most the models are the previous time variables (y_{t-1} and y_{t-2}). While y_{t-1} has always a positive influence, y_{t-2} have higher coefficients in most of the nine models, even if it does not influence the outcome in the same way for all the model (in the second one it has a slightly negative coefficient).

On the other hand, the *Native Population*, as well as the *Internal Migration*, does not seem to be discriminant at all or in the analysis.

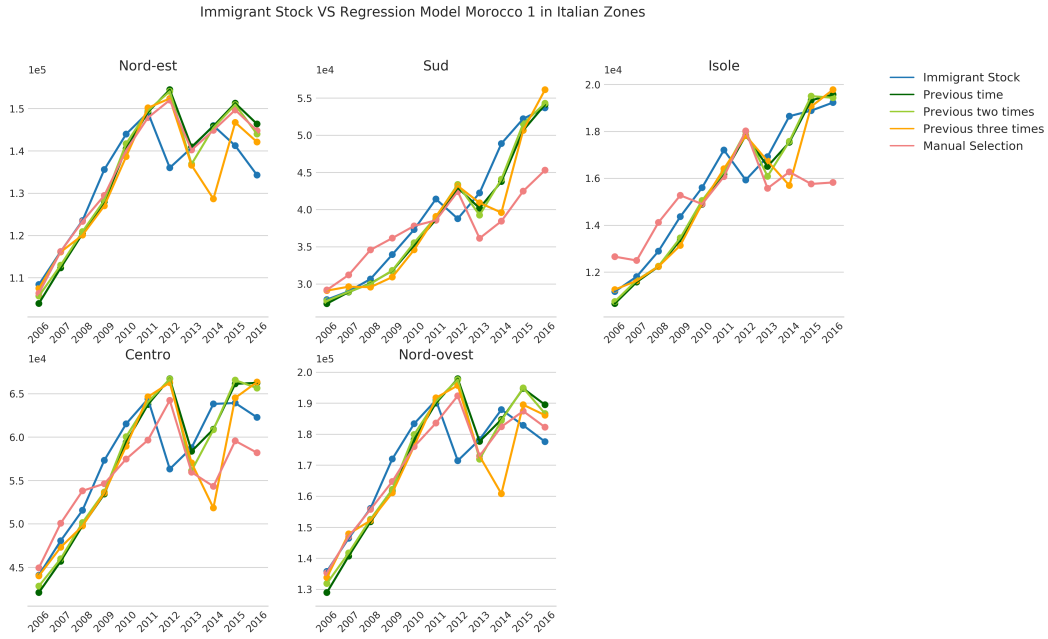
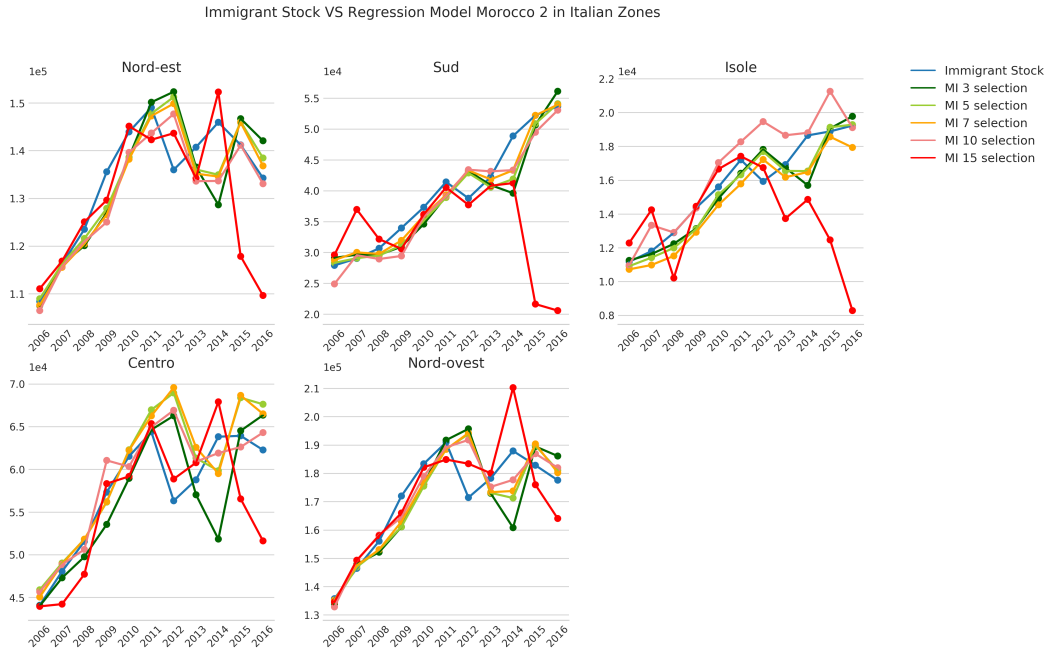


Table 4.2. Regression Results Morocco

Indep. Var	I	II	III	IV	V	VI	VII	VIII	IX
y_{t-1}	1.04	1.24	1.01	0.963	1.01	1.07	1.13	1.18	1.16
y_{t-1}		-0.209	0.693	0.103	0.693	0.209	0.206	-0.0594	-1.47
y_{t-3}			-0.727		-0.727	-0.383	-0.402	-0.254	0.964
Native population - Total				0.0043			0.0005	-0.0016	-0.0047
Free activities in voluntary associations				-3.34		1.04	-6.51	-3.13	3.28
Meetings in cultural, recreational or other associations				-21			-4.85	32.7	22.2
Disposable Income				-2.38					5.53
Average monthly expenditure for housing				524					
Unemployment - Total				-26.4					
Reach Difficulty - Emergency room				-2.6					
Internal Migration - Foreign country						0.173	0.227	0.207	0.365
Internal Migration - Italy								0.094	0.196
Free activities in non voluntary associations								-43.6	-83
Pay money to an association								-7.78	11.6
Reach Difficulty - Post offices									-15.7
Other goods and services									33
Non food									-1.1
Political Info - Every day									-7.33
MAE	4.49e+03	4.1e+03	7.1e+03	6.01e+03	7.1e+03	4.81e+03	4.39e+03	3.27e+03	1.41e+04
MSE	3.99e+07	3.21e+07	1e+08	4.55e+07	1e+08	4.13e+07	3.46e+07	2.31e+07	2.91e+08
RMSE	6.32e+03	5.66e+03	1e+04	6.74e+03	1e+04	6.42e+03	5.88e+03	4.81e+03	1.71e+04



As for Germany, even for Morocco, the previous time stocks are quite important. Here, the most influent between the three *time variables* is y_{t-1}, while y_{t-2} and y_{t-3} lose importance.

The variables that seem to affect the most (in terms of absolute coefficients) are *Free activities in voluntary associations* and *Meetings in cultural, recreational or other associations*. Even if the two variables have high coefficients, they are not influencing, in the same way, the models, i.e. the coefficients have a different sign in the nine models.

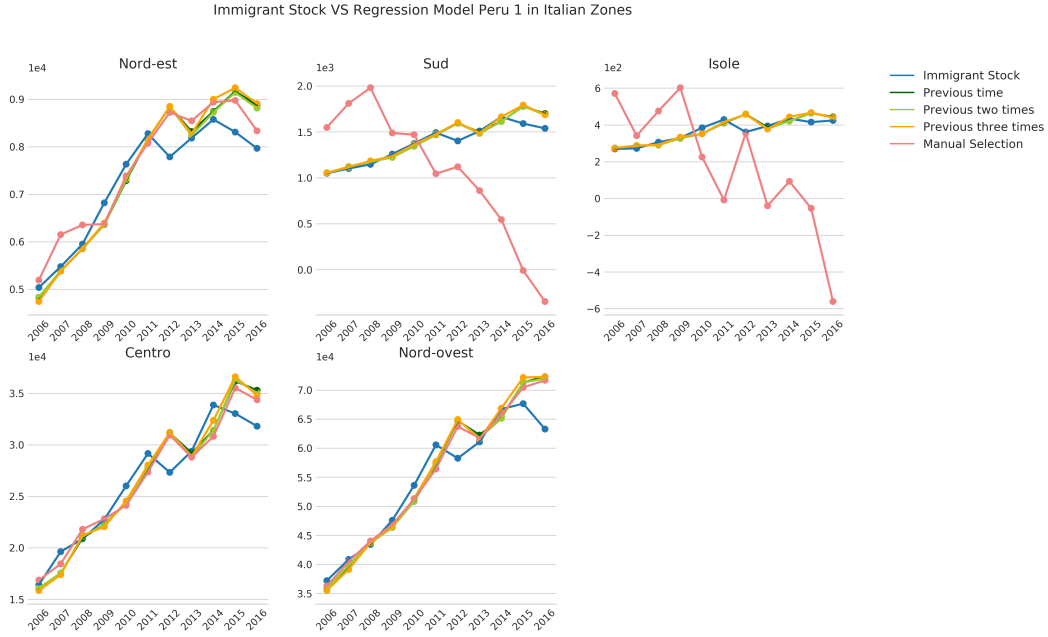
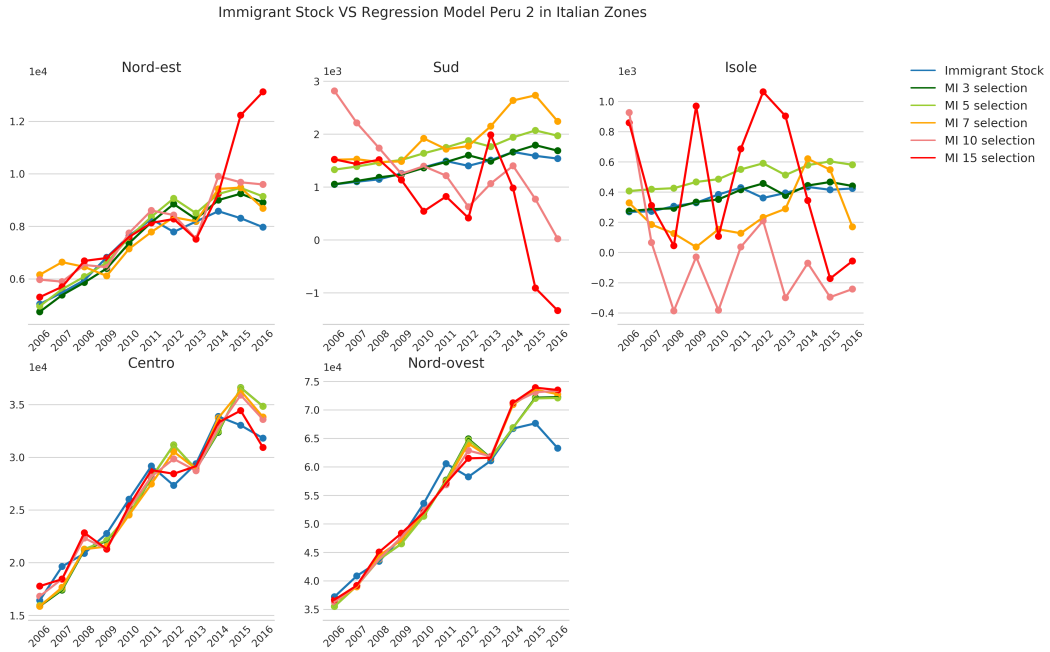


Table 4.3. Regression Results Peru

Indep. Var	I	II	III	IV	V	VI	VII	VIII	IX
y_{t-1}	1.07	1.13	1.15	1.04	1.15	1.14	1.09	0.864	0.818
y_{t-2}		-0.0706	-0.275	0.0434	-0.275	-0.282	-0.58	-0.406	-0.818
y_{t-3}			0.205		0.205	0.22	0.641	0.671	1.02
Native population - Total				0.001		0	0.0009	0.0012	-0.0001
Free activities in voluntary associations				-2.4		0.0326	-2.41	-0.195	3.36
Meetings in cultural, recreational or other associations				-1.33				-9.83	-10.7
Disposable Income				-0.307			-1.1	-0.509	0.364
Average monthly expenditure for housing				61.8					
Unemployment - Total				-3.4					
Reach Difficulty - Emergency room				-2.21					
Internal Migration - Foreign country							0.0155	-0.0463	-0.058
Reach Difficulty - Post offices								-4.76	-5.06
Pay money to an association								4.55	6.31
Internal Migration - Italy									0.0473
Reach Difficulty - Supermarket									-1.58
Other goods and services									4.94
Food and non-alcoholic beverages									4.69
Non food									-1.77
MAE	1.71e+03	1.69e+03	1.63e+03	1.87e+03	1.63e+03	1.73e+03	2.08e+03	2.3e+03	2.72e+03
MSE	8.43e+06	8.04e+06	8.46e+06	7.37e+06	8.46e+06	8.3e+06	1.09e+07	1.16e+07	1.5e+07
RMSE	2.9e+03	2.84e+03	2.91e+03	2.72e+03	2.91e+03	2.88e+03	3.29e+03	3.41e+03	3.87e+03



Also in Peru, most of the dependent variable can be explained just by looking at the previous time variables. *Reach Difficulty* variables, as for Germany, influences negatively a lot the outcome.

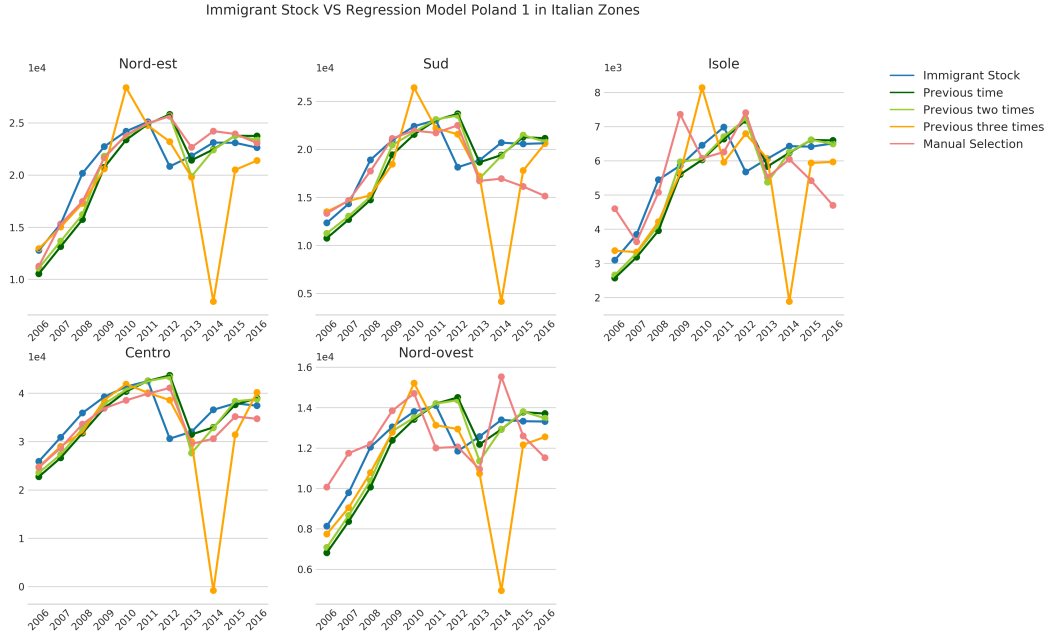
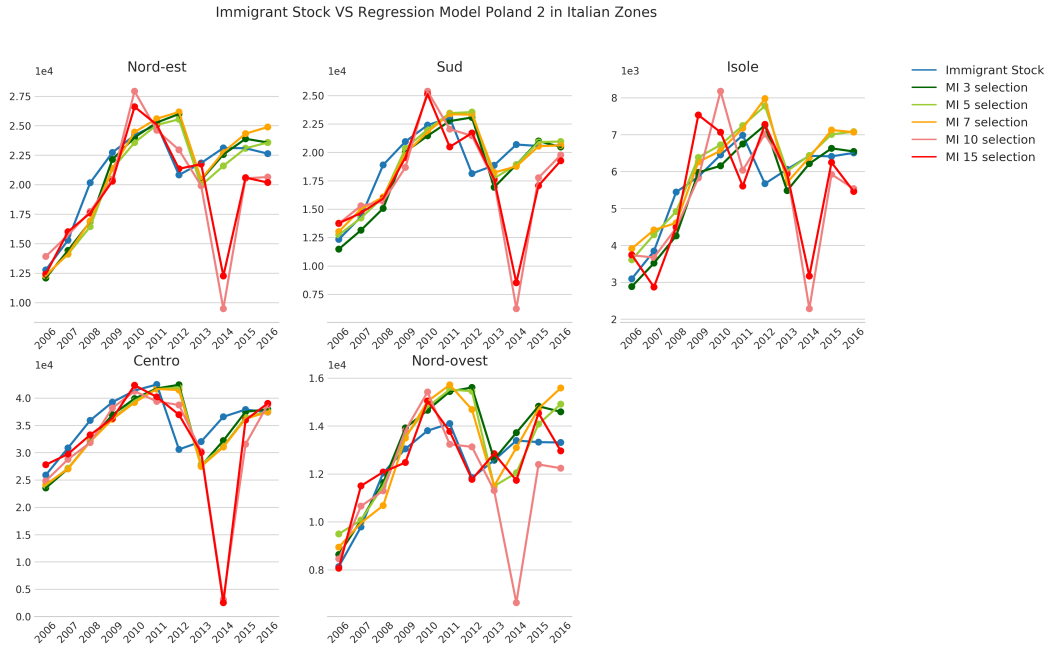


Table 4.4. Regression Results Poland

Indep. Var	I	II	III	IV	V	VI	VII	VIII	IX
y_{t-1}	1.03	1.3	0.558	0.96	1.21	1.07	1.07	0.591	0.56
y_{t-2}		-0.284	2.45	0.0205	-0.244	-0.122	-0.113	2.1	2.2
y_{t-3}			-2.21					-1.91	-2.09
Native population - Total				0.003		0	0.0008	0.0003	0.0018
Free activities in voluntary associations				-0.212				-3.69	2.38
Meetings in cultural, recreational or other associations				-12.2					-24.8
Disposable Income				-0.495					
Average monthly expenditure for housing				285					
Unemployment - Total				-11.1					
Reach Difficulty - Emergency room				-6.9					-7.01
Pay money to an association					0.648	2.2	2.09	0.676	4.94
Internal Migration - Foreign country						-0.0526	-0.0802	0.0237	0.0698
Reach Difficulty - Post offices							-4.59	-4.43	8.79
Political Info - Every day							-0.453	-0.18	0.936
Reach Difficulty - Pharmacy								4.82	7.28
Reach Difficulty - Supermarket									-17.1
Political Info - Never									-0.0733
Political Info - Some times in a week									5.38
MAE	817	773	6.73e+03	2.36e+03	905	1.12e+03	1.22e+03	6.12e+03	5.21e+03
MSE	1.4e+06	1.38e+06	1.38e+08	8.62e+06	1.92e+06	2.89e+06	3.47e+06	1.09e+08	9.82e+07
RMSE	1.18e+03	1.18e+03	1.17e+04	2.94e+03	1.39e+03	1.7e+03	1.86e+03	1.05e+04	9.91e+03

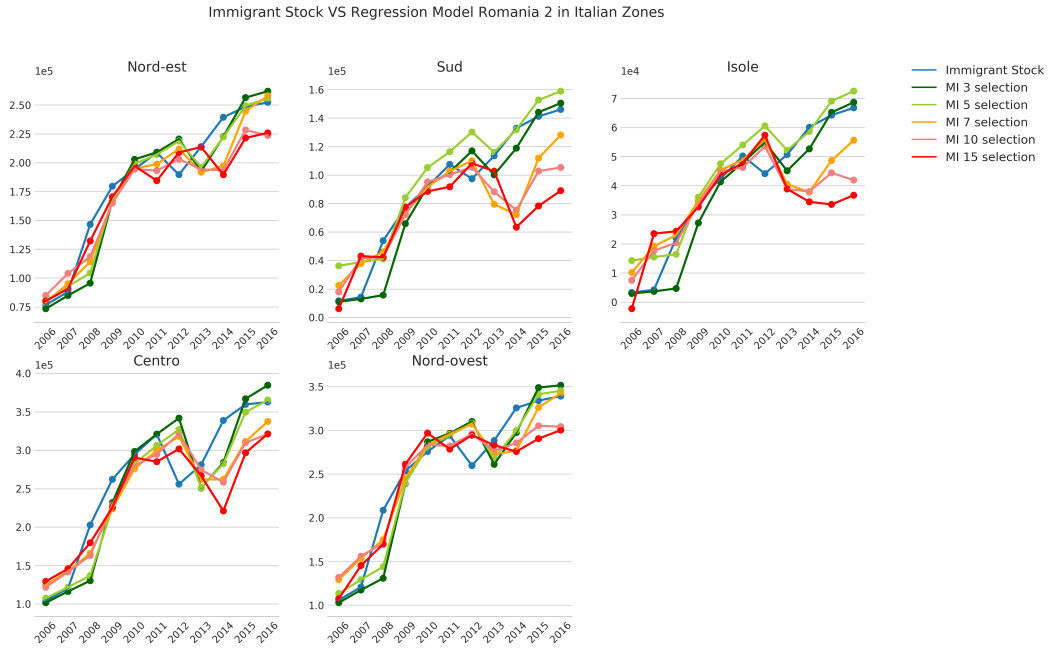


As for Germany, the models including just the first two *time varibles* are better than the other ones with additional features. In Poland, the y_{t-3} variable has always a negative effect on the actual stock.

Reach Difficulty variables still have a large impact on the outcome variable, having high overall negative coefficients.

Table 4.5. Regression Results Romania

Indep. Var	I	II	III	IV	V	VI	VII	VIII	IX
y_{t-1}	1.08	1.33	1.28	0.799	1.28	1.09	0.708	0.748	0.549
y_{t-2}		-0.282	-0.081	0.0691	-0.081	-0.0817	0.156	0.127	0.352
y_{t-3}			-0.172		-0.172	-0.0894	-0.109	-0.172	-0.564
Native population - Total				0.0021					
Free activities in voluntary associations				15.6				-79.7	-101
Meetings in cultural, recreational or other associations				-72.5					
Disposable Income				17.9		3.05	13.3	13.8	43
Average monthly expenditure for housing				1.68e+03					
Unemployment - Total				-154			-105	-96	-163
Reach Difficulty - Emergency room				-18.3					
Average age of fathers at birth						-69.9	-19.4	-1.09e+04	-8.05e+04
Communications							-84.6	-48.6	132
Internal Migration - Foreign country								1.17	1.62
Average age of mothers at birth								1.26e+04	9.05e+04
Reach Difficulty - Pharmacy									-1.17
Born alive									-1.63
Clothing and footwear									-95.9
Other goods and services									-121
Transport									92.1
MAE	1.23e+04	1.16e+04	1.37e+04	3.63e+04	1.37e+04	1.11e+04	2.79e+04	3.83e+04	4.89e+04
MSE	2.38e+08	2.1e+08	3.62e+08	1.77e+09	3.62e+08	3.14e+08	1.26e+09	1.71e+09	2.93e+09
RMSE	1.54e+04	1.45e+04	1.9e+04	4.21e+04	1.9e+04	1.77e+04	3.55e+04	4.13e+04	5.41e+04



Also in Romania, the coefficients of y_{t-3} are negative and y_{t-1} is the variable that explains, on its self, most of the dependent variable.

4.1.4 Spatial Error Model

As for the regression model, years 2005-2013 are used for training, 2014-2016 for testing.

To implement the model, it necessary to build a spatial weights matrix W. To do that, useful information are available on the ISTAT website.

ISTAT releases the origin-destination matrices of distances in meters and travel times (in minutes) between all Italian municipalities. The matrices are grouped by Region. The files are provided in text format and all Italian municipalities. The islands are treated separately. The matrices of Sicily and Sardinia contain only the between-region distances (the distances only of the municipalities of the regions). In a separate excel file are available the distances for the main ports that connect the islands with respect to Peninsular Italy. This allows you to add the travel time by ship to the routes calculated from the main ports of connection to the islands.

There are some problems with the different tables:

- the column names are not always the same
- column are index may be inverted: for all the regions but Lombardy the region-provinces are in the "Destination" column and the other provinces in the "Origin" column. In the Lombardy table the two data are inverted (thus, the need to transpose the matrix)
- it was not possible to find the Sicily and Sardinia ports distances

To compute the distances between Sicily-Sardinia and the other Italian regions, it is possible to manually detect (GoogleMaps) some useful distances between the main ports and compute the distances as the sum of multiple component.

In particular:

- from Sicily:

$$\begin{cases} d_{(\text{O, Trapani})} + d_{(\text{Trapani, Cagliari})} + d_{(\text{Cagliari, D})}, & \text{D in Sardinia} \\ d_{(\text{O, Messina})} + d_{(\text{Messina, Villa San Giovanni})} + d_{(\text{Villa San Giovanni, D})}, & \text{otherwise} \end{cases}$$

- from Sardinia:

$$\begin{cases} d_{(\text{O, Cagliari})} + d_{(\text{Cagliari, Trapani})} + d_{(\text{Trapani, D})}, & \text{D in Sicily} \\ min\left(d_{(\text{Messina})}, d_{(\text{Livorno})}, d_{(\text{Civitavecchia})}\right), & \text{otherwise} \end{cases}$$

where

- O and D stay for origin and destination
- $d_{(\text{Messina})} = d_{(\text{origin, Cagliari})} + d_{(\text{Cagliari, Trapani})} + d_{(\text{Trapani, Messina})} + d_{(\text{Messina, Villa San Giovanni})} + d_{(\text{Villa San Giovanni, destination})}$
- $d_{(\text{Livorno})} = d_{(\text{origin, Olbia})} + d_{(\text{Olbia, Livorno})} + d_{(\text{Livorno, destination})}$
- $d_{(\text{Civitavecchia})} = d_{(\text{origin, Olbia})} + d_{(\text{Olbia, Civitavecchia})} + d_{(\text{Civitavecchia, destination})}$

The regions distance matrix is computed as the mean distance between all the province pairs belonging to the two regions considered.

The zones distance matrix is obtained by taking the mean distance between all the regions pairs belonging to the two zones considered.

As spatial weights the inverse of the squared of the distances are used. The result matrix W is a symmetric, non-negative matrix with $w_{ij} \geq 0$ and $w_{ii} = 0$.

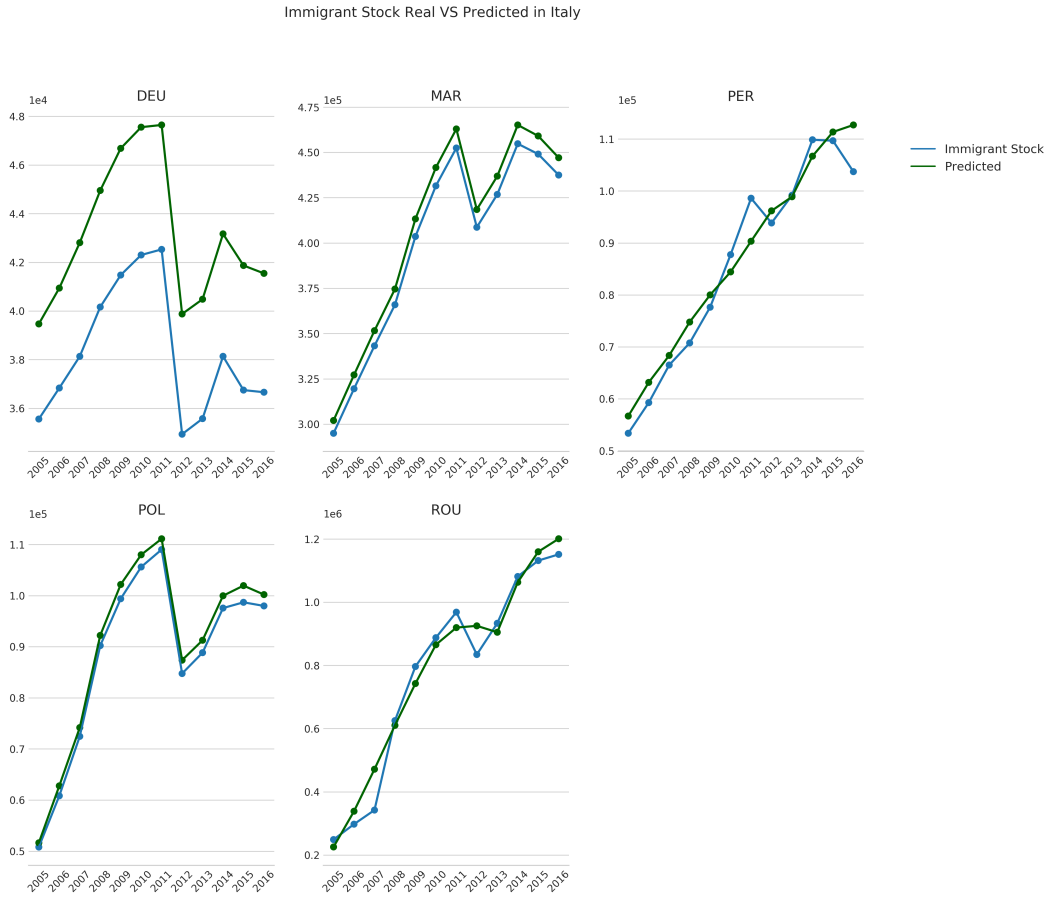
The row-normalized W is used for ease of interpretation. It is defined as $\sum_{j=1}^{n} w_{ij} = 1 \quad \forall i = 1, \ldots, n$. This ensure that all weights are between 0 and 1. Each row-normalized weight, w_{ij}, can be interpreted as the fraction of all spatial influence on unit i attributable to unit j.

The model specified is used to predict the immigrant stock at Italian zones level. Same features, times period and origin countries used in the previous regression model are considered.

All the features are assumed to be time invariant, thus as reference period the 2013 is considered.

The model also requires a reference territory, the estimation of the Italy stock at the current time is used.

The prediction of the overall Italian flow is obtained through ridge regression models whose independent variables are chosen using an automatic feature selection based on mutual information. Follows the results and the plot of the prediction.



Once the flow of the referred territory is estimated, for each country different models are trained using different examples: from 1 to 7 independent variables selected using MI criterion.

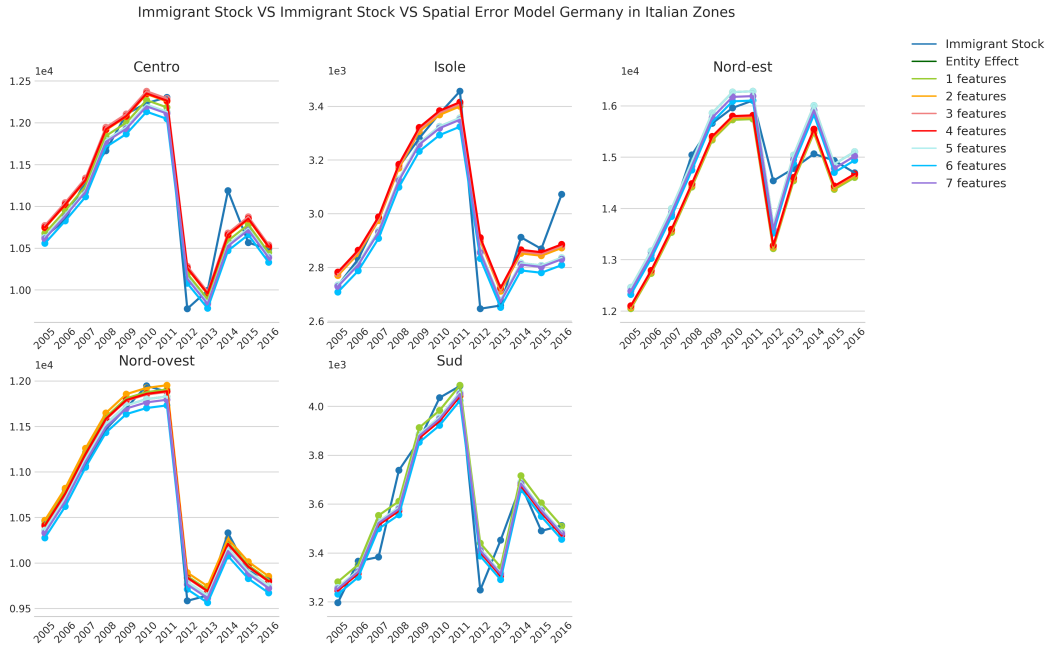
Country	β	ρ
Germany	0.981500	7.035800
Morocco	0.814100	7.041200
Peru	0.513300	7.023900
Poland	0.857800	7.032500
Romania	0.926300	6.865800

The table reports the results of the first stage of the model for the five origin countries, i.e. the network effect β and the spatial autocorrelation coefficient ρ.

The presence of spatial autocorrelation seems to be confirmed by positive coefficients.

The estimated network effect β is also positive for all the specific nationalities. It is reasonably high, ranging from 0.98 (Germany) to 0.51 (Peru).

As result from the first stage also the *fixed effects* are obtained. They represent the attractiveness of the zone discarding the network effect, the five measures are considered with respect to the reference territory Italy. The *fixed effects* are shown in the first column of the following results tables.


Immigrant Stock VS Immigrant Stock VS Spatial Error Model Germany in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Entity Effect
1 features
2 features
3 features
4 features
5 features
6 features
7 features

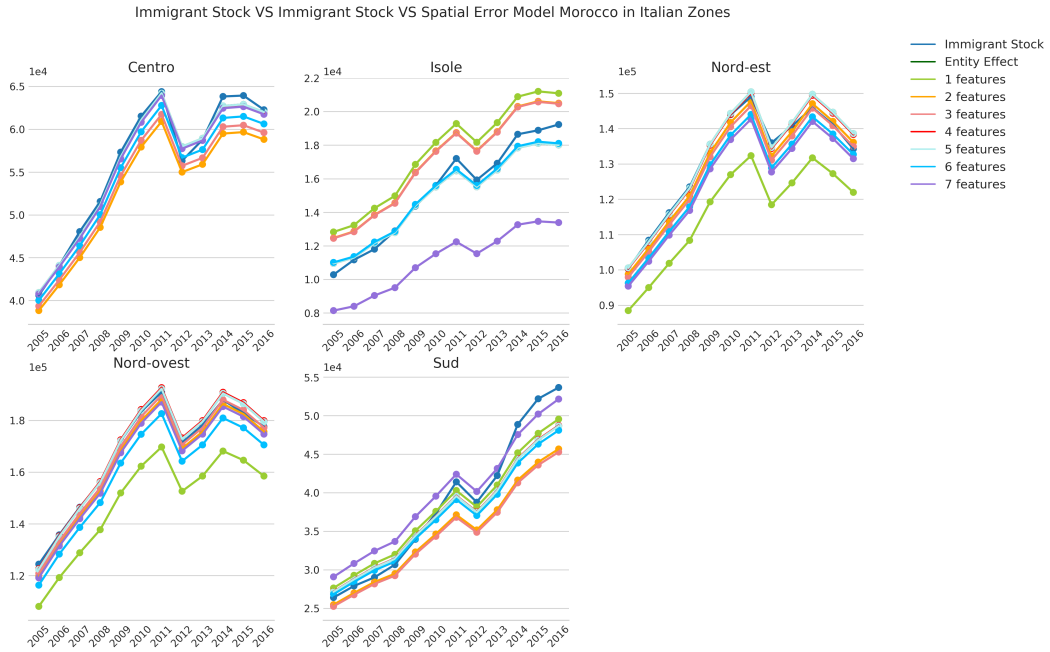

Immigrant Stock VS Immigrant Stock VS Spatial Error Model Morocco in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Entity Effect
1 features
2 features
3 features
4 features
5 features
6 features
7 features

Table 4.6. H. Jayet et al. paper Results Germany

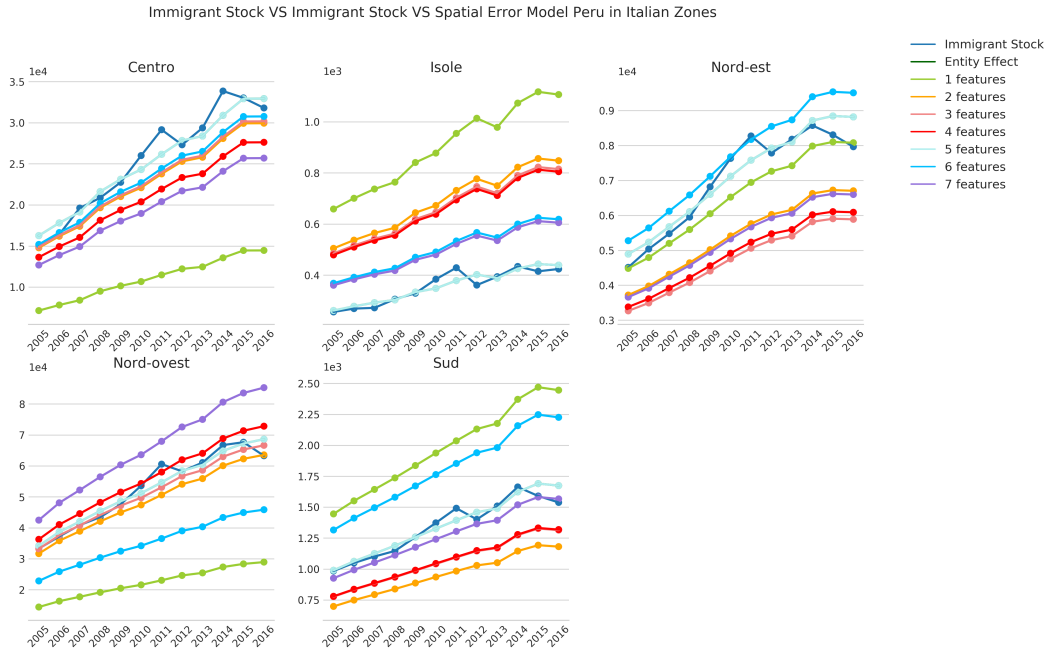
Indep. Var	I	II	III	IV	V	VI	VII	VIII
a_{Center}	-0.0358							
$a_{Islands}$	-0.0559							
$a_{NorthEast}$	-0.0022							
$a_{NorthWest}$	-0.0375							
a_{South}	-0.0421							
Free activity for a union		0.0001	0.0002	0.0001	-0	0.0613	0	0
Political Info - Never			-0	-0	-0	0.0089	-0	-0
Reach Difficulty - Municipal offices				0	0	-0.0027	0.0001	0.0001
Political Info - Some times in a year					0.0001	-0.0524	-0.0001	-0.0002
Average age of mothers at birth						0.1769	-0.0007	-0.0016
Communications							0.0002	0.0001
Total fertility rate								0.0273
MAE	205.9	167.3	165	164.3	160.2	205.9	220.8	207.3
MSE	1.084e+05	6.436e+04	6.092e+04	5.988e+04	6.003e+04	1.084e+05	9.607e+04	9.703e+04
RMSE	329.2	253.7	246.8	244.7	245	329.2	310	311.5

Table 4.7. H. Jayet et al. paper Results Morocco

Indep. Var	I	II	III	IV	V	VI	VII	VIII
a_{Center}	-0.3706							
$a_{Islands}$	-0.6142							
$a_{NorthEast}$	-0.2113							
$a_{NorthWest}$	-0.163							
a_{South}	-0.4477							
Internal Migration - Foreign country		0	0	0	-0	-0	-0	0
Free activity for a union			-0.0017	-0.0016	0.0113	0.0155	0.0018	0.0013
Recreation, shows and culture				-0	0.0009	0.0011	0.0002	0.0003
Reach Difficulty - Municipal offices					-0.0006	-0.0009	-0.0001	0.0002
Work Satisfaction - Not						0.004	-0.0074	-0.0308
Pay money to an association							0.0001	0
Education								0.001
MAE	2644	7932	3190	2777	2681	2644	3479	2814
MSE	9.781e+06	1.136e+08	1.675e+07	1.615e+07	9.777e+06	9.781e+06	1.702e+07	1.069e+07
RMSE	3128	1.066e+04	4093	4019	3127	3128	4126	3269

Table 4.8. H. Jayet et al. paper Results Peru

Indep. Var	I	II	III	IV	V	VI	VII	VIII
a_{Center}	-0.6077							
$a_{Islands}$	-2.679							
$a_{NorthEast}$	-1.217							
$a_{NorthWest}$	-0.242							
a_{South}	-2.03							
Internal Migration - Foreign country		0	0	0	0	-0.0003	-0	0
Furniture, articles and services for the house			-0.0026	-0.0045	-0.0036	0.1077	-0.0029	-0.0079
Communications				0.0027	0.002	-0.0052	0.0081	0.0093
Free activities in non voluntary associations					-0.0012	0.0651	0.0103	0.0078
Education						-1.114	-0.0547	-0.0225
Work Satisfaction - Not							-0.0366	0.1484
Average monthly expenditure for housing								-0.3672
MAE	913.3	1.165e+04	2024	1919	2770	913.3	5190	5389
MSE	2.893e+06	3.559e+08	8.575e+06	6.072e+06	1.568e+07	2.893e+06	9.36e+07	7.505e+07
RMSE	1701	1.887e+04	2928	2464	3959	1701	9675	8663


Immigrant Stock VS Immigrant Stock VS Spatial Error Model Peru in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Entity Effect
1 features
2 features
3 features
4 features
5 features
6 features
7 features


Immigrant Stock VS Immigrant Stock VS Spatial Error Model Poland in Italian Zones
Centro
Isole
Nord-est
Nord-ovest
Sud
Immigrant Stock
Entity Effect
1 features
2 features
3 features
4 features
5 features
6 features
7 features

Table 4.9. H. Jayet et al. paper Results Poland

Indep. Var	I	II	III	IV	V	VI	VII	VIII
a_{Center}	-0.1554							
$a_{Islands}$	-0.3788							
$a_{NorthEast}$	-0.1935							
$a_{NorthWest}$	-0.2834							
a_{South}	-0.2217							
Free activity for a union		0.0006	0.0018	0.0032	0.0037	0.0289	0.0031	0.0044
Internal Migration - Italy			-0	-0	-0	-0	-0	-0
Recreation, shows and culture				0.0003	0.0003	-0.0029	0.0004	0.0005
Average monthly expenditure for housing					-0.0059	-3.542	0.0391	0.0117
Total fertility rate						32.34	-0.4688	-0.1875
Reach Difficulty - Emergency room							0.0001	-0
Internal Migration - Foreign country								-0
MAE	520.9	1311	839.7	602.1	560.4	520.9	583.4	713.3
MSE	4.86e+05	3.086e+06	8.514e+05	5.867e+05	5.097e+05	4.86e+05	5.949e+05	1.094e+06
RMSE	697.2	1757	922.7	766	713.9	697.2	771.3	1046

Table 4.10. H. Jayet et al. paper Results Romania

Indep. Var	I	II	III	IV	V	VI	VII	VIII
a_{Center}	-0.1097							
$a_{Islands}$	-0.1102							
$a_{NorthEast}$	-0.1368							
$a_{NorthWest}$	-0.1128							
a_{South}	-0.0863							
Free activity for a union		0.0002	0.0005	0.0004	-0.0008	0.0018	0.0005	0.0008
Free activities in voluntary associations			-0	-0.0001	0.0001	-0.0003	0	-0.0001
Meetings in cultural, recreational or other associations				0	-0.0003	0.002	-0.0001	0
Meetings in ecological associations, for civil rights, for peace					0.002	-0.0049	0.0004	0.0002
Free activities in non voluntary associations						-0.0026	0.0001	-0.0001
Political Info - Some times in a year							-0	0
Recreation, shows and culture								0
MAE	7352	6731	6313	6194	6651	7352	6137	7675
MSE	9.999e+07	7.942e+07	6.861e+07	6.79e+07	7.712e+07	9.999e+07	6.67e+07	8.439e+07
RMSE	1e+04	8912	8283	8240	8782	1e+04	8167	9186

Looking at the tables and the figures it is possible to say that decomposing the fixed location effects into observable location factors is not really improving the estimation. It is not helping in terms of interpretation either.

Moreover, in general *Islands,* is the less attractive zone while *North West*, *North East* and *Center* are the more ones. This is true for all the origin nationalities but Romania, where there is not a clear difference between the five zones fixed effect.

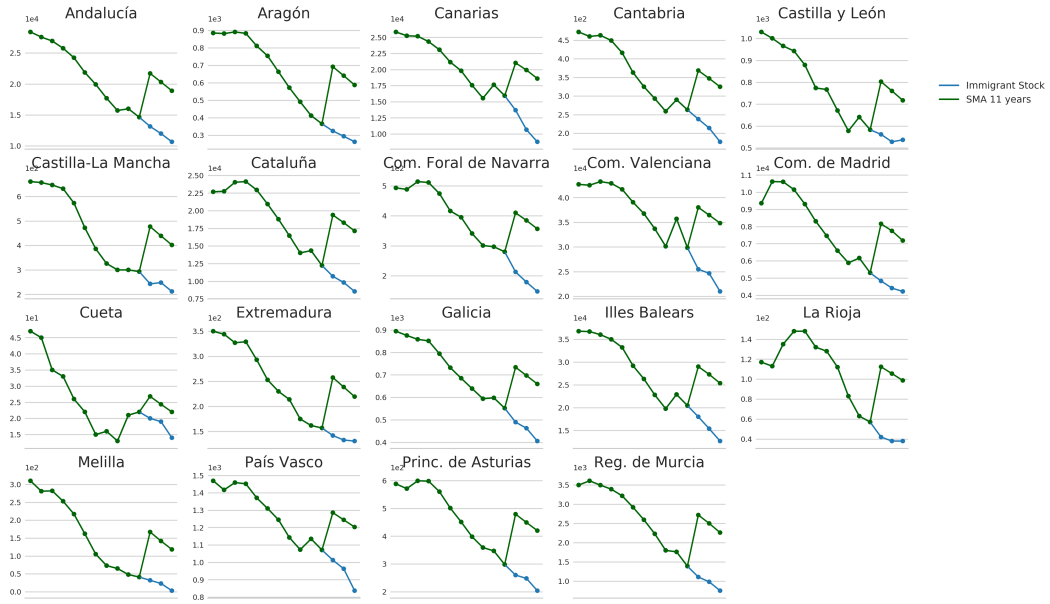
4.2 Spain

For Spain, the period 2003-2016 is studied. Same procedures used for Italy are applied here.

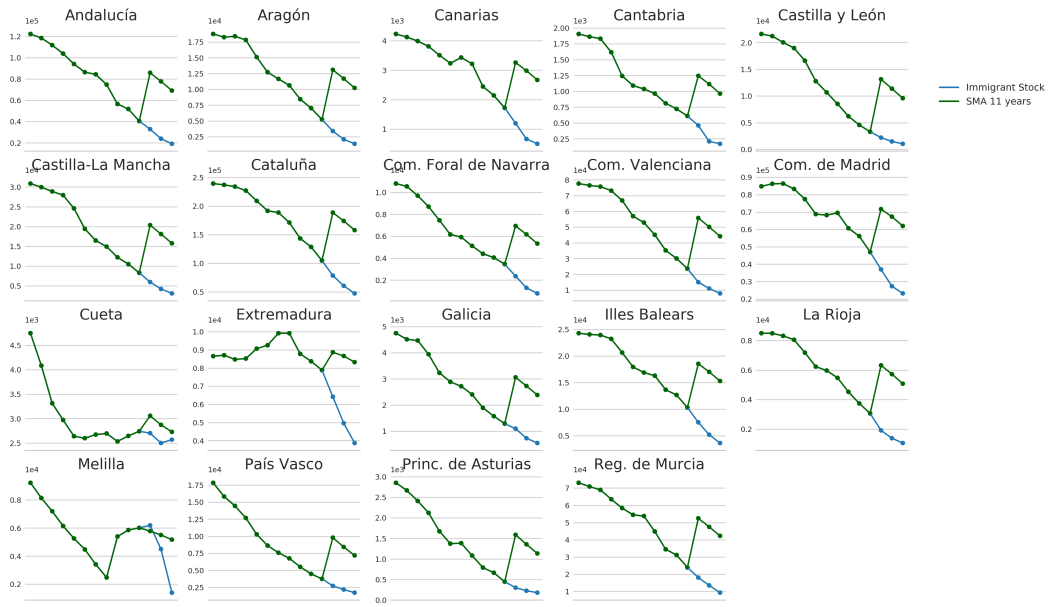
4.2.1 Simple Moving Average

Country	MAE	MSE	RMSE
Germany	3014	2.839e+07	5328
Morocco	1.815e+04	1.042e+09	3.228e+04
Peru	3279	5.646e+07	7514
Poland	2152	1.618e+07	4023
Romania	2.205e+04	1.437e+09	3.79e+04

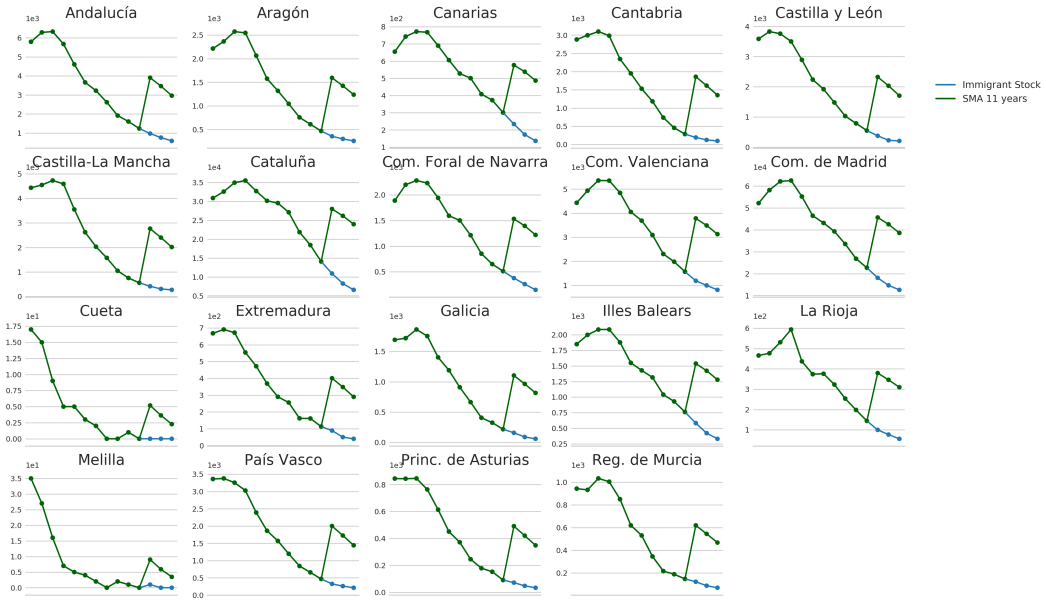
Immigrant Stock VS SMA in Germany in Spanish Autonomous Communities - years 2003 - 2016



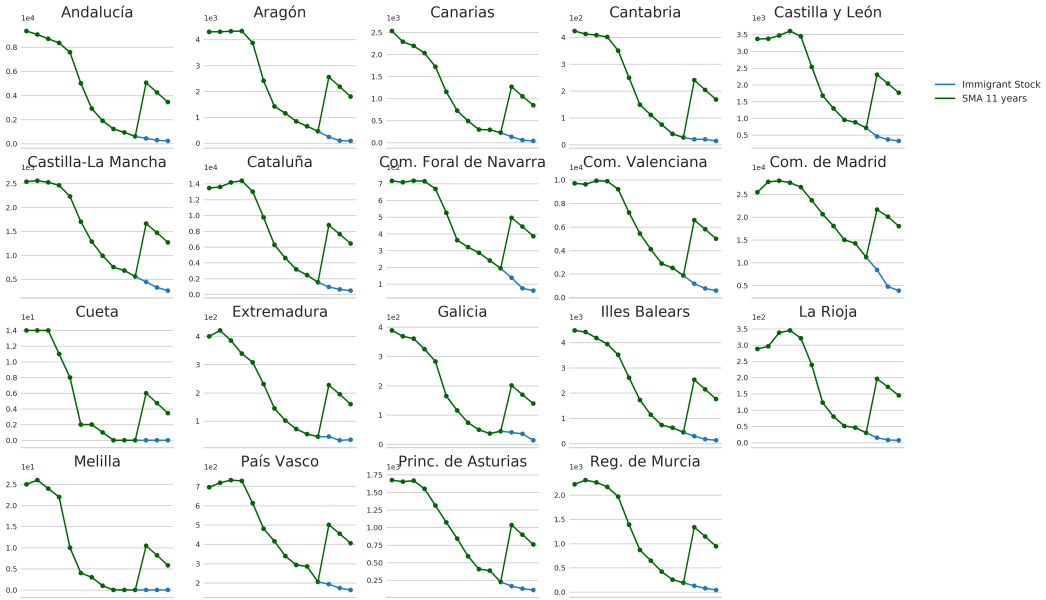
Immigrant Stock VS SMA in Morocco in Spanish Autonomous Communities - years 2003 - 2016

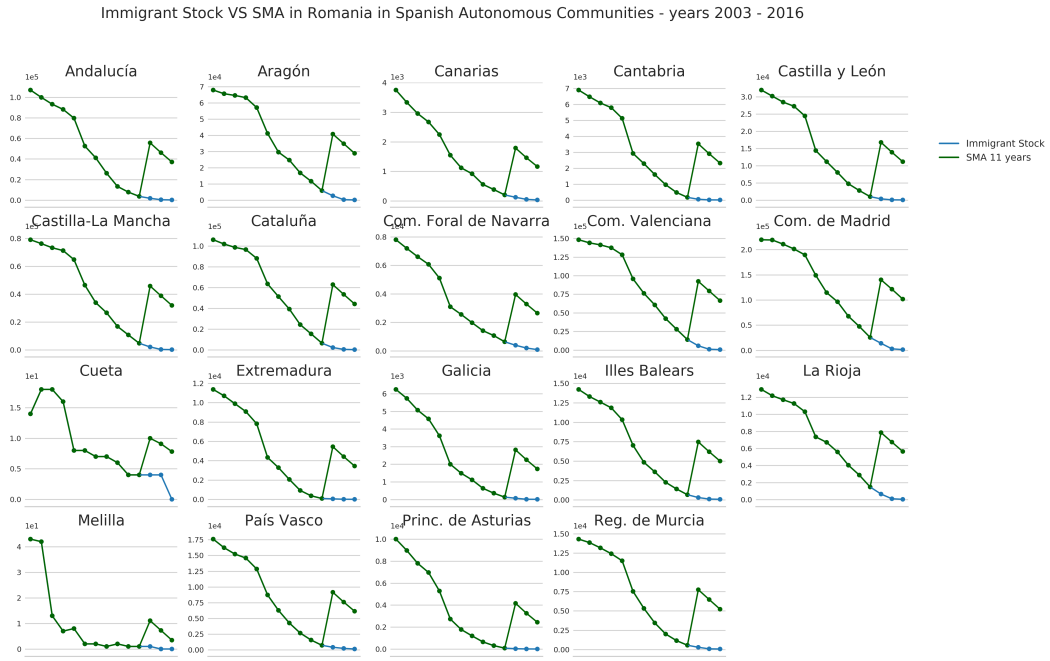


Immigrant Stock VS SMA in Peru in Spanish Autonomous Communities - years 2003 - 2016



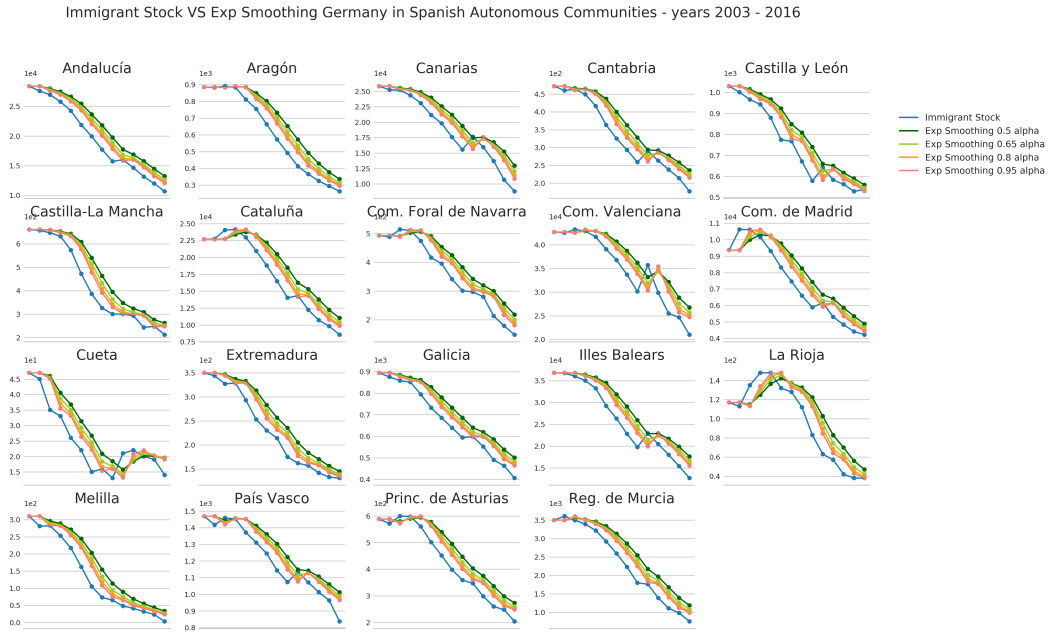
Immigrant Stock VS SMA in Poland in Spanish Autonomous Communities - years 2003 - 2016



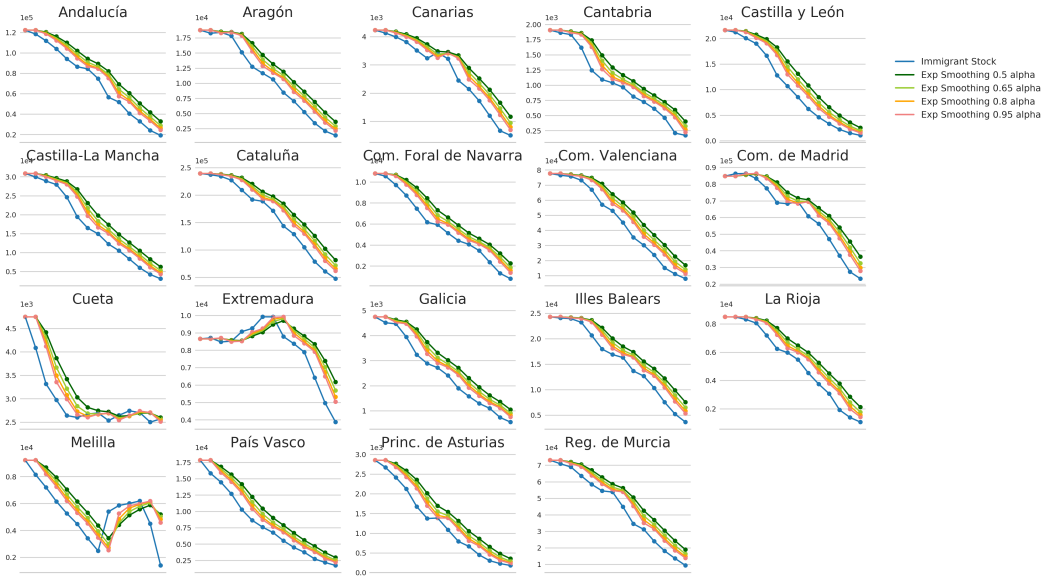


Also for the Spanish case, SMA is not able to forecast the immigrant stock values over the autonomous communities.

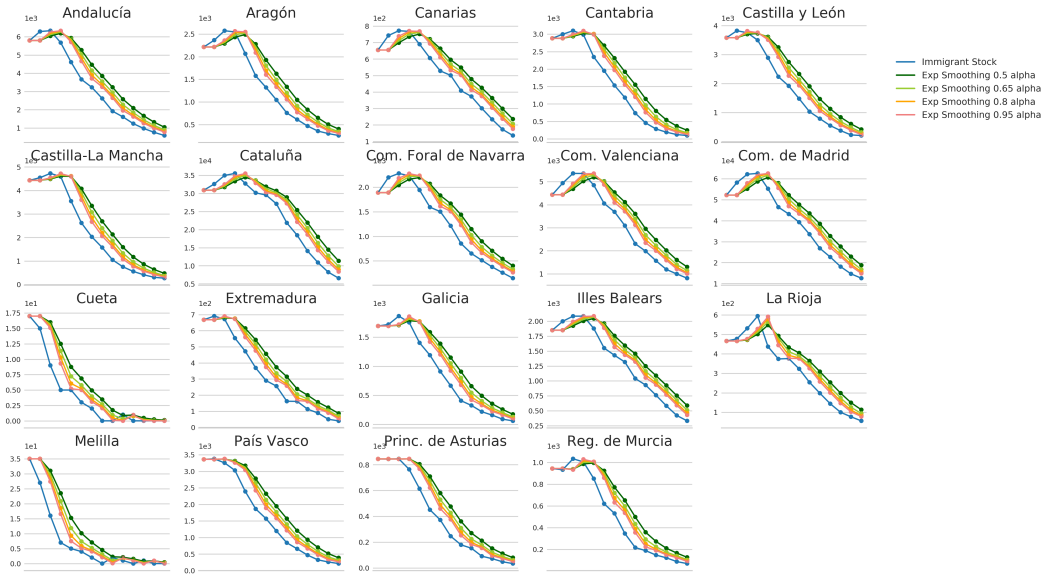
4.2.2 Exponential Smoothig

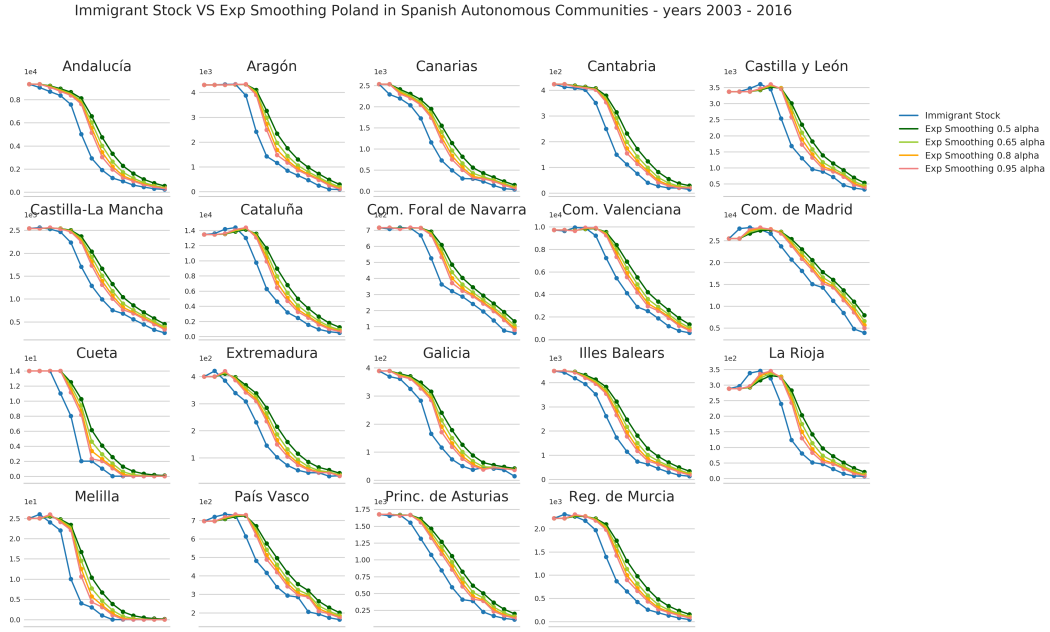


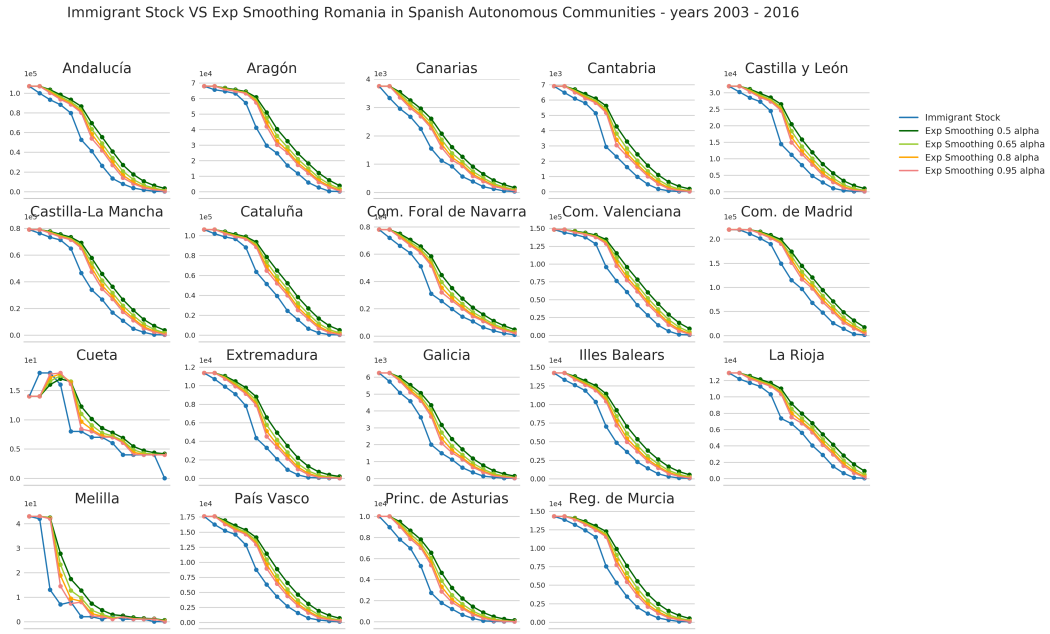
Immigrant Stock VS Exp Smoothing Morocco in Spanish Autonomous Communities - years 2003 - 2016



Immigrant Stock VS Exp Smoothing Peru in Spanish Autonomous Communities - years 2003 - 2016




Immigrant Stock VS Exp Smoothing Poland in Spanish Autonomous Communities - years 2003 - 2016
Andalucía
Aragón
Canarias
Cantabria
Castilla y León
Castilla-La Mancha
Cataluña
Com. Foral de Navarra
Com. Valenciana
Com. de Madrid
Cueta
Extremadura
Galicia
Illes Balears
La Rioja
Melilla
País Vasco
Princ. de Asturias
Reg. de Murcia
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha


Immigrant Stock VS Exp Smoothing Romania in Spanish Autonomous Communities - years 2003 - 2016
Andalucía
Aragón
Canarias
Cantabria
Castilla y León
Castilla-La Mancha
Cataluña
Com. Foral de Navarra
Com. Valenciana
Com. de Madrid
Cueta
Extremadura
Galicia
Illes Balears
La Rioja
Melilla
País Vasco
Princ. de Asturias
Reg. de Murcia
Immigrant Stock
Exp Smoothing 0.5 alpha
Exp Smoothing 0.65 alpha
Exp Smoothing 0.8 alpha
Exp Smoothing 0.95 alpha

Metric	α	Germany	Morocco	Peru	Poland	Romania
MAE	0.5	840	4.83e+03	962	633	6.8e+03
	0.65	716	3.88e+03	788	505	5.32e+03
	0.8	632	3.24e+03	662	419	4.35e+03
	0.95	574	2.78e+03	568	357	3.67e+03
MSE	0.5	2.93e+06	8.29e+07	5.32e+06	1.69e+06	1.67e+08
	0.65	1.66e+06	3.78e+07	2.65e+06	8.11e+05	7.48e+07
	0.8	1.659e+06	3.78e+07	2.65e+06	8.11e+05	7.48e+07
	0.95	1.41e+06	2.84e+07	2.04e+06	6.15e+05	5.59e+07
RMSE	0.5	1.71e+03	9.11e+03	2.31e+03	1.3e+03	1.29e+04
	0.65	1.45e+03	7.33e+03	1.9e+03	1.06e+03	1.03e+04
	0.8	1.29e+03	6.15e+03	1.63e+03	901	8.65e+03
	0.95	1.19e+03	5.33e+03	1.43e+03	784	7.47e+03

Similar results to the ones obtained for Italy are achieved for Spain: ES performs much better than SMA and the value at the previous time $(t-1)$ is more important than the ones before it $(1, \ldots, t-2)$.

4.2.3 Regression Model for Panel Data

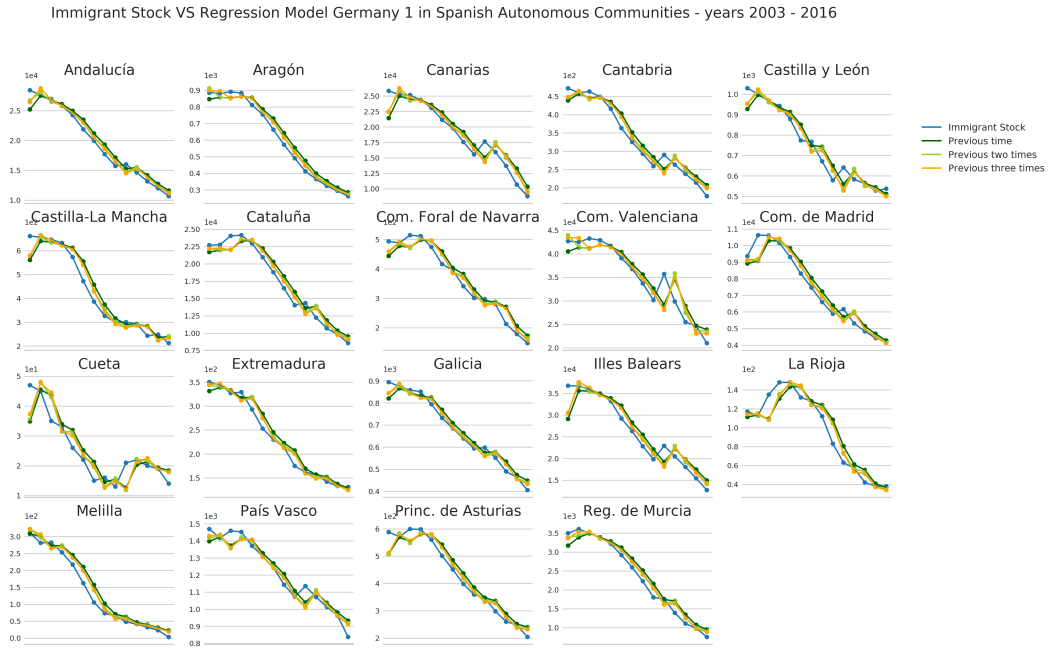
Here, the folling models are implemented:

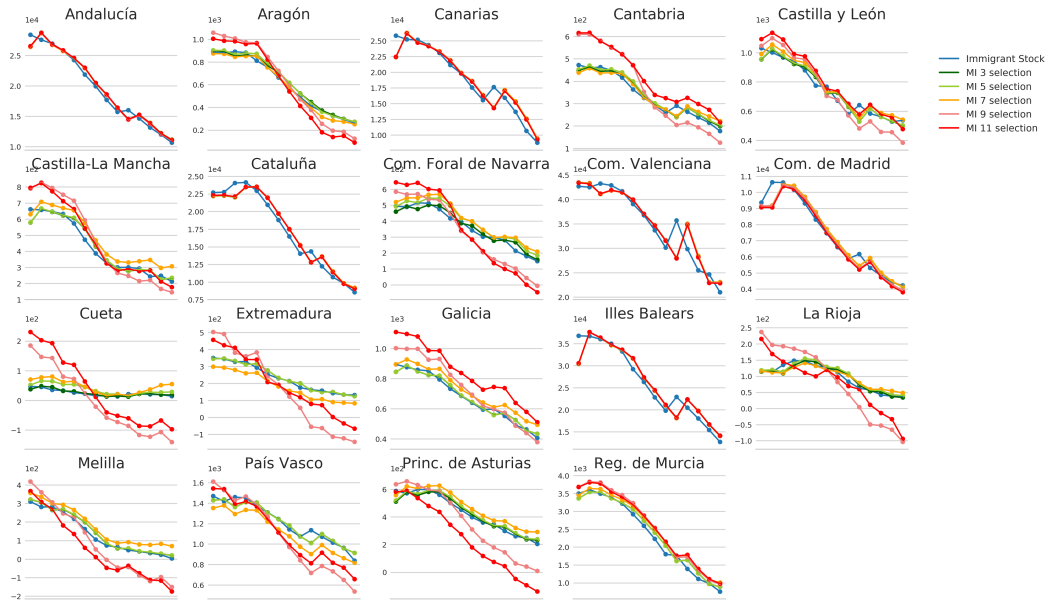
- only the previous time stock
- the previous two times stock
- the previous three times stock
- a set of 3, 5, 7, 9, 11 variables selected through the Mutual information criterion.

For each origin country, the time period 2003-2013 is used for training and 2014-2016 for testing.

Table 4.11. Regression Results Germany

Indep. Var	I	II	III	IV	V	VI	VII	VIII
y_{t-1}	0.968	1.2	1.15	1.15	1.15	1.15	1.14	1.14
y_{t-2}		-0.237	-0.0536	-0.0536	-0.0539	-0.053	-0.0613	-0.0601
y_{t-3}			-0.132	-0.132	-0.132	-0.133	-0.118	-0.118
Native Population					-1.34e-05	-3.73e-05	-3.28e-05	-0.000154
Employed					0.04	0.0953	0.0836	0.382
Mortality Rate						14.5	61.5	60.1
Birth Rate						-10.3	209	169
Fertility Rate							-53.9	-45.2
Activity Rate							-7.38	-0.273
Per capita income								-0.0189
Unemployment Rate								3.53
MAE	448	323	338	338	339	370	399	411
MSE	8.75e+05	4.63e+05	5.4e+05	5.4e+05	5.39e+05	5.57e+05	5e+05	5.05e+05
RMSE	935	680	735	735	734	746	707	710

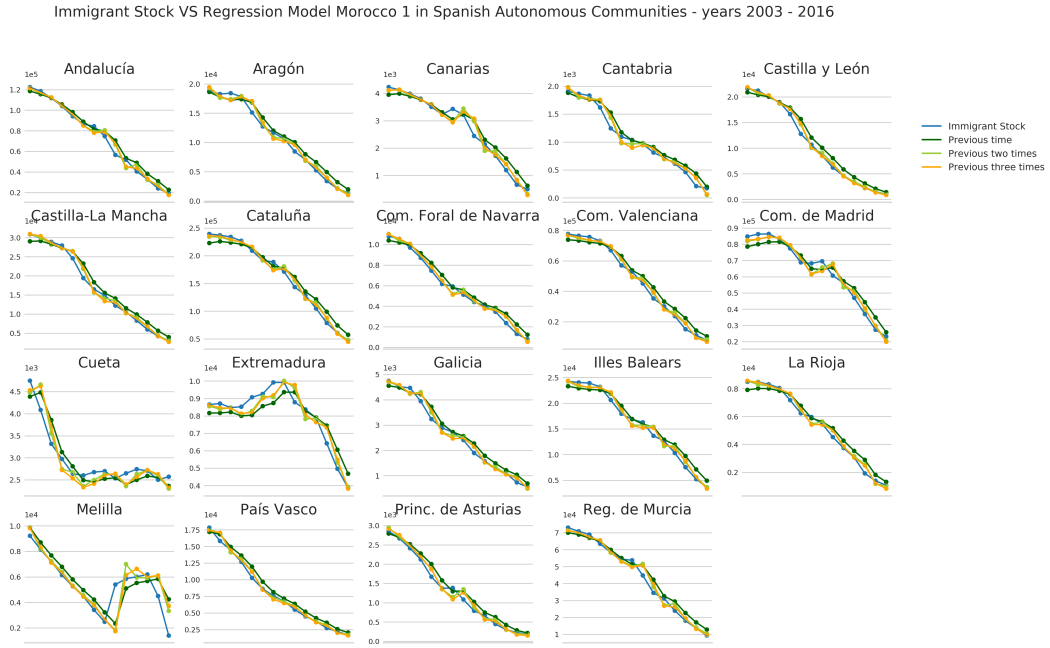


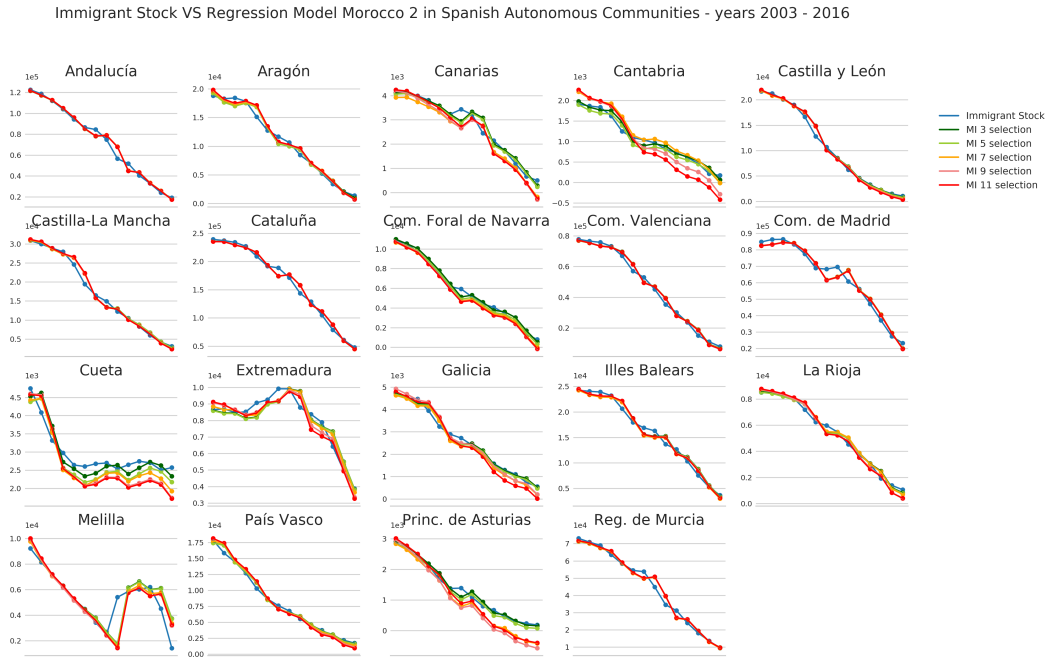


It seems that all the flow of Germany-born immigrants could be explained by looking at the flows of the previous two years. Considering only the variable y_{t-1} is not enough, i.e. there a remarkable difference, in terms of error, between the first model and all the others.

An interesting result is that both y_{t-2} and y_{t-3} have a slightly negative effect in every model they are considered in, while the coefficient of y_{t-1} is always positive.

Other variables such as *Mortality Rate*, *Birth Rate* and *Fertility Rate* have very high coefficients, their presence in the models does not improve the performances.





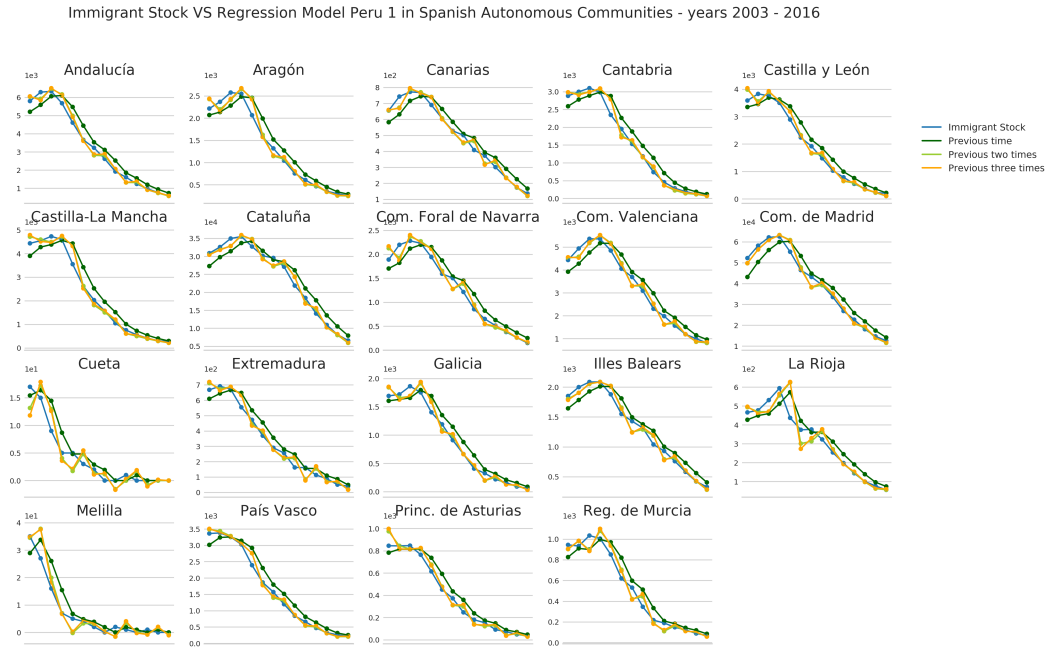
Similiar observations can be done for Morocco.

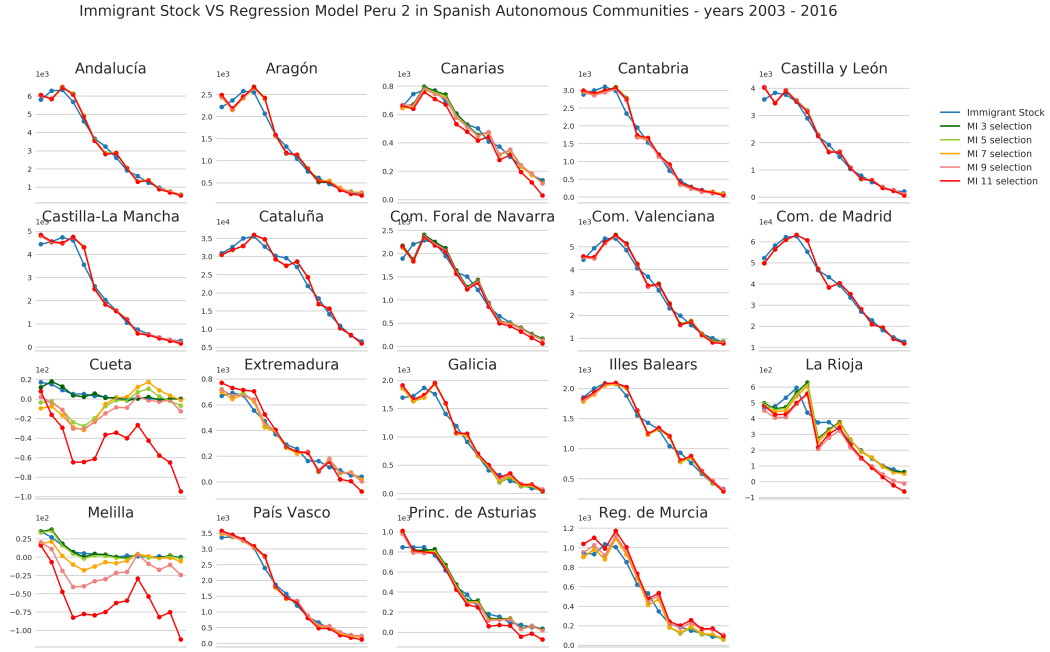
Table 4.12. Regression Results Morocco

Indep. Var	I	II	III	IV	V	VI	VII	VIII
y_{t-1}	0.944	1.57	1.37	1.37	1.37	1.36	1.36	1.36
y_{t-2}		-0.605	-0.135	-0.135	-0.135	-0.136	-0.139	-0.141
y_{t-3}			-0.266	-0.266	-0.263	-0.258	-0.252	-0.249
Native Population					-2.87e-05	0.000144	0.000103	0.000288
Birth Rate					-2.37	286	481	471
Fertility Rate						-72.4	-115	-111
Employed						-0.437	-0.315	-0.765
Mortality Rate							103	91.4
Activity Rate							-19.4	-20.8
Per capita income								0.0137
Unemployment Rate								-11.9
MAE	2.31e+03	769	869	869	879	946	980	992
MSE	1.84e+07	2.42e+06	2.9e+06	2.9e+06	2.91e+06	2.88e+06	2.89e+06	2.88e+06
RMSE	4.29e+03	1.55e+03	1.7e+03	1.7e+03	1.7e+03	1.7e+03	1.7e+03	1.7e+03

Table 4.13. Regression Results Peru

Indep. Var	I	II	III	IV	V	VI	VII	VIII
y_{t-1}	0.964	1.75	1.88	1.88	1.88	1.88	1.88	1.88
y_{t-2}		-0.777	-1.06	-1.06	-1.06	-1.05	-1.05	-1.06
y_{t-3}			0.147	0.147	0.147	0.146	0.146	0.149
Native Population					2.63e-05	1.54e-05	-7.89e-06	-1.74e-05
Employed					-0.071	-0.0422	0.0222	0.0368
Birth Rate						-2.22	4.94	109
Unemployment Rate						1.1	2.69	6.84
Mortality Rate							13.4	31
Activity Rate							-3.88	-8.76
Fertility Rate								-27.1
Per capita income								0.00573
MAE	314	87.6	85.8	85.8	86.5	85.8	95.1	130
MSE	6.84e+05	5.32e+04	5.64e+04	5.64e+04	5.65e+04	5.63e+04	5.62e+04	6.26e+04
RMSE	827	231	237	237	238	237	237	250

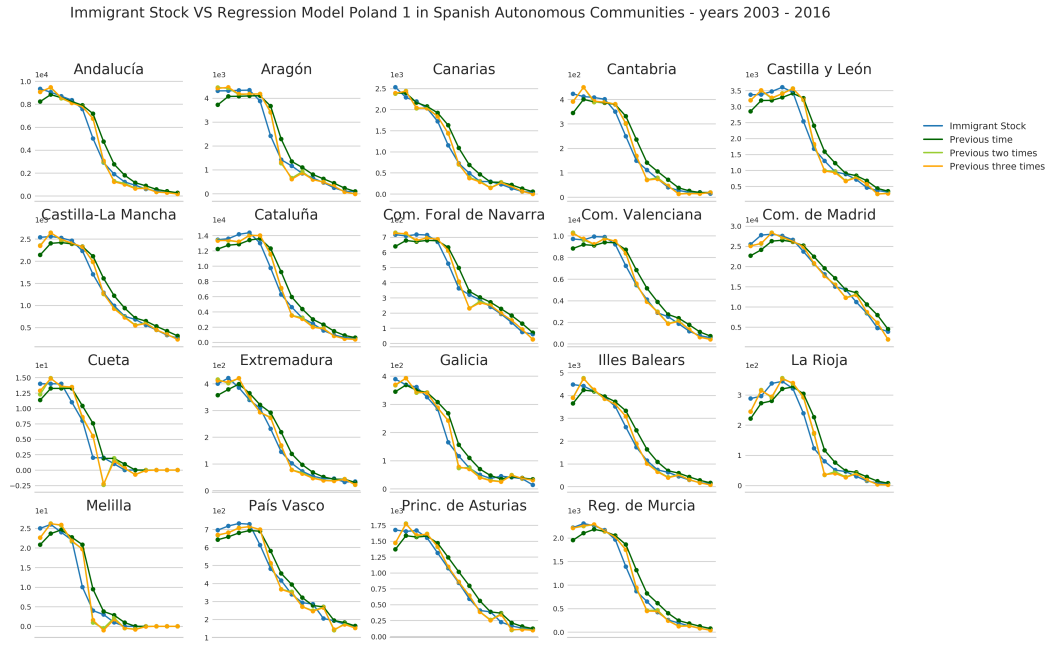


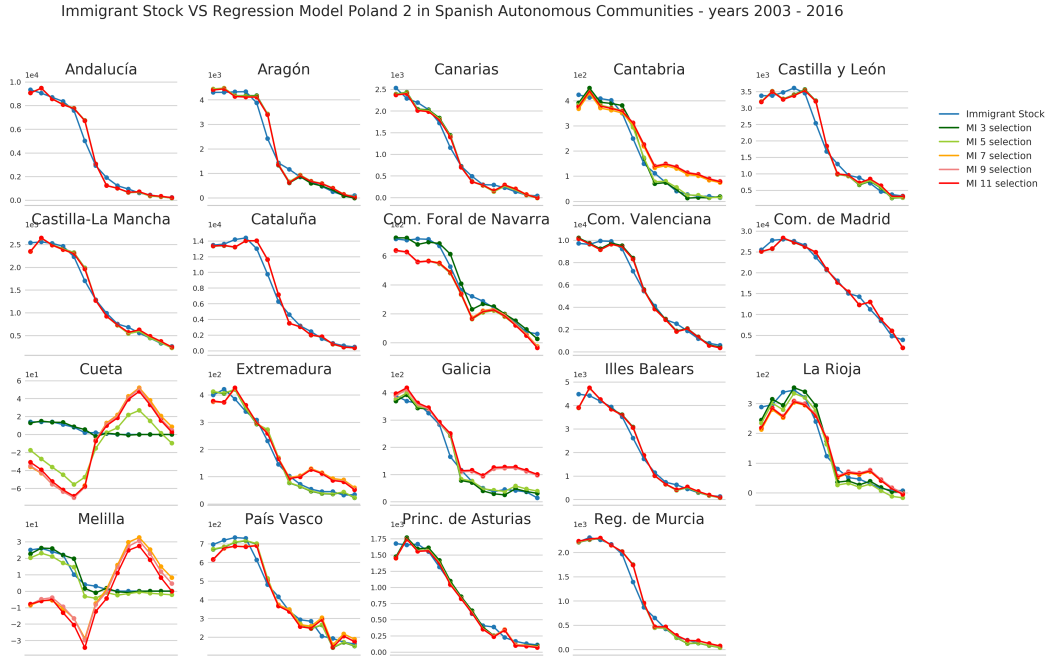


Also for Peru, the same results hold true. The only significant difference is that as before y_{t-2} coefficients are negatives whereas y_{t-3} values are slightly positives.

Table 4.14. Regression Results Poland

Indep. Var	I	II	III	IV	V	VI	VII	VIII
y_{t-1}	0.946	1.69	1.62	1.62	1.62	1.61	1.61	1.61
y_{t-2}		-0.723	-0.582	-0.582	-0.573	-0.558	-0.558	-0.559
y_{t-3}			-0.0762	-0.0762	-0.0798	-0.0917	-0.091	-0.09
Native Population					6.32e-05	2.06e-05	2.13e-05	1.68e-05
Employed					-0.167	-0.0694	-0.0726	-0.0612
Birth Rate						-5.41	0.157	1.89
Unemployment Rate						4.21	4.21	4.28
Fertility Rate							-1.71	-2.11
Activity Rate							0.263	0.265
Mortality Rate								1.45
Per capita income								-0.000745
MAE	175	93.5	96.6	96.6	103	114	114	115
MSE	2.89e+05	1e+05	9.74e+04	9.74e+04	9.85e+04	9.97e+04	9.99e+04	1e+05
RMSE	538	316	312	312	314	316	316	316



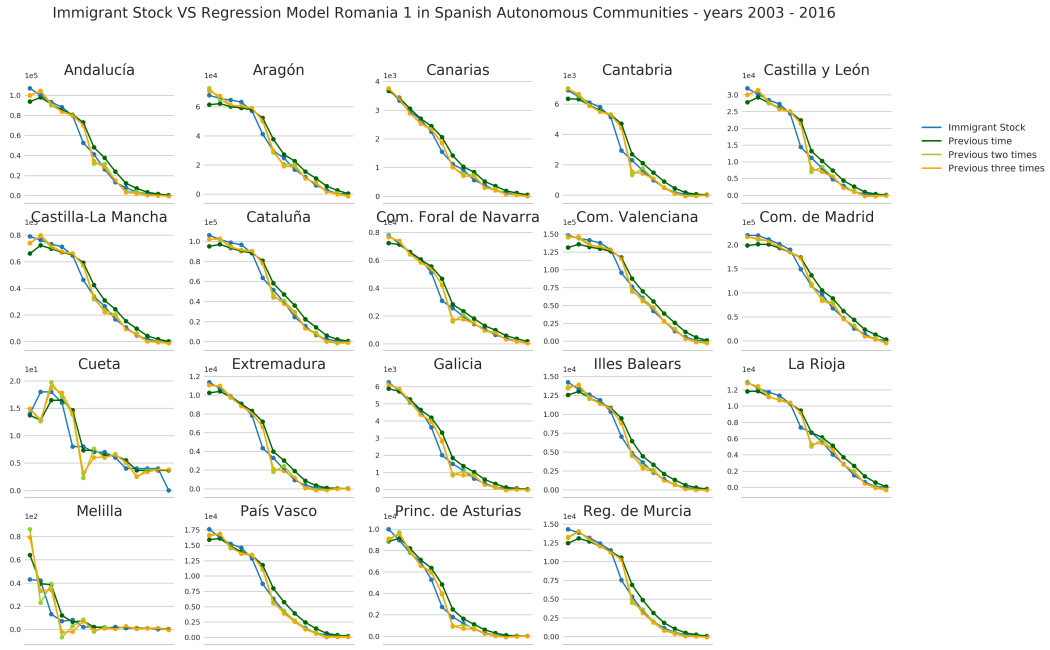


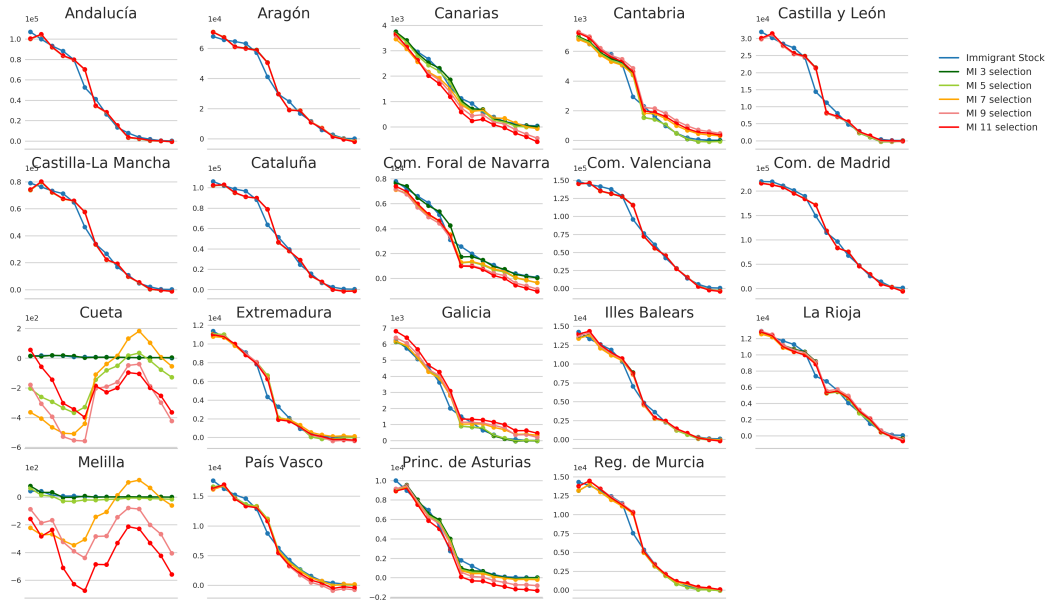
As for Peru, even for Poland, the difference in terms of coefficients between y_{t-1} and y_{t-2} and y_{t-3} is quite important.

Here is also evident the difference between the first model including just y_{t-1} and all the others.

Table 4.15. Regression Results Romania

Indep. Var	I	II	III	IV	V	VI	VII	VIII
y_{t-1}	0.915	1.61	1.45	1.45	1.45	1.45	1.45	1.44
y_{t-2}		-0.664	-0.307	-0.307	-0.305	-0.302	-0.304	-0.305
y_{t-3}			-0.197	-0.197	-0.196	-0.2	-0.194	-0.189
Native Population					0.000273	-7.29e-06	-5.19e-05	-0.000331
Employed					-0.769	-0.107	-0.0765	0.664
Fertility Rate						-11.6	-149	-98.1
Unemployment Rate						30.2	34.7	28.5
Mortality Rate							83.9	104
Birth Rate							486	310
Per capita income								-0.0723
Activity Rate								18.4
MAE	948	574	714	714	759	742	982	1.04e+03
MSE	5.19e+06	1.51e+06	2.03e+06	2.03e+06	2.14e+06	2.13e+06	2.57e+06	2.87e+06
RMSE	2.28e+03	1.23e+03	1.43e+03	1.43e+03	1.46e+03	1.46e+03	1.6e+03	1.69e+03





Same observations done on Poland can be made for Romania.

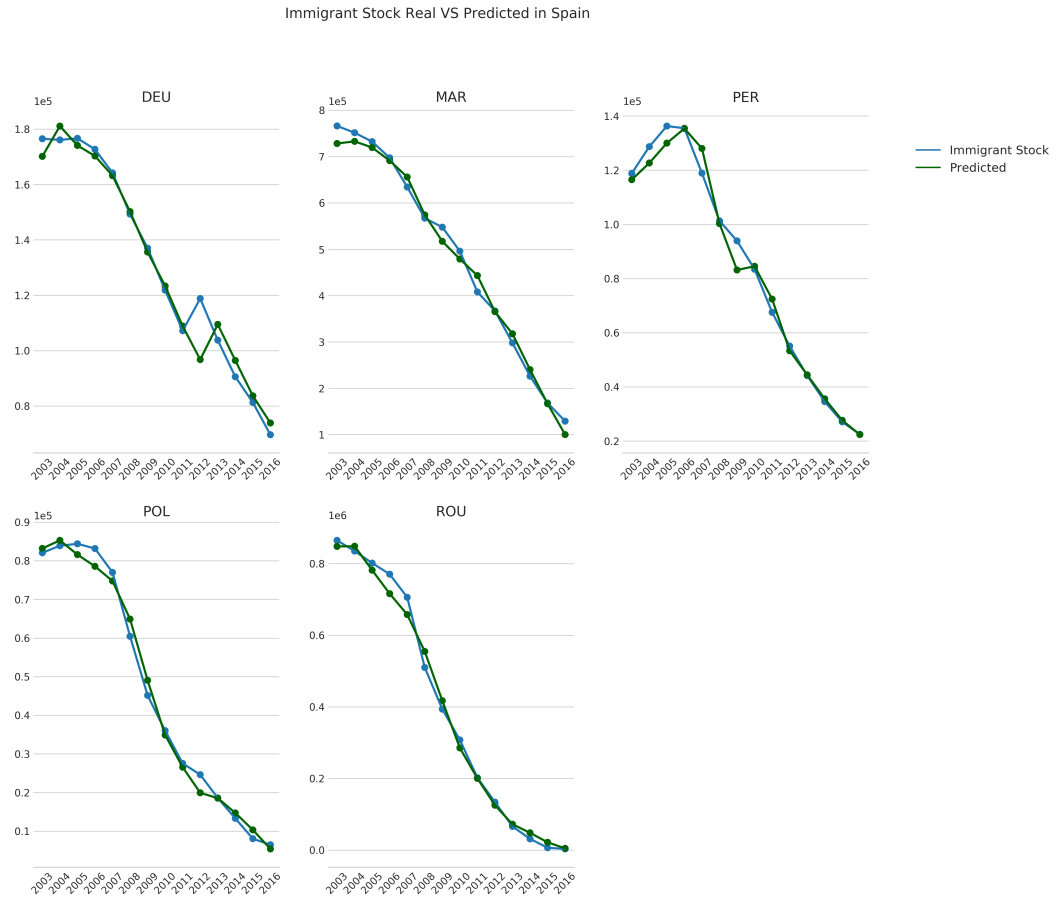
4.2.4 Spatial Error Model

The same process followed for Italy is here used for Spain.

Thus, as first the spatial weights matrix is needed. Since it is not possible to access this information from the INE website, the matrix is manually built using GoogleMaps.

2013 is again used as refered period for the *time invariant independent variables*.

Follows the plot of the estimated Spanish overall flow.

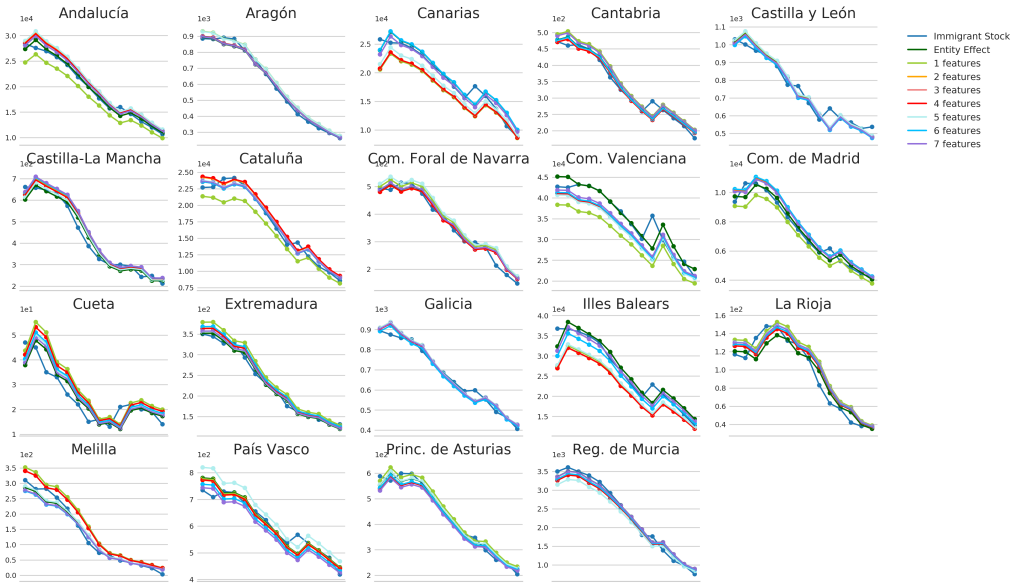


Again, the model is now trained using from 1 to 7 independent variables and the Spain as reference territory.

Country	β	ρ
Germany	0.956700	-0.258700
Morocco	0.970700	-0.138900
Peru	0.850600	-0.642600
Poland	0.888700	-0.377100
Romania	0.888600	-0.287800

As before from the first stage it is possible to get the estimated network effect and spatial autoregressive coefficient. The β values are very similar: positive and fairly high (from 0.97 of Morocco to 0.85 of Peru) for all the five origin countries. The ρ parameter instead is always negative.

Immigrant Stock VS Immigrant Stock VS Spatial Error Model Germany in Spanish Autonomous Communities - years 2003 - 2016



Immigrant Stock VS Immigrant Stock VS Spatial Error Model Morocco in Spanish Autonomous Communities - years 2003 - 2016

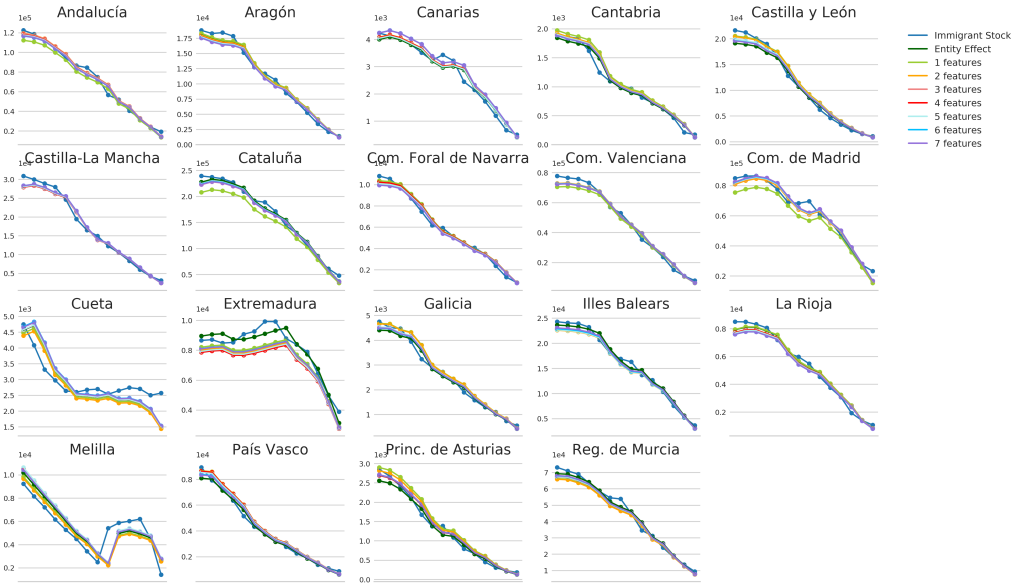


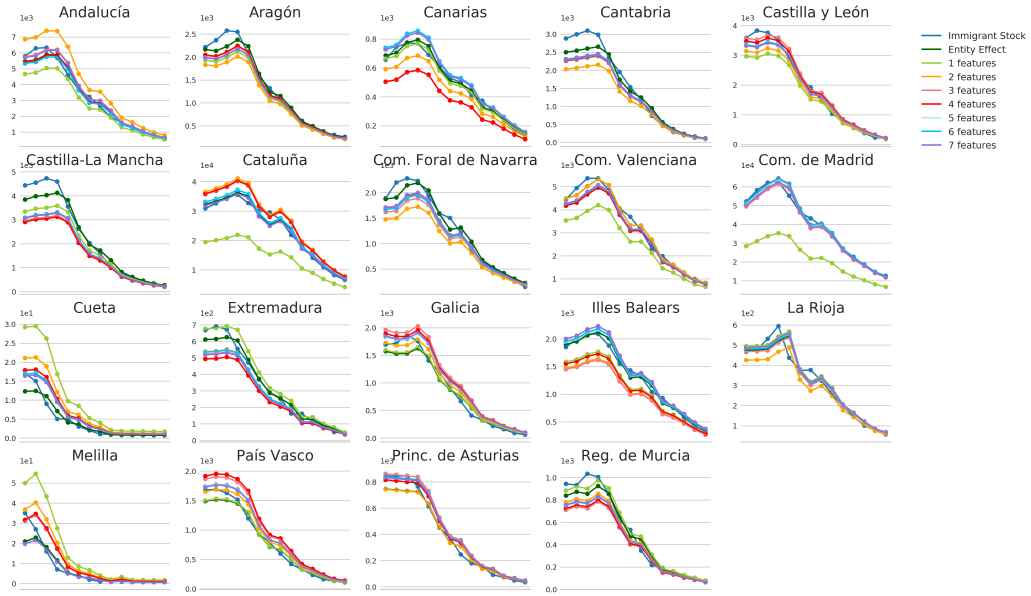
Table 4.16. H. Jayet et al. paper Results Germany

Indep. Var	I	II	III	IV	V	VI	VII	VIII
$a_{Andalucía}$	-0.0829							
$a_{Aragón}$	-0.2589							
$a_{Canarias}$	-0.0623							
$a_{Cantabria}$	-0.2567							
$a_{CastillayLeón}$	-0.2186							
$a_{Castilla-LaMancha}$	-0.2587							
$a_{Cataluña}$	-0.0903							
$a_{Com.ForaldeNavarra}$	-0.2475							
$a_{Com.Valenciana}$	-0.0371							
$a_{Com.deMadrid}$	-0.1232							
a_{Cueta}	-0.3683							
$a_{Extremadura}$	-0.2933							
$a_{Galicia}$	-0.2163							
$a_{IllesBalears}$	-0.0536							
$a_{LaRioja}$	-0.3227							
$a_{Melilla}$	-0.4308							
$a_{PaísVasco}$	-0.2104							
$a_{Princ.deAsturias}$	-0.2472							
$a_{Reg.deMurcia}$	-0.1982							
Native Population		0	0	0	0	0	0	0
Mortality Rate			-0.007	-0.007	-0.0071	-0.0098	-0.0203	-0.0187
Employed				0	0	-0.0002	-0.0001	-0.0001
Per capita income					0	0	0	-0
Fertility Rate						-0.0041	-0.0081	-0.0073
Activity Rate							0.0077	0.0139
Unemployment Rate								-0.0081
MAE	326.5	303.9	260.2	258.3	258.1	240.1	247.9	265.8
MSE	5.197e+05	5.186e+05	3.29e+05	3.238e+05	3.236e+05	3.056e+05	3.392e+05	3e+05
RMSE	720.9	720.1	573.5	569.1	568.8	552.8	582.4	547.7

Table 4.17. H. Jayet et al. paper Results Morocco

Indep. Var	I	II	III	IV	V	VI	VII	VIII
$a_{Andalucía}$	-0.0799							
$a_{Aragón}$	-0.153							
$a_{Canarias}$	-0.1539							
$a_{Cantabria}$	-0.2088							
$a_{CastillayLeón}$	-0.2063							
$a_{Castilla-LaMancha}$	-0.1389							
$a_{Cataluña}$	-0.0279							
$a_{Com.ForaldeNavarra}$	-0.1563							
$a_{Com.Valenciana}$	-0.0987							
$a_{Com.deMadrid}$	-0.0346							
a_{Cueta}	-0.1233							
$a_{Extremadura}$	-0.0531							
$a_{Galicia}$	-0.1948							
$a_{IllesBalears}$	-0.1015							
$a_{LaRioja}$	-0.1452							
$a_{Melilla}$	-0.1079							
$a_{PaísVasco}$	-0.2054							
$a_{Princ.deAsturias}$	-0.2665							
$a_{Reg.deMurcia}$	-0.0943							
Employed		0	0	0	0.0001	0.0001	0.0001	0.0001
Mortality Rate			-0.0036	-0.0115	-0.0114	-0.0089	-0.0129	-0.0097
Fertility Rate				0.0018	0.0018	0.0021	0.0005	-0.0036
Native Population					-0	-0	-0	-0
Per capita income						-0	-0	-0
Unemployment Rate							0.0032	0.0023
Birth Rate								0.0178
MAE	994.5	1132	987.3	992.6	999.5	995.6	1006	1009
MSE	4.798e+06	6.826e+06	4.854e+06	4.84e+06	4.833e+06	4.812e+06	4.85e+06	4.861e+06
RMSE	2190	2613	2203	2200	2198	2194	2202	2205

Immigrant Stock VS Immigrant Stock VS Spatial Error Model Peru in Spanish Autonomous Communities - years 2003 - 2016



Immigrant Stock VS Immigrant Stock VS Spatial Error Model Poland in Spanish Autonomous Communities - years 2003 - 2016

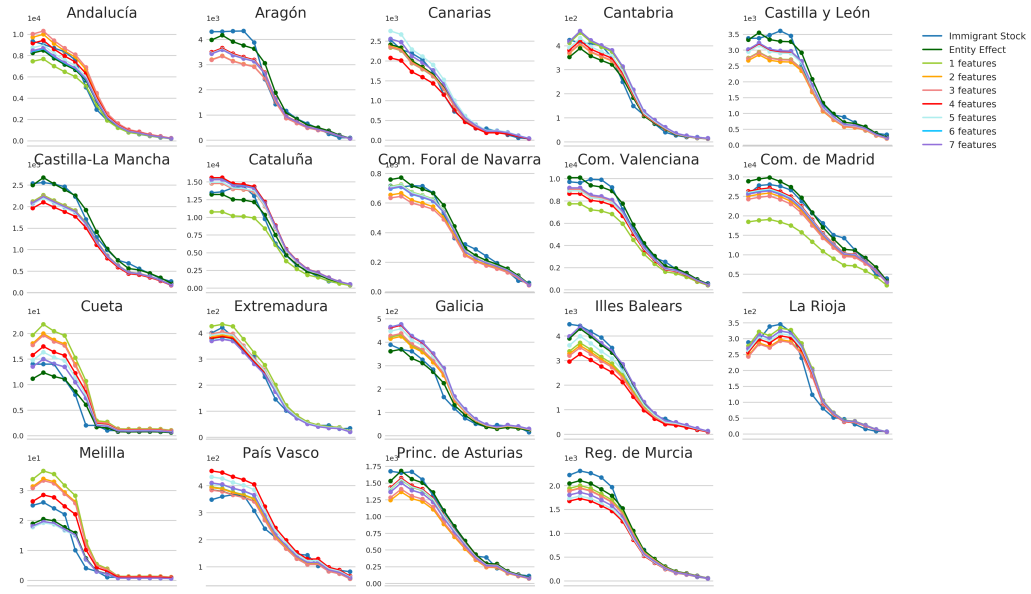
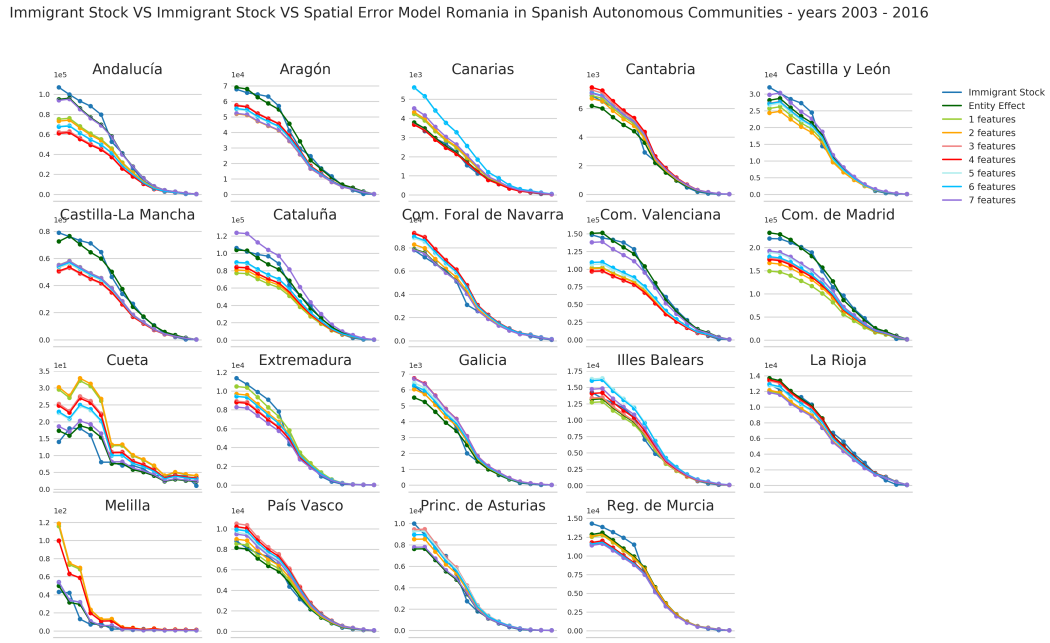


Table 4.18. H. Jayet et al. paper Results Peru

Indep. Var	I	II	III	IV	V	VI	VII	VIII
$a_{Andalucía}$	-0.5377							
$a_{Aragón}$	-0.6771							
$a_{Canarias}$	-0.7492							
$a_{Cantabria}$	-0.7242							
$a_{CastillayLeón}$	-0.6503							
$a_{Castilla-LaMancha}$	-0.6426							
$a_{Cataluña}$	-0.1683							
$a_{Com.ForaldeNavarra}$	-0.6547							
$a_{Com.Valenciana}$	-0.4928							
$a_{Com.deMadrid}$	-0.0958							
a_{Cueta}	-1.68							
$a_{Extremadura}$	-0.9019							
$a_{Galicia}$	-0.7813							
$a_{IllesBalears}$	-0.6126							
$a_{LaRioja}$	-0.8498							
$a_{Melilla}$	-1.685							
$a_{PaísVasco}$	-0.7823							
$a_{Princ.deAsturias}$	-0.9178							
$a_{Reg.deMurcia}$	-0.8447							
Employed		0	0.0003	0.0003	0.0003	0.0002	0.0002	0
Unemployment Rate			-0.0101	-0.0212	-0.0256	-0.02	-0.0208	-0.0255
Mortality Rate				0.0356	0.0234	-0.0073	-0.0007	0.0038
Activity Rate					0.0038	0.02	0.0192	0.0203
Fertility Rate						-0.0195	-0.0271	-0.0358
Birth Rate							0.0331	0.0723
Native Population								0
MAE	60.38	550.8	136.4	136.4	119.8	75.23	72.61	64.8
MSE	1.907e+04	2.999e+06	1.592e+05	1.341e+05	1.264e+05	3.456e+04	3.258e+04	2.039e+04
RMSE	138.1	1732	399.1	366.2	355.5	185.9	180.5	142.8

Table 4.19. H. Jayet et al. paper Results Poland

Indep. Var	I	II	III	IV	V	VI	VII	VIII
$a_{Andalucía}$	-0.3812							
$a_{Aragón}$	-0.4036							
$a_{Canarias}$	-0.5098							
$a_{Cantabria}$	-0.712							
$a_{CastillayLeón}$	-0.3461							
$a_{Castilla-LaMancha}$	-0.3771							
$a_{Cataluña}$	-0.2614							
$a_{Com.ForaldeNavarra}$	-0.4953							
$a_{Com.Valenciana}$	-0.2419							
$a_{Com.deMadrid}$	-0.0282							
a_{Cueta}	-1.135							
$a_{Extremadura}$	-0.6716							
$a_{Galicia}$	-0.6888							
$a_{IllesBalears}$	-0.4067							
$a_{LaRioja}$	-0.6865							
$a_{Melilla}$	-1.145							
$a_{PaísVasco}$	-0.565							
$a_{Princ.deAsturias}$	-0.4723							
$a_{Reg.deMurcia}$	-0.4979							
Employed		0	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001
Mortality Rate			-0.0157	-0.0073	0.0156	-0.022	-0.0146	-0.015
Per capita income				-0	0	0	-0	-0
Activity Rate					-0.0099	0.0087	0.0228	0.0229
Birth Rate						-0.0666	-0.0582	-0.0581
Unemployment Rate							-0.018	-0.0183
Native Population								0
MAE	94.97	109.8	83.05	86.28	79.01	79.89	83.35	83.69
MSE	8.779e+04	1.756e+05	4.757e+04	5.003e+04	5.273e+04	4.969e+04	4.95e+04	4.91e+04
RMSE	296.3	419.1	218.1	223.7	229.6	222.9	222.5	221.6



Same results obtained for Italy hold true also for Spain: most of the dependent variable can be explained by considering just the *network effect* and the *location fixed effects*. The observable location factors obtained from the second stage of the method are almost everywhere around 0 and they are not improving the results.

Andalucía, *Cataluña* and *Comunidad de Madrid* can be considered the most attractive autonomous commmunities for the immigrant nationalities considered, while *Cueta* and *Melilla* the less ones.

Table 4.20. H. Jayet et al. paper Results Romania

Indep. Var	I	II	III	IV	V	VI	VII	VIII
$a_{Andalucía}$	-0.3285							
$a_{Aragón}$	-0.2708							
$a_{Canarias}$	-0.6717							
$a_{Cantabria}$	-0.6686							
$a_{CastillayLeón}$	-0.4655							
$a_{Castilla-LaMancha}$	-0.2878							
$a_{Cataluña}$	-0.2524							
$a_{Com.ForaldeNavarra}$	-0.5411							
$a_{Com.Valenciana}$	-0.1669							
$a_{Com.deMadrid}$	-0.1019							
a_{Cueta}	-1.094							
$a_{Extremadura}$	-0.7369							
$a_{Galicia}$	-0.7143							
$a_{IllesBalears}$	-0.5187							
$a_{LaRioja}$	-0.424							
$a_{Melilla}$	-1.408							
$a_{PaísVasco}$	-0.5921							
$a_{Princ.deAsturias}$	-0.7554							
$a_{Reg.deMurcia}$	-0.5354							
Per capita income		0	0	0	0	0	0	-0
Mortality Rate			-0.0203	0.0058	0.005	-0.0459	-0.0736	-0.0919
Unemployment Rate				-0.012	-0.0182	-0.0045	-0.0033	-0.0249
Activity Rate					0.0063	0.02	0.0253	0.0443
Birth Rate						-0.0953	-0.2372	-0.4444
Fertility Rate							0.0327	0.0881
Native Population								0
MAE	599.6	290.4	279.7	289	284.7	346.4	332.7	461.4
MSE	2.043e+06	4.182e+05	4.523e+05	5.16e+05	4.98e+05	6.936e+05	6.244e+05	1.181e+06
RMSE	1429	646.7	672.6	718.4	705.7	832.8	790.2	1087

Chapter 5

Conclusion

To better understand the behavior of the methods used and to compare them, let's see the prediction error values within each selected nationalities over the different models adopted.

5.1 Italy

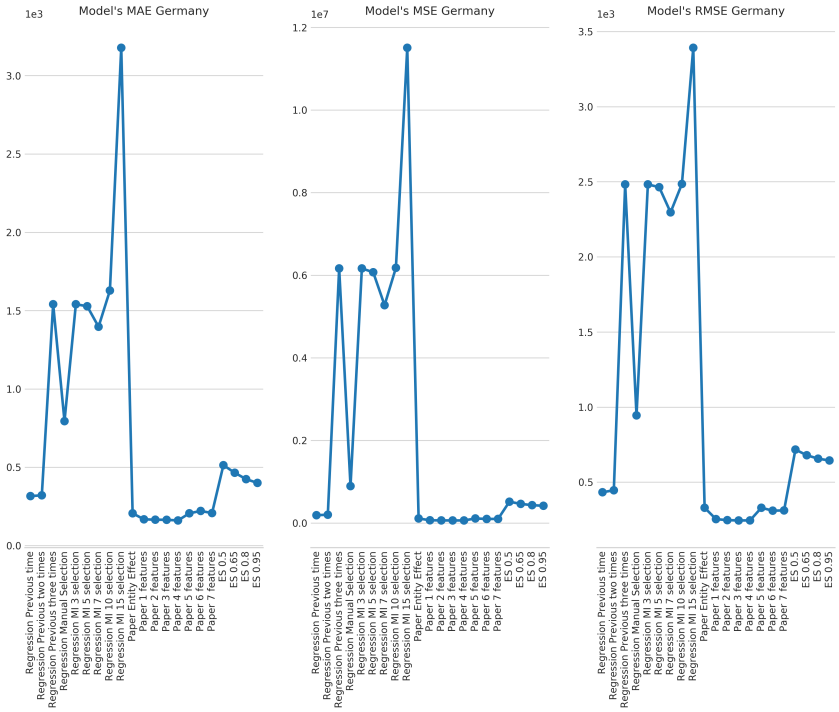


Figure 5.1. Error Study - Germany

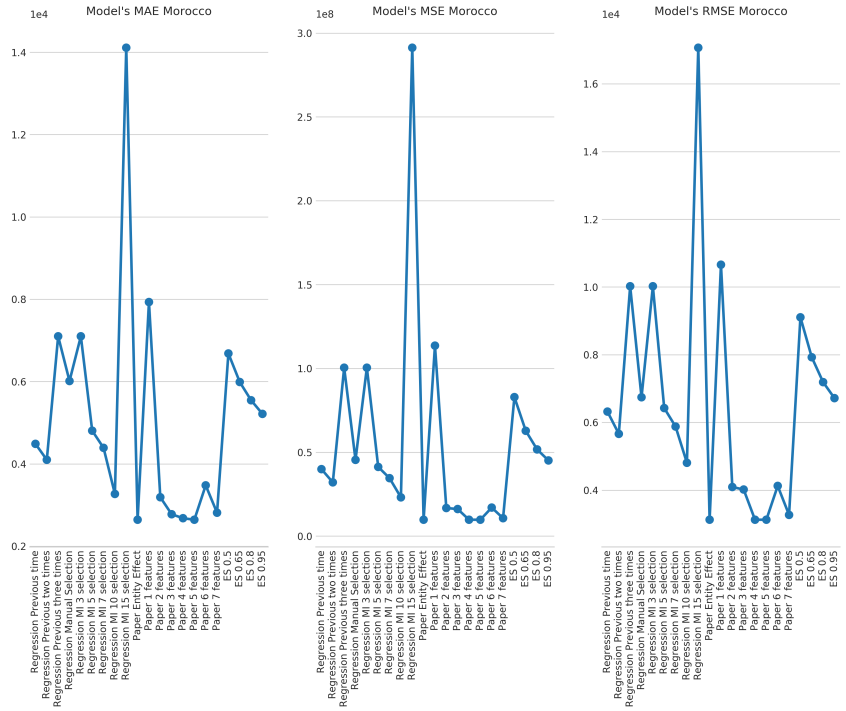


Figure 5.2. Error Study - Morocco

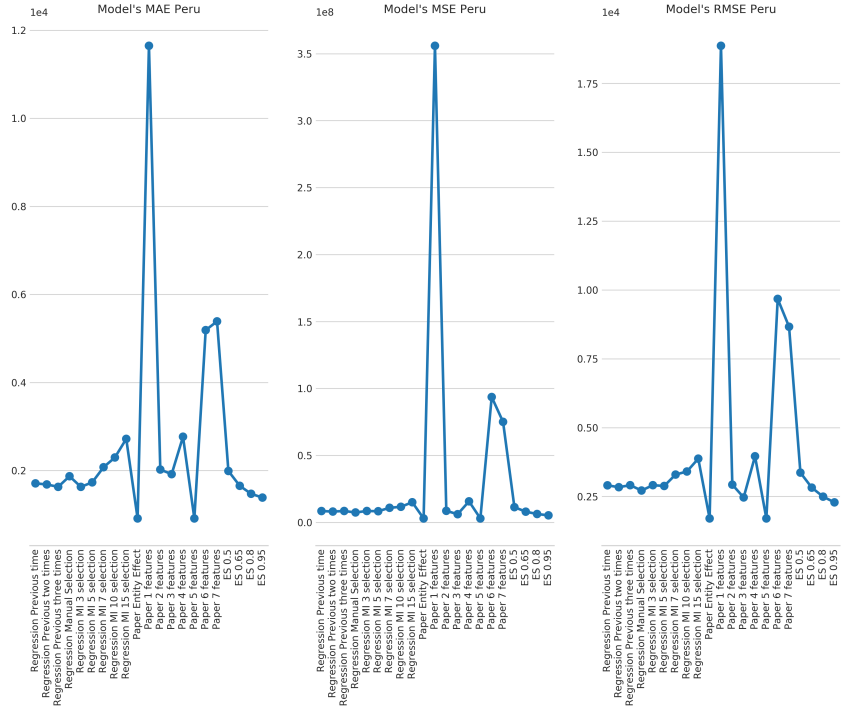


Figure 5.3. Error Study - Peru

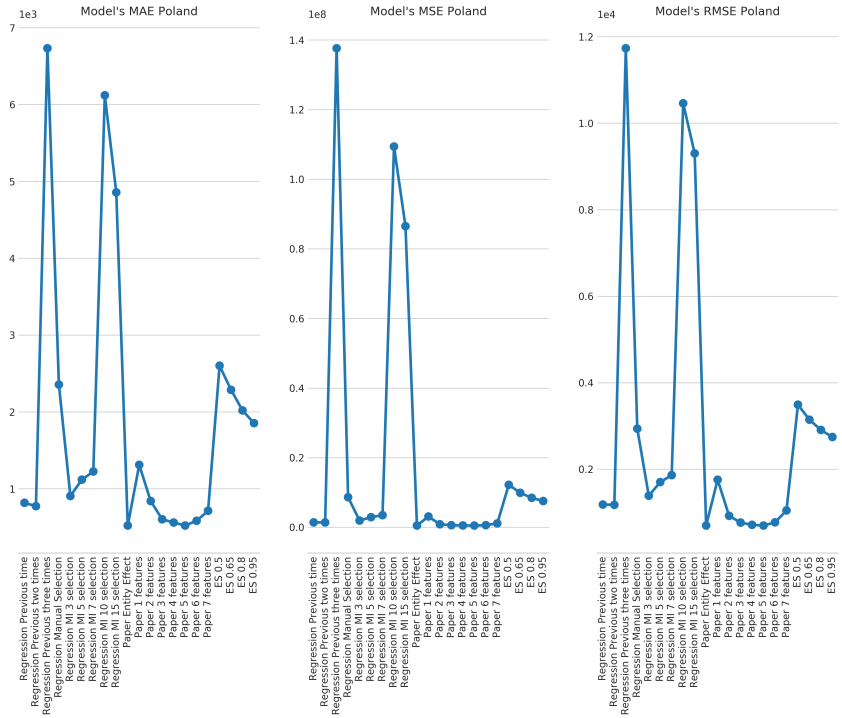


Figure 5.4. Error Study - Poland

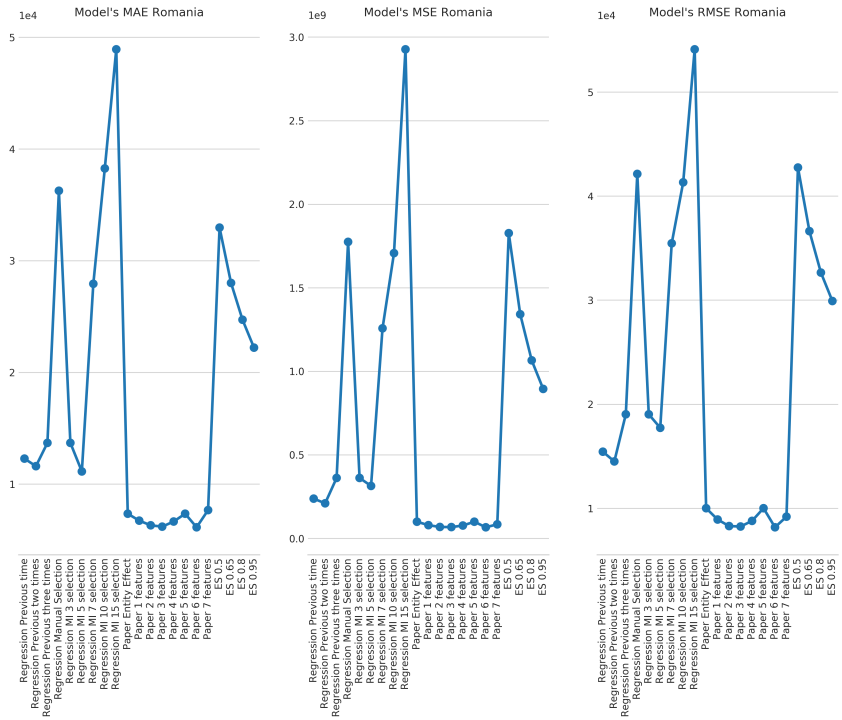


Figure 5.5. Error Study - Romania

5.2 Spain

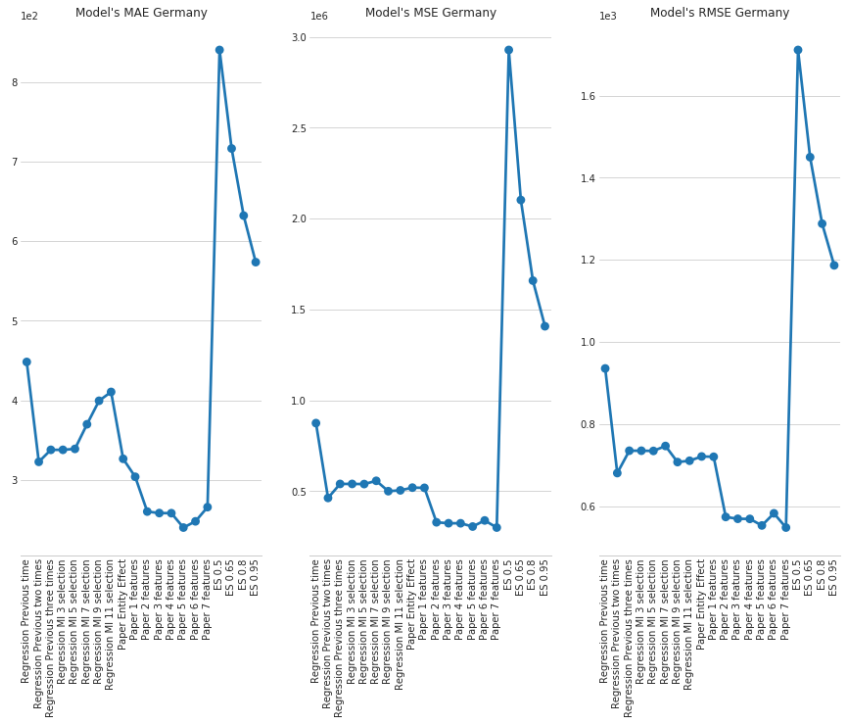


Figure 5.6. Error Study - Germany

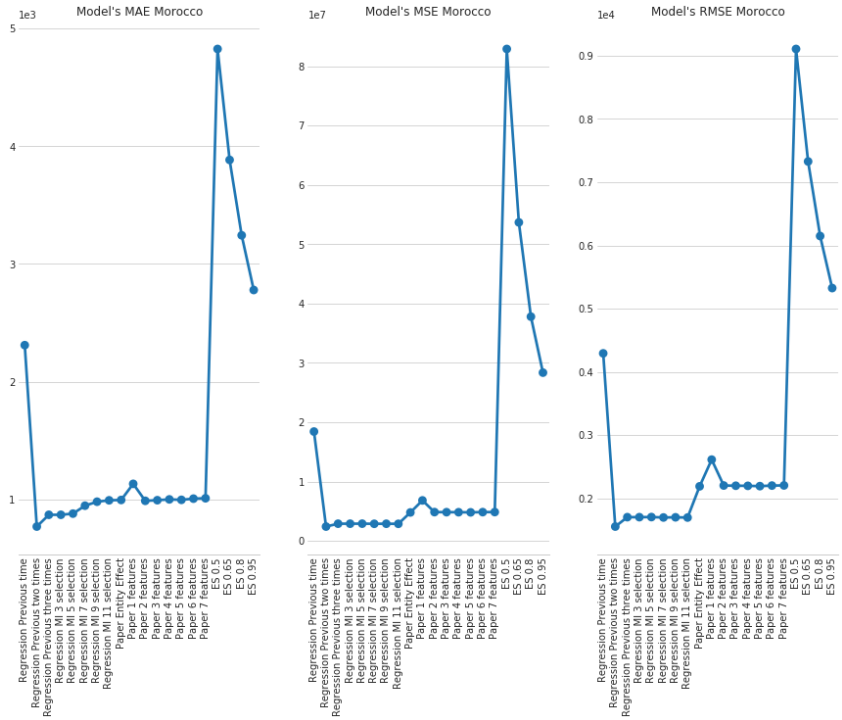


Figure 5.7. Error Study - Morocco

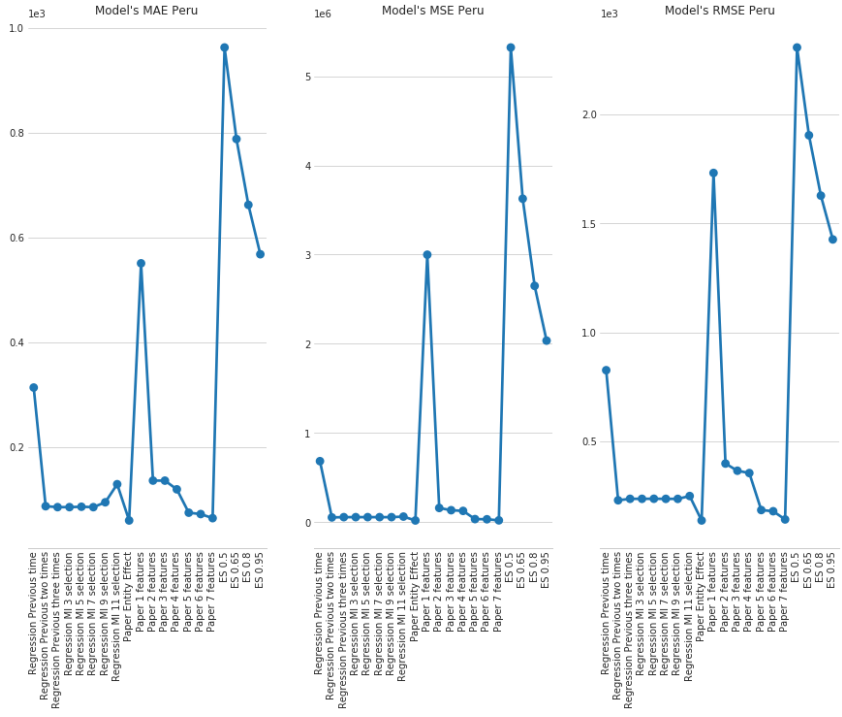


Figure 5.8. Error Study - Peru

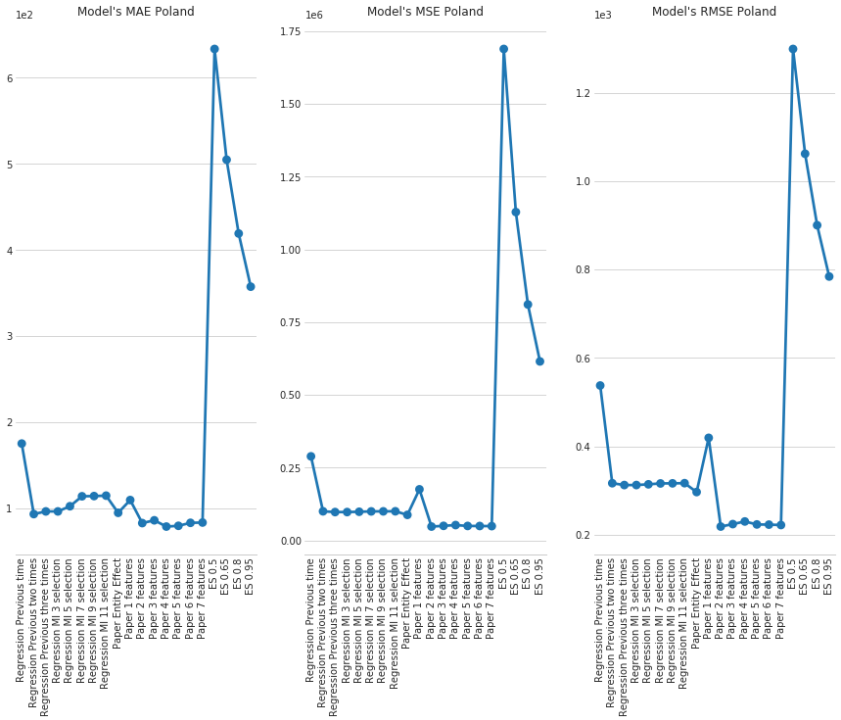


Figure 5.9. Error Study - Poland

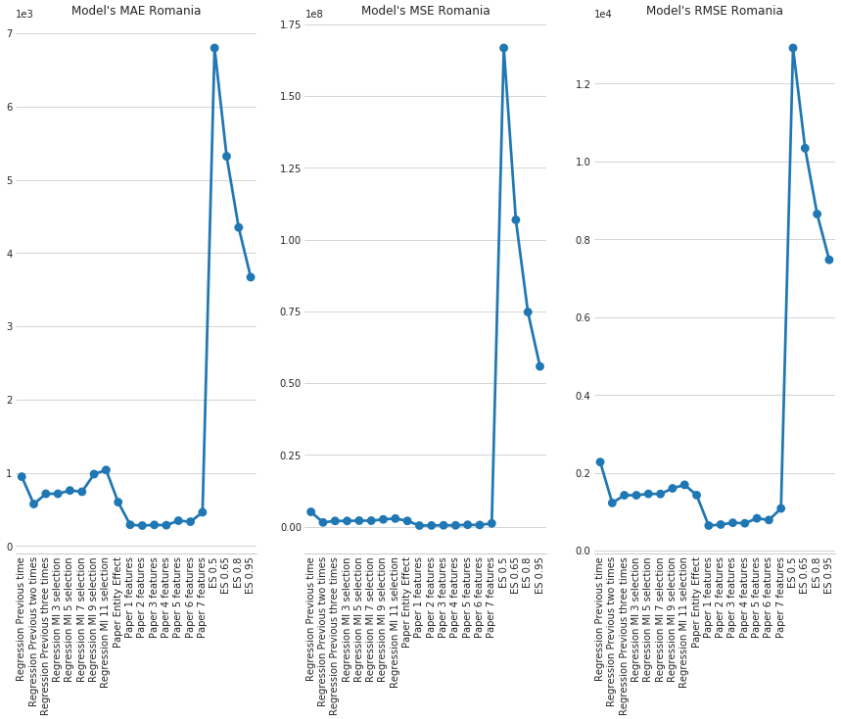


Figure 5.10. Error Study - Romania

The conclusions are the same for both the country Italy and Spain.

From the previous section, it is possible to say that adding features or using more sophisticated models does not improve significantly the performance in terms of prediction nor in terms of interpretation.

In particular:

- In the regression model, increase the number of features does not imply a notable decrease of the model error. In some cases, like *Cueta* and *Melilla*, adding independent variables worsen the prediction for all the origin countries but Morocco.
- As aforementioned, decomposing the *fixed location effects* into observable location factors is not really improving the estimation. In many cases, e.g. from Morocco to Italy, from Peru to Spain and from Romania to Spain, the output of the first stage model performs better than the one of the second stage. Just in the case Germany-Spain the second stage seems to be relevant.
- For the following origin-destination pairs there is no real difference between simple regression and the Jayet model, the errors of the two are almost equivalent: Peru-Italy, Poland-Italy, Peru-Spain, Poland-Spain, Romania-Spain.
- The regression outperform the Jayet model in the Germany-Spain case.

The study seems to emphasize the fact that most of the migration flow can be captured only by the network effect. So it can be identified by the number of previous arrivals in the same location.

Andrea Vandambrini. *The attractiveness represented by Europe and the West is now carried directly through the internet. There is a well-established narrative of the migrant who succeeded, and that sends selfies to friends - perhaps near a monument to a European city, or a luxury car pretending to be the owner - posting them on social media.*. Il Fatto Quotidiano, 4 July 2018. [1]

To conclude, we propose two major future topics to extend the present work. First, reproduce the study with data about illegal immigrant flow for both Italy and Spain. Second, analyze separately the migrants based on instruction level, age range, and gender.

[1] L'attrattiva rappresentata dall'Europa e l'Occidente è ormai veicolata in modo diretto attraverso internet. Esiste una ben consolidata narrativa del migrante che ce l'ha fatta, e che invia selfie agli amici – magari vicino a un monumento di una città europea, o a una macchina di lusso che finge sia la sua – postandoli sui social. [11]

Bibliography

[1] Anselin, L. , AK Bera, L. (1998). Introduction to Spatial Econometrics. *Handbook of applied economic statistics, 237.*

[2] Campomori, F. (2016). Le politiche per i rifugiati in Italia: dall'accoglienza all'integrazione. Missione impossibile? *Social Coherence Paper, N.02/2016.*

[3] Docquier, F. , Peri, G. and Ruyssen, I. (2014). The Cross-country Determinants of Potential and Actual Migration. *Int Migr Rev, 48: S37-S99.*

[4] Griffith, D. A. (1980), Towards a Theory of Spatial Statistics. *Geographical Analysis, 12: 325-339.*

[5] Jayet, H., Ukrayinchuk, N., De Arcangelis, G. (2010). The Location of Immigrants in Italy: Disentangling Networks and Local Effects. *Annals of Economics and Statistics, (97/98), 329-350. doi:10.2307/41219121.*

[6] Jayet, H., Rayp, G., Ruyssen, I. et al. (2016). Immigrants' location choice in Belgium. *Ann Reg Sci 57: 63.*

[7] Ord, K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association, 70(349), 120-126.*

[8] Valero-Matas, J.A.; Coca, J.R.; Valero-Oteo, I. (2014). Análisis de la inmigración en España y la crisis económica. *Pap. poblac [online]. 2014, vol.20, n.80, pp.9-45. ISSN 2448-7147.*

[9] Zeng G. (2015). A Unified Definition of Mutual Information with Applications in Machine Learning. Mathematical Problems in Engineering, vol. 2015, Article ID 201874.

[10] `http://ec.europa.eu/eurostat/documents/3217494/8787947/KS-05-17-100-EN-N.pdf/f6c45af2-6c4f-4ca0-b547-d25e6ef9c359`

[11] `https://www.ilfattoquotidiano.it/premium/articoli/i-selfie-degli-amici-e-il-sogno-del-lavoro-cosi-si-convincono-che-e-meglio-partire/`

[12] `http://www.interno.gov.it/sites/default/files/sub-allegato_n._25_-_intesa_conferenza_stato_regioni_del_10_luglio_2014.pdf`

[13] `http://www.repubblica.it/solidarieta/immigrazione/2011/03/06/news/1991_il_primo_grande_esodo_dall_albania_verso_l_italia-13263392/`

[14] http://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=03005.px&L=0

[15] http://stra-dati.istat.it/Index.aspx#

[16] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc42.htm

[17] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc43.htm

[18] https://github.com/SaraR-1/Immigration

[19] https://github.com/SaraR-1/Immigration-Models