Researchers' Migration Across Europe

Nonparametric Statistics Project

Simone Vantini, Francesca Ieva
Matteo Fontana, Andrea Cappozzo

Academic Year 2021/2022
Mathematical Engineering, Politecnico di Milano

Carlo Ghiglione, Jacopo Ghirri,
Marta Mastropietro, Alfredo Gimenez Zapiola

26 January 2022

Contents

1 Introduction

The goal of this project is to investigate the high skilled workers and researchers' migration flows in European Union and continental Europe.
This analysis is of high interest because this phenomenon is particularly difficult to capture due to the lack of coherent and widespread collection of data about it. Indeed, there is not an international uniform practice for tracking these movements: some countries do not even record them and the majority do it without standard and exhaustive procedures.
Nevertheless, a knowledge of the researchers migration flows would be of great utility to European policymakers. For instance, during Brexit negotiations, it was a compelling question to know the impact on the European researchers community of the hard border between United Kingdom and continental Europe, but there was not any available data to provide a satisfactory answer.
We propose to tackle the problem exploiting unconventional data, estimating the migration flows on top of *ORCiD* researchers international register.
The analysis we carry out is composed of the following steps:

1. reconstruction of migration flows from *ORCiD* register and collection of socio-economic factors of European countries;

2. individuation of subgroups of countries and their characterization with respect to socio-economic factors using *ad-hoc* designed techniques (based on *Nonparametric Combinations* ANOVA) to deal with heteroscedasticity issues;

3. highlight of the principal determinants that influence flows in and out of countries, finding appropriate models through nonparametric regression techniques;

4. analysis of the temporal properties of the phenomenon through longitudinal and functional approaches.

2 Dataset

2.1 ORCiD dataset

ORCiD (Open Researcher and Contributor ID) is an organization for all people involved in research, scholarship and innovation. It provides a unique identifier to each enrolled member, keeping track of all contributions and affiliations. In particular, it provides information about where and when each *ID* has studied or worked.

We use these raw data to reconstruct movements of affiliated researchers through space and time and build an estimator of the total number of people travelling from one country to the others, year by year. Consequently, the statistical units of this analysis are the total number of researchers respectively entering or leaving a country: we refer them as *InFlow* and *OutFlow*.

For the analysis, we consider just European countries (EU with United Kingdom and Switzerland, Norway, Russia, Iceland, Belarus, Ucraine and Turkey, for a total of 35 countries) and we select 2006-2015 as time horizon, being ORCiD a relatively young platform.

The underlying assumption of this work lies in ORCiD providing a representative sample of the global population of researchers. It is reasonable to assume that, in the restricted context of this analysis, representativeness holds.

2.2 Socio-economic determinants

In order to characterize each country we consider the following socio-economical indicators:

- *GDPCap*: GDP per capita is the Gross Domestic Product divided by midyear population;
- *LPPI*: Local Purchasing Power Index shows relative purchasing power in buying goods and services for the average net salary;
- *Education*: Global education level index, composed of the average of mean years of schooling (of adults) and expected years of schooling (of children);
- *HDI*: Human Development Index measures average achievement in three basic dimensions of human development (health, knowledge and a standard of living);
- *ResGDP*: Public research and development expenditure as a proportion of GDP.
- *GGGI*: Global Gender Gap Index examines the gap between men and women in terms of Economic Participation and Opportunity, Educational Attainment, Health and Political Empowerment;
- *PS*: Public Services Indicator refers to the presence of basic state functions, such as health, education, water and sanitation, transport infrastructure, electricity and power, internet and connectivity;

- *SA*: Security Apparatus considers the security threats to a state, such as bombings, attacks and battle-related deaths, rebel movements, mutinies, coups, or terrorism;
- *Patents*: The overall number of patents applications in that country;
- *StudStaff*: Average number of students per academic staff member;
- *CitDoc*: Average number of citations per document.
- *NumStaff*: Average total number of staff in academic institutions.

2.3 University Score

In this section, we introduce the University Score (*UniScore*), a hand-crafted index meant to be a global indicator of the overall prestige of the academic institutions of a country.

It is defined starting from the *QS World University Ranking*, the annual higher education ranking assigned by the company Quacquarelli Symonds that is internationally recognized as one of the most influential.

In particular, taking the 2018 overall global ranking of top 1000 universities, $UniScore$ is computed in the following way:

1. each university is assigned to class $c \in 1, ..., 10$ telling if its ranking is in position [1-100], [101-200], ..., [801-900] or [901-1000] respectively;
2. for each country i, the number of universities in each class $n_i(c)$ for $c \in 1, ..., 10$ is computed;
3. $UniScore_i = \sum_{c=1}^{10} \frac{1}{c} n_i(c)$

In this way, a country with an high number of academic institutions in the top positions of QS Rankings receives an high *UniScore* that, coherently, represents the overall international prestige of its universities.

A normalized version of the *UniScore* divided by the number of universities in the QS rankings is also developed trying to capture only the reputation of academic institutions of a country independently on their numerosity.

In the following analysis, the original version of *UniScore* proves to be very useful to explain the research flows contrary to the second that, as a consequence, is abandoned.

3 Group Analysis

The goal of this section is to uncover a qualitative representation of the phenomenon, by defining grouping structures among different countries and understanding what these groups differ in.

3.1 Groups definition

The first step is uncovering said groups. To do this we set ourselves in a Euclidian space, where the i^{th} country is represented by $Flow_i$, *i.e.* the total number of people moving in and out of it over the whole time horizon; hence we represent countries as:

$$Flow_i := (\sum_{year=2006}^{2015} InFlow_i^{year}, \sum_{year=2006}^{2015} OutFlow_i^{year}) \in \mathbb{R}^2$$

In this setting we perform hierarchical clustering with Ward's Approach, providing the following partition:

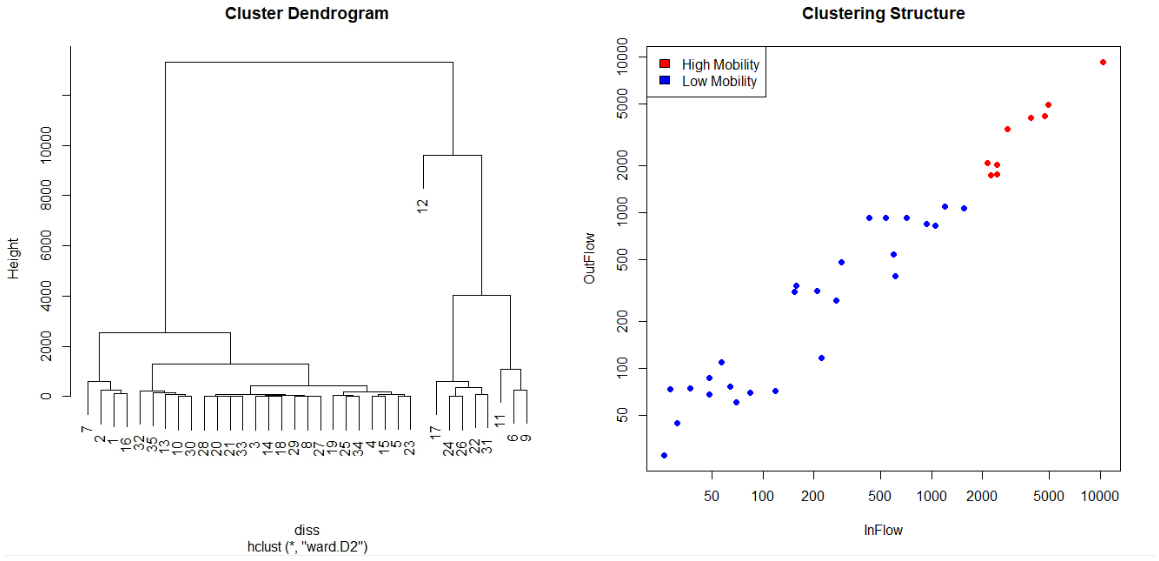

Figure 1: Clustering dendogram (sx) and partition induced on embedding space (dx).

This classification is interpretable as *High Mobility* countries versus *Low Mobility* countries:

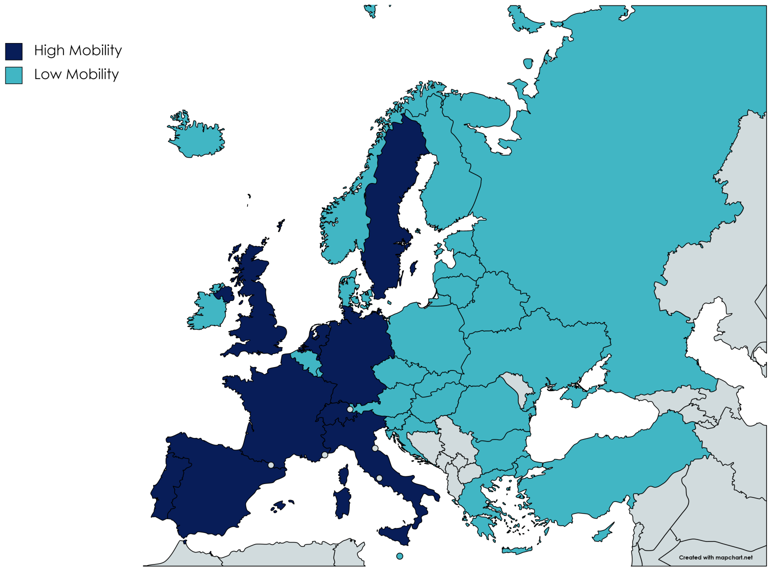

Figure 2: A visualization of the partition.

3.2 Multi-Aspect Testing

The next step is understanding with respect to which variables the two groups differ. The standard way to perform such a task is ANOVA testing but in this case the groups are characterized by both a non gaussian behaviour and strong heteroscedasticity with respect to most of the determinants of interest.

For this reason even the nonparametric version of the standard ANOVA test is not directly appliable. Given the univariate regressor divided in the two groups $\{X_i\}_{i=1}^{n_X}$ and $\{Y_i\}_{i=1}^{n_Y}$, the test is:

$$\begin{aligned} \{X_i\}_{i=1}^{n_X} \overset{iid}{\sim} X,\ \{Y_i\}_{i=1}^{n_Y} \overset{iid}{\sim} Y \\ H_0 : X = Y \\ H_1 : X \neq Y \end{aligned}$$

but this amounts to testing $\{X_i\}_{i=1}^{n_X} \overset{d}{=} \{Y_i\}_{i=1}^{n_Y}$ that is violated by the difference of scale between the two groups. Indeed, the objective is to detect if there is a difference in the locations of two populations regardless of the differences in scale.

A possible solution to this problem is offered by *Multi-Aspect Testing*, that is able to test differences in distribution conducting two simultaneous tests for the two types of violation of the null hypothesis: differences in location and in scale. Due to the presence of possible outliers, we decide to make those tests robust.

The two simultaneous test have the following definition:

$$\begin{aligned}&\textbf{Test 1}\\ &H_0 : location_X = location_Y\\ &H_1 : location_X \neq location_Y\\ Test\,statistic : T_1 &= (Med(\{X_i\}_{i=1}^{n_X}) - Med(\{Y_i\}_{i=1}^{n_Y}))^2\end{aligned}$$

$$\begin{aligned}&\textbf{Test 2}\\ &H_0 : scale_X = scale_Y\\ &H_1 : scale_X \neq scale_Y\\ Test\,statistic : T_1 &= (MAD(\{X_i\}_{i=1}^{n_X}) - MAD(\{Y_i\}_{i=1}^{n_Y}))^2\end{aligned}$$

However, in order to consider the two tests simultaneously, a *pvalue* correction is needed. Since the *Bonferroni* correction would lead to a too conservative results, we decide to embed this problem in the framework of *Nonparametric Combinations* (NPC).

3.3 NPC ANOVA

3.3.1 Theoretical introduction

The idea of *Nonparametric Combinations* is to account for the dependence between two or more marginal tests by having a *pvalue* correction based on the result of other tests, obtained as the combination of the marginal tests themselves.
Rigorously, given a set of marginal tests $\{T_j\}_{j\in J}$ and a test T_i, $i \in J$, we call $T_{i\cap A}$ the test obtained combining T_i with $\{T_k\}_{k\in A}$, $A \subseteq J\backslash\{\text{i}\}$ that tests:

$$H_0 : H_{0_i} \cap (\bigcap_{k\in A} H_{0_k})$$

$$H_1 : H_{1_i} \cup (\bigcup_{k\in A} H_{0_k})$$

Let's call $p_{i,A}$ its *pvalue*. Now the corrected *pvalue* $\tilde{p}_i$ for T_i is

$$\tilde{p}_i = \underset{A\subseteq J\backslash\{i\}}{max} \{p_{i,A}\}$$

It should be noted that, if $A = \emptyset$, $T_{i\cap A}$ is just T_i and $p_{i,A} = p_i$. In other words, the correction is based on testing the marginal null hypothesis when combined with all possible combinations of all other null hypothesis, hence keeping into account their possible dependence.
Such a procedure provides control over the *Family-Wise Error Rate* (FWER), allowing in our case the tests to be considered simultaneously, possibly in a more powerful way than simple *Bonferroni*.

3.3.2 Implementation of NPC

Summing up, we need to build the joint test:

$$H_0 : location_X = location_Y \wedge scale_X = scale_Y$$

The power of this approach lies in the choice of a test statistic for the joint test.
A first solution suggested by the literature is the *Fisher Combining* function, which introduces as test statistic $T_{1,2}$ for the joint test:

$$T_{1,2}^{Fisher} = -2log(p_1 \, p_2)$$

where p_1 and p_2 are the *pvalues* of the two marginal tests.
Moreover, we design a second *ad-hoc* test statistics specific for this problem. Calling U_1, U_2 the two *Mann-Whitney U-statistics* computed via sum of ranks of the direct observations of the two groups, and V_1, V_2 the two *Mann-Whitney U-statistics* computed via sum of ranks over the square deviation of the observations with respect to their group median, the test statistic for the joint test is:

$$T_{1,2}^{Rank} = max\{max(U_1, U_2), max(V_1, V_2)\}$$

In this way, the test statistic is equally susceptible to differences in location (detected by the U-test over observations, hence checking if the medians of the two populations are the same) and scale (detected by the U-test over squared deviations, checking if the median squared deviation is the same between the two populations). *Mann-Whitney* statistics are based on sum of ranks, so the obtained global test statistic is not sensitive to the plausible difference in magnitude of the two individual test statistics for scale and location.
This allows to perform ANOVA, checking for differences in the location of two population while keeping into account the differences in scale.

3.3.3 Development of NPC

We implement this procedure in R and we inspect its performance on simulated gaussian data for differences in scale and location.
For difference in scale, we set a grid of $\delta \in [0, 2]$:

$$X_1, X_2, ...X_{50} \overset{iid}{\sim} \mathcal{N}(0, (1 + \delta)^2)$$
$$Y_1, Y_2, ...Y_{50} \overset{iid}{\sim} \mathcal{N}(1, 1)$$

For difference in location, we set a grid of $\delta \in [0, 2]$:

$$X_1, X_2, ...X_{50} \overset{iid}{\sim} \mathcal{N}(0 + \delta, 1)$$
$$Y_1, Y_2, ...Y_{50} \overset{iid}{\sim} \mathcal{N}(0, 2)$$

In the following plots, we see the power of the procedure compared to the respective marginal tests (Test 1 and Test 2 of section 3.2) respectively with and without *Bonferroni* correction.

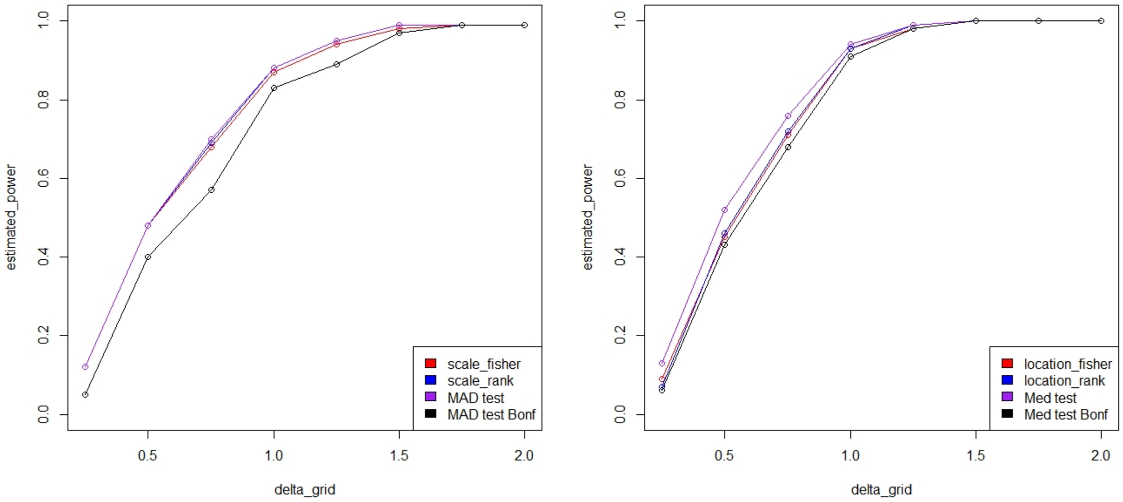

Figure 3: Estimated power for scale (sx) and location (dx) difference.

From this and other analysis with different distributions the NPC based correction retains significantly more power than *Bonferroni* and the *ad-hoc* designed test statistic seems to be slightly better than Fisher's.

3.4 Results

NPC based *Multi-Aspect* test allows to perform ANOVA on our data. We remind that we are just interested in the location results since the scale test is exclusively performed to apply the right *pvalue* correction to deal with heteroscedasticity.
First of all, we test if the division of continental Europe countries between EU vs non-EU nations is reflected in a significant difference of total migration flow:

$$H_0 : location(TotFlow_{EU}) = location(TotFlow_{non-EU})$$

$$H_1 : location(TotFlow_{EU}) \neq location(TotFlow_{non-EU})$$

This difference is not significant ($pvalue = 1$).
Then, we try to find significant differences between the two mobility-based groups (section 3.1) *w.r.t.* the socio-economic determinants. The conducted test has the following form:

$$H_0 : location(X_{HighMobility}) = location(X_{LowMobility})$$
$$H_1 : location(X_{HighMobility}) \neq location(X_{LowMobility})$$

with X a generic regressor of the dataset.
The results of the *one-at-a-time* ANOVA tests are summarized in the following table:

Determinants	*pvalue* with *ad-hoc* statistic	*pvalue* with Fisher
UniScore	0.000	0.000
CitDoc	0.100	0.163
NumStaff	0.011	0.011
StudStaff	0.945	0.945
GDPCap	0.097	0.201
ResGDP	0.048	0.142
Education	0.759	0.452
LPPI	0.045	0.102
HDI	0.042	0.059
GGGI	0.178	0.127
PS	0.031	0.031
SA	0.407	0.201
Patents	0.015	0.015

From these results, we can identify that *High Mobility* countries generally tend to have:

- *high reputation and quality of universities*: higher QS Ranking (*UniScore*) and in the sheer number of staff involved in their functioning.
- *comfortable daily life*: elevated levels of Public Services and Human Development indexes;
- *active entrepreneurial environment*: more yearly patent applications;
- *public investments*: huge public expense in R&D proportionally to GDP.

It is interesting to notice how there seems to be no evidence to say that people living in *High Mobility* countries are richer or have higher salaries, since both GDP per capita and Local Purchasing Power Index have no significant difference between the two groups.

4 Determinant Analysis

The goal of this analysis is to identify the relation between the academic migration flows and the socio-economic determinants. To do so, for each country both the *total flow* (*inflow* + *outflow*) and the *normalized net flow* (*inflow* - *outflow* divided by the number of researches), computed over the whole time horizon, are associated to one or more covariates presented in section 2.2.

4.1 Models for Total Flow

From the ORCiD data, for the i^{th} country the logarithm of the total flow ($LogTotFlow_i$) is defined:

$$LogTotFlow_i = \log(InFlow_i + OutFlow_i)$$

This quantity is analyzed *w.r.t.* all the available socio-economic determinants, showing any possible relation through both parametric and non-parametric regression techniques. The logarithm of the total flow was performed in order to recover homoscedasticity of residuals and to normalize the huge scale differences in the phenomenon.

4.1.1 Covariate selection

Variable selection for this model is carried out through the use of GAMs with Cubic B-Splines. The very limited amount of data (since we only have 35 countries) and the high numbers of degrees of freedom involved require the use of a forward procedure. The procedure leads to multiple admissible models where the best one in terms of adjusted R^2 and interpretability is the one based on *UniScore* alone.
Hence the selected form of a model is:

$$LogTotFlow = f(UniScore) + \varepsilon$$

Where f is fitted via nonparametric regression methods.

4.1.2 Model selection

Once the *UniScore* is identified as the most impactful factor, many different regression model are performed to explain the relation with *LogTotalFlow*.
For each model, *hyper-parameter* tuning (*i.e.* bandwidth for *Kernel Regression*, number of knots and degrees of freedom for *Spline Regression* and smoothing parameter for *Smoothing Splines*) is performed according to *Akaike Information Criterion (AIC)* and *Stratified K-Fold Cross-Validation w.r.t* the MSE.
This second method is designed *ad-hoc* to compensate the low numerosity of the dataset. It consist in composing the folds so that each fold contains units uniformly distributed according to the quantiles of *UniScore* regressor. In this way, each model of the *K-Fold Cross-Validation* is less biased by the lack of data, that in such a context can have a huge impact.
Once each Regression technique provides its best model, all the models are compared

in term of *Root Mean Square Error* (RMSE) to find the one that best fits the available data.
In the following tables, the performances in terms of RMSE of the best model selected by each technique are reported.

Regression Technique	RMSE for best $w.r.t$ AIC	RMSE for best $w.r.t$ K-Fold
Linear	1.519333	1.181639
Gaussian Kernel	0.580424	0.569026
Gaussian Kernel - KNN	0.537544	0.537544
Cubic B-Spline	0.526651	0.526651
Natural Cubic B-Spline	0.521259	0.521259
Smoothing Splines	0.515446	0.781174

4.1.3 Final model

In terms of RMSE, the best model results to be the *Smoothing Splines Regression* selected with AIC. As we can see in the below image, the resulting model lacks of interpretability due to its irregular shape.

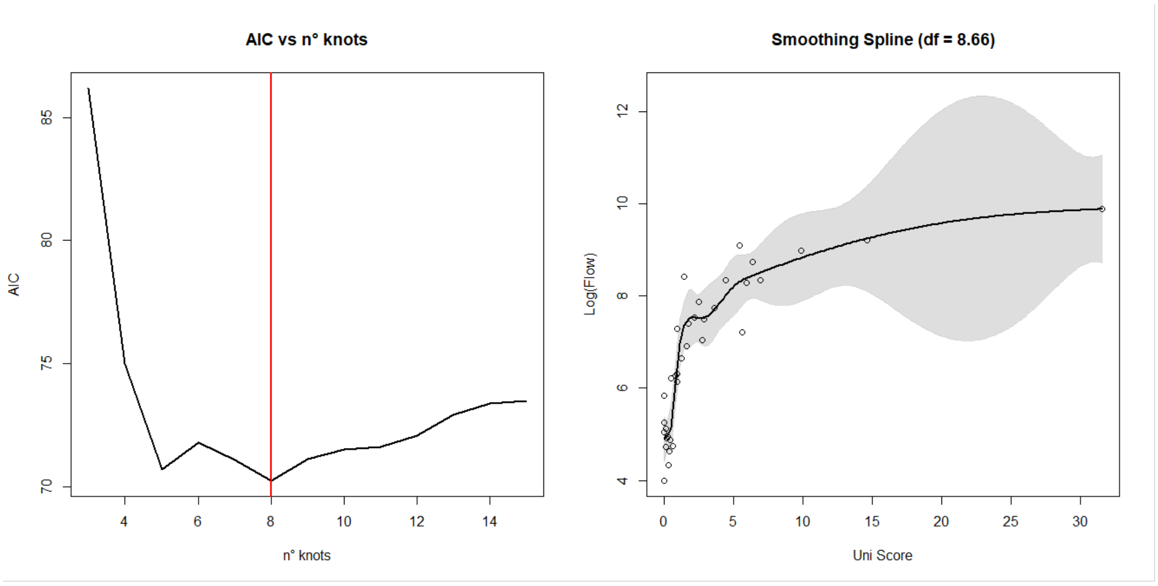

Figure 4: Model selection with AIC (sx) and corresponding fitted model with 95% confidence interval (dx).

As a consequence, the *Smoothing Spline* selected with *Stratified K-Fold* is chosen as best model. Despite its lower performances in terms of RMSE, it is preferable for its regular shape.

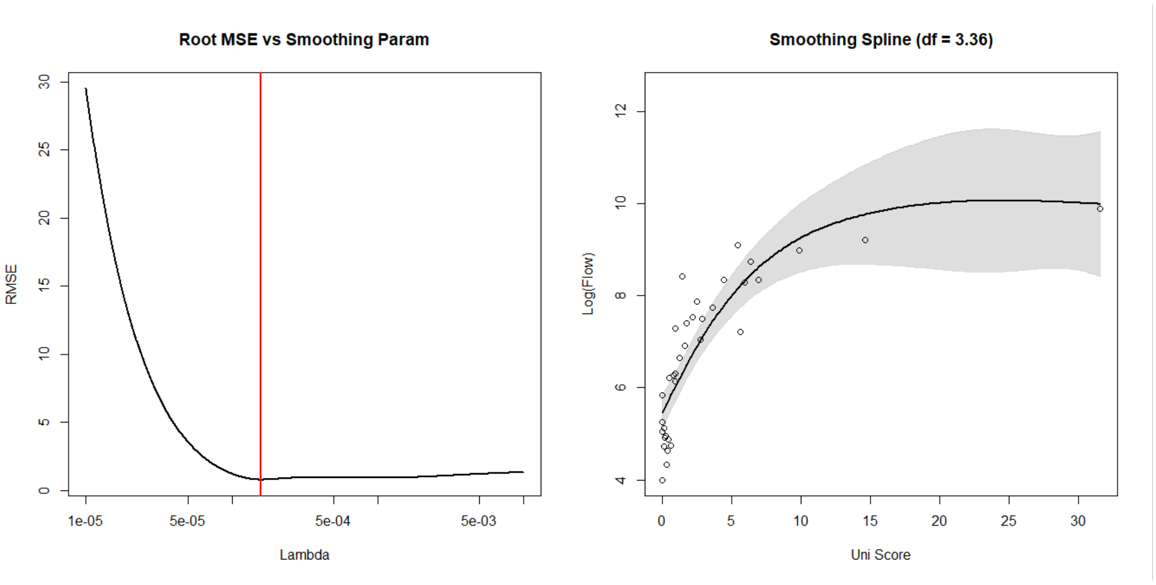

Figure 5: Model selection with Stratified K-Fold (sx) and corresponding fitted model with 95% confidence interval (dx).

This result shows that having prestigious academic institutions (*i.e.* an high *UniScore*) is strongly positively correlated to the total research mobility of a country. This relation is not linear, since for units with already high mobility the impact of *UniScore* is almost null.
This analysis shows that the increment the international reputation of the academic institutions of a country with low mobility levels could play a fundamental role to raise them. On the contrary, a nation with already huge academic flows would not benefit of a further augmentation of the prestige of its universities with the same efficiency.

4.2 Models for Normalized Flow

In this section, for each country i the net flow normalized *w.r.t.* the total number of researchers is analyzed:

$$NormFlow_i = \frac{InFlow_i - OutFlow_i}{Researchers_i}$$

In this way, it is captured the capacity of a nation to attract researchers independently on its volumes and population. As before, this quantity is analyzed in relation with the socio-economic factor presented in section 2.2.

4.2.1 GAM

Expecting nonlinear dependencies, covariate selection with *Generalized Additive Model* (GAM) is performed.

The best fitted model (in terms of $adjustedR^2$) has the following structure:

$$\begin{aligned} NormFlow = 0.186 - 0.0031\, StudStaff + f_1(CitDoc) + \\ + f_2(ResGDP) + f_3(UniScore) + \varepsilon \\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

being, for each country:

- *StudStaff*: average student per academic staff ratio in universities;
- *CitDoc*: average number of citations per document;
- *ResGDP*: average percentage of *Gross Domestic Product* spent in R&D sector;
- *UniScore*: university score introduced in section 2.3.

and f_1, f_2 and f_3 the three nonlinear functions fitted by the GAM in the below image.

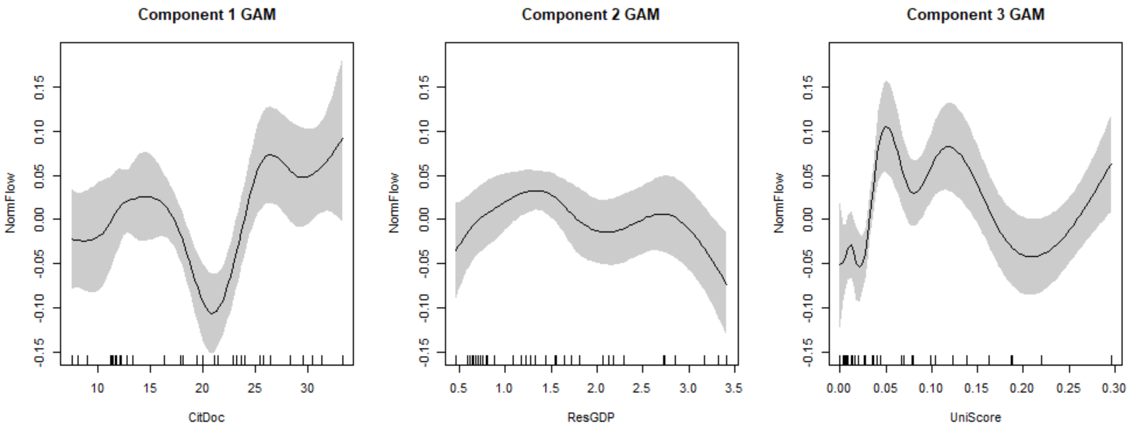

Figure 6: Nonlinear functions f_1 (sx), f_2 (center) and f_3 (dx) with 95% confidence interval fitted by GAM.

The model is able to explain almost 90% of variability of the data ($adjusted\ R^2 = 0.88$) having all hypotesis satisfied and the regressor statistically significant. Despite this, the three nonlinear components are very irregular and are not interpretable, so this model is discarded.

4.2.2 Composite Index definition

In order to avoid the issues encountered with GAM (section 4.2.1), the four regressors *StudStaff*, *CitDoc*, *ResGDP* and *UniScore* are synthesized in a unique composite index *CIndex*.
CIndex is defined as the first component of *Principal Component Analysis* applied to the four scaled regressors, able to explain 61.6% of their total variability.
The result is the following:

$$CIndex = -0.17\,StudStaff + 0.59\,CitDoc + 0.58\,ResGDP + 0.54\,UniScore$$

CIndex can be interpreted as a global score of the overall quality of the research sector of a country since it is a weighted average of four indicators of good state of academic institutions and investment. Coherently, *StudStaff* has a negative weight since an high ratio of students per teacher is generally associated to a poor didactic. Hence the selected form of a model is:

$$NormFlow = f(CIndex) + \varepsilon$$

Where f is fitted via nonparametric regression methods.

4.2.3 Model selection

Analogously to section 4.1, the relation between the normalized net flows (*NormFlow*) and the overall quality of research sector (*CIndex*) is analyzed with different regression models, performing *hyper-parameter* tuning both with AIC and *Stratified K-Fold Cross-Validation w.r.t.* MSE (see section 4.1.2).
In the following tables, the performances in terms of RMSE of the best model selected by each technique are reported.

Regression Technique	RMSE for best *w.r.t* AIC	RMSE for best *w.r.t* K-Fold
Linear	0.00403848	0.00403848
Gaussian Kernel	0.00373254	0.00363847
Gaussian Kernel - KNN	0.00375968	0.00353355
Cubic B-Spline	0.00351689	0.00325356
Natural Cubic B-Spline	0.00360410	0.00360410
Smoothing Splines	0.00365155	0.00363356

4.2.4 Final model

In terms of RMSE, the best model results to be the *Cubic B-Spline Regression* selected with *Stratified K-Fold.* As we can see in the image, the fitted model has a very irregular shape, possibly due to overfitting. This makes it not interpretable and generalizable.

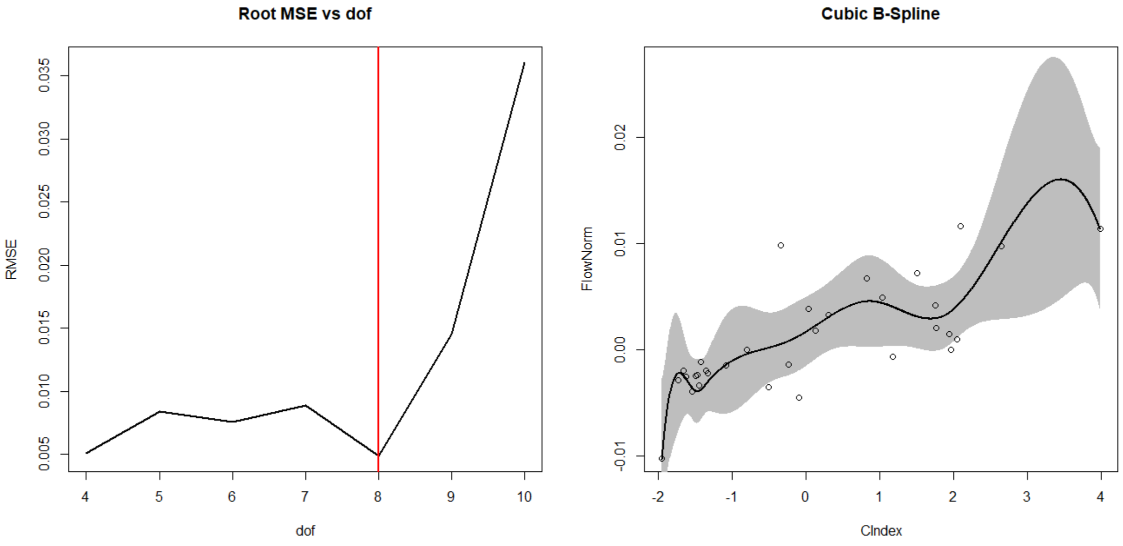

Figure 7: Model selection with Stratified K-Fold (sx) and corresponding fitted model with 95% confidence interval (dx).

As a consequence, the *Natural Cubic B-Spline Regression* model selected with *Stratified K-Fold* is chosen as best one. On the one hand, it has one of the lowest RMSE among all the candidates and, on the other one, it has a very regular and smooth shape.

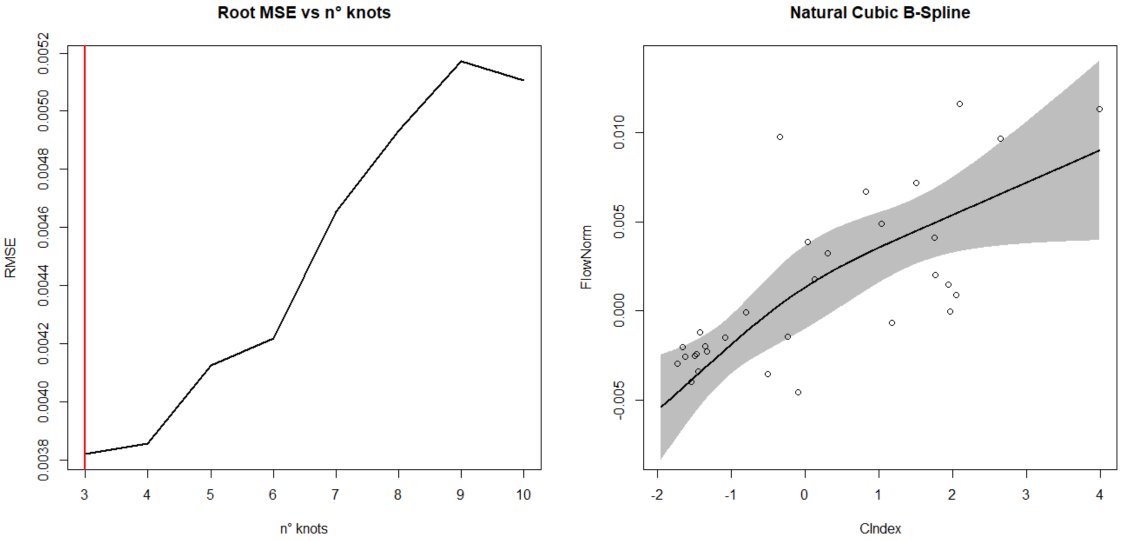

Figure 8: Model selection with Stratified K-Fold (sx) and corresponding fitted model with 95% confidence interval (dx).

The model shows an evident positive correlation between *CIndex* and *FlowNorm*. This suggests that countries with high quality of the academic sector are more able

to attract researchers *w.r.t.* the ones with few investments and poor academic activity and institutions.
Moreover, this result shows that the relation is slightly nonlinear with a concave shape. This means that an overall improvement of the R&D sector would be particularly important for the countries that want to attract more researches but are not able to.

5 Temporal Analysis

In this section the flows of researchers are no longer aggregated with respect to the countries involved in the migration only, but are also tied with the specific year the movement occurred in. The goal is to determine whether there is a temporal characterization of the phenomenon itself or in the way that we model it.

5.1 Phenomenon evolution

A crucial aspect of the research migration phenomenon is its behaviour along time. We recall the model of section 4.1, where we have as target variable the logarithm of the total flow for each country and as single regressor the university score. We approximate the nonlinear behaviour with a piecewise linear model featuring a single cut point in order to employ a linear setting for parametric inference, as it is a way easier approach that still retains a good enough predictive performance. The underlying assumption is that *UniScore* is stable year by year, indeed we want to check if its effect varies over time.
We thus pose the following question: if we were to gather the total flows separately each year (i.e. from 2006 to 2015), and to fit the same model respectively, would the estimated parameter β for the covariate *uniscore* be the same along time?
The model we then work on this:

$$\begin{aligned} &log(TotFlow_{i,t}) = \beta_0 + f_t(UniScore_i) + \varepsilon_{i,t} \quad \forall t \in \mathit{Year}, \quad \forall i \in \mathit{Country} \\ &f_t(x) = \beta_{1,t}x + \beta_{2,t}(x - x_0)\mathbb{I}_{\{x>x_0\}} \end{aligned}$$

where the Total Flow of researchers is divided not only country by country, but also year by year, and x_0 is the cut-off point ($x_0 = 5$).
Therefore, the null hypothesis is formulated as:

$$\begin{aligned} &H_0 : (\beta_{1,t_i}, \beta_{2,t_i}) = (\beta_{1,t_j}, \beta_{2,t_j}) \quad \forall\ i, j \in \{2006, ...2015\} \\ &H_1 : \exists\ i, j\ s.t.\ (\beta_{1,t_i}, \beta_{2,t_i}) \neq (\beta_{1,t_j}, \beta_{2,t_j}) \end{aligned}$$

Consequently, the problem reduces to determine if the interaction of the regressor *UniScore* with time (i.e. the year) is significant or not.
We should however state that this approach leads to a model where residuals are not *iid*, since we expect a strong correlation between units observed for the same country but at different years. So, we have to find a way to account for this correlation.

5.1.1 Random Effects Model - Framework

The way we account for the dependence amongst observation is with a *Random Effect Model* that, contrary to the standard practice of ARMA models, allows to exploit permutation tests, without needing any parametric assumption.

Therefore, considering now the vector form $\underline{y}_i = \{LogTotFlow_{i,t}\}_{t \in Years}$, the model becomes:

$$\begin{aligned} \underline{y}_i &= \mathbf{X}_i \underline{\beta} + \mathbf{Z}_i \underline{b}_i + \underline{\varepsilon}_i \quad && \forall i \in Country \\ \underline{b}_i &\overset{ind}{\sim} \mathcal{N}(\underline{0}, \mathbf{Z}_i \mathbf{D}_i \mathbf{Z}_i^T) && \forall i \in Country \\ \underline{\varepsilon}_i &\overset{iid}{\sim} \mathcal{N}(\underline{0}, \sigma^2 \mathbf{I}) && \forall i \in Country \end{aligned}$$

where now $\mathbf{X}_i$ is the design matrix for i^{th} Country, and $\mathbf{Z}_i$ the design matrix for the relative random term. In particular, $\mathbf{X}_i$ is the design matrix for the above model, we are just changing the regression parameters amongst different years.
It should be noted moreover that, since we are working in a Nonparametric framework, we are not rigid with respect to the Gaussianity assumptions over $\{\underline{\varepsilon}_i\}_{i \in Country}$, indeed we can even assume that $\varepsilon_{i,t} \overset{iid}{\sim} F_\varepsilon$, where F_ε is a generic unknown distribution.
As a last remark, since we are modeling $\underline{\varepsilon}_i$ with a diagonal covariance matrix, this model is a *conditional-independence model*, indeed, conditionally on $(\underline{\beta}_i, \underline{b}_i)$, all component of all $LogTotFlow_i$ are independent too.
For what concerns the modeling of the random effect part, we decide to build each $\mathbf{Z}_i$ as a country dependent random intercept. This allows to model time dependencies while guaranteeing independence between countries.

5.1.2 Random Effects Model - Fitting and Testing

The model is fitted via *Restricted Maximum Likelihood* as it is known it provides better estimates than *Maximum Likelihood* for the within group covariance matrix, hence the one that we exploit to account for time dependence amongst our observations.
The test we are performing, in terms of its implicit expression with respect to the random component, can be equivalently stated as:

$$\begin{aligned} &H_0 : LogTotFlow_{i,t} = \beta_0 + f(UniScore_i) + \tilde{\varepsilon}_{i,t} \quad \forall t \in Year, \quad \forall i \in Country \\ &H_1 : LogTotFlow_{i,t} = \beta_0 + f_t(UniScore_i) + \tilde{\varepsilon}_{i,t} \quad \forall t \in Year, \quad \forall i \in Country \end{aligned}$$

where $\tilde{\varepsilon}_{i,t}$ is now the random variable obtained by the sum of the actual residuals $\varepsilon_{i,t}$ and the random component, f has the same structure of f_t, but is constant in time.
To perform the test, we adopt the following permutation scheme:

1. Estimate the full model, hence leaving the time dependence of the regression coefficients.

2. Obtain the test statistic T_0, given by the maximum of the absolute values of the t-statistics of each of the interaction terms, hence calling t_0 a reference year:

$$T = \underset{s \in Years}{max} [|t(\beta_{1,s} - \beta_{1,t_0})|, |t(\beta_{2,s} - \beta_{2,t_0})|]$$

3. Estimate the reduced model, hence the one under H_0, obtaining its fitted values and residuals.

4. Repeat B times:

(a) Permute the residuals $\varepsilon_{i,t}$ of the reduced model.

(b) Compute the permuted responses:

$$y_i^{perm} = y_i^{fitted} + \hat{\varepsilon}_{\pi(i)}$$

(c) Fit the full model on the permuted responses, store its test statistic given by the above formula

5. The *pvalue* is given by the proportion of test statistics higher or equal than the original one yielded by the full model (T_0)

5.1.3 Robust Random Effects Model

To protect ourselves from the presence of outliers, we decided to fit a robust version of the Random Effects model, using the R package *robustlmm*.
The model structure is the same as before, but its fitting is different.
Firstly, the smoothed Huber $\phi - function$ is applied over the residuals and the same procedure is done to estimate the variance. The parameters of these functions are such that we still gain an efficiency of the estimators of 95%.
Secondly, for the random term, we apply the same weights for all units and no smoothing is performed, we indeed just want to cover ourselves from the presence of outliers in the residuals.
The test statistic and the method applied for computing the estimate of its distribution is the one applied in Section 5.1.2.

5.1.4 Results

For what concerns the result of the permutation test for the Random Effects model (section 5.1.2), with $pvalue = 0$ we do reject the null hypothesis; however, the residuals of the model under H_0 do not show a regular behaviour due to the presence of outliers. This is the reason why we implement a robust procedure.

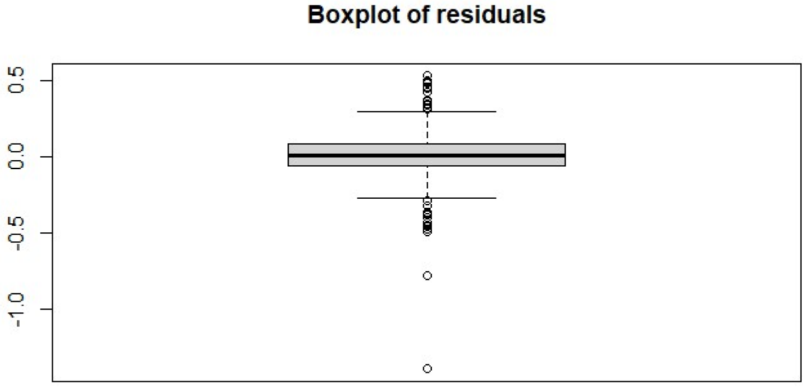

Figure 9: Boxplot of the residuals of the fitted model

For what concerns the permutation test for the Robust Linear Random Effects model, $pvalue = 0$, confirming the rejection of the null hypothesis.

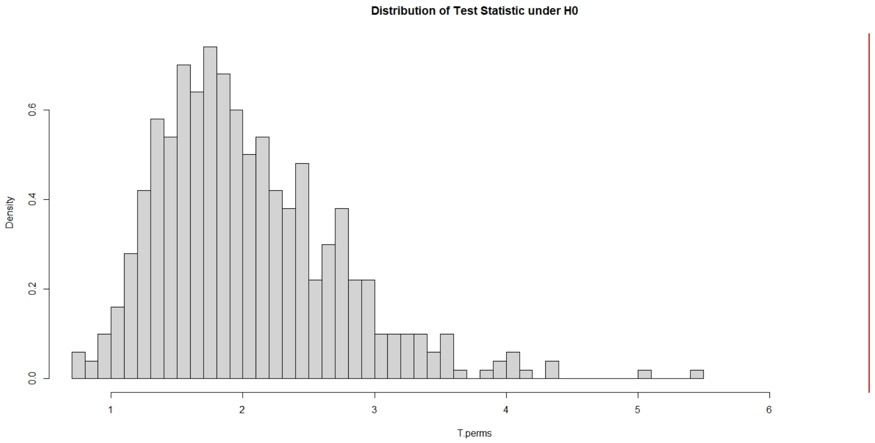

Figure 10: Empirical distribution of test statistics T with the observed value in red.

Indeed, this procedure successfully reduces the impact of extreme observations on the model fitting procedure, giving such outlying points a lower weight. A visualization of this aspect can be shown by looking at the fact that the observations with a lower weight are those that differ from the distribution of the bulk of data, which in this case is gaussian:

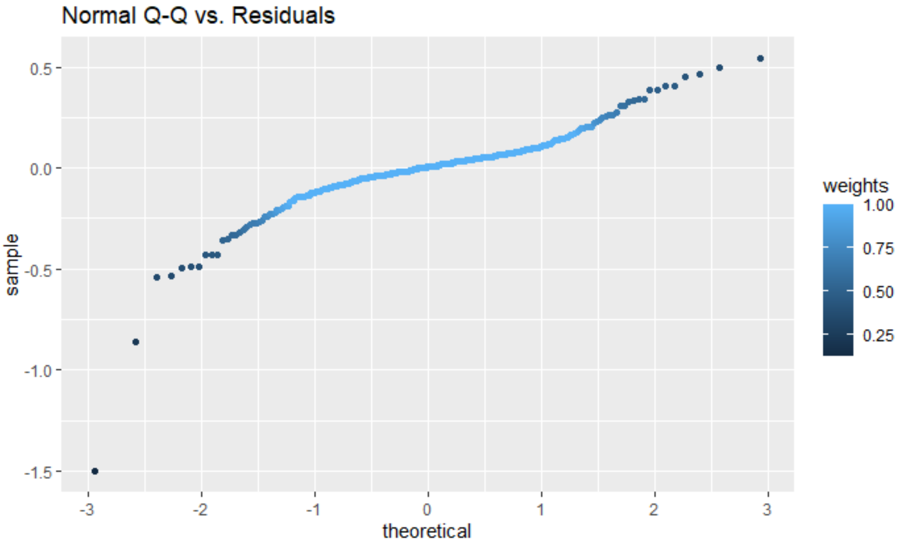

Figure 11: QQplot of the residuals of the robust model with respective weights.

In conclusion, the robust procedure provides statistical evidence to state that the impact of *UniScore* changes over time.

This result, however interesting, should be tested on a less rough approximation of the nonlinear model discussed in section 4.1. Nevertheless, we pose the attention on the fact that our time horizon touches the beginning of Brexit discussions, so it is possible that this led to a change in the UK's mobility, which is undoubtedly a leverage point.

5.2 InFlow and OutFlow correlation

In this section, it is analyzed the eventual concordance or discordance of researchers in and out flows for all countries, considering them as functions of time in the period 2005-2015.

To do so, first of all the year by year *InFlow* and *OutFlow* data are normalized *w.r.t.* the number of researchers in the country to avoid scale variability issues and dependence on the volumes. Then, for each i^{th} country data are smoothed with *Cubic B-Splines* to obtain a functional bivariate dataset:

$$\{\mathbf{X}(t), \mathbf{Y}(t)\} = \{X_i(t), Y_i(t)\}_{i=1}^{35} \quad t \in [2006, 2015]$$

with $X(t)$ representing the *InFlow* and $Y(t)$ the *OutFlow* as function of time.

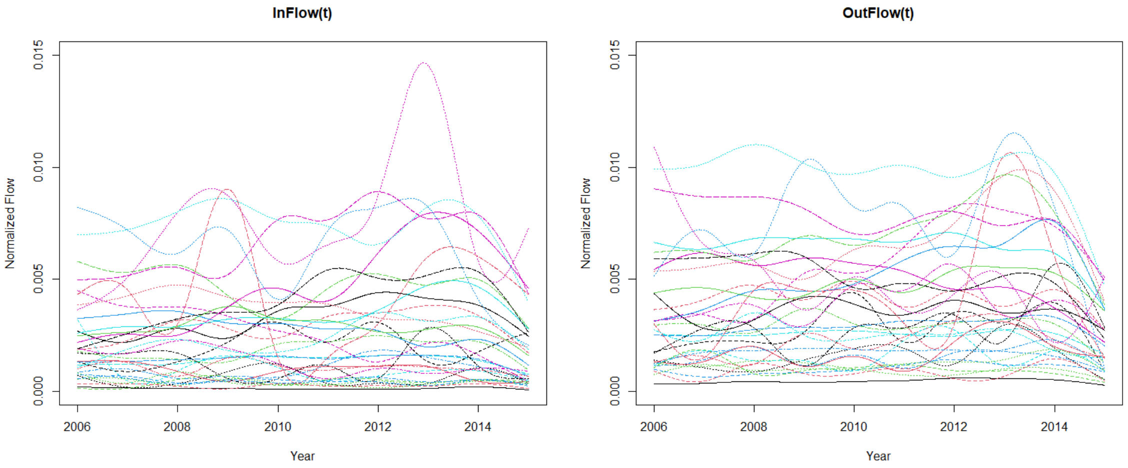

Figure 12: Bivariate functional data: $InFlow(t)$ (sx) and $OutFlow(t)$ (dx).

5.2.1 Independece Test with Spearman Index

In order to understand if generally *InFlow(t)* and *OutFlow(t)* behave in a concordant way or not, an *Independence Test* with *Spearman Correlation Index* is performed. The test has the following structure:

$$H_0: \mid \rho_s\left(X(t), Y(t)\right) \mid = 0$$
$$H_1: \mid \rho_s\left(X(t), Y(t)\right) \mid \neq 0$$

using as test statistics T the absolute value of the *Spearman Correlation Index* $\rho_s\left(X(t), Y(t)\right)$ for the functional bivariate dataset $\{\mathbf{X}(t), \mathbf{Y}(t)\}$ based on *Modified Epigraph Index* (MEI).

To perform the test, the following permutation scheme is used:

1. $T_0 =\mid \rho_s\left(\mathbf{X}(t), \mathbf{Y}(t)\right) \mid$ for the observed data is computed;
2. $\{Y(t)\}_{i=1}^{35}$ observed units are randomly shuffled B times in order to obtain B new permuted datasets $\{\mathbf{X}(t), \mathbf{Y}_b(t)\}_{b=1}^{B}$
3. for each permuted dataset $\{\mathbf{X}(t), \mathbf{Y}_b(t)\}$, the corresponding test statistics is computed:
$$T_b =\mid \rho_s\left(\mathbf{X}(t), \mathbf{Y}_b(t)\right) \mid$$
4. the collection of the T_b provides the empirical distribution of the test statistics;
5. the *pvalue* of the test is computed as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left[T_0 > T_b\right]$.

5.2.2 Results

The test is performed shuffling the dataset $B = 10000$ times. In the below image, the empirical distribution of $T =\mid \rho_s\left(X(t), Y(t)\right) \mid$ and the corresponding value T_0 for the observed data is shown.

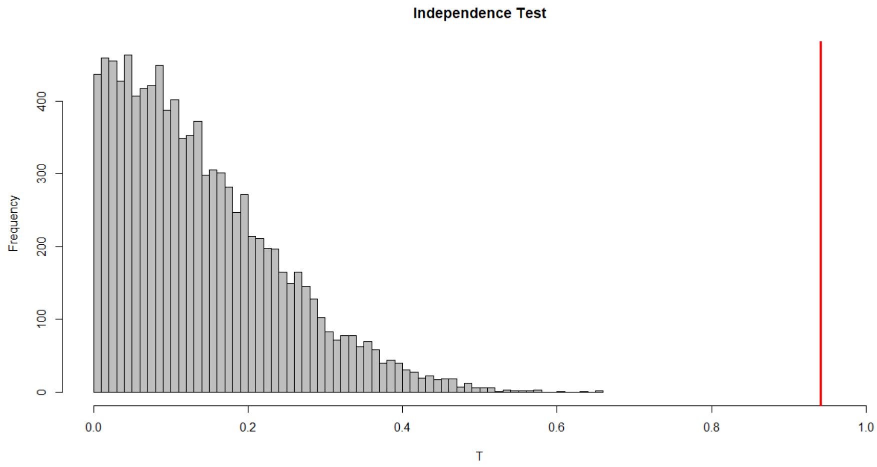

Figure 13: Empirical distribution of test statistics T with the corresponding value of T_0 in red.

The procedure provides for the test *pvalue* $= 0$, supporting the rejection of null hypotesis H_0.
This allows to affirm with high statistical evidence that *InFlow(t)* and *OutFlow(t)* do not behave in an uncorrelated way.
In particular, since $\rho_s\left(\mathbf{X}(t), \mathbf{Y}(t)\right) = 0.94$, for each country the two flows behave in a positive concordant way across the years, suggesting that if the inward mobility increases, the outward does too and viceversa.

6 Conclusions

During this analysis, the use of *ORCiD* dataset allows to treat the researchers migration phenomenon in a coherent way, overcoming the issue of lack of international uniform data.

Preliminary analyses show that is possible to partition European countries in high and low mobility groups (section 3.1). This division is reflected into different characterization according to socio-economic factors (section 3.4).

We observe that is is useful to divide the analysis of mobility in two parallel paths:

1. total migration flows allow to focus on overall richness and activity of the research environment of countries (section 4.1);
2. net migration flows allow to focus on the attractiveness of the R&D sector of countries and to eventually detect brain drain phenomena (section 4.2).

In both cases, global indicators of the prestige of academic environment and of the overall health of R&D sector proves to be effective to explain these trends.

In particular, we benefit from treating the mobility phenomenon as a whole, since we observe that in and out flows behave in a strongly concordant and positively correlated way across the years (section 5.2).

From the temporal analysis (section 5.1), we note that results could not be the same across all years so, when approaching such a phenomenon, one should always be aware of the specific situation that the time horizon is linked with.

It is of our opinion that there results could be of interest on a practical standpoint for what concerns policy-making. In particular, for the countries characterized by low levels of mobility the strategy may be to improve its prestige in R&D sector. On a EU level, this could be obtained with a more targeted allocation of already existing funds on this direction, which should then be focused on academic initiatives aimed at getting more international recognition.

For the countries that, on the contrary, already have a well internationally established research environment, opportunities and improvements can still be found in widespread investments in this sector.

References

[1] Caughey, Devin, Dafoe, and Seawright. Nonparametric Combination (NPC): A framework for testing elaborate theories. *The Journal of Politics*, 79(2):688–701, April 2017.

[2] N. M. Laird and J. H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):936–974, December 1982.

[3] J. C. Pinheiro and James H. Ware. Conditional versus Marginal Covariance representation for Linear and Nonlinear models. *Austrian Journal of Statistics*, 35(1):31–44, April 2016.

[4] F. Ieva, F. Palma, and J. Romo. Bootstrap-based Inference for Dependence in Multivariate Functional Data. *Submitted*, 2017.

[5] M. Koller. robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software*, 75(6):1–24, December 2016.