Research project in Econometrics

Social connections and views in Europe[*]

Semen Zhizherin[†] Danila Kochnev[‡]

Moscow, 2021

Abstract

This work is devoted to identifying and quantifying the impact of social connections on the views of Europeans. Among the studied variables, two proxies of euroscepticism are considered, as well as some other variables reflecting the views of Europeans. The estimation is carried out using models of spatial econometrics, where, among other things, a matrix based on the Social Connectedness Index between regions is used as the matrix of spatial weights. The main result is that spatial econometrics models turn out to be better than OLS for all dependent variables. Additionally, it confirms the hypothesis that social connections better describe the preferences of Europeans than geographic distances. The results obtained are robust to the removal of a part of the sample.

Keywords: social connectedness, spatial econometrics, euroscepticism

[*] We thank our fellow students Alina Kudanova, Taisia Kulikova, and Nikolay Markov for their contribution to the work on the project. Our particular thanks go to Dmitry Artamonov and Olga Suchkova for their advice and guidance.

[†] Moscow State University, Faculty of Economics, 1(46) Leninskiye Gory, 119991, Moscow, Russia, e-mail: samzhizherin@gmail.com

[‡] Moscow State University, Faculty of Economics, 1(46) Leninskiye Gory, 119991, Moscow, Russia, e-mail: danila.kochnev2000@mail.ru

Contents

1 Introduction

A classic assumption of cross-sectional econometric analysis is sample observations' independence: in contrast to time series or panel data, here it is hard to imagine, how different observations collected at the same period of time can be dependent. But, once studying particular preferences is considered, it becomes clear that such an assumption seems to be unrealistic. One can argue that people's political and economic views and values can be influenced by their friends' and acquaintance's opinions – opinions of those communicated most frequently. Today, when a significant share of social interaction is conducted online, it turns out to be particularly interesting to study the effect of social networks on the preferences of their active users. The objective of this paper is to expose and estimate this effect by the Europeans' preferences analysis.

Currently existing studies on the topic converge to social connections' being a statistically significant determinant of various preferences from tourism and attitude towards work to scientific co-authorship and euroscepticism. However, in the majority of works the influence of this factor is estimated by adding some social connectedness proxies as a regressor. Such an approach does not pay attention to the spatial structure of the data, thereby ignoring observations' dependence. Here, it seems promising to use spatial econometrics models that would take into account the influence of observation in one region on observations in other ones. The relevance of this work lies in the application of these methods to refine the previously obtained estimates and check the significance of the described effect. While the most common approach of spatial models' construction in the literature understands the proximity of countries in terms of geographic distance, this work focuses on proximity in terms of social connectedness. As shown below, in a number of cases this approach significantly outperforms the geographic approach in terms of estimation accuracy.

Within the framework of this work, various Europeans' preferences are analyzed. First, political views and attitudes towards the European Union are examined. The motivation is that social connections contribute to the dissemination of political information both through the communication of individual users and as a result of the activities of entire communities. We then discuss more general preferences related to attitudes and values. The issues of tolerance and trust both in individual social groups (for example, homosexuals) and in institutions (for example, the government or the church) should also be viewed through the prism of social ties. The motivation here is quite the same: personal communication and the activities of groups on the network help to disseminate information, thereby making spatial observations interdependent.

2 Literature overview

Over the past year, a large number of works has been published on the impact of social connections on various socio-economic indicators. A significant share of such studies is based on a series of articles by M. Bailey and co-authors, where the SCI (Social Connectedness Index) indicator is introduced, calculated on the basis of Facebook users active in the summer, 2019, using the formula

$$SCI_{ij} = \frac{FB\_Connections_{ij}}{FB\_Users_i \times FB\_Users_j} \tag{1}$$

Here $FB\_Connections_{ij}$ means the number of friendships in Facebook network between residents of regions i and j, and FB_Users_i and FB_Users_j stand for the number of different users of the network in regions i and j respectively. In Bailey et al. (2020) and other works, this indicator is interpreted as the probability of friendship between residents of the regions. There are several arguments in favor of such understanding of SCI_{ij}, including the maximum number of friends limit, the "symmetry" of friendships (as opposed to, for example, Twitter subscriptions), as well as the results of studies that reveal that this social network is used by 40% to 67% of Europeans, depending on the region, and often friendship on Facebook corresponds to real acquaintance.

Bailey et al. (2020) provides a comprehensive analysis of the determinants of social connectedness. The authors calculate pairwise SCIs between NUTS2 regions of the EU countries[1] and conduct an econometric analysis of these data. As shown in the article, social connectedness is largely determined by geographic distance and also increases remarkably if residents of the regions have a common religion or language and similar demographic indicators. It is also interesting that SCI turns out to be the higher, the greater the difference between the incomes of the regions is, which the authors associate with migration flows. In addition, the historical borders of countries and political unions, both relatively recent (for example, Czechoslovakia and Yugoslavia) and those that ceased to exist at the beginning of the 20th century (for example, Austria-Hungary), have a great influence on the value of social connectedness. Since a significant proportion of SCI variance is due to exogenous factors, this variable can be considered uncorrelated with random errors in regressions describing different preferences of Europeans.

[1] NUTS2 regions are territorial units of the EU countries with a population of 800 thousand to 3 million people, often coinciding with in-country administrative divisions.

Finally, the referenced article Bailey et al. (2020) estimates the relationship between SCI and political preferences. The authors use two proxies of euroscepticism: the results of the Eurobarometer survey[2], conducted in the fall of 2018, and the share of votes for anti-European parties in the parliamentary elections from 2000 to 2017. On the basis of SCI, the authors calculate the shares of social connections of the region's population with residents of other EU countries – $EU\_friends\_abroad$, and then with the help of OLS find out that higher shares of residents of other EU countries among friends on Facebook are associated with greater trust in the European Union and less euroscepticism. However, when country fixed effects are added to the regression for the second proxy, the estimate loses its statistical significance. It seems that this connection still exists and is strong enough, so it can be assumed that more advanced methods will help to reveal it more clearly.

An earlier article, Bailey et al. (2018), does a similar analysis of the determinants of social connectedness, but this time for the United States. The authors come to the conclusion that SCI is closely related to trade flows between states, a more detailed analysis of which is conducted in the work Bailey et al. (2021). When SCI is added to the standard regression of the gravity model of international trade, the geographic distances' coefficients lose their significance. It can be assumed that social connections are not only closely related to trade flows, but also accumulate all information about the distance between countries. Despite the fact that the authors do not claim that the identified relationship implies causal effect, at the end of the work a set of arguments in favor of this is presented. In other words, SCI_{ij} can be considered proportional to the probability that trade relations will be established between the regions i and j.

SCI is also used by many other researchers as a proxy for social connectedness. So, for example, results similar to the analysis of trade flows are obtained in Diemer and Regan (2020) when analyzing the citations of US patents. When both social connectedness and geographic distance are included in the regression, the latter one loses its significance, while SCI remains significant at 1% level even after the introduction of control variables reflecting professional connections.

The experience of the above-mentioned studies suggests that preferences in one region may influence preferences in others, thereby causing spatial autocorrelation. Spatial econometrics models that use the weighting matrix W can be used to analyze such data. Intuition lies in the fact that in the model the indicated autocorrelation can be associated with both the dependent variable y (in a sense, an analogue of the AR model for time series), and with random errors (analogue of the MA model).

[2] The respondents were asked: "I would like to ask you a question about how much trust you have in the European Union. Could you please tell me if you tend to trust it or tend not to trust it?".

In other words, there are two specifications to consider:

$$y = \rho W y + X\beta + \varepsilon \tag{2}$$

$$y = X\beta + \varepsilon, \qquad \varepsilon = \lambda W \varepsilon + \zeta \tag{3}$$

Where y is the dependent variable, W is the spatial connectedness matrix (it is assumed that $w_{ii} = 0$), and X is the regressor matrix. Following the terminology of LeSage and Pace (2009) and Croissant and Millo (2019), we will call the model (2) SAR (Spatial Autoregressive), and the (3) model SEM (Spatial Error Model). Such methods are successfully used to construct econometric models with a pronounced spatial structure. For example, Elhorst et al. (2019) provides an example of the application of such models to the analysis of oligopolized markets. Another example of successful use of spatial econometrics is the Ivanova (2019) study devoted to the ecological Kuznets curve (EKC), which describes the relationship between per capita GRP and emissions into the environment. By comparing three different ways of forming the W matrix (based on inverse geographic distances, their squares, and contiguity), the author confirms the key hypothesis about the U-shaped form of the EKC for Russian regions. To compare the weight matrices, the article uses Moran's spatial autocorrelation index, calculated by the formula

$$I(W) = \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\frac{1}{N}\sum\limits_{i=1}^{N}(y_i - \overline{y})^2 \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} w_{ij}} \tag{4}$$

Where y_i is the indicator under study. The index has asymptotically normal distribution, which allows it to be tested for significance.

Speaking directly about preferences, we can mention the work Hsieh and van Kippersluis (2019), devoted to the estimation of the influence of peer effects on smoking. The authors have data on students from several schools and use a modified SAR model to analyze how student's smoking is affected by the fact that his friends smoke. An innovation of this work is that the authors divide students into emotionally stable and unstable (depending on the results of the school psychological test) and suggest that the peer effect may differ depending on the emotional stability of both the student and his friends. Thus, four parameters ρ are estimated at once, each controlling for the influence of a different block of the matrix W. Then, the authors eliminate obvious endogeneity of the spatial connectedness matrix using the logit regression of the probability of friendship on a set of instrumental variables. The results show that emotionally unstable students are indeed significantly more likely to be influenced by friends than their more stable peers.

Following the motivation above, it can be assumed that spatial econometrics models will describe the formation of preferences among residents of European NUTS2 regions better than OLS regressions. It is reasonable to hypothesize that SAR model will be a better choice than SEM model, since it reflects the direct influence of observations on each other. Finally, in terms of political preferences, it seems that the $EU\_friends\_abroad$ regressor introduced in Bailey et al. (2020) should not lose significance in the new setting, since friendly connections within European countries should contribute to greater confidence in the EU and less euroscepticism. These questions lead to a final list of hypotheses tested:

1. spatial econometrics models are better suited for describing the preferences of EU residents than OLS;

2. spatial lag parameter ρ in SAR formulation helps to explain the variance of considered preferences and is statistically significant;

3. the preferences of Europeans in terms of their spatial structure are better described by social connectedness than by geographic distances;

4. in terms of political preferences, a greater number of social connections within the Eurozone contributes to greater confidence in the EU and less euroscepticism.

3 Main part

3.1 Data

To test hypotheses, two sets of cross-sectional data collected by Bailey et al. (2020) for NUTS2 European regions are primarily used, **trust_in_EU** and **anti_EU_votes**. These datasets contain a significant number of control variables that are used for further analysis. In addition to these data, we use the data of the seventh wave of the World Values Survey (2017 - 2020), aggregated by NUTS2 regions, on the preferences and views of Europeans, **world_values_survey**. For regressions, where the variable from this dataset acts as the dependent variable, additional controls are added to the equation, characterizing the structure of the region's population belonging to various associations. A complete list of variables is presented in Table 1 below.

Table 1: Variable description

Variable	Description
Dependent variables	
Trust_in_EU	Share of respondents in the region who answered "Tend to trust" to the question of trust in the European Union
Anti_EU_vote	Share of votes for anti-European parties in the region
Conf_in_civil_services	Aggregate variable of trust in government institutions
Homo_neighbours	Aggregate variable of attitude towards neighborhood with homosexuals
Religion	Aggregate variable of the importance of religion in life
Trust	Aggregate variable of trust in people
Variables of interest	
EU_friends_abroad	Share of social connections of the inhabitants of the region with Europeans from other countries
$W \times y$	Spatial lag of the dependent variable

Variable	Description
Control variables	
ED0_2	Share of people with less than secondary education in region's population
ED3_4	Share of people with secondary education in region's population
ED5_8	Share of people with higher education in region's population
median_age	Median age of region's population
Y0_9	Share of region's population aged 0-9 years
Y10_19	Share of region's population aged 10-19 years
Y20_29	Share of region's population aged 20-29 years
Y30_39	Share of region's population aged 30-39 years
Y40_49	Share of region's population aged 40-49 years
Y50_59	Share of region's population aged 50-59 years
Y60_69	Share of region's population aged 60-69 years
Y70_MAX	Share of region's population over the age of 70
Health	Respondents' subjective assessment of their state of health
share_other_EU	Share of residents of the region born in another European country
unemp_rate	Unemployment rate in the region in % by 2018
B	Share of residents of the region engaged in mining
C	Share of residents of the region employed in production
D	Share of residents of the region employed in electricity, gas and steam supply
E	Share of residents of the region engaged in water supply, sewerage, reclamation
F	Share of residents of the region engaged in construction
G	Share of residents of the region engaged in trade and transport repair
I	Share of residents of the region engaged in cargo transportation and storage

Variable	Description
Control variables	
J	Share of residents of the region employed in accommodation and food services
L	Share of residents of the region employed in the field of information technology
M	Share of residents of the region employed in real estate
N	Share of residents of the region engaged in scientific and technical activities
income_ thous	Average income of residents of the region, thousand euros
Member_ control_ 1	Aggregate variable of belonging to religious organizations
Member_ control_ 2	Aggregate variable of belonging to organizations related to education or the arts
Member_ control_ 3	Aggregate variable of affiliation with trade union organizations
Member_ control_ 4	Aggregate variable of political party membership
Member_ control_ 5	Aggregate variable of belonging to environmental organizations
Member_ control_ 6	Aggregate variable of belonging to professional associations
Member_ control_ 7	Aggregate variable of belonging to sports and entertainment organizations
Member_ control_ 8	Aggregate variable of consumer group affiliation
Member_ control_ 9	Aggregate variable of charity affiliation
Member_ control_ 11	Aggregate variable for support group membership

Since SAR and SEM models are used to test hypotheses, the question arises of creating a spatial matrix W. For each of the datasets used, we construct three spatial connectedness matrices. $W^{(d)}$ and $W^{(d^2)}$ are matrices based on the reciprocal of geographic distances and their squares, respectively, and $W^{(SCI)}$ is a matrix based on SCI. The diagonal of each of the matrices is filled with zeros, normalization is performed[3]. Using geographic distances between region centroids is one of the standard approaches to constructing a spatial matrix. This approach has a significant advantage: the spatial matrix is clearly exogenous. However Bailey et al. (2020) show that in some cases SCI is better at explaining spatial relationships in the data than distance-based measures. At the same time, as noted earlier, there are substantial grounds for considering $W^{(SCI)}$ to be exogenous. The differences between geographic distances and SCI are clearly visible in the graph of social connectedness of European countries (Fig. 1)[4].

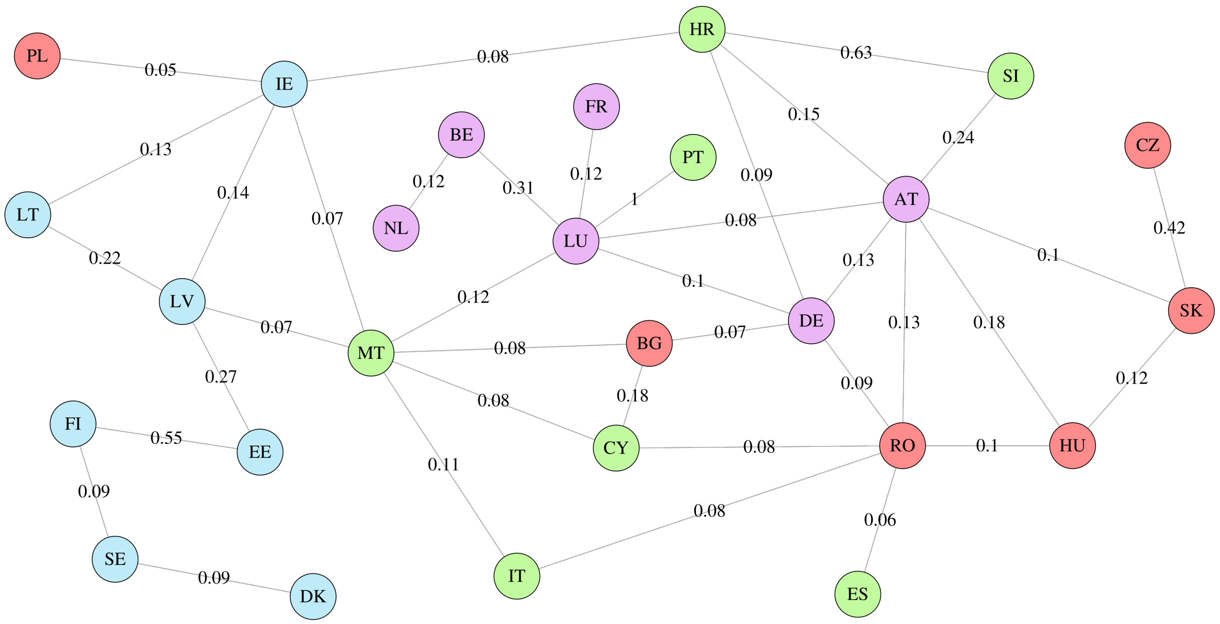

Figure 1: Graph of social connectedness of European countries

[3] Matrices are row-standardized. This approach is generally accepted in the literature on the topic, since it helps to avoid problems in estimating the model and interpret the spatial lag coefficient.

[4] The weakest connections are omitted for clarity.

Groups of countries are highlighted in color on the graph by geographical principle: blue – the countries of Northern Europe, purple – the countries of Western Europe, green – the countries of Southern Europe, red – the countries of Eastern Europe. The weights of the graph edges are the SCI between countries, normalized to the SCI between Luxembourg and Portugal. The graph illustrates important features of social connections in Europe. First, there is a clear correlation between SCI and geographic proximity: it is easy to see that, for example, in the Nordic countries, intra-group connections are quite strong, while external ties are few and relatively weak. Secondly, it is obvious that the SCI structure does not replicate the geographic one: there is a number of strong intergroup connections. Thus, we can conclude that it is reasonable to use $W^{(SCI)}$ for the analysis of preferences.

An important element of our empirical strategy is adding fixed effects to the model based on the intuition of Croissant and Millo (2019). The datasets used have a feature that makes it difficult to use traditional fixed effects: for a number of countries there are only one or two observations due to their small areas[5]. Thus, a situation arises when particular variables are added to the equation explaining a single observation, which hardly has a positive effect on the quality of the model. To avoid this problem and at the same time introduce fixed effects, we use the obtained by Bailey et al. (2020) clustering of European regions (Fig. 2)[6]. The number 19 was chosen for several reasons. First, such a number of clusters ensures a relatively even distribution of regions across clusters in the datasets used. Second, when more fixed effects are added, the estimates for variables of interest in the main models change insignificantly. Third, such a degree of sampling seems reasonable in terms of the number of dummy variables generated.

The cluster map also clearly shows the features of social connections. The contours of the clusters often coincide with the political borders of countries, however, there are clusters, parts of which do not share a common land border (for example, the cluster of Portugal and Switzerland). We believe that, due to these features, it is quite reasonable to add regional clusters fixed effects instead of country fixed effects to the model to explain the preferences.

[5] Also, for some countries, observations are available not for NUTS2 regions, but only for larger territorial units – NUTS1.

[6] The authors of the article use the hierarchical agglomerative linkage clustering algorithm.

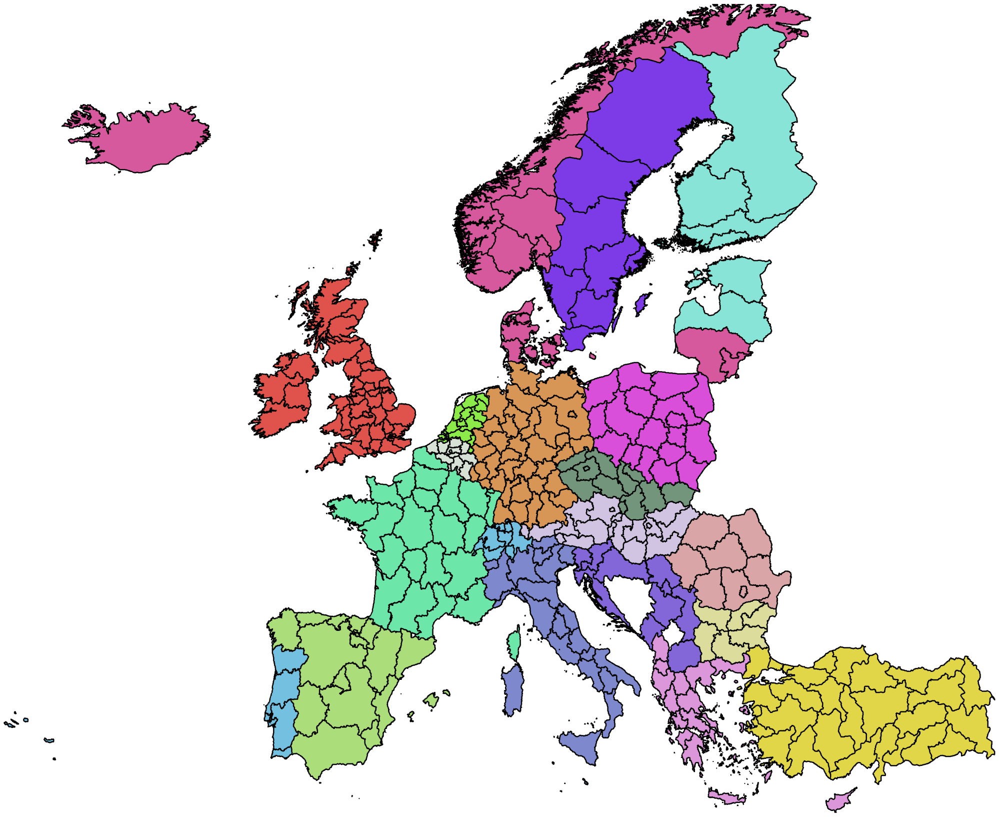

Figure 2: 19 clusters of European regions based on SCI

Proceeding on the description of data, it is pertinent to provide a table with Moran's spatial correlation indices for main dependent variables[7].

Table 2: Spatial correlation indices

Variable	Dataset	N	$W^{(d)}$	$W^{(d^2)}$	$W^{(SCI)}$
Trust_in_EU	**trust_in_EU**	198	0,096	0,292	0,482
Anti_EU_vote	**anti_EU_votes**	215	0,062	0,201	0,459
Conf_in_civil_services	**world_values_survey**	190	0,117	0,321	0,443
Homo_neighbours	**world_values_survey**	190	0,259	0,549	0,553
Religion	**world_values_survey**	190	0,216	0,505	0,623
Trust	**world_values_survey**	190	0,217	0,510	0,643

[7] N is the number of observations in the dataset.

All indices are statistically significant at 1% level. This allows us to speak about the spatial correlation of the studied preferences and provides evidence in favor of the rationality of using spatial econometrics models. Note also that the Moran's index for $W^{(SCI)}$ is usually larger than the indices for distance-based matrices. This result can be interpreted as an argument in favor of hypothesis 3 that SCI better reflects the spatial structure of preferences than geographic distances.

Since the data has a pronounced spatial structure, it makes sense to visualize it using a map. Thus, Figure 3 shows clusters of regions with a similar level of EU trust. The contrast between the UK and the countries of Northern and Central Europe is especially evident.

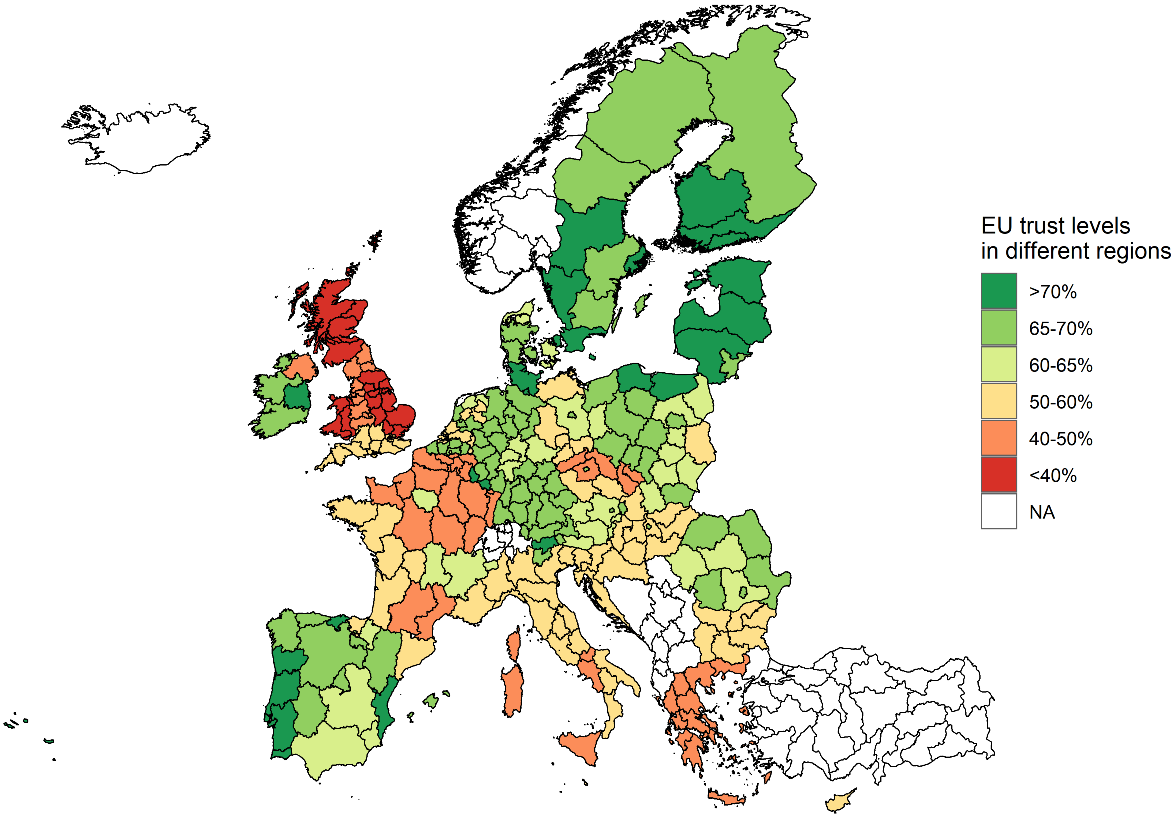

Figure 3: EU confidence levels in regions

If we turn to another proxy of euroscepticism, the share of votes for the parties of eurosceptics, a similar picture can be seen (Fig. 4).

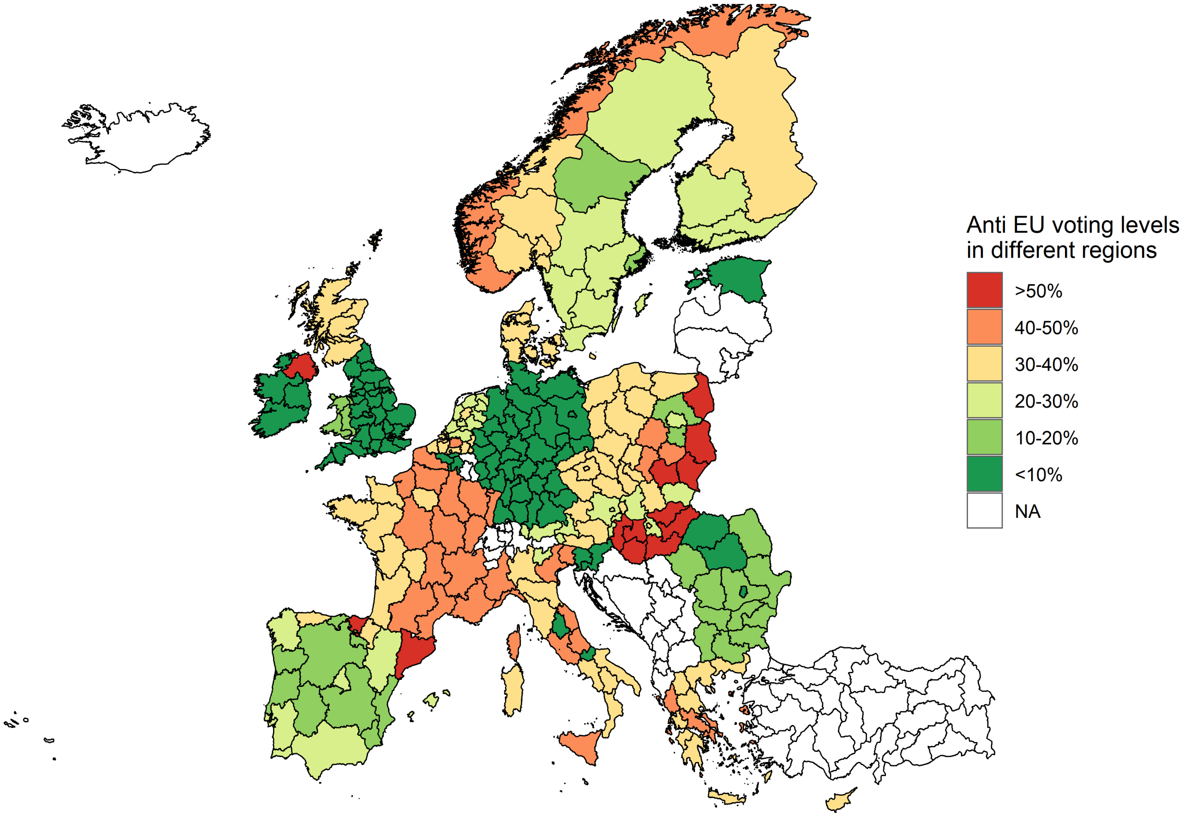

Figure 4: Shares of votes for eurosceptic parties in the regions

It is noticeable that there is a negative correlation between the two indicators. At the same time, the indicators characterize qualitatively different expressions of euroscepticism: the EU confidence levels are estimated on the basis of aggregate survey data, whereas the share of votes for eurosceptic parties is measured directly and is a good estimate of the percentage of active opponents of the EU. Thus, a joint analysis of these indicators is particularly interesting, since in the case of obtaining similar results, one can talk about their stability and make more confident conclusions about the degree of various factors' influence on Euroscepticism.

Spatial models were estimated using the quasi-maximum likelihood (QML) method. Data manipulation, model estimation and plotting are done in the R package, R Core Team (2020), using the splm library described in Millo and Piras (2012).

3.2 Estimation results

All regression models were considered in two settings: short and long. The evaluated equations have the same structure, but the short version differs from the long one by the absence of some control variables, which in most cases turned out to be insignificant[8] The desire to reduce the number of regressors is due to the rather small number of observations per estimated parameter. Hereinafter, the estimation results for long formulations are given, since it turns out that they are robust to removal of observations from the sample (see the section on robustness). This alleviates concerns about models being overly parameterized.

The model comparison scheme recommended in Croissant and Millo (2019) and used in Ivanova (2019) is as follows: at the first stage of testing, two Lagrange multipliers tests for the combined model are performed. The first of them tests the hypothesis $\lambda = 0$ under the condition $\rho = 0$ (error-test), and the second – in a sense, the symmetric hypothesis $\rho = 0$ under the condition $\lambda = 0$ (lag test). If exactly one of the above tests accepts the null hypothesis, then the corresponding parameter is set equal to zero. If both tests reject the null hypothesis, two robust tests are performed with the hypotheses $\rho = 0$ under the condition $\lambda \neq 0$ and $\lambda = 0$ under the condition $\rho \neq 0$ and the decision on the choice of the model is made on their basis (in the case of repeated rejection of both null hypotheses, the selection is made according to the smallest p-value).

As for the goodness of fit measure used to compare the models, the Akaike information criterion (AIC) and Nagelkerke pseudo-R^2 are used, the latter calculated by the formula

$$R^2 = \frac{1 - \left(\frac{L_0}{L}\right)^{2/N}}{1 - L_0^{2/N}} \tag{5}$$

Here L and L_0 are the values of the likelihood function when estimating the full model and the intercept regression model, respectively, and N is the sample size. This indicator coincides with the standard R^2 for OLS estimation, so its use for comparison purposes is quite justified.

For the **trust_in_EU** dataset, LM tests (Table 3) indicate that SAR model is preferable for matrices $W^{(SCI)}$ and $W^{(d^2)}$. As for $W^{(d)}$, spatial models with it are inferior to OLS models.

[8] Long regression includes all control variables, except for *share_other_EU*, as well as regressors that cause pure multicollinearity. The short regression from educational control includes only $ED5\_8$, from industry controls only E, G and L. Also in the short regression *unemp_rate* is omited.

Table 3: LM tests results for *Trust_ in_ EU*

	$W^{(SCI)}$	$W^{(d)}$	$W^{(d^2)}$
	$p - value$	$p - value$	$p - value$
error	0,259	0,484	0,831
lag	0,001	0,416	0,056
error (robust)	0,005	0,017	0,069
lag (robust)	0,000	0,016	0,009

For **anti_EU_votes** tests (Table 4) indicate the preference of the spatial model only for the matrix $W^{(SCI)}$, and in the SAR version. The rest of the regressions again turn out to be worse than OLS and therefore are not shown in Table 5, which presents the estimation results.

Table 4: LM tests results for *Anti_ EU_ vote*

	$W^{(SCI)}$	$W^{(d)}$	$W^{(d^2)}$
	$p - value$	$p - value$	$p - value$
error	0,386	0,787	0,197
lag	0,025	0,961	0,337
error (robust)	0,010	0,665	0,387
lag (robust)	0,001	0,732	0,940

The fact that the spatial model turned out to be better than OLS for at least one of the matrices under consideration confirms the first hypothesis put forward for both **trust_in_EU** and **anti_EU_votes** datasets: the SAR model can indeed better describe the political preferences of Europeans. It should be noted that although robust standard errors were used, the Breusch – Pagan test for heteroscedasticity did not reject the null hypothesis for both data sets. This allows us not to doubt the effectiveness of the estimates obtained and their significance. The best models for political preference are presented in more detail in Table 5.

Table 5: Analysis of Euroscepticism

	Trust_in_EU		*Anti_EU_vote*
	$W^{(d^2)}$	$W^{(SCI)}$	$W^{(SCI)}$
EU_friends_abroad	0,393*** (0,089)	0,380*** (0,086)	-0,098 (0,202)
ρ	0,254* (0,138)	0,469*** (0,106)	0,363*** (0,122)
Cluster Fixed Effects	Y	Y	Y
Socioeconomic controls	Y	Y	Y
Membership controls	-	-	-
Number of observations	198	198	215
R^2	0,798	0,807	0,700
AIC	-592,1	-601,5	-339,1
Note:	$^{*}p < 0,1$; $^{**}p < 0,05$; $^{***}p < 0,01$		

The spatial lag ρ turns out to be significant in each of the selected settings. This confirms the second of the hypotheses put forward. Finally, since models with matrices based on geographic distances lose even to OLS for the **anti_EU_votes** dataset, and for the **trust_in_EU** data they have lower R^2 and higher Akaike criterion, the third hypothesis can also be considered confirmed. Social connections actually explain political preferences better than geographic distance.

The results of the analysis correspond to the conclusions of Bailey et al. (2020): for the dataset **trust_in_EU**, the regressor *EU_friends_abroad* turns out to be statistically significant. Moreover, the numerical value of its coefficient, estimated in Bailey et al. (2020), is 0.326, while in the SAR model this coefficient equals 0.380. There is no such similarity in the estimates of the coefficient for *EU_friends_abroad* for the **anti_EU_votes** dataset, but this is not surprising since it is insignificant both in the SAR setting and in the original work of Bailey et al. (2020). It should be noted here that the interpretation of the coefficients in spatial models differs from the standard interpretation in OLS, therefore the difference (albeit small) is quite natural and does not indicate the omission of significant variables or any other form of endogeneity.

Due to the specification of the SAR model, the coefficients of the variables on the right side cannot be interpreted in the same way as in standard linear regression. The result of changing any regressor will be decomposed into two effects – direct and indirect. The direct effect is the effect of the change in the regressor in the region i on the dependent variable in the same region, while the indirect effect is the effect of this change on the other region j. This can be illustrated by calculations from LeSage and Pace (2009):

$$\begin{gathered} y = \rho W y + X\beta + \varepsilon \quad \Rightarrow \quad (I_n - \rho W)y = X\beta + \varepsilon \\ \frac{\partial y_i}{\partial x_{jr}} = [(I_n - \rho W)^{-1} I_n \beta_r]_{ij} \end{gathered} \tag{6}$$

Here x_{jr} is an observation of the r-th regressor in the region j. The formula shows that the direct and indirect effects are really different, that is, $\frac{\partial y_i}{\partial x_{jr}}$ for different j, generally speaking, do not coincide. For a quantitative interpretation, following the example of Ivanova (2019), we use the average of these effects. In particular, for the variable $EU\_friends\_abroad$ in the **trust_in_EU** dataset, the direct effect is 0.390, the indirect effect is 0.325, and the average is 0.357. Thus, an increase in the share of European friends from other countries by 1% will be associated with an increase in confidence in the European Union by 0.39% in this region and by 0.33% in other regions.

Everything said about long and short specifications is true for regressions with dependent variables from the World Values Survey[9]. Almost all the data here are homoscedastic, according to Breusch-Pagan test[10]. The specification tests for the *Religion* variable are presented in Table 6 below.

Table 6: LM tests results for *Religion*

	$W^{(SCI)}$	$W^{(d)}$	$W^{(d^2)}$
	$p - value$	$p - value$	$p - value$
error	0,784	0,377	0,834
lag	0,009	0,213	0,073
error (robust)	0,000	0,027	0,096
lag (robust)	0,000	0,017	0,015

[9] Also long regressions additionally include all control variables related to participation in organizations, and short ones only include $Member\_control\_2$, $Member\_control\_4$ and $Member\_control\_7$.

[10] Except for the *Trust* variable, but robust standard errors should account for this.

These results are similar to the test results for the **trust_in_EU** dataset and indicate that SAR models with matrices $W^{(SCI)}$ and $W^{(d^2)}$ are the most preferable. For the variable $Conf\_in\_civil\_services$, surprisingly better than OLS estimates can be obtained only from the SAR model with $W^{(d^2)}$ (see p-values of tests in Table 7).

Table 7: LM tests results for $Conf\_in\_civil\_services$

	$W^{(SCI)}$	$W^{(d)}$	$W^{(d^2)}$
	$p-value$	$p-value$	$p-value$
error	0,868	0,363	0,752
lag	0,439	0,425	0,038
error (robust)	0,014	0,001	0,000
lag (robust)	0,010	0,001	0,000

The final comparison of the best models for the variables $Conf\_in\_civil\_services$ and *Religion* is presented in Table 8.

Table 8: Analysis of attitudes towards religion and trust in government institutions

	Religion		$Conf\_in\_civil\_services$
	$W^{(d^2)}$	$W^{(SCI)}$	$W^{(d^2)}$
$EU\_friends\_abroad$	-0,103 (0,140)	-0,013 (0,138)	0,340** (0,135)
ρ	0,254** (0,109)	0,416*** (0,109)	0,379*** (0,139)
Cluster Fixed Effects	Y	Y	Y
Socioeconomic controls	Y	Y	Y
Membership controls	Y	Y	Y
Number of observations	190	190	190
R^2	0,876	0,879	0,660
AIC	-531,1	-535,8	-543,7
Note:	$^{*}p < 0,1$; $^{**}p < 0,05$; $^{***}p < 0,01$		

Speaking about the results of the *Religion* study, it can be noted that again SAR models turn out to be better than OLS, and the coefficient ρ is significant at 1% level. Moreover, R^2 in the model with $W^{(SCI)}$ is higher. All three main hypotheses are confirmed for this variable. The same cannot be said about $Conf\_in\_civil\_services$: here the last hypothesis cannot be accepted, since the best spatial model is the SAR model with $W^{(d^2)}$ matrix. It turns out that geography explains the variation of attitudes towards state institutions better than social connectedness.

Interesting results are obtained by analysing the $Trust$ variable. LM specification tests for all three matrices indicate the superiority of the SAR model and are therefore omitted. In this case, the coefficients ρ turn out to be significant at 5% level and, interestingly, negative (see the estimates of the models in Table 9). The interpretation of this could be as follows: less gullible and more suspicious people tend to maintain social connections only with more gullible and naive people. This is quite natural, because it seems that a community consisting entirely of people who do not trust each other cannot be sustainable. It is also interesting that the $EU\_friends\_abroad$ variable has a significant positive effect on trust. It turns out that more trusting people have a higher proportion of friends among residents of other European regions.

Table 9: Analysis of trust in people

	Trust		
	$W^{(d)}$	$W^{(d^2)}$	$W^{(SCI)}$
EU_friends_ abroad	0,319*** (0,111)	0,295*** (0,110)	0,375*** (0,105)
ρ	-1,005** (0,399)	-0,308** (0,147)	-0,687*** (0,138)
Cluster Fixed Effects	Y	Y	Y
Socioeconomic controls	Y	Y	Y
Membership controls	Y	Y	Y
Number of observations	190	190	190
R^2	0,886	0,885	0,895
AIC	-625,2	-623,2	-639,4
Note:	$^{*}p < 0,1$; $^{**}p < 0,05$; $^{***}p < 0,01$		

As for $Homo\_neighbors$, spatial models again prevail. However, here for the first time tests point to the SEM model, and for two matrices at once – $W^{(SCI)}$ and $W^{(d^2)}$ (Table 10). As an interpretation of this choice, Croissant and Millo (2019) suggests omission of a factor that explains the spatial autocorrelation of the dependent variable. Model estimates are given in Table 11.

Table 10: LM tests results for *Homo_ neighbours*

	$W^{(SCI)}$	$W^{(d)}$	$W^{(d^2)}$
	$p - value$	$p - value$	$p - value$
error	0,014	0,097	0,070
lag	0,826	0,044	0,086
error (robust)	0,000	0,000	0,000
lag (robust)	0,003	0,000	0,000

Table 11: Analysis of attitudes towards homosexual neighbours

	Homo_ neighbours		
	$W^{(d)}$	$W^{(d^2)}$	$W^{(SCI)}$
EU_friends_ abroad	0,478** (0,237)	0,549** (0,230)	0,483** (0,225)
ρ	0,658*** (0,200)	-	-
λ	-	-0,796*** (0,207)	-0,964*** (0,215)
Cluster Fixed Effects	Y	Y	Y
Socioeconomic controls	Y	Y	Y
Membership controls	Y	Y	Y
Number of observations	190	190	190
R^2	0,861	0,864	0,869
AIC	-330,1	-334,2	-341,2
Note:	$^{*}p < 0,1$; $^{**}p < 0,05$; $^{***}p < 0,01$		

3.3 Robustness check

To check the robustness of the results, two series of additional tests were performed. The first series of tests was devoted to the resistance to the part of the sample removal. For each setting, 10 random observations were removed from the full dataset, and the best spatial model was estimated on the remaining subsample. For each dataset, 1000 iterations were performed, and then the empirical distribution density $\hat{\rho}$ or $\hat{\lambda}$ (depending on the model) was compiled. Examples of the obtained densities for **trust_in_EU** and **anti_EU_votes** are shown in Figures 5 and 6 (the bold line marks the estimate of the ρ parameter for the full sample, and the dashed lines – 95% confidence intervals). Similar plots for other variables can be found in Appendix A.

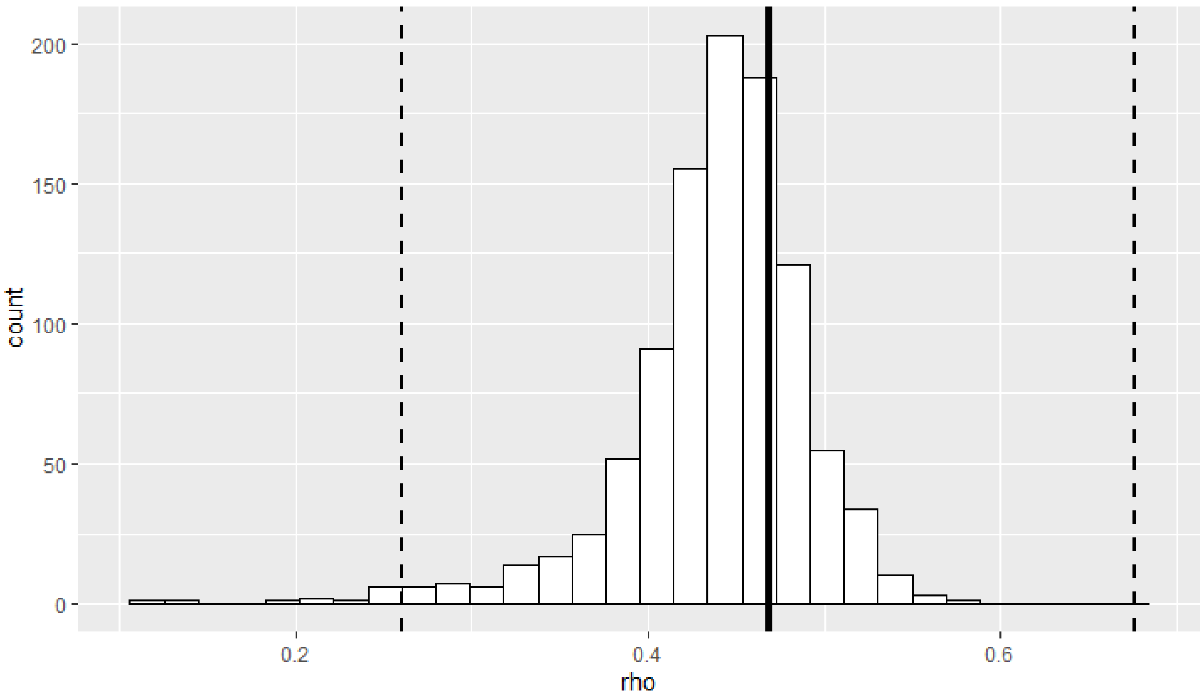

Figure 5: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable *Trust_in_EU*

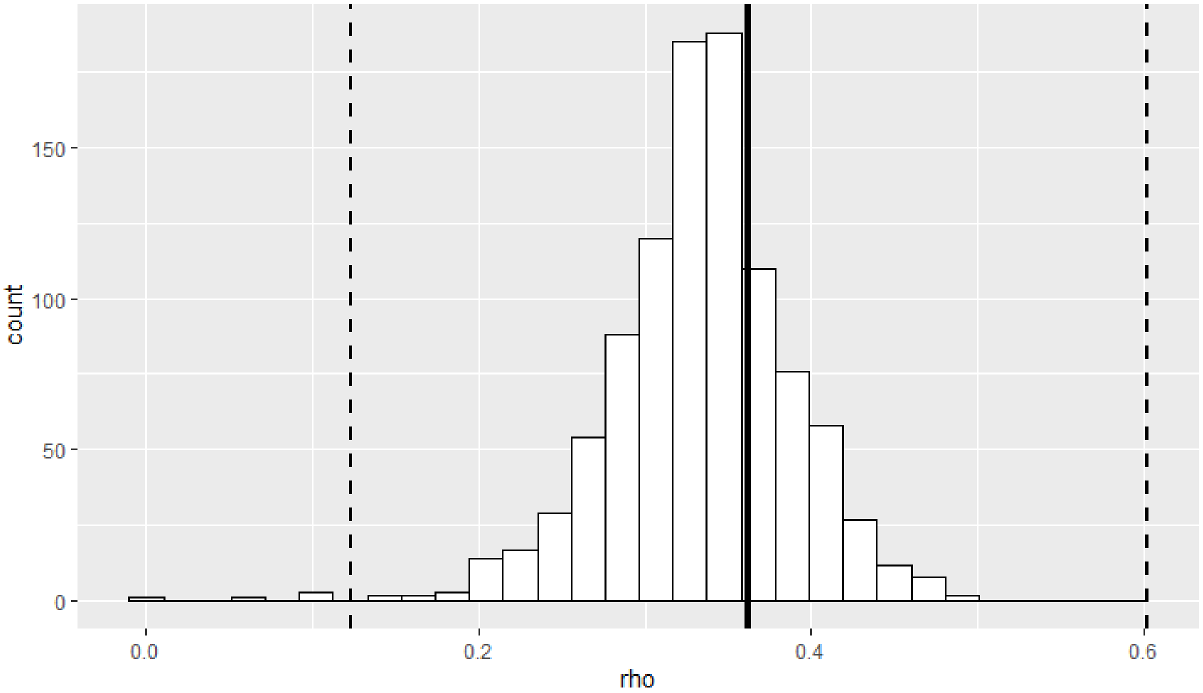

Figure 6: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable $Anti\_EU\_vote$

Then, for all models, a placebo test was carried out, which consisted in checking the significance of the ρ coefficient for the SAR model (and, accordingly, the λ coefficient for the SEM model) using a random matrix W. Randomness is understood in the sense of the uniformity of the distribution of its elements before the normalization. In other words, a model was built where $w_{ii} = 0$ and for $i \neq j$

$$w_{ij} = \frac{\xi_{ij}}{\sum\limits_{j \neq i} \xi_{ij}}, \qquad \xi_{ij} \sim U[0;1] \tag{7}$$

For all the above regressions, the parameters ρ and λ with a random matrix W are insignificant and the models built on their basis turn out to be worse than OLS.

3.4 Links

Bailey et al. (2020) follow the principles of reproducible scientific research and publish their code in a ***github-repository***, which allows us to check and supplement their results. We consider this approach best practice, so we also place the code written by us for this study in a ***repository***.

4 Conclusion

Summing up the research conducted, it should be noted that all the proposed hypotheses were accepted to a certain extent. First, the models of spatial econometrics, as shown by LM tests, for all considered preferences turned out to be better than the OLS for at least one of the three proposed weights matrices. It turns out that the studied dependent variables are indeed characterized by noticeable spatial autocorrelation, ignoring which leads to inefficiency of the estimates.

Secondly, in all constructed regressions, except for the setting with matrices $W^{(SCI)}$ and $W^{(d^2)}$ for the variable $Homo\_neighbors$, the spatial lag parameter ρ turned out to be statistically significant. Here we can conclude that the SAR model is much better suited to describe the preferences of European countries' residents than the alternative SEM model.

Thirdly, as the R^2 coefficients calculated for spatial models and the Akaike AIC criteria indicate that $W^{(SCI)}$ matrix, reflecting the "proximity" of regions in the sense of social connectedness, in most cases is better suited for describing the considered preferences than weight matrices based on geographic distances.

Finally, the regressions for the variables related to EU confidence and Euroscepticism are in good agreement with the results obtained in Bailey et al. (2020). Moreover, the estimates of the coefficients for the variable $EU\_friends\_abroad$ in those equations where this regressor is significant are quite close in numerical value. This suggests that the share of friends living in other European countries have a similar effect on people's preferences and are therefore an important factor in their analysis.

This work uses both methods that are relatively new for the subject area – models of spatial econometrics – and recent data on SCI. Together, this combination allows to achieve interesting results from an academic point of view and certainly has great potential for further development.

A Appendix: figures

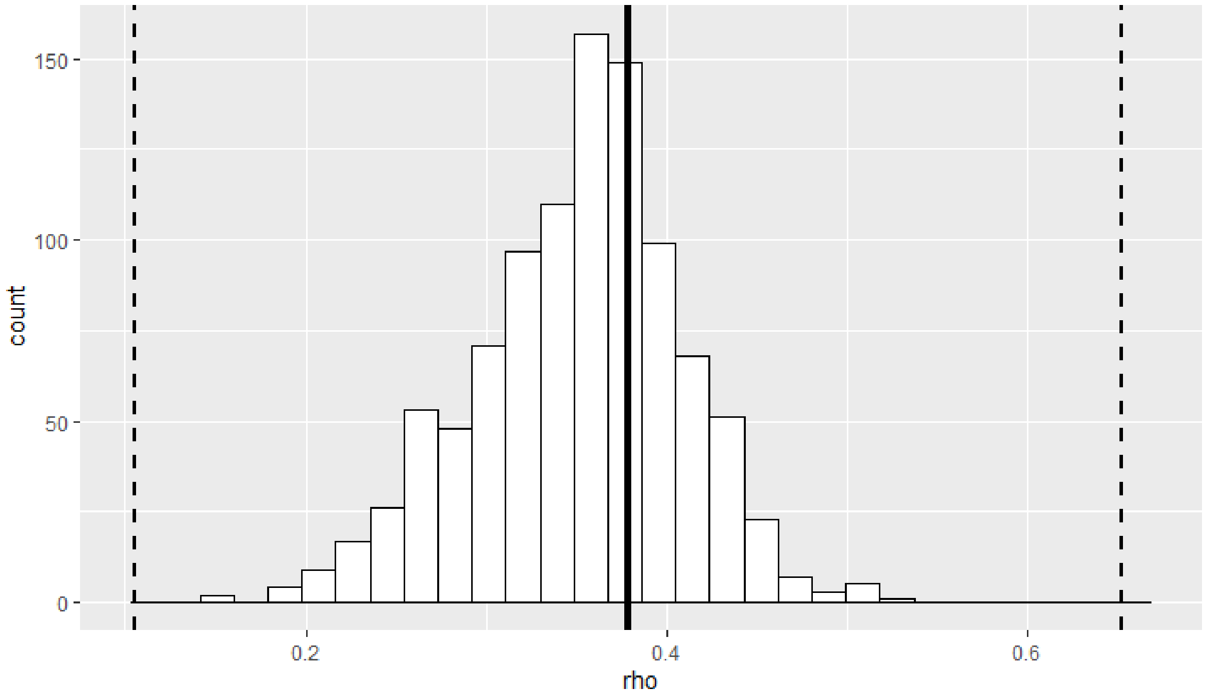

Figure 7: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable $Conf\_in\_civil\_services$

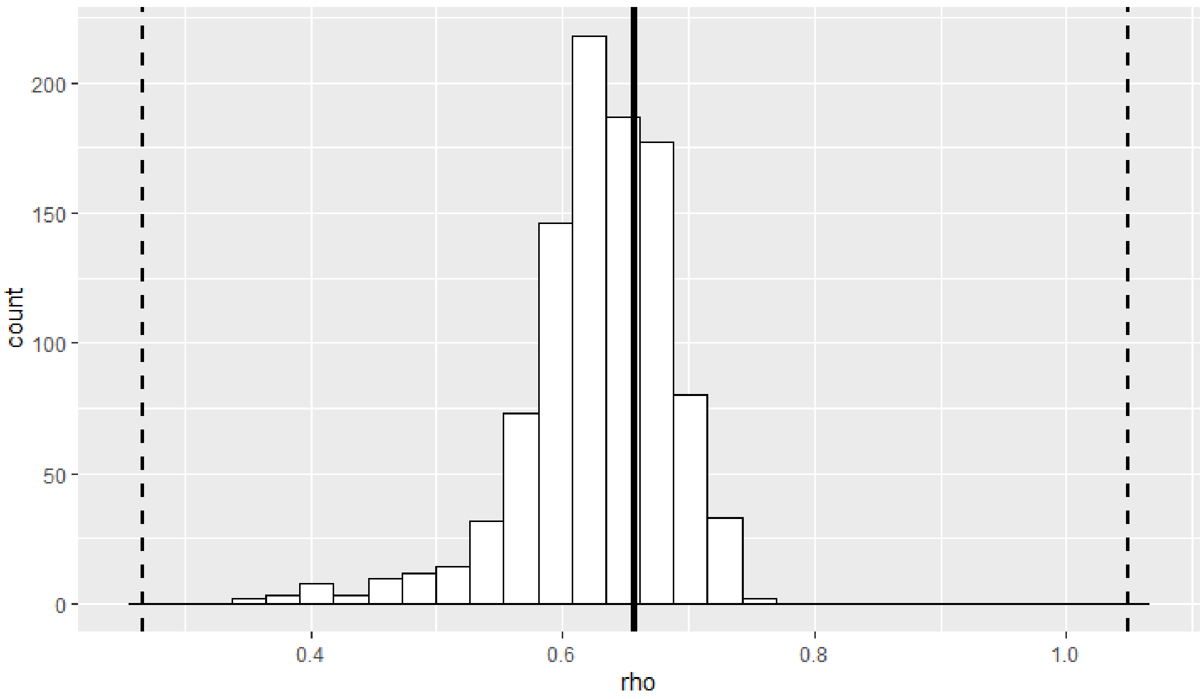

Figure 8: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable *Homo_ neighbours*

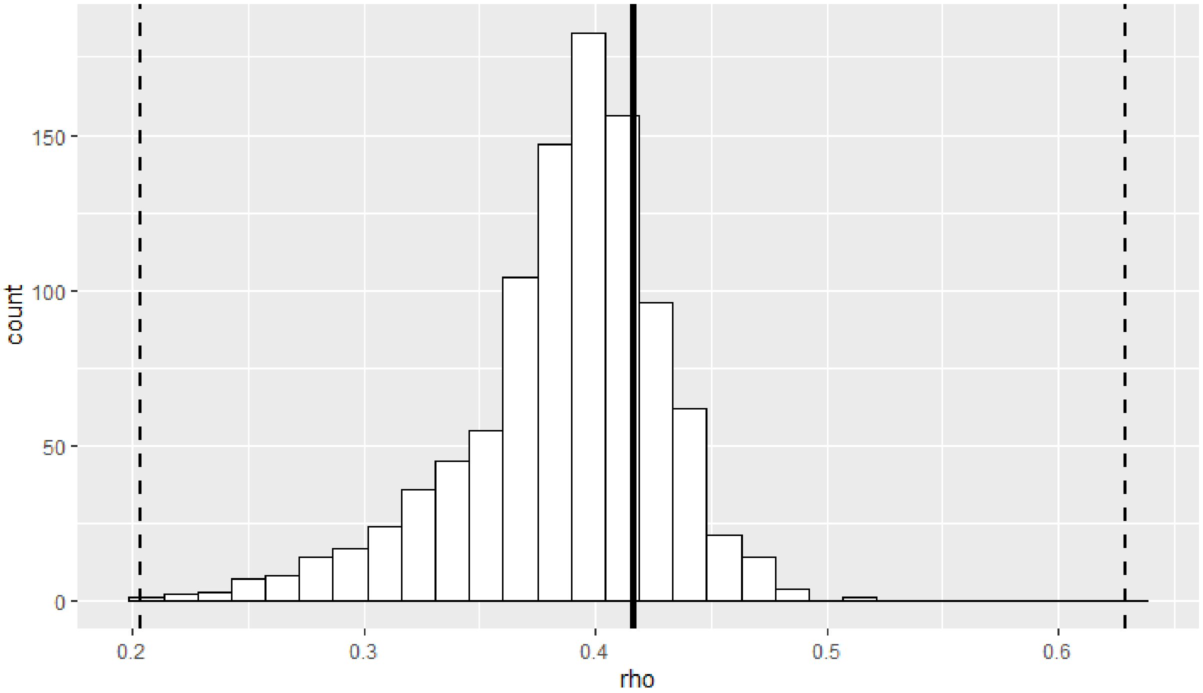

Figure 9: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable *Religion*

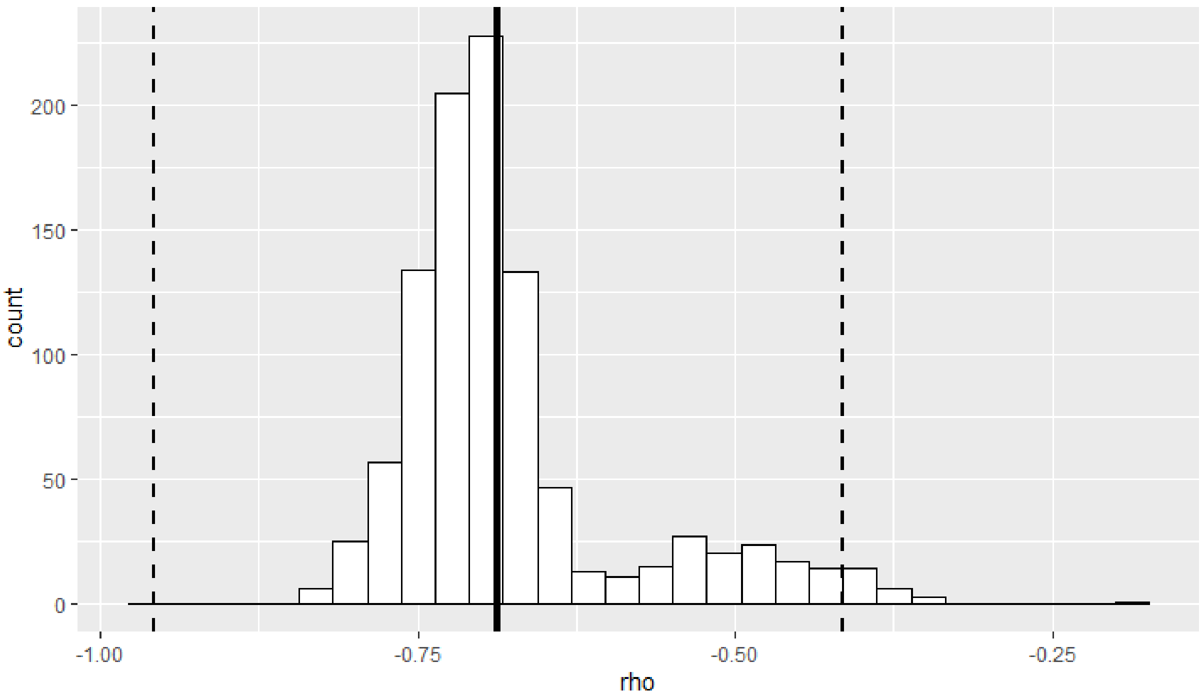

Figure 10: Empirical distribution density of $\hat{\rho}$ in robustness tests with respect to the variable *Trust*

Bibliography

Bailey, M., Cao, R., T Kuchler, J. S., and Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(2):259–280.

Bailey, M., Gupta, A., Hillenbrad, S., Kuchler, T., Richmond, R., and Stroebel, J. (2021). International trade and social connectedness. *Journal of International Economics*, 129(103418).

Bailey, M., Johnston, D., Kuchler, T., Russel, D., State, B., and Stroebel, J. (2020). The determinants of social connectedness in Europe. *Social Informatics. SocInfo 2020. Lecture Notes in Computer Science*, 12467(8310):1–14.

Croissant, Y. and Millo, G. (2019). *Panel Data Econometrics with R.* John Wiley Sons Ltd.

Diemer, A. and Regan, T. (2020). No inventor is an island: social connectedness and the geography of knowledge flows in the US. *CEP Discussion Papers, Centre for Economic Performance, LSE.*, (1731).

Elhorst, P., Abreu, M., Amaral, P., Bhattacharjee, A., Corrado, L., Doran, J., Fuerst, F., Gallo, J. L., McCann, P., Monastiriotis, V., Quatraro, F., and Yu, J. (2019). Raising the bar (11). *Spacial Economic Analysis*, 14(2):129–132.

Hsieh, C. and van Kippersluis, H. (2019). Smoking initiation: Peers and personality. *Quantitative Economics*, 9(2):825–863.

Ivanova, V. (2019). GRP and environmental pollution in Russian regions: spatial econometric analysis. *Quantile*, (14):53–62.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics.* Chapman Hall/CRC.

Millo, G. and Piras, G. (2012). splm: Spatial panel data models in R. *Journal of Statistical Software*, 47(1):1–38.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.