THE EFFECT OF EU-FUNDS ON CORRUPTION RISK

A Cross-Regional Quantitative Analysis of Public Procurement Data in the European Union (2014-2020)

By
Alexandra Fehér

Submitted to Central European University - Private University
Department of Public Policy

In partial fulfilment of the requirements for the degree of Master of Arts in Public Policy

Supervisor: Professor Mihály Fazekas

Vienna, Austria
2023

AUTHOR'S DECLARATION

I, the undersigned, Alexandra Fehér, candidate for Master of Arts in Public Policy, declare herewith that the present thesis is exclusively my own work, based on my research.

All sources have been properly credited in the text, notes, and the bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. Furthermore, I declare that no part of this thesis has been generated using artificial intelligence (ChatGPT).

I also declare that no part of the thesis has been submitted in this form as coursework for credits or to another institution of higher education for an academic degree.

Date: Vienna, 02. June 2023

Name: Alexandra Fehér

Signature

ABSTRACT

There has been a sustained interest among Public Policy and Big Data scholars regarding the measurement of corruption. So far, no consensus was established, however, as new corruption detection tools emerged, the scientific discourse rekindled. This thesis wishes to contribute to the new wave of corruption research by using public procurement data from opentender.eu. The thesis creates a novel Corruption Risk Index measuring the lack of competition, to uncover whether EU-funded tenders have a higher risk of corruption than national tenders. The index comprises three integrity indicators: single bid, procedure type and whether a call for tender has been published. The research utilizes Random Forest from the field of Big Data, Coarsened Exact Matching, and regression analysis to estimate the causal impact. The result of the analysis suggests that EU-funding has a statistically significant effect on corruption risk. However, further research is needed to assess the direction of the impact.

ACKNOWLEDGEMENTS

I'm very thankful for Viktor Gyetvai, a certified data analyst who helped immensely improve my analytical models. His advice guided me to a better Corruption Risk Index and drew my attention to the weaknesses and sensitivity of the applied matching process.

I wish to express my gratitude towards Sanjay Kumar, my academic writing instructor, who helped improve the style and language of the thesis. His kind words and early deadlines helped me keep my work on track. I'm also grateful for his support throughout the whole academic year, Sanjay is an excellent teacher.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ACA	Anti-Corruption Agency
CEE	Central Eastern Europe
CEM	Coarsened Exact Matching
CoC	Control of Corruption
CPI	Corruption Perception Index
CPV	Common Procurement Vocabulary
CRI	Corruption Risk Index
EPPO	European Public Prosecutor's Office
ESS	Effective Sample Size
EU	European Union
GRECO	Group of States Against Corruption
IMF	International Monetary Fund
MSE	Mean Squared Error
NE	Northern Europe
NUTS	Nomenclature of Territorial Units for Statistics
OLAF	European Anti-Fraud Office
PSM	Propensity Score Matching
RMSE	Root Mean Squared Error
SE	Southern Europe
TED	Tenders Electronics Daily
TI	Transparency International
WB	World Bank
WE	Western Europe

INTRODUCTION

Corruption is detrimental to society. It slows down economic growth, erodes trust in political institutions and in politicians (Lima et al., 2020). It only favors a small group of elites, who get rich from public funds (Thompson, 2005). The problem of corruption is still unresolved, even though it is as old as mankind. A big part of the problem is that it remains hidden, and corrupt politicians and businessmen are not held accountable. Moreover, there are no sufficient means by which scholars could detect or even forecast the presence of corruption. This puzzle still arouses the curiosity of Public Policy scholars, which resulted in the emergence of new data-based approaches (Petheram et al., 2019; Aldana et al, 2022). This research wishes to contribute to the new wave of corruption research by using quantitative methods like big data and matching. Within this realm, the scientific contribution is the newly created Corruption Risk Index (CRI) that attempts to measure the likelihood of corruption. The proposed CRI focuses on European countries and European regions (see Figure 1).

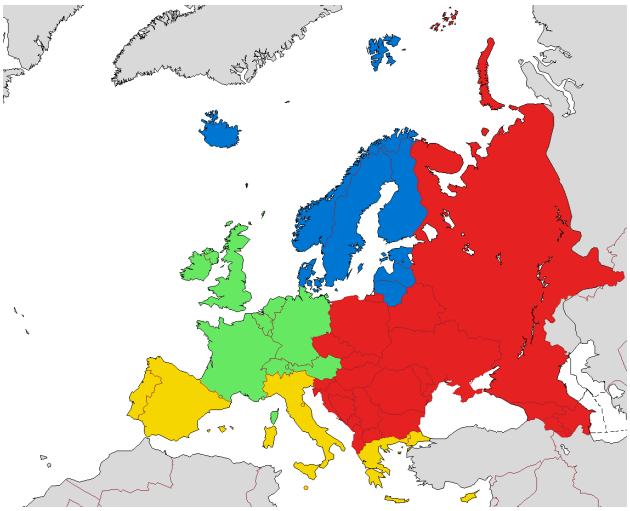

Figure 1: Classification of European countries, EuroVoc 2014

Central Eastern Europe (CEE): Bulgaria, Croatia, Czech Republic, Hungary, Poland, Romania, Slovenia, Slovakia (8 countries, in red)

Northern Europe: Denmark, Estonia, Finland, Latvia, Lithuania, Sweden (6 countries, in blue)

Southern Europe: Cyprus, Greece, Italy, Malta, Portugal, Spain (6 countries, in yellow)

Western Europe: Austria, Belgium, France, Germany, Ireland, Luxembourg, Netherlands (7 countries, in green)

The primary objective of this research is to devise a novel Corruption Risk Index that can predict corruption with a high external validity at both national and European levels. With the help of CRI – that combines three integrity indicators: single bid, procedure type and call for tender publication – the study seeks to answer the question whether EU-funded tenders have a higher risk of corruption than national ones. The research question contains at least two implicit

assumptions. One of them suggests that there is a notable difference in the level of corruption between European and domestic tenders, and the other assumes that EU-funded tenders are more prone to corruption. The research question was inspired by the study of Fazekas et al. (2014), who analyzed three Central and Eastern European (CEE) countries, namely Hungary, Slovakia and the Czech Republic in the time period of 2009-2012. Their preliminary findings indicate that EU-funds positively impact institutionalized grand corruption. To explain the mechanism behind this phenomenon, the authors provide three reasons, which will also serve as the basis of the thesis' theoretical assumptions.

I. Eastern Europe is historically less developed than Western Europe, so it received greater proportions of discretionary funding from the EU Cohesion and Structural Funds. These funds are not earmarked for fixed expenses such as debt payments or pensions, and can be used more freely on various projects, including education, healthcare, national defense, and scientific research. However, as these funds are not linked to a specific project, there is a higher risk of corruption, as argued by Mauro (1998) and Tanzi and Davoodi (2001).

II. Large-scale EU funding projects are perceived by corrupt politicians as a prime opportunity for rent-seeking. Such projects contribute to the pool of public resources allocated on a particularistic basis, which can become a significant source of rent extraction. Corrupt elites can maintain their power and strengthen their political and business standing through the supplementary resources obtained from EU funds (Mungiu-Pippidi, 2013).

III. EU funding can create a discrepancy between government revenues and services, similar to external funding in developing countries. The availability of more money than the country's economic output can lead to skilled bureaucrats being drawn away from domestic institutions to aid organizations. Moreover, the extra administration associated with external funding may reduce the quality of governmental services, leading to weakened administrative capacity and making rent extraction easier (Fazekas et al., 2014).

Furthermore, the EU's control mechanisms to prevent corruption are deemed inadequate (Martin-Russu, 2022, De Sousa, 2010, Batory, 2018). Despite the establishment of several anti-corruption authorities in 1999, the situation has not improved significantly.

Based on these theoretical considerations, the null and alternative hypothesis are as follows:

H_0 = EU-funds do not affect corruption risk at all *(expected in NE and WE)*
H_1 = EU-funds increase corruption risk *(expected in CEE and SE)*
H_2 = EU-funds decrease corruption risk *(theoretical option)*

The theoretical assumptions and the literature review have a strong focus on CEE. The reason for the academic interest is fourfold. First – as indicated in assumption one – CEE receives far the most EU-funding. As a result, the dataset opentender.eu has a large sample of EU and non-EU funded tenders, which makes the comparison more reliable. Second, based on Corruption Perception Indexes (CPIs) and scholarly literature (Batory, 2018, Fazekas and Kocsis, 2017), CEE has the highest level of corruption. Third, the higher the corruption risk, the more important it is to identify clear corruption red flags. The measurement and understanding of corruption can contribute to a better policy response. Fourth, the thesis assumes the strongest relationship between EU-funded tenders and corruption risk in CEE.

The thesis is centered on the last complete financial cycle of the EU, spanning from 2014 to 2020. Examining the entire cycle is crucial as the yearly distribution of EU funds can differ within the cycle. The utilized methods are Random Forest – which is a machine learning algorithm in the field of big data – and Coarsened Exact Matching developed by Gary King (2009). The combination of the two methods contribute to a more grounded estimate on the actual level of European corruption. Random Forest chooses the most important variables for the final regression model, while matching ensures covariate balance between the treated (EU-funded tenders) and control group (nationally funded tenders). If the two groups are in perfect balance, researchers can make a causal estimate.

The structure of the thesis is as follows. First, it introduces core corruption literature, narrows down the scope to Europe, and reviews the possibilities of measuring corruption in the field of big data. Then, it proceeds by presenting the database and the applied methods. After that, it discusses the composition of Corruption Risk Index, the selection of key variables by random forest, the matching process and the regression analysis. Finally, it concludes with an evaluation of the results in the context of the initial assumptions and some proposals for further research.

LITERATURE REVIEW

What is Corruption?

Scholars agree on the conceptualization of corruption as the abuse of public power over private gains (Klitgaard, 1988; Rose-Ackerman, 1996; Shleifer & Vishny, 1993, Mackey et al. 2016, Aldana et al, 2022). Petheram (2019) conceptualizes corruption as a 'fiscal leakage', which refers to governmental money loss due to unfair redirecting of public funds to private accounts. Or as Thompson (2005, p. 4) famously stated, "the pollution of the public by the private". Corruption is the act of engaging in illegal payments to gain unmerited personal gains, which in turn inflict damage upon society. Such damage may manifest itself in various forms, such as reduced economic growth, diminished business performance, decreased private investment, unnecessary expenditure, unequal distribution of resources, ineffective regulations, and diminished public trust (Colonnelli et al., 2020; Lima et al., 2020).

In the case of public procurements, corruption can include practices such as the presentation of fake documentation and receipts, not delivering the goods or services stated in the contract, and greatly overpricing the offered services, so the price and the quality of project is not in proportion (Aldana et al, 2022). It is also a common practice, that friendly companies secretly return parts of the money to government officials (Petheram, 2019). Tanzi (1998) points out, that corruption is often calculated in the budget as a planned expenditure. In these cases, instead of choosing the best price-quality combination, corrupt officials choose the 'best briber' (Søreide, 2002). The selected companies are rewarded with a de facto monopoly situation in business and can access classified information easily. Despite their poor performance, they might get better licenses, lower taxes and favorable law modifications. Most likely, without corruption, these companies would not have been contracted (Søreide, 2002).

According to the IMF (2019), a comparison of governments with varying corruption perceptions revealed that the least corrupt governments can generate tax revenues equal to 4% of their GDP more than corrupt governments. A reduction in corruption worldwide (without complete eradication) could lead to a global savings of around $1 trillion, equivalent to roughly 1.25% of the world's GDP. However, the World Bank (2022) presents a much larger estimate of the economic impact of corruption worldwide, with a loss of $2.6 trillion, equivalent to 5% of global GDP, annually. The RAND Corporation (2016) estimates that the overall cost of

corruption in Europe could be as high as €990 billion, with public procurement corruption alone accounting for €5 billion.

Based on these figures, it is evident that many countries are not doing enough to combat corruption. However, there are plenty of policy recommendations to minimize the risk of corruption (see the 33 policy recommendations by Søreide, 2002). An example from her regards the regular rotation of public officials. Another advice is to be transparent about government finances and expenditure. Søreide emphasizes the clarity and simplicity of public procurements, so the public can keep an eye on the decision makers. The importance of open data recurs in the study of IMF (2019), which found that countries with higher levels of transparency are perceived to be less corrupt by the public.

European Institutions and Corruption

Corruption in the EU is particularly interesting, because these developed countries have sufficient knowledge and funds to prosecute corruption, but they continuously fail to do so. A widely used practice to tackle corruption is the establishment of Anti-Corruption Agencies (ACAs). The European Union has several supranational agencies, such as the European Anti-Fraud Office (OLAF), which possesses the power of gathering evidence, investigating, providing support and expertise, and recommending sanctions such as fines and suspension from receiving EU funds to the European Commission. However, it lacks the authority to prosecute on its own and intervene on the national level. If OLAF uncovers corruption during an investigation, it can bring attention to the issue but cannot initiate court proceedings in any European countries (de Sousa, 2009).

The Group of States Against Corruption (GRECO) is an intergovernmental group under the Council of Europe and it has comparable powers to OLAF. GRECO assesses the compliance of member states with the Council of Europe's anti-corruption standards and offers recommendations on how to improve their anti-corruption frameworks and practices (Council of Europe, n.d). Nevertheless, these recommendations are non-binding, and similarly to OLAF, it can be easily disregarded. A further concern can be that member states appoint experts for GRECO's evaluation, which might lead to bias towards the experts' own country. Since GRECO experts are mandated by the government, it might imply some political connections and good relations to the government.

The establishment of the European Public Prosecutor's Office (EPPO) in 2021 indicates that previous efforts to combat corruption were unsuccessful. Unlike OLAF or GRECO, EPPO has broader powers and can prosecute criminal cases before national courts of EU member states. Nonetheless, participation in this authority is voluntary, allowing the most corrupt politicians in power to avoid it. Reviewing the competencies of these ACAs were important because it becomes clear why they are not fulfilling their role. The fundamental issue with all of these endeavors is that they assume the state desires to combat corruption (Martin-Russu, 2022). However, in most cases, public officials are the ones who engage in corrupt transactions.

Moreover, the study of Rand Corporation (2016) found that EU countries are just as corrupt as before joining the European Union. Even though the EU has certain moral and legal obligations, as well as institutions that were established to help, the level of corruption stuck around the year of 1995. Once countries met the admission criteria, the EU could no longer put strong pressure on them. This tendency is in sharp contrast with the associated positive change of EU-membership. Batory, (2018) described these mechanisms in CEE as a post-communist heritage, where citizens have a strong preference for a 'culture of corruption'. In deeply corrupt systems the elite is empowered to bend the rules to their liking. Furthermore, there is a naïve assumption, that the adopted democratic and anti-corruption improvements before accession remain in place. The dismantlement of laws is perceived to be too costly, that states are unwilling to pay (Martin-Russu, 2022).

Big Data in Corruption Research

The outlined problems have at least one thing in common, which is the inadequate response to corruption. The inaction for detecting corrupt practices may be attributed to a lack of political will or inadequate measures. Researchers can address the latter issue. As corruption is not directly observable, scholars use proxy indicators that signal corruption red flags. It is a common approach to construct simple linear models and correlate them to established corruption perception indexes (CPIs), where a high correlation coefficient indicates a good proxy measure. Chang and Golden (2007) correlated Transparency CPI to the countries' electoral system, Lipset and Man (1960) to the levels of education of society, and Wachs et al (2021) and Fazekas et al (2014) used the variable single-bid, that indicates the lack of competition in public tenders.

More recently, researchers in the field of big data have attempted to estimate corruption by utilizing machine learning algorithms to predict and prevent corruption. In 2019, Petheram published an article titled "The next generation of anti-corruption tools: Big data, open data & artificial intelligence", which highlights the new direction of corruption detection. The author suggests that the digitalization of public procurement data provides a great opportunity to develop anti-corruption tools, and with the use of machine learning algorithms, researchers can accurately predict instances of corruption within government.

According to studies conducted by Colonnelli et al. (2020) and Lima et al. (2020), tree-based models such as random forest outperform other machine learning methods like LASSO, neural networks, and support vector machines. By combining random forest with a multi-source database, relevant corruption related information can be revealed, leading to the most accurate predictions.

The most commonly utilized feature of random forest is predicting variable importance. According to Lima et al. (2020) in their analysis of country-level data, the most significant variables for predicting corruption are government integrity, property rights, judicial effectiveness, and education index. Colonnelli et al. (2020) found that private sector and human capital characteristics are the most powerful indicators for predicting corruption, while financial development is also a relevant factor. Aldana et al (2022) came to the conclusion that the relationship between buyers and suppliers is the most critical predictor of corruption, not the features of individual contracts.

It is clear from the big data literature, that random forest is one the most useful tools in corruption research, so this thesis will follow a similar approach on a European dataset by applying random forest with its function of plotting variable importance.

RESEARCH DESIGN

In this section, first, the thesis describes the general characteristics of the dataset. Second, it introduces the widely used Corruption Perception Indexes (CPIs) and third, the basic ideas of Random Forest and CEM. Fourth, the dataset will be prepared for the analysis by data cleaning, based on theoretical and technical considerations.

Dataset

This study uses a free and openly accessible dataset from the website opentender.eu. It contains country level procurement data from European countries, and it is updated twice a year. The dataset was created in the frames of the DIGIWHIST Project, which received funding from the European Union. It gathers officially reported details of public tenders, both EU and nationally funded, from two main different data sources: Tenders Electronic Daily (TED) – the online version of an official EU journal dedicated to European public procurement – and national public procurement websites. Regarding the latter source, the researchers applied a big data technic called web scraping, by which they could transfer the information from the websites to neatly organized tables.

Since it is an observational dataset, some variables have a high missing rate. This might be caused by the differing country-specific regulations regarding the substance of publishable content and the required publication threshold. The dataset aims to make public procurement tenders more transparent and helps to measure the degree of integrity. It is a large-scale structured database, which has altogether 46,734,888 observations and 138 variables ranging from the time period 2009-2022. This study utilizes those observations that have been acquired from TED, because of better data quality and reporting standards, as well as lower rate of missing values.

Corruption Perception Index (CPI)

The most widely used CPI is developed by Transparency International (TI), which measures the yearly change of country-level corruption. More precisely, it measures corruption in the public sector, such as bribery, the diversion of public funds, effective prosecution of corruption, legal frameworks, access to information, legal protection of whistleblowers, journalists and investigators. TI does not conduct primary research; it merely harmonizes the 13 external

sources of corruption perception (CP). Then, TI takes the average of the standardized 13 sources to get the well-known Transparency CPI. It has a scale of 0-100, 0 indicating a 'highly corrupt', and a 100 'very clean' country (Transparency International, 2021). For easier interpretability the thesis turns the scale (0 - very clean, 100 - highly corrupt).

Since Transparency CPI is a complex indicator, scholars often use it as a reference point (Lima et al, 2020, Petheram et al, 2019). However, there might be other indexes, that are also useful for the analysis. The Eurobarometer CPI targets its surveys to bidding companies, who contested in a public tender. Eurobarometer has one question which is particularly important for this study. '*In the last three years, do you think that corruption has prevented you or your company from winning a public tender or a public procurement contract?'* Even though Eurobarometer is a good addition to Transparency CPI, it only operates with a small sample (number of country-level observations are between 40 and 140). Furthermore, data collection happens in every second year, so for the analysis the years 2015, 2017 and 2019 are used.

Other concerns with Eurobarometer regard a potential selection bias, as those companies who have a significant market share but have not applied for a tender, are not included in the sample. Not contesting might occur because of self-regulation, as those who think the tender is inherently corrupt, will not contest because they do not expect to win. In cases of single bidding, there is only one contestant, and it will not complain about corruption (winners rarely do). So, tenders that are meant to be corrupt, might not seem to be corrupt when only this dimension counts. The point here is, that there are subtle ways of corruption, that are incredibly hard to measure.

Another CPI, which is often used as a reference is the World Bank Control of Corruption. It measures the perceived extent to which public power is used for individual gain (World Bank, DataBank). However, this yearly indicator is also incorporated into TI's composite indicator, which is also visible in the correlation heatmap (see Figure 2 in the section 'Selecting the Best Corruption Risk Index'). The major problem with all these CPIs is that they are based on corruption perception, which is highly subjective. CRI aims to take out the subjective dimension of corruption measurement and focus solely on the aspects restricting competition. The newly created Corruption Risk Index (CRI) will be exposed to an external validity check with CPIs. This means that the country-level scores of CRI will be correlated with each above-mentioned CPIs. As CPIs cannot perfectly capture corruption, an exact match is neither expected nor

desired. What matters here is that there should be a strong or at least moderate positive correlation between the indexes.

Random Forest and Coarsened Exact Matching

After finding a good enough CRI, an unsupervised machine learning algorithm, random forest will be used to select the most important variables, which then will be incorporated into the regression. More generally, random forest can be used for both classification and regression analysis. The name "random forest" comes from the fact that it generates multiple decision trees. These trees are created using randomly selected subsets of features and data points, which together make up a "forest." By using this random selection method, the algorithm is able to avoid selection bias and produce more accurate predictions. The random subsets of data are intended to be representative of the overall population, which helps to improve prediction accuracy.

The final output of the model is produced by combining the results of all the trees in the forest. One of the key advantages of using the random forest method is that it can help reducing overfitting, which can occur when a single decision tree is too closely tailored to the training data. By combining the results of multiple trees, the random forest method can produce a more robust estimate that is less prone to overfitting. A robust estimate is key in a quantitative analysis, that is why it is preferred by for example Colonnelli et al. (2020) and Lima et al. (2020).

However, the essence of the research lies in the difference between EU-funded and national funded tenders. Coarsened Exact Matching (CEM) aims to establish a causal relationship between a treatment (EU-funded) and a control group (non-EU-funded). Both groups need to be balanced, which means the covariates in the two groups should be roughly similar to each other. By dividing Europe into four regions, balancing and matching the covariates facilitates the analysis. Based on the observed covariates, the treatment variable will be paired with a closely-matched control variable. The assumption is that if the observed covariates are in balance, the unobservable covariates will be in balance as well.

Ideally, the two groups should be similar in all aspects, except for the treatment. Matching assumes that there are no other treatments happening concurrently and the difference between the groups is solely due to the treatment. All kinds of matching technique imply data pruning, which is not problematic in this case because the database is exceptionally large. After pruning,

there will be sufficient number of matches to estimate causal inference, so no valuable information will be lost. CEM is different from other types of matching because it requires covariates to match within a specific interval. In contrast to exact matching (one-to-one matching), CEM permits more flexibility in the covariates. As per Gary King et al. (2009), CEM is regarded as more reliable than propensity score matching (PSM), which is often deemed a "black box" because of its opacity.

Data Cleaning

This thesis utilized an exceptionally large dataset comprising of 6.5 million observations and 137 variables. The data spanned the time period between 2014 to 2020 and covered all countries within the European Union. Due to the extensive size of the dataset, substantial computational resources and time were required for processing. Prior to analysis, meticulous data cleaning was performed, employing the methodology outlined by Fazekas and Kocsis (2017). The authors removed non-competitive markets, meaning markets with high entry costs and low profitability, such as water collection and purification and certain segments of the oil and gas industry. The low number of bidders in these markets did not necessarily indicate corruption but rather the presence of naturally monopolistic market structures.

Additionally, observations with fewer than ten bidders throughout the seven-year period were eliminated from the dataset. However, this approach resulted in a higher number of deleted observations for smaller countries with lower competition, such as Malta and Cyprus. Furthermore, another pruning practice was implemented to remove tenders that fell below the official reporting threshold established by the EU. Such tenders usually pertain to small-scale projects where the likelihood of corrupt practices is assumed to be minimal. As a result of these data cleaning measures, a more manageable dataset was obtained, consisting of 5.37 million observations and 49 variables.

Following the theoretically based data cleaning procedures, certain technically required data cleaning steps also needed to be taken. Random forest assumes that the variables under examination cannot contain any missing fields. Consequently, all observations with missing values for variables included in the random forest model had to be removed. This led to a final sample size of 2,740,419 observations for the random forest analysis. The basis for removing missing observations was the assumption that they were randomly distributed throughout the sample.

RESULTS

The Results section explains in detail how the Corruption Risk Index was created and how it correlates with Corruption Perception Indexes. Then with the help of Random Forest, the most relevant variables will be identified and used for regression analysis. For the matching, two approaches will be utilized: Model 1, which embodies the main model, and Model 2, as a sort of robustness check. Both types of matching are conducted on a regional basis. CEM combined with Random Forest improves the prediction of the model that results in a more reliable estimate.

Selecting the Best Corruption Risk Index

Corruption Risk Index is an instrument to estimate the actual level of corruption within the European Union. In order to select the best CRI, the correlation coefficient of two external Corruption Perception Indexes have been checked with several CRI combinations. CRIs were acquired by calculating the country-level arithmetical average of some corruption indicators (red flags) available in the dataset and comparing them to country-level CPIs. While calculating the means, only the available indicators have been taken into account for each observation, meaning that NAs were simply omitted. As a result, indicators with lots of missing values had significantly less impact on the outcome of the CRI than indicators with less or almost no missing values.

First, the total mean of all tender indicators (according to DIGIWHIST standards) was checked. These include integrity, transparency and administrative indicators and the result was a very weak positive correlation with Transparency CPI and a weak or moderate correlation with Eurobarometer CPI (see Table 1). Since the core corruption indicators are included among integrity indicators (administrative and transparency indicators are rather complementary ones giving information on administrative capacity and missing information of some key indicators), and the two other type of indicators also tend to have much more missing values, the next step was to look at different subsets of integrity indicators.

Corruption Risk Index	Correlation with Transparency CPI	Correlation with Eurobarometer CPI
Total average of all tender indicators	0,0666	0,2302

6-component integrity indicator [1]	0,0925	0,1572
5-component integrity indicator [2]	0,1093	0,1802
4-component integrity indicator [3]	**0,4376**	0,3979
3-component integrity indicator [4]	**0,4346**	**0,4039**
Single bid	**0,4449**	0,3393

Table 1: Comparison of correlation coefficients of different Corruption Risk Indexes with Transparency and Eurobarometer CPI

Therefore, the correlation of country-level means of available integrity indicators with the two external CPIs was checked starting from the whole set of integrity indicators and moving towards only the most important integrity indicators. As a result, an almost continuously increasing correlation coefficient was acquired with Transparency CPI and Eurobarometer CPI. The three highest coefficients with Transparency CPI were the 4- and 3-component integrity indicators and single bid; whereas the two highest coefficients with Eurobarometer CPI proved to be the 4- and 3-component integrity indicator.

The 4-component integrity indicator (that contains the information of the winner company being registered in country classified as tax haven by opentender.eu) had high level of missing values. As the indicator tax haven changed neither the CRI values nor the correlation coefficients significantly, this option was ruled out. Similarly, as single bid (the one-component indicator) leaded to a lower coefficient with Eurobarometer and would cause more significant loss in available red flags, the 3-component integrity indicator was selected for the analysis. This indicator measures both if there was only one candidate for a public procurement and whether the buyer of the tender has made some efforts to decrease the chance of competition by selecting a non-open procedure type and/or not publishing a call for tender.

The advantages of this CRI are that it has less components (easier interpretation) and all of its components are available for most observations. Furthermore, it has a moderate positive

[1] 6-component: single bid, procedure type, call for tender publication, tax haven, decision period, advertisement period

[2] 5-component: single bid, procedure type, call for tender publication, tax haven, decision period

[3] 4-component: single bid, procedure type, call for tender publication, tax haven

[4] 3-component: single bid, procedure type, call for tender publication

correlation coefficient with both external CPIs, and it probably does not overestimate the extent of corruption as it emphasizes on a few clear signs of the lack of competition in public procurement. Additionally, it has the highest correlation coefficient with Eurobarometer CPI, which is the most directly linked to corruption in public procurement.

Its main disadvantage is that it may underestimate the extent of corruption by not being able to measure more sophisticated forms of corruption. An example could be when there is a theoretical possibility for competition (open tenders with a call for tender published and more than one bidder), but corrupt practices are present in the decision procedure.

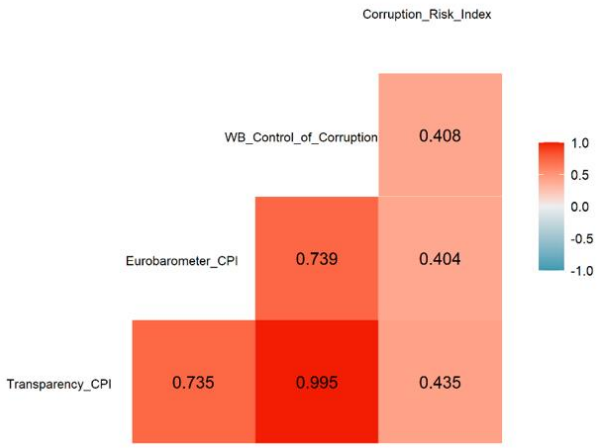

Figure 2: Correlation matrix for External Validity Check

During the selection of the best CRI, two external validity checks have been conducted through correlations. The correlation matrix of the final CRI, the two initial CPIs and an additional CPI can be found in Figure 2. The heatmap shows the correlation coefficient of all used CPIs and the new CRI. The shades of red indicate a positive correlation coefficient and the darker the color the stronger the correlation.

All the established indicators highly correlate with each other (WB Control of Corruption (CoC) and Eurobarometer 73,9%; Eurobarometer and Transparency 73,5%), however, a 99,5% correlation coefficient between WB CoC and Transparency is unexpectedly high. As discussed in the section Corruption Perception Index, Transparency CPI is comprised of 13 external sources of corruption perception. One of the sources is WB CoC, which partly explains the high value. Because of the great overlap, the WB CoC was not used to pick the best combination of CRI, as that would have resulted in an overrepresentation of two very similar indexes.

Final Corruption Risk Index

The final CRI was constructed by taking the mean of three indicators, namely single bid, procedure type, and call for tender publication. For better interpretability, the final CRI was inverted from the original scale to its opposite. As a result, 0 indicates no corruption risk and 100 indicates very high corruption risk. Its constituent indicators were considered the most explicit signs of non-competitive tenders.

Single bid pertains to a bidding process where there is only one participant. This circumstance leaves the buyer with no alternative but to select the sole contender. In markets that are competitive and abundant with suppliers, it is highly improbable for only one company to express interest in participating in a bid. Consequently, it is assumed that such tenders lack transparency and fairness. Fazekas and Kocsis (2017) utilized the variable "single bid" as a dependent variable to gauge the likelihood of corruption, and it exhibited the strongest correlation coefficient among other indicators of corruption perception.

Procedure types are classified by DIGIWHIST standards, which can encompass options such as: Open, Restricted, Restricted with publication, Negotiated without publication, Competitive dialog, Design contest, Minitender, DPS purchase, Outright award, Approaching bidders, Public contest, Negotiated, Innovation Partnership, Concession, and Other (national type). To facilitate analysis, the string variables were converted to numerical values on a scale of 0-100. A corruption risk of 0 represents no corruption risk, such as in the case of an open tender, while a score of 100 signifies a high corruption risk, such as limited access to the tender process. The specific numerical assignments were carefully determined, taking into account country-specific characteristics. Consequently, the assigned values for a procedure type like "approaching bidders" may differ between two different countries.

Call for tender publication is a binary indicator that assesses whether tender announcements were made publicly or not. When tenders are not announced, only politically connected firms, who have been informed in advance, can submit their bids. This selective practice limits competition and exhibits a positive correlation with corruption risks. In their analysis, Fazekas and Kocsis (2017) incorporated procedure type and call for tender publication as components of their dependent variable. They supplemented these indicators with two additional measures that identified advertisement and decision periods that were suspiciously either too long or too short, thus further contributing to the assessment of corruption risk.

Visual representation of CRI correlation with CPIs

This section is about the visualization of correlation coefficients between established CPIs and the newly created CRI. The following graphs (Figure 3, 4, 5) show a moderate a positive correlation with different CPIs. The graphs give a more nuanced picture of the accurate positioning of countries and their positions within the European regions. Not only the countries but the country groups can be ranked as well. The least corrupt country group is Northern

Europe, then comes Western Europe, followed by Southern Europe and finally Central-and Eastern Europe. This sequence seems to be stable across all CPIs and it is accurately captured by the CRI as well.

Even though CRI in general provides a lower estimate for corruption than CPI, it can be observed that countries above the linear line (depicting the correlation coefficient) seem to be proportionately more corrupt in CRI compared to countries below the line.

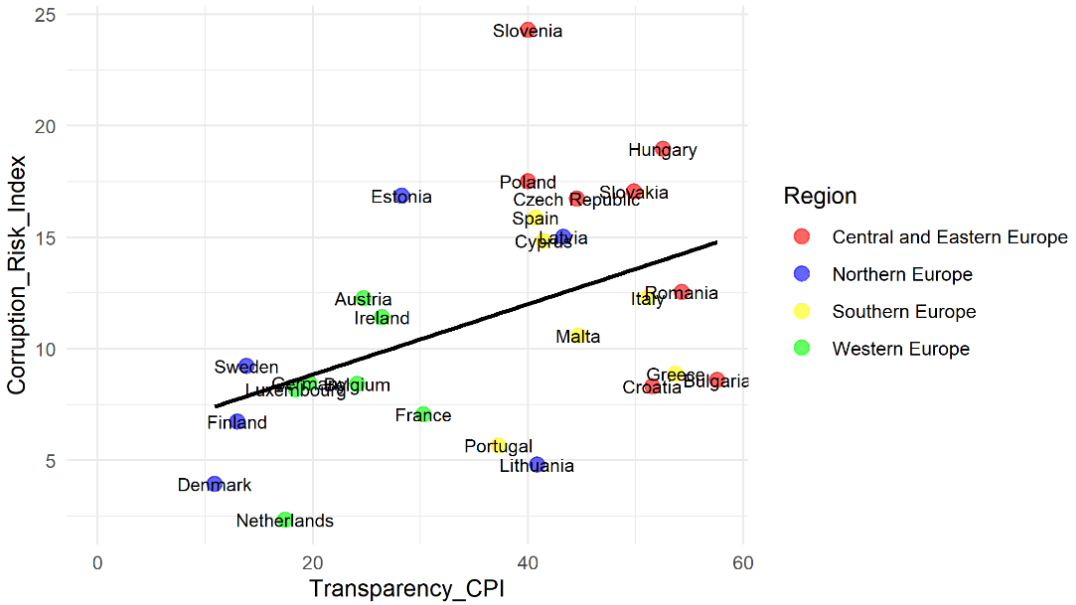

Figure 3: Comparing the Corruption Risk Index to Transparency CPI (Correlation is 43,5%)

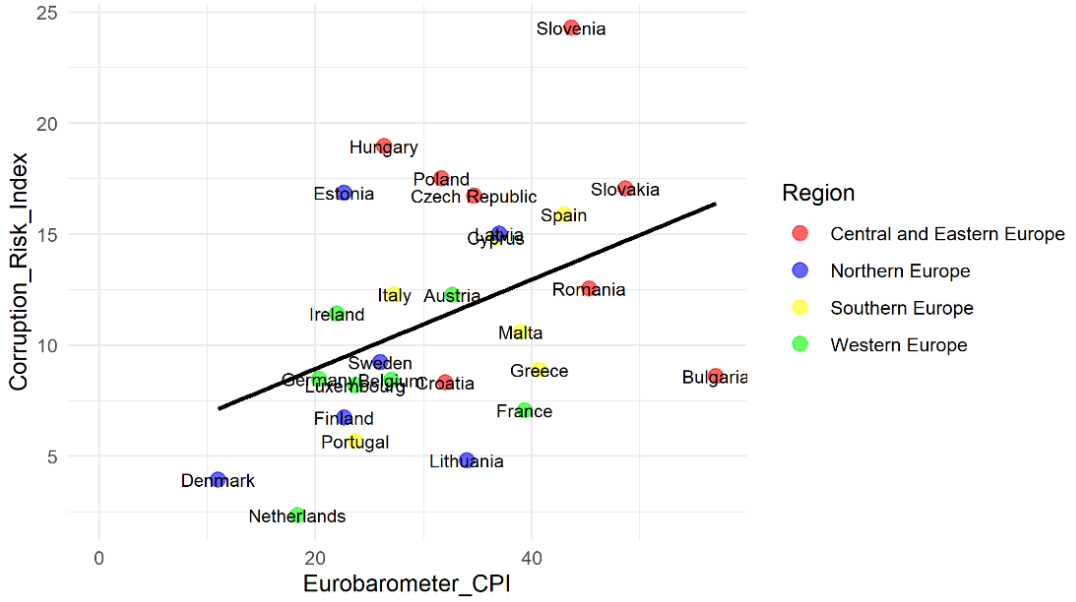

Figure 4: Comparing the Corruption Risk Index to Eurobarometer CPI (Correlation is 40,4%)

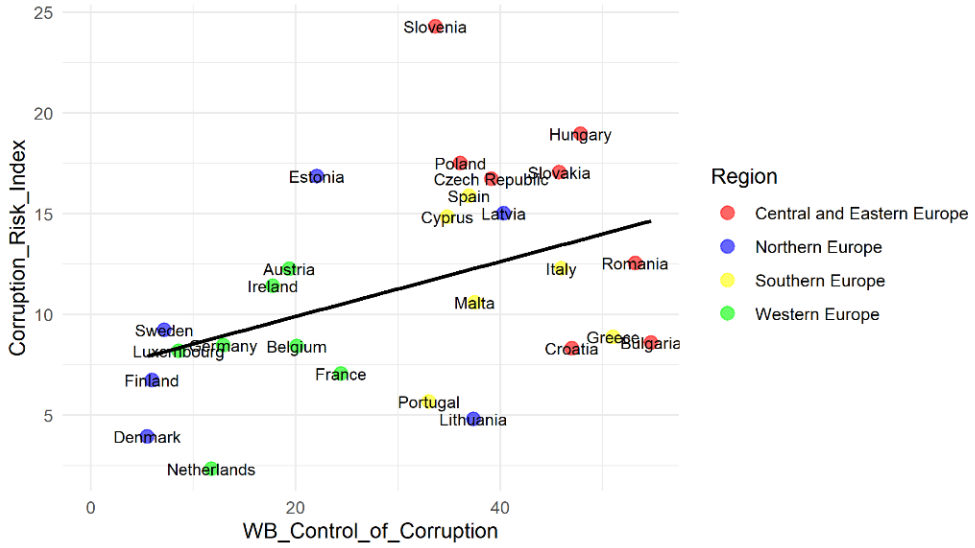

Figure 5: Comparing the Corruption Risk Index to World Bank Control of Corruption (Correlation is 40,8%)

Descriptive statistics of CRI and EU funding

In this section, a few visualizations of the two most important variables (CRI and whether a tender received EU funding) will follow.

As per Figure 6, the number of EU-funded tenders identified in the database is constantly decreasing each year between 2014 and 2020. It is also associated with a decreasing rate of information about tenders being or not being supported by EU funding, as the number of missing values is sharply increasing.

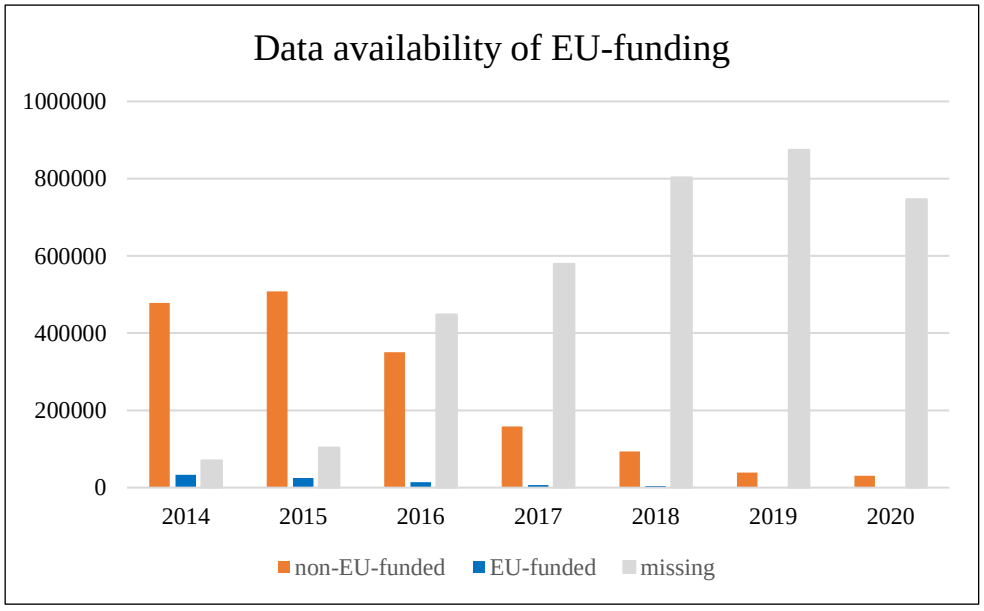

Figure 6: Comparing the data availability of EU funding and the number of EU-funded tenders, 2014-2020

The availability of tender data collected during the initial 7-year period will influence the overall results. Tenders from the latter part of the period will probably have a smaller effect on the results, primarily because of the limited availability of observations and lower number of EU funded tenders. Out of the total 5.37 million observations that remained after data cleaning, there are only 1.74 million observations with information about the presence or absence of EU funding provided for the tender.

After the data cleaning procedure and before matching, in all regions, a slightly higher CRI average is observable among the EU funded tenders compared to non-EU-funded ones. It is crucial to highlight that there are already quite significant deviations between the regions, especially between CEE and the remaining three regions regarding the average CRI in non-EU-funded tenders (see Figure 7). This cross-regional difference is also observable among EU-funded tenders.

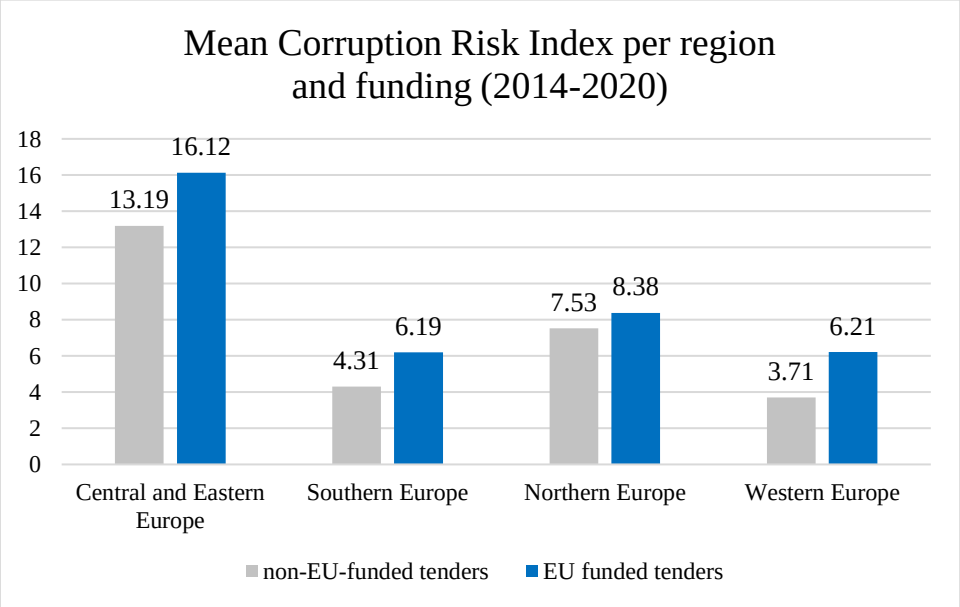

Figure 7: Comparing the mean CRI of EU funded and non-EU-funded tenders

Figure 8 represents the annual average CRI for each region, without splitting between EU funded and non-EU-funded tenders. The cross-regional difference remains visible, with a slight increase over time in the level of CRI in most regions. In 2019 and 2020, the average CRI figures in NE, SE and WE are not included in the diagram due to the limited number of observations remaining in the dataset after undergoing data cleaning.

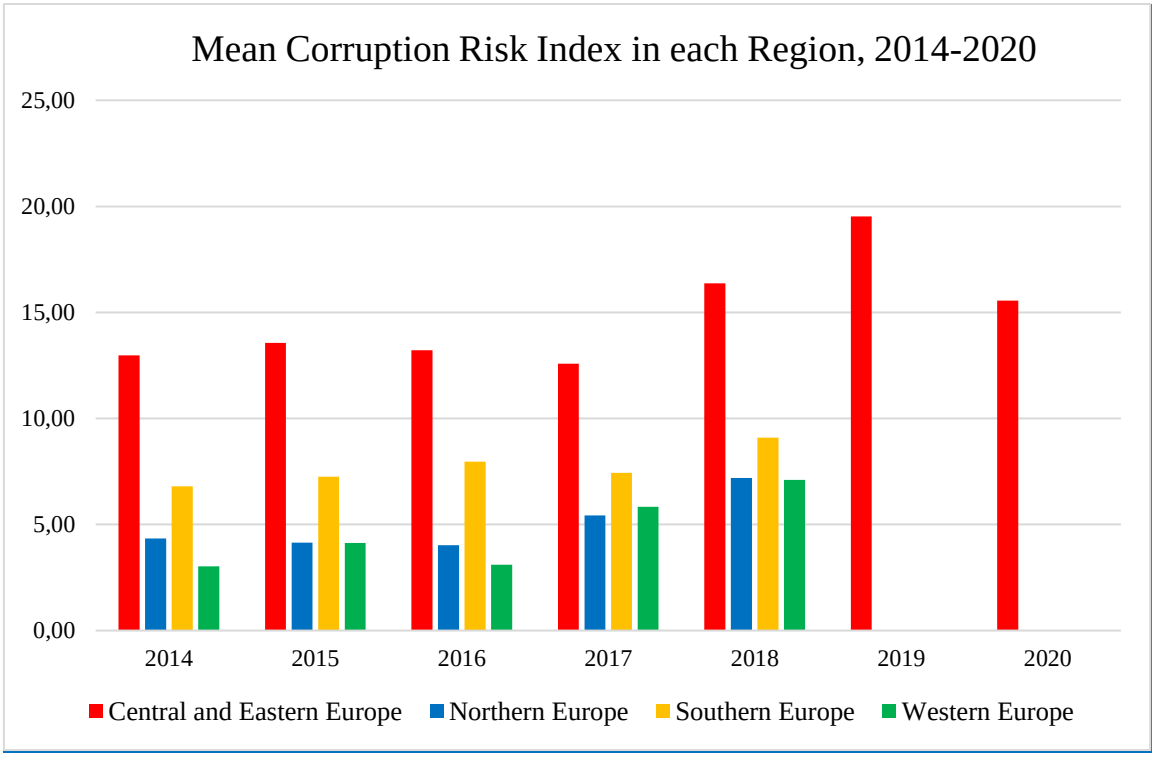

Figure 8: Regional division of Corruption Risk Index per year

Random Forest

Random Forest is a machine learning method utilized to choose the most important variables for the analysis. The variables with the highest explanatory power were put into the final regression model. However, Random Forest only runs if there are no missing values in a variable. Therefore, only covariates with low or close to zero missing rate could be included. Missing observations for these variables were omitted in the hope that missing observations are randomly distributed in the sample, and it will not lead to selection bias.

Here is the list of potential covariates with good data availability that will be tested by Random Forest:

- **Country**
- **Tender year**: the year in which the tender was awarded. In this paper years 2014-2020 are relevant.
- **Division**: variable created by using the first two digits of CPV. Encompasses 46 industries.
- **Tender has lots**: a lot is part of a tender that can be awarded separately. If a tender has multiple lots, it is assumed to be less corrupt.

- **Buyer NUTS**: shows the regional code of the buyer authority. There are several NUTS within a country.
- **Buyer type**: it signifies the type of buyer authority, that can have the following values: national authority, national agency, regional authority, regional agency, public body, European agency, utilities, other.
- **Tender supply type:** type of the purchase, it has values such as supplies, services and public works.
- **Tender description length**: it counts the number of characters. If the description is suspiciously short (not giving enough details for companies) or long (being too specific that the description only fits one company), the tender is considered to be corrupt.
- **Tender final price**: the larger the price, the greater the room for corruption, which means it could attract more corrupt officials.

Random Forest will take a random subset of these variables, in other words, it will create random trees. Since this grouping process is purely random, the same values can be selected multiple times. The more trees in a forest, the more accurate the prediction. Therefore, scholars aim to have big forests with plenty of trees, but the theory is often confronted with reality. Having a big forest requires serious computing capacity. The biggest possible forest that could be run with this dataset consists of 20 trees (see Figure 9). Around 15 trees, the mean squared error (MSE) is 80, which translates to roughly 9 root mean squared error (RMSE). An RMSE value of 9 signifies that, on a scale of 0-100, the average error of the model is approximately 9. An even lower error term is achieved by a forest with 20 trees, indicating that the prediction of the model is quite close to the actual value. By adding extra trees, the model prediction would only experience a slight improvement, as the curve nicely flattens around 20.

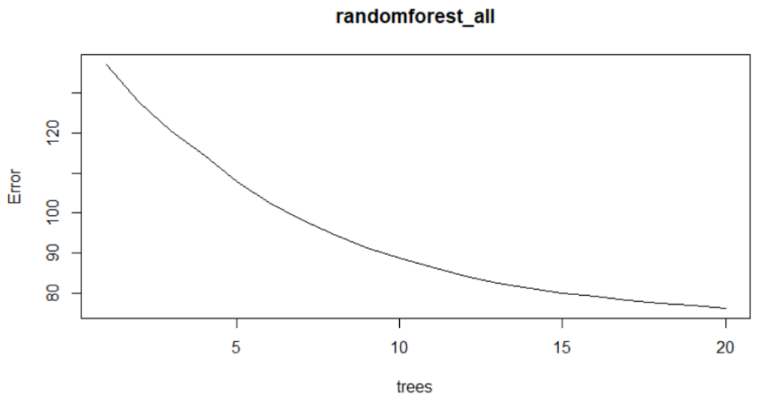

Figure 9: Random Forest with 20 trees, run on 60% of the total sample (training dataset is 60%). Total sample: n= 2,740,419; training sample: n = 1,644,251

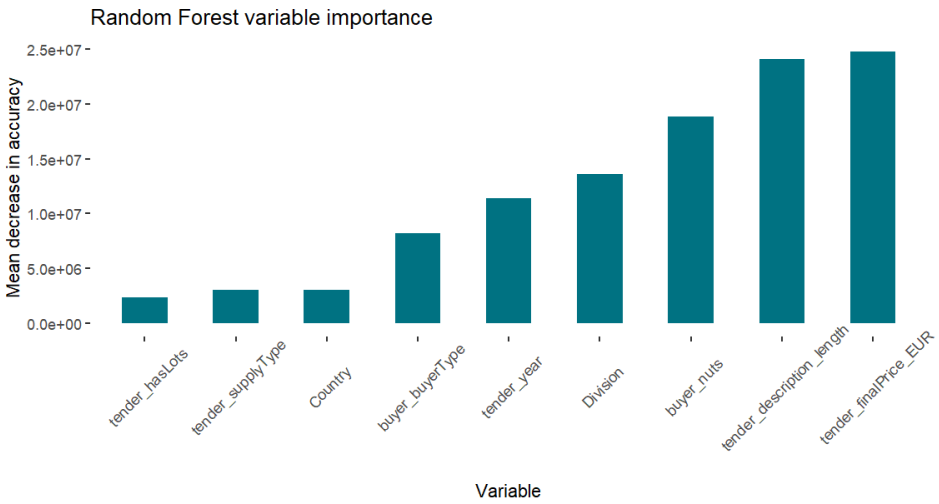

Figure 10: Random Forest variable importance run on all country groups.

After ensuring that the predictive power of the model is satisfactory, random forest was run on both the total sample (see Figure 10) and each region separately (see Figure 11). Figure 10 gives a general and aggregated overview of how variable importance is ranked in the total sample, while Figure 11 focuses on regional differences. In both cases, the final price of the tender is at the first place, and shortly after comes tender description length. It is a notable difference that in Southern Europe and in Central Eastern Europe buyer NUTS is taking the third place, while in Northern Europe and Western Europe it is the variable division.

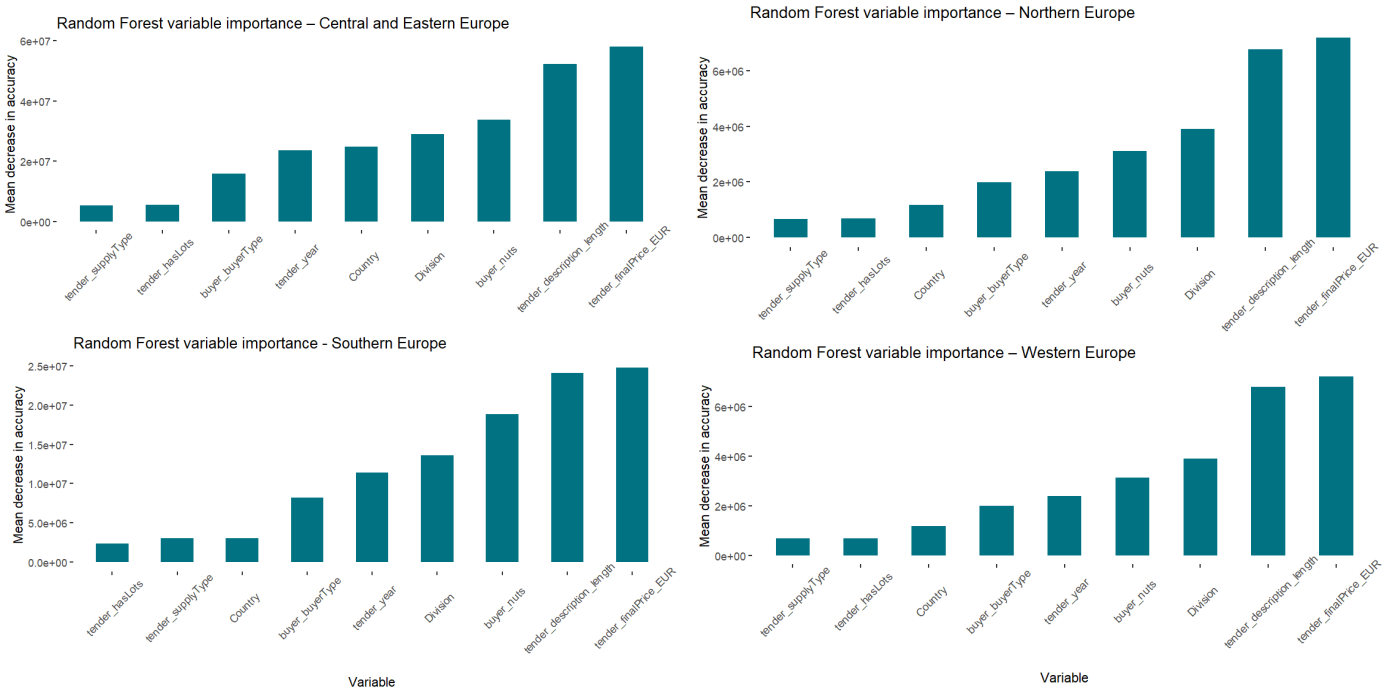

Figure 11: Random Forest variable importance run on European regions CEE, SE, NE, WE

Because SE and CEE are more populous and they received more EU-funding (as they are less developed), in the overall ranking their results carry more weight. Thus, these two regions have a bigger influence on the general impression of variable importance. However, there could be significant regional differences in the dependent and main independent variable, so the unit of analysis remain on a regional basis.

All in all, based on the results of the random forest analysis, the five most important covariates chosen to be included in the regression model are: final price of the tender, description length, NUTS of the buyer, division of the main product code (CPV) of the tender and year.

Coarsened Exact Matching

Matching is used to create a balance between the treatment (EU-funded) and control variables (national funded tenders). A balanced sample is required for a reliable causal inference model. In the matching process, two different approaches are applied. The first one (Model 1) aims to reduce confounder bias by including variables that are associated with the tenders being EU-funded. This is done without losing significant number of observations, which is important as there is already a tremendous loss of observations due to the missing values in the treatment variable (EU funded, 67.5% missing rate). The second approach (Model 2) attempts to decrease confounder bias by an even larger extent by including an additional potential confounder (Division) and sacrificing some more observations. After conducting the matching according to both approaches, the same regression model will be run, and the results will be compared also as a robustness check. The post-matching observations and average CRI figures are presented in Appendix A-D.

Potential variables to be used in the matching process are the following:

- **Country**
- **Tender year**: the year in which the tender was awarded. In this paper years 2014-2020 are relevant.
- **Division**: variable created by using the first two digits of CPV. Encompasses 46 industries.
- **Tender has lots**: a lot is part of a tender that can be awarded separately. If a tender has multiple lots, it is assumed to be less corrupt.

- **Buyer NUTS**: shows the regional code of the buyer authority. There are several NUTS within a country.
- **Buyer type**: it signifies the type of buyer authority, that can have the following values: national authority, national agency, regional authority, regional agency, public body, European agency, utilities, other.
- **Tender supply type:** type of the purchase, it has values such as supplies, services, public works.
- **Tender description length**: it counts the number of characters. If the description is suspiciously short (not giving enough details for companies) or long (being too specific that the description only fits one company), it may contribute to higher corruption risk.
- **Tender final price**: the larger the price, the greater the room for corruption, which means it could attract more corrupt officials.
- **Tender is central procurement**: whether the tender is a centralized procurement, which shows good administrative capacity that may reduce the risk of corruption (Hrubý et al, 2018).

Model 1: Matching by Country, Buyer type, Lots

Following the first approach, three variables have been selected for matching: country, buyer type, and tender_haslots.

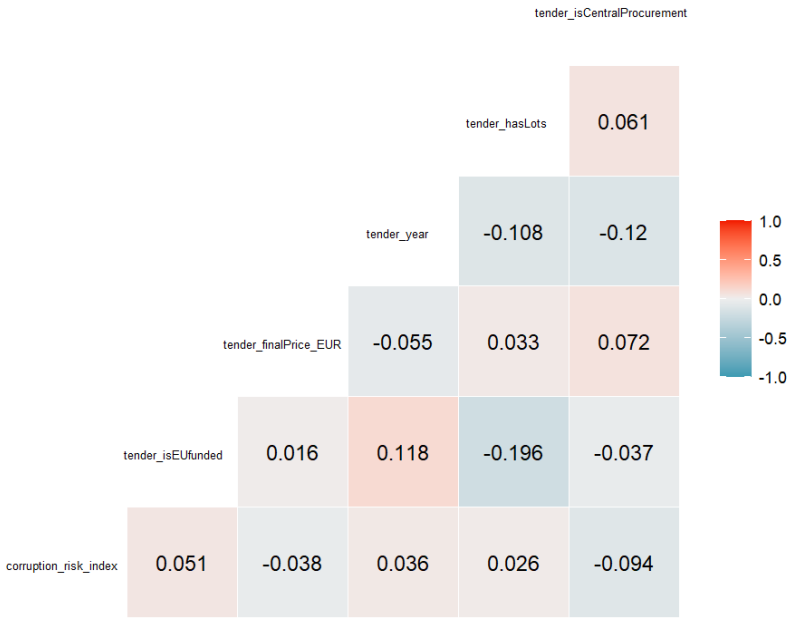

Figure 12: Correlation matrix to select variables for matching

The variable tender_haslots has been selected as a result of a correlation matrix including the treatment variable and all potential confounders that are not categorical and thus can be used for computing pairwise correlations. Out of all coefficients, tender is EU funded and tender has lots showed the strongest correlation of -0.196 – even though it is still a rather weak association (see Figure 12).

Out of the remaining potential confounders that are categorical variables and cannot be put into a correlation matrix (Country, Division, buyer NUTS, buyer type, tender supply type), two more have been selected following a check of covariate balances: country and buyer type. They showed a significant imbalance in most of their sub-categories but had not too many distinct values to prevent a further loss of observations. According to bivariate linear regression models conducted between country, buyer type, and the treatment variable (see Table 2), they are important predictors of tenders receiving EU funding, and their imbalance have been strongly reduced by the matching process.

Model 2: Matching by Country, Buyer type, Division

The second matching approach aims to decrease confounder bias to an even further extent by explaining a larger proportion of the variance in the treatment variable (EU funding). Table 2 presents all potential confounder variables that were one-by-one put into linear regression with the treatment variable (EU-funded). R^2 shows what percentage of the variance of tenders being EU funded is explained by each variable. The variables with the highest explanatory power were the following: Division[5], 2-digit NUTS, Country and Buyer type. Since Division and NUTS have dozens of categories, including them would result in a further loss of hundreds of thousands of observations.

However, including only Division, which has stronger relationship with tenders being EU funded, already strongly contributes to reducing confounder bias and requires "merely" sacrificing tens of thousands of observations. Fortunately, a good substitute for the 2-digit NUTS is country, which has a naturally strong correlation with NUTS (as NUTS refers to regions within a country). Including this variable in the matching process increases the explanatory power towards the treatment variable without sacrificing most of remaining observations. The last variable which explains at least 1% of the variance of tenders being EU

[5] Division refers to the industry based on the first two digits of main CPV.

funded is buyer authority, which is also the last variable to be included in the matching process with the second approach.

Variable	R^2 of bivariate regression with EU funded
Industry (division by 2-digit CPV)	11.57%
2-digit NUTS of buyer authority	10.07%
Country	6.96%
Type of buyer authority	1.28%
Description length of tender	0.069%
Supply type of tender	0.031%
Tender has lots	0.029%
Year	0.19%
Final price of tender in million €	0.00013%

Table 2: explanatory power of potential confounders regarding EU funding

In the following sections, the pre- and post-matching covariate balances are presented for each region with both matching approaches (Model 1 and Model 2).

Central and Eastern Europe

Model 1

In CEE, no substantial loss of observations has happened during the matching process (see Table 3). Countries with the highest imbalance were Poland, Slovenia, and Czech Republic. The most imbalanced buyer authorities (variable buyertype) were public body and national authority (see Appendix E).

	Control (non-EU-funded)	Treated (EU funded)
All	365,053	27,602
Matched (ESS)	64,079.77	27,601
Matched (unweighted)	364,546	27,601
Unmatched	507	1

Table 3: Effective sample sizes of control and treatment group (CEE, Model 1)

Model 2

Table 4 shows that during the matching process CEE experience a minimal loss of observations among EU-funded tenders, while non-EU-funded tenders have a moderate level of observation loss. The division with the highest pre-matching imbalance is clearly "Medical equipment, pharmaceuticals and personal care products". The highest pre-matching imbalance among buyer authority types was in the category public body; among countries it was Poland, Slovenia, and Czech Republic (see Appendix F).

	Control (non-EU-funded)	Treated (EU funded)
All	402,398	30,467
Matched (ESS)	965.67	29,830
Matched (unweighted)	335,256	29,830
Unmatched	67,142	637

Table 4: Effective sample sizes of control and treatment group (CEE, Model 2)

Southern Europe

Model 1

No substantial loss of observations has happened during the matching process in SE (see Table 5). Buyer authority with the highest imbalance was regional authority, nevertheless this variable was significantly more balanced in SE than in CEE. Countries with the highest imbalance were Italy and Greece (see Appendix G).

	Control (non-EU-funded)	Treated (EU funded)
All	117,209	7,997
Matched (ESS)	31,140.11	7,993
Matched (unweighted)	117,107	7,993
Unmatched	102	4

Table 5: Effective sample sizes of control and treatment group (SE, Model 1)

Model 2

In SE, there is a moderate loss of observations among both EU-funded and non-EU-funded tenders (see Table 6). Just like during the first matching in SE, buyer authority type with the highest pre-matching imbalance is regional authority, and among countries it is Italy and Greece. As for division, the most imbalanced category is the same as in CEE, which is medical equipment (see Appendix H).

	Control (non-EU-funded)	Treated (EU funded)
All	137,244	9,534
Matched (ESS)	3,522.78	8,643
Matched (unweighted)	107,434	8,643
Unmatched	29,810	891

Table 6: Effective sample sizes of control and treatment group (SE, Model 2)

Northern Europe

Model 1

As per Table 7, there is absolutely no loss of observations in NE among treated observations and only a little loss in the control group. Buyer authority types with the highest imbalance were public body, national agency, and national authority; among countries it was Finland, Latvia, and Lithuania (see Appendix I).

	Control (non-EU-funded)	Treated (EU funded)
All	82,064	3,925
Matched (ESS)	16,184.25	3,925
Matched (unweighted)	77,158	3,925
Unmatched	4,906	0

Table 7: Effective sample sizes of control and treatment group (NE, Model 1)

Model 2

In NE, there is significant loss of observations in both the control and treated group (see Table 8). Buyer authority type with the highest pre-matching imbalance is public body; among countries it is Finland again. Among divisions, "Repair and maintenance services" as well as "Medical equipment, pharmaceuticals and personal care products" were especially imbalanced (see Appendix J).

	Control (non-EU-funded)	Treated (EU funded)
All	89,599	5,302
Matched (ESS)	980.34	4,452
Matched (unweighted)	52,317	4,452
Unmatched	37,282	850

Table 8: Effective sample sizes of control and treatment group (NE, Model 2)

Western Europe

Model 1

In WE, similarly to NE, there is absolutely no loss of observations among EU-funded tenders and a low level of loss among non-EU-funded tenders (see Table 9). In general, covariate imbalances were the lowest in this region. Buyer authority types with the highest pre-matching imbalance are utilities, national authority, and other. The country with the highest imbalance was Netherlands (see Appendix K).

	Control (non-EU-funded)	Treated (EU funded)
All	65,889	2,939
Matched (ESS)	29,707.19	2,939
Matched (unweighted)	62,939	2,939
Unmatched	2,950	0

Table 9: Effective sample sizes of control and treatment group (WE, Model 1)

Model 2

In WE, there is a slight loss of observations among EU-funded tenders and a significant loss among non-EU-funded tenders. Covariate imbalances of country and buyer type are again the lowest in this region. Among divisions, "Transport services (excluding Waste transport)", "Repair and maintenance services" as well as "Medical equipment, pharmaceuticals and personal care products" were especially imbalanced before conducting the matching (see Appendix L).

	Control (non-EU-funded)	Treated (EU funded)
All	80,226	3,615
Matched (ESS)	1367.04	3,533
Matched (unweighted)	43,260	3,533
Unmatched	36,966	82

Table 10: Effective sample sizes of control and treatment group (WE, Model 2)

Regression analysis

Using matched data from the first matching approach (Model 1), the following regression model will be used to estimate CRI with main explanatory variable EU funding and covariates selected by random forest:

$$CRI \sim \alpha + \beta_1 * EUfunded + \beta_2 * final\ price\ of\ tender + \beta_3 * description\ length + \beta_4 * division + \beta_5 * buyer\ NUTS\ 2 * \beta_6 * year + u$$

Equation 1: Regression model on matched data (first matching approach)

Before conducting the analysis, Hypothesis 1 (finding a positive relationship between EU-funding and the extent of corruption risk) was expected to hold in Central and Eastern Europe and Southern Europe. In the remaining two regions, a non-significant (H_0) effect was expected.

After running the regression model for each region, the results presented in Table 11 are obtained.

	Central and Eastern Europe	Southern Europe	Northern Europe	Western Europe
EU-funded	1.718 *** (0.080)	0.654 *** (0.152)	-0.852 *** (0.162)	1.995 *** (0.190)
Observations	392,147	125,100	81,083	65,878
Adjusted R-squared	0.209	0.122	0.137	0.129

Table 11: Regional level regression results of Model 1

Both the number of observations (where information about EU and non-EU funded tenders is available) and the number of EU funded tenders is the highest in Central and Eastern Europe, which contributes to a better reliability of results in this region. As for goodness of fit, the explanatory power of the model is the highest in CEE (R^2 = 20.9%), followed by Northern Europe (R^2 = 13.7%), Western Europe (R^2 = 12.9%) and Southern Europe (R^2 = 12.2%). Both factors are advantageous for the analysis, as CEE is the region with the most EU funded tenders and the highest mean of CRI, making it also the most interesting region regarding the potential effect of EU funding on CRI.

According to the results of Model 1 (see Table 11), there is a statistically significant effect of EU-funding on the Corruption Risk Index in public procurement in all regions. The coefficient of 1.718 suggests that in CEE, EU-funded tenders bring, on average, 1.718 points higher corruption risk on a 0-100 scale, compared to tenders without any EU funding. Similarly, the coefficient of 0.654 suggests that in Southern Europe, EU-funded tenders bring, on average, 0.654 points higher corruption risk on a 0-100 scale, compared to tenders without any EU funding. At the same time, in Northern Europe, Model 1 indicates a negative effect: the coefficient of -0.852 suggests that in NE, EU-funded tenders bring, on average, 0.852 points lower corruption risk on a 0-100 scale. Contrary to expectations, in Western Europe, Model 1 finds a positive effect: the coefficient of 1.995 suggests that in WE, EU-funded tenders bring, on average, 1.995 points higher corruption risk on a 0-100 scale.

Nevertheless, the magnitude of the effect found in each region is very limited: none of them leads to more than 2 percentage points increase or decrease in the 0-100 scale corruption risk

index. Furthermore, when including EU-funding as the only independent variable in the regression model, the explanatory power (R^2) of the model goes below 1%, showing that the main source of the model's explanatory power comes from other covariates, such as industry (division based on 2-digit CPV codes), and 2-digit NUTS codes.

Country-level findings

When running the same regression (Model 1) on matched data for each country separately, there is a strong variance in the country-level sample sizes (see Table 12) mainly driven by the allocation of EU funds and less driven by country-level population.

	Austria	Belgium	Bulgaria	Croatia	Cyprus	Czech Republic	Denmark	Estonia
EU-funded	1.576 (1.440)	-0.289 (0.531)	0.634** (0.317)	0.405 (0.598)	2.060 (1.763)	2.033 *** (0.240)	-2.026 ** (0.808)	-1.114 *** (0.352)
Observations	1,107	8,958	23,360	13,937	2,044	21,510	3,681	7,317
Adjusted R-squared	0.254	0.211	0.236	0.157	0.217	0.254	0.138	0.183

	Finland	France	Germany	Greece	Hungary	Ireland	Italy
EU-funded	2.513 ** (1.026)	1.708 *** (0.214)	3.099 *** (0.442)	4.017 *** (0.281)	1.672 *** (0.338)	3.939 ** (1.619)	-2.075 *** (0.378)
Observations	4,202	32,932	19,284	21,991	17,752	1,059	47,392
Adjusted R-squared	0.140	0.080	0.161	0.131	0.189	0.357	0.135

	Latvia	Lithuania	Luxembourg	Malta	Netherlands	Poland	Portugal
EU-funded	-1.900 *** (0.422)	1.348 *** (0.248)	1.307 (1.634)	0.360 (1.446)	0.287 (1.565)	2.540 *** (0.115)	1.257 *** (0.483)
Observations	12,574	51,932	414	557	1,584	234,512	4,672
Adjusted R-squared	0.193	0.145	0.066	0.110	0.127	0.105	0.189

	Romania	Slovakia	Slovenia	Spain	Sweden
EU-funded	4.609 *** (0.188)	-2.507 *** (0.579)	- 7.631 *** (0.656)	-0.218 (0.243)	0.328 (2.256)
Observations	48,508	3,558	29,010	48,444	1,377
Adjusted R-squared	0.327	0.153	0.229	0.141	0.110

Table 12: Country-level regression results of Model 1

This is in line with the first theoretical assumption already indicated in the Introduction, that countries in Eastern Europe receive significantly greater proportions of discretionary funding from the EU. As a result, some samples from smaller countries of Western and Northern Europe are probably too small to acquire meaningful conclusions.

However, sample sizes in countries of the key region (CEE) are mostly sufficient to report results. First, countries with more than 5,000 observations and high levels of explanatory power ($R^2 \geq 0.2$) include Bulgaria, Czech Republic, Romania, and Slovenia. Out of all these countries except Slovenia, there is a significant positive effect of EU-funded tenders on CRI. Second, countries with more than 5,000 observations and significant positive coefficients include Czech Republic, Poland and Romania.

Significance of covariates

Final price of tender has a significant negative effect on CRI in CEE, and a significant positive effect in SE and NE – however, the magnitude of this effect is limited. Tender description length has a significant negative effect in CEE, and a significant positive effect in SE and WE.

	Central and Eastern Europe	Southern Europe	Northern Europe	Western Europe
Final price of tender (million €)	-0.033 *** (0.001)	0.0003 *** (0.0001)	0.005 *** (0.001)	-0.001 (0.001)
Tender description length	-0.0002 *** (0.00000)	0.0001 ** (0.00005)	0.00003 (0.0001)	0.001 *** (0.0001)

Table 13: Significance of covariates in regional level regressions (Model 1)

Divisions which have a significant positive effect on CRI in every region are „Education and training services", and „IT services: consulting, software development, Internet and support". In addition, divisions with a significant positive effect on CRI in nearly every region include „Research and development services and related consultancy services", „Security, fire-fighting, police and defence equipment", „Recreational, cultural and sporting services", „Software package and information systems", as well as „Health and social work services".

Robustness check

In order to test the robustness of the results, the same regression model with matched data (second matching approach, Model 2), where industry (division) is also included in the matching process, as an important predictor of the treatment variable (EU-funding). This way, at the cost of some further loss of observations, confounder bias may be reduced to an even further extent compared to the previous model.

$$CRI \sim \alpha + \beta_1 * EUfunded + \beta_2 * final\ price\ of\ tender + \beta_3 * description\ length + \beta_4 * division + \beta_5 * buyer\ NUTS\ 2 * \beta_6 * year + u$$

Equation 2: Regression model on matched data (second matching approach)

Running the same regression model after a differently designed matching process led to different results, which are also less in line with intuition and the findings of former research. Surprisingly, Model 2 finds a statistically significant negative effect of EU-funding on CRI in public procurement in three regions (CEE, SE, and NE); and no significant effect in WE (see Table 17).

	Central and Eastern Europe	Southern Europe	Northern Europe	Western Europe
EU-funded	-0.904*** (0.114)	-3.551*** (0.315)	-4.082*** (0.253)	-0.221 (0.448)
Observations	365,086	116,077	56,769	46,793
Adjusted R-squared	0.132	0.099	0.594	0.165

Table 14: Regional level regression results (Model 2 – same regression model, second matching approach)

Summary of results

The final CRI used for the analysis consists of three indicators: single bid, procedure type, and call for tender publication. It has a correlation coefficient of 43.5% with Transparency CPI and 40.4% with Eurobarometer CPI. As an outcome of random forest analysis, the most important variables (with regard to CRI) to be included in the regression model are the tender's final price, description length, division, NUTS of buyer authority, and year.

The regression results of Model 1 show statistically significant effect of EU funded tenders on corruption risk in three regions, including Central and Eastern Europe, Southern Europe and Western Europe**.** However, the magnitude of these impacts is small, none of them proposes more than 2 percentage points increase in corruption risk. EU-funding slightly improved the regression's goodness of fit, but it is not the main source of the model's explanatory power. Geographical location (Country or buyer authority NUTS) and industry (Division) explains a significantly higher proportion of the variance of the target variable (CRI) compared to whether the tender is EU-funded. The most surprising component of the result is that it identifies a higher corruption risk in WE.

When conducting robustness check with the same regression analysis applied after a different matching process, Model 2 shows different results, with EU funding having a statistically significant negative impact of EU funding on corruption risk in CEE, SE, and NE. These counter-intuitive results in CEE and SE shows that the findings of Model 1 are not robust, and some of the additional corruption risk observed among EU funded tenders may be caused by some industries that receive more EU funding being more corrupt.

The advantage of Model 1 is that most of its findings are in line with the theoretical assumptions of the thesis (outlined in the Introduction). Another beneficial point is that during the matching process, less observations were lost compared to Model 2. However, a limitation of this model is that it cannot rule out the possibility of some confounder bias, as not all variables are included in the matching process that influence EU funding.

The variable division – measuring the industry of the tender using the first two digits of CPV – can point out some industries, that may pose an additional corruption risk. Based on the CRI, which measures the lack of proper competition, seven industries emerged (see page 32). These areas were present in at least three of all four regions with a statistically significant positive effect on CRI at a 5% significance level. For further research, looking deeper into these areas may be beneficial.

Overall, the results of Model 1 and Model 2 make room for a few different interpretations and assumptions. Firstly, the overall magnitude of the effect of EU-funding on CRI is small in both models, in contrast to some key industries, which bring significantly higher levels of CRI compared to other industries. Second, the difference in Model 1 and Model 2 findings indicate

a lack of robustness. Third, even if one believes that EU-funded tenders bring a higher corruption risk, there are some potential factors which may hinder proving that hypothesis:

1) national or municipal authorities might not want EU institutions to find out about corrupt practices in EU-funded public procurement tenders, leading to more careful administration and more sophisticated forms of corruption;

2) the currently used CRI mostly measures the lack of competition and is not sensitive to more sophisticated forms of corruption.

In effect, in further research, it would be crucial to find even more detailed indicators to compose a CRI that may more successfully measure more sophisticated corrupt practices.

CONCLUSION

The research question of the thesis intended to uncover whether EU-funded tenders pose a higher corruption risk than national tenders. Even though the findings are mixed, there is a statistically significant difference between the level of corruption among EU and national tenders. The direction of this relationship depends on the model of choosing. The advantage of Model 1 is that most of its findings are in line with the theoretical assumptions of the thesis (outlined in the Introduction). Both Model 1 and the theoretical considerations suggest that CEE misuses EU-funds more than its own resources. The three assumptions are firstly, that CEE receives more EU-funding (which is clearly visible in the number of EU-funded observations), and the nature of these fundings are not earmarked therefore more susceptible to corruption (Mauro, 1998; Tanzi and Davoodi, 2001). Secondly, an additional pool of funding is appealing to rent-seeking politicians because it strengthens their political and business standing (Mingui-Pippidi, 2013). Thirdly, an external source of money can change the bureaucratical setup of a country for the benefit of rent-extracting politicians.

Furthermore, the results of Model 1 regarding CEE are confirmed by the cited literature in the section Literature Review. Batory (2018) emphasizes the deterministic importance of historical heritage in post-communist countries (which is contemporarily known as CEE region). Batory argues that despite the EU accession, the 'culture of corruption' is still existent, and citizens have a strong preference for a system where they can arbitrarily bend the rules. This observation is in line with the finding of Rand Corporation (2016), which found that the level of corruption has not changed much before and after the EU accession. This leads to a conclusion that ACAs could not fulfil their mandate, in other words, could not mitigate corruption. Additionally, the findings of Model 1 regarding CEE and SE do not contradict external CPIs (like TI and Eurobarometer), which measure the corruption perception of experts in their own countries.

What theories and CPIs cannot explain is the finding of Model 1 regarding greater misuse of EU-funds in WE. WE is generally perceived to be less corrupt, and countries in WE are net contributors to the EU budget (Buchholz, 2020). So even though EU-funds are considered to be external sources, the funds ultimately come from these countries' own national budget. The intuition would be that there should be no significant difference between EU and nationally funded corruption. Therefore, this finding cannot be explained by the same mechanism

operating in CEE and SE, where countries are net beneficiaries and regard EU-funds as an additional pool of corruption (as introduced in the second theoretical assumption).

The pitfalls of the analysis could lie in the created Corruption Risk Index. CRI captures – presumably well – the lack of competition, but it cannot detect more sophisticated forms of corruption. The mixed results may arise from more disguised forms of corruption in EU funded tenders. It might include instances where buyer authorities create an impression of real competition but select an overpriced bid and/or a winner that has common economic interests with the buyer of the tender. Another factor that hindered a more detailed analysis is the high rate of missing values (for instance 67.5% missing rate on the information about EU-funding and 99% missing rate for the variable new company). It narrows down the list of corruption red flags and the number of tenders that can be potentially used in the analysis. If missing values are not randomly distributed in the sample, the removal of observations with missing values in key variables can lead to a biased estimate, which might explain the mixed results.

The biggest potential for further research may lie in finding even more detailed indicators to compose a CRI that could signal more complex corrupt practices. A potential way to measure corruption in a more refined way is to enrich the official database with additional data points. Finding out more about the relationship between bidder company and buyer authority (as suggested by Aldana et al, 2022) as well as a realistic bid price for each tender may be key steps to better understand the drivers of corruption in public procurement. Broadening the database is certainly a time- and resource-intensive process, which means that it probably needs to be approached from a local level and be continuously extended to make more general findings.

BIBLIOGRAPHY

Aldana, Andrés, Andrea Falcón-Cortés, and Hernán Larralde. "A machine learning model to identify corruption in M\'exico's public procurement contracts." *arXiv preprint arXiv:2211.01478* (2022).

Batory, Agnes. "Chapter 10 Corruption in East Central Europe: has EU membership helped?". In *Handbook on the Geographies of Corruption*, (Cheltenham, UK: Edward Elgar Publishing, 2018) accessed May 7, 2023, https://doi.org/10.4337/9781786434753.00016

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. "cem: Coarsened exact matching in Stata." The Stata Journal 9, no. 4 (2009): 524-546.

Buchholz, Katharina. „Which Countries are EU Contributors and Beneficiaries?" Statista. (2020). Retrieved from: https://www.statista.com/chart/18794/net-contributors-to-eu-budget/

Chang, Eric CC, and Miriam A. Golden. "Electoral systems, district magnitude and corruption." *British journal of political science* 37, no. 1 (2007): 115-137.

Colonnelli, Emanuele, Jorge A. Gallego, and Mounu Prem. "What predicts corruption?." *Available at SSRN 3330651* (2020).

Council of Europe. "Group of States against Corruption. How does GRECO work?" (n.d.). Retrieved from: https://www.coe.int/en/web/greco/about-greco/how-does-greco-work

De Sousa, Luís. "Anti-corruption agencies: between empowerment and irrelevance." *Crime, law and social change* 53 (2010): 5-22

Eurobarometer. Businesses' Attitudes Towards Corruption in the EU. (2015) Retrieved from: https://europa.eu/eurobarometer/surveys/detail/2084

Eurobarometer. Businesses' Attitudes Towards Corruption in the EU. (2017) Retrieved from: https://europa.eu/eurobarometer/surveys/detail/2177

Eurobarometer. Businesses' Attitudes Towards Corruption in the EU. (2019) Retrieved from: https://europa.eu/eurobarometer/surveys/detail/2248

EuroVoc. "European sub-regions (according to EuroVoc, the thesaurus of the EU).png" Picture from Wikipedia. (2014). Retrieved from: https://en.wikipedia.org/wiki/File:European_sub-regions_%28according_to_EuroVoc,_the_thesaurus_of_the_EU%29.png

Fazekas, Mihály, and Gábor Kocsis. "Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data." British Journal of Political Science 50, no. 1 (2020): 155-164.

Fazekas, Mihály, Jana Chvalkovska, Jiri Skuhrovec, István János Tóth, and Lawrence P. King. "Are EU funds a corruption risk? The impact of EU funds on grand corruption in Central and Eastern Europe." The anticorruption frontline. The ANTICORRP project 2 (2014): 68-89.

Hafner, Marco, Jirka Taylor, Emma Disley, Sonja Thebes, Matteo Barberi, Martin Stepanek, and Mike Levi, The Cost of Non-Europe in the area of Organised Crime and Corruption: Annex II - Corruption. Santa Monica, CA: RAND Corporation, 2016. https://www.rand.org/pubs/research_reports/RR1483.html.

Hartmann Till, Johannes, Ferreyra Carlos. "What are the costs of corruption?". World Bank. (2022). Retrieved from: https://blogs.worldbank.org/governance/what-are-costs-corruption

Hlavac, Marek. Stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. (2022). https://CRAN.R-project.org/package=stargazer

Hrubý Jan, Pošepný Tomáš, Krafka Jakub, Toth Bence, Skuhrovec Jiří. "D2.8 Methods Paper." The Digital Whistleblower: Fiscal Transparency, Risk Assessment and the Impact of Good Governance Policies Assessed. (2018). Retrieved From: https://digiwhist.eu/wp-content/uploads/2018/03/D2.8-revised-version-FINAL.pdf

IMF (2019). Fiscal Monitor: Curbing Corruption. Retrieved from https://www.imf.org/en/Publications/FM/Issues/2019/03/18/fiscal-monitor-april-2019

Klitgaard, Robert. *Controlling corruption*. Univ of California Press, 1988.

Lima, Marcio Salles Melo, and Dursun Delen. "Predicting and explaining corruption across countries: A machine learning approach." *Government Information Quarterly* 37, no. 1 (2020): 101407.

Lipset, S. M. and Man, P. "The social bases of politics". Baltimore: The Johns Hopkins UniversityPress. (1960).

Mackey, Tim K., and Bryan A. Liang. "Combating healthcare corruption and fraud with improved global health governance." *BMC international health and human rights* 12 (2012): 1-7.

Martin-Russu, Luana. "Introduction: The European Paradox of Expecting Corrupt Political Elites to Lead the Fight Against Corruption." In *Deforming the Reform: The Impact of Elites on Romania's Post-accession Europeanization*, pp. 1-17. Cham: Springer International Publishing, 2022.

Mauro, P. Corruption and the composition of government expenditure. Journal of Public Economics, 69, (1998). 263–279.

Mungiu-Pippidi, A. (Ed.). Controlling Corruption in Europe. The Anticorruption Report 1. Berlin: Barbara Budrich Publishers. (2013).

Opentender Portals. Making public tenders more transparent. DIGIWHIST Project. (n.d.) Retrieved from: https://opentender.eu/start 11

Petheram, André, Walter Pasquarelli, and Richard Stirling. *The next generation of anti-corruption tools: Big data, open data & artificial intelligence*. Tech. rep., Oxford Insights, 2019.

Rose-Ackerman, Susan. "Altruism, nonprofits, and economic theory." *Journal of economic literature* 34, no. 2 (1996): 701-728.

Shleifer, A. and Vishny, R. W. (1993). Corruption, The quarterly journal of economics 108(3): 599–617.

Søreide, Tina. *Corruption in public procurement. Causes, consequences and cures*. Chr. Michelsen Intitute, 2002. https://www.cmi.no/publications/file/843-corruption-in-public-procurement-causes.pdf

Tanzi, V., & Davoodi, H. "Corruption, growth, and public finances." In A. K. Jain (Ed.), The Political Economy of Corruption New York: Routledge. (2001). 89–110.

Thompson, Dennis F. "Two concepts of corruption: Making campaigns safe for democracy." *Geo. Wash. L. Rev.* 73 (2004): 1036.

Transparency International. "Corruption Perception Index". (n.d.) Retrieved from: https://www.transparency.org/en/cpi/2021

Transparency International. THE ABCS OF THE CPI: HOW THE CORRUPTION PERCEPTIONS INDEX IS CALCULATED. (2021). Retrieved from: https://www.transparency.org/en/news/how-cpi-scores-are-calculated

Wachs, Johannes, Mihály Fazekas, and János Kertész. "Corruption risk in contracting markets: a network science perspective." *International Journal of Data Science and Analytics* 12 (2021): 45-60.

World Bank. DataBank. (nd) Retrieved from: https://databank.worldbank.org/databases/control-of-corruption

APPENDIX

Appendix A: Post-matching annual CRI figures in Southern Europe (Model 1)

Figure A1 shows that tenders with EU funding have roughly equal mean CRI in 2014 compared to non-EU-funded tenders, while they seem to be more corrupt in 2015, 2017, and 2018. However, their mean CRI is smaller in 2016. The majority of observations stem from the earlier years, among which there is only one year with a lower mean CRI among EU-funded tenders.

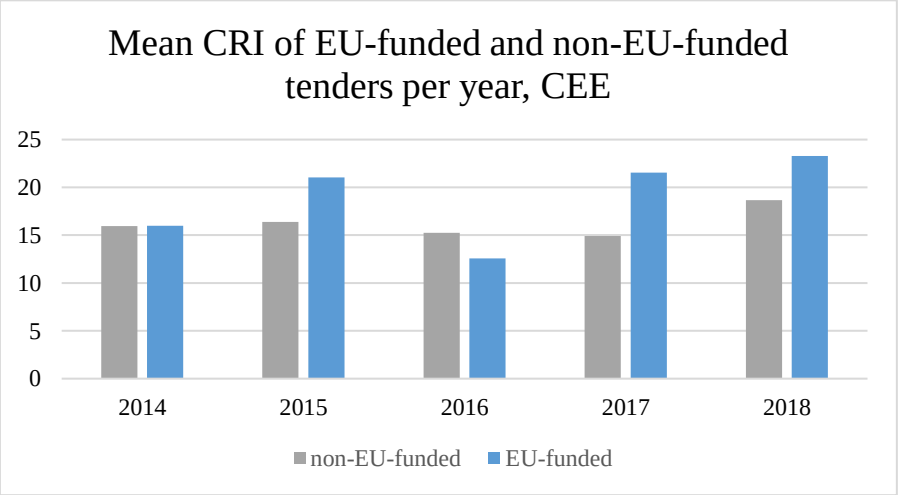

Figure A1: Average CRI of EU funded and non-EU-funded tenders after matching in CEE (2014-2018)

The number of observations in 2019 and 2020 is too few to report an adequate mean CRI – for instance, only 334 EU-funded tenders are in the matched sample from 2020 (see Figure A2).

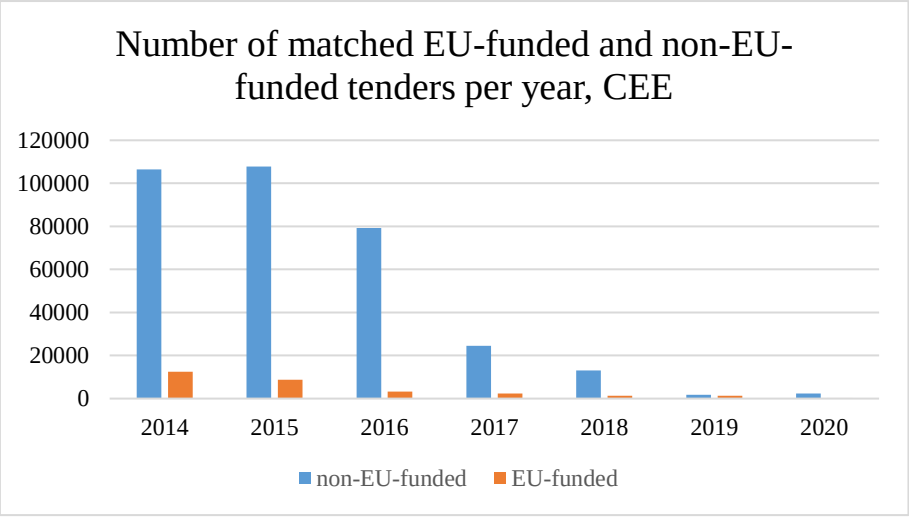

Figure A2: Number of EU funded and non-EU-funded tenders in CEE after matching, 2014-2020

Appendix B: Post-matching annual CRI figures in Southern Europe (Model 1)

Figure B1 illustrates that EU-funded tenders have slightly lower mean CRI in 2014 and 2017 than non-EU-funded ones, significantly lower in 2016, strongly higher in 2015, and slightly higher in 2018.

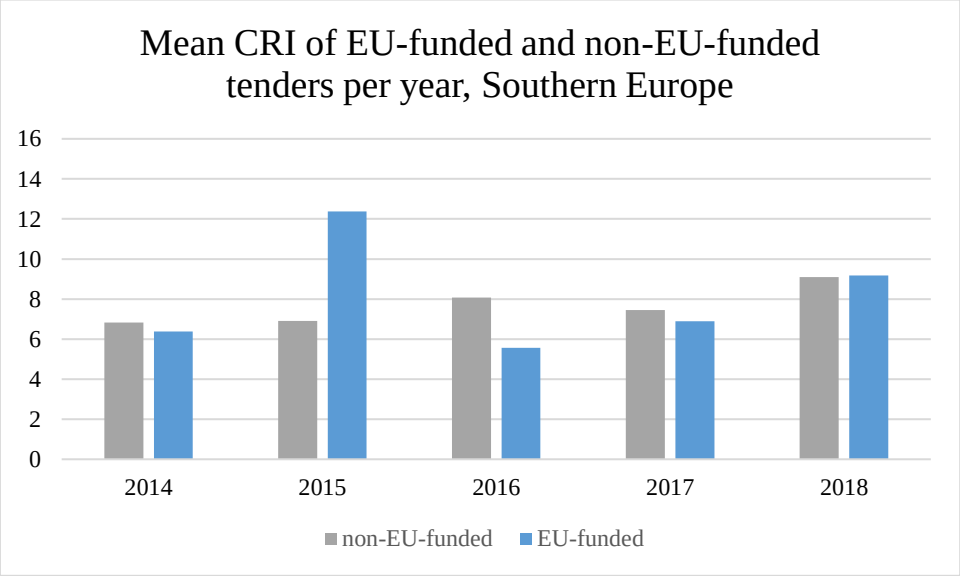

Figure B1: Average CRI of EU funded and non-EU-funded tenders after matching in SE (2014-2018)

In SE too, the level of observations in 2019 and 2020 is too low for reporting a proper mean CRI – for instance, only 22 EU-funded tenders are in the matched sample from 2020 (see Figure B2).

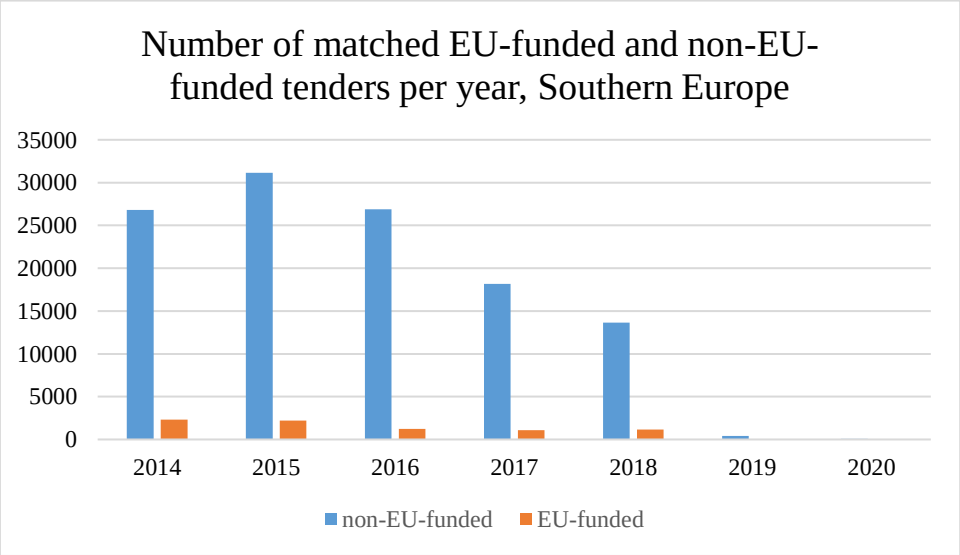

Figure B2: Number of EU funded and non-EU-funded tenders in SE after matching, 2014-2020

Appendix C: Post-matching annual CRI figures in Northern Europe (Model 1)

EU-funded tenders show higher mean CRI in 2014, 2015 and 2017 compared to non-EU-funded tenders; and slightly lower in 2016 (see Figure C1).

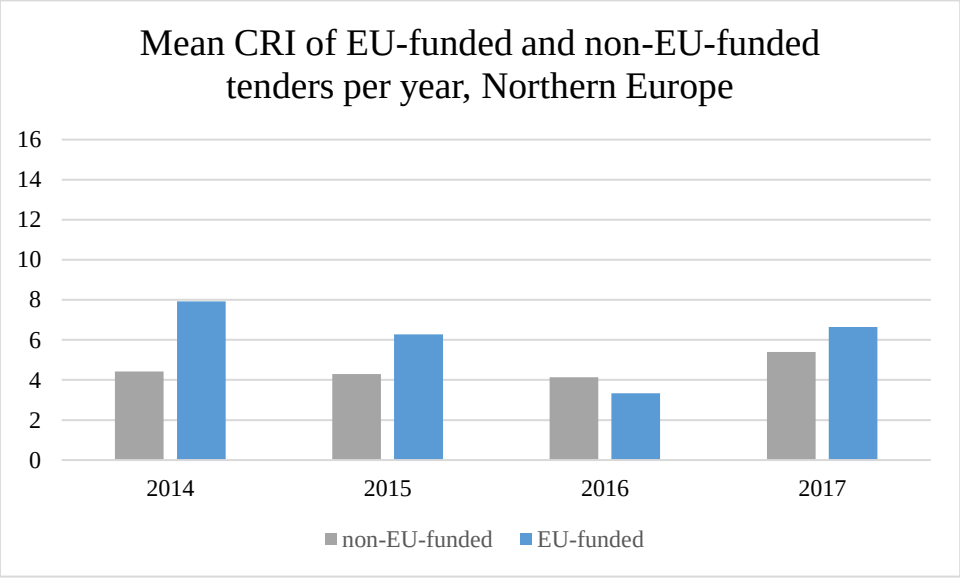

Figure C1: Average CRI of EU funded and non-EU-funded tenders after matching in SE (2014-2017)

The number of observations between 2018-2020 is too low in NE for reporting a proper mean CRI – for instance, only 3 EU-funded tenders are in the matched sample from 2018 (see Figure C2).

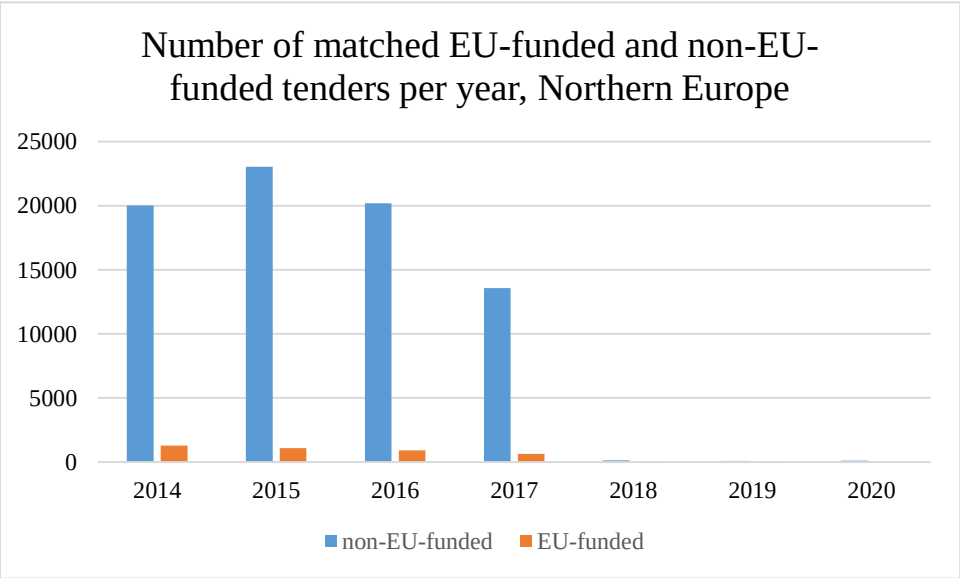

Figure C2: Number of EU funded and non-EU-funded tenders in NE after matching, 2014-2020

Appendix D: Post-matching annual CRI figures in Western Europe (Model 1)

EU-funded tenders show significantly higher WE mean CRI in 2014 and 2016 compared to non-EU-funded tenders, significantly lower in 2017, and slightly higher in 2015 (see Figure D1).

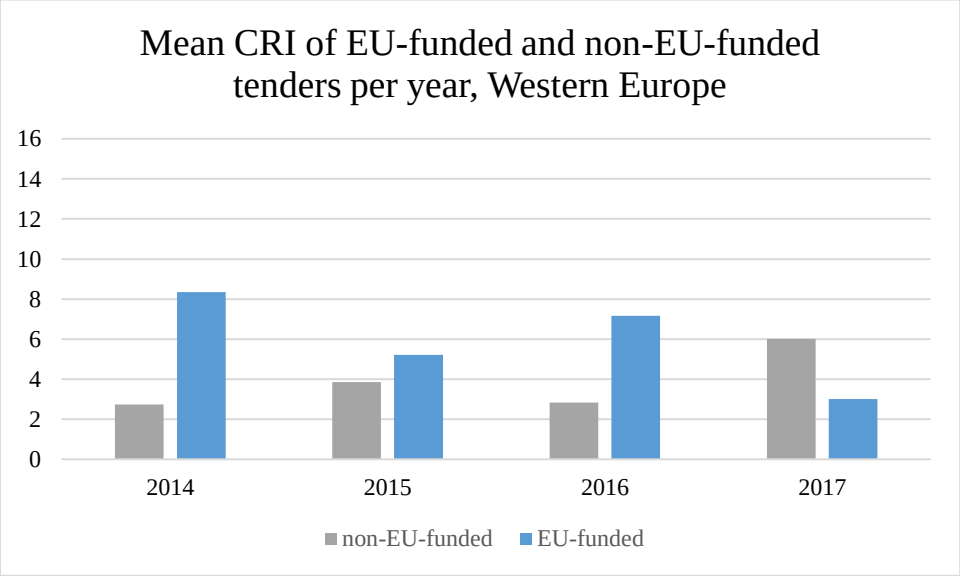

Figure D1: Average CRI of EU funded and non-EU-funded tenders after matching in WE (2014-2017)

The number of observations between 2018-2020 is too low in WE for reporting an adequate mean CRI (see Figure D2).

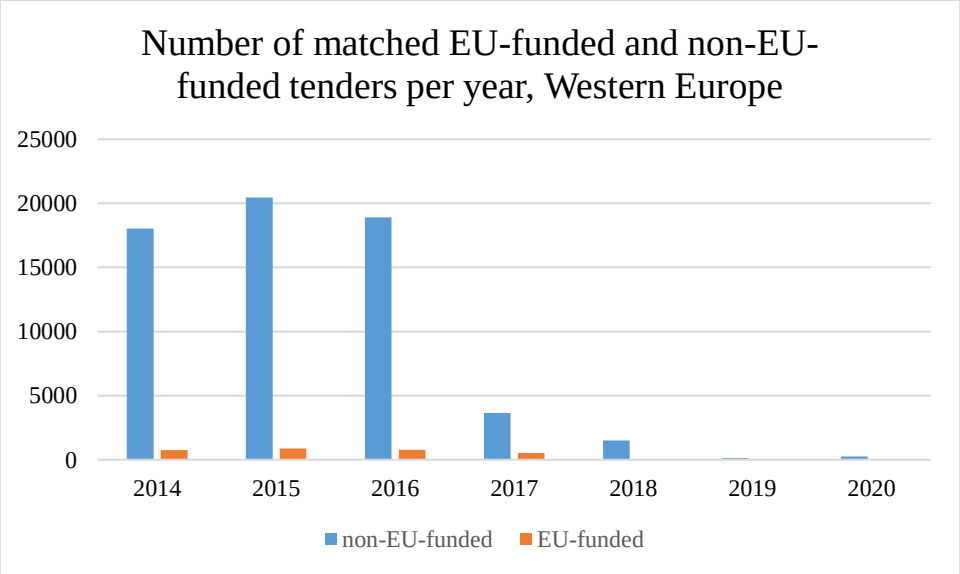

Figure D2: Number of EU funded and non-EU-funded tenders in WE after matching, 2014-2020

Appendix E: Covariate balance in Central Eastern Europe before and after matching (Model 1)

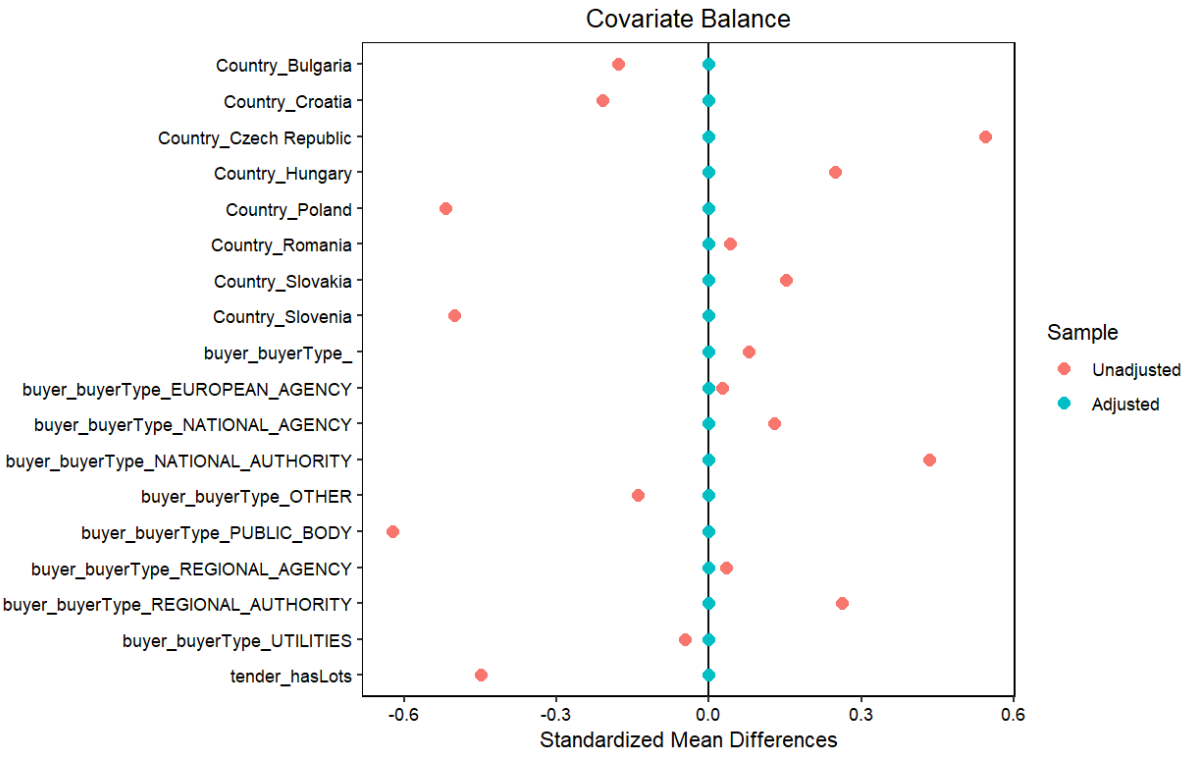

Figure E1: Covariate balance in Central Eastern Europe before and after matching (Model 1)

Appendix F: Covariate balance in Central Eastern Europe before and after matching (Model 2)

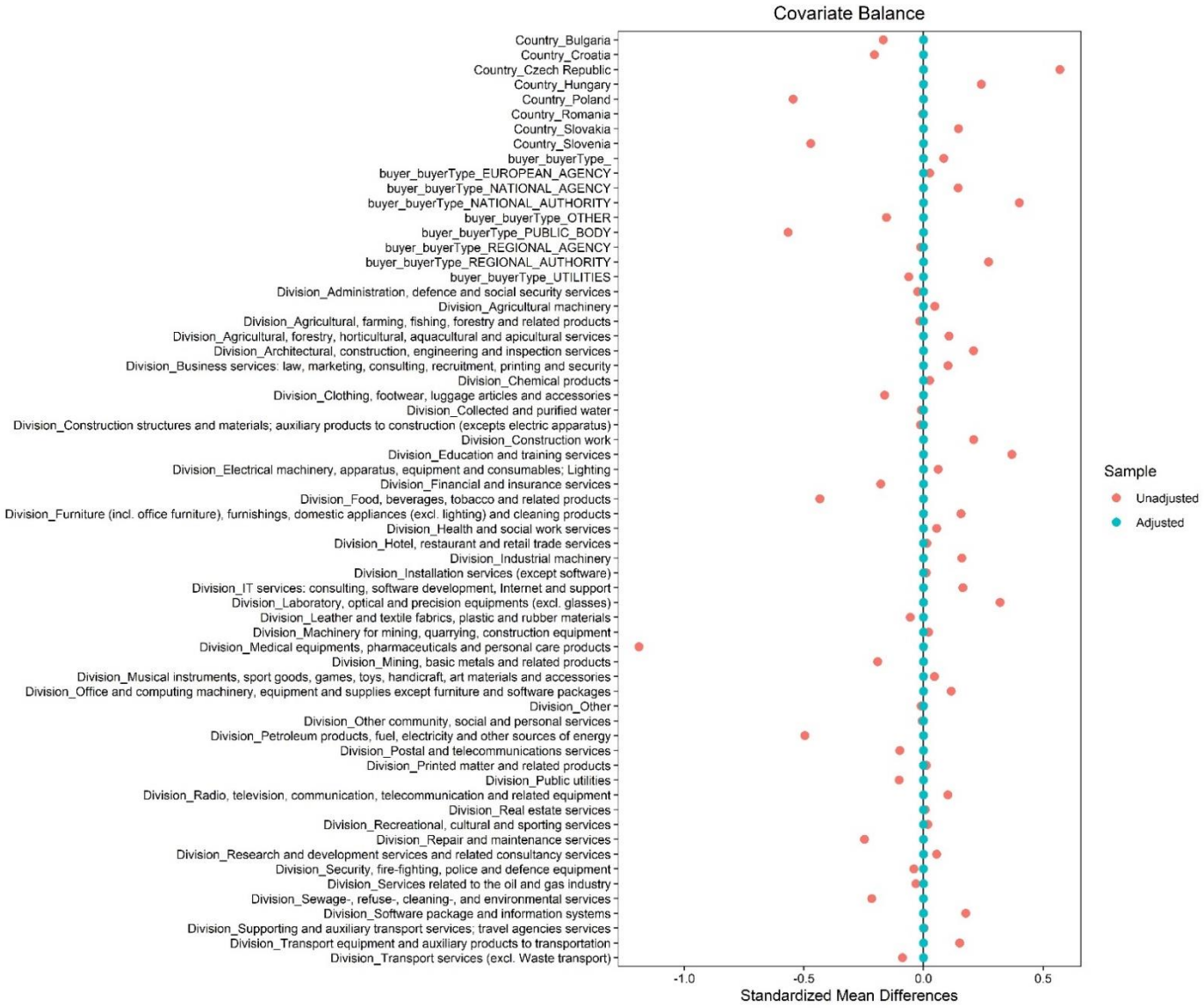

Figure F1: Covariate balance in Central Eastern Europe before and after matching (Model 2)

Appendix G: Covariate balance in Southern Europe before and after matching (Model 1)

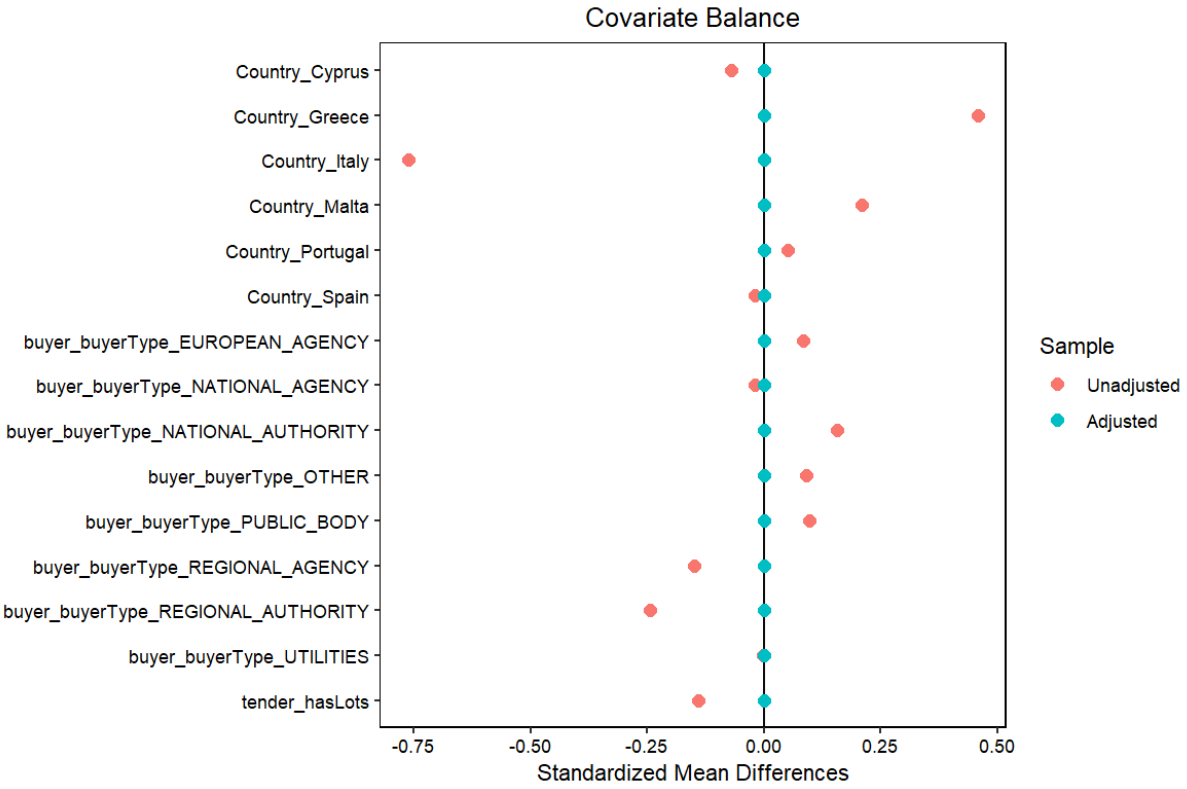

Figure G1: Covariate balance in Southern Europe before and after matching (Model 1)

Appendix H: Covariate balance in Southern Europe before and after matching (Model 2)

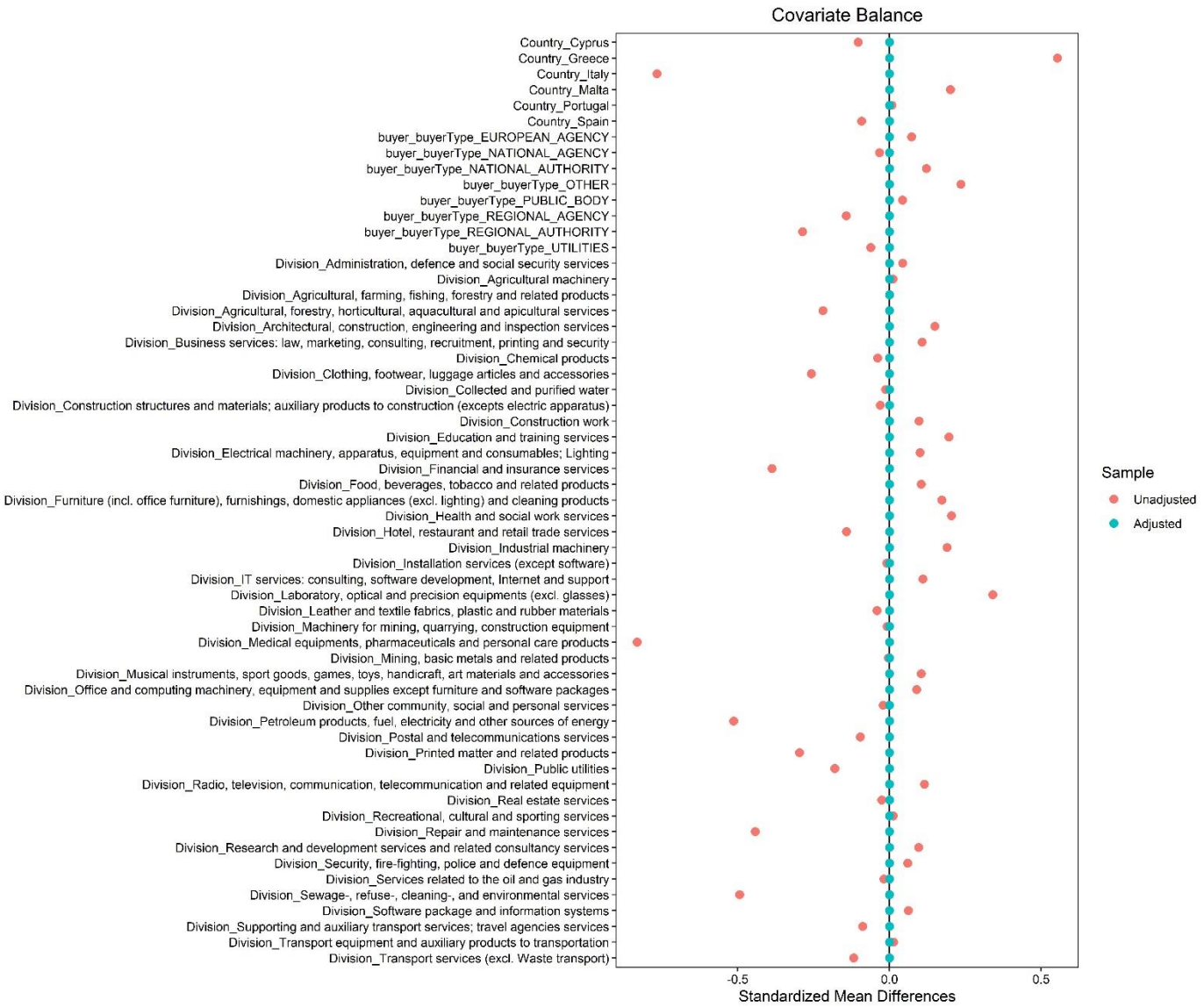

Figure H1: Covariate balance in Southern Europe before and after matching (Model 2)

Appendix I: Covariate balance in Northern Europe before and after matching (Model 1)

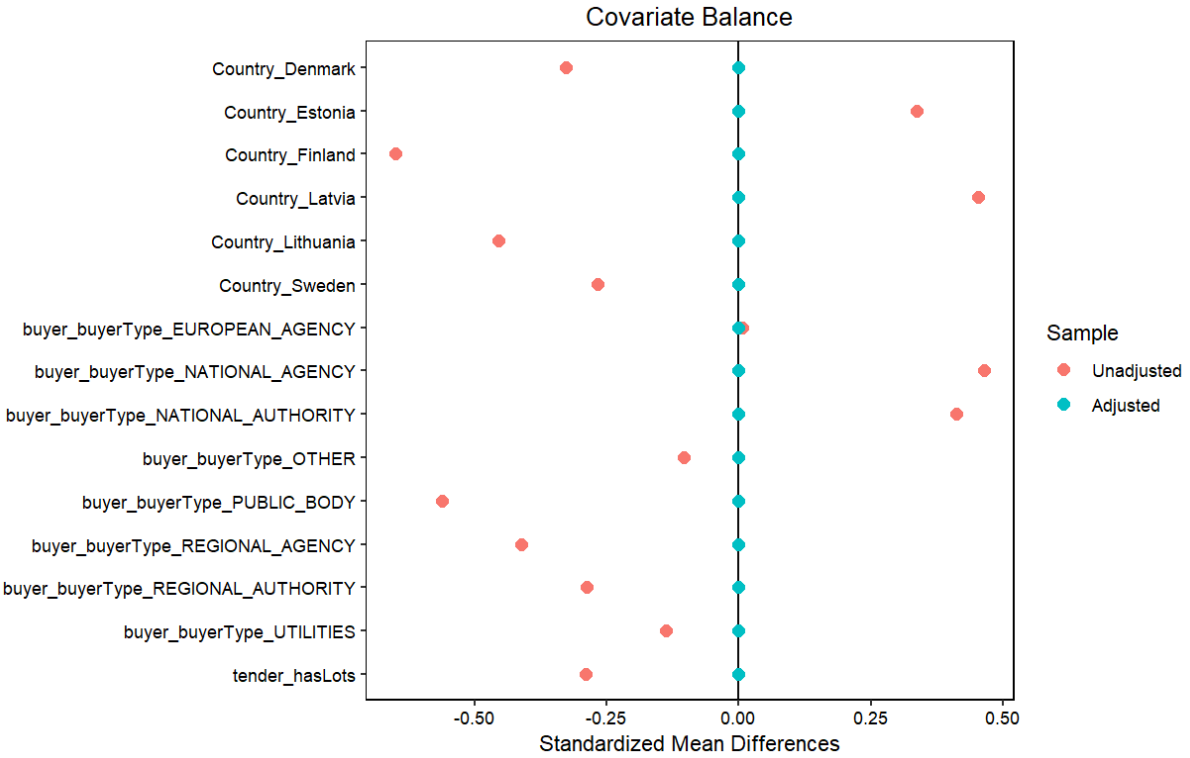

Figure I1: Covariate balance in Northern Europe before and after matching (Model 1)

Appendix J: Covariate balance in Northern Europe before and after matching (Model 2)

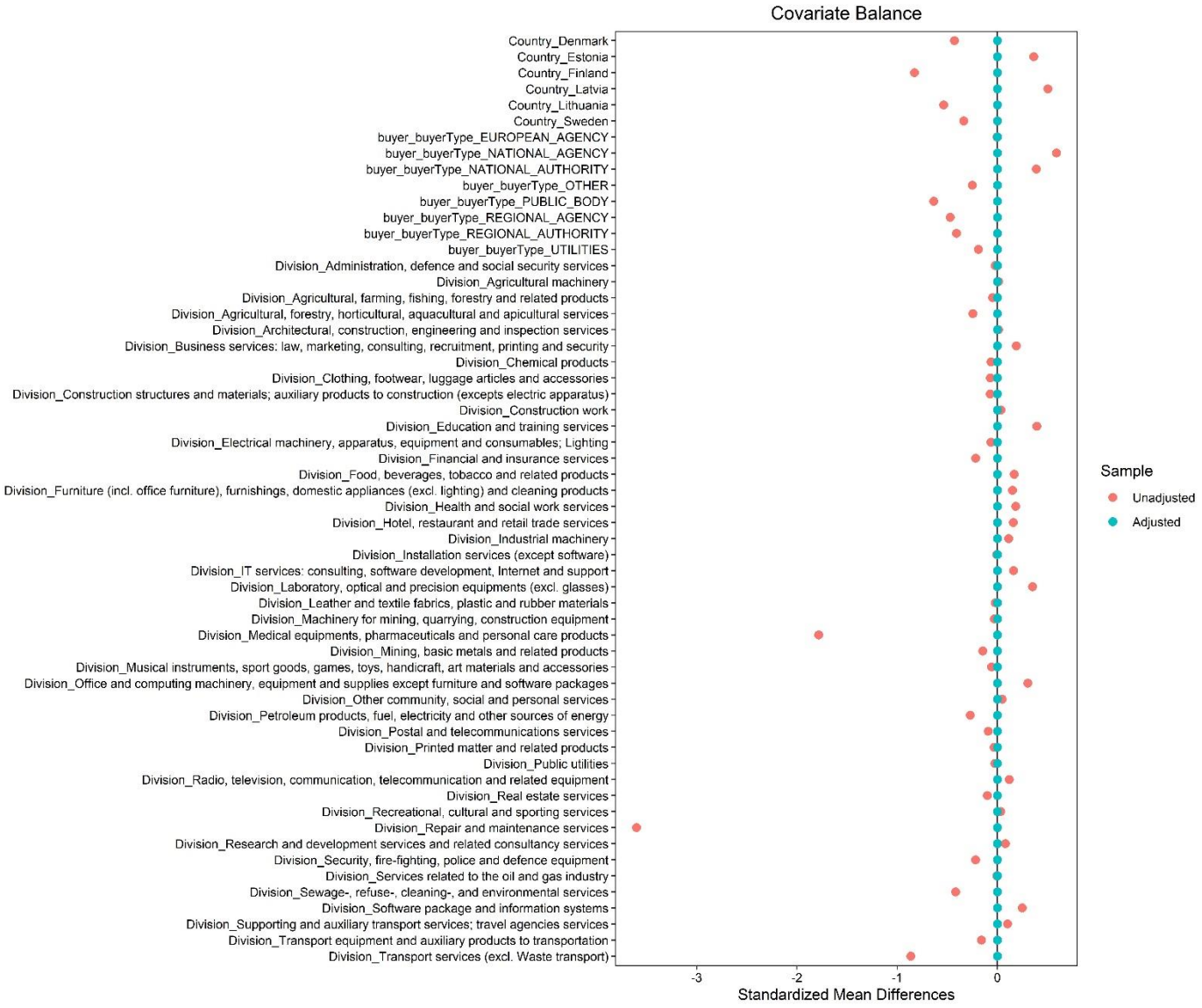

Figure J1: Covariate balance in Northern Europe before and after matching (Model 2)

Appendix K: Covariate balance in Western Europe before and after matching (Model 1)

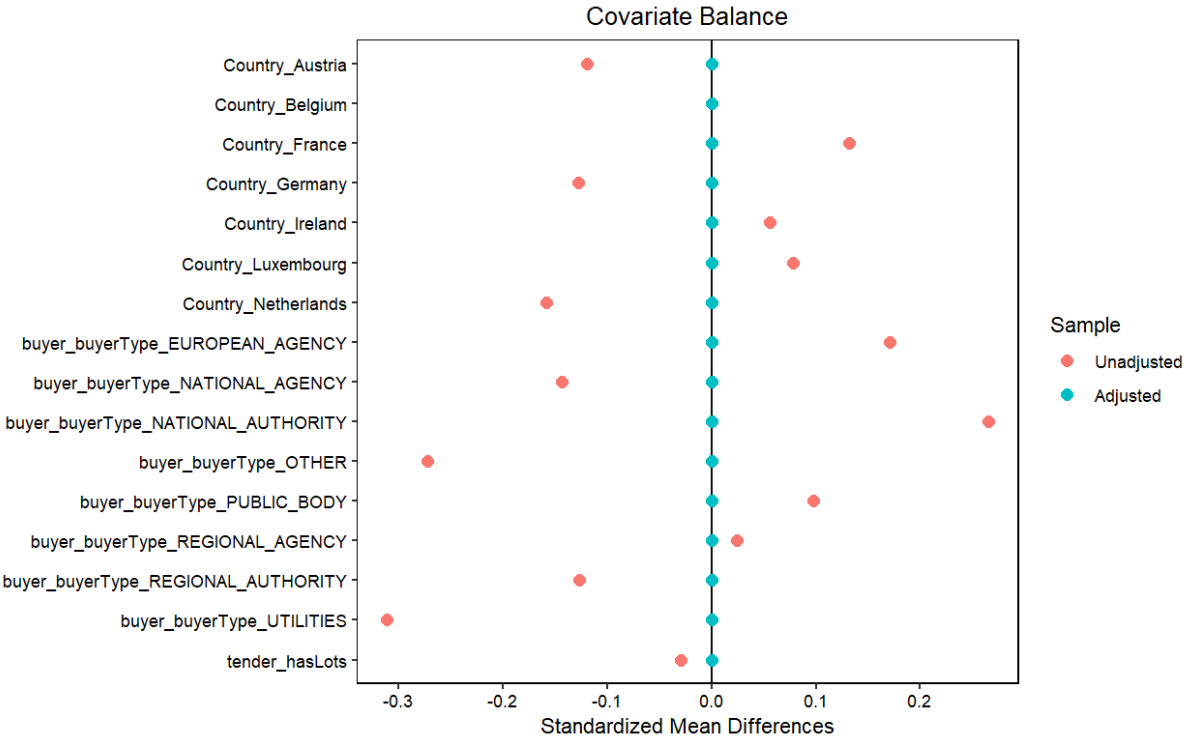

Figure K1: Covariate balance in Western Europe before and after matching (Model 1)

Appendix L: Covariate balance in Western Europe before and after matching (Model 2)

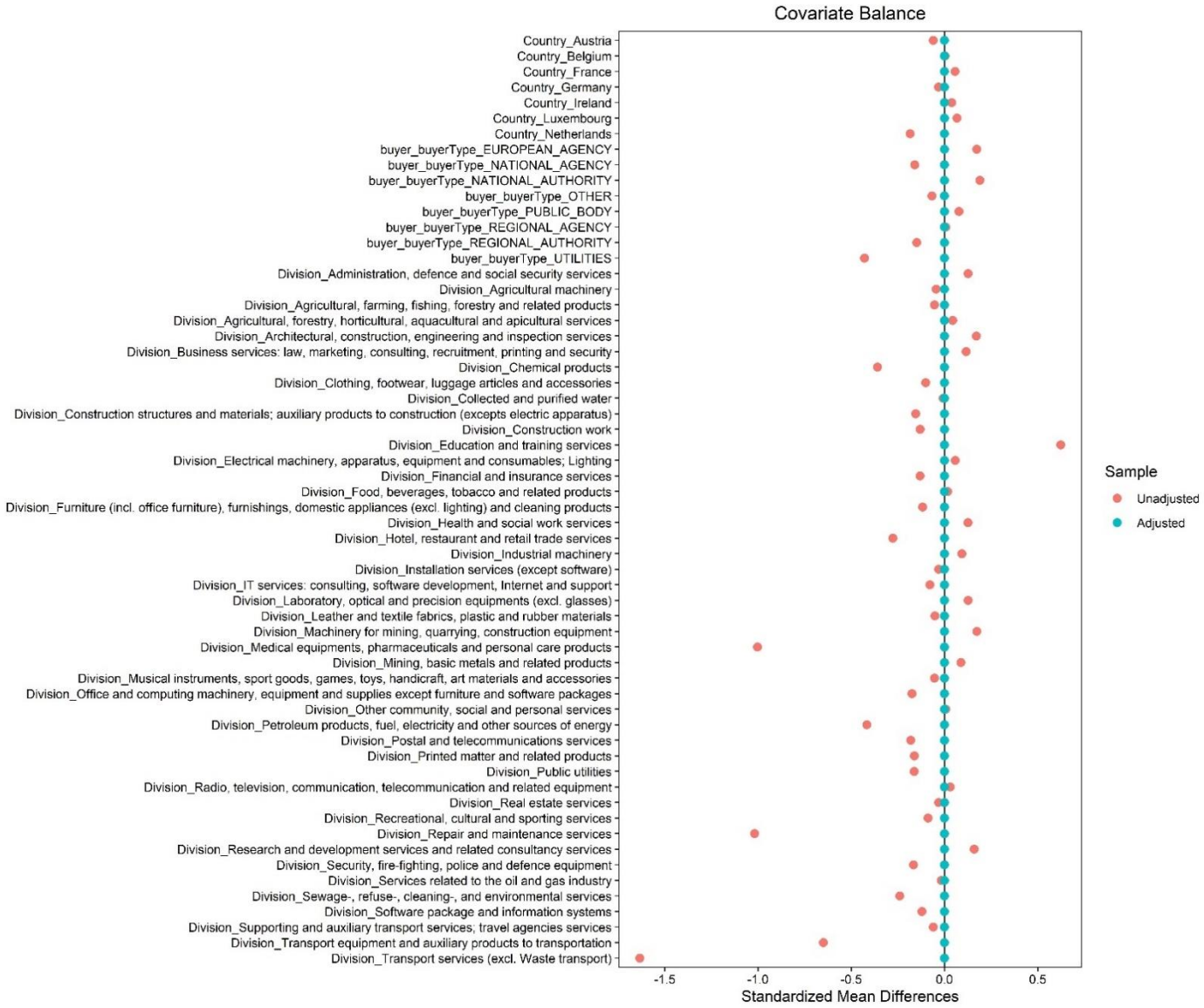

Figure L1: Covariate balance in Western Europe before and after matching (Model 2)