International
Institute of
Social Studies

Determinants of Measurement Errors in Immigration Flow Statistics:

A Case of Nine Central & Eastern European Countries

A Research Paper presented by:

Arpan Sagar

(India)

in partial fulfilment of the requirements for obtaining the degree of

MASTER OF ARTS IN DEVELOPMENT STUDIES

Major:

Economics of Development

Specialization:

Econometric Analysis of Development Policies

Members of the Examining Committee:

Matthias Rieger

Zemzem Shigute Shuka

The Hague, The Netherlands
December 2022

Disclaimer:

This document represents part of the author's study programme while at the International Institute of Social Studies. The views stated therein are those of the author and not necessarily those of the Institute.

Inquiries:

International Institute of Social Studies
P.O. Box 29776
2502 LT The Hague
The Netherlands

t:	+31 70 426 0460
e:	info@iss.nl
w:	www.iss.nl
fb:	http://www.facebook.com/iss.nl
twitter:	@issnl

Location:

Kortenaerkade 12
2518 AX The Hague
The Netherlands

Contents

List of Tables

List of Maps & Figures

List of Appendices

Acronyms

BG	Bulgaria
CEECs	Central and Eastern European Countries
CEPII	Centre d'Etudes Prospectives et d'Informations Internationales
CZ	Czech Republic
EE	Estonia
EU	European Union
EU-LFS	EU Labour Force Survey
GDP	Gross Domestic Product
GERD	Gross expenditure on R&D
HU	Hungary
IMEM	Integrated Modelling of European Migration
IMME	Implicit Minimal Measurement Error
LT	Lithuania
LV	Latvia
ME	Undercount Indicator
NSO	National Statistical Office
OLS	Ordinary Least Squares
PIN	Personal identification number
R&D	Research & Development
REWB	Within-Between Random Effects Model
RO	Romania
SEEMIG	Managing Migration and its Effects in South-East Europe
SI	Slovenia
SK	Slovakia
THESIM	Towards Harmonised European Statistics on International Migration
UN	United Nations
VDEM	Variety of Democracies
WDI	World Development Indicators
WGI	Worldwide Governance Indicators
WPP	World Population Prospects database

Abstract

This study analyses the determinants of measurement errors in immigration flow statistics in nine Central and Eastern European Countries (CEECs) between 2007 and 2019. This research is based on immigration flow statistics reported to Eurostat and a synthetic dataset of immigration flows generated as a part of this study. Based on these, a within-between random effects (REWB) model is employed to analyse the effects of public sector corruption, government effectiveness and research and development (R&D) in the government sector on the level of undercount and accuracy in available migration statistics. Additionally, an auxiliary analysis is conducted using a random effect model to identify if administrative register maintenance costs, dynamic migration processes and financial allocations can be attributed to missing disaggregated immigration flow data.

This study identifies that there exists a non-linear relationship between public sector corruption and the measurement errors. Initially, the measurement errors drop with rising levels of corruption and increase beyond a specific threshold, partially supporting the perception of public sector corruption as an act of 'greasing the wheels'. While public sector corruption is shown to decrease the measurement error, the directionality is attributed to overall low levels of corruption amongst the CEECs and the fact that the inflection point lies in the tail end of our sample. An additional conflicting finding of the study is that R&D expenditure seems to increase the level of inaccuracy and that government effectiveness has no effect on the measurement errors in the given context. However, the analysis of missingness reveals that increases government sector R&D expenditure does reduce the probability of missing data, though, the results hold only if countries with no missing data are excluded from the sample.

Relevance to Development Studies

The topic is highly relevant to the subject of Development Studies given the dominance of migration in the European policy discourse. Policymaking in the European Union (EU) is dependent on the available migration statistics which often suffer from measurement errors. Making decisions and designing policies based on these possibly inaccurate figures can result in ineffective migration governance. Furthermore, the importance of the relationship between migration and development is also widely recognised across academia with a large body of literature focussed on identifying the underlying determinants of migration. As a result, conceptualizing the types of measurement errors prevalent in migration statistics and identifying their determinants can help improve the robustness of

current migration research in addition to contributions towards more effective migration policymaking.

Keywords

Migration, migrants, measurement error, undercounting, accuracy, missing data, determinants

Chapter 1: Introduction

Over the past three decades, the subject of migration has assumed centre stage in the European Union's policy discourse (Willekens, 2019). Given that effective policymaking relies on the quality of available data, the demand for timely and accurate migration statistics has steadily increased (Willekens, 2019). However, migration processes, especially immigration flows, are quite dynamic, making them extremely hard to measure, which has resulted in the production of statistics lacking in quality (Raymer et al., 2013). A variety of interventions aimed at streamlining data production systems and improving technical skills of government officials have been undertaken but errors in immigration flow statistics still persist (Raymer et al., 2013). A key reason behind the same is the poor quality of administrative data sources used to derive flow statistics (Willekens, 2019).

Consequently, a large volume of demographic research has been undertaken to generate error-free datasets of immigration flows, by making *a priori* assumptions about the prevalent measurement errors or by using expert opinions and reconciliation techniques (see Bijak & Wiśniowski, 2010; Keilman & Aristotelous, 2021; Rampazzo et al., 2021; Raymer et al., 2013; Wiśniowski et al., 2019). However, there is no literature that we know of, that tests the mechanisms by which the quality of administrative data (upstream data) might change and their subsequent impact on the reported statistics.

The need for harmonized and error-free migration statistics has only recently gained prominence in Central and Eastern European countries (CEECs), which makes the region an interesting subject of analysis. The demand has been driven by the recognition of reliability issues in population registers (and amongst other administrative data sources) and a shift towards proactive migrant integration policymaking (Raymer et al., 2013; Gárdos and Gödri, 2014; Solano & Huddleston, 2020; Keilman & Aristotelous, 2021). This has led to the implementation of projects such as the Managing Migration and its Effects in South-East Europe[1] (SEEMIG) project in 2012 which aimed to outline the limitations of domestic statistical systems and provided policy recommendations for their improvement (Gárdos & Gödri, 2014). Certain countries are excluded from our analysis due to geography, migration dynamics, and data limitations, which are discussed in more detail in Chapter 2. The specific research questions of this paper are as follows:

[1] Despite the name suggesting that the project focuses only on South-eastern and Eastern Europe, Central European countries such as Slovakia, Slovenia, Austria, and Hungary were also included within the project's scope.

- What are the types of measurement errors prevalent in immigration flow statistics?
- How do determinants of upstream data quality affect the reported immigration flow statistics in CEECs?

This study starts by contextualizing the dynamics of immigration flows to CEECs and the processes undertaken to produce migration statistics. Post which we scrutinize various techniques that can be utilized to generate alternative flow estimates, but in the end, settle for a combination of the simplistic stock difference and Dennett's (2015) migration rate approaches. This is followed by the conceptualization of the forms of measurement errors based on existing literature and the indicators required to quantify them. A Within Between Random Effects (REWB) model, as developed by Bell and Jones (2015) and Allison (2009), is employed to analyze the institutional determinants of measurement errors on a panel dataset from 2007 to 2019. We find that public sector corruption leads to a 4.2 and 1.4 percent reduction in undercounting of migrants and inaccuracy of available statistics, respectively. Although the overall results might seem counter-intuitive, we observe a non-linear relationship between the variables as initial increases in public sector corruption decreases the measurement errors, however, beyond a specific threshold, the measurement errors begin to rise. We do not find any effect for government sector R&D expenditure, or the quality of public services administered (government effectiveness). Our results are robust to sample restrictions and model extensions bar a few marginal changes. Lastly, we conduct an analysis of missing data using a linear probability model with random effects and find a large negative effect of gross expenditure on government sector R&D on the probability of missing flow statistics. However, the results hold only when the sample is restricted to countries with some missing data.

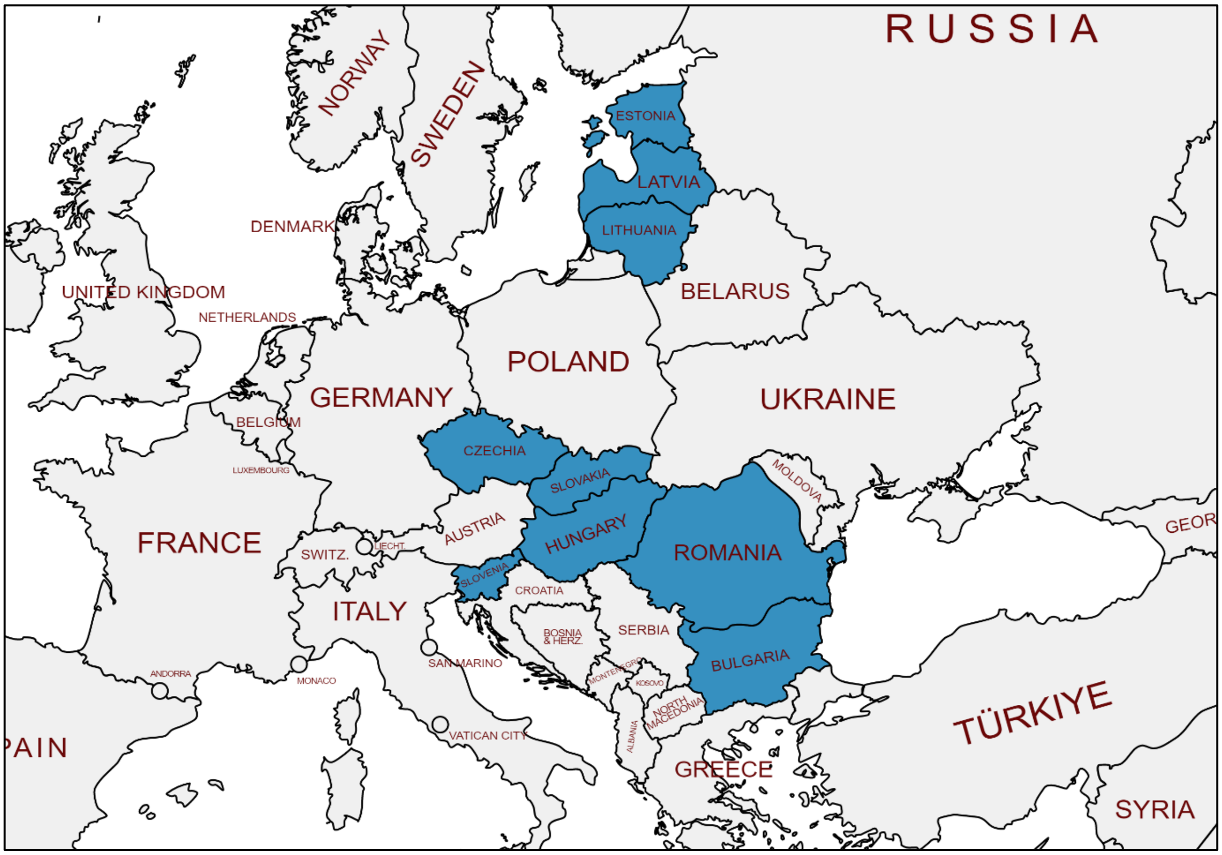
The rest of this paper is structured as follows: Chapter 2 provides contextual insights into the migration dynamics, country selection criteria, and the data production process of immigration flow statistics for CEECs. Chapter 3 lays out the conceptual framework, which delves into the types of measurement errors and identifies their potential causes. Chapter 4 sheds some light on existing literature in the field of measurement errors in demographic data, while Chapter 5 provides insights into the methodology. A brief description of the data is provided in Chapter 6, with Chapter 7 presenting and discussing the study's results. Finally, Chapter 8 concludes.

Chapter 2: Context

2.1 Dynamics of Immigration Flows

As a precursor to the discussion on the types of measurement errors, we must contextualize the dynamics of migration flows and the data production systems employed by the various CEECs. Map 1 presents the CEECs that fall within the geographical scope of this study. These countries can be subdivided into two regions, which are Central and Eastern Europe respectively. Given the diversity in their socio-political history and migration processes, we further divide the Eastern European countries into the Baltic States and Southeast Europe.

Map 1: Map of Europe outlining Geographical Scope of the Study



Source: Author's Elaboration

2.1.1 Eastern Europe

2.1.1.1 The Baltic States

To better understand the dynamics of migration flows to the region, we must historically and geographically place the three Baltic States, namely, Estonia, Latvia, and Lithuania. The Baltic States are located in North-Eastern Europe and are bounded by the Baltic Sea on their western border, which lends them their name. The southern and eastern borders, however, are entirely shared with Russia, Belarus and Poland.

As a part of the western frontier of the Soviet Union, the Baltic states were a key component of the post-World War 2 Soviet industrialization project (Kovalenko et al., 2010). This led to large volumes of immigration of ethnic Russians and other Russian speakers to the three countries. As a result, the percentage of the indigenous population shrank rapidly (Kovalenko et al., 2010). Often classified as old immigrants or recognized non-citizens, these population subsets play a crucial role in the region's minority discourses. Post the dissolution of the Soviet Union, immigration would not gain political relevance until the countries' accession to the EU in 2004 (Kovalenko et al., 2010). The accession led to a marked increase westward movement of domestic populations and the utilization of the Baltic States as transit points for incoming migrants and asylum seekers. While emigration from the three countries was initially described as temporary and circular, the 2008 financial crisis incited fears amongst diasporas due to the rising unemployment levels in their home countries, which in turn, led to greater permanent emigration (Birka, 2019).

Even in the wake of a shrinking domestic population, the socio-political responses towards migration have been varied and polarizing. Out of the three Baltic States, Estonia seems to have taken the most proactive approach towards encouraging immigration as a potential solution. In this regard, the country has revised its migration governance legislation five times since 2008, with the intent of making Estonia a viable option in the migrant labour market. (Birka, 2019). The prospects of migrating to the country have also improved due to a shift in domestic perceptions of migrants. This shift has been led by policy reforms for migrant integration and the expansion of the technology sector (Solano & Huddleston, 2021). This has helped introduce more jobs for migrant workers and has established the country as a global leader in the technology industry.

Latvia, on the other hand, despite encouraging foreign investments through investor visa programs and recognizing the need for reforms in the migration policy, has consistently focused on engaging with its diaspora and return migration over immigration of non-nationals as a potential solution to the demographic and economic downturn (Birka, 2019). Lithuania, similarly, has focused on fostering return migration and diaspora engagement. However, foreseeing potential benefits of immigration, the country has also taken an active stance on migrant integration by developing a multitude of action plans and task forces, such as the Action Plan for Integration of Foreigners in Lithuanian Society 2018-20 (Birka, 2019). However, despite the shift in policies interventions, migratory flows to Lithuania are fairly small and originate from neighbouring countries.

Table 1 shows the values of the largest migrant flows by country of origin for each of the three Baltic States, starting 2007. As previously mentioned, a majority of the flows to the three countries originate from neighbouring or post-Soviet states. In the case of Estonia, Russians comprise the largest volume of migratory flows to the country (approximately 17.2 percent), followed by Ukraine and Finland. Although, the number of incoming migrants is quite low, with an average of only 1239 Russian migrants moving to Estonia each year. Lithuania reports relatively larger flows, of which, 70 percent can be traced to Ukrainian and Belarusian nationals. Russians also frequently migrate to the country, though they comprise a smaller proportion of the total flows at approximately 7.7 percent. Latvia reports the smallest overall immigration flows amongst the Baltic States. Around 40 percent of incoming migrants to the country are either Ukrainian or Russian nationals, with inflows of Indians presenting themselves as a surprisingly large proportion of the total flows at 11.9%. The gradual increase in Indian immigration to the country is attributable to amendments made to the immigration laws, which allow for employment for foreign nationals on long-term visas (OECD, 2020). The policy measure has proved to be more affordable and flexible than drawing up residence permits, thus generating an increased demand for skilled labour.

Table 1: Major Countries of Origin for Migrants to the Baltic States

Nationality	Average Number of Migrants per Year	Percentage of Total Flows
Estonia		
Russia	1239	17.2%
Ukraine	1170	16.2%
Finland	516	9.8%
Lithuania		
Ukraine	4334	41.8%
Belarus	2781	26.8%
Russia	803	7.7%
Latvia		
Ukraine	1099	21.4%
Russia	983	19.1%
India	613	11.9%

Source: Author's Elaboration based on Eurostat (2022e)

2.1.1.2 South-eastern Europe

We restrict South-eastern European countries included in our study to Romania and Bulgaria. Even though Greece and Turkey fall within the same geographic category, they present severe data related limitations which prevent us from including them in our analysis[2]. Like most post-Soviet states, Bulgaria is perceived as a point of transit and net emigration (Bobeva-Filipova, 2017; Krasteva, 2019). However, as of 2019, the annual immigration flows had increased by approximately 169 percent since 2012 depicting a steady increase in inflows (Eurostat, 2022e). Further, the size of foreign population has also increased by approximately 143 percent over the same period. Thus, displaying a relative decrease in transit migration as migrants tend to settle in the country than use it as an access point to migrate westward (Eurostat, 2022i; Mancheva & Troeva, 2011). This shift can be attributed to policy measures undertaken to improve migrant integration. For example, the Labor Migration and Labor

[2] Greece does not report disaggregated migration statistics, and Turkey does not supply any migration statistics to Eurostat as it's not an EU member state (Mooyart et al., 2021).

Mobility Act was amended in 2018 and allowed the family members of migrants to enjoy the same social rights as that of Bulgarian citizens (Solano & Huddleston, 2020). The amendment also helped improve access to education and language support for migrants' children (Solano & Huddleston, 2020).

Table 2 highlights that Russia, Turkey and Syria are the three major nationalities of immigrants moving to Bulgaria. While Russian inflows are driven by high mobility amongst post-Soviet states, immigration from Turkey and Syria can be attributed to political unrest and geographic proximity of the two countries of origin which also leads to high volumes of irregular migration. However, irregular Turkish and Syrian migration to Europe is centered around the Turkey-Greece border the Bulgarian border primarily acts as a secondary route for the same (Frontex, 2022).

Like Bulgaria, Romania is also primarily discussed as a subject of emigration and transit migration. However, while outflows of native populations are still extremely high, migration of foreign nationals to the country has been increasing. This is highlighted by an increase in immigration flows of non-nationals by 21 percent in the 2008 to 2019 period (Eurostat, 2022j). Additionally, migrant have shown a tendency to stay in the country as the migrant stocks have increased by 64 percent over the same timeframe (Eurostat, 2022i).

Migration to Romania is comparatively diverse with individuals moving to the country from Western Europe, Southeast Asia and even North America. However, Moldova, China and Turkey compose the most frequent countries of origin and constitute approximately 22 percent of the overall immigration flows to the country in the 2008 to 2019 period (see Table 2). The three are immediately followed by Italy, France and Germany which jointly comprise a further 15 percent of the total migration flows. While migration from Moldova and Turkey can be attributed to geographical proximity, flows from China are driven by the positive domestic perception of Chinese immigrants considering the large imports of Chinese goods to the country post the Soviet era (Radavoi, 2015).

Table 2: Major Countries of Origin for Migrants to Chosen South-eastern European Countries

Nationality	Average Number of Migrants per Year	Percentage of Total Flows
Bulgaria		
Russia	3074	23.3%
Turkey	2842	21.5%
Syria	1847	14.0%
Romania		
Moldova	1982	10.6%
China	1072	5.7%
Turkey	1001	5.4%

Source: Author's Elaboration based on Eurostat (2022e)

2.2.1 Central Europe

The Central European countries included in our study consist of the Czech Republic, Hungary, Slovakia, and Slovenia. We exclude the Bosnia-Herzegovina, Montenegro, Serbia, North Macedonia, and Albania from our discussion as even though, they are a part Central Europe, they are not EU members and as a result, have potentially different definitions of migrants and do not report statistics harmonized with other EU countries.

Further, we do not include Poland, Croatia and Germany[3] due to limited data availability (Mooyart et al., 2021) . Lastly, Austria is excluded from this discussion as the nature of migratory flows to the country are vastly different from other Central European countries selected. Moreover, in line with the EuroVoc's (2022) definitions of Europe's geographic divisions, Austria is classified as a part of Western Europe.

Migratory flows to the four Central European countries selected present greater diversity, both in volume and migrants' origins in comparison to the Eastern European countries. Major flows to Central European countries starting 2007 are reported in Table 3. Out of these, the Czech Republic reports the highest volume of incoming migrants per year. The relatively larger flows can be attributed to the country's comprehensive approach towards migration policies and considerably higher levels of migrant integration when compared to other CEECs (Solano and Huddleston, 2020). Ukrainians

[3] Germany only reports aggregated migration statistics to Eurostat as noted by Mooyaart et al. (2021)

comprise a majority of the total flows to the Czech Republic with approximately 13378 individuals moving to the country each year in the 2007 to 2019 period. Further, migrants from Vietnam and Slovakia jointly comprise an additional 24 percent of the migratory flows to the country.

Meanwhile, migration to Hungary can be characterized as short distance as it can be traced to countries with high cultural and linguistic proximity (Rédei, 2009). A majority of migrants moving to Hungary are citizens of Romania, Ukraine and Germany, which when combined, account for 40 percent of the total migratory flows to the country (Table 3). In the case of Slovenia, a majority of migrants hold a non-EU citizenship and move from the Yugoslavia's successor states. Bosnia & Herzegovina, Serbia and North Macedonia account for approximately 70 percent of all migratory flows to the country (Table 3). Out of the three, Bosnia & Herzegovina has the largest volume of flows at approximately 7591 incoming migrants per year.

The subject of migration suffers from heavy politicization in Slovakia (Filipec and Vargová, 2019).This polarization has resulted in the country having the smallest stock of migrants amongst the Central European countries, with only 76116 foreigner and stateless migrants residing in the country as of 2019 (Eurostat, 2022e). Transit migration from the country is not high either: as 2486 non-nationals immigrated to the country in 2019 and only 39 non-nationals emigrated in the same time period (Eurostat, 2022c; Eurostat, 2022f). As for composition, most of the migrants are from neighbouring EU countries, particularly, the Czech Republic, Hungary and Romania, which collectively account for approximately 47 percent of the immigration flows to Slovenia with an almost equal division of flows amongst them (Table 3).

Table 3: Major Countries of Origin for Migrants to Chosen Central European Countries

Nationality	Average Number of Migrants per Year	Percentage of Total Flows
Czech Republic		
Ukraine	13378	26.5%
Slovakia	8658	17.2%
Vietnam	3876	7.7%
Hungary		
Romania	4866	16.6%
Ukraine	4651	15.9%

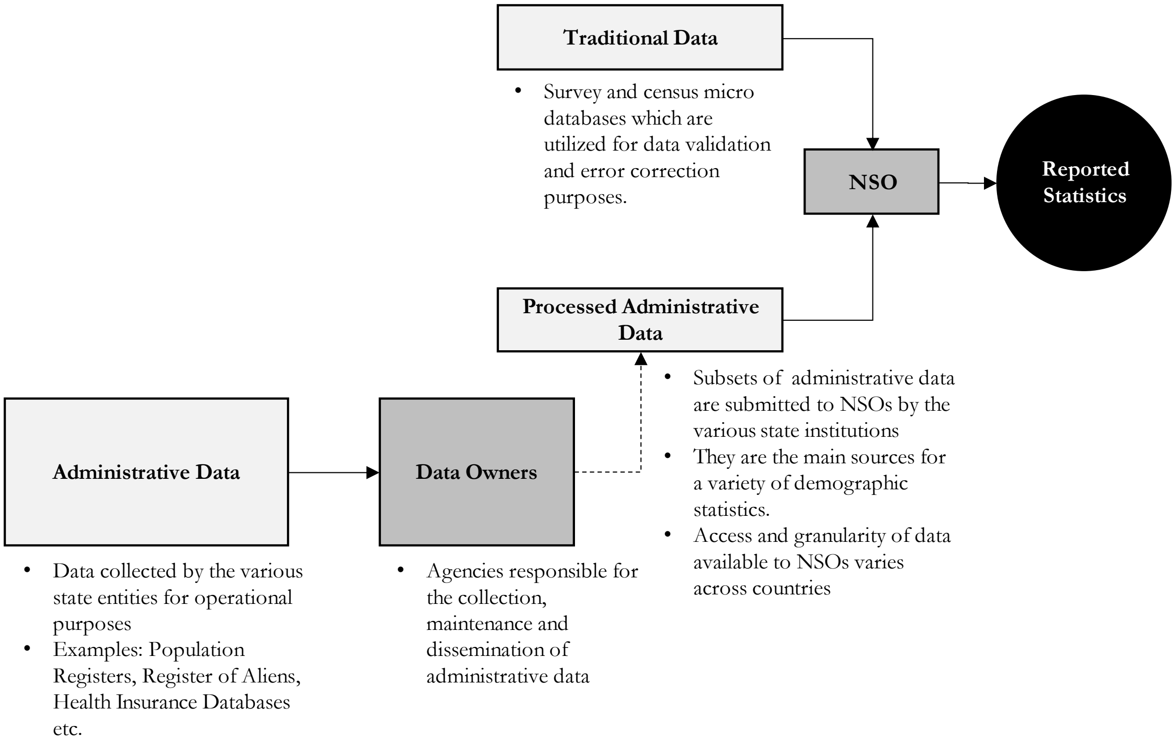
Germany	2259	7.7%
Slovenia		
Bosnia and Herzegovina	7591	46.5%
Serbia	2437	14.9%
North Macedonia	1522	9.3%
Slovakia		
Czech Republic	603	16.5%
Hungary	589	16.1%
Romania	514	14.0%

Source: Author's Elaboration based on Eurostat (2022e)

2.2 Production of Migration Statistics

This section focuses on the data production processes undertaken, and the upstream data utilized by the CEECs to generate official migration statistics, with Figure 1 presenting an outline of the same.

Figure 1: Production Process of Migration Statistics



- Survey and census micro databases which are utilized for data validation and error correction purposes.

- Subsets of administrative data are submitted to NSOs by the various state institutions
- They are the main sources for a variety of demographic statistics.
- Access and granularity of data available to NSOs varies across countries

- Data collected by the various state entities for operational purposes
- Examples: Population Registers, Register of Aliens, Health Insurance Databases etc.

- Agencies responsible for the collection, maintenance and dissemination of administrative data

Source: Author's Elaboration and Gárdos and Gödri (2014)

There is a general sense of agreement amongst EU member states that key aspects of the migration processes are costly to measure and cannot be estimated by solely relying on traditional data sources (Eurostat, 2018b). As a result, countries utilize a variety of administrative data sources to supplement and produce migration and other vital demographic statistics.

Administrative data refers to information collected by the various state entities for operational purposes. For example, local population registers, social security databases, tax databases etc (CROS, 2022). This data, however, is scattered across a multitude of institutions and cannot be used for statistical purposes without transformation (Hand, 2018). Certain countries utilize personal identification numbers (PIN) to integrate micro-data across multiple administrative domains (Gárdos and Gödri, 2014). While greater integration allows for consistency across data sources, privacy issues and low inter-institutional data sharing prevent national statistical offices (NSOs) from having unrestricted access to administrative data. Thus, NSOs are often supplied with truncated data files which are then utilized to generate official statistics (Eurostat, 2018a).

The lack of accessible raw data prevents statistical agencies and even other state entities from undertaking data validation and reconciliation practices. Thus, the quality of official statistics is highly dependent on the available administrative data and the institutions that collect them (Gárdos and Gödri, 2014). The quality of administrative sources also depends on the geographical size and population of the country. As Gárdos and Gödri (2014, p.14) note, "The use of administrative data is much more effective in small countries where keeping central registers is much easier than in countries with large territories and populations." Thus, smaller, and less dense countries tend to have higher quality administrative data purely because the cost and effort required for the maintenance of administrative data systems is lower.

Population registers, registers of foreigners, registers of asylum seekers along with the social security and health insurance databases act as the principal administrative data sources utilized by CEECs to produce migration statistics (Eurostat, 2022h). While social security and health insurance databases are used to compile information on livelihoods and socio-economic conditions, the various registers provide insights into population composition and the count of incoming/outgoing foreign nationals (Gárdos and Gödri, 2014).

The usability and quality of population registers varies across CEECs as certain individuals/populations may or may not be included in the registers. In the case of Romania, for example, information on non-nationals is not recorded in the population registers and, as a result, the NSO does not identify the population register as a key data source for reported migration statistics (Gárdos and Gödri, 2014). On the other hand, population registers managed by government agencies in the Baltic States are frequently used as one of the principal sources of administrative data for migration statistics, as they contain information on foreign populations.

Alternatively, various registers of foreigners are utilized by the CEECs to produce migration statistics. These registers may be stand-alone data sources or may act as feeders into the primary population registers (Gárdos and Gödri, 2014). In addition to records of residence permits, these registers also contain personal data such as educational attainment, occupation, marital status etc. Similar to the other administrative sources, variations in linkage and access to upstream data can limit the efficacy of published statistics derived from these sources. This is the case in Romania, as the NSO does not have direct access to the register of foreigners and the concerned ministry produces their own independent estimates of migrant populations[4] (Gárdos and Gödri, 2014). The NSO as a result, acts as an intermediary between Eurostat and the data production agencies. A relatively less severe case is that of Slovakia, where aggregated data is provided to the NSO (Gárdos and Gödri, 2014). On the opposite end of the spectrum is Bulgaria, where the NSO and various data owners have a fluid approach to data sharing. This facilitates greater linkages across different sources and in turn improves the quality of statistics produced.

While administrative data sources allow for greater coverage and do not suffer from sampling error, they often have issues of completeness and are prone to undercounting migrants, the reasons for which we discuss in the coming sections (Gárdos & Gödri, 2014). For this reason, certain countries continue to utilize more traditional sources such as census and survey-based databases to reduce errors in official statistics (Gárdos & Gödri, 2014). Census data is also useful for comparisons across migrants and native populations as the data is collected through a uniform process and can be utilized for data validation purposes (Gárdos & Gödri, 2014). A key limitation of the same is the lack of comparability across time, as census databases are only collected every ten years and are not subject to consistent changes and revisions to maintain their quality. Thus, these databases cannot quantify dynamic changes in the population structure of a country. While countries across the EU have identified a uniform set of definitions and methods, comparability of census data on migrants across countries is

[4] The General Inspectorate for Immigration (IGI) is a specialized public institution coordinated by the Ministry of Internal Affairs and is the responsible agency for the development and dissemination of migration statistics in Romania.

still quite low due to differing levels of integration and subsequent under/over coverage of different population subsets. For example, the 2011 Hungarian census data understated foreign population stocks by approximately 60000 due to issues of coverage and low migrant integration, which might lead us to misappropriate Hungary as a relatively less attractive country for migrants (Gárdos & Gödri, 2014).

Lastly, we discuss survey-based data sources used for producing migration statistics. The EU Labour Force Survey (EU-LFS) is the most comprehensive and standardized survey-based instrument which captures the resident population and can be utilized for the construction of migration statistics (Gárdos and Gödri, 2014). It includes a large sampling frame (the entire labour force within the age range of 15 to 74) and has a high frequency of data collection (quarterly), in addition to common definitions and concepts across all EU countries (Martí and Ródenas, 2007). However, questions related to migration in the survey are somewhat limited. Further, the data collected cannot be used for the estimation of the actual volume of migratory flows as the survey does not account for non-working populations (Martí and Ródenas, 2007).

Alternatively, countries conduct targeted surveys to collect data on migration. An example of the same is Bulgaria's implementation of mobility related surveys in parallel to the 2001 and 2011 census[5]. The country also undertook specialized monthly surveys known as Sample Survey on Bulgarian and Foreign Citizens Departing from Bulgaria at border checkpoints over the same timeframe (Gárdos and Gödri, 2014). Slovakia, on the other hand, has adopted a more generalized approach to high frequency surveys by administering monthly population surveys for the estimation of migrant population (International Monetary Fund, 2022). Overall, while accurate sources of migration statistics, surveys have categorical issues in the form of under coverage and non-response which reduces their validity. Hence, the survey data is largely utilized to supplement administrative data sources (except for the case of Slovakia due to an extremely high survey frequency) (Martí and Ródenas, 2007).

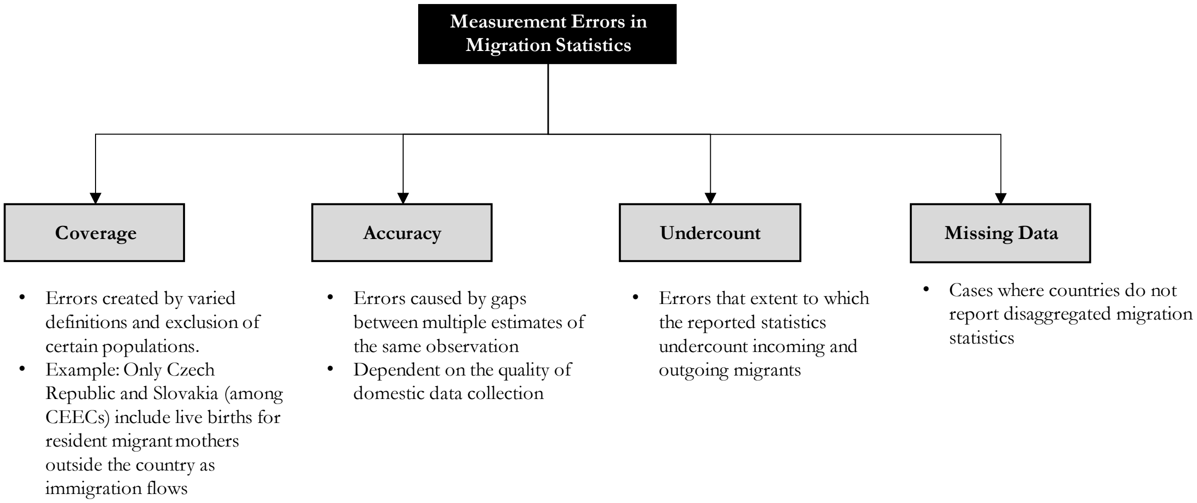
[5] The Territorial Mobility of the Population Survey was implemented along with the 2001 census and the Migration and Migration Behaviour of the population survey was rolled out along with the 2011 census

Chapter 3: Conceptual Framework

3.1 Types of Measurement Errors in Migration Statistics

This section intends to conceptualize the various types of measurement errors that persist in migration flow statistics. Figure 2 presents a brief outline of the various dimensions of the same. The figure is partially derived from Raymer et al. (2013) who identify undercounting, accuracy, and coverage as the three components of measurement errors which they then utilize to generate debiased migration flows as a part of the Integrated Modelling of European Migration (IMEM) project. In addition to the three components, we also conduct a brief analysis of missing data.

Figure 2: Types of Measurement Errors



Coverage
- Errors created by varied definitions and exclusion of certain populations.
- Example: Only Czech Republic and Slovakia (among CEECs) include live births for resident migrant mothers outside the country as immigration flows

Accuracy
- Errors caused by gaps between multiple estimates of the same observation
- Dependent on the quality of domestic data collection

Undercount
- Errors that extent to which the reported statistics undercount incoming and outgoing migrants

Missing Data
- Cases where countries do not report disaggregated migration statistics

Source: Author's Elaboration and Raymer et al. (2013)

Errors of coverage are generated by varied definitions and exclusions of certain subpopulations from the reported statistics (Aristotelous et al., 2022; Raymer et al., 2013). Examples of such variations are the exclusion of asylum seekers and refugees, live births outside the reporting country, varied duration of stay definitions etc. (Aristotelous et al., 2022). Coverage errors, however, have considerably reduced since the introduction of Council Regulation (EEC) No. 311/76 in 2007, which mandated the adoption of a uniform taxonomy and reporting standards for community statistics on migration and international protection across the EU (European Parliament and Council of the European Union, 2007).

Accuracy, refers to the gaps between multiple estimates of the same observation. In the context of migration statistics, this relates to the quality of data collection and production systems utilized by the reporting countries (Raymer et al., 2013). In the cases where mirror statistics[6] are used, accuracy relies on the systems used by the country of previous residence i.e., the country of previous residence. Accuracy also depends on the type of data sources utilized. For example, Keilman and Aristotelous (2021) argue that usage of sample surveys for demographic purposes (especially for the estimation of migrant population) generates statistics of lower accuracy when compared to register-based statistics due to survey errors. As previously discussed, countries such as Slovakia cope with these issues by combining high-frequency surveys that cover all migration events and residence changes with register data to produce migration statistics.

Undercounting, on the other hand, refers to the degree by which the reported flows are understated when compared to a latent true flow value (Keilman & Aristotelous, 2021). This type of measurement error may be caused by a variety of reasons ranging from the state capacity, quality of data collection systems, and under-registration by migrants. A solution often proposed to reduce the error is to utilize mirror statistics for immigration flows. However, mirror statistics pose problems of their own as emigration data suffer from much higher levels of undercounting considering that individuals often have little incentive to declare their move away from the country of previous residence (Keilman & Aristotelous, 2021).

While undercounting and accuracy act as the key components of measurement errors in demographic statistics, we posit that missing data is also a form of measurement error. Blackwell et al. (2017) state that measurement errors in available data can be treated as partially missing information, and completely missing data needs to be treated as an extreme form of measurement error. Thus, we identify missing disaggregated flow data as a form of measurement error. Although, a few caveats must be placed on this characterization. To begin with, missing data originating from a lack of legal requirements cannot be qualified as a case of measurement error. This was the case for missing migration statistics on Eurostat prior to 2007, as the data submission regulations (2007 Council Regulation (EEC) No 311/76) that obligate EU member states to regularly submit migration and associated statistics to Eurostat were not yet put in place. Further, some non-EU countries such as North Macedonia, Turkey, and Serbia voluntarily submit their data to Eurostat, however, missing data from these countries cannot be qualified as a type of measurement error as they are not subject to the

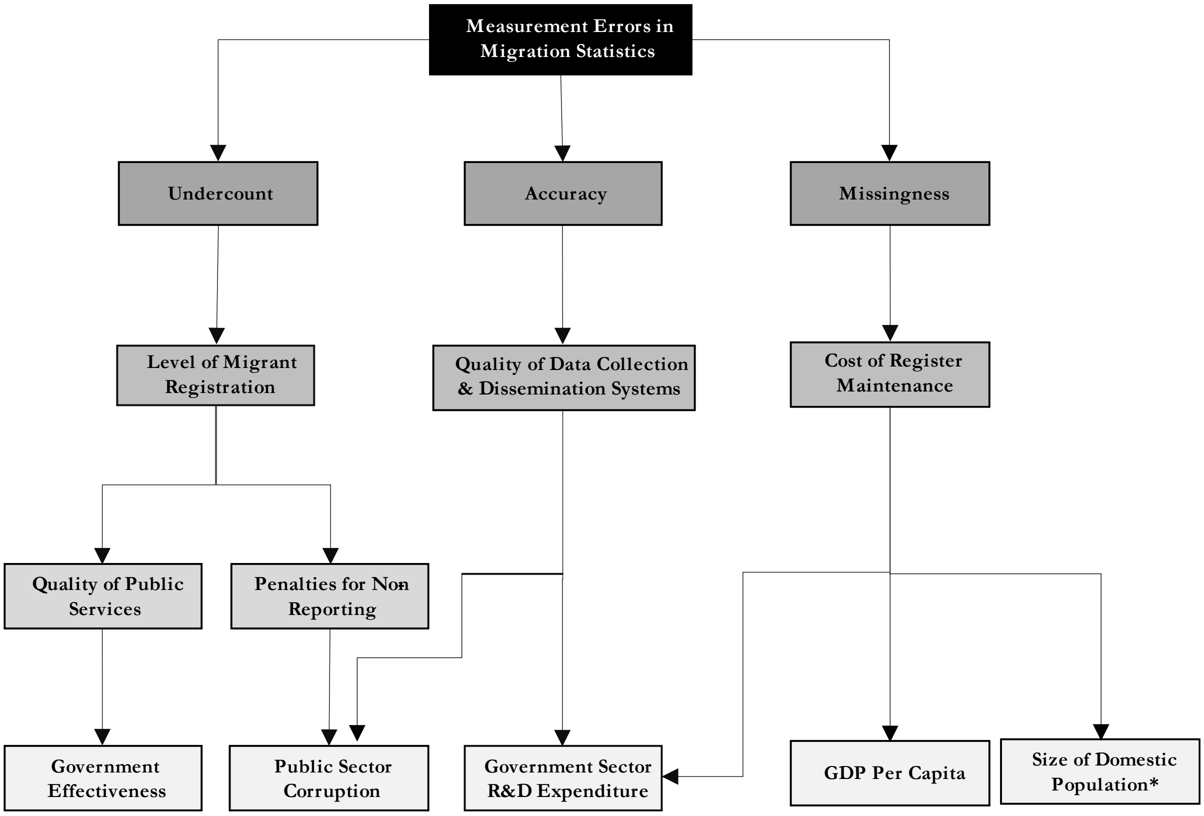
[6] Mirror immigration statistics refer to the counts of outgoing individuals as reported by the country of previous residence (UNECE, 2021). For example, Italian migration to Romania can also be measured as the number of individuals leaving Italy and moving to Romania as per Italian population statistics (Eurostat, 2022h)

EU data submission regulations. Hence, missing data can only be considered as an extreme form of measurement error when data submission regulations are in place and the countries with missing data fall within the purview of those regulations in the given time period.

3.2 What Causes Measurement Errors

Based on the classifications presented in the previous section, this section discusses the underlying mechanisms by which measurement errors might arise. These factors and their associated mechanisms are presented in the figure below:

Figure 3: Factors Affecting Measurement Errors in Migration Statistics


Measurement Errors in Migration Statistics
Undercount
Accuracy
Missingness
Level of Migrant Registration
Quality of Data Collection & Dissemination Systems
Cost of Register Maintenance
Quality of Public Services
Penalties for Non Reporting
Government Effectiveness
Public Sector Corruption
Government Sector R&D Expenditure
GDP Per Capita
Size of Domestic Population*

*Domestic Population includes both migrants and non-migrants

Source: Author's Elaboration

The first row in the Figure 3 outlines the types of measurement errors. The second row explains the factors that affect the measurement errors and lastly, rows three and four highlight the key drivers of these factors (Row 4 consists of the explanatory variables which are explained later on). Undercounting in the context of our study occurs largely due to low levels of self-reporting of migration (Willekens, 2019).

Since migrants must self-report their move, deterrents such as monetary penalties and even imprisonment, are often put in place by the relevant state authorities for individuals who fail to disclose their move (European Union Agency for Fundamental Rights, 2016). Such deterrents, as a result, aim to reduce irregular entry and stay in addition to facilitation of the same. Despite the prevalence of deterrents to discourage non-disclosure in the EU, their implementation is patchy, and information regarding the same among resident migrant populations is quite limited (Aristotelous, Smith and Bijak, 2022). Thus, we expect access to public services such as healthcare, education, and social care to act as larger incentives for migrant registration. As a result, countries which require civil registration (and subsequent inclusion in the population registers) for accessing public services are expected to have relatively lower levels of undercounting. This mechanism also relies on the quality of public services administered and the extent to which migrants can access them. Hence, we assert that the quality of public service provision would foster registrations amongst incoming migrants and consequently reduce the level of undercounting. However, the impact of registration requirements and quality of public services is not expected to be uniform across migrant sub-groups. For example, migrants from Schengen member states may have less of an incentive to register due to their high mobility and access to public services in their countries of origin.

With regards to accuracy, as previously mentioned, the quality of data collection systems and interlinkage of data across institutions play a key role in ensuring consistency and validation across multiple data sources. Thus, the prevalence of public sector corruption and the poor quality of state institutions can have a detrimental effect on the official statistics produced. However, the relationship is not expected to be linear. We argue that initial increases in public sector corruption result in lower levels of measurement error. This can be seen as a 'greasing the wheels' phenomenon as bureaucratic and administrative corruption in the public sector, to a certain extent, can help enhance efficiency by skirting data production and dissemination protocols (Aidt, 2003). However, beyond a specific threshold, the extent of corruption starts deteriorating the overall quality of the data infrastructure. This reduces the quality of upstream data utilized for statistical purposes and results in increasing levels of inaccuracy. These arguments of non-linearities also hold for undercounting as initial increases in corruption might make the process of civil registration more accessible for migrants. However, beyond a specific threshold, the costs become too high, and incoming migrants respond by choosing to not register themselves. Our arguments are embodied by Méon & Weill (2010), who argue that initial increases in corruption helps in the reduction of red-tapism, although, overall, the effects of corruption are expected to still be detrimental.

R&D also plays a key role in reducing the limitations of administrative data sources as it fosters innovation for improvement of available data and helps design new coping strategies for measurement errors (Anderson and Whitford, 2017; Bosco et al. 2022; Santamaria and Vespe, 2018). These may consist of projects undertaken to harmonize internal data sources, introduction of technological improvements in data infrastructure and the development of new estimation techniques to enhance available statistics. While there have been a multitude of country level endeavours focusing on the same, the IMEM, QuantMig, SEEMIG and Towards Harmonised European Statistics on International Migration (THESIM) projects are some of the key EU-funded projects focused on identifying gaps in current demographic data production systems and developing estimation techniques to minimize errors in migration statistics. Alternatively, R&D expenditure can be foster the usage of non-traditional sources such as big data, as their high velocity and volume can help increase the timeliness and accuracy of the official statistics produced (Braaksma, Zeelenberg and de Broe, 2020). Hence, we expect that higher levels of investment in R&D would lead to a decline in all forms of measurement errors through a reduction in production costs, greater data availability and the introduction of new estimation methods.

Lastly, we discuss the causes of missing data. These can be challenging to pinpoint and quantify. Gárdos and Gödri (2014) suggest that extremely low rates of migrant registration may contribute to missing data, but data-related limitations prevent us from assessing this relationship. The income levels of a country are indicative of its data infrastructure, as increases in GDP can help foster investment in the country's administrative data infrastructure in addition to the provision of more funding for statistical institutions (Kim, 2022). Thus, increasing income levels is expected to decrease the probability of missing data (assuming that there is demand). Further, we can assert that high costs of register maintenance and a lack of funding to address data-related issues may also lead to increased missing data. While we cannot determine the precise cost of administrative data maintenance, previous discussions based on the findings of Gárdos and Gödri (2014) can be used to argue that increases in domestic and foreign populations can put pressure on register maintenance costs, which may contribute to the lack of available disaggregated data. Therefore, we propose that domestic population growth, whether through immigration or natural increase, can be used as a proxy for register maintenance costs and a potential factor affecting missing data.

Chapter 4: Literature Review

Literature about errors in migration statistics is often restricted to demographic studies that seek to overcome the limitations of reported migration data by accounting for measurement errors, availability issues and definitional inconsistencies (Wiśniowski, Zagheni & Fava, 2019). These studies make *a priori* assumptions about the types of measurement errors and utilize methods such as Delphi surveys[7] to capture the extent of measurement errors across countries (see Bijak & Wiśniowski, 2010; Keilman & Aristotelous, 2021; Rampazzo et al., 2021; Raymer et al., 2013; Wiśniowski et al., 2019). These studies then incorporate the covariates of measurement errors obtained, in addition to traditional drivers of migration, into a hierarchal model to generate estimates of bilateral migration flows.

Alternatively, empirical literature on the subject of measurement errors focuses on conducting a comparative analysis of statistics produced using multiple data sources and base themselves on the assumption that a specific data source is more accurate than the other. Gomez & Glaser (2006), for example, seek to examine potential misclassification errors in administrative data. As a part of their study, they find that minorities in the US Cancer Registry data were more susceptible to misclassification with Native Americans and Hispanics being racially/ethnically misclassified in 83% and 30% of the cases respectively. The issue of misclassification also persists within the context of migration as Saarela & Weber (2017) argue that research associated with recently arrived immigrants is likely to be affected by measurement errors caused by misclassification and limited information. The authors conclude the same based on a comparison of the level of educational attainment of Finnish immigrants in Sweden as reported by Sweden upon arrival and Finland prior to leaving. The authors further argue that the tendency to misreport in the destination country is likely to decrease over time i.e., Misreporting in the Swedish statistics on the level of educational attainment of Finnish migrants is relatively higher for migrant flows than migrant population stocks. As a result, issues of misclassification can generate measurement errors for immigration flows as it directly affects the composition and count of incoming migrants at a given point in time.

Lanzieri (2018) conducts a comparative analysis between Eurostat's migrant stocks and residence permit statistics. Based on this, they conclude that the gap between the two available estimates cannot be clearly decomposed and as a result, there exists a measurement error in the reported migration

[7] Delphi surveys are utilized to gather expert opinions on a specific subject through sequential questionnaires

statistics. Martí & Ródenas (2007) on the other hand assess the viability of using the EU-LFS data as a proxy for migrant statistics by drawing a comparison with population register data. Based on a comparative analysis, they argue that LFS data suffers from a variety of survey errors such as non-response and inconsistent updating of the sampling frame, which can introduce bias and by association measurement errors. The authors also argue that variation of these errors across countries and time prevents the LFS data from being an accurate proxy for capturing migration processes.

Despite a large body of literature recognizing the prevalence of measurement errors in migration statistics, unfortunately there is little empirical research on the underlying mechanisms of the measurement errors themselves. This makes it difficult to theoretically ground potential determinants of measurement errors in migration statistics and distinguish them from circumstantial factors that might alter the overall dynamics of migration flows. As a result, we also provide a brief discussion on studies exploring the factors affecting overall statistical capacity of countries.

Kim (2022) argues that a variety of technological, financial, and political factors affect the statistical capacity of countries. The author based on their study covering 135 countries, argues that increased levels of democratization and financial allocations lead to higher statistical capacity. The levels of technological advancements, however, are shown to have little effect on statistical capacity. Lokshin (2021) argues that efforts of increasing statistical capacity are associated with the diversity of data sources utilized and the overall sophistication of the data production processes. A key contributor to the same, however, is the alignment with local actors as they act as the primary consumers for the statistical outputs. Taylor (2016), on the other hand, argues that increased capacity of NSOs spills over to administrative data production processes leading to a higher level of overall statistical capacity in the country. Additionally, Taylor (2016) argues that institutional independence along with de facto and de jure protections which ensure NSOs' independence from political pressures also enable the improvement of statistical capacity.

Overall, the literature presented is highly disjointed and does not seek to answer the questions that form the crux of this paper. However, they provide some insights into measurement errors in migration statistics and the determinants of statistical capacity which have an indirect impact on the quality of national statistics produced. Thus, empirically linking the various types of measurement errors specific to migration data to factors that affect them marks the start of a completely new body of work.

Chapter 5: Methods & Data

This paper takes a two-step methodological approach, which initially involves discussing the construction of the dependent variables and is followed by a discussion on the empirical strategy utilized. This chapter focuses on the former while the next chapter delves into the latter.

5.1 Period of Study

Before delving into the methods utilized, we must contextualize the period of analysis. This paper utilizes estimated and reported migration flows for CEECs from 2007 to 2019. 2007 is chosen as the starting point for this study as prior to the specific year, there weren't any EU regulations present for a consistent time oriented duration of stay definition for migrants (Aristotelous, et al., 2022). Thus, prior to 2007 countries opted for varied time definitions for qualification of non-nationals as migrants. In this regard, a few countries opted for 3-6 months of stay, while others opted for the UN 12-month definition or even permanent residence. The regulations introduced in 2007 mandated the utilization of the 12-month UN definition of migrants for the foreign population statistics supplied to Eurostat (European Parliament and the Council of the European Union, 2007). The regulations also called for improved data quality checks and added granularity for demographic data collection (European Parliament and the Council of the European Union, 2007). Hence, utilization of data post 2007 allows to ensure that migrant populations are estimated with a harmonized time definition. Further, the study is restricted to 2019 to reduce the potential impact of the COVID–19 pandemic on migration flow estimates.

5.1 Migration Flow Estimates

5.1.1 A Review of Flow Estimation Techniques

A variety of approaches have been developed to estimate migration flows. These techniques range from qualifying the variations in stock values over time as migration flows to Raymar et al.'s (2013) approach of utilizing a hierarchal model that integrates reported data, covariate information and expert opinions to generate consistent and reliable migration flow estimates.

Despite the variety in estimation methods available, a majority of them primarily rely on stock data. Utilization of population stock tables for the estimation of immigration flows has a few key advantages. First, Rees (as cited in Abel 2013, p.506) argues that the implementation of an accounting framework which includes both stocks and flows can help with data validation and match available data with conceptual models. Second, stock data is static, i.e., it is measured at a specific point in time. This makes stocks easier to measure, when compared to flow data which is dynamic and is measured over a period of time (Abel, 2013). This relative ease of measurement makes migrant stocks less prone to measurement errors. Thus, migrant stocks can be used to produce more reliable and consistent flow statistics in comparison to the reported flow values (Abel, 2010, 2013; Abel & Cohen, 2019; Azose & Raftery, 2019; Saarela & Weber, 2017).

Having outlined the advantages of utilizing stocks for the generation of migration flow estimates, we now delve into the discussion of the potential estimation methods. The flow estimation techniques can be divided into three distinct categories:

5.1.1.1 Stock differencing

As the name suggests, the method utilizes differenced values of migrant population stocks to estimate immigration flows (Abel and Cohen, 2019). However, there are variations in the treatment of negative flow values generated by the reduction in stocks. On the one hand, Beine et al. (2011) treat negative stock differences as zero inflows by setting all negative values to zero. On the other, Beine and Parsons (2015) treat negative flow values as reverse migration flows i.e., emigration back to the country of origin.

5.1.1.2 Migration Rates

This estimation technique, initially proposed by Dennett (2015) uses migrant stocks to generate flow rates or the probability of migration. This flow rate is then multiplied by the volume of global migration flows to obtain country-pair wise migration flows for each year. Migration rates are relatively more demanding than the stock differencing approach as they require country-pair wise migrant stock populations in addition to the global stocks and flows of foreign population for each year.

5.1.1.3 Demographic Accounting Frameworks

Certainly, this is the most demanding approach from a data requirement perspective. All demographic accounting frameworks of migration rely on sequential tables of bilateral migrant stocks but vary in their estimation methods (Abel and Cohen, 2019). The method requires the creation of a square matrix of bilateral migrant stocks for each time-period (Abel and Cohen, 2019). All diagonal elements in a matrix represent native populations. Migration flows are then estimated by assessing changes in the global migrant stocks between two points in time (comparing across matrices) within the accounting system (Berlemann, Haustein and Steinhardt, 2021).

These frameworks can be further divided into open and closed. Within open demographic accounting frameworks, flows outside of the specified system are assumed to not exist, while closed demographic accounting frameworks seek to rescale the input data to generate flows in and out of the specified system (Abel, 2010; Abel & Sander, 2014).

A key assumption of Abel's (2010) demographic accounting frameworks is that they maximize the number of individuals that stay within their country of current residence at any given time period. Azose & Raftery (2019) propose a pseudo-Bayesian demographic accounting framework that relaxes this stayer maximization assumption by taking the weighted average of the open and closed frameworks to produce migration flow estimates.

While the demographic accounting frameworks provide stock-derived migration flow estimates of the highest quality, they require a complete matrix of global migrant stocks for each time period in question (Abel, 2010). Abel (2010) deals with the data availability problem by utilizing the UN migrant stock data which is recorded over five-year intervals starting in 1990. However, the lack of temporal granularity limits the usability of the data for our research as certain events that drive migration might happen in the middle of the five-year intervals and we cannot capture the effects of the same through the data. For example, increases in inflows of migrants caused by the Syrian civil war in 2011 might not captured by flow estimates derived from UN's 5-year data. Similar issues persist when Beine and Parsons' (2015) approach of reversing migration flows is considered.

5.1.2 Flow Estimation Techniques Employed

We settle for a combination of stock differencing and migration rates approaches for flow estimation purposes. Starting with migration rates, we first must generate bilateral migration rates or the stock probabilities relative to the total migrant stock population (Dennett, 2015), which is given by:

$$PS_{odt} = \frac{Stock_{odt}}{Stock_t} \qquad (1)$$

Where PS_{odt} is a ratio of the stock of migrants from country *o* living in country *d* at time *t* to the total stock of foreign population in the system at the given time[8]. Hence, PS_{odt} represents the

[8] Total migrant stocks are given by $Stock_t = \sum_o \sum_d Stock_{odt}$

probability that a migrant belonging to country of origin *o,* resides in country of destination *d* at time *t*.

Similarly, Dennett (2015) develops the probability of immigration flow which is given by PF_{odt} as highlighted by Equation 2. Where $Flow_{odt}$ is the number of migrants moving from country of origin *o* to country of destination *d* at time *t* and $Flow_t$ represents the total inflows of foreign population at time *t* [9].

$$PF_{odt} = \frac{Flow_{odt}}{Flow_t} \tag{2}$$

Post which, Dennett (2015) conducts a comparison of the PS_{odt} and PF_{odt} using World Bank's bilateral migrant stock database and the IMEM project's flow estimates and observes that the two are roughly equal. We assume the same based on Dennett's (2015) findings as highlighted by equation 3.

$$PF_{odt} \approx PS_{odt} \tag{3}$$

Given equation 3, substituting PF_{odt} in equation 2 and solving for $Flow_{odt}$ allows to generate a set of migration flow estimates which are given by equation 4. Where $MR_Flow_{odt}$is the migration flow from country *o* to country *d* at time *t*.

$$MR\_Flow_{odt} = Flow_t \times PS_{odt} \tag{4}$$

Dennett (2015), like Abel (2010), utilizes UN migrant stock and flow data to produce flow estimates. A key issue with the same is that of varied time dimensions across the stock and flow databases. UN's aggregated flow data is reported annually, however, the bilateral stock data is only reported every five years (Dennett, 2015). As a result, the author assumes that the bilateral migration rates i.e., PS_{odt}, for each country pair remain constant over a five-year period. However, it was not necessary for us to make such assumptions or harmonize time periods as we are able to source annual stock and flow data from Eurostat (Discussed in Chapter 6).

Estimation using migration rates has a few limitations. First, the method relies on the assumption that the determinants of migration processes have a roughly uniform impact on PF_{odt} and PS_{odt} . In

[9] Total migrant flows are given by $Flow_t = \sum_o \sum_d Flow_{odt}$

cases of non-uniform effects on the two probabilities, the assumption of approximate equivalency between the two breaks down, thus generating a bias within the estimated flows.

Another key limitation of the approach is its inability to account for bursts in migration flows. This limitation is highlighted by Dennett (2015) who takes the example of Italian immigration to Australia which rose as a direct consequence of the second World War but experienced a sudden drop post 1970 as per UN data. The sudden decrease in immigration flows is not recognized in the migration rate estimates which continue to present inflated estimates of incoming Italian migrants.

Bertoli and Fernández-Huertas Moraga (2015) difference the migrant stocks and drop negative values (set all negative values to zero) to estimate migration flows. as a part of their study on the effects of migration policies on bilateral migration flows. Flows derived from this approach are called SD_Flow_{odt} and are given by the following equation:

$$SD\_Flow_{odt} = \begin{cases} Stock_{odt} - Stock_{od(t-1)} \ \ if \ Stock_{odt} \geq Stock_{od(t-1)} \\ 0 \ if \ Stock_{odt} < Stock_{od(t-1)} \end{cases} \tag{5}$$

Where $Stock_{odt}$ and $Stock_{od(t-1)}$ are the stocks of migrants from country *o* living in country *d* at time *t* and *t-1* respectively. Bertoli & Fernández-Huertas Moraga (2015) argue that negative flow values can be attributed to shifts in patterns of emigration and natural population change, and as a result, they are replaced with zeros. This assumption can be quite reductive as reduction in stocks does not necessarily mean zero inflows. For example, countries with a negative net migration might suffer from reducing migrant stocks irrespective of the volume of inflows.

Overall, both estimation methods are reliant on the validity and availability of stock data, which in turn makes them susceptible to non-inflow related factors that affect population stocks. For example, bias in the estimated flows can be driven by shifts in factors such as the natural rate of change, emigration, and acquisition of citizenship. Going forward, we term this potential bias as method dependency.

5.1.3 Comparing Estimated and Reported Migration Flows

Having outlined the techniques selected, we now draw a comparison between the estimated and reported migration flows. The estimated flows are generated using Eurostat data. We utilize migration statistics (for both stocks and flows) disaggregated by citizenship (country of origin) for the purpose of our analysis.

However, data on migration flows to the CEECs suffers from issues of completeness (Mooyart et al., 2021). We take a two-fold approach to cope with the same. First, in cases where stocks or flow data disaggregated by citizenship is unavailable, we impute stock and flow values disaggregated by country of birth[10]. Second, we drop observations where both citizenship and country of birth wise flow data is missing as they cannot be used to generate the measurement error indicators.

We start out by assessing the differences in the number of migration events reported[11], which are presented in Table 5. We observe that reported statistics for the CEECs contain 9335 migration events between 2007 and 2019, implying that there are over nine thousand country pair and year combinations in our sample with non-zero reported flow values. Stock difference estimates report the lowest number of migration events across the board. However, this can be attributed to zero inflation caused by the reduction in stock values due to high volumes of emigration in CEECs. On the other hand, migration rate estimates contain 11427 migration events between 2007 and 2019, which is roughly 22.4 percent higher than the reported statistics. The gap between the migration rate estimates and reported values persists across all nine CEECs, with Slovakia and Slovenia displaying the largest gap. The gap also persists over time and reaches its peak in 2015, with the migration rate estimates containing 255 more migration events than the reported flows (see Appendix Table 6).

Table 4: Comparison of Non-Zero Flows (Migration Events) Across Estimates

Country of Residence	Reported Flows	Stock Difference	Migration Rates
Bulgaria	923	704	1,170
Czech Republic	1,831	1,464	2,075
Estonia	626	573	768
Hungary	1,849	1,291	2,060
Latvia	375	262	430
Lithuania	496	357	548
Romania	865	522	931
Slovak Republic	1,018	1,006	1,730
Slovenia	1,352	971	1,715
Total	9,335	7,150	11,427

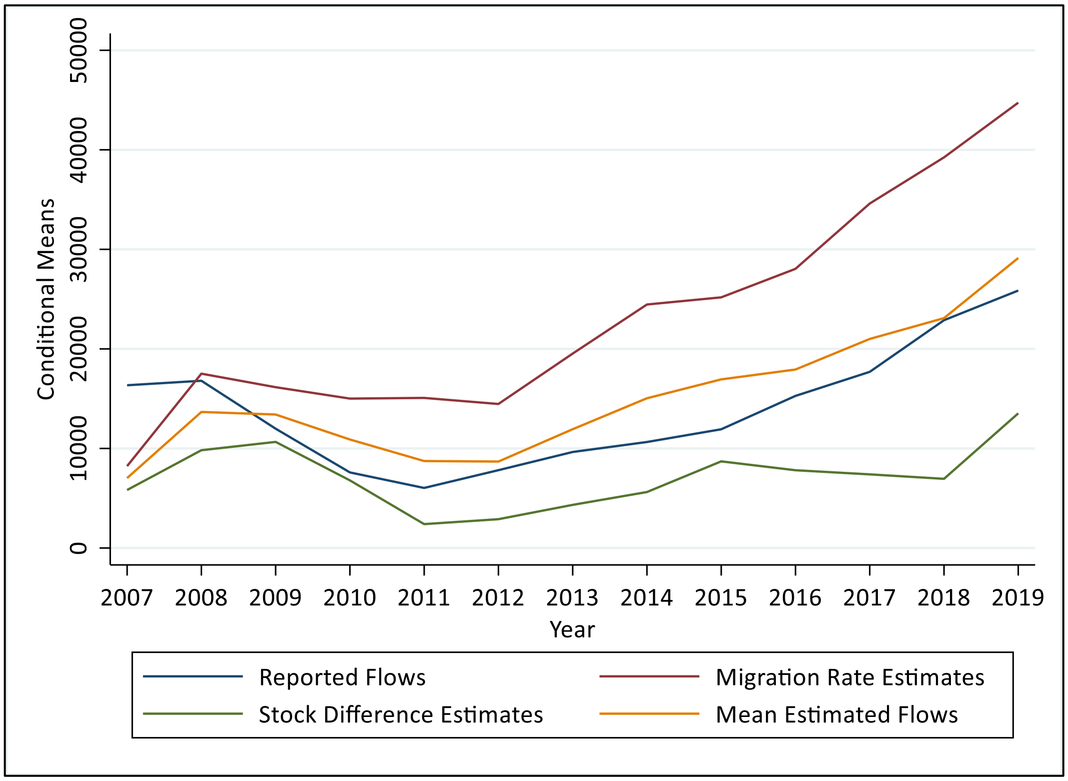
***Note:** The number of non-zero values for the mean of the estimated flows are not reported as they are the same as the migration rate estimates.

[10] Imputed values largely consist of small stocks and flows of migrants. 88 imputations for migrant stocks and 276 for migrant flows

[11] Migration events in this context refer to the movement of *n* number of individuals between a country pair in a given year where *n* is any non-zero value.

In addition to differences in the number of migration events, we observe variations in the total volume of immigration flows across the estimated and reported flows. Figure 4 presents a time evolution of total reported and estimated migration flows. Migration rate estimates consistently have the highest volumes of total immigration flows per year, whereas the stock difference estimates generate the lowest values on average, with the reported flows lying in between the two estimated flows. The average of the estimated flows (utilized for generating the undercount indicator) is much closer to the reported flows but is consistently higher, except for 2007 and 2008. It is, however, important to note that Figure 4 only presents an annual aggregate measure of immigration flows, and there might be cases where the reported flows are smaller than the stock difference estimates, or the reported flows might even be larger than any of the estimated flows. For example, Eurostat (2022e) reports that only 2 Austrians moved to Romania in 2013, but stock difference and migration rate estimates indicate 182 and 214 incoming migrants, respectively.

Figure 4: Temporal Evolution of Total Reported & Estimated Immigration Flows



Source: Author's Elaboration based on Eurostat (2022e) data

Note: The country pair values are summed over country year and then plotted

5.2 Measurement Error Indicators

5.2.1 Indicator Design

Having outlined techniques utilized to generate alternative flow values, we must identify indicators to quantify the measurement errors in line with the conceptual framework. Starting the discussion with errors of undercounting, we identify that the true value of migration flows, in this case given by the average of the two estimated flows (Discussion on why the average is assumed to be the true value is provided in the Appendix A.1), is a function of the reported flows and an added measurement error. This is highlighted by the equation below:

$$\frac{(MR\_Flow_{odt} + SD\_Flow_{odt})}{2} = (1 + ME_{odt})Reported\_Flow_{odt} \tag{6}$$

Thus, ME_{odt} captures the extent to which the reported migration flows are smaller than the estimated flow values for migrants from country of origin *o* currently living in country of destination *d* at time *t*. Given the nature of our data, we must undertake a few considerations. It is common for migration data like all other forms of count data to suffer from zero inflation (Tu and Liu, 2016). Zero inflation refers to the presence of excessive zero values and can generate computational difficulties. Researchers often deal with zero information by imputing information. Our undercount indicator carries over some of these issues. To begin with, ME_{odt} takes the value 1 by default if the reported flows are zero and the estimated flows are non-zero. The variable takes the value -1 for all cases where the reported flows are non-zero and the estimated flows are both zero. Further, the indicator takes the value zero when both, the reported and synthetic flows are zero, which can generate zero inflation. The presence of zero inflation in the flow estimates and the reported statistics, as a result, can generate frequency spike as -1, 0 and 1.

Measuring accuracy, on the other hand, requires comparison across multiple estimates and checking for variations amongst them. As a result, we do not utilize the mean values of the estimated flows to measure inaccuracy but instead compare across the two estimated flows and reported statistics. We undertake multiple considerations to choose an indicator that captures the same. Initially, we consider Tsao and Wright's (1983) concept of the maximum ratio, which checks for the distance between the maximum and minimum observed values proportional to the value of the lowest estimate. We find that the indicator does not allow us to compare more than two estimates at a time and, in certain cases, may drop the reported flows from the comparison altogether. Thus, we opt for the

Implicit Minimal Measurement Error (IMME) developed by Van Bergeijk (1995). This indicator was designed to capture the extent of inaccuracy in bilateral trade data and assumes that all available estimates are inaccurate to a certain degree. While intended to capture errors in differenced variables that take both positive and negative values, the IMME can also be utilized to measure the extent of inaccuracy present in population counts.

The IMME for immigration flow statistics of migrants from country *o* living in country *d* at time *t*, in the context of our study is given by:

$$IMME_{odt} = \frac{|SD\_Flow_{odt} - \mu_{odt}| + |MR\_Flow_{odt} - \mu_{odt}| + |Reported\_Flow_{odt} - \mu_{odt}|}{|SD\_Flow_{odt} + MR\_Flow_{odt} + Reported\_Flow_{odt}|} \tag{7}$$

Where μ_{odt} is the mean of SD_Flow_{odt}, MR_Flow_{odt} and $Reported\_Flow_{odt}$. The IMME also suffers from issues carried over by zero inflation of flow data. To begin with, the indicator takes a value of 1.33 by default if only one estimate is non-zero. Further, the indicator tends to wrap around 0.66 in cases where one of the three estimates is zero. This results in three large frequency spikes at 0.66, 1.33 and 0 which can generate misleading results.

Lastly, we utilize a dummy variable called Missing to capture cases of missing disaggregate flow statistics. The variable takes the value 1 if the country does not report any disaggregated migration flow statistics in a given year and 0 if country of origin wise migration flow statistics are reported to Eurostat. We must note that inconsistent disaggregation of flow statistics are not captured by the variable. Thus, even if a CEEC reports even one country of origin to Eurostat and aggregates the rest, the variable Missing takes the value 0.

5.2.2 Data Description of Measurement Error Indicators

Having outlined the methods utilized to develop the measurement error indicators, we now discuss the indicators in depth. The descriptive statistics for the dependent variables are provided in Appendix Table 7. Starting with our indicator of inaccuracy, the IMME has a possible range of negative infinity to infinity, however, comparing count data restricts its range from zero to infinity (van Bergeijk, 2017). ME, on the other hand, has a range of -1 to infinity. Table 6 presents the country wise mean and standard deviations of ME and IMME. Based on which we observe that CEECs report an average ME and IMME of 0.467 and 0.598 respectively. This implies that the reported flows to CEECs, on average, are approximately 47 percent lower than the mean of the estimated flows.

Additionally, the minimum possible measurement error (IMME) amongst the reported and estimated flows is approximately 59.8 percent. It is important to note that the standard deviation of ME is almost five times as high as that of IMME, which points towards greater variation in the undercount indicator. The Slovak Republic reports the largest ME and IMME, on average at 148 and 75.3 percent respectively. On the other hand, Latvia reports the lowest ME and IMME, on average, at 6.2 and 45.2 percent respectively.

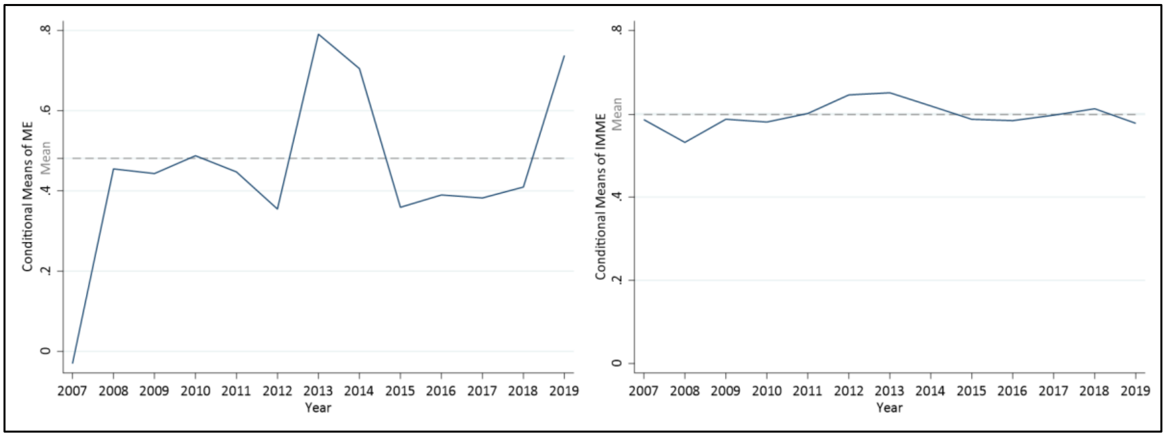
Table 5: Country of Residence-wise Mean Values of ME & IMME

Country of Residence	ME		IMME		Freq.
	Mean	Std. dev.	Mean	Std. dev.	
Bulgaria	0.353	1.156	0.607	0.425	1367
Czech Republic	0.572	1.383	0.596	0.352	2158
Estonia	0.332	1.518	0.467	0.429	1026
Hungary	0.138	2.743	0.568	0.338	2173
Latvia	0.062	0.828	0.462	0.452	684
Lithuania	0.074	1.482	0.544	0.489	855
Romania	0.818	5.975	0.651	0.380	1026
Slovak Republic	1.486	2.529	0.753	0.447	2018
Slovenia	0.088	0.766	0.584	0.451	2212
Total	0.467	2.357	0.598	0.420	13519

Source: Author's Elaboration

Figure 6 outlines the temporal evolution of IMME and ME, where the blue line traces their yearly mean value, and the grey lines represent their overall mean. Consistent with the observations from Table 6, we notice higher volatility in ME compared to IMME, with the latter loosely wrapping around its mean value. The shocks in ME might be driven by the unbalanced nature of our panel. For example, the sharp upward shift of ME from -0.03 in 2007 to 0.45 in 2008 might be down to the inclusion of the Slovak Republic, as our flow data for the country is only available post 2008. A secondary shock that we must discuss is the spike in ME in 2012-13 which causes the ME to rise from 0.35 to 0.79. This upward shift can be attributed to the inclusion of Romania in our sample, and the exclusion of the country causes the ME to drop by 20 percentage points (ME reduces from 0.79 to 0.6). On the other hand, the largest spike in IMME is much smaller in volume and takes place when IMME jumps from 0.53 in 2008 to 0.58 in 2009 (5 percentage points).

Figure 5: Time Evolution of ME and IMME



Source: Author's Elaboration

Table 7 outlines the cases where the dummy variable Missing takes the value one i.e., cases where the CEEC only reports aggregated immigration flow statistics. In all, there are 15 country year combinations where disaggregated flow data is unavailable, which comprise approximately 13 percent of all the country year combinations in our study. Our initial observations from Table 7 indicate that Czech Republic, Estonia Hungary, Lithuania, and Slovenia have no years with missing data. Latvia, on the other hand, has the highest occurrence of missing data as it reports disaggregated statistics for only four years within the selected time-period of our study. Our observations are further validated by Mooyart et. al. (2021), who find that the level of completeness of immigration flow data for Latvia is consistently below 2.5%. In contrast, a majority of the CEECs report with a completeness level above 97.5%, except for Bulgaria, which doesn't report any migration flow statistics between 2008 and 2011 and displays a relatively high level of completeness after that (Mooyaart, et al., 2021).

Table 6: Missing Annual Migration Flow Data in Eurostat Statistics

	Country									
Year	BG	CZ	EE	HU	LT	LV	RO	SI	SK	Total
2007	0	0	0	0	0	1	0	0	1	3
2008	1	0	0	0	0	1	0	0	0	2
2009	1	0	0	0	0	1	0	0	0	2
2010	1	0	0	0	0	1	0	0	0	2
2011	1	0	0	0	0	1	0	0	0	2
2012	0	0	0	0	0	1	0	0	0	1
2013	0	0	0	0	0	1	0	0	0	1
2014	0	0	0	0	0	0	0	0	0	0
2015	0	0	0	0	0	1	0	0	0	1
2016	0	0	0	0	0	1	0	0	0	1
2017	0	0	0	0	0	0	0	0	0	0
2018	0	0	0	0	0	0	0	0	0	0
2019	0	0	0	0	0	0	1	0	0	1
Total	4	0	0	0	0	9	1	0	1	15

Source: Author's Elaboration based on Eurostat (2022e)

***Note:** The variable takes the value 0 for Romania in 2007 despite missing flows as the country acceded to the EU in 2008

Chapter 6: Empirical Strategy

6.1 Econometric Model

The panel data models utilized for this study are divided into two distinct parts. We initially model the key determinants of ME and IMME using a REWB model as outlined by Allison (2009)[12] and Bell & Jones (2015). Post which, we employ a simple linear probability model with random effects to conduct an analysis of the determinants of missing migration data. The various equations of the econometric models and a brief description of the variables are presented in the following sections (Justification for variable selection is provided in Chapter 6).

6.1.1 Pre-Analysis

Prior to delving into the institutional determinants, we check for method dependency of the measurement error indicators. As previously mentioned, this involves regressing the IMME and ME on factors that might generate a bias in the estimated flows. The equations for which are as follows:

$$\begin{aligned} ME_{odt} = {} & \beta_0 + \beta_1 Natural\_Pop\_Change_{odt} + \beta_2 Naturalization_{odt} \\ & + \beta_3 Emigration_{odt} + \beta_4 Migrant\_Stock_{odt} + \beta_5 \bar{\alpha}_{od} + \delta_t + \alpha_{od} \\ & + \varepsilon_{odt} \end{aligned} \tag{8}$$

$$\begin{aligned} IMME_{odt} = {} & \beta_0 + \beta_1 Natural\_Pop\_Change_{odt} + \beta_2 Naturalization_{odt} \\ & + \beta_3 Emigration_{odt} + \beta_4 Migrant\_Stock_{odt} + \beta_5 \bar{\alpha}_{od} + \delta_t + \alpha_{od} \\ & + \varepsilon_{odt} \end{aligned} \tag{9}$$

We must note that there is potential endogeneity between our measurement indicators and the factors used for addressing method dependency as they are affected by and have an impact on immigration flows, which serves as the input data for ME and IMME. Due to this, regressions from our pre-analysis and the core analysis should viewed as exploratory regressions and the results must be interpreted as correlations and not causal effects.

[12] While conceptually same as Bell and Jones (2015), Alisson (2009) refers to the REWB model as a hybrid model.

Note: Standard errors are clustered at the country pair level for the analysis of ME and IMME and robust standard errors are utilized for the analysis of missingness

6.1.2 Core Analysis

Post analysing factors of method dependency, we extend the model to include our level three indicators to test for the mechanisms that affect source data as discussed on Chapter 3. These consist of public sector corruption, government effectiveness and the log transformed value of the gross expenditure on government sector R&D. The equation is given as follows:

$$\begin{aligned} ME_{odt} = \beta_0 &+ \beta_1 Natural\_Pop\_Change_{odt} + \beta_2 Naturalization_{odt} \\ &+ \beta_3 Emigration_{odt} + \beta_4 Migrant\_Stock_{odt} \\ &+ \beta_5 R\&D\_Exp_{dt} + \beta_6 PS\_Corruption_{dt} \\ &+ \beta_7 SqPS\_Corruption_{dt} + \beta_8 Gov\_Effectiveness_{dt} \\ &+ \ \beta_9 \bar{\alpha}_{od} + \beta_9 \bar{\gamma}_d + \delta_t + +\alpha_{od} + \varepsilon_{odt} \end{aligned} \tag{10}$$

$$\begin{aligned} IMME_{odt} = \beta_0 &+ \beta_1 Natural\_Pop\_Change_{odt} + \beta_2 Naturalization_{odt} \\ &+ \beta_3 Emigration_{odt} + \beta_4 Migrant\_Stock_{odt} \\ &+ \beta_5 R\&D\_Exp_{dt} + \beta_6 PS\_Corruption_{dt} \\ &+ \beta_7 SqPS\_Corruption_{dt} + \beta_8 Gov\_Effectiveness_{dt} \\ &+ \ \beta_9 \bar{\alpha}_{od} + \beta_9 \bar{\gamma}_d + \delta_t \ + \alpha_{od} + \varepsilon_{odt} \end{aligned} \tag{11}$$

6.1.2 Missingness Analysis

Having outlined our core models, we must also specify the random effects model utilized for our missingness analysis, the equation for which is as follows:

$$\begin{aligned} Missing_{dt} = \beta_0 &+ \beta_1 R\&D\_Exp_{dt} + +\beta_2 GDP\_pc_{dt} + \beta_3 TotalFlow_{dt} \\ &+ \beta_4 Population_{dt} + \alpha_d + \varepsilon_{dt} \end{aligned} \tag{12}$$

6.1.4 Description of Explanatory Variables

We now provide a discussion on the explanatory variables utilized as a part of the econometric models. Table 4 presents a list of the explanatory variables utilized along with a brief description of the same. This section also provides a discussion on the descriptive statistics which are reported in Appendix Table 7.

Table 7: List of Explanatory Variables

Variable Name	Description
$Natural\_Pop\_Change_{odt}$	The difference between the annual births and deaths divided by the total mid-year population.
$Naturalization_{odt}$	Log transformed value of the number of individuals from country of origin *o*, acquiring the citizenship of country of destination *d* at time *t*
$Emigration_{odt}$	Log transformed value of the number of non-nationals from country of origin *o*, moving out of country *d* at time *t*
$Migrant\_Stock_{odt}$	Log transformed value of the stock of migrants
$R\&D\_Exp_{dt}$	Log transformed value of gross expenditure on R&D in the government sector in millions of euros
$PS\_Corruption_{dt}$	VDEM Public Sector Corruption Index which is continuous within the range of 0-1
$SqPS\_Corruption_{dt}$	Squared value of VDEM Public Sector Corruption Index
$Gov\_Effectiveness_{dt}$	WGI on Government Effectiveness which is continuous within the range of -2.5 to 2.5 and is measured in standard normal units
GDP_pc_{dt}	Log transformed value GDP Per Capita measured in millions of current US $
$TotalFlow_{dt}$	Log transformed value of the total immigration flows to a country in a given year
$Population_{dt}$	Log transformed value of the total domestic population at the start of each year
$\bar{\alpha}_{od}$	Vector of cluster means of variables that vary over country pair and year
$\bar{\gamma}_{d}$	Vector of cluster means for variables that vary over country of current residence and year
δ_t	Vector of within and between year effects

6.1.4.1 Demographic Indicators

Starting with demographic indicators, the natural rate of population change or the natural population change is sourced from the UN World Population Prospects (WPP) database. However, the fertility and mortality patterns of migrants are extremely heterogeneous and often differ from country of origin and current residence (Aldridge et al., 2018; Desiderio, 2020; Kraus and González-Ferrer, 2021). As a result, we proxy the rate of natural increase for migrants by taking the average of the country of origin and current residence. We observe that the rate of natural increase for migrants in CEECs is 6.09 percent on average, in contrast to -1.68 percent for the nine CEECs, i.e., the rate of natural increase for migrants in CEECs is 7.07 percentage points higher than the overall rate for CEECs.

The statistics on emigration and naturalization of non-nationals are both sourced from Eurostat. As previously discussed, the CEECs are countries of high emigration. This is highlighted by an average outward movement of approximately 46500 non-nationals from the nine CEECs in a given year (Eurostat, 2022c). On the other hand, a very small number of non-nationals acquire the citizenship of one of the nine CEECs. On average, approximately 12 foreign nationals from a specific country acquire the citizenship of one of the CEECs. This is considerably lower than the European average of 127 over the same time period (Eurostat, 2022b). Appendix Table 7 shows that the indicator has a standard deviation of 215 which points towards varied integration and naturalization policies across the CEECs.

6.1.4.2 Economic Indicators

For the purpose of this study, we utilize two key economic indicators. To begin with, we source data on Research and Development expenditure from Eurostat's Gross Domestic Expenditure on Research and Development (GERD) database. GERD is measured at the country year level (Eurostat, 2022l). However, since we seek to identify innovation expenditure for administrative data collection systems and NSOs, we utilize GERD specific to the government sector. Appendix Table 7 indicates that the nine CEECs spend approximately 169 million euros on average on government sector research and development which is considerably lower than the EU average of 2604 million euros over the same time period (Eurostat, 2022e). Further, we source data on GDP per capita from World Bank's World Development Indicators (WDI) database. The variable is primarily utilized for the analysis of missingness.

6.1.4.3 Indicators of Institutional Quality

The Public Sector Corruption Index is sourced from the VDEM database and identifies the extent to which public sector officials grant favours in exchange for material inducements and the frequency of misutilization of state resources (Mcmann *et al.*, 2016). The index has a range from 0 to 1, with higher scores representing higher levels of public sector corruption. However, considering that the index is continuous within a given range, we can interpret it in terms of percent increases. The average indicator value for CEECs in the selected period is 0.23 which is considerably higher than the EU average of 0.13 (VDEM, 2022).

Secondly, we utilize Government Effectiveness indicator from World Bank's Worldwide Governance Indicators (WGI). The indicator captures the latent values of government effectiveness by aggregating data on the perceptions of the quality of public services provision and the bureaucratic efficiency of state institutions (Kaufmann, Kraay and Mastruzzi, 2010). The index also captures the state's commitment to good policymaking and the freedom of public services from political pressures. The indicator functions within a range of -2.5 to 2.5 with zero mean value and is to be interpreted in terms of standard normal units (Kaufmann, Kraay and Mastruzzi, 2010). Appendix Table 7 indicates that the average value of the government effectiveness index is 0.68, which implies that the perceived government effectiveness in the nine CEECs is 0.68 standard deviations higher than the global mean.

6.1.4.4 Geography & Policy Agreements

The variables on geography and policy agreements are not mentioned in Table x as they are only used for robustness checks. *Contiguity* is a dummy variable sourced from the CEPII GeoDist database that takes the value 1 if a country pair shares a border and zero if not. The definition of shared borders is restricted to land and rivers and, as a result, sea and lake borders are excluded (Mayer and Zignago, 2011). Additionally, we construct a dummy variable for Schengen membership based on the European Commission's (2022) definition of Schengen member states. The variable as a result takes the value 1 if both the country of origin and current residence are Schengen members in the given year and zero otherwise.

6.2 Estimation Technique

As previously mentioned, we utilize the Within Between Random Effects (REWB) model to analyse the determinants of undercount and accuracy. The REWB model allows for controlling of panel and time related heterogeneity as it is algebraically equivalent to a fixed effects model (Alisson, 2009; Bell & Jones, 2015; Schunk & Perales, 2017). However, the REWB model has a few key advantages that must be discussed. To begin with, fixed effects models get rid of the heterogeneity bias, but it comes at the cost of being unable to measure the effects of factors that have zero within panel variation. For example, shared land borders, distances between countries or other higher level variables (Bell & Jones, 2015). The REWB model, on the other hand, allows for the introduction of variables that vary over a higher level by explicitly modelling the within and between effects at the lowest level of variation, as a way to deal with the heterogeneity bias[13]. The between effects in this regard are represented by the panel level cluster means, and while of little interest in our specific study, their explicit modelling can make for interesting observations. Hence, the REWB demeans the level one variables and adds their panel/cluster level means in the model as explanatory variables. The model can simply be stated by:

$$y_{it} \ = \beta_0 + \beta_1(x_{it} - \bar{x}_i) + \beta_2\bar{x}_i + \beta_3 z_i + \alpha_i + \varepsilon_{it} \tag{12}$$

Where, x_{it} is a level one explanatory variable that varies over panel variable *i* and time *t* and $\bar{x}_i$ is the panel/cluster/group level mean of x_{it}. Additionally, z_i represents a higher level variable that only varies over the panel variable *i*. Considering that the REWB model is just a random effects model with additional covariates, the error is the same as that of a standard random effects model. Hence, the unexplained part of the model is a combination of the unobserved random effects α_i and the error term ε_{it} (Bell & Jones, 2015).

Further, we can establish an equivalency with the two-way fixed effects model through the incorporation of the within-between decomposed time effects (Bell and Jones, 2015). However, the data utilized for this study is more complex as it varies over three levels which requires further extrapolation. To begin with, all variables that vary over country pair and year represent level one indicators and as a result are within transformed and their cluster means are added to the model in addition to the within-between time effects (δ_t). The vector of country pair cluster means is presented as $\bar{\alpha}_{od}$ as highlighted by equation 8, 9, 10 and 11.

However, level two variables such as public sector corruption and government effectiveness only vary over country of current residence *d* and time *t*. As a result, they are not demeaned, but instead are added in their normal form in addition to their group level means i.e. means at country of current

[13] Most commonly the Hausman test is utilized to check if the between effects are biasing the results by comparing fixed and random effects estimates of the same model

residence level, in the form of the vector $\bar{\gamma}_d$, as highlighted in equations 10 and 11. Adding $\bar{\gamma}_d$ to the model allows us to get rid of the assumption that the two indicators are completely random and account for heterogeneity at the country of residence level (Wooldridge, 2021). Lastly, we utilize two level three variables for robustness checks, which are time invariant. These variables consist of shared land borders and country pairs with Schengen membership. Since, the level three variables are time invariant, we do not include their cluster means into the model and assume that they do not suffer from heterogeneity bias.

Shifting the focus towards the analysis of missingness, we must choose from three alternatives of estimation techniques, which are, pooled OLS, random effects and fixed effects. Making a sound choice implies that several tests must be undertaken (All the test results are provided in Appendix A.2). To begin with, we conduct the Breusch-Pagan Lagrange Multiplier Test to check for heteroskedasticity, where the null hypothesis is that there are no significant differences across the nine CEECs (Breusch and Pagan, 1980). Based on the results of the test provided in Appendix Table 3, we reject the null hypothesis and, as a result, rule out the possibility of using a pooled OLS approach. Secondly, we undertake the Hausman Specification Test to check for endogeneity and omitted variable bias (Wooldridge, 2012). The test can also be interpreted as a method to check for significant differences in the within and between effects. Not accounting for the between effects in case they are significantly different from the within effects, can bias the estimation outcomes (Bell and Jones, 2015). The test's null hypothesis is that random effects is the preferred method, which we fail to reject in line with the results provided in Appendix Table 4. Thus, we choose a random effects model for estimation purposes. Lastly, we check for potential time-oriented heterogeneity, by conducting the Wald test of joint significance (Wooldridge, 2012). The test checks if the effect of a set of explanatory variables is jointly equal to zero (Wooldridge, 2012). We undertake the test for all the year dummies (Wooldridge, 2012). In line with the results from Appendix Table 5, we argue that our model does not require the inclusion of year dummies as there is no evidence of time related heterogeneity. In conclusion, our analysis of missingness is conducted using a simple random effects model with robust standard errors.

Chapter 7: Results & Discussion

7.1 Pre-Analysis

The results of our panel data analysis for the REWB model that tests for method dependency are presented in Table 8, where Columns 1 and 2 present the results for ME and IMME respectively.

Table 8: Estimation Results with Factors that Generate Potential Method Dependency

	(1)	(2)
Variables	ME	IMME
Natural Population Change	-0.0536	-0.00841
	(0.0439)	(0.00932)
Naturalization (Log)	0.144	-0.00129
	(0.160)	(0.0123)
Emigration (Log)	-0.231***	-0.0186***
	(0.0787)	(0.00558)
Migrant Stock (Log)	0.607***	-0.187***
	(0.0850)	(0.0118)
Observations	10,725	10,725

Note: Robust standard errors are presented in parentheses and the coefficient for the constant and time effects are not reported
*** p<0.01, ** p<0.05, * p<0.1

According to our results, the natural rate of population change[14] and the reduction in migrant stocks caused by naturalization does not have a statistically significant relationship with either of measurement error indicators. On the other hand, we observe that migrant stocks and emigration have a statistically significant effect on both the measurement error indicators. To begin with, a percent increase in migrants leaving the country i.e., emigration of non-nationals leads to a reduction in ME by 0.23 percentage points. Further, Column 2 highlights that a percent increase in emigration results in a 0.019 percentage point reduction in IMME.

The negative effect of emigration on our measurement error indicators is as expected considering that migrant stocks can often be inflated due to low levels of deregistration. This in turn, causes our

[14] The natural rate of change is sensitive to the inclusion of migrant stocks, which causes it to lose statistical significance (see Appendix Table 2)

estimated flows to also be inflated, artificially increasing the size of the measurement error. Thus, as the number of observed emigrants increases, potential method dependency in our estimated flows decreases causing both measurement error indicators to decrease.

The results also indicate that a percent increase in migrant stocks leads to a 0.6 percentage point increase in ME. This implies that the extent to which migration flows to a CEEC are undercounted increases by roughly 0.6 percentage points for a percent increase in migrant stocks. This specific correlation is driven by the strong positive relationship between migrant stocks and flows, and points to the idea that larger migrant stocks are closely related to higher immigration flows which are more prone to undercounting. With regards to IMME, a percent increase in migrant stocks results in a 0.19 percentage point reduction in the levels of inaccuracy. Like ME, the relationship between IMME and migrant stocks also relates to the link between the volume of migration flows and stocks. That is, as migration stocks increase, so do the estimated and reported flows. As a result, the IMME declines with increasing migration flows due to decreasing absolute gaps relative to the overall flow size. For example, in 2018, 18491 Ukrainians moved to the Czech Republic as per Eurostat, (stock difference estimates reported 0 inflows and migration rates estimated 32179 incoming Ukrainian migrants) and the IMME when comparing the three was around 0.66. In the same year, only 10 Libyans were reported to have migrated to the country (stock difference estimates reported 0 inflows and migration rates estimated 55 incoming Libyans), but the low flow volume inflated the IMME to around 1.02. While both represent cases of high inaccuracy, the smaller flows from countries such as Libya can inflate IMME even if the absolute difference amongst the estimates is not as large. This is why increases in migrant stocks ends up reducing the IMME.

Overall, our results indicate that there is a case to be made for potential method dependency in our measurement error indicators, which is largely driven by emigration and the overall size of the migrant stocks. However, the effect sizes for the two variables are quite small in magnitude and are not expected to generate large biases in our measurement error indicators. Despite this, going forward, we identify these variables as potential controls for method dependency.

7.2 Core Analysis

The results for our core analysis are presented in Table 9. Starting with expenditure on government sector R&D, we observe no significant effect on ME. However, the variable has a positive and statistically significant relationship with IMME. Based on Column 10, we observe that a percent increase in expenditure on government sector R&D increases the IMME by around 0.11 percentage points. The direction of the coefficient contrasts our expectations, as we initially hypothesized that increases in government sector R&D might help foster a reduction in measurement errors through technological improvements in the data infrastructure or the development of new coping strategies to overcome limitations in the source data.

Second, we observe a statistically significant relationship between our measurement error indicators and public sector corruption. However, the inclusion of the quadratic term implies that they must be interpreted together. For this purpose, we differentiate with respect to public sector corruption and then impute the mean value of the index to generate the effect size[15]. The sample mean of the public sector corruption index is roughly 0.22 (See Appendix Table 7) and the sample mean of ME is 0.47. Thus, a percentage point increase in public sector corruption reduces ME by 4.2 percent at the mean (100*(12.13*2*0.22–- 7.3)/47). Given the non-linearity, we must also discuss the point of inflection by maximizing our differential equation. This yields an inflection point of roughly 0.3. Thus, as the index approaches the value of roughly 0.3, the negative effect of corruption peters out. Post which, public sector corruption increases the level of ME. The sample mean of IMME is roughly 0.6. Thus, a percentage point increase in public sector corruption results in 1.4 percent reduction in IMME at its mean (100*(1.791*2*0.22-1.65)/60)[16]. Similar to ME, we must also identify the point of inflection for IMME. Maximizing the differential equation yields an inflection point of 0.46 i.e. public sector corruption improves the accuracy of flow values until the point where the index reaches 0.46, post which, it leads to decreasing accuracy of flow values.

The results are in line with our expectations. From an institutional perspective, increasing levels of corruption initially lead to an increase in efficiency as the bureaucratic processes such as migrant

[15] Differentiating the equation with respect to Public Sector Corruption gives us the following:

$$\frac{\partial ME}{\partial Public\_Sector\ \_Corruption} = 12.13 * 2 * Public\_Sector\_Corruption - 7.3$$

[16] Differentiating the equation with respect to Public Sector Corruption gives us the following:

$$\frac{\partial IMME}{\partial Public\_Sector\ \_Corruption} = 1.791 * 2 * Public\_Sector\_Corruption - 1.652$$

registration are sped up in exchange for nominal gains in addition to the lack of adherence to data production protocols. . However, beyond a specific threshold the quality of state entities deteriorates to a point where the data collection and dissemination systems are affected. This drop in institutional quality in turn, reduces the accuracy of migration statistics.

Our results are also consistent from the perspective of the migrants' decision making. Initial increases in corruption can be seen as the erosion of inaccessible state institutions and a reduction in the costs of migrant registration, which reduces the level of undercount. However, beyond a specific level of corruption, the cost of reporting one's move and adherence to rules becomes too high and migrants respond by reducing their level of registration.

Lastly, we do not identify an effect for government effectiveness, which implies that the quality of public services provided, and migration related policymaking do not have a statistically significant effect on measurement errors in migration statistics.

Overall, we observe varied results for the institutional determinants of measurement errors, some of which contrast our expectations. Expenditure on government sector R&D, to begin with, does not significantly affect undercounting, however, it increases the level of inaccuracy prevalent in flow statistics. Public sector corruption is shown to decrease undercount and inaccuracy, as the incentives for migrants' self-reporting and the efficiency gained by skirting bureaucratic rules and regulations for data production are higher than the cost in the context of our study. The overall negative sign of public sector corruption can also be attributed to the inflection points lying at the tail end of our sample. Lastly, we do not observe any significant effect on measurement errors for the quality of public services (government effectiveness).

Table 9: Hybrid Model with All Core Determinants

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	ME	ME	ME	ME	ME	IMME	IMME	IMME	IMME	IMME
Natural Population Change	-0.0536	-0.0534	-0.0484	-0.0386	-0.0484	-0.00841	-0.00878	-0.0104	-0.00878	-0.00890
	(0.0439)	(0.0440)	(0.0410)	(0.0388)	(0.0397)	(0.00931)	(0.00925)	(0.00919)	(0.00914)	(0.00920)
Naturalization (Log)	0.144	0.144	0.143	0.152	0.143	-0.00129	-0.00169	-0.00119	0.000349	0.000229
	(0.160)	(0.160)	(0.159)	(0.159)	(0.161)	(0.0123)	(0.0124)	(0.0126)	(0.0126)	(0.0126)
Emigration (Log)	-0.231***	-0.232***	-0.233***	-0.230***	-0.233***	-0.0186***	-0.0165***	-0.0160***	-0.0155***	-0.0155***
	(0.0786)	(0.0788)	(0.0773)	(0.0776)	(0.0779)	(0.00558)	(0.00553)	(0.00558)	(0.00552)	(0.00554)
Migrant Stock (Log)	0.607***	0.606***	0.608***	0.608***	0.615***	-0.187***	-0.186***	-0.187***	-0.187***	-0.187***
	(0.0849)	(0.0848)	(0.0837)	(0.0835)	(0.0852)	(0.0118)	(0.0117)	(0.0117)	(0.0116)	(0.0118)
R&D Expenditure (Log)		-0.0780	-0.0690	-0.179	-0.156		0.132***	0.129***	0.111***	0.112***
		(0.131)	(0.131)	(0.128)	(0.128)		(0.0304)	(0.0303)	(0.0305)	(0.0305)
Public Sector Corruption Index			1.174	-6.473***	-7.269***			-0.384**	-1.653***	-1.652***
			(1.360)	(1.800)	(1.913)			(0.172)	(0.387)	(0.391)
Public Sector Corruption Index^2				10.78***	12.13***				1.788***	1.791***
				(3.763)	(3.879)				(0.527)	(0.537)
Government Effectiveness Index					-0.474					-0.00531
					(0.323)					(0.0547)
Observations	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725

Note: Robust standard errors are presented in parentheses and the coefficient for the constant and time effects are not reported

*** p<0.01, ** p<0.05, * p<0.1

7.3 Analysis of Missingness

We now pivot towards our auxiliary analysis and focus on the missingness of information. As outlined in the previous chapters, the variable Missing takes the value 1 for cases where countries do not report disaggregated immigration flow statistics and 0 otherwise. The results of the random effects linear probability model with Missing as the dependent variable are provided in Table 10. Upon preliminary observation of Table 6, we identify that certain countries have no missing data during the time period of our study. Thus, we present two sets of estimates. In Column 1, the sample consists of all CEECs and Column 2 where the sample is restricted only to countries with at least one year of missing data.

Table 10: Estimation Results of Missing Data

	(1)	(2)
Variables	Missing	Missing
R&D Expenditure	-0.142	-0.530***
	(0.0893)	(0.163)
GDP Per Capita (Log)	-0.0666	0.415
	(0.409)	(0.287)
Total Immigration Flow (Log)	0.0109	0.0990
	(0.0293)	(0.0668)
Population (Log)	0.0567	0.157
	(0.136)	(0.216)
Observations	112	47
R-squared	0.114	0.398

Note: Robust standard errors are presented in parentheses and the coefficient for the constant and time effects are not reported

*** p<0.01, ** p<0.05, * p<0.1

Based on Column 1 we observe that none of our explanatory variables are statistically significant in the full sample, however, sample restriction to the four countries with at least one year of missing data causes gross expenditure on government sector R&D to gain statistical significance. The coefficient indicates that a percent increase in R&D expenditure leads to a 53 percentage point reduction in the probability of missing data. The lack of statistical significance of other explanatory variables also makes for an interesting point of analysis, as we cannot argue that missing data across either of the samples is driven by the size of domestic population, volume of incoming migrants or the GDP per capita of the country. As a result, in the given context, the costs of maintaining registers (which is proxied by the size of the domestic population) or the dynamic nature of migration processes cannot be held responsible for missing disaggregated flow data. However, we must note that these effects, or the lack of, cannot be generalized as the model suffers from an exceedingly small sample size.

7.4 Robustness Checks

This section is concerned with robustness checks undertaken to test whether our results hold across varied specifications and subsamples. We only conduct robustness checks for our core analysis due to data related limitations in the missingness analysis.

For this purpose, we initially propose a subsample to exclude small immigration flows. Migration data, especially in the context of CEECs is heavily skewed towards the right due small volume of flows. Low values of immigration flows can exaggerate the measurement error and potentially render our results spurious. In order to test the same, we identify an exclusion criterion to drop the country with the smallest number of incoming migrants. Slovakia falls within this category, as the country reports the lowest number of incoming migrants, with approximately 6500 non-nationals moving to the country on average in our period of study (Eurostat, 2022f). Thus, we drop the country from our sample and estimate our model for ME and IMME to check for potential differences in our results. Table 11 presents our results of the first robustness check where the results from the unrestricted sample are presented in Column 1 and 3 and Columns 2 and 4 present our results after dropping Slovakia.

Our estimates are robust to the sample restriction bar a few exceptions. To begin with, the natural rate of population change gains statistically significance for ME, although we observe no such change when the dependent variable is swapped for IMME. Further, emigration's effect size marginally reduces for ME, but loses all statistical significance in the case of IMME. Similarly, R&D expenditure in government sectors loses its statistical significance for IMME post sample restriction. On the other hand, government effectiveness gains statistical significance for ME but is only significant at the ten percent level.

Table 11: Robustness Check Results with Restricted Sample

Variables	(1) ME	(2) ME	(3) IMME	(4) IMME
Natural Population Change	-0.0484	-0.0952**	-0.00890	-0.00920
	(0.0397)	(0.0440)	(0.00920)	(0.0107)
Naturalization (Log)	0.143	-0.0549	0.000229	-0.00413
	(0.161)	(0.155)	(0.0126)	(0.0138)
Emigration (Log)	-0.233***	-0.174**	-0.0155***	-0.00452
	(0.0779)	(0.0849)	(0.00554)	(0.00570)
Migrant Stock (Log)	0.615***	0.568***	-0.187***	-0.216***
	(0.0852)	(0.0867)	(0.0118)	(0.0121)
R&D Expenditure (Log)	-0.156	0.164	0.112***	0.000510
	(0.128)	(0.241)	(0.0305)	(0.0438)
Public Sector Corruption Index	-7.269***	-6.537***	-1.652***	-1.663***
	(1.913)	(2.012)	(0.391)	(0.409)
Public Sector Corruption Index2	12.13***	10.37***	1.791***	1.823***
	(3.879)	(3.958)	(0.537)	(0.554)
Government Effectiveness Index	-0.474	0.606*	-0.00531	0.0552
	(0.323)	(0.347)	(0.0547)	(0.0602)
Observations	10,725	8,940	10,725	8,940

Note: Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Moving towards our second robustness check, we now seek to ensure if our results hold when additional factors that might affect the measurement error are accounted for. As previously discussed, border contiguity (shared borders) and Schengen membership play a key role in the migration dynamics of CEECs, as they encourage the flow of highly mobile populations. To account for the same, we introduce two dummy variables – Contiguity and Schengen. Contiguity takes the value one when countries share a border and zero if they don't. Schengen, on the other hand, captures country pairs where both, the country of migrant origin and current residence are members of the Schengen area in the given year.

The results of our extended model are presented in Table 11. Columns 1 and 3 present the results from our standard model while Columns 2 and 4 present the extended results. Overall, we conclude that our analysis is robust to the inclusion of the two indicators as there is barely any change in any of the core results. However, we still interpret the coefficients of the two new dummies. Contiguity does not have statistically significant effect on ME in contrast to IMME, as on average countries sharing a border report a 10 percentage points higher IMME than the countries that do not share a border after controlling for method dependency and institutional factors. The results, imply that while contiguity might not affect the extent of undercounting, it affects the accuracy of official statistics for migrants moving from neighbouring countries.

Further, Schengen membership only has a statistically significant relationship with ME. In this regard, migration between country pairs that are Schengen members report a 22.5 percentage points lower ME than the country pairs where either one or both countries are not Schengenn members. Although the sign of the coefficient contrasts our expectations given that highly mobile populations are expected to generate higher levels of undercount, the directionality can be attributed to the proactive approach taken by the EU towards ensuring consistency in statistical reporting amongst members states.

Table 12: Robustness Check Results with Model Extension

Variables	(1) ME	(2) ME	(3) IMME	(4) IMME
Natural Population Change	-0.0484	-0.0484	-0.00890	-0.00890
	(0.0397)	(0.0397)	(0.00920)	(0.00920)
Naturalization (Log)	0.143	0.143	0.000229	0.000233
	(0.161)	(0.161)	(0.0126)	(0.0126)
Emigration (Log)	-0.233***	-0.233***	-0.0155***	-0.0155***
	(0.0779)	(0.0779)	(0.00554)	(0.00554)
Migrant Stock (Log)	0.615***	0.615***	-0.187***	-0.187***
	(0.0852)	(0.0852)	(0.0118)	(0.0118)
R&D Expenditure (Log)	-0.156	-0.156	0.112***	0.112***
	(0.128)	(0.128)	(0.0305)	(0.0305)
Public Sector Corruption Index	-7.269***	-7.262***	-1.652***	-1.651***
	(1.913)	(1.909)	(0.391)	(0.391)
Public Sector Corruption Index^2	12.13***	12.12***	1.791***	1.791***
	(3.879)	(3.875)	(0.537)	(0.537)
Government Effectiveness Index	-0.474	-0.474	-0.00531	-0.00515
	(0.323)	(0.323)	(0.0547)	(0.0547)
Contiguity		0.366		0.105***
		(0.267)		(0.0407)
Schengen		-0.225*		-0.0300
		(0.133)		(0.0208)
Observations	10,725	10,725	10,725	10,725
Number of groups	1,518	1,518	1,518	1,518

Note: Robust standard errors in parentheses and the constant is not reported

*** p<0.01, ** p<0.05, * p<0.1

7.5 Implications & Lessons Learnt

This sub-section delves into the lessons learnt and implications from the study. Overall, this paper has been an empirical exercise aimed at contextualizing the various forms of measurement errors in migration flow statistics and identify the determinants of the same.

To begin with, there are a multitude of methods that can be used to estimate migration flows. One of the main advantages of using these methods is that they can be used to validate reported flow estimates and help bridge the gaps in reported values where necessary. In the case of CEECs, the need to bridge this gap has been identified by governments due to sub-par administrative data (Willekens, 2019). Hence, we believe that making use of these techniques could help to produce error-free migration statistics

Further, knowledge about the types of measurement errors prevalent in migration statistics can have a variety of implications in the context of CEECs. From a research perspective, accounting for measurement errors in traditional migration modelling and demographic research can help generate more robust results. From a policymaker's perspective, however, accounting for measurement errors can help design more efficient migration policies and better understand the migration processes of the country (Sales 2022).

This study observes that public sector corruption reduces errors of inaccuracy and undercounting in migration flow statistics. However, we argue that this highlights the rigidity of state institutions more than the gains from corruption. Such institutions deter migrants from registering themselves, while also relying on data related protocols that are not conducive for statistical production. Thus, the policy responses aimed at reducing the measurement errors should be focused on making registration and deregistration accessible for migrants and ensuring that the internal protocols for statistical production are directed towards improving the country's statistical capacity.

Lastly, expenditure on research and development can provide mixed results, as it might only achieve the desired effect when it comes to reducing the probability of presenting missing data. However, our findings on the subject are limited to completely missing data and cannot be generalized to cases of partially missing information, which is a larger and more complex issue (Aristotelous et al., 2022).

Chapter 8: Conclusion

This paper sought to conceptualize the types of measurement errors prevalent in immigration flow statistics and identify the determinants of upstream data quality affects the same. Nine CEECs are chosen as the focus of this study due to a relatively recent increase in the demand for harmonized and error-free immigration statistics.

After discussing the dynamics of immigration flows to the selected CEECs, we contextualize the data sources and processes used to produce migration statistics in these countries. Given the availability and low costs, the countries largely rely on administrative data sources to produce population and other vital statistics (Eurostat, 2018b). The administrative data is obtained from a variety of sources such as local population registers, social security databases and tax databases (CROS, 2022). However, NSOs, often do not have complete access to administrative data sources, but instead are supplied with truncated files by the respective state entities for generating official statistics (Eurostat, 2018a). As a result, the factors affecting administrative data collection systems have a direct impact on the quality of official statistics.

We conceptualized the forms of measurement errors in flow statistics in line with Raymer et al.'s (2013) approach undertaken as a part of the IMEM project. Based on which, measurement errors in migration statistics can be classified as errors of coverage, undercounting and inaccuracy. However, errors of coverage can be highly contextual and difficult to quantify, which is why we exclude them from our study. Additionally, we identify missing data as an extreme form of measurement error (Blackwell et al., 2017).

Conducting an analysis of measurement errors, requires the construction of alternative migration flows and designing indicators to quantify the measurement errors. To begin with, we utilize the stock difference and migration rate approaches as developed by Beine et al. (2011) and Dennett (2015) to generate two alternative sets of migration flow estimates. Based on which, we construct an indicator to measure undercount (ME). The indicator captures the difference between the mean of the estimated flows and the reported flow values as a ratio of the reported flows. Further, we utilize the IMME as designed by Van Bergeijk (1995) to assess the accuracy of all available flow values. The indicator allows for the computation of the minimal possible error across the two generated and the reported estimates of immigration flows. Lastly, we generate a dummy variable called Missing to capture cases of missing disaggregated migration flow statistics.

As a part of our analysis, we initially test for method dependency and find that the indicators of undercount and accuracy have a statistically significant relationship with migrant stocks and emigration, however, the effects are small in magnitude and not noteworthy. The results of our core empirical analysis support our initial arguments regarding a potential non-linear relationship between public sector corruption and measurement errors, as increases in public sector corruption is seen to initially decrease the measurement error and, beyond a specific threshold, the relationship between the two variables changes and the measurement error indicators begin to increase. The inflection points for ME and IMME, in this regard, are 0.3 and 0.46 respectively.

However, we find that the overall impact of public sector corruption is positive[17], which seems to contrast Aidt's (2003) arguments of an overall detrimental effect of corruption. To be specific, a percentage point increase in the public sector corruption index is associated with a 4.4 percent and 1.4 percent reduction in ME and IMME respectively. The positive effect can be attributed to the inflection points lying in the tail end of our sample, and generally low levels of corruption in the nine CEECs. This is to say that, in the context of our study, the efficiency gains are still larger than the drop in the quality of administrative data systems leading to higher accuracy (reduction in IMME). Thus, skirting of data production protocols and associated regulations can help increase the accuracy of flow estimates. Further, the benefits of bribing state officials for migrant registration are higher than the costs leading to reducing levels of undercount.

Nevertheless, we hold reservations towards this positive impact of rising corruption levels as larger samples might end up showing the detrimental effects. Additionally, the directionality of the effect is expected to be more representative of the restrictive nature of the state institutions and their regulations than potential gains from corruption. Thus, policy responses aimed at reducing measurement errors should rather focus on making registration and deregistration accessible for migrants and ensuring that the data related regulations are conducive for statistical production.

With regards to missing flow data, we do not observe any statistically significant effects in the unrestricted sample. However, restricting the sample only to countries that had some years of missing data causes gross government sector R&D expenditure to significantly reduce the probability of missing data. However, it is important to consider that the results are affected by a small sample size prior to drawing any conclusions about the same.

To check for the validity of our results (for ME and IMME) across varied sub-samples and model extensions, we conduct two robustness checks. First, we drop Slovakia from our sample to deal with potential biases in our measurement error indicators caused by low flow volumes. Second, we extend our core models to include additional controls that account for high mobility

[17] A positive impact refers to the reduction in measurement errors

migration channels. These consist of immigration amongst countries with a shared land border and country pairs where the country of origin and destination are both Schengen member states in the given time period. Based on Table 10 and 11 we conclude that our results for undercounting, as well as inaccuracy, are robust to the sample restriction and model extensions.

The paper has a few key limitations that require consideration. First is the assumption that the migrant flows derived from stock data are less prone to measurement errors than the reported flow values (see Abel (2010,2013), Abel & Cohen (2019), Raymar et al (2013), Azose & Raftery (2019)). The assumption is based on the idea that static populations are easier to measure in comparison to dynamic demographic processes. However, if the assumption breaks down, then the results of our study would be rendered spurious as the estimated flows would be just as, if not more, susceptible to measurement errors. Second, this study fails to capture demographic variations with and across migrant sub-populations that might affect the measurement error. For example, variations in age and gender composition can influence the measurement error. Decomposition of the data along these demographic lines as well as levels of integration might help explain why specific migrant groups are more prone to measurement errors. However, we are unable to test for the same due to data related limitations (Mooyart et al., 2021). Further, measurement errors might be influenced by factors affecting statistical capacity beyond the mechanisms tested in this study. For example, the sophistication of data infrastructure, funding for NSOs and independence of official statistics from political pressures. Lastly, this paper only treats missing data as a measurement error in cases where there is no disaggregated flow data available. However, a larger and more prominent issue is that of inconsistent disaggregation and partially missing migration flow data, which this study fails to account for (Mooyart et al., 2021).

It is also important to emphasize, that the results from this study are exploratory and do not seek to make any causal claims about the determinants of measurement errors. However, there are a few additional policy and research implications that we must acknowledge. First, the findings of this study can be utilized to account for measurement errors in traditional modelling of migration which can help generate more robust results. Second, from a policymaking perspective, adoption of a critical lens towards measurement errors in migration statistics can help improve the efficacy of migration management and integration policies in Central and Eastern Europe. Lastly, this paper makes an indirect contribution towards shifting the narrative surrounding CEECs from countries of net emigration and transit migration to that of potential destinations for migrants.

References

Abel, Guy J. (2010) "Estimation of international migration flow tables in Europe", *Journal of the Royal Statistical Society: Series A (Statistics in Society)* , 173(4), pp. 797–825. Available at: https://doi.org/10.1111/J.1467-985X.2009.00636.X.

Abel, G.J. (2013) "Estimating global migration flow tables using place of birth data", *Demographic Research*, 28, pp. 505–546. Available at: https://doi.org/10.4054/DemRes.2013.28.18.

Abel, G.J. and Sander, N. (2014) "Quantifying global international migration flows", *Science*, 343(6178), pp.1520-1522. Available at: https://doi.org/10.1126/science.1248676

Abel, G.J. and Cohen, J.E. (2019) "Bilateral international migration flow estimates for 200 countries", *Scientific Data*, 6(1). Available at: https://doi.org/10.1038/s41597-019-0089-3.

Aidt, T.S. (2003) "Economic analysis of corruption: A survey", *Economic Journal*, 113(491). Available at: https://doi.org/10.1046/j.0013-0133.2003.00171.x.

Aldridge, R.W. et al. (2018) "Global patterns of mortality in international migrants: a systematic review and meta-analysis", *Lancet (London, England)*, 392(10164), p. 2553. Available at: https://doi.org/10.1016/S0140-6736(18)32781-8.

Allison, P.D. (2009) "Fixed effects regression models", *SAGE Publications*, Available at: https://dx.doi.org/10.4135/9781412993869

Anderson, D.M. and Whitford, A.B. (2017) "Developing knowledge states: Technology and the enhancement of national statistical capacity", *Review of Policy Research*, *34*(3), pp.400-420, Available at: https://doi.org/10.1111/ropr.12230

Aristotelous, G., Smith, P.W.F. and Bijak, J. (2022) "Technical report: Estimation methodology"

Azose, J.J. and Raftery, A.E. (2019) "Estimation of emigration, return migration, and transit migration between all pairs of countries", *Proceedings of the National Academy of Sciences of the United States of America*, 116(1), pp. 116–122. Available at: https://doi.org/10.1073/pnas.1722334116.

Bell, A. and Jones, K. (2015) "Explaining Fixed Effects: Random Effects Modelling of Time-Series Cross-Sectional and Panel Data", *Political Science Research and Methods*, 3(1), pp. 133–153. Available at: https://doi.org/10.1017/psrm.2014.7.

Bergeijk, P., Bergeijk and Peter (2017) "Making Data Measurement Errors Transparent: The Case of the IMF", *World Economics*, 18(3), pp. 133–154. Available at: https://EconPapers.repec.org/RePEc:wej:wldecn:679 (Accessed: 21 October 2022).

Berlemann, M., Haustein, E. and Steinhardt, M.F. (2021) "From Stocks to Flows: Evidence for the Climate-Migration-Nexus". Available at: https://ssrn.com/abstract=3865474 (Accessed: 28 October 2022).

Bertoli, S. and Fernández-Huertas Moraga, J. (2015) "The size of the cliff at the border", *Regional Science and Urban Economics*, 51, pp. 1–6. Available at: https://doi.org/10.1016/J.REGSCIURBECO.2014.12.002.

Bijak, J. and Wisniowski, A.W. (2010) "Bayesian forecasting of immigration to selected European countries by using expert knowledge", *J. R. Statist. Soc. A*, pp. 775–796.

Blackwell, M., Honaker, J. and King, G. (2017) "A Unified Approach to Measurement Error and Missing Data: Overview and Applications", *Sociological Methods & Research*, 46(3), pp. 303–341. Available at: https://doi.org/10.1177/0049124115585360.

Bobeva-Filipova, D. (2017) "Migration: Recent Developments in Bulgaria", *SSRN Electronic Journal*, Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2929544

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F. and Spyratos, S. (2022) "Data Innovation in Demography, Migration and Human Mobility", *Publications Office of the European Union, Luxembourg*, Available at: https://doi:10.2760/027157

Braaksma, B., Zeelenberg, K. and de Broe, S. (2020) "Big Data in Official Statistics: A Perspective from Statistics Netherlands", *Big Data Meets Survey Science: A Collection of Innovative Methods*, pp. 303–338. Available at: https://doi.org/10.1002/9781118976357.CH10.

Breusch, T.S. and Pagan, A.R. (1980) The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. Available at: https://about.jstor.org/terms.

Buettner, T. (2021) "Stocktaking of Migration Data", *KNOMAD Working Papers*, 42. Available at: www.knomad.org

Council of the European Union. (2022) "The Schengen Area.", Available at: https://www.consilium.europa.eu/en/policies/schengen-area/#members [Accessed: 15 Aug. 2022]

CROS (2022) "Administrative Data", Available at: https://ec.europa.eu/eurostat/cros/content/administrative-data-0_en[Accessed: 15 Aug. 2022]

Dennett, A. (2015) "Estimating an Annual Time Series of Global Migration Flows–- An Alternative Methodology for Using Migrant Stock Data", *Global Dynamics: Approaches from Complexity Science*, pp. 125–142. Available at: https://doi.org/10.1002/9781118937464.CH7.

Desiderio, R. (2020) "The Impact of International Migration on Fertility: An Empirical Study.", *KNOMAD Working Papers*, Available at: www.knomad.org . (Accessed: 22 December 2022).

European Commission (2022), "The Schengen visa" [online], *European Commission*, Available at: https://home-affairs.ec.europa.eu/schengen-visa_en [Accessed 15 Nov. 2022]

European Parliament and Council of the European Union (2007) "Regulation (EC) No 862/2007 on Community statistics on migration and international protection", *Official Journal of the European Union*, Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007R0862

European Union Agency for Fundamental Rights (2016) "Criminalisation of migrants in an irregular situation and of persons engaging with them", Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2014-criminalisation-of-migrants-1_en.pdf

Eurostat (2018a) "Good practices in accessing, using and contributing to the management of administrative data", *Eurostat*, Available at: https://ec.europa.eu/eurostat/cros/system/files/admin-wp1.2_good_practices_final.pdf

Eurostat (2018b) "Power from Statistics: Data, Information and Knowledge Outlook Report" *Eurostat*, Available at: https://doi.org/10.2785/09698.

Eurostat (2022a) "Acquisition and Loss of Citizenship Reference Metadata in Euro SDMX Metadata Structure (ESMS)", Available at: https://ec.europa.eu/eurostat/cache/metadata/en/migr_acqn_esms.htm
Eurostat (2022b), Acquisition of citizenship by age group, sex and former citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_ACQ/default/table?lang=en&category=migr.migr_cit.migr_acqn [Accessed 25 Nov. 2022]

Eurostat (2022c), Emigration by age group, sex and citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_EMI1CTZ/default/table?lang=en&category=migr.migr_cit.migr_emi [Accessed 25 Nov. 2022]

Eurostat (2022d), Emigration by age group, sex and country of birth [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_EMI4CTB/default/table?lang=en&category=migr.migr_cit.migr_emi [Accessed 25 Nov. 2022]

Eurostat (2022e), GERD by sector of performance [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/rd_e_gerdtot/default/table?lang=en [Accessed 25 Nov. 2022]

Eurostat (2022f), Immigration by age group, sex, and citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_IMM1CTZ/default/table?lang=en&category=migr.migr_cit.migr_immi [Accessed 25 Nov. 2022]

Eurostat (2022g), Immigration by age group, sex, and citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_IMM3CTB/default/table?lang=en&category=migr.migr_cit.migr_immi [Accessed 25 Nov. 2022]

Eurostat (2022h), Immigration by age group, sex, and citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_POP1CTZ/default/table?lang=en&category=demo.demo_pop [Accessed 25 Nov. 2022]

Eurostat (2022i) "Immigration: Reference Metadata in Euro SDMX Metadata Structure (ESMS)", Available at: https://ec.europa.eu/eurostat/cache/metadata/en/migr_immi_esms.htm

Eurostat (2022j), Population on 1 January by age group, sex, and citizenship [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_POP1CTZ/default/table?lang=en&category=demo.demo_pop [Accessed 25 Nov. 2022]

Eurostat (2022k), Population on 1 January by age group, sex, and country of birth [online] *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/MIGR_POP3CTB/default/table?lang=en&category=demo.demo_pop [Accessed 25 Nov. 2022]

Eurostat (2022l), “Research and development (R&D): Reference Metadata in Euro SDMX Metadata Structure (ESMS)”, Available at: https://ec.europa.eu/eurostat/cache/metadata/en/rd_esms.htm, [Accessed 25 Nov. 2022]

Eurovoc (2022) “Western Countries”, Available at: https://op.europa.eu/s/xlZx [Accessed on 20 Nov. 2022]

Filipec, O. and Vargová, N. (2019) “Perception of Migration from Non-EU Countries in Slovakia: The Case of Nitra Region”, *EJTS European Journal of Transformation Studies*, 7(2), pp. 165–175. Available at: https://czasopisma.bg.ug.edu.pl/index.php/journal-transformation/article/view/5017/4402

Frontex (2022) Detections of Illegal Border-Crossings Statistics [online] *Frontex*. Available at: https://frontex.europa.eu/we-know/migratory-map/ [Accessed 25 Nov. 2022]

Gárdos, É. and Gödri, I. (2014) “Analysis of existing migratory data production systems and major data sources in eight South-East European countries”, *Hungarian Demographic Research Institute, Budapest*. Available at: https://www.demografia.hu/en/downloads/Projects/SEEMIG/outputs/SEEMIGWorkingPapers2.pdf

Geddes, A. and Scholten, P. (2018) “In the Shadow of the “Fortress”? Migration Dynamics in Central and Eastern Europe”, *The Politics of Migration and Immigration in Europe*, pp. 195–215. Available at: https://doi.org/10.4135/9781473982703.N9.

Georgiou, A. V. (2021) “The manipulation of official statistics as corruption and ways of understanding it”, *Statistical Journal of the IAOS*, 37(1), pp. 85–105. Available at: https://doi.org/10.3233/SJI-200667.

Gomez, S.L. and Glaser, S.L. (2006) “Misclassification of race/ethnicity in a population-based cancer registry (United States).”, *Cancer Causes and Control : CCC”*, 17(6), pp. 771–781. Available at: https://doi.org/10.1007/S10552-006-0013-Y.

Hand, D.J., (2018) “Statistical challenges of administrative and transaction data.”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3), pp.555-605. Available at: https://doi.org/10.1111/rssa.12315

International Monetary Fund (2022) “Special Data Dissemination Standard Plus: Slovakia Republic Population” [online] *International Monetary Fund*, Available at: https://dsbb-imf-org.eur.idm.oclc.org/sddsplus/dqaf-base/country/SVK/category/POP00

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2011) "The Worldwide Governance Indicators: Methodology and Analytical Issues" *Hague journal on the rule of law*, 3(2), pp.220-246. Available at: https://papers-ssrn-com.eur.idm.oclc.org/abstract=1682130 (Accessed: 23 October 2022).

Keilman, N. and Aristotelous, G. (2021) "Expert opinion on migration data. QuantMig Deliverable D6.1", Available at: https://www.duo.uio.no/bitstream/handle/10852/88033/1/QuantMig%2BDeliverable%2B6.1%2BExpert%2Bopinion%2Bon%2Bmigration%2Bdata.pdf

Kim, E.Y. (2022) "Statistical capacity building in developing countries: essays on aid effectiveness, sustainability, and measurement (Doctoral dissertation)", University of Texas at Austin, Available at: http://dx.doi.org/10.26153/tsw/43865

Kovalenko, J., Mensah, P., Leončikas, T. and Žibas, K. (2010) "New Immigrants in Estonia, Latvia, and Lithuania", *Legal Information Centre for Human Rights*, Available at: https://ec.europa.eu/migrant-integration/library-document/new-immigrants-estonia-latvia-and-lithuania_en

Krasteva, A. (2019) "The Bulgarian Migration Paradox: Migration & Development in Bulgaria", *Cáritas Bulgaria* , Available at: https://ec.europa.eu/migrant-integration/sites/default/files/2019-06/Migracionen-paradoxEng.pdf

Kraus, E.K. and González-Ferrer, A. (2021) "Fertility Differences Between Migrants and Stayers in a Polygamous Context: Evidence from Senegal", *Journal of International Migration and Integration* [Preprint]. Available at: https://doi.org/10.1007/s12134-020-00802-0.

Mancheva, M. and Troeva, E. (2011) "Migrations to and from Bulgaria: the State of Research", *Migrations, gender and intercultural interactions in Bulgaria*, pp.13-61, Available at: https://www.researchgate.net/publication/305402899_Migrations_Gender_and_Intercultural_Interactions_in_Bulgaria

Martí, M. and Ródenas, C., 2007. "Migration estimation based on the Labour Force Survey: An EU-15 perspective.", *International Migration Review*, *41*(1), pp.101-126. Available at: https://journals.sagepub.com/doi/10.1111/j.1747-7379.2007.00058.x

Mayer, T. and Zignago, S. (2011) "Notes on CEPII"s distances measures: The GeoDist database" Available at: http://www.cepii.fr/anglaisgraph/bdd/distances.htm.

McMann, K.M., Pemstein, D., Seim, B., Teorell, J. and Lindberg, S.I. (2016) "Strategies of Validation: Assessing the Varieties of Democracy Corruption Data". Available at: www.v-dem.net.

Mooyaart, J., Dańko, M.J., Costa, R. and Boissonneault, M. (2021) "Quality assessment of European migration data. Changes". Available at: www.quantmig.eu.

OECD (2015) "Frascati Manual 2015", *OECD (The Measurement of Scientific, Technological and Innovation Activities)*. Available at: https://doi.org/10.1787/9789264239012-en.

OECD (2020) "Latvia, in International Migration Outlook", *OECD (International Migration Outlook)*. Available at: https://doi.org/10.1787/ec98f531-en.

Pemstein, D., Marquardt, K.L., Tzelgov, E., Wang, Y.T., Medzihorsky, J., Krusell, J. and Römer, J.V. (2022), "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data", *V-Dem Working Paper*, 21, Available at: https://ssrn.com/abstract=3595962.

Radavoi, C.N. (2015) "Factors Countervailing Immigrant Phobia: A Paradoxically Successful Case of Chinese Migration", *Romanian Review of Political Science and International Relations*, 12, pp.3–20. Available at: https://web.archive.org/web/20180421015401id_/http://journal.ispri.ro/wp-content/uploads/2012/03/1Radavoi3-20.pdf

Rampazzo, F. *et al.* (2021) "A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom", *Demography*, 58(6), pp.2193–2218. Available at: https://doi.org/10.1215/00703370-9578562 .

Raymer, J. *et al.* (2013) "Integrated Modelling of European migration", *Journal of the American Statistical Association*, 108(503), pp.801–819. Available at: https://doi.org/10.1080/01621459.2013.789435.

Rees, P.H. (1980) "Multistate demographic accounts: measurement and estimation procedures" *Environment and Planning A: Economy and Space*, *12*(5), pp.499-531. Available at: https://journals-sagepub-com.eur.idm.oclc.org/doi/10.1068/a120499

Rédei, M.L. (2009) "Foreigners in Budapest", (13), pp.31–49. Available at: http://www.migrationinformation.org/Feature/display.cfm?ID=167.

Sales M.I. (2022) "Big Data at the Crossroads: Seizing the Potential of Big Data to Guide the Future of EU Migration Policy", *European Institute of the Mediterranean*, Available at: https://www.iemed.org/wp-content/uploads/2022/02/Policy-Brief-No116.pdf

Santamaria, C., Vespe, M. (2018) "Towards an EU Policy on Migration Data, *Publications Office of the European Union*, Luxembourg, Available at: https://policycommons.net/artifacts/2163259/towards-an-eu-policy-on-migration-data/2918795/

Schunck, R. and Perales, F. (2017) "Within-and between-cluster effects in generalized linear mixed models: A discussion of approaches and the xthybrid command" *The Stata Journal*, *17*(1), pp.89-115. Available at: https://journals.sagepub.com/doi/abs/10.1177/1536867X1701700106

Solano, G. and Huddleston, T. (2020) "Migrant Integration Policy Index", *Migration Policy Group*. Available at: www.mipex.eu.

Tu, W. and Liu, H. (2016) "Zero-Inflated Data", *Encyclopedia of environmetrics*, pp.1–7. Available at: https://doi.org/10.1002/9781118445112.stat07451.pub2.

UNDESA (2022a) "World Population Prospects 2022: Methodology of the United Nations population estimates and projections". Available at: https://population.un.org/wpp/Publications/Files/WPP2022_Methodology.pdf

UNDESA (2022b), World Population Prospects Database [online] *UNDESA*. Available at: https://population-un-org.eur.idm.oclc.org/wpp/Download/Standard/MostUsed/ [Accessed 25 Nov. 2022]

UNECE (2021) "International Migration: A Practical Guide.", *UNECE,* Available at: https://unece.org/fileadmin/DAM/stats/publications/International_Migration_Practical_Guide_ENG.pdf

V-Dem (2022), V-Dem Dataset [online]. Available at: https://www.v-dem.net/data/the-v-dem-dataset/country-year-v-dem-fullothers/ [Accessed 25 Nov. 2022]

Willekens, F. (2019) "Evidence-based monitoring of international migration flows in Europe", in *Journal of Official Statistics*, 35(1), pp.231–277. Available at: https://doi.org/10.2478/jos-2019-0011.

Wiśniowski, A., Zagheni, E. and Fava, E. del (2019) "Modelling international migration flows by integrating multiple data sources" Available at: https://doi.org/10.31235/osf.io/cma5h.

Wooldridge, J.M. (2012) "Introductory Econometrics: A Modern Approach",*Mason* , 5. Available at: https://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf

Wooldridge, J.M. (2021) "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators", *SSRN*, Available at: http://dx.doi.org.eur.idm.oclc.org/10.2139/ssrn.3906345

World Bank (2022a), World Development Indicators [online] *World Bank*. Available at: https://databank-worldbank-org.eur.idm.oclc.org/reports.aspx?source=2&series=NY.GDP.PCAP.KD&country=# [Accessed 25 Nov. 2022]

World Bank (2022b), Worldwide Governance Indicators [online] *World Bank*. Available at: https://info-worldbank-org.eur.idm.oclc.org/governance/wgi/Home/Reports [Accessed 25 Nov. 2022]

Appendix A: Methodological Justifications and Model Selection Tests

Appendix A.1

This section is focused on the justification for using the average of the two estimated flows to generate the undercount indicator (*ME*). For this purpose, we utilize the bilateral immigration flow database published by Abel and Cohen (2019). The flow database is derived from UN's migrant stock database which provides origin destination wise migrant stocks for each country in the world over a five year interval. The database contains flow estimates for all flow estimation techniques discussed in Chapter 5. To begin with, we restrict the global database to the nine CEECs and conduct a comparison of the mean estimated flow i.e. the average of stock difference drop negative and migration rates estimates with flow values generated using other techniques.

Appendix Table 1 and 2 present the average and total immigration flows disaggregated by estimation technique and time period. Appendix Table 1 indicates that the mean estimated flows on average are closest to the stock difference reverse negative and closed demographic accounting approaches as designed by Beine & Parsons (2015) and Abel & Sander (2014) respectively. Additionally, we make similar observations for total immigration flows as highlighted by Appendix Table 2. These approaches are considered to be more reliable estimation techniques in comparison to migration rates and stock difference drop negative based on validation exercises undertaken by Abel and Cohen (2019). They conduct the same by calculating Pearson correlations between various migration indicators based on the estimates and the equivalent reported values (For example, immigration rates, emigration rates and the number of migrants proportional to the total population) and conclude that the closed demographic accounting framework has highest correlation across the board making it the most reliable method of bilateral flow estimation. This is closely followed by the Pseudo Bayesian demographic accounting framework developed by Azose & Raftery (2019), Abel's (2010) open demographic accounting framework and Beine & Parsons' (2015) stock difference approach. Given the proximity in average and total immigration flow values and relatively higher reliability of Beine & Parsons and Abel & Sander's (2014) approaches, we assume that the mean estimated flows are a better approximation of the true flow values than the other available estimates

Appendix Table 1: Average Immigration Flows Disaggregated by Technique and Time Period

Period	Stock Difference Drop Negative	Stock Difference Reverse Negative	Migration Rates	Demographic Accounting - Open	Demographic Accounting - Closed	Demographic Accounting - Pseudo Bayesian	Mean Estimated Flows
1990-1995	101.366	157.6418	454.8878	119.445	166.137	484.995	278.127
1995-2000	132.229	193.8505	344.3407	155.207	199.374	542.539	238.285
2000-2005	151.165	243.1461	330.2679	172.408	182.197	589.764	240.717
2005-2010	165.413	250.7592	392.0085	193.865	275.257	760.853	278.711
2010-2015	198.9	307.3078	341.9322	218.632	260.664	817.252	270.416
2015-2020	461.334	526.76	259.4183	499.053	500.786	1111.43	360.376
Total	201.735	279.9109	353.8092	226.435	264.069	717.805	277.772

Source: Author's Elaboration based on Abel & Cohen (2019)

Appendix Table 2: Total Immigration Flows Disaggregated by Technique and Time Period

Period	Stock Difference Drop Negative	Stock Difference Reverse Negative	Migration Rates	Demographic Accounting - Open	Demographic Accounting - Closed	Demographic Accounting - Pseudo Bayesian	Mean Estimated Flows
1990-1995	19969.1	31055.44	89612.89	23530.7	32728.9	95544	54791
1995-2000	26049.1	38188.56	67835.11	30575.8	39276.7	106880	46942.1
2000-2005	29779.6	47899.78	65062.78	33964.3	35892.9	116183	47421.2
2005-2010	32586.3	49399.56	77225.67	38191.3	54225.7	149888	54906
2010-2015	39780	61461.56	68386.44	43726.4	52132.9	163450	54083.2
2015-2020	92266.8	105352	51883.67	99810.7	100157	222285	72075.2
Total	40071.8	55559.48	70001.09	44966.5	52402.4	142372	55036.5

Source: Author's Elaboration based on Abel & Cohen (2019)

Appendix A.2

The results from the Breusch Pagan Lagrange Multiplier Test are provided in Table Appendix Table 3. Based on the LM p value presented in the last row of the regression table, we must reject the null hypothesis that the random effects are zero. Thus, we opt for the random effects model over the pooled OLS approach.

Appendix Table 3: Estimation Results for Missingness with Random Effects and LM Test Results

	(1)
Variables:	Missing
R&D Expenditure (Log)	-.142
	(.117)
GDP Per Capita (Log)	-.067
	(.23)
Total Immigrant Flows (Log)	.011
	(.044)
Population (Log)	.057
	(.175)
Constant	.425
	(4.006)
Observations	112
R^2	.114
LM (chibar2(01))	123.55705
LMp (Prob > chibar2)	0

Note: Standard errors are in parentheses
*** p<.01, ** p<.05, * p<.1

The results of the Hausman Specification Test are presented in Appendix Table 4. Based on the P value, we fail to reject H_0 which states that the differences in the fixed effects and random effects estimates is not systematic. Thus, the random effects model is more appropriate.

Appendix Table 4: Hausman Specification Test Results for Analysis of Missingness

	(b) Fixed Effects	(B) Random Effects	(b - B) Difference	sqrt (diag(V_b -V_B)) Standard Errors
R&D Expenditure (Log)	-0.05306	-0.1416	0.088541	0.045618
GDP Per Capita (Log)	-0.08836	-0.06665	-0.02172	0.101368
Total Immigrant Flows (Log)	0.063736	0.010931	0.052805	0.025809
Population (Log)	2.363869	0.056694	2.307175	0.98032

Test of H_0: Difference in estimates is not systematic	
$chi^2(4)$	= (b-B)'[(V_b-V_B)^(-1)](bB)
	= 6.34
Prob > chi^2	= 0.1749

The results for the Wald Test for Joint Significance are provided in Appendix Table 5. The results indicate that there is no time related heterogeneity present in the data and as a result, we fail to reject the null hypothesis which states that the time effects are jointly equal to zero.

Appendix Table 5: Results for Wald's Test for Joint Significance of Time Effects

(1) 2008.year = 0
(2) 2009.year = 0
(3) 2010.year = 0
(4) 2011.year = 0
(5) 2012.year = 0
(6) 2013.year = 0
(7) 2014.year = 0
(8) 2015.year = 0
(9) 2016.year = 0
(10) 2017.year = 0
(11) 2018.year = 0
(12) 2019.year = 0
$chi^2(12)$ = 7.95
Prob > chi^2 = 0.7887

Appendix B: Descriptive Statistics

Appendix Table 6 highlights the differences between the year wise migration events across the reported flows and the two flow estimates. Based on the same, we observe that stock difference estimates report the lowest number of non-zero flow values, while migration rate estimates report the highest. The migration rate estimates, consistently report a higher number of migration events than the reported flows, with the gap between the two reaching it's peak in 2014, where migration rate estimates contained 255 more migration events than the reported flows.

Appendix Table 6: Number of non-zero flows (Migration events) by year

Year	Reported Flows	Stock Difference	Migration Rates
2007	352	261	413
2008	479	374	557
2009	472	371	579
2010	473	381	586
2011	450	372	567
2012	569	334	732
2013	700	541	886
2014	864	621	1119
2015	938	761	1144
2016	951	770	1134
2017	1059	811	1265
2018	1090	740	1291
2019	938	813	1154
Total	9335	7150	11427

***Note:** The number of non-zero values for the mean of the estimated flows are not reported as they are the same as the migration rate estimates.

Appendix Table 7: Descriptive Statistics

Variable	N	Mean	SD	Min	Max
ME	13519	.467	2.35	-1	98.086
IMME	13519	.598	.42	0	1.333
Missing	20088	.128	.334	0	1
Reported Flows	13519	120.198	829.591	0	39572
Total Immigrant Flow[18]	19232	38188.168	46424.711	1561	202422
Total Migrant Stock	19061	171601.24	138532.09	18682	557458
Mean Estimated Flows[19]	13519	131.499	866.738	0	28280.143
Migration Rate Flows	13519	201.235	1391.805	0	41489.285
Stock Difference Flows	13519	61.763	480.545	0	19798
Migrant Stocks	13519	867.282	5891.049	0	130933
Natural Population Change	13519	6.097	5.713	-6.557	20.02
Naturalization	17913	11.807	215.237	0	15658
Emigration	16289	37.32	291.59	0	15391
R&D Expenditure	20088	169.851	159.795	15.038	707.854
Public Sector Corruption Index	20088	.219	.137	.033	.556
Government Effectiveness Index	20088	.686	.418	-.36	1.19
Contiguity	20088	.026	.159	0	1
Schengen	20088	.14	.347	0	15391

[18] Total reported stocks refer to the total number of migrants moving to a specific country in a given year unlike migrant stocks and flows which are country pair and year specific

[19] Average of stock difference and migration rate flow estimates

Appendix C: Estimation Results

Appendix Table 8: Pre-Analysis Tables with Gradual Variable Addition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	ME	ME	ME	ME	IMME	IMME	IMME	IMME
Natural Population Change	-0.0835**	-0.0847**	-0.0795*	-0.0536	-0.00241	-0.00239	-0.000411	-0.00841
	(0.0414)	(0.0411)	(0.0424)	(0.0439)	(0.00947)	(0.00947)	(0.00939)	(0.00932)
Naturalization (Log)		0.152	0.151	0.144		-0.00279	-0.00334	-0.00129
		(0.160)	(0.159)	(0.160)		(0.0135)	(0.0131)	(0.0123)
Emigration (Log)			-0.131*	-0.231***			-0.0495***	-0.0186***
			(0.0723)	(0.0787)			(0.00584)	(0.00558)
Migrant Stock (Log)				0.607***				-0.187***
				(0.0850)				(0.0118)
Observations	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725
R-squared	0.383	0.383	0.384	0.392	0.542	0.542	0.546	0.570

Note: Robust standard errors are presented in parentheses and the coefficient for the constant and time effects are not reported

*** p<0.01, ** p<0.05, * p<0.

Appendix Table 9: Extended Hybrid Model Between Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variables	ME	ME	ME	ME	ME	IMME	IMME	IMME	IMME	IMME
Between Effects										
Natural Population Change	0.00860	0.00480	0.00464	0.00551	0.00353	0.00115	-0.000155	0.000303	0.000557	0.000280
	(0.00733)	(0.00784)	(0.00785)	(0.00785)	(0.00788)	(0.00174)	(0.00176)	(0.00177)	(0.00178)	(0.00177)
Naturalization (Log)	0.108	0.106	0.107	0.112	0.117	0.0356	0.0332	0.0302	0.0318	0.0329
	(0.205)	(0.206)	(0.205)	(0.205)	(0.201)	(0.0223)	(0.0223)	(0.0224)	(0.0222)	(0.0217)
Emigration (Log)	-0.474***	-0.470***	-0.470***	-0.461***	-0.456***	-0.0249***	-0.0234***	-0.0195**	-0.0172**	-0.0167**
	(0.0595)	(0.0595)	(0.0588)	(0.0587)	(0.0579)	(0.00758)	(0.00758)	(0.00778)	(0.00783)	(0.00778)
Migrant Stock (Log)	0.370***	0.356***	0.356***	0.354***	0.351***	0.00563	0.000789	-0.00106	-0.00167	-0.00202
	(0.0353)	(0.0362)	(0.0364)	(0.0364)	(0.0365)	(0.00638)	(0.00655)	(0.00658)	(0.00657)	(0.00661)
R&D Expenditure (Log)		0.280*	0.282*	0.233	-0.122		-0.0655*	-0.0992***	-0.125***	-0.169***
		(0.144)	(0.150)	(0.174)	(0.188)		(0.0342)	(0.0356)	(0.0408)	(0.0419)
Public Sector Corruption Index			-1.273	10.42***	16.29***			0.739***	3.136***	3.789***
			(1.562)	(2.275)	(2.442)			(0.220)	(0.671)	(0.702)
Public Sector Corruption Index^2				-18.16***	-17.80***				-3.846***	-3.528***
				(4.466)	(4.516)				(1.114)	(1.114)
Government Effectiveness Index					3.088***					0.366***
					(0.577)					(0.101)
Observations	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725	10,725

Note: Robust standard errors are presented in parentheses and the coefficient for the constant and time effects are not reported

*** p<0.01, ** p<0.05, * p<0.1