Romanian Statistical Review
Revista Română de Statistică

THE JOURNAL OF NATIONAL INSTITUTE OF STATISTICS

1/2022

www.revistadestatistica.ro

CONTENTS 1/2022

Online Job Advertisements for Labour Market Statistics using R

Andrea ASCHERI[1]
Eurostat, European Commission, Luxembourg

Gabriele MARCONI
Sogeti, Luxembourg

Matyas MESZAROS
Eurostat, European Commission, Luxembourg

Fernando REIS
Eurostat, European Commission, Luxembourg

ABSTRACT

This paper introduces the implementation through R of the methodology used to calculate a labour market concentration (Herfindahl-Hirschman) index for European urban areas, based on a database of over 100 million online job advertisements. After introducing the broader context and the motivation for the analysis, the authors describe the overall processing workflow. In addition, the paper presents in more detail the solutions provided to two main challenges encountered: addressing computational efficiency by using parallel computing and cloud data querying; and a custom-built machine learning model to classify an important variable for the study (company name). Finally, the paper discusses the main rationales for using R and for sharing the code in a public repository.

Keywords*: R, Big data, Online Job Advertisements, Labour market*

INTRODUCTION

The last years have seen a steep increase in the use of web data as new and non-traditional data source to complement official statistics. The Web Intelligence Hub is an initiative led by Eurostat and aimed at fostering cooperation and building the necessary capabilities in the European Statistical System (ESS) to collect and process web data to derive meaningful information for statistical purposes.

1. Corresponding autor: andrea.ascheri@ec.europa.eu

The experimental work presented in this paper stems from the activities performed by various statistical offices of the ESS within the recent ESSnet Big Data II project (ESS, 2020). The objective of this paper, after introducing the topic, is to present more in detail the R code behind this experimentation and to describe the overall workflow as well as a few solutions implemented. In particular, this paper focuses on a two-step algorithm proposed to process the company names, which are critical for calculating the competition between employers in the hiring market.

Section 1 *Background and motivation* presents the rationale behind this experimental study and the importance of the data source used. Section 2 *Methodology* describes in detail the dataset and the applied methodology. This section focuses on the main issues encountered and their solutions, including the problem of dealing with big data and the dedicated text processing algorithm to deal with company names. Section 3 *Results* briefly presents the experimental results of this study. Finally, Section 4 *Conclusions* sums up the main achievements and lessons learned of the proposed approach.

1. BACKGROUND AND MOTIVATION

1.1. Big data and open sourcing

The use of the internet for publishing job advertisements has increased in recent times, spurred by the improvement of network connections across Europe and by increased information and communication technology literacy. As a result, public interest in using these data for labour market analysis has increased at the international level (Cedefop; European Commission; ETF; ILO; OECD and UNESCO, 2021). Online Job Advertisements (OJAs) can offer high degrees of timeliness (if harvested regularly), relevance (e.g. they are a unique source of information on skill requirements) and granularity (e.g. detailed location data), providing them with a great potential to complement official statistics (Cedefop, 2019; ESS, 2020). Altogether, these projects succeeded in developing methods to collect OJAs on a regular basis in all European Union Member States. Eurostat and Cedefop joined forces to augment the most advanced of their OJA data collection systems and start producing more indicators at the European and national level (Descy, Kvetan, Wirthmann, & Reis, 2019).

The result of this is a new, publicly available dataset on OJAs covering the whole European Union. The current production system is still being developed and improved, with the important goals of increasing the availability of final and intermediate data, and of making the underlying algorithms open-source. In addition, the European Statistical System Committee (ESSC)

discussed the principles of Trusted Smart Statistics (TSS) and priority areas for producing European statistics from new data sources (Eurostat, 2019; Ricciato, Wirthmann, Giannakouris, Reis, & Skaliotis, 2019). This includes the creation of a Web Intelligence Hub (WIH) that collects data from the web and make them publicly available to enhance statistical information in various domains.

As it is already clear since a decade, open source architecture will expand in the future and will be used more and more by Statistical Offices in Europe and elsewhere (Beat Hulliger, 2012). This is in line with one of the Trusted Smart Statistics principles (Ricciato, Wirthmann, Giannakouris, Reis, & Skaliotis, 2019): Statistical Offices should systematically ensure that all the software code along the whole data processing chain are made fully open and auditable by default for stepping up transparency and accountability. For this purpose, R as a free Open Source software product (R Core Team, 2021), is an ideal environment for sharing, collaborating and automatizing the production of official statistics. R has become a 'lingua franca' for statisticians, methodologists and data scientists worldwide thanks to an active worldwide community of users and its vast amount of functionalities for data preparation, methodology, visualisation and application building. To achieve this, the integral code behind the experimental work presented in this paper is shared on an open GitHub repository (Eurostat GitHub, 2021).

1.2. Motivation for the analysis

As the first statistical application of the new database, we chose to analyse labour market concentration across all occupations and functional urban areas (FUAs) in the EU for the biennium 2019-2020. This choice was motivated by its policy relevance and by its potential to build upon new statistical developments and complement existing statistics.

In advanced economies, cities play a fundamental role in job creation and the enhancement of innovation (OECD & European Commission, 2020). They also account for a large part of existing employment: 72% of employed people in the European Union are found in predominantly urban regions (Ascheri, et al., 2021). Urban labour markets, and particularly large urban conglomerates, can provide workers with opportunities for labour specialisation and learning through knowledge spill over (Gordon & Turok, 2005). For this to happen, however, the labour market must be "thick", meaning that sufficient job options are available for workers in their occupations and within a reasonable distance (Brown & Scott, 2012). In contrast, "thin" labour markets with limited job alternatives within commuting distance give firms more market power, with negative effects on wages and working conditions

(Manning, 2003; OECD, 2020). Therefore, the degree of concentration (or thinness) in labour market hiring is of great policy relevance.

Despite its policy relevance, measuring labour market hiring concentration has until recently remained outside the scope of traditional statistics, due to the detailed information needed on job offers (location, occupation, and employer) and commuting options. An opportunity for change has been offered by two concurrent developments in international statistics:

- the availability of granular and detailed web-scraped data on OJAs;
- the development of new international definitions of geographic areas based on commuting distance (OECD, 2012).

The availability of detailed location identifiers in OJA data makes it possible to aggregate OJAs based on their commuting areas (functional urban areas, or FUAs). Measures of demand-side competition in each urban labour market can therefore be generated, by using additional information on company name and occupation contained in OJAs. Following Azar, Marinescu, Steinbaum and Taska (2020), who first applied this approach to the US, the HHI concentration index can be calculated as a measure of thinness of the urban labour market. A higher HHI level corresponds to a more concentrated demand in the labour market, implying that fewer companies compete for recruiting workers or (in other words) that the labour market is thinner.

A more detailed discussion of the rationales for the analysis is given by Ascheri et al. (2021). In the same paper, the authors explain the underlying statistical methodology and present the results of the analysis. This paper focuses on the implementation of the analysis through R (even though a very brief summary of the results is contained in Section 3).

2. METHODOLOGY

2.1. **Dataset**

The dataset used for the calculation of the labour market concentration index is composed by 116 851 363 distinct online jobs ads collected in all 27 EU countries from 316 distinct sources. Although data for the UK are also available in the dataset, they have not been used for this analysis. The data sources from which OJAs are scraped have been identified by statisticians and labour market experts from all the countries involved in the data collection, through an agreed procedure described by Cedefop (2019). These sources are usually either job search engines (e.g. Indeed, Monster) or public employment services' websites listing job advertisements.

OJAs refer to advertisements published on the World Wide Web revealing an employer's interest in recruiting workers with certain

characteristics for performing certain work. It is important to note that this does not coincide with vacancies, as employers can publish job ads for various reasons (e.g. to fill a current vacancy, but also to explore potential recruitment opportunities).

Currently, new OJA data are released each quarter. The analysis of labour market concentration is based on the most recent dataset version (v9) available at the time at which it was carried out. That version of the OJA dataset contains data from the third quarter of 2018 to the end of 2020. Given the lower coverage across countries and sources of 2018 data, only data from 2019 and 2020 have been used.

OJAs usually include information on job characteristics (e.g. location, occupation), employer characteristics (e.g. name, economic sector) and job requirements (e.g. required skills). Most of the variables contained in the dataset are categorical, and have been cleaned and classified following international classifications (e.g. the ISCED classification for education, or the ESCO classification for occupations). Nonetheless, some pieces of information (in particular the company name, a crucial variable for this analysis) are only available as unstructured data (natural language text), increasing the burden on the analyst.

The large size and the presence of some unstructured data are characteristics of the OJA dataset that can be found in other sources of "big data" (UN Statistical Commission, 2013). These characteristics mean that big data presents not only important opportunities, but also serious challenges for their applications in official statistics (Yongdai, 2013; Hackl, 2016; Tam, 2015). Fortunately, more and more computing and processing resources are becoming available that help dealing with the complex challenges posed by this type of data. More research on the utilisation of these resources, and more sharing of good practices, are needed to increase the confidence in using these new data sources (UN Statistical Commission, 2013). By describing the implementation through R of the main processes leading from data querying until the production of the statistical indicator, this paper contributes to this stream of research and to the advancement of the use of big data in international statistics.

2.2. Implementation of the analysis in R: General structure

The R code implementing the methodology, based on the original code developed by De Lazzer (2020), is written in R and can be divided in the following macro steps, also shown in Figure 1. This code is openly shared on GitHub (Eurostat GitHub, 2021). The main code is contained in one single script *lmci_v1.R*, which uses several custom functions defined in a separate

file *hhi_functions.R*. Other files contain code used for particular phases of the analysis. For example, *Other%20scripts/globalmodelevaluation_sampling.R* contains the codes running the sampling for the machine learning model evaluation, while other scripts and files related to this evaluation are in the folder *Other%20scripts/imputation_model_evaluation*.

Macro-steps of the R code

Figure 1

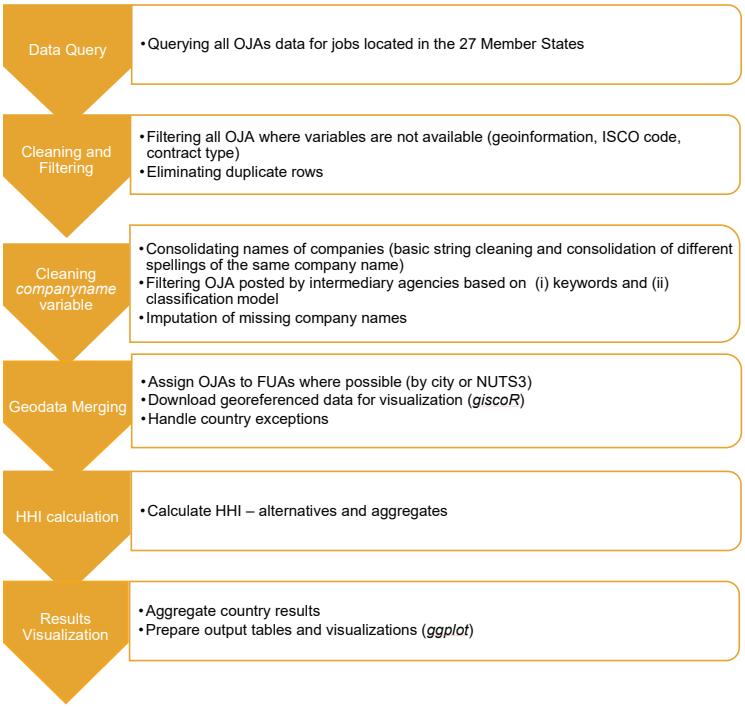


Figure 1 shows the main steps that we followed to produce the results:

- **Data Querying**: the code queries the online job advertisements from the dataset. The queries are run in parallel for all the 27 EU Member States. The OJAs corresponding to each country are saved in various *.Rds* files. The function *get_data()* uses the *noctua* (Jones, 2021) and DBI package (R-SIG-DB R Special Interest Group on Databases, Wickham, & Müller, 2021) to connect to the AWS Athena database and to query the job ads.

```
get_data <-   function(query){
                con <- DBI::dbConnect(noctua::athena(),
                                      s3_staging_dir=readLines("~/.aws/s3_staging_dir"),
                                      work_group=readLines("~/.aws/work_group")
                )
                my_data <- noctua::dbGetQuery(con, query)
                dbDisconnect(con)
                return(my_data)
              }
```

The code snippet below illustrates an example of a query from the dataset, with all the main retrieved variables of a job advertisement, where the version of the data (data_table) and the country (countrycode) are two parameters for the query.

```
query <-   paste0("SELECT general_id, grab_date, lang, idesco_level_4,
           esco_level_4,  idcity, city, idprovince, province, idregion, region, idcountry,
           country, idcontract, contract, idsector, sector, sourcecountry, source, site,
           companyname  ", "FROM ", data_table, " ","WHERE idcountry = '", countrycode,"' AND
           idprovince != '' AND contract != 'Internship' AND grab_date > 17896 ", ";")
data <- get_data(query)
saveRDS(data,filename)
```

- **Data Cleaning and Filtering:** OJAs are filtered to eliminate the advertisements that cannot be use for computing the index. The excluded ads do not contain information on the location of the job (information on occupation is always present), or they advertise explicitly a stage/internship.
- **Cleaning the variable *companyname*:** this part of the process deals with one specific variable of the dataset that is crucial to the computation of the index: the name of the company advertising the job. The string variable *companyname* contains the name as it is extracted from the ad. Therefore, it is possible to find multiple denominations of the same company (e.g. “XYZ”, “XYZ s.a.r.l.”, “XYZ company”) or to find characters, spaces or cases in strings referring to the same company. Moreover, local branches of franchisees of the same company might post ads under their own names.

To tackle this issue we applied some basic string cleaning operations (e.g. convert to lower case, delete punctuations, symbols and white spaces).

```
companynames_sep<-unlist(
       parallel::mclapply(tolower(dframe$companyname),sep2,mc.cores=hhi_cores))
 dframe[,companyname:=trimws(gsub(" ","_",ascii(companynames_sep)))]
```

Moreover we created a dictionary of company names where several variants of the names are listed together with the "master" version that will replace the different spelling modes. The list has a generic part which is applied for all countries (`clean_names$country=="EU"`) and a part which can be specific for each country (`clean_names$country==countrycode`). The generic part of the dictionary (*companies_to_clean_EU.csv*) contains the companies with the largest number of ads and it can be incrementally improved with smaller company names or with national companies, as it has been done for the country-specific dictionaries for Italy, Portugal, Romania and Slovenia. A consolidation of the company names (i.e., replacing name variants with the master version) in the data frame is applied using this dictionary.

```
clean_names <- read.csv("companies_to_clean_EU.csv" , sep = ",")
clean_names<- clean_names[clean_names$country=="EU"|clean_names$country==countrycode , ]
dframe_names<-
data.table(rn=dframe[companyname!="",which=T],dframe[companyname!="",c("companyname")])
  f_clean_names<-function(cl,dframe) { dframe[(grepl(cl[[1]][3],companyname) &
companyname!=cl[[1]][5]) |companyname==cl[[1]][4] ,companyname:=cl[[1]][2]][]
    }
all<-
rbindlist(unique(lapply(as.list(as.data.frame(t(clean_names))),f_clean_names,dframe=dframe
_names)))
  dframe[all$rn,companyname:=all$companyname]
```

In addition, the nature of the variable *companyname* poses a series of challenges related to the type of company posting the ads that are described in detail in Section 2.4.

- **Merging Geodata:** The index is computed at urban level, taking Functional Urban Areas (FUAs) as the main unit of analysis. However, the location of the job extracted from the advertisements is classified using the LAU (Local Administrative Unit) code (for municipalities) and NUTS classification. Hence, this section of the code assigns the online job ads to a specific urban area based on the information contain in the entry and downloading a correspondence table between LAUs, NUTS and FUAs from the Eurostat website and using the *giscoR* package (Hernangómez, 2021). Assigning an advertisement to an urban area is always possible when the LAU of the job offer is known. If this is not the case, the matching can be done

using the NUTS 3 code in cases where NUTS 3 areas are equivalent to FUAs. When ads cannot be matched to a FUA, they are excluded from the dataset. Due to some inconsistencies between the LAU code used in the OJA dataset, some country specific adjustments have been implemented by various string manipulations.

```
filename <- "EU-28-LAU-2018-NUTS-2016.xlsx"
  if(!file.exists(filename))
  {download.file("https://ec.europa.eu/eurostat/documents/345175/501971/EU-28-LAU-2018-
  NUTS-2016.xlsx", destfile=filename)}
geoinfo <- giscoR::gisco_get_nuts(year = 2016,epsg = 3035, nuts_level = 0, country =
countrycode,spatialtype = "RG", resolution = "01")
sfile <-giscoR::gisco_get_urban_audit(year = 2020,epsg = 3035,country = countrycode, level
= "FUA", spatialtype = "RG", update_cache = TRUE)
```

- **Computing the Index:** in this part, the Herfindahl-Hirschman Index (HHI) is computed. The HHI is a commonly accepted measure of market concentration calculated as the sum of the squares of the market shares of each firm competing in a market. The HHI ranges from close to 0 under perfect competition to 10 000 in monopoly/ monopsony (i.e., 100% market share). The lower the index, the more competitive (or less monopolistic) the market is. The code snippet below highlights the function developed for the calculation of the index.

```
calculate_hhi <- function (dframe,cores=2) {
  grid <- expand.grid(esco = unique(dframe$idesco_level_4), geo = unique(dframe$fua_id),
qtr = unique(dframe$qtr), stringsAsFactors = FALSE)
…
  f_calc_hhi<-function(gr,subset){
    subset <- unique(dframe[idesco_level_4 == gr[1] & fua_id == gr[2] & qtr == gr[3],
c("idesco_level_4", "fua_id", "qtr", "mshare", "ms2", "companyname", "ncount"), with =
FALSE])
    subset$hhi <- sum(subset$ms2)
    subset[1, !c("companyname") ]
   }
  hhi<-
rbindlist(parallel::mclapply(as.list(as.data.frame(t(grid))),f_calc_hhi,subset=dframe,mc.c
ores=cores))

  return (hhi)
}
```

- **Results Aggregation and Visualisation:** The results are aggregated and visualisations are produced using the *ggplot2* package (Wickham, et al., 2019) for all the quarters of the analysis. An example of the map generated with the following lines of code is shown in the results section (Figure 3).

```
hhigeo_plot_tot<-function(qrtr,hhigeo_q,geoinfo,resultspath){
ggplot(eval(parse(text=paste0("hhigeo_q$`",qrtr,"`")))) +
  geom_sf( aes(fill = mean),lwd=0) + theme_void() +
  theme(panel.grid.major = element_line(colour = "transparent")) +
  labs(title = paste("Labour market concentration index", qrtr,"\naverage over all
occupations")) +
  scale_fill_continuous(name = "Labour market concentration index",low="blue",
high="orange") +
  #geom_sf_text(aes(label = label), size = 2.5, colour = "black")+
  geom_sf(data=geoinfoTOT,alpha = 0)+
  coord_sf(xlim = c(2700000, 5850000),ylim = c(1390000, 5400000)) + theme_bw()
  #including cyprus coord_sf(xlim = c(2700000, 7050000),ylim = c(1390000, 5400000)) +
theme_bw()
  ggsave(paste0(EU_resultspath,"/HHI_",qrtr,".png"), width = 15, height = 10, units =
"cm")
}
```

The following sections will deal in detail with two topics of particular interest: i) the use of R for dealing with large amounts of data and the efficiency issues thereof and ii) the algorithm developed to classify the company names extracted from the job advertisement between employers and non-employers.

2.3. Efficiency issues

One of the most evident advantages of R is the possibility to work seamlessly with different kinds of sources. The R code developed for this study merges elements from AWS Athena and geodata from the Eurostat GISCO services (GISCO, 2021) and MS Excel files.

Overall, the starting dataset contained more than 100 million records with 12 variables. To handle this amount of date in an acceptable time, first the extensive use of base R functions and secondly the use of packages based on C/C++ are sensible choices. As an example, for the first case *grepl/gsub* is more efficient compared to the functions of the *stringr* package. For the second case the *data.table* package that is based on C++, and is faster to merge and aggregate large datasets as *data.table* as it uses less memory and parallel processing in the background than the base R *data.frame*.

One of the challenges of this implementation was to select the best method to query the AWS Athena® database. As a first option, the *RJDBC* package was chosen. Subsequently, due to the increased security requirements which needed a rotating access token, the *RAthena* package was used. Because of some compatibility issues with the underlying python miniconda environment and the standard python installation used in Jupyter notebooks in the final version of the code, to overcome these issues the *noctua* package is used.

The *noctua* package uses AWS API that allows scripting authentication, which was used to retrieve the access token automatically without user intervention and without including any sensitive information in the code.

```
con <- DBI::dbConnect(noctua::athena(),
                      s3_staging_dir=readLines("~/.aws/s3_staging_dir"),
                      work_group=readLines("~/.aws/work_group")
)
```

Finally, the code makes ample use of parallel computation with the *mclapply()* function. Parallel computing is used in different phases of the code: from the data query by country, to the computation of the index up until the creation of the plots for each quarter. Several attempts were made at finding the optimal number of cores to speed up as much as possible the processing time while avoiding memory overload. Finally, the number of cores was set in two steps. First the general numbers of countries to process parallel were set to 3 due to the limitation to concurrent connection to AWS Athena®. In the second step, for the calculation of the HHI, a new set of cores were used which in the final version were 5. Therefore, up to 15 cores were used in total at the same time, which allowed reducing the overall processing time, which now takes around 3 hours to run for all the 27 countries.

2.4. Classification of company names through machine learning

2.4.1. Rationale

The labels (hereafter, "names") contained in the variable "companyname" are used to calculate the ad shares that define the index. However, this is not a clean variable: it contains whatever was extracted from the field "company" or "employer" or "name" (or similar fields, depending on the exact configuration of the data source) on the website. This could correspond to the name of the prospective employer seeking to recruit the worker(s), recruiting agencies, job portals or generic strings. Some examples are drawn in Table 1 for Italy, Portugal, Romania and Slovenia, countries for which the variable "companyname" has been analysed more in depth to drive subsequent data processing.

Most common names by category

With number of ads, found in a sample of one million, in brackets

Table 1

		Italy	Romania	Portugal	Slovenia
Most common:	Job portal	Jobtome (34964)	Tiptopjob.com, 477	Tiptopjob.com (619)	M servis
	Agency	Manpower (19163)	Adecco (2181)	Randstad (10444)	Adecco
	Employer	Axcent (1040)	Secretariatul General Guvernului (9381)	BNP Paribas (2672)	Mercator
	Generic string	Azienda (3230)	Professional (470)	Anonimo (165)	NA

Note: For Slovenia, the numbers are not reported because the sampling was done at a later stage and with a different sampling scheme. No generic string were found among the 1000 Slovenian company names, sampled with probabilities roughly proportional to their number of ads.

For the calculation of the HHI index, it is necessary to identify which names belong to actual employers and which ones do not. Once known, names that do not belong to employers can be treated statistically (Ascheri et al., 2021, use this information to calculate upper and lower bounds for the index). Therefore, a strategy was implemented through R to identify names that do not belong to prospective employers. Besides the querying functions described in Section 2, this strategy uses functions from Base R and from the Tidyverse package (Wickham, et al., 2019).

2.4.2. Main challenge in identifying non-employer names

The proportion of job ads coming from non-employer names changes substantially across countries. For example, based on the procedure described in this section, we estimated it at 70% for Italy and 11% for Romania. This makes it very important, for the purpose of comparability, to identify non-employer names.

The main challenge to overcome, in choosing a procedure for an automatic identification of non-employer names, has been the lack of a proper training set. Only limited resources were available for human coding, which were employed to code manually names with a relatively large number of ads (because their impact on the estimation is larger than for names with smaller number of ads) for the three pilot countries (Italy, Portugal and Romania). Some input was also provided by the German and Slovenian statistical offices for their countries. For names with fewer ads and all other countries covered in the dataset, no human-coded training data set was available.

The need to overcome the lack of a proper training dataset led to a two-stage model for the classification of names. This model is composed of an ontology model that classifies companies based on a set of keywords; and of a machine learning model for names that have not been classified based on the ontology. This approach is described in detail in the following sections, and summarized in the list below:

1. Manually code a sample of names as either employer or non-employer in some pilot countries.
2. Drafting keywords (e.g. "human resources", "jobs", "adecco") to identify non-employer names in these countries.
3. Applying the keywords of Step 2 to identify non-employer names in other portions of the data. It appears that there are many recurrent patterns in the names of recruiting agencies and job portals, so this step worked well (see model evaluations in the following sub-sections). As a result, a set of non-employer names was identified among smaller and larger names in every country.
4. Identifying through visual inspection data patterns allowing to discriminate between employers and non-employers in the pilot countries based on the data gathered in Steps 1 and 2. Each pattern was then captured through an R function that flags a name as non-employer if it is very similar to other non-employers in terms of some observed relationships/characteristics.
5. Applying the functions of Step 4 to discriminate between employers and non-employers in other countries. In order to do so, these functions needed to be re-parameterised for each country. That is why these functions were only made dependent on characteristics of non-employers (note that a set of non-employer names was identified for each country in Step 3).

2.4.3. Identifying non-employer names in three pilot countries through human coding

To identify the set of keywords, we first drew a random sample of 1 million job ads from the dataset for three pilot countries (Italy, Portugal and Romania) through the get_data() function seen in Section 2.

```
companies_names_query <- query_athena("SELECT companyname, general_id
FROMestat_dsl2531b_oja.ft_document_en_v9WHEREidcountry='",country,"'
ORDER BY RAND() LIMIT 1000000")
```

From this sample of ads, we identified the sample of names to code manually. First, we extracted the list of names appearing at least 100 times in

the sample. To get an idea of the size of this sample, this yielded 319 company names for Romania, 254 for Italy and 117 for Portugal.

```
companies_names_dataframe <- as.data.frame(table(companies_names_
query$companyname))
write.csv2(companies_names_dataframe[companies_names_
dataframe$Freq>99,], "Data/companies_list_atleast100ads.csv")
```

In addition, we added to this a sample of 160 company names per country appearing between 20 and 99 times, stratified by the number of ads.

```
subsample <- companies_names_dataframe[companies_names_
dataframe$Freq<100 & companies_names_dataframe$Freq>19 , ]
subsample1 <- subsample[duplicated(subsample$Freq)==FALSE , ]
subsample <- subsample[duplicated(subsample$Freq)==TRUE , ]
subsample2 <- subsample[duplicated(subsample$Freq)==FALSE , ]
subsample <- merge(subsample1 , subsample2, all.x=TRUE)
write.csv2(subsample, "Data/subsample_companynames.csv")
```

We looked at these names one by one, and determined (usually based on an internet search) if they were names of prospective employers or other names (recruiting agencies, smaller job portals or generic words like "confidential"). We identified a name as belonging to a recruiting agency when, from our internet search, it appeared that the primary activity of the organisation with that name are services for the recruitment of personnel on behalf of other companies.

We compiled a "blacklist" of English-language keywords likely to be included in the names of recruiting agencies, such as "manpower", "personnel" and "hire". We added this to a list of keywords developed for German names by the German National Statistical Institute, and saved it on a csv-format file. This csv file was imported in the R code and used to "filter out" non-employers.

```
staff_agencies <- read.csv("staff_agencies_EU.csv" , sep = ",")
   staff_agencies <- staff_agencies[staff_agencies$country=="EU"|staff_
agencies$country==countrycode , ]

  blacklist <- staff_agencies[staff_agencies$exact != "exact" , 2]
  blacklist_exact <- staff_agencies[staff_agencies$exact == "exact" , 2]
filteredout <- filter(dframe, str_detect(dframe$companyname,
paste(blacklist, collapse = '|')) | sub(paste(blacklist_exact,
collapse = '|'),"",dframe$companyname) == "" )
```

To give an idea of the extent of this filtering process, the first completed round of filtering (without using automatic flagging) yielded a list of 218, 256

and 95 keywords (for Italy, Portugal and Romania, respectively) to identify names different from prospective employers (almost all keywords referred to job agencies). These, together with 89 keywords provided by the German Statistical Institute, constituted the blacklist used as a filter for the dataset. This blacklist was then applied to filter out observations from the whole data, leading to 9232 (Italy), 3444 (Portugal) and 1847 (Romania) names that were identified as non-employers. This compares to 69878 (Italy), 43151 (Portugal) and 49532 (Romania) names that were not identified as agencies based on the blacklist.

2.4.4. Automatic classification of non-employer names

Three discrimination rules were defined to automatically classify names that had not been already identified as non-employers through the keyword matching described in the previous section. Each of these three rules is based on an Ordinary Least Squares (OLS) estimate of the relationship between a dependent variable and a set of independent variables, carried out for each country based on the sample of non-employer names identified through the keyword matching. The OLS estimate is carried out twice, with the half of the observations with the worse fit excluded in the second estimation. Each rule flags as non-employers all the names that lie sufficiently far away from the estimated curve, i.e. those names for which the absolute value of the difference between their dependent variable's fitted and observed values exceeded 1.96 times the standard error of the fitted value.

The discrimination rules, based on regression parameters estimated for each single country, were applied to names that had at least 20 ads in the 1-million ads sample (because the indicators used at the name level are aggregate indicators and are therefore less robust for small numbers of ads). Only if a name was flagged as a non-employer name by all three rules, we categorized that name as non-employer. Therefore, the resulting model is a decision-tree machine learning model where a name is classified as non-employer if it is flagged as such by all three regression-based rules.

Figure 2 exemplifies how visual inspection helped identifying the empirical rules to distinguish between employers and non-employers. It has been generated as follows:

```
plotdata <- sumstats_by_company[sumstats_by_company$tot_n>15 &
sumstats_by_company$filteredout != -1, ]
ggplot(data = plotdata) +
    geom_point(mapping = aes(x = ln_undup_n, y = ln_esco3,
colour=filteredout))
```

Figure 2 represents the relationship between the log of the total (unduplicated) number of job ads posted under a certain name (horizontal axis) and the log of the number of distinct job occupation codes observed in the job ads posted under that name (vertical axis). This chart has been drawn based on a sample of 1000000 job ads in Italy, Portugal and Romania. Names identified as non-employers are in light blue, while names identified as prospective employers are in dark blue. The chart shows that names identified as agencies tend to display a larger number of distinct job occupations per job ad than other names, a pattern used in Rule 1.

Relationship between number of distinct job occupation codes and (unduplicated) job ads (in logs) across names in Italy, Romania and Portugal (2018 to 2020)

Figure 2

Note: "non-employer" stands for agency names, job portals and generic strings.

Based on the visual inspection and on the assessment on the Gini impurity index for various discrimination rules (see Table 2 below), three empirical rules have been set up for the automatic classification of names. Each rule is based on the estimation for each country of a curve by linear regression, based on names already coded as non-employers. The regressions have been run in two stages, with the half of the observations with the worse fit excluded in the second stage. The following relationships have been estimated:

- Rule 1: flag name as non-employer if it does not lie significantly (at least 1.96 standard deviations) below the curve represented in Figure 2 (log distinct occupational categories as function of log unduplicated ads), estimated through a quartic polynomial function
- Rule 2: flag name as non-employer if it does not lie significantly below the curve relating log total ads (including duplicates which are recorded multiple times in the data set) with log unduplicated ads

- Rule 3: flag name as non-employer if it does not lie significantly below the curve relating the log number of distinct NUTS3 regions for a name's ads with the log of distinct economic activities (2-digits NACE) and the log unduplicated job ads.

The automatic classification based on the machine learning model is implemented through two custom-made functions available on the project's Github repository and called automflag() and automflag_combine().

```
testflag1 <- automflag(mydata=sumstats_by_company[sumstats_by_company$ln_undup_n>3,],
       xvar2="sqln_undup_n", xvar3="culn_undup_n", xvar4="quln_undup_n")
testflag2 <- automflag(mydata= sumstats_by_company[sumstats_by_company$ln_undup_n>3 ,],
       yvar="ln_n", xvar1="ln_undup_n", xvar2="sqln_undup_n", flag_above=FALSE,
       flag_below=TRUE)
testflag3 <- automflag(mydata= sumstats_by_company[sumstats_by_company$ln_undup_n>3,
       ],yvar="ln_sector", xvar1="ln_prov", xvar2="ln_undup_n", xvar3="ln_undup_prov",
       flag_above=TRUE, flag_below=FALSE)
automflag_output <- automflag_combine(mydata=
       sumstats_by_company[sumstats_by_company$ln_undup_n>3, ], automflag1= testflag1,
       automflag2= testflag2 )
automflag_output <- automflag_combine(mydata=
       sumstats_by_company[sumstats_by_company$ln_undup_n>3, ], automflag1=
       automflag_output, automflag2= testflag3 )
filterlist <- c(filterlist,as.character(automflag_output[[5]]))
```

Table 2 reports the Gini impurity index by rule and pilot country. To calculate the index, each rule has been used to split the whole training dataset.

Gini impurity index of decision tree rules, by country

Table 2

	Rule 1	Rule 2	Rule 3	Average
Italy	0.268	0.287	0.29	0.282
Portugal	0.163	0.199	0.195	0.186
Romania	0.218	0.233	0.208	0.22
Average	0.216	0.24	0.231	0.229

Classifying company names: The example of France

In the following paragraphs, we describe the application of the classification procedure for the case of France. This country belongs to the majority of countries for which no human coding was available for a training data set.

After matching the set of non-employer keywords (developed through the human coding of data for Italy, Germany, Portugal and Romania), a total of 15352 names were identified as non-employers in France. These included, for example, "bennett_game_ltd_recruitment", "besancon_randstad", "parisjob" and "pro_talento" (matching, respectively, the keywords "recruit", "randstad", "job" and "talent"). A new name-level dataset was created for these names,

with one row per name and a set of aggregate variables such as the number of jobs advertised under each name and the number of distinct occupations or NUTS3 regions for jobs advertised under this name.

The three OLS regressions underlying the discrimination rules were fit using this name-level dataset. This yielded three fitted curves to which all other French company names could be compared. An additional 650 names were found that lied sufficiently close (according to the criterion defined by our model and reported above) to each of the three curves. These names, including for example "abalone" and "emploi_team", were then added to the set of names classified as non-employers. For example, these two names were flagged as a non-employer by the first rule because their number of advertised occupational categories were 53 and 48, respectively. This is not too far from the numbers predicted by the trained model (38 and 36, as predicted by their total of 388 and 333 unduplicated ads), and it is suspiciously large for a regular employer of that size.

2.4.5. Evaluation of the model for all countries in the OJA dataset

To evaluate the performance of the model outside the three pilot countries, a sample of 400 names has been drawn using the function sample() from all other countries in the OJA dataset. The probability for each name to be included in the sample has been set proportionally to its number of ads.[1] This has yielded a diverse sample including names in 18 different countries, which number of ads ranges from 1 to 746572 (median: 1101).

```
sample1 <- as.data.frame(sample(samplingframe$id, size=200, prob=samplingframe$N))
sample2 <- as.data.frame(sample(samplingframe$id, size=200,))
```

Two confusion matrices (see Table 3) have been developed to evaluate separately the fit of the first part of the model (ontology) and the overall model (ontology plus machine learning). Table 3 includes the proportions based on the numbers of (in)correctly classified names in the sample, as well as those weighted by the number of ads of each name. The weighted matrix allows calculating performance metrics based on the ads (instead of the names) in the sample.

1. As noticed by Stubner (2018), there is some evidence that R base function sample() introduces a bias in the calculation of the sampling probabilities. The function dqsample(), available via Drat (Eddelbuettel, 2021), provides an alternative that should be free of bias. However, this function does not support weighted sampling as of 28 February 2022, so it was not possible to use it to draw a new sample for comparison purposes. Therefore, while our sample remains a random sample valid to carry out the model evaluation, our estimates of the accuracy rate may carry a bias, which direction is difficult to establish a priori.

Confusion matrix for the assessment of classification model

Table 3

	Ontology model				Overall model (including ontology and decision tree)			
	True positives	*True negatives*	*False positives*	*False negatives*	*True positives*	*True negatives*	*False positives*	*False negatives*
Un-weighted (name level)	27.0%	44.5%	0.0%	28.5%	32.0%	40.0%	4.5%	23.5%
Weighted (ad level)	60.3%	17.8%	0.0%	22.0%	64.0%	17.4%	0.4%	18.2%

Notes: Positive stands for non-employer. The use of the base R function sample() may induce a bias in the weighted sampling probabilities for each companyname.

Overall, we considered the model performance satisfactory. The accuracy rate (i.e. the proportion of cases correctly classified as either employers or non-employers – true positives plus true neg atives) is 72% for the un-weighted matrix, and 81% when weighting by the number of ads under each name. This implies that 81% of ads and 72% of names in the sample are correctly classified. The proportion of false positives is below 5% for names, and almost negligible (0.4%) for ads.

A striking feature of the ontology model is that it has not led to any false positives in this evaluation sample. This is a sign that the keywords work well, but also that the ontology model on its own may be too conservative an approach (some error is tolerable if it is compensated by an increase in the number of correctly classified cases). Applying the machine learning model after the ontology increases the accuracy rate (weighted by the number of ads) from 78% to 81%.

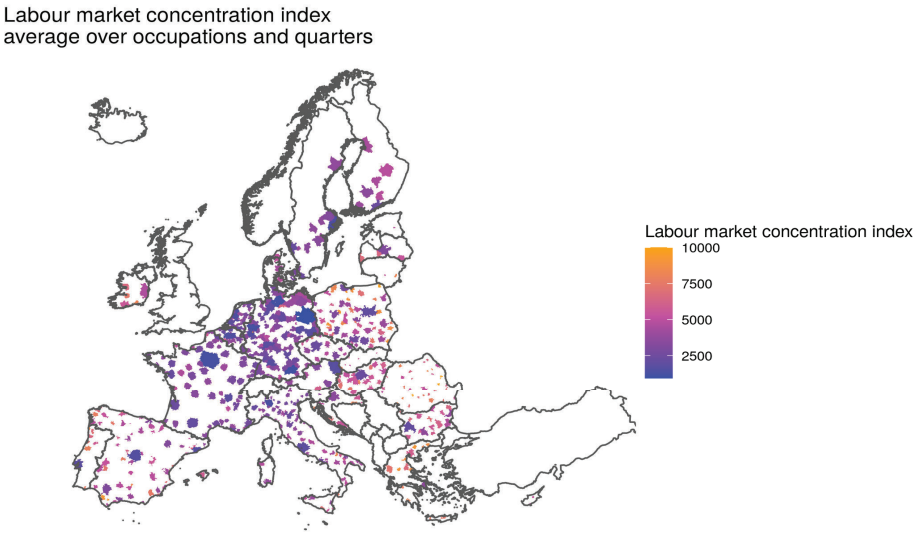
3. RESULTS

The algorithm developed in R allowed to calculate the HHI based on the share of vacancies of all the firms that post vacancies in that market, with a market defined as a triplet of time (quarter), occupation type (based on the ISCO classification code extracted from the job ad) and urban area. The market share of a firm in a given market is defined as the number of vacancies posted by the given firm in that market divided by the total number of vacancies posted in that market. The HHI is then aggregated through arithmetic averages at levels suitable for a descriptive analysis. A detailed presentation of the results from this descriptive analysis is provided in (Ascheri, et al., 2021), while only a brief summary will be given here.

As part of the results, the concentration indices are correlated with other variables collected by Eurostat at the urban-area and country level, namely the change in population / migration, employment rates and survey data satisfaction with personal job situation in urban areas. In addition, the evolution of the aggregate HHI over the considered timeframe (2019-2020) is analysed, and a glimpse of the recurring occupations (at 4-digit ISCO code level) is provided, together with their corresponding average concentration index. The goal of this descriptive analysis is not to uncover causal relationships or effects. It is rather to perform some external content validation, by showing that the newly calculated indicator is related to geography, time and other variables in a way roughly consistent with expectations.

Visual representation (ggplot2) of the concentration index (HHI) across European functional urban areas (average 2019-2020) across all occupation types

Figure 3



The results indicate that the largest urban areas in Europe tend to have lower level of concentration of the hiring market, indicating more competition among employers and more job opportunities for workers across all occupations. The index, averaged across all occupations, shows very low levels of concentration in large urban areas with thriving labour markets like

Berlin, Milan and Paris (indicating that job-seekers in these urban areas tend to have more online ads to consider). In contrast, there tends to be less choice of employers (as indicated by a higher level of concentration) all along the southern and eastern periphery of the European Union (Greece, Lithuania, Romania, Portugal and other countries and regions), particularly in smaller towns. This is also confirmed by migration trends that show how these urban areas attract more people in search for better job conditions. Some occupation types appear to be more concentrated than others on average, but this may be partly due to the fact that some occupations are more frequently advertised online than others. With respect to time series, an average increase in labour market concentration can be seen in the second quarter of 2020, when the pandemic crisis hit Europe stronger.

Despite its potential for official statistics, OJAs come with some limitations as they do not represent the entire job ads population. Some occupations and economic activities are less well represented than other and in some regions, digital tools may be less widespread leading to fewer advertisements published online. Moreover, the data source presents currently several quality issues, coming from the extraction pipeline that result in missing data, inadequate classifications or interruption of the scraping from some of the sources due to the instability of the web as a source of data. Despite these limitations the quality of the data source is expected to increase thanks to the ongoing activities on improving the data extraction pipeline.In addition, at several stages of the process some records are discarded due to missing information (e.g. geocoding, no contract or occupation type info) or a failed match with the functional urban areas. This results in an average 60% of ads used for calculating the index out of the total OJAs. The presence of some ads with a failed match was expected, because both of data limitations and of the fact that ads for jobs in rural areas cannot (by definition) be matched to an urban area. However, for some countries (e.g. IE, LT, SK) this percentage drops significantly, indicating some problems in matching the ads to a functional urban areas possibly due to changes in recent changes in the territorial units or an inadequate classification of the ads.

4. CONCLUSIONS

In this paper, we present one of the first Eurostat's attempts of producing statistical outputs from a non-traditional data source using R programming. A dataset of more than 100 million online job advertisements is used to calculate an index to measure the competition among employers in hiring markets of European urban areas. For this purpose, a methodology implemented in R is optimised to work with large amount of data for all the

27 EU Member States. Some of the main challenges to improve the efficiency of the calculation are discussed. Moreover, the authors present an algorithm developed using common R functions to deal with the classification of the names of the companies advertising jobs. This issue is discussed both from a conceptual point of view, highlighting the importance for the study, and from an implementation point of view, explaining the solutions adopted.

Following the open source nature of R, the code is shared publicly on GitHub, allowing replicating and further developing this research for those who have access to the OJA dataset. In addition, the public sharing of the code makes the results auditable and provides further details and additional knowledge of what is behind the graphs and tables.

Given the novel nature of the data source and the methodological improvements still under way, the nature of this work must be considered as experimental. This implies that the study can be improved in several ways. Besides the possible improvements regarding the quality of the input data source and the specific labour market analysis, which are discussed in Ascheri et al. (2021), there is room for improvements concerning the methodology implemented in R.

In particular:

- Production ready workflow: preparing the current methodological workflow in R for the regular production of statistics.
- Generalizing some of the functions and distributing them in an R package format.
- Improving the cleaning and classification of employer names. As for all machine learning models, a crucial input is human-coded data. In this case, this covered only a portion of data. The focus was Italy, Portugal and Romania, with additional input for Germany and Slovenia provided by the National Statistical Institutes of these countries. Expanding the set of human-coded data will improve the model input and allow to use better algorithms.

5. REFERENCES

1. Ascheri, A., Kiss Nagy, A., Marconi, G., Meszaros, M., Paulino, R., & Reis, F. (2021). *Competition in urban hiring markets: Evidence from online job advertisements*. Retrieved from https://doi.org/10.2785/667004
2. Azar, J., Marinescu, I., Steinbaum, M., & Taska, B. (2020). Concentration in US labor markets: Evidence from online vacancy data. *Labour Economics, 66*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0927537120300907
3. Beat Hulliger, R. L. (2012). *Analysis of the future research needs for Official Statistics.* Luxembourg: Eurostat Methodologies and Working Papers. Retrieved from https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-ra-12-026
4. Brown, W., & Scott, M. (2012). Human Capital Location Choice: Accounting for Amenities and Thick Labor Markets. *Journal of Regional Science, 52*, 787-808. Retrieved from . https://doi.org/10.1111/j.1467-9787.2012.00772.x

5. Cedefop. (2019). *Online job vacancies and skills analysis: a Cedefop pan-European approach.* Luxembourg: Publications Office. Retrieved from https://www.cedefop.europa.eu/files/4172_en.pdf
6. Cedefop; European Commission; ETF; ILO; OECD and UNESCO. (2021). *Perspectives on policy and practice: tapping into the potential of big data for skills policy.* Luxembourg: Luxembourg: Publications Office. Retrieved from http://data.europa.eu/doi/10.2801/25160
7. De Lazzer, J. (2020). *Labour market concentration index from CEDEFOP data*. Retrieved from https://github.com/OnlineJobVacanciesESSnetBigData/Labour-market-concentration-index-from-CEDEFOP-data
8. Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. *Statistical Journal of the IAOS, 35*(4), 669-675. Retrieved from https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547
9. Eddelbuettel, D. (2021). Retrieved from CRAN-R: https://cran.r-project.org/web/packages/drat/index.html
10. ESS. (2020). *ESSNet big data II*. Retrieved from WPB Online job vacancies: https://ec.europa.eu/eurostat/cros/content/WPB_Online_job_vacancies_en
11. ESS. (2020). *ESSnet Big Data II - Work Package B Online job vacancies*. Retrieved from https://ec.europa.eu/eurostat/cros/content/WPB_Online_job_vacancies_en
12. Eurostat. (2019, May 16). Trusted Smart Statistics Strategy and Roadmap, implementation of the Bucharest Memorandum on “Official Statistics in a datafied society (Trusted Smart Statistics)”. *40th meeting of the European Statistical System Committee*. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/item_02_-_background_document_-_essc_2019_40_07_tsssr.pdf
13. Eurostat GitHub, A. M. (2021, 12 01). *oja_hhi*. Retrieved from https://github.com/eurostat/oja_hhi
14. GISCO, E. (2021). *Eurostat GISCO services*. Retrieved from https://ec.europa.eu/eurostat/web/gisco/
15. Gordon, I., & Turok, I. (2005). How Urban Labour Markets Matter. In I. H. Gordon, *Changing Cities - Buck I* (pp. 242-264). London: Palgrave.
16. Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS, 32*, 43–52. doi:10.3233/SJI-160965
17. Hernangómez, D. (2021). *giscoR: Download Map Data from GISCO API - Eurostat*. Retrieved from https://cran.r-project.org/web/packages/giscoR/index.html
18. Jones, D. (2021). *noctua: Connect to ‘AWS Athena’ using R ‘AWS SDK’ ‘paws’ (‘DBI’ Interface)*. Retrieved from https://cran.rstudio.com/web/packages/noctua/index.html
19 . Manning, A. (2003). The real thin theory: monopsony in modern labour markets. *Labour Economics, 10*, 105-131. Retrieved from https://doi.org/10.1016/S0927-5371(03)00018-6
20. OECD & European Commission. (2020). *Cities in the World: A New Perspective on Urbanisation.* (P. OECD Publishing, Ed.) OECD Urban Studies. Retrieved from https://doi.org/10.1787/d0efcbda-en
21. OECD. (2012). *Redefining “Urban”: A New Way to Measure Metropolitan Areas.* Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264174108-en
22. OECD. (2020). *Competition in Labour Markets.* Retrieved from http://www.oecd.org/daf/competition/competition-concerns-in-labour-markets.htm
23. R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

24. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., & Skaliotis, M. (2019). Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS, 35*, 589–603. doi:10.3233/SJI-190584
25. R-SIG-DB R Special Interest Group on Databases, Wickham, H., & Müller, K. (2021). *DBI: R Database Interface*. Retrieved from https://cran.r-project.org/web/packages/DBI/index.html.
26. Stubner, R. (2018). *dqsample: A bias-free alternative to base::sample()*. Retrieved from R-bloggers: https://www.r-bloggers.com/2018/10/dqsample-a-bias-free-alternative-to-basesample/
27. Tam, S.-M. C. (2015). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. *International Statistical Review, 83*(3), 436-448. Retrieved from https://doi.org/10.1111/insr.12105
28. UN Statistical Commission. (2013). *Big data and modernization of statistical systems.* Report of the Secretary-General , Economic and Social Council. Retrieved from http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf
29. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., . . . Vaughan. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 1686.
30. Yongdai, K. K. (2013). Big data and statistics. *Journal of the Korean Data and Information Science Society, 24*, 959-974.

Autocoding based Multi-Class Support Vector Machine by Fuzzy c-Means

Yukako Toko (ytoko@nstac.go.jp)
National Statistics Center, 19-1 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan,

Mika Sato-Ilic (mika@risk.tsukuba.ac.jp)
University of Tsukuba, Tennodai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan,

ABSTRACT

This paper proposes a new autocoding method for the coding task of the Family Income and Expenditure Survey. The data of the Family Income and Expenditure Survey included text descriptions extracted from digital receipts which have been getting large and complex in recent years. This paper proposes a new autocoding method to obtain stable results of discrimination as coding with high generalization performance dealing with cognitive uncertainty for text description data. This method is a combination of multi-class Support Vector Machine (SVM) by fuzzy c-means and the previously developed reliability score based classification method. The proposed method utilizes both SVM, a machine learning method known as high generalization performance, and fuzzy c-means that is a computational intelligence method known as high performance dealing with cognitive uncertainty. Also, the proposed method utilizes the previously developed classification method based on reliability score. A numerical example shows a better performance of the proposed method with the Family Income and Expenditure Survey compared with the previously proposed classification method. The proposed method is developed in python utilizing python libraries, and also it can be easily run in R, which is a popular language in the official statistics field.

Keywords: *Coding, Machine Learning, Word2Vec, Support Vector Machine, Fuzzy c-Means, Reliability Score*

JEL Classification: *C38*

1. Introduction

For the coding task on official statistics, though coding is originally performed manually, the studies of automated coding have made progress with the improvement of computer technology. For example, Hacking and Willenborg (2012) introduced coding methods, including autocoding. Gweon et al. (2017) illustrated methods for automated occupation coding based on statistical learning. For the coding task of *the Family Income and Expenditure Survey*, we have developed an autocoding method which is the reliability score

However, it is well known that the Bernoulli type simple Bayes model does not perform well for a large amount of complex data, whereas the data of the *Family Income and Expenditure Survey* data, included text descriptions that were extracted from receipts digitally, is getting large and complex. In order to obtain stable results of discrimination as coding with high generalization performance dealing with cognitive uncertainty for text description data, this paper proposes a new autocoding method which is a combined method of Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000) and fuzzy c-Means (Bezdek, 1981).

In our previous study (Toko and Sato-Ilic, 2021a), we developed a hybrid method of SVM utilizing Word2Vec (Mikolov et al., 2013), which is a method to produce word embeddings, and the previously developed classification method based on reliability score. However, as the previously proposed method simply applied SVM to a whole given data, there was room for more efficiently classifying of those data to improve the classification accuracy. Therefore, we have proposed a hybrid method of multi-class SVM and k-means based on reliability score (Toko and Sato-Ilic, 2021b). The method utilized k-means before applying SVM to capture significant features of data. Based on the captured features, SVM is applied individually to each data included in the corresponding cluster. The merit of obtaining several groups before applying SVM is that it allows applying SVM individually to each group considering each group's discriminable ability. In the previously method, we combined such a function with the previously proposed hybrid method of SVM utilizing Word2Vec and our previously developed classification method based on reliability score.

However, our recent data is obtained as receipts digitally, which include complex representations of text description. Since the k-means is one of the methods of hard clustering in which each data is classified to only one cluster, the k-means tends to need many clusters to obtain a better solution. In this case, the result has less solution robustness, and a large amount of computation is necessary for obtaining a better solution.

Therefore, the purpose of this paper is the inclusion of fuzzy clustering as fuzzy c-means in order to obtain stable results dealing with cognitive uncertainty for text description data. For this purpose, this paper proposes a new autocoding method which is a combined method of SVM and a reliability score based fuzzy c-means in computational intelligence, which is linguistically motivated computational paradigms, theory and design of fuzzy logic, neural networks, and evolutionary computation.

The rest of this paper is organized as follows: Fuzzy c-means method, Word2Vec, and SVM are explained in sections 2, 3, and 4. The method of autocoding based reliability score is described in section 5. The multi-class SVM by fuzzy c-means is proposed in section 6. The numerical examples are illustrated in section 7. Conclusions are described in section 8.

2. Fuzzy c-means

In fuzzy clustering, each object has a degree of belongingness to clusters which can range from any value from 0 to 1. In hard clustering, each object has only two values for the degree of belongingness which is 0 or 1. If an object belongs to a cluster, then the degree of belongingness of the object to the cluster is 1, otherwise it is 0. However, if the obtained data has complexity which causes boundary uncertainty situation of the belongingness to the certain number of clusters, then without increasing the number of clusters, it will be difficult to explain the complex classification situation of the data. Therefore, fuzzy clustering is useful to explain the complex data into the clustering.

In this study, we utilize fuzzy c-means (Bezdek, 1981), which is one fuzzy clustering method. The purpose of fuzzy c-means is to obtain U and V which minimize the following objective function:

$$J(U,V) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ik})^m \|\boldsymbol{x}_i - \boldsymbol{v}_k\|^2, (1)$$

where, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$ is a vector of i-th object, $V = (v_{ka})$ shows a matrix of cluster centers consisted of a vector $\boldsymbol{v}_k = (v_{k1}, \dots, v_{kp})$ which is a center of cluster k, $U = (u_{ik})$ is a matrix of clustering results where u_{ik} is a degree of belongingness of i-th object to a cluster k which satisfy the conditions $u_{ik} \in [0,1], \sum_{k=1}^{K} u_{ik} = 1$. m is a parameter which controls fuzziness of the fuzzy clustering result and satisfies $m \in (1, \infty)$. K is a number of clusters, n is a number of objects, and p is a number of variables. Local optimum solutions which minimizes (1) are obtained as follows:

$$u_{ik} = \left[\sum_{j=1}^{K}\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{v}_k\|}{\|\boldsymbol{x}_i - \boldsymbol{v}_j\|}\right)^{\frac{2}{m-1}}\right]^{-1}. \quad (2)$$

$$v_k = \frac{\sum_{i=1}^{n}(u_{ik})^m \boldsymbol{x}_i}{\sum_{i=1}^{n}(u_{ik})^m}. \quad (3)$$

Using (2) and (3), the algorithm of fuzzy c-means is shown as follows:

Step1. Initialize the degree of belongingness of objects to clusters

Step2. Calculate the cluster centers by using (3)

Step 3. Update the degree of belongingness of objects to clusters by using (2)

Step 4. Stop if the difference of the degree of belongingness of objects to clusters and the degree calculated in the previous iteration is smaller than ε; otherwise, iterate steps 2 and 3

3. Word2Vec

Word2Vec was developed based on an idea of a neural probabilistic language model in which words are embedded to a continuous space by using distributed representations of the words (Mikolov et al., 2013). The algorithm of Word2Vec learns word association from a given dataset utilizing a neural network model based on an idea of a neural probabilistic

given dataset utilizing a neural network model based on an idea of a neural probabilistic language model (Bengio et al., 2003). It produces a vector space and each word in the given dataset is assigned a corresponding numerical vector of a word in the produced vector space. The essence of the idea is to avoid the curse of dimensionality by distributed representations of words.

Word2Vec utilizes continuous bag-of-words (CBOW) model and continuous skip-gram model (Mikolov et al., 2013) to distributed representations of words. The CBOW model predicts the current word based on the context. The skip-gram model uses each current word to predict words within a certain range before and after the current word. It gives less weight to the distant context words. In this study, the skip-gram model is applied.

4. Support Vector Machine

SVM (Cristianini and Shawe-Taylor, 2000) is a supervised machine learning algorithm for classification.

If $\boldsymbol{w}$ is the weight vector realizing a functional margin of 1 on the positive point $\boldsymbol{x}^+$ and negative point $\boldsymbol{x}^-$, a functional margin of 1 implies

$$\langle \boldsymbol{w} \cdot \boldsymbol{x}^+ \rangle + b = +1,$$
$$\langle \boldsymbol{w} \cdot \boldsymbol{x}^- \rangle + b = -1,$$

while $\boldsymbol{w}$ is normalized. Then the margin γ is the functional margin of the resulting classifier

$$\gamma = \frac{1}{2}\left(\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \cdot \boldsymbol{x}^+ \rangle - \langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \cdot \boldsymbol{x}^- \rangle\right)$$
$$= \frac{1}{2\|\boldsymbol{w}\|_2}(\langle \boldsymbol{w} \cdot \boldsymbol{x}^+ \rangle - \langle \boldsymbol{w} \cdot \boldsymbol{x}^- \rangle) = \frac{1}{\|\boldsymbol{w}\|_2}.$$

Therefore, the resulting margin will be equal to $1/\|\boldsymbol{w}\|_2$ and the following can be written. Given a linearly separable training sample

$$S = \left((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\right)$$

the hyperplane $(\boldsymbol{w}, b)$ that solves the optimization problem

$$\min_{\boldsymbol{w},b} \langle \boldsymbol{w} \cdot \boldsymbol{w} \rangle, \tag{4}$$

$$\text{subject to } y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1, i = 1, \dots, l,$$

realizes the maximal margin hyperplane with geometric margin $\mathrm{M} = 1/\|\boldsymbol{w}\|_2$. Then, slack variables are introduced to allow the margin constraints to be violated

$$\text{subject to } y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, i = 1, \dots, l,$$
$$\xi_i \geq 0, i = 1, \dots, l.$$

From the above, the optimization problem shown in (4) can be written as

$$\min_{\xi,w,b}\langle \boldsymbol{w}\cdot\boldsymbol{w}\rangle + C\sum_{i=1}^{l}\xi_i, \tag{5}$$

$$\text{subject to } y_i(\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + b) \geq 1-\xi_i, i=1,\dots,l,$$

$$\xi_i \geq 0, i=1,\dots,l.$$

where C is the cost parameter that will give the optimal bound as it corresponds to finding the minimum of $\|\xi\|_1$ in (5) with the given value for $\|\boldsymbol{w}\|_2$. This is soft-margin linear SVM.

Also, SVM performs a non-linear classification transforming input data into higher dimensional spaces and calculating the inner product between the data in higher dimensional space using kernel trick. SVM uses kernel functions to enable it to obtain the inner product of data in higher dimensional space (kernel trick), which is represented as follows:

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x}_j),$$

where φ is a mapping function from an observational space to a higher-dimensional space. The conditions for $k(\boldsymbol{x},\boldsymbol{x}')$ to be a kernel function are as follows:

- Symmetry: $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x}',\boldsymbol{x})$.
- Gram matrix is Positive semi-definite.

As there are many possible choices for the kernel function, the radial basis function is applied in this paper.

- Radial basis function kernel

$$k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2), \left(\gamma = \frac{1}{2\sigma^2}\right), \tag{6}$$

The mapped feature space of the radial basis kernel function has an infinite number of dimensions.

For multi-class SVM, there are two approaches: one-versus-the-rest and one-versus-one. In one-versus-the-rest, SVM builds binary classifiers that discriminate between one class and the rest, whereas it builds binary classifiers that discriminate between every pair of classes in one-versus-one.

5. Classification method based on reliability score

The classification method based on reliability score performs the extraction on objects and retrieval of candidate classes from the object frequency table provided by using the extracted objects. Then, it calculates the relative frequency of j -th object to a code k defined as

$$p_{jk} = \frac{n_{jk}}{n_j}, n_j = \sum_{k=1}^{K} n_{jk}, j=1,\dots,J, k=1,\dots,K,$$

where n_{jk} is the number of occurrence of statuses in which an object *j* assigned to a code k in the training dataset. J is the number of objects and K is the number of codes.

However, this classifier has difficulty correctly assigning codes to text descriptions for complex data included uncertainty. To address the problem, we developed the overlapping classifier that assigns codes to each text description based on the reliability score (Toko et al., 2018a, Toko et al., 2018b, Toko et al., 2019, Toko and Sato-Ilic, 2020). Then, the classifier arranges $\{p_{j1}, \cdots, p_{jK}\}$ in descending order and creates $\{\tilde{p}_{j1}, \cdots, \tilde{p}_{jK}\}$, such as $\tilde{p}_{j1} \geq \cdots \geq \tilde{p}_{jK}, j = 1, \cdots, J$. After that, $\left\{\tilde{\tilde{p}}_{j1}, \cdots, \tilde{\tilde{p}}_{j\tilde{K}_j}\right\}, \tilde{K}_j \leq K$ are created. That is, each object has a different number of codes. Then, the classifier calculates the reliability score for each code of each object. The reliability score of j-th object to a code k is defined as

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \qquad j = 1, \ldots, J, k = 1, \ldots, \tilde{K}_j.$$

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2\right), \quad j = 1, \ldots, J, k = 1, \ldots, \tilde{K}_j.$$

These reliability scores were defined considering both probability measure and fuzzy measure (Bezdek, 1981, Bezdek et al., 1999). That is, $\tilde{\tilde{p}}_{jk}$ shows the uncertainty from the training dataset (probability measure) and $1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}$ or $\sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2$ shows the uncertainty from the latent classification structure in data (fuzzy measure). These values of the uncertainty from the latent classification structure can show the classification status of each object; that is, how each object is classified to the candidate codes. *T* shows *T*-norm in statistical metric space (Menger, 1942, Mizumoto, 1989, Schweizer and Sklar, 2005). We generalize the reliability score by using the idea of *T*-norm, which is a binary operator in statistical metric space. Furthermore, to prevent an infrequent object having significant influence, sigmoid functions $g(n_j)$ were introduced to the reliability score. The reliability score considering the frequency of each object over the codes for each object in the training dataset as follows (Toko et al., 2019, Toko and Sato-Ilic, 2020):

$$\bar{\bar{p}}_{jk} = g(n_j) \times \bar{p}_{jk}.$$

In this study, algebraic product is taken as *T*-norm and $n_j/\sqrt{1 + n_j{}^2}$ is taken as a sigmoid function for the reliability score.

6. Autocoding method of SVM and fuzzy c-means method with the reliability score

The proposed autocoding method is a combined method of multi-class SVM by fuzzy c-means and classification method based on reliability score.

First, the proposed method tokenizes each text description by Sudachi (Takaoka et al., 2018). Then, it obtains numerical vectors corresponding words utilizing Word2Vec and normalizes each feature of the obtained set of numerical vectors. After that, it produces

sentence vectors for each text description based on the normalized vectors. Then, it applies fuzzy c-means to sentence vectors to classify them into several clusters. After applying fuzzy c-means, the proposed method assigns corresponding codes applying SVM for each dataset of each cluster. After that, it extracts dataset that assigned codes with low classification accuracy by SVM. Then, the proposed method applies fuzzy c-means and SVM to the extracted dataset. In fact, it iteratively applies fuzzy c-means and SVM to sub-clusters that has low classification accuracy at previous iteration until obtaining sufficiently better result. After code assignment by the combined method of fuzzy c-means and SVM, it extracts unmatched data and re-assign corresponding codes to the extracted data by the reliability score based classifier.

The detailed algorithm of the proposed method is the following:

Step 1. The proposed algorithm tokenizes each text description into words by Sudachi.

Step 2. It obtains numerical vectors corresponding to words utilizing Word2Vec: First, it produces a dataset concatenating all tokenized words consecutively. Then, it trains Word2Vec model using the produced dataset. Each unique word in the dataset will be assigned a corresponding numerical vector. The following are determined by trial and error:

- Type of model architecture: CBOW model or skip-gram model
- The number of vector dimensions
- The number of training iterations
- Window size of words considered by the algorithm

Step 3. It normalizes each feature of the obtained set of numerical vectors.

Step 4. It produces sentence vectors for each text description based on the normalized vectors at Step 3: First, it obtains a corresponding numerical vector for each word in each text description from the set of normalized numerical vectors. Then, it calculates the sum of obtained numerical vectors for each text description as sentence vectors.

Step 5. It applies the fuzzy c-means method: First fuzzy c-means method is applied to sentence vectors produced in step 4 to classify them into K clusters. For implementing the fuzzy c-means method, we determine the following by trial and error:

- Number of clusters K
- Error rate appeared in sect. 2 as ε
- m parameter appeared in sect. 2 as m
- Maximum number of iterations allowed

Step 6. It assigns corresponding codes applying SVM: It trains a Support Vector Machine and predicts a corresponding code for each target text description. For training a Support Vector Machine, we determine the following:

- Cost parameter appeared in (5) as C
- Kernel function to be applied
- Gamma parameter appeared in (6) as γ if radial basis function kernel is applied as a kernel function

- Type of methods: one-versus-the-rest or one-versus-one

In this study, a radial basis function as a kernel function is applied. We apply the one-versus-one method. Cost parameter C and gamma parameter γ are determined by grid search.

Step 7. It extracts datasets that assigned codes with low classification accuracy in step 6. The dataset that assigned codes with high classification accuracy in step 6 are accepted as classification results.

Step 8. It performs step 5 thorough step 7 iteratively to the extracted dataset in step 7 until obtaining sufficiently better result.

Step 9. It extracts unmatched data.

Step 10. It implements re-classification based on reliability score to the extracted unmatched data.

7. Numerical example

For the numerical example, the proposed method is applied to the *Family Income and Expenditure Survey* dataset. The *Family Income and Expenditure Survey* is a sampling survey related to a household's income and expenditure conducted by the Statistics Bureau Japan. This survey dataset includes short text descriptions related to a household's daily income and expenditure (receipt items name and purchase items name in Japanese) and their corresponding codes. In this numerical example, the target data is only data related to household expenditure in the dataset. The total number of codes related to household expenditure is around 520. Approximately 810 thousand text descriptions were used for this numerical example.

The proposed method was developed in python, applying the following python libraries: For training the Word2Vec model, we used "gensim" (Rehurek and Sojka, 2010). We selected the skip-gram model as a type of model architecture and set the number of vector dimensions as 100, the number of training iterations as 10,000, and the window size as 2. For implementing fuzzy c-means clustering, we used "skfuzzy" (Warner et al., 2019, scikit-fuzzy development team). We set the number of clusters as 2, m parameter appeared in sect. 2 as 1.1, error rate ε appeared in sect. 2 as 0.0001, and the maximum number of iterations as 10,000. For normalization of each feature of the set of numerical vectors and training support vector machines, we used "scikit-learn" (Pedregosa et al., 2011). We applied radial basis function kernel as the kernel function and selected the cost parameter C appeared in (5) and the gamma parameter γ appeared in (6) by grid search.

Table 1 shows the classification accuracy of SVM for each dataset obtained from the result of fuzzy c-means at the first iteration. The classification accuracy of dataset in C1 cluster is 0.858, whereas the classification accuracy of the dataset in C2 cluster is 0.917. Then, we accept the classification result of C2 cluster as classification result, and implement the second iteration to data in C1 cluster that has lower classification accuracy. Table 2 shows

classification accuracy of each cluster in the second iteration. The classification accuracy of C1_1 cluster is 0.788, whereas the classification accuracy of C1_2 cluster is 0.898, which is higher than the previously obtained score, 0.858. This means that we improve the accuracy of the data in C1 cluster partially. However, the remained data in C1, which is data in C1_1 is still a lower score at 0.788. Therefore, our next target to be improved accuracy is the data in C1_1. Then, we implement the third iteration to data in C1_1 cluster. Table 3 shows the classification accuracy of each cluster in the third iteration. The classification accuracy of C1_1_1 cluster is 0.837, whereas the classification accuracy of C1_1_2 cluster is 0.764. Again, we could obtain a better score which is 0.837 compared with 0.788. For the remained data, which has lower accuracy, has been implemented for further iteration to obtain a better accuracy. Table 4 shows successfully obtained the better score, which is 0.905 for the classification accuracy of the fifth iteration for the data in C1_1_2_1.

Table 1. First iteration: classification accuracy of each cluster

Cluster label	Number of text descriptions				Accuracy	SVM	
	Total	Train	Test	Correctly assigned		Cost	Gamma
C1	566,355	509,719	56,636	48,596	0.858	10	0.0012
C2	248,136	223,322	24,814	22,762	0.917	100	0.001
Total	814,491	733,041	81,450	71,358	0.876		

Table 2. Second iteration: classification accuracy of each cluster

Cluster label	Number of text descriptions				Accuracy	SVM	
	Total	Train	Test	Correctly assigned		Cost	Gamma
C1_1	249,261	224,334	24,927	19,635	0.788	10	0.0012
C1_2	317,094	285,384	31,710	28,488	0.898	10	0.0012
Total	566,355	509,718	56,637	48,123	0.850		

Table 3. Third iteration: classification accuracy of each cluster

Cluster label	Number of text descriptions				Accuracy	SVM	
	Total	Train	Test	Correctly assigned		Cost	Gamma
C1_1_1	88,614	79,752	8,862	7,420	0.837	10	0.0012
C1_1_2	160,647	144,582	16,065	12,275	0.764	10	0.0012
Total	249,261	224,334	24,927	19,695	0.790		

Table 4. Fourth iteration: classification accuracy of each cluster

Cluster label	Number of text descriptions				Accuracy	SVM	
	Total	Train	Test	Correctly assigned		Cost	Gamma
C1_1_2_1	61,398	55,258	6,140	5,556	0.905	10	0.0012
C1_1_2_2	99,249	89,324	9,925	6,700	0.675	10	0.0012
Total	160,647	144,582	16,065	12,256	0.763		

Table 5 shows the summary of classification accuracy of the proposed multi-class SVM by fuzzy c-means method. The total classification accuracy of the proposed method is 0.871 finally.

Table 5. Summary of classification accuracy of the proposed multi-class SVM by fuzzy c-means

Cluster label	Number of text descriptions				Accuracy
	Total	Train	Test	Correctly assigned	
C2	248,136	223,322	24,814	22,762	0.917
C1_2	317,094	285,384	31,710	28,488	0.898
C1_1_1	88,614	79,752	8,862	7,420	0.837
C1_1_2_1	61,398	55,258	6,140	5,556	0.905
C1_1_2_2	99,249	89,324	9,925	6,700	0.675
Total	814,491	733,040	81,451	70,926	0.871

In addition, table 6 compares the classification accuracy of the proposed method that is a combined method of multi-class SVM by fuzzy c-means method and the reliability score and the previously proposed method that is a combined method of multi-class SVM by k-means method and the reliability score (Toko and Sato-Ilic, 2021b). When combined multi-class SVM by fuzzy c-means and the reliability score, the classification accuracy is 0.922. This is better than the classification accuracy of the previously proposed method, as almost 250 data was increased for the correctly assigned. Although the number of increased text descriptions is not so large, the classification accuracy of the previously proposed method has already over 0.9. This means, therefore, text descriptions which easily classified to codes have already been assigned, and only difficult text descriptions to be classified have remained. Considering this matter, successfully classifying such difficult 250 text descriptions to the correct codes shows better results by the use of fuzzy clustering. Note that, for tokenizing text descriptions, the previously proposed method applies MeCab (Kudo et al., 2004), whereas this study applies Sudachi (Takaoka et al., 2018).

Table 6. Comparison of classification accuracy of the proposed method and the previously proposed method

Classification method	Accuracy
Combined method of multi-class SVM by **fuzzy c-means method** and the reliability score (Propsed method)	0.922
Combined method of multi-class SVM by **k-means method** and the reliability score (Previously propsed method)	0.919

Furthermore, when comparing the classification accuracy of multi-class SVM by fuzzy c-means and the classification accuracy of multi-class SVM by k-means, k-means in the previously proposed method requires many clusters to obtain a better result. Table 7 shows the classification accuracy of multi-class SVM by k-means. When comparing table 1 and table 7, the classification accuracy of the multi-class SVM by fuzzy c-means method is 0.876 with the number of clusters as 2, whereas the classification accuracy of the multi-class SVM

by k-means is 0.862 with the number of clusters as 10. From this comparison, it can be seen that SVM by fuzzy c-means allows us to reduce the number of clusters while retaining higher classification accuracy when compere with the multi-class SVM by k-means. In other words, when we utilize fuzzy c-means instead of k-means, data in each cluster tend to remain homogeneously, and the effectiveness of utilizing the cognitive uncertainty for text description data by fuzzy c-means can be seen in this result.

Table 7. Classification accuracy of SVM by k-means (previously proposed method)

Cluster	Number of text descriptions				Accuracy	SVM	
	Total	Training	Evaluation	Correctly assigned		cost	gamma
Cluster 1	4,568	4,111	457	375	0.821	30	0.0001
Cluster 2	37,454	33,708	3,746	3,719	0.993	100	0.0010
Cluster 3	137,157	123,441	13,716	12,068	0.880	30	0.0010
Cluster 4	148,585	133,726	14,859	12,909	0.869	10	0.0064
Cluster 5	38,003	34,202	3,801	3,082	0.811	10	0.0010
Cluster 6	31,288	28,159	3,129	3,116	0.996	100	0.0010
Cluster 7	275,852	248,266	27,586	22,929	0.831	90	0.0255
Cluster 8	47,421	42,678	4,743	4,243	0.895	90	0.0001
Cluster 9	48,332	43,498	4,834	4,093	0.847	100	0.0010
Cluster 10	45,831	41,247	4,584	3,672	0.801	10	0.0010
Total	814,491	733,036	81,455	70,206	0.862		

8. Conclusion

This paper proposes a new autocoding method which is a combined method of multi-class SVM by fuzzy c-means and the previously developed classification method based on reliability score to obtain stable result of discrimination as coding with high generalization performance dealing with cognitive uncertainty for text description data. SVM, a machine learning method known as high generalization performance, is utilized for classification based on the numerical vectors obtained by Word2Vec. Fuzzy c-means, a computational intelligence method known as high performance dealing with cognitive uncertainty with linguistically motivated computation, is utilized as a fuzzy clustering method to capture significant features of data before applying SVM. In addition, the previously developed classification method based on reliability score is applied to improve classification accuracy. The numerical example shows a better performance of the proposed method with *the Family Income and Expenditure Survey* data. From the result of the numerical example, it seems that data in each cluster tend to remain homogeneously, and the effectiveness of utilizing the cognitive uncertainty for text description data by fuzzy c-means. The proposed method is developed in python utilizing python libraries. Also, it can be run in R that is a popular language in the official statistics field utilizing the “reticulate” package (Ushey et al., 2021), which provides a comprehensive set of tools for interoperability between R and python. As

the “reticulate” package provides the “source_python( )” function that enable us to source a python script as we would source an R script, all functions in the python script become directly available to the R session after sourcing the script. Therefore, all functions and objects in our developed python script can be easily called from R utilizing the “reticulate” package. In future work, we would like to consider making the code open source in the framework in R.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003) “A neural probabilistic language model”, Journal of Machine Learning Research, 3, pp. 1137-1155.
2. Bezdek, J.C. (1981), Pattern recognition with fuzzy objective function algorithms, Plenum Press.
3. Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R. (1999), Fuzzy models and algorithms for pattern recognition and image processing, Kluwer Academic Publishers.
4. Cristianini, N., Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press.
5. Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., Steiner, S. (2017), “Three methods for occupation coding based on statistical learning”, Journal of Official Statistics, Vol. 33, No. 1, pp. 101-122.
6. Hacking, W., Willenborg, L. (2012). “Coding; interpreting short descriptions using a classification”, Statistics Methods, Statistics Netherlands, The Hague, Netherlands, Available at: https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/throughput/throughput/coding (accessed December 2020).
7. Kudo, T., Yamamoto, K., Matsumoto, Y. (2004), “Applying conditional random fields to Japanese morphological analysis”, in the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230-237.
8. Menger, K. (1942), “Statistical metrics”, in Proceedings of the National Academy of Sciences of the United States of America, Vol. 28, pp. 535-537.
9. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013), “Efficient estimation of word representations in vector space”, arXiv preprint arXiv:1301.3781.
10. Mizumoto, M. (1989), “Pictorical representation of fuzzy connectives, Part I: Cases of T-norms, t-Conorms and Averaging Operators”, Fuzzy Sets and Systems, Vol. 31, pp. 217-242.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011), “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830.

12. Rehurek, R., Sojka, P. (2010), “Software Framework for Topic Modelling with Large Corpora”, in Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45-50.
13. Schweizer, S., Sklar, A. (2005), Probabilistic metric spaces, Dover Publications.
14. Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., Matsumoto, Y. (2019), “Sudachi: a Japanese Tokenizer for Business”, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), May 2018, Miyazaki, Japan, pp. 2246-2249, European Language Resources Association.
15. Toko, Y., Wada, K., Iijima, S., Sato-Ilic, M., (2018a), “Supervised multiclass classifier for autocoding based on partition coefficient”, Czarnowski, I., Howlett, R.J., Jain, L. C., and Vlacic, L. (Eds.), Intelligent Decision Technologies 2018, Smart Innovation, Systems and Technologies, Springer, Vol. 97, pp. 54-64.
16. Toko, Y., Iijima, S., Sato-Ilic, M. (2018b), “Overlapping classification for autocoding system”, Journal of Romanian Statistical Review, Vol. 4, pp. 58-73.
17. Toko, Y., Iijima, S., Sato-Ilic, M. (2019), “Generalization for improvement of the reliability score for autocoding”, Journal of Romanian Statistical Review, Vol. 3, pp. 47–59.
18. Toko, Y., Sato-Ilic, M., (2020), “Improvement of the training dataset for supervised multiclass classification”, Czarnowski, I., Howlett, R.J., Jain, L. C. (Eds.), Intelligent Decision Technologies, Smart Innovation, Systems and Technologies, Springer, Singapore, Vol. 193, pp. 291-302.
19. Toko, Y., Sato-Ilic, M., (2021a),”Efficient Autocoding Method in High Dimensional Space”, Romanian Statistical Review, Vol. 1, pp. 3-16.
20. Toko, Y., Sato-Ilic, M., (2021b), “A Hybrid Method of Multi-Class SVM and Classification Method Based on Reliability Score for Autocoding of the Family Income and Expenditure Survey”, Czarnowski, I., Howlett, R.J., Jain, L. C. (Eds.), Smart Innovation, Systems and Technologies, Vol. 238, pp. 403-414. Springer, Singapore.
21. Ushey, K., Allaire JJ, Tang Y. (2021), reticulate: Interface to ‘python’, R package version 1.22, https://CRAN.R-project.org/package=reticulate.
22. Warner, J., Sexauer, J., scikit-fuzzy, twmeggs, alexsavio, Unnikrishnan, A., et al. (2019) JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2, https://doi.org/10.5281/zenodo.3541386.
23. scikit-fuzzy development team, “skfuzzy” Available at: https://pythonhosted.org/scikit-fuzzy (accessed November 2021)

Determining the Business Cycle of Turkey

Muhammed Fatih Tüzen, Phd (fatih.tuzen@tuik.gov.tr)
Turkish Statistical Institute, Ankara, Turkey

Fatma Aydan Kocacan Nuray (aydan.kocacan@tuik.gov.tr)
Turkish Statistical Institute, Ankara, Turkey

İlayda Kuru (ilayda.kuru@tuik.gov.tr)
Turkish Statistical Institute, Ankara, Turkey

ABSTRACT

In this study, it is aimed to examine the basic characteristics of the cyclical fluctuations in the Turkish economy and to determine the business cycles (contraction and expansion). By using the Bry and Boschan (1971) algorithm, the turning points in the business cycles were obtained. In order to determine the business cycle, Turkey's monthly Industrial Production Index, monthly and quarterly Gross Domestic Product data were examined and analyzed. As a result of the analysis, the average business cycle for the Turkish economy was calculated as 5 years. It has been observed that this result is compatible with the related studies in the literature and the cycle characteristics of developing countries. Peak and trough points were obtained with the algorithm named "Harding-Pagan (Quarterly Bry-Boschan) Business Cycle Dating Procedure" in the BCDating R package released in 2019.

Keywords: *business cycle, Bry-Boschan procedure, economic crises, Tramo-Seats, temporal disaggregation, national accounts, industrial production*

JEL codes: *C22, E32 and P44*

1. INTRODUCTION

Business cycles can be expressed as repetitive and fluctuating movements in a country's economic activities. "The business cycle is the periodic but irregular up-and-down movements in economic activity measured by fluctuations in real GDP and other macroeconomic variables." (Parkin and Bade, 2015). The business cycle is critical to policymakers as it gives information about the state of the economy. Policymakers determine policies that will stabilize the fluctuations in the economy. Therefore, the business cycle analysis is an essential indicator for monetary policy. (Luvsannyam et al., 2019)

There are two primary approaches used to define the economy's cyclical behavior: classical and growth (or deviation) cycle approaches. The classical business cycles approach defines cycles in terms of absolute declines and increases of macroeconomic time series (Schumpeter, 1939). On the contrary, growth business cycles approach, turning points are

defined concerning deviations of the rate of growth of macroeconomic time series from their long-term trend component (Kydland and Prescott, 1990).

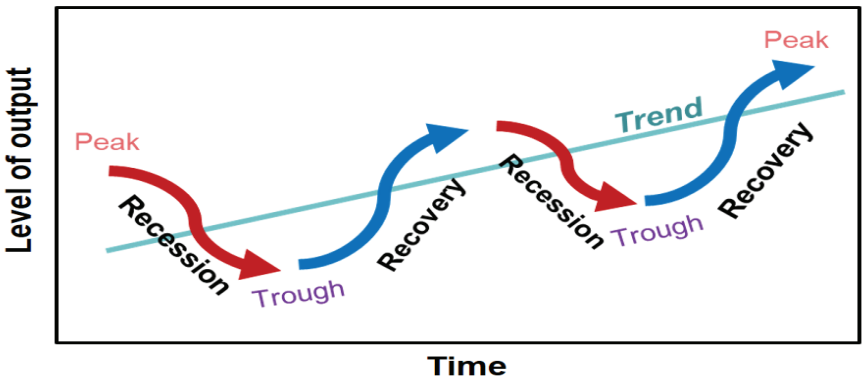
The essential reference to the business cycle is the turning points determined by NBER (National Bureau of Economic Research) for the US. Aims of the NBER dating process are to find large fluctuations, determine maximum and minimum points, and capture specific dates as a turning point. Nevertheless, consistency is lost over time since it is a non-programmed approach. (Christoffersen,2000) Although there are many different definitions, it can be said that there is a consensus in the literature about some features of business cycle fluctuations (Christiano and Fitzgerald (1998), Banerji and Hiris (2001), Klein and Moore (1982), Stock and Watson (1989), Zarnowitz (1987). These are;

- Business cycle fluctuations are defined as the joint movement of many economic variables.
- The cyclical fluctuations are repetitive but non-periodic.
- Business cycle fluctuations must be precise and continuous; small and short movements do not reflect this situation.

There are four stages of the business cycle;

- In the **Expansion** stage, time series have increasing movement and are above the trend.
- In the **Recession** stage, the time series decreases even though it is above the trend.
- In the **Depression** stage, the time series is below the trend and has decreasing movement.
- In the **Recovery** stage, the time series increases despite being below the trend.

Figure 1



The classical analysis of business cycle fluctuations was first made by Burns and Mitchell (1946). Their studies made some suggestions, such as using seasonally adjusted series while determining the business cycle. The other one is that the business cycle should last for at least 15 months and at most 10-12 years. They also used "reference dates" to determine the turning points of the business cycle. The method developed by Burns and Mitchell for calculating turning points was translated into a computer algorithm in 1971 by Bry and Boschan. The Bry-Boschan algorithm has been developed and applied in many studies. The most well-known works are by Stock and Watson (1989), Artis (1995).

There are many studies on Turkey's business cycle calculation in the literature. Leading indicators were estimated for Turkey by Selçuk (1994) by using the business cycle approach. Atabek et al. (2005) obtained the turning points using the Bry and Boschan algorithm. In their study, Özkan and Erden (2007), analyzed the monthly industrial production index to learn the basic characteristics of business cycles. By using the Bry-Boschan algorithm to determine the turning points of the cycles, it has been revealed that the cycle time is between two and six years. Alp et al. (2011) used two different methods to estimate the smoothing parameter in the Hodrick-Prescott (HP) filter used in business cycle analysis. The average cycle length in Turkey was approximately four years were calculated.

This study, it is aimed to examine the main features of the cyclical fluctuations in the Turkish economy and, to date or in other words, to determine the cycle duration (contraction and expansion). Since the studies conducted in the national field are in the past, the data used in this study contribute to the literature as it covers a long period and is up-to-date. Since the Gross Domestic Product (GDP) data has been used with quarterly frequencies in the studies carried out in this context so far, a special feature of this study is the monthly Gross Domestic Product (how to obtain the data is explained in the next section) data used.

2. LITERATURE

Cyclical fluctuations have been the subject of many scientific studies and discussions. There is extensive literature attempting to find reliable estimating tools for the business cycle, from the first landmark study by Burns and Mitchell (1946) to the more complex Stock and Watson (1989). In the international literature, Stock and Watson examined the leading indicators of the business cycles of the USA, and Artis (1995) examined the OECD countries. The detection of turning points begins with defining the concept of a cycle. In the classical cycle, fluctuations in the absolute level of the series are identified. The early NBER approach identified cycles as recurrent sequences of alternating phases of expansion and contraction in the levels of a large number of economic time series (Burns and Mitchell, 1946; Bry and Boschan, 1971).

With the general move towards estimating turning points in business cycles, much attention has been given to which models the best estimate these points. Such models include linear, non-linear (including Markov switching) parametric, and non-parametric models. However, some schools of thought suggest that turning point determination should instead be based on the business cycle's fundamental (theoretical) definition, as defined by Burns and Mitchell (1946), as opposed to model-based approaches. Burns and Mitchell (1946) defined business cycles as a type of fluctuation found in the aggregated economic activity of nations that organize their work mainly in business enterprises:

- a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle,
- the sequence of change is recurrent but not periodic,
- business cycles vary from more than one year to ten or twelve years,
- business cycles are not divisible into shorter cycles of similar character with amplitudes approximating their own.

Bry and Boschan (BB) (1971) replicated the Burns and Mitchell approach to determining turning points and later introduced a method for working with quarterly data. They coded the BB procedure into an algorithm that could easily be applied. Harding and Pagan

(2002) developed a BBQ (Quarterly Bry-Boschan) version of this method. Similarly, Moore (1980) noted that expansions and contractions should reflect an absolute rise and fall in trend-adjusted aggregate economic activity. Moore and Zarnovitz (1986) used a weighted average of several series rather than a single series. Burns and Mitchell (1946) also did not have a GDP series available to them at the time and instead extracted a reference cycle from many series to determine turning points. They dated turning points based on where the data clustered during peaks and troughs.

3. METHODOLOGY

3.1. Bry and Boschan Procedure

Studying the duration and width of the determined business cycle fluctuations requires a separate analysis study. The method developed by Burns and Mitchell for calculating the trough and peak point was transferred to a computer algorithm with Bry and Boschan in 1971. Bry and Boschan formulated it with computer codes using the business cycle criteria determined by NBER. In this method, which is widely used in the literature in order to determine the turning points in a series, the following path is basically followed.

a) The peak is determined based on consecutive decreases in the absolute level of the series.
b) The trough is determined based on successive increases in the series level.
c) Peaks and troughs need to change in different cycles. For this, multiple observed equal peaks and troughs will not be considered.
d) In order for consecutive increases (decreases) to be defined as an expansion (contraction) phase, they must complete at least two months in quarterly series and at least five months in monthly series.
e) For the cycles to be defined as a business cycle, the contraction and expansion phases should last at least five months or longer in quarterly series and at least 15 months or longer in monthly series.

Additionally, the Bry-Boschan (BB) and Harding Pagan (H-P) algorithms end the turning points as follows:

- The data is smoothed after outlier adjustment by constructing short-term moving averages.
- The preliminary set of turning points are selected for the smoothed series subject to the criterion described later.
- Turning points in the raw series are identified, taking results from smoothed series as the reference (Pandey et al., 2017).

3.2. Tramo-Seats

TRAMO and SEATS, a time series decomposition method, are two programs developed by Victor Gomez and Agustin Maravall with the support of Gianluca Caporello for the analysis of monthly, quarterly, semi-annual and annual data. (Gomez and Maravall, 1996). It is a method based on econometric model estimation and developed by the Bank of Spain, also recommended by EUROSTAT.[1]

This method estimates effects with a parametric method and separates them from the statistically significant data. Therefore, a seasonal or calendar effect that is not statistically significant is not excluded from the data. TRAMO (Time Series Regression with ARIMA Noise, Missing Observations and Outliers) is used to estimate and predict a regression model,

[1] http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf

taking into account non-stationary (ARIMA) error terms and missing observations. The program detects and corrects several types of outliers by estimating missing observations by interpolation. It also predicts special effects such as the calendar effect and Easter or exogenous variables. Fully automatic pattern detection and outlier correction procedures are also available. SEATS (Signal Extraction in ARIMA Time Series) is a program used to predict components in the time series that cannot be directly observed using the ARIMA model obtained in the TRAMO stage. The two programs are configured to be used together.

3.3. Temporal Disaggregation

The time series to be used in the analysis may not always be at the desired frequency. Temporal disaggregation methods are used to disaggregate low frequency time series to higher frequency series. The most used methods are Denton (Denton, 1971), Chow-Lin (Chow and Lin, 1971), Fernandez (Fernández, 1981) and Litterman (Litterman, 1983). Chow-Lin's (Chow and Lin, 1971) method executes a regression on the low frequency series which are stationary or cointegrated. (Sax and Steiner, 2013)

4. ECONOMIC CRISES OF TURKEY

The meanings of words such as "danger", "distress" and "depression" can be loaded into the crisis. The economic crisis appears as a phenomenon in which some external and internal factors lead to adverse developments suddenly or unexpectedly, causing severe damage to both firms and the country (Yılmaz,2005).

In Turkey, the period of 1988-1993 was when the growth strategy based on the expansion of domestic demand and the use of external resources was maintained, and growth was preferred to stability. However, regular and stable growth could not be achieved in GDP, the instability in the economy deepened, and annual inflation settled around 60% on average. There was severe stagflation in 1988, and this stagflation was tried to be overcome with monetary and fiscal measures on February 4, 1989. As a result of applications, the crisis in financial markets was partially stopped, but inflation in the economy increased from 50% to 60% on average (Şahin, 2014). The unfavorable conjuncture caused by the Gulf War in 1991 had a high cost to Turkey, and revival in the economy in 1990 stopped again in 1991. After the problematic and uncontrolled growth in 1992-1993, the economy experienced a severe crisis in 1994, and the GDP shrank by 6.1%.

As a result of the deepening crisis in 1994, the economic stabilization program known as the "April 5 Decisions" was put into practice. With the April 5 Decisions put into practice, stability in the money and foreign exchange markets in the short term were achieved, economic growth gained momentum, the exchange rate was brought under control, and confidence in the Turkish Lira (TL) increased. Even if the measures of April 5 Decisions regarding internal and external imbalances were seen as successful in the short term, they could not reach their goals in the long term (Toprak,1996).

In order to end macroeconomic instability, Turkey implemented the new economic program at the beginning of 2000. Based on the stand-by agreement with the IMF, the program suffered a strong shock in November 2000 and collapsed in February 2001. While the interest rate and exchange rate are rising, The Central Bank of the Turkish Republic's (CBRT) foreign exchange reserves has melted (Özatay, 2013).

The 2001 February Crisis was a continuation of the currency crisis of November 2000. Exchange rate losses in the November Crisis increased the imbalances in macroeconomic indicators in the economy, which led to political tensions in the country and the deepening of the crisis (Boratav, 2000). The long-term stabilization program, which was initiated with the

"Transition to a Strong Economy Program" after the February 2001 crisis, included a series of structural reforms as well as inflation targeting, exchange rate, foreign trade, monetary, fiscal and income policies.[2]

The cyclical fluctuations and the growth rates of the national economies are closely related. One of the general characteristics of the crisis periods of the Turkish economy was the decline in GDP. After the 2001 crisis, economic growth was achieved due to the transition to a Strong Economy Programme. In the last quarter of 2008, it was interrupted by the world's global crisis, and a contraction was observed until the first quarter of 2009. The growth that started with the "Transition to a Strong Economy Program" has become sustainable in the long run. In times of crisis, the contraction in the economy brought along high inflation and an increase in unemployment rates. With the transition to inflation targeting, which includes a more extended period, together with the "Transition to a Strong Economy Program", inflation rates were reduced to single digits, limited interventions were made to the exchange rate in the short term, and inflation assumed the anchor role. However, in this period, the value of TL increased at high levels in the floating exchange rate (Fırat and Demirtaş, 2012).

While the wounds inflicted by the economic crises experienced in the past years could not be healed yet, a global crisis broke out in August 2007. The starting point of this crisis; was caused by the Subprime Mortgage (high risk and high-interest loan) scandal in the USA. The 2008 global economic crisis affected developing and developed countries and caused recession and unemployment (Savaş, 2012).

The 2008 global economic crisis started in the housing market in the USA, but the reflection of this crisis on Turkey was on the industrial sector rather than the housing market. Because the private sector had a large amount of foreign currency debt, it was caught in this crisis (Susam and Bakkal, 2008). While the economic crisis caused a decrease in the capacity utilization rate and GDP, it also increased the uncertainty. Increasing uncertainty has caused investment decisions to be postponed, or the capital to shift to the finance sector (Demircan, 2018). Other macroeconomic balances have also deteriorated, but the 2008 global crisis did not shake the Turkish economy to the expected depth compared to the crises in previous years (Kaba, 2019).

While Turkey was already dealing with problems in economic dynamics in 2018, the Pastor Brunson crisis exacerbated these problems and ignited the economic crisis of 2018. While the Turkish Lira depreciated largely, inflation was stuck in double digits. With the increase in interest rates, growth has lost its momentum, high inflation, economic contraction, and increasing unemployment have become inevitable. With the Pastor Brunson crisis, the current exchange rate suddenly increased. In many sectors, especially in construction, bottlenecks have been experienced, and bankruptcies have followed. The Central Bank's foreign exchange reserves also decreased significantly in 2018. While the net foreign exchange reserve was 77.9 billion dollars in January 2018, it decreased to 33.9 billion dollars as of January 2020. The 2018 economic crisis, with its current conditions, is classified as a long and sticky economic crisis in the literature.

The effects of the economic contraction due to Covid-19 will also be clearly seen in the coming years.

[2] Republic of Turkey Ministry of Treasury and Finance, www.treasury.gov.tr

5. DATA

In our analysis, three different data sets were used to determine Turkey's average cycle. These are;

- Quarterly GDP (Gross Domestic Product) data between 1987Q1 and 2021Q2
- Monthly IPI (Industrial Production Index) data between 1986M1and 2021M6
- Monthly GDP data between 1987M1 and 2021M6

Since the GDP variable is thought to reflect the overall economic performance, this series is widely used in business cycle studies. However, since GDP is published at quarterly frequencies, it was found appropriate to analyze the timing of the movements of economic activity with series measured at monthly frequencies. Therefore, to determine the average cycle, also monthly IPI data were analyzed. In addition, the GDP data was converted into monthly series and used as a third series to determine the average cycle by using the monthly IPI series as a regressor variable. Chow-Lin, one of the most common methods, was used in temporal disaggregation.

Although the main purpose of our study is to determine the cycle duration, what we need is to obtain the cycle component for the Business Cycle Monitor of Turkey. For this reason, first of all, the trend-cycle components of the series should be obtained. We used the TRAMO-SEATS procedure to adjust the seasonality of GDP and IPI to get a smoother estimation of the cycle. Because when defining the business cycle, noisy estimation is not efficient. The Bry-Boschan algorithm was used to determine the peak-trough points and the average cycle by using the trend-cycle component of the series.

All computations regarding performed analyses were carried out in an R environment (R Core Team, 2020) by using BCDating (Einian, 2019) and RJDemetra (Quartier-la-Tente et al., 2021) packages. BCDating package uses the Harding and Pagan algorithm that creates a quarterly dating from a univariate time series. The peak-trough points of the series were determined by using the BBQ function included in this package. The *mincycle* (Minimum length of a cycle) and *minphase* (Minimum length of a phase) arguments in the function can take different values. In this study, as recommended in the Bry-Boschan procedure, the mincycle argument is set to 15 for monthly series and 5 for quarterly series, and the minphase argument to 2 for monthly series and 5 for quarterly series. We also used RJDemetra package in order to obtain seasonally adjusted series. This package is an interface for JDemetra+, the seasonal adjustment software officially recommended to the members of the European Statistical System (ESS) and the European System of Central Banks.

6. RESULTS

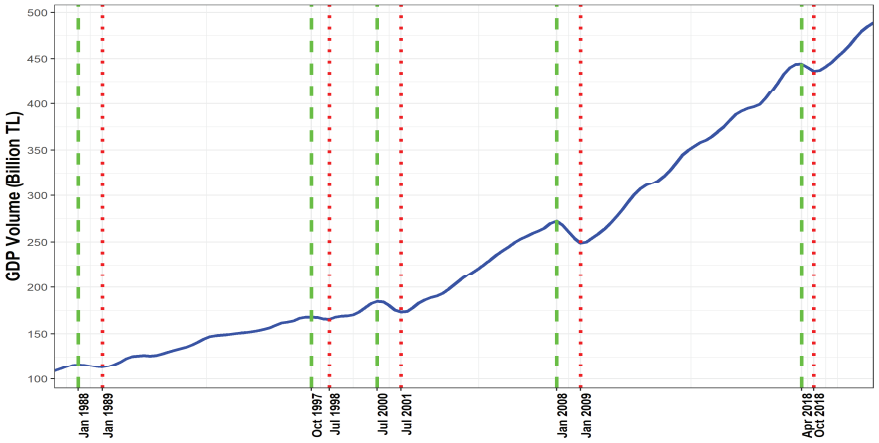
The results were examined separately for each data and were evaluated especially in terms of the cycle durations of the data, the length of the expansion and contraction phases, and how much the determined low points reflect the known crisis periods in Turkey.

6.1. Quarterly GDP

The peaks and troughs determined by the trend-cycle component of quarterly GDP are shown in Figure 1, and the results of the cycles are shown in Table 1.

Peak and Trough dates based on quarterly GDP

Figure 2



The red dotted lines in figure 1 show the trough points, and the green dotted lines show the peaks. As seen in the figure, four complete business cycles were observed in terms of the periods covered by the data used. The first recession point is observed in the first quarter of 1989. In 1988, the expected growth in sectors was not achieved, and the liberalization of imports of all goods, except for exceptional goods, appealed to industrialists. With the increase in the exchange rate, stagflation was experienced in Turkey in 1989. While the crisis was not yet fully overcome, the Berlin Wall collapsed, the Iraq crisis, and the USSR's disintegration took place. These events have caused radical transformations in economic policies globally and in Turkey. In 1998, was determined as the second crisis point, the Asian-Russian crisis occurred, and Turkey was also affected by this crisis. After the monetary crisis in the third quarter of 2001, the Turkish economy entered a long period of expansion. Turkey was also under the influence of the global economic crisis that emerged in the last months of 2008. Lastly, the lowest point was the foreign exchange and debt crisis, effective in the last quarter of 2018. In this period, TL depreciated significantly.

Business cycle chronology based on the quarterly GDP

Table 1

Phase	Start	End	Duration
Expansion	-	1988Q1	-
Recession	1988Q1	1989Q1	4
Expansion	1989Q1	1997Q4	35
Recession	1997Q4	1998Q3	3
Expansion	1998Q3	2000Q3	8
Recession	2000Q3	2001Q3	4
Expansion	2001Q3	2008Q1	26
Recession	2008Q1	2009Q1	4
Expansion	2009Q1	2018Q2	37
Recession	2018Q2	2018Q4	2
Expansion	2018Q4	-	-

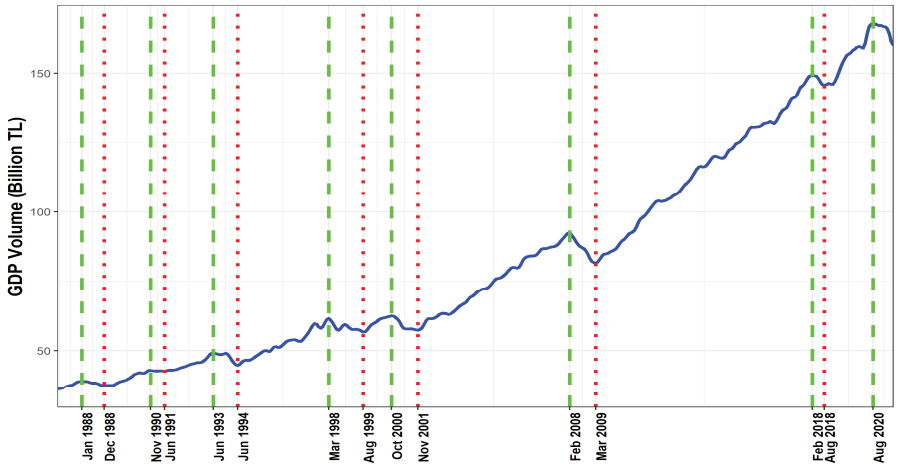
According to Table 1, while the average expansion is 26.5 quarters, the average recession is 3.4 quarters. These results mean the average expansion period is about eight times longer than the average recession period. It is an expected result for a developing country's economy (Rand and Tarp, 2002). Also, the low average recession period indicates that Turkey has quickly come out of the crisis. According to these findings, cycles vary between 11 and 41 quarters. The average cycle length in terms of peaks is 30.25 quarters, while it is 29.75 quarters in terms of troughs. According to both peaks and troughs, we can state that the cycle is 7.5 years.

6.2. Monthly GDP

The peaks and troughs determined by the trend-cycle component of monthly GDP are shown in Figure 2, and the results of the cycles are shown in Table 2.

Peak and Trough dates based on monthly GDP

Figure 3



Monthly GDP results show that six complete business cycles were observed in troughs. In addition to the results obtained from the quarterly GDP data, crises were also detected in 1991 and 1994. Contrary to the quarterly GDP data, the periods known as the crisis period for Turkey could be determined with the monthly GDP data. The first crisis of the Turkish economy shaped by external effects was the Gulf Crisis in 1990. The Gulf War, shaped by the United Nations' interventions in Iraq and Kuwait, is one of the crucial events of this crisis. Turkey felt the Gulf crisis that started in the late 1990s for seven months. Turkey experienced its most profound crisis in 1994. This economic crisis started in the middle of 1993 for about a year. Before 1994, public sector primary expenditures had a more significant deficit than public revenues. The public sector spent more than it earned. As a result of financing public debts with the Central Bank, Turkey experienced hyperinflation for the first time. With partial but insufficient improvements, Turkey was able to get out of the crisis.

Business cycle chronology based on the monthly GDP

Table 2

Phase	Start	End	Duration
Expansion	-	1988M1	-
Recession	1988M1	1988M12	11
Expansion	1988M12	1990M11	23
Recession	1990M11	1991M6	7
Expansion	1991M6	1993M6	24
Recession	1993M6	1994M6	12
Expansion	1994M6	1998M3	45
Recession	1998M3	1999M8	17
Expansion	1999M8	2000M10	14
Recession	2000M10	2001M11	13
Expansion	2001M11	2008M2	75
Recession	2008M2	2009M3	13
Expansion	2009M3	2018M2	107
Recession	2018M2	2018M8	6
Expansion	2018M8	2020M8	24
Recession	2020M8	-	-

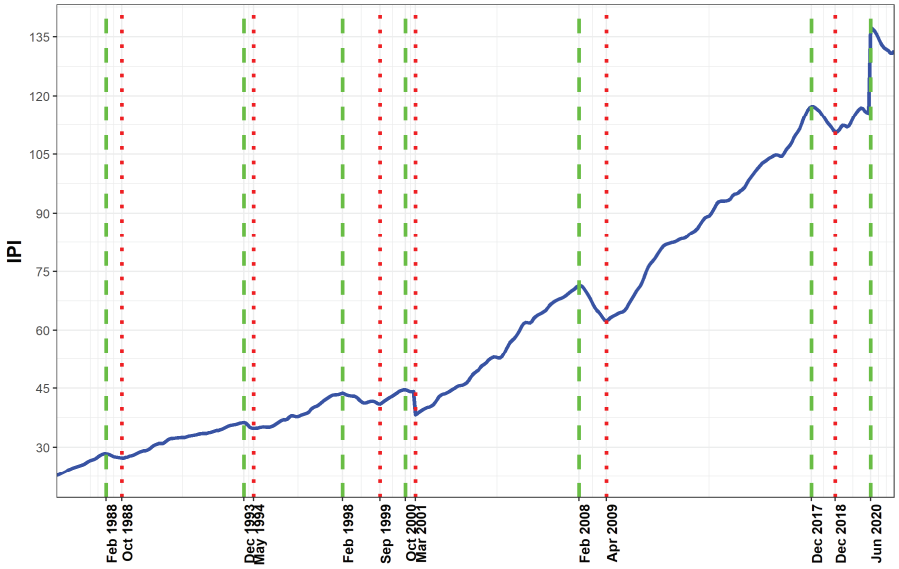
We found that the average expansion time is 44.6 months, and the average recession time is 11.3 months. The average expansion period is approximately four times longer than the average recession period. According to these findings, cycles vary between 30 and 120 months. The average cycle length obtained from both the peak and the trough was calculated as close to each other. The average cycle length for peaks is 55.85 months, while 59.33 months for troughs. We can say that the average duration of the cycle is about five years.

6.3. Monthly IPI

The peaks and troughs determined by the trend-cycle component of monthly IPI are shown in Figure 3, and the results of the cycles are shown in Table 3.

Peak and Trough dates based on monthly IPI

Figure 4



The results of monthly IPI data show that there are five complete business cycles were observed in terms of troughs. Unlike the monthly GDP results, the crisis in 1991 was not detected with this data. It is thought to be due to the scope of the data. Because the IPI series represents only a specific part of the economy, but GDP covers the entire economy. Therefore, the crisis may have affected sectors other than industry more deeply. In addition, the outflow of foreign capital in Turkey during the Gulf Crisis in 1991 caused a significant decrease in tourism revenues.

Business cycle chronology based on the monthly IPI

Table 3

Phase	Start	End	Duration
Expansion	-	1988M2	-
Recession	1988M2	1988M10	8
Expansion	1988M10	1993M12	62
Recession	1993M12	1994M5	5
Expansion	1994M5	1998M2	45
Recession	1998M2	1999M9	19
Expansion	1999M9	2000M10	13
Recession	2000M10	2001M3	5

Phase	Start	End	Duration
Expansion	2001M3	2008M2	83
Recession	2008M2	2009M4	14
Expansion	2009M4	2017M12	104
Recession	2017M12	2018M12	12
Expansion	2018M12	2020M6	18
Recession	2020M6	-	-

According to our findings, the average expansion time is 54.2 months, and the average recession time is 10.5 months. The average expansion period is approximately five times longer than the average recession period. Similar to the results obtained from the monthly GDP data, the range of cycles varies between the monthly IPI data varies between 32 and 118 months. The average cycle length in terms of peaks is 64.66 months, while 72.4 months in troughs. As a result, we can say that the average duration of the cycle is about 5.5 years.

7. DISCUSSION OF RESULTS

Average Business Cycle Duration

Table 4

Data	Full Cycles	Expansions	Recessions	Number of Cycles
Quarterly GDP	30 quarters	26.5 quarters	3.4 quarters	4
Monthly GDP	57.46 months	44.6 months	11.3 months	6
Monthly IPI	68.18 months	54.2 months	10.5 months	5

As can be seen from Table 4, different cycle lengths were obtained due to the analyses made with different data. Therefore, we need to assess which cycle length should be used. It is possible to make this evaluation from two different perspectives. The first perspective is the frequency of the series. Quarter-frequency data make it difficult to follow the timing of the movements of economic activity.

On the other hand, high-frequency (monthly) series are more advantageous in terms of the changes they exhibit, as they allow the detection of cycles to be revealed more clearly (Artis, 2002). The second perspective is how many cycles reflect the changes in the Turkish economy. According to our results, monthly GDP data is thought to reflect better the crisis periods that Turkey has experienced. Also, the monthly IPI series represents only a specific part of the economy, but GDP covers the entire economy. As a result of these evaluations, we found it appropriate to prefer the monthly GDP results, and we can say that the cycle length of Turkey is five years.

8. CONCLUSION

In this study, the contraction and expansion periods in the Turkish economy were examined with the Bry-Boschan algorithm, and the average cycle duration was tried to be determined. Quarterly GDP, monthly GDP, and monthly IPI data were used in this study, and different results were obtained regarding average cycle durations. The results obtained with three series used to determine the cycle duration for Turkey vary according to the frequency of the series and how much they reflect the changes in the economy. As a result of the evaluations, the 5-year cycle time obtained with the monthly GDP series was appropriate, considering that the Turkish economy better revealed the crises experienced in the past. It has been observed that this result is compatible with the related studies in the literature and the cycle characteristics of developing countries. It is thought that this study will make an essential contribution to the literature since the studies conducted in the national field are in the past, the data used covers a long period and is up-to-date.

REFERENCES

1. Alp H., Başkaya Y.S., Kılınç, M., Yüksel, C. (2011) "Türkiye için Hodrick-Prescott Filtresi Düzgünleştirme Parametresi Tahmini", TCMB Ekonomi Notları, Sayı: 2011-3.
2. Artis, M.J., Kontolemis, Z.G., Osborn, D.R. (1995), "Classical Business Cycles for G7 and European Countries," CEPR Discussion Paper No. 1137.
3. Artis,M.J. (2002). "Dating the business cycle in Britain", National Institute Economic Review, 182, 90-95.
4. Atabek, A., Coşar, E.E., Şahinöz, S. (2005). "A New Composite Leading Indicator for Turkish Economic Activity", Emerging Markets Finance and Trade, 41 (1), 45-64.
5. Banerji, A., Hiris, L. (2001). "A multidimensional framework for measuring business cycles". International Journal of Forecasting, 17, 333-348.
6. Boratav, K. (2000). "Dış Borca Yöneliş Bir Tuzak". In TOBB Ekonomik Forum Dergisi (Vol. 7, No. 2).
7. Bry, G., Boschan, C. (1971), "Cyclical Analysis of Time Series: Selected Procedures and Computer Programs," Technical Paper 20, NBER, Columbia University
8. Burns, A. F., & Mitchell, W. C. (1946). "Measuring business cycles" (No. burn46-1). National bureau of economic research.
9. Chow, G. C., Lin, A. L. (1971). "Best linear unbiased interpolation, distribution, and extrapolation of time series by related series". The review of Economics and Statistics, 372-375.
10. Christiano, L. J., Fitzgerald, T. J. (1998). "The business cycle: it's still a puzzle. Economic Perspectives", Federal Reserve Bank Of Chicago, 22, 56-83.
11. Christoffersen, P. (2000). "Dating the turning points of Nordic business cycles".
12. Demircan, C. (2018). "2001 ve 2008 Ekonomik Krizlerinde Türkiye'nin Kriz Yönetim Uygulamaları ve Karşılaştırması", (Master's thesis, İstanbul Gelişim Üniversitesi Sosyal Bilimler Enstitüsü).
13. Denton, F. T. (1971). "Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization". Journal of the american statistical association, 66(333), 99-102.
14. Einian, M. (2019). BCDating: Business Cycle Dating and Plotting Tools. R package version 0.9.8. https://CRAN.R-project.org/package=BCDating

15. Fernandez, R. B. (1981). “A methodological note on the estimation of time series”. The Review of Economics and Statistics, 63(3), 471-476.
16. Fırat, E., Demirtaş, C. (2012). “Konjonktürel Teoriler Işığında Türkiye’de Yaşanan 2000-2001 Krizinin Değerlendirilmesi”. Ekonomi Bilimleri Dergisi, 4(1), 23-32.
17. Gómez, V., Maravall Herrero, A. (1996). “Programs TRAMO and SEATS: instructions for the user (beta version: September 1996)”. Banco de España. Servicio de Estudios.
18. Harding, D., Pagan, A. (2002). "Dissecting The Cycle: A Methodological Investigation". Journal of Monetary Economics, 49(2). 365-381.
19. Kaba, G. (2019). “Türkiye’de 1980 Yılı Sonrası Yaşanan Ekonomik Krizler Ve Türk Siyasal Karar Birimlerinin Bu Krizlerle Mücadele Politikaları”, (Master's thesis, Dicle Üniversitesi Sosyal Bilimler Enstitüsü).
20. Klein, P. A., Moore, G. H. (1982). “The leading indicator approach to economic forecasting- retrospect and prospect”. NBER Working Paper(w0941).
21. Kydland, F. E., Prescott, E. C. (1990). “Business cycles: Real facts and a monetary myth”. Federal Reserve Bank of Minneapolis Quarterly Review, 14(2), 3-18.
22. Litterman, R. B. (1983). “A random walk, Markov model for the distribution of time series”. Journal of Business & Economic Statistics, 1(2), 169-173.
23. Luvsannyam, D., Batmunkh, K., Buyankhishig, K. (2019). “Dating the business cycle: Evidence from Mongolia”. Central Bank Review, 19(2), 59-66.
24. Moore, G. H. (1980). “Business cycles, inflation, and forecasting”. NBER Studies in Business Cycles, 24. Cambridge, Mass: Ballinger Publishing Company.
25. Moore, G. H., Zarnowitz, V. (1986). The Development and Role of the National Bureau of Economic Research Business Cycle Chronologies, in: Gordon, R. A., ed., The American Business Cycle: Continuity and Change, University of Chicago Press for NBER, Chicago.
26. Özatay, Fatih (2013), Parasal İktisat Kuram ve Politika, Ankara: Efil Yayınevi.
27. Özkan, İ. ve L. Erden (2007). “Türkiye Ekonomisinde İş Çevrimlerinin Tarih ve Süre Aralıklarının Tespiti” , Akdeniz İ.İ.B.F. Dergisi 14, sf. 1–19.
28. Pandey, R., Patnaik, I., Shah, A. (2017). “Dating business cycles in India”. Indian Growth and Development Review.
29. Parkin, M., Bade, R. (2015). Economics: Canada in the Global Environment. Pearson Canada.
30. Quartier-la-Tente, A., Michalek, A., Palate, J., Baeyens, R. (2021), RJDemetra: Interface to 'JDemetra+' Seasonal Adjustment Software. R package version 0.1.7. https://CRAN.R-project.org/package=RJDemetra
31. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
32. Rand, J., Tarp, F. (2002). Business cycles in developing countries: are they different? World development, 30(12), 2071-2088.
33. Savaş, V. F. (2012). Küresel finans ve makro İktisat. Efil Yayınevi.
34. Sax, C., Steiner, P. (2013). Temporal disaggregation of time series.
35. Schumpeter, J. A. (1939). Business cycles (Vol. 1). New York: McGraw-Hill.
36. Selçuk, F. (1994) “TÜSİAD öncü göstergeler indeksi”, Ekonomiyi İzleme Sempozyumu, H.Ü, Ankara
37. Stock J. H. , Watson M. W. (1989). “New Indices of Coincident and Leading Economic Indicators”, Macro Economics Annual, NBER (U.S.).

38. Susam, N., Bakkal, U. (2008). “Kriz Süreci Makro Değişkenleri ve 2009 Bütçe Büyüklüklerini Nasıl Etkileyecek?”. Maliye dergisi, (155), 72-88.
39. Şahin, Hüseyin (2014), Türkiye Ekonomisi, Bursa: Ezgi Kitapevi.
40. Toprak, M. (1996). Türk ekonomisinde yapısal dönüşümler 1980-1995. Turhan Kitabevi.
41. Yılmaz, Ö., Kızıltan, A., Vedat, K. A. Y. A. (2005). “İktisadi Kriz Kuramları, Finansal Küreselleşme Ve Para Krizleri”. Erciyes Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi, (24), 77-97.
42. Zarnowitz, V., National Bureau of Economic, R. (1987). “The Regularity of Business Cycles”. National Bureau of Economic Research.

Selective Editing Using Contamination Model

Ieva Burakauskaitė (ieva.burakauskaite@stat.gov.lt)
Statistics Lithuania,

Vilma Nekrašaitė-Liegė (vilma.nekrasaite-liege@stat.gov.lt, vilma.nekrasaite-liege@vilniustech.lt)
Statistics Lithuania, Vilnius Gediminas Technical University

ABSTRACT

Results of an outlier detection study with a focus on selective editing are presented in the paper. The aim of selective editing is to identify observations affected by errors that have a major impact on the quality of sample estimates. This way the data editing process can be focused on the corresponding observations therefore allocating excess human resources and reducing time costs though maintaining the quality of sample estimates. These objectives are especially important for national statistical institutions such as Statistics Lithuania seeking to optimize the data editing process.

A few different versions of selective editing were applied to the data editing process of the quarterly statistical survey on service enterprises (turnover indicator) of Statistics Lithuania. Predictions of the target variable were obtained using the contamination model. An impact of a potential error on a sample estimate was evaluated using a score function with a standard structure – a difference between the observed value of the target variable and its prediction multiplied by a sample weight and a suspicion component. Two types of the suspicion component (discrete and continuous) were used and an impact of the suspicion component on the effectiveness of selective editing was investigated. Efficiency of the continuous suspicion component supported its advantage over the discrete suspicion component, and therefore turned out to be a major factor in optimizing the data editing process.

Keywords: *selective editing; contamination model; data validation; statistical survey; official statistics.*

1. Introduction

An appropriate accuracy of sample estimates is one of the most important objectives to be achieved using sampling methods in official statistics. The quality of statistical data together with the sampling strategy (a sampling plan and an estimator) have a major impact on the accuracy of sample estimates. Commonly, an erroneous part of statistical data is unknown and therefore may only be detected by either logical or mathematical methods using some kind of known additional information. Previously the data editing process was usually focused on editing all of the detected errors. However, according to various studies, in order to achieve the desired accuracy of sample estimates, it is unnecessary to edit all errors. The main idea of selective editing is to identify and sort errors according to the influence they have on sample estimates (Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000). For this purpose, a score function is constructed that portrays the influence possibly erroneous observation has on sample estimates (Latouche and Berthelot, 1992; Hedlin, 2003). As the error detection procedure is usually carried out before the calculation of sample estimates, the best versatile selective editing model for identifying only the most influential part of erroneous data has to be chosen in advance.

Although selective editing is already implemented as one of the methods for outlier detection and data validation procedure in Statistics Lithuania, it still remains an important, uncommon topic for research in Lithuania. The currently used selective editing model identifies a high number of outliers as a result of its similarity to the default selective editing model accessible through the package "SeleMix" of the programming language R. In order to adopt a more suitable selective editing model for a specific statistical survey of Statistics Lithuania, an outlier detection study was carried out. Results of the latter study have been briefly presented at Summer School on Survey Statistics 2021 (Burakauskaitė and Nekrašaitė-Liegė, 2021) and The Use of R in Official Statistics (uRos2021) conferences. This paper provides a closer look into two cases of the carried out study, and indicates the main features that have to be taken into consideration while choosing the most suitable selective editing model.

Section 2 of the paper introduces the contamination model and the selective editing method that form a base for the practical study of the outlier detection. Section 3 presents the study that was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. During the study some randomly selected values of statistical data were replaced with errors. The detection of randomly introduced errors was then carried out using a few cases of selective editing. The comparison of results as well as its summary are presented in Section 4. Calculations were carried out using the statistical programming language R and its package "SeleMix" that has been designed to execute the selective editing method (Guarnera and Buglielli, 2013).

2. Methodology on Selective Editing

2.1 Contamination Model

Suppose that true (unobserved) data are independent realizations of p-variate random vectors $\mathbf{Y}_i^* = (\mathbf{Y}_{i1}^*, \dots, \mathbf{Y}_{ip}^*)'$, $i = 1, \dots, n$, with a Gaussian distribution with mean vectors $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip})'$ and a common covariance matrix $\boldsymbol{\Sigma}$. Also, a set of q covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})'$ exists for every sampled unit i and $\boldsymbol{\mu}_i = \mathbf{B}'\boldsymbol{x}_i$ where $\mathbf{B}$ is a $q \times p$ matrix of unknown coefficients (Di Zio and Guarnera, 2013). The corresponding true data model can be expressed as

$$\mathbf{Y}^* = \mathbf{XB} + \mathbf{U} \qquad [1]$$

where $\mathbf{Y}^*$ is the $n \times p$ true data matrix, $\mathbf{X}$ – $n \times q$ covariate matrix and $\mathbf{U}$ – $n \times p$ matrix of normal residuals. Rows of the matrix $\mathbf{U}$ are independent realizations of Gaussian random vectors with mean equal to $\mathbf{0}$ and a covariance matrix $\mathbf{\Sigma}$.
Generic marginal probability distributions of the ith sampled unit of matrices $\mathbf{Y}^*$ (true data) and $\mathbf{U}$ (residuals) are denoted as

$$f(\boldsymbol{y}_i^*) = N(\boldsymbol{y}_i^*;\ \boldsymbol{\mu}_i, \mathbf{\Sigma}),\ \ f(\boldsymbol{u}_i) = N(\boldsymbol{u}_i; \mathbf{0}, \mathbf{\Sigma}),\ \ i = 1, \dots, n. \tag{2}$$

In general, form $N(\boldsymbol{y};\ \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a marginal probability distribution of the p-variate random vector $\mathbf{Y}$ with mean equal to $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.
It is assumed that the presence of errors in data are described by independent Bernoulli random variables. Therefore, the observed (erroneous) data can be expressed as

$$\mathbf{Y} = \mathbf{Y}^* + \mathbf{I}\boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{I}$ is a diagonal $n \times n$ matrix with its diagonal elements equal to Bernoullian variables $I_1, \dots, I_n$ ($I_i = 1$ if the corresponding sampled unit is erroneous, and $I_i = 0$ otherwise, $i = 1, \dots, n$). A marginal probability distribution of the p-variate random vector $\boldsymbol{\epsilon}_i$ (random noise) can be expressed as

$$f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \mathbf{0}, \mathbf{\Sigma}_\epsilon),\ \ \mathbf{\Sigma}_\epsilon = (\alpha - 1)\mathbf{\Sigma}, \tag{4}$$

with a numeric constant $\alpha > 1$.
$f(\boldsymbol{y}|\boldsymbol{y}^*)$ denotes a conditional marginal probability distribution of random variables $\mathbf{Y}$ and $\mathbf{Y}^*$. Therefore, [3] can be expressed equivalently:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1 - \pi)\delta(\boldsymbol{y} - \boldsymbol{y}^*) + \pi N(\boldsymbol{y};\ \boldsymbol{y}^*, \mathbf{\Sigma}_\epsilon) \tag{5}$$

where π is "a priori" probability of contamination and $\delta(\boldsymbol{y} - \boldsymbol{y}^*)$ is the Dirac delta function with mass at $\boldsymbol{y}^*$.
Furthermore, a marginal probability distribution of the observed data can be expressed as

$$\begin{aligned} f(\boldsymbol{y}_i) &= \int_0^\infty f(\boldsymbol{y}_i, \boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^* \\ &= \int_0^\infty f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^* \\ &= (1 - \pi) N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \mathbf{\Sigma}) + \pi N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \alpha\mathbf{\Sigma}). \end{aligned} \tag{6}$$

Coefficients of the latter observed data model can be obtained by the maximum likelihood estimation.

2.2 Selective Editing

Selective editing is based on the comparison between the observed data and predictions of the true (unobserved) data. The latter can be obtained from a conditional marginal probability distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ (Di Zio and Guarnera, 2013). An application of the Bayes formula provides:

$$\begin{aligned} f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) &= \frac{f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)}{\int_0^\infty f(\boldsymbol{y}_i^*) f(\boldsymbol{y}_i|\boldsymbol{y}_i^*)\, d\boldsymbol{y}_i^*} \\ &= \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i) N(\boldsymbol{y}_i^*;\ \tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{\Sigma}}) \end{aligned} \tag{7}$$

where

$$\tilde{\boldsymbol{\mu}}_i = \frac{y_i + (\alpha - 1)\boldsymbol{\mu}_i}{\alpha},$$

$$\widetilde{\boldsymbol{\Sigma}} = \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Sigma},$$

$\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i)$ is the Dirac delta function with mass at $\boldsymbol{y}_i$, $\tau_1(\boldsymbol{y}_i)$ and $\tau_2(\boldsymbol{y}_i)$ are posterior probabilities that the ith sampled unit with observed values $\boldsymbol{y}_i$, $i = 1, \dots, n$, is not erroneous and that it is contaminated respectively:

$$\tau_1(\boldsymbol{y}_i) = P(\boldsymbol{y}_i = \boldsymbol{y}_i^* | \boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i;\ \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})},$$
$$\tau_2(\boldsymbol{y}_i) = P(\boldsymbol{y}_i \neq \boldsymbol{y}_i^* | \boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i). \quad [8]$$

Posterior probabilities [8] are defined in terms of the conditional expected value $\widetilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$, $i = 1, \dots, n$. Therefore, the expected error can be defined as

$$\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_i = \tau_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \widetilde{\boldsymbol{\mu}}_i). \quad [9]$$

In practice, [9] is usually applied by using maximum likelihood estimates instead of the corresponding true data values.

2.2.1 Definition of the Score Function

Hereinafter $\hat{p}$ denotes a maximum likelihood estimate of some parameter p.

Suppose one seeks to estimate a sum of the variable Y_j^*, $j = 1, \dots, p$, with a sampling weight w_i of the ith sampled unit, – $T_j^* = \sum_{i=1}^n w_i y_{ij}^*$. A ratio between the expected error [9] with a sampling weight w_i multiplied by a suspicion component s_{ij} (probability that the ith sampled unit is erroneous) and the target parameter estimate $\hat{T}_j = \sum_{i=1}^n w_i \hat{y}_{ij}$ represents a conditional error of the ith sampled unit:

$$r_{ij} = \frac{s_{ij} w_i (y_{ij} - \hat{y}_{ij})}{\hat{T}_j}. \quad [10]$$

A local score function for the variable Y_j is denoted as $S_{ij} = |r_{ij}|$. Separate local scores can be combined into one global score GS_i in a few different ways:

$$GS_i = \max_j S_{ij} \quad \text{or}$$
$$GS_i = \sum_j S_{ij}.$$

In order to identify an optimal number of observations to be edited, the corresponding sampled units are sorted in descending order according to the GS_i. First $\tilde{k}$ observations are then chosen for the editing procedure:

$$\tilde{k} = \min\left\{k^* \in 1, \dots, n \,\middle|\, \max_j R_{kj} < \eta,\ \ \forall\, k > k^*\right\} \quad [11]$$

where $R_{ij} = \left|\sum_{k \geq i}^n r_{kj}\right|$ with an accuracy level η.

The suspicion component s_{ij} might have a discrete form (e.g., $s_{ij} \in \{0, 1\}$) or a continuous form ($s_{ij} \in [0, 1]$). In the paper the latter continuous suspicion component is defined according to Norberg et al. (2010). An additional test variable should be defined prior to defining the suspicion component:

Definition 1 (Test variable) *Test variable* $\boldsymbol{t}$ can be a combination of variables from a statistical survey and (or) some additional information. Statistical errors might then be identified by

checking whether a value of the test variable $\boldsymbol{t}_{j'}$, $j' = 1, \dots, p'$, for the ith sampled unit falls into a chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$.

The above-mentioned statistical error might be an observation that stands out compared to the rest of observations in the corresponding data set, to observations of a previous round of the same statistical survey or to some other additional information (e.g., administrative data). A few examples of a non-statistical error might be inconsistent answers the same respondent provides to the same question over different periods of time (e.g., a variable does not equal to the sum of its summands), disallowed values (e.g., observations that do not fall into a previously defined quantitative interval), item non-response, etc.

Definition 2 (Discrete suspicion component) *Discrete suspicion component*

1. $s_{ij} = 1$ if a value of the jth ($j = 1, \dots, p$) survey variable of the ith ($i = 1, \dots, n$) sampled unit y_{ij} is a non-statistical error;
2. $s_{ij'} = 1$ if a value of the j'th ($j' = 1, \dots, p'$) test variable of the ith sampled unit $t_{ij'}$ is a statistical error, i.e., $t_{ij'} \notin \left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$. In this case, $s_{ij} = 1$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$;
3. $s_{ij} = 0$ otherwise.

Nonetheless, it is important to take into consideration a different distance between observations that do not fall into the chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$ and the corresponding bounds of the region. The continuous suspicion component might convey the information on the latter distance more effectively.

Definition 3 (Continuous suspicion component) *Continuous suspicion component*

1. $s_{ij} = 1$ if a value of the jth ($j = 1, \dots, p$) survey variable of the ith ($i = 1, \dots, n$) sampled unit y_{ij} is a non-statistical error;
2. $\tilde{s}_{ij'} = \frac{\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) - t_{ij'}}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} < \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right)$;
3. $\tilde{s}_{ij'} = \frac{t_{ij'} - \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} > \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$;
4. $\tilde{s}_{ij'} = 0$ if $\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) < t_{ij'} < \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$.

The continuous suspicion component then equals to $s_{ij'} = \tilde{s}_{ij'} / \left(\tau + \tilde{s}_{ij'}\right)$ with parameters $\kappa \geq 0, \alpha > 0$ and $\tau > 0$ that regulate the size of the acceptance region. $s_{ij} = \max_{j'} s_{ij'}$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$.

3. Results of the Outlier Detection Study

The outlier detection study was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. The purpose of the statistical survey is to prepare and publish statistical information on the sales income (turnover) and their indices of service enterprises and provide the data for users on short term statistics. Enterprise turnover[1] of the accounting period was the target variable of the study.

In order to obtain the most suitable yet flexible enough selective editing model, four different predictor variables were chosen for comparison purposes – turnover from value-added tax (hereinafter referred to as VAT) declarations, turnover from the quarterly F-01 questionnaire[2], average number of employees and total hours worked. It is known that the first two predictor variables (turnover indicators) tend to have a high correlation with the target variable. On the contrary, the last two predictor variables (labour statistics indicators) usually have a lower

correlation with the target variable of the study. The latter characteristic will be explored more in the following part of the paper. Every chosen predictor variable was used separately providing four data sets and therefore four different cases of selective editing application.
For simplicity, the outlier detection study was carried out using only units with non-missing values that are greater than 0 for both the target variable and the corresponding predictor variable. In practice, an acceptance region, e.g., $(Q_1 - 3IQR, Q_3 + 3IQR)$ with the first quartile Q_1, the third quartile Q_3 and an interquartile range IQR, is usually constructed for the outlier detection in other units. As four data sets contain different number of non-missing values that are greater than 0, the number of observations left for the further study varies. The corresponding number of observations in data sets (primary populations) according to the predictor variable is given in Table 1 below.

Table 1: Number of observations in statistical data sets

Predictor variable	**Number of observations**
Turnover from VAT declarations	4085
Turnover from the quarterly F-01 questionnaire	574
Average number of employees	4867
Total hours worked	4931

The data contamination process was an important part of the outlier detection study as it contributed to selecting a level of accuracy for the further study (see
Table ***2***), and finding the most suitable outlier detection procedure given the chosen accuracy. In order to control the data contamination process, influential outliers in primary populations had to be replaced with plausible values. Therefore, default outlier detection procedure was performed using the statistical programming language R and its package "SeleMix", and the detected outliers in primary populations were replaced with contamination model predictions. The following algorithm was then applied to every modified primary population:

1. The target variable was contaminated in 3 different ways:
    a. 1.5 percent of observations were multiplied by 100,
    b. 2 percent of observations were trimmed leaving only the first and the last digits,
    c. 20000000 was added to 1.5 percent of observations;
2. Estimation of model coefficients and outlier (potential error) detection were carried out using the statistical programming language R and its package "SeleMix" (function "ml.est");
3. Values of the target variable were sorted in descending order according to estimates of the global score function obtained using the statistical programming language R and its package "SeleMix" (function "sel.edit"). An estimate of the global score function is close to 0 when a value of the target variable is not identified as an outlier and therefore has no major impact on the quality of sample estimates, and greater than 0 when a value of the target variable is identified as an outlier;
4. The part of outliers that have a major impact on the quality of sample estimates (influential errors) was chosen for the editing procedure.

The latter influential error detection procedure was repeated in two different ways – by calculating estimates of the score function (1) with a discrete suspicion component that stays the same among all observations ($s_i = 1$), and (2) with a continuous suspicion component. The latter suspicion component was designed using an acceptance region between the first and the third quartiles $\left(\hat{t}^{(L)}, \hat{t}^{(U)}\right)$ where $\hat{t}_i = \hat{y}_i \ \ (i = 1, \dots, n)$. Parameters κ and τ varies ($\kappa \in \{0, 0.5, 1, 1.5\}$, $\tau \in \{0.1, 0.5, 1, 1.5, 2\}$) and specific values are chosen depending on the lowest number of identified influential errors, $\alpha = 0.05$.

Selective editing with different levels of accuracy gives a different number of influential errors. If all of the detected influential errors were introduced by the above-mentioned data contamination procedure, the corresponding level of accuracy was chosen for the further study (see
Table ***2*** below).

Table 2: Levels of accuracy (threshold values) for statistical data sets

Predictor variable	**Level of accuracy**
Turnover from VAT declarations	0.011
Turnover from the quarterly F-01 questionnaire	0.004
Average number of employees	0.027
Total hours worked	0.026

As it can be observed from the above-given
Table ***2*****:** Levels of accuracy (threshold values) for statistical data sets, levels of accuracy fall into two groups – below 0.02 and above 0.02. It gives an insight that selective editing models with either turnover indicator as a predictor variable might perform better and result in higher quality sample estimates. Whereas labour statistics indicators chosen as predictors might provide lower quality selective editing models. The same two groups of predictor variables were mentioned earlier in the paper while comparing the correlation with the target variable of the study. Having the four specific statistical data sets, correlation coefficients between predictor and target variables may be calculated easily (see Table 3 below). Correlation coefficients separate predictor variables into the same two groups – turnover indicators have a higher correlation with the target variable (both above 0.9) while labour statistics indicators tend to have a lower correlation with the target variable (both below 0.6).

Table 3: Correlation coefficients between target and predictor variables

Predictor variable	**Correlation coefficient**
Turnover from VAT declarations	0.96
Turnover from the quarterly F-01 questionnaire	0.97
Average number of employees	0.52
Total hours worked	0.52

Keeping in mind the chosen levels of accuracy (see
Table ***2***), the results of different selective editing approaches were then compared by estimating the relative absolute bias after every sequential edit of each influential error. The latter procedure helped determining a number of influential errors that have to be edited in order to achieve desired levels of accuracy. The combined results are provided in Table 4 below.

Table 4: Number of influential errors in statistical data sets

Predictor variable	**Total number of influential errors**	**Number of influential errors to be edited**
(1) Selective editing with the discrete suspicion component		
Turnover from VAT declarations	134	92
Turnover from the quarterly F-01 questionnaire	23	14
Average number of employees	90	>90
Total hours worked	111	>111
(2) Selective editing with the continuous suspicion component		

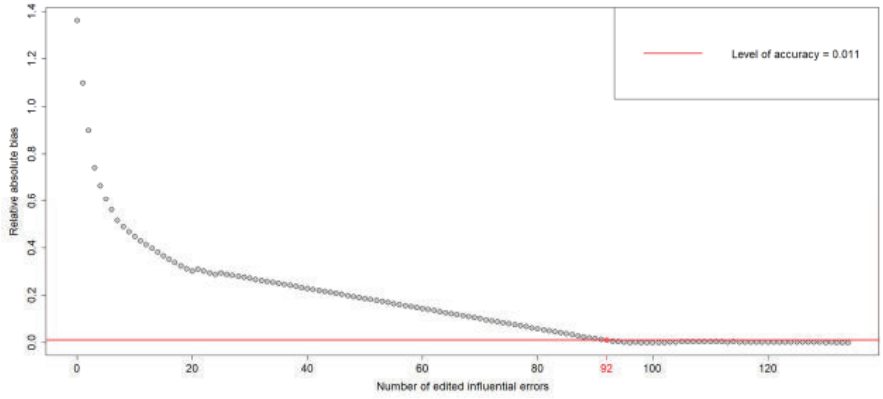
Turnover from VAT declarations	93	92
Turnover from the quarterly F-01 questionnaire	15	14
Average number of employees	136	121
Total hours worked	124	123

Results observed applying selective editing with different approaches may be compared in two different ways – the impact of the discrete / continuous suspicion component (Subsection 3.1) and the impact of the predictor variable (Subsection 3.2).

3.1 The Impact of the Suspicion Component

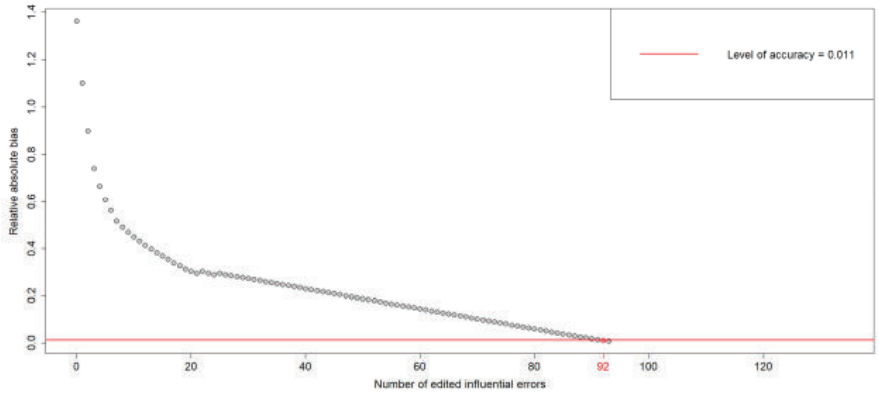
Consider the case when a predictor variable is turnover from VAT declarations. As observed in Table 4, a total number of influential errors differ significantly depending on a type of the suspicion component.
When the discrete suspicion component is used in the selective editing model, the same value of suspicion is used for every observation ($s_i = 1$). Therefore, no additional impact is put on determining whether the corresponding observation is an influential error or not. Although a total of 134 errors are identified as influential, the relative absolute bias calculation shows that only 92 of the identified influential errors have to be edited in order to achieve the desired level of accuracy (0.011), see Figure 1 below.

Figure 1: Relative absolute bias dependency on the number of edited influential errors using the discrete suspicion component with turnover from VAT declarations as the predictor variable



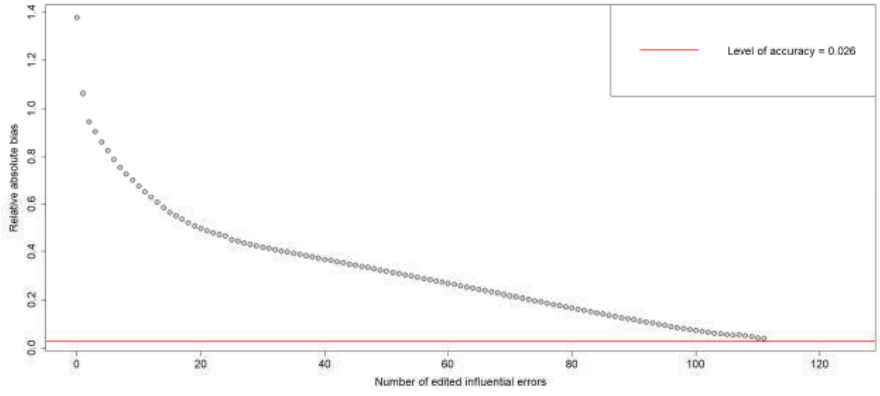
On the contrary, the use of the continuous suspicion component lets to take into consideration distances between observations that do not fall into the chosen acceptance region and the corresponding bounds of the region. With this additional impact on the selective editing model, the calculation of relative absolute bias shows that almost every identified influential error (92 out of 93) has to be edited in order to achieve the desired level of accuracy, see Figure 2 below. The latter example illustrates an advantage of the continuous suspicion component over the discrete suspicion component. In practice, the use of the continuous suspicion component could prevent statisticians from the overediting of data while preserving the preferable quality of sample estimates.

Figure 2: Relative absolute bias dependency on the number of edited influential errors using the continuous suspicion component with turnover from VAT declarations as the predictor variable



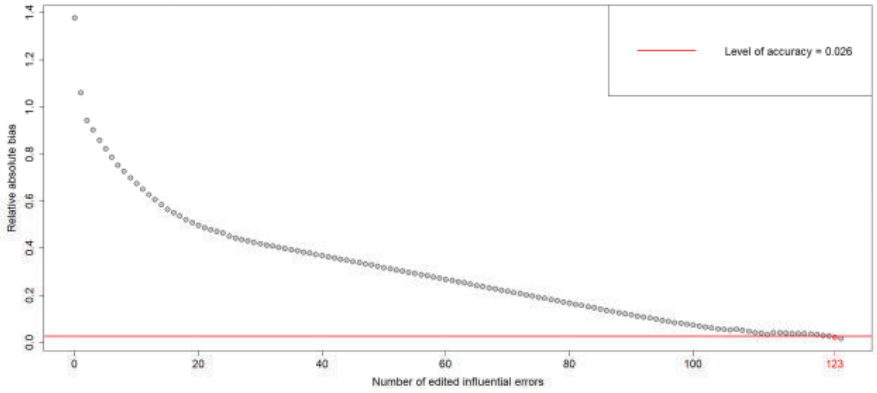
Now consider the case when a predictor variable is one of labour statistics indicators – total hours worked. Table 4 indicates a significant difference between a total number of influential errors while applying selective editing with different types of the suspicion component.
In the case when the discrete suspicion component is used, relative absolute bias calculation demonstrates that editing all the identified 111 influential errors does not result in sample estimates with the preferable level of accuracy (0.026), see Figure 3 below.

Figure 3: Relative absolute bias dependency on the number of edited influential errors using the discrete suspicion component with total hours worked as the predictor variable



Similar to the previously seen example with turnover from VAT declarations as the predictor variable, the use of the continuous suspicion component increases the chance of achieving the preferable level of accuracy. The calculation of relative absolute bias shows the need to edit almost every identified influential error (123 out of 124), see Figure 4 below.

Figure 4: Relative absolute bias dependency on the number of edited influential errors using the continuous suspicion component with total hours worked as the predictor variable



As it was illustrated in Table 4 and Figure 1 through Figure 4, applications of selective editing with different predictor variables have shown an effectiveness of the continuous suspicion component on the outlier detection procedure. This approach to selective editing lets to identify only the most important influential errors, and therefore prevents from the overediting or, in other cases, insufficient data editing.

3.2 The Impact of the Predictor Variable

Another important factor in choosing the most suitable selective editing model besides the suspicion component is the predictor variable. As it was previously noted in the paper, a weak dependency between the predictor variable and the target variable of the study might be considered as a drawback of a selective editing model. The latter model would result in a lower level of accuracy and therefore lower quality of sample estimates. Moreover, the use of the discrete suspicion component might be inconsistent with the preference to achieve the previously chosen level of accuracy. As it was shown in Table 4 and Figure 3, a previously chosen, i.e., "desired", level of accuracy was not achieved using the discrete suspicion component when one of labour statistics indicators was the predictor variable.
In contrast, a strong dependency between the predictor variable and the target variable of the study results in a better selective editing model. A previously chosen level of accuracy may be achieved using both the discrete and the continuous suspicion components. However, as it was illustrated in Subsection 3.1, the continuous suspicion component might be a more efficient choice.

4. Conclusions

After calculations of the relative absolute bias dependency on the number of edited influential errors, selective editing with the continuous suspicion component was determined to be an optimal method for the outlier detection procedure. The latter version of selective editing prevents from the overediting or, in some cases, insufficient statistical data editing.
Turnover from VAT declarations and turnover from the quarterly F-01 questionnaire were identified as the most suitable predictor variables for the outlier detection procedure. The main property of a suitable predictor variable turned out to be a high correlation between the latter predictor variable and the target variable of the study.

Main findings of the outlier detection study contributed to the outlier detection procedure currently implemented at Statistics Lithuania. The construct of the continuous suspicion component lets to improve the existing selective editing model and focus statistical data editing on the most important influential errors while preserving the quality of sample estimates.
In the paper, all four cases of selective editing were modelled using undivided data sets. A possible extension of the carried out study could focus on the search of the most suitable selective editing model when data sets are separated into groups according to some factor, e.g., the economic activity of enterprises.

5. References

Burakauskaitė, I. and Nekrašaitė-Liegė, V. (2021), "Selective Editing Using Contamination Model", SUMMER SCHOOL ON SURVEY STATISTICS 2021, BNU Network on Survey Statistics, Statistics Lithuania, Vilnius, Lithuania, 40–45.
Di Zio, M. and Guarnera, U. (2013), "A Contamination Model for Selective Editing", Journal of Official Statistics, 29, 4, 539–555.
Guarnera, U. and Buglielli, M. T. (2013), "SeleMix: an R Package for Selective Editing", Italian National Institute of Statistics, Rome, Italy, 2013-12-12.
Hedlin, D. (2003), "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics", Journal of Official Statistics, 19, 177–199.
Latouche, M. and Berthelot, J. M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys", Journal of Official Statistics, 8, 389–400.
Lawrence, D. and McDavitt, C. (1994), "Significance Editing in the Australian Survey of Average Weekly Earnings", Journal of Official Statistics, 10, 437–447.
Lawrence, D. and McKenzie, R. (2000), "The General Application of Significance Editing", Journal of Official Statistics, 16, 243–253.
Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P. and Nordberg, L. (2010), "A General Methodology for Selective Data Editing", Statistics Sweden, Stockholm, Sweden, 2010-02-04.
Official Statistics Portal (2021), "The sales income (turnover) and indices of servise enterprises", Metadata, Statistics Lithuania, Vilnius, Lithuania, 2021-11-26, retrieved from https://osp.stat.gov.lt/documents/10180/5118910/Paslaug%C5%B3+%C4%AFmoni%C5%B3+rodikliai+%5BEN%5D+636.html/62ebe741-2e13-4398-8630-34f666003eb0, accessed 2022-01-14.
Official Statistics Portal (2022), "Financial indicators of enterprises", Metadata, Statistics Lithuania, Vilnius, Lithuania, 2022-01-05, retrieved from https://www.stat.gov.lt/documents/10180/5118910/%C4%AEmoni%C5%B3+finansiniai+rodikliai+%28AB%2C+UAB%2C+V%C4%AE%2C+S%C4%AE+ir+kiti%2C+i%C5%A1skyrus+I%C4%AE+ir+fizinius+asmenis%29+%5BEN%5D+5101.html, accessed 2022-01-14.

[1] *Enterprise sales income (turnover)* refers to income from selling goods and / or providing services received by an economic entity in the reporting period (VAT excluded). Income from selling long-term fixed assets, financial and investment activity, dividend, etc., as well as funding from the budget, are excluded. The services provided do not include the acquisition value of customer's materials, products, spare and component parts (Official Statistics Portal, 2021).
[2] The quarterly F-01 questionnaire was a tool for collecting statistical data for the quarterly statistical survey on financial indicators of enterprises. The objective of statistical information was to provide for users quarterly statistical information on the performance and financial indicators (employees, income, expenses, profit, equity, liabilities, assets, etc.) and efficiency ratios (profitability, liquidity, turnover, etc.) of non-financial corporations (Official Statistics Portal, 2022).

Reshaping jobs in healthcare sector based on digital transformation

Bunduchi Elena
Institute of National Economy, Romanian Academy, Romania; „G.E.Palade" University of Medicine, Pharmacy, Sciences and Technology of Târgu Mureş, Romania

Vasile Valentina
Institute of National Economy, Romanian Academy, Romania

Ștefan Daniel
Institute of National Economy, Romanian Academy, Romania; „G.E.Palade" University of Medicine, Pharmacy, Sciences and Technology of Târgu Mureş, Romania

Comes Călin-Adrian
Institute of National Economy, Romanian Academy, Romania; „G.E.Palade" University of Medicine, Pharmacy, Sciences and Technology of Târgu Mureş, Romania

ABSTRACT

Digitization in the health sector is unequally distributed by activities and specializations, but it remains a trend that will change the employment model, with jobs disruption and infusion of financial capital and associated technologies. The facilities offered by digitalization not only offer solutions to adapt medical services to the challenges / restrictions of Covid-19 but also offer multiple possibilities to access expert services or reduce waiting times on the value chain of services, allowing to increase quality in perspective. patient-centered treatments. The aim of the research is to identify to what extent factors such as the level of economic development, the financing of the health sector and the external mobility of specialists influence the digital reform in the health sector. The results confirm the significant influence of the level of economic development and health spending on the potential for digitization of jobs in the health sector. It also highlights that e-health services have a reverse impact on the migration of doctors.

Keywords*: health, digital reform, jobs, migrant health worker*

INTRODUCTION

The impact of digitalization is felt in all spheres of life, from processing industries (Çela et al., 2021; Herman, 2020) to services sector with activities like education (Timus et al., 2020), health and culture. Practically, digitalization is transforming our lives as we knew it before. Although many industries, such as retail, tourism, and entertainment have begun to grow with the advent of the Internet, health is a sector in which digitalization has failed to exert rapid

influence due to financial, technological, and legislative, but also cultural, differences. For example, until 2018, teleconsultations in Germany were not allowed (Olesch, 2021), unlike in Switzerland, where remote treatment has been widely used for more than seventeen years (Behringer, 2018). However, healthcare is currently undergoing a digital revolution. Big Data, mobile devices, surgeries assisted by artificial intelligence (AI) and other innovations, including in health technological processes management are opening up new frontiers in medicine, a trend that has become especially visible during the COVID-19 pandemic. According to some studies (Siemens Healthineers, 2021), as a result of the COVID-19 pandemic, more and more medical institutions are investing in digital tools and AI. At the same time, they find that most of the investments have been made in equipment and programs that contribute to improving communication between medical staff and patients. Smartphone applications are tracking the spread of the virus, provide preliminary recommendation for treatment route and alternatives for access the proper services and AI as facilitator for better and faster diagnosis is helping doctors to increase the quality and efficiency of treatment.

The connections between patients and their caregivers can be secure, easier, and faster due to the use of technologies and digital tools, which aim to reduce physical distances and can give sometimes better result than usual face-to face medical services (Oliviera Hashiguchi, 2020). Consequently, the relational mechanism between patients and medical staff is starting to be based more and more on the concept of limiting face-to-face interaction with patients and faster communication through digital platforms. Monitoring the health of patients at home through telemedicine technology will allow them to benefit from diagnosis and quality outpatient treatments while they are at home, without the need for their physical presence in hospital units (Sundberg et al., 2015). This shift to home care is only possible with the help of digital decision support tools, including big data on similar cases, that can identify the optimal approach for treatment model and the right people/cases for outpatient home care. Through distance and telehealth care, patients will gain more transparency in their own care and become more active and aware of their own health and symptoms and take needed measures to prevent diseases (Player et al., 2018). The use of digital mHealth tools promotes patient empowerment, while enhancing the connection between patients and health personnel (Qudah & Luetsch, 2019). In addition, telemedicine has a significant impact on people with disabilities and mental disorders (Toquero, 2021; Groyer & Campbell, 2018). Thus, patient monitoring and symptom surveillance can be performed remotely and treatment can be much more beneficial (Berrouiguet et al., 2016), with patients finding themselves in a familiar environment contributing to a positive response to the treatment. Farrell (2016) illustrates the use of

smartphones by health personnel in acute care and it has shown that its are a very important digital tool used in communication between patients, nurses, and other healthcare professionals.

In addition to the benefits of digital communication between patients with various diseases and medical staff, it improves access to health services for people in rural or remote areas, especially due to the fact that worldwide 43.85% of the world's population still lives in countryside (World Bank, 2021). van Dis (2002) results show that people in rural areas are more prone to health problems, such as poor dental hygiene, high risk of mental illness, chronic illness, substance abuse, alcohol, and tobacco use, all of which lead to poor general health and low quality of life. Health status disparities for people in rural areas compared to urban ones are due to lack of access to health services. One way to address the issue of access to health care for people in those areas, but also for people from urban area would be to implement telemedicine and communication between patients and healthcare professionals through digital tools. These services would allow early diagnoses, prescribing treatments and disease prevention actions, without the need to travel to medical units, which are sometimes too far from home, or if those units do not have enough qualified staff.

Besides the impact that digitalization has on the relations between patients and medical staff, the technology has made it easier for the health staff to communicate between them beyond geographic borders, being possible to organize teleconferences, webinars, and videos, or even to perform operations or other medical treatments / interventions together, saving time with the physical movement of specialist doctors.

However, it is important to carefully address patients' distrust of digital tools and to explain their benefits, without creating communication barriers between patients and healthcare professionals, both digital and face-to-face. Studies show that although telemedicine contributes to better patient – doctor communication, some patients are not satisfied with these services due to the fact that they do not have adequate training on the use of digital tools, they do not have a high level of digital readiness, problems with internet connection or even concerns about the confidentiality of data and discussions with medical staff (Bagchi et al., 2018; Parker et al., 2018). Even if there are people who would prefer face-to-face meetings with their doctors, the long waiting time makes them finally to accept online consultations (Collins et al., 2004).

The problems identified by patients are not the only barriers, the medical staff is not fully prepared for the digitization of medicine as well. For example, in Romania it can be difficult to digitize medical services, especially due to the fact that many young medical

graduates have migrated to practice abroad and the remaining doctors are older and have difficulties in relation to new technologies, a problem identified and by Reßing (et al., 2018).

Besides improving communication and connectivity between patients and healthcare professionals, technological development and the digitization process have a major impact on changing the content of services and the use of Artificial Intelligence in health sector becomes a common denominator for health care model transformation. Official data (Statista, 2021) show that in 2020, medical institutions that used AI and have automation strategies reached over 90% compared to 53% in 2019. AI is used both to improve management for some categories of services and for several professions or fields of medicine, and the incidence of digitization is different. The most common use of AI is in laboratory medical services, which aims to reduce the number of errors in diagnosing patients, compare results with AI-database query, optimize AI decision and AI assistance in interpreting laboratory results (Apostu et al., 2021). Digitalization and AI have a significant impact not only in medical services, but in medical intervention and treatment as well. Studies revealed that AI can be successfully implemented from preoperative planning (Hashimoto, Rosman, Rus, & Meireles, 2018) and guidance operations during surgery to its integration into surgical robots (Zhou, Guo, Shen, & Yang, 2020), intended to assist physicians in the first instance, and to replace them in the easiest surgeries in the future, all for the highest quality patient care. At the same time Perkins (et al., 2020) observed that AI improve the ability of surgeons to decide the need for an acute surgery. As other example, in the field of ophthalmology, AI technology has the ability to detect cataracts or other eye diseases early, and AI laser-assisted surgery is 93% more accurate than unassisted (Jayadev & Shetty, 2020).

Although there are conflicting opinions and fears that AI will replace medical specialists (Shuaib et al., 2020), digitization will actually contribute to changing the job content of highly qualified doctors, requiring not only professional skills but also technological knowledge for digital assisted procedures/activities. Practically, digital disruption in healthcare has a predominance of enriching the content of work through digital knowledge and skills than the total replacement of some jobs with AI. Technologicalization and digitalization of medical services must be seen as tools used to increase the quality of medical services that will benefit patients, aiming to better tailor treatments in a patient-centered approach. And e-health services by definition are patient-centered medical services, which can be practiced both in preventive or curative / maintenance medicine, and in various subfields of medicine. The connection range between digitization and health services is very wide, from the activities in the psychiatric / psychological counselling offices, (where a face-to-face connection is important in order to

increase the quality of services, but not indispensable) where is a high degree of digitization as a result of the computerization of the entire system of evidence of the patient's progress (treatment sheet, diagnosis, progress, remissions) and implementation of new digital tools such as, chatbots, digital apps and virtual reality (Torous et al., 2021) to the services of family doctors. And in family medicine, digitalization has changed traditional relationships, not only from the perspective of communication and connectivity with patients, but also from the perspective of centralizing information on a single basis. This allows to keep a health history of each patient and, most importantly, by creating databases that combine all the electronic health records of patients from different specialists, any interference in the treatment administered over a period of time can be tracked (Atasoy et al., 2019), thus preventing adverse medical conditions.

The transformation of medical services through digitalization affects all categories of staff in the medical sector, digital skills being just as important with the digital inclusion of patients. Basically, the medical service is redefined, it changes radically from a technological point of view, from the perspective of going through the value chain, from investigation and diagnosis to post-intervention surveillance and establishing the strategy of preserving the post-treatment regained health status.

So, the digital transformation of health jobs is the direction to follow in the future, reconfirmed and shaken by the Covid-19 crisis, but the dynamics depend on technological, economic, social, and cultural factors. In this paper we will analyze the extent to which the dynamics of digitization is facilitated or not, directly or indirectly, by a) the financial resources allocated to the health sector, as a support for technological adaptation; b) depends on the level of economic development of the country, as the basis for medical inclusion (health services for all) and the acceptance by the beneficiaries of the e-health services model and c) is obstructed / slowed or not by the migration of human resources, health specialists.

METHODOLOGY AND DATA DESCRIPTION

The aim of the research is to identify to what extent factors such as the level of economic development, the financing of the health sector and the external mobility of specialists influence the digital reform in the health sector.

In order to achieve the purpose of the research, we set out to test the following economic hypotheses:

H_1 - the level of economic development of the countries has a positive influence on the financing rate of the health sector;

H_2 - the migration of doctors from the countries of origin is determined by the investments in health in the countries of destination and their level of digitalization and the use of AI.

To test the proposed hypotheses, we will use both descriptive statistics and econometric data analysis.

The empirical study is based on OLS method with fixed and random effects for panel data, with the following general matrix form:

$$Y_{it} = \beta_0 + \beta_1 * X_{1it} + \beta_n * X_{nit} + \epsilon_{it}$$

Where,

Y_{it} – the value of dependent variable;

β_0 – the scalar;

β_1 – represents a $k * 1$-dimensional vector;

X_{1it} – the value of the independent variable;

ϵ_{it} – the discrepancy variable or deviation;

i and t indicate the analyzed countries and time period, respectively.

To test the proposed economic hypotheses, we will apply the following econometric models:

H_1:

$$health\ exp_{it} = \beta_0 + \beta_1 * \log(gdp/cap_{it}) + \epsilon_{it} \quad (1)$$

H_2:

$$\log(migr\ doct_{it}) = \beta_0 + \beta_1 * health\ exp_{it} + \epsilon_{it} \quad (2)$$

$$\log(migr\ doct_{it}) = \beta_0 + \beta_1 * health\ exp_{it} + \beta_2 * DESI_{it} + \epsilon_{it} \quad (3)$$

The explanatory and dependent variables are presented in Annex 1.

The data used in this research include EU Member States and OECD countries, except for some countries due to lack of data availability.

Regarding the econometric model (1) and (2), the database used covered all OECD countries for the period 2007-2019, except Australia, Chile, Colombia, Costa Rica, Iceland, Japan, Korea, Mexico, Norway, Switzerland, and Turkey. We selected the OECD database, both due to the availability of data on migrant workers employed in the labor market, and in health sector, in the OECD destination countries, and due to the fact that Romanian doctors traditionally migrate mainly to these countries (Apostu et al., 2022; Boboc et al., 2011), as can be seen also from the Table 2.

The econometric model (3) includes the following EU Member States for the period 2016-2019: Austria, Belgium, the Czech Republic, Denmark, Estonia, France, Germany, Hungary, Italy, Latvia, the Netherlands, Poland, Slovenia, and Sweden.

The lack of data availability for certain states or short periods of time is the limitation of the research, and this aspect will be remedied in subsequent research.

Econometric analysis was performed using the R programming language.

RESULTS AND DISCUSSION

In order to be able to carry out the technologically assisted medical activity and AI, it is important to invest financial resources, both by the national public authorities and by the private sector, in order to develop and digitize it. According to Chart 1, we notice significant differences in the European Union in terms of investments in health, their share ranging between 5% and 12% of total GDP.

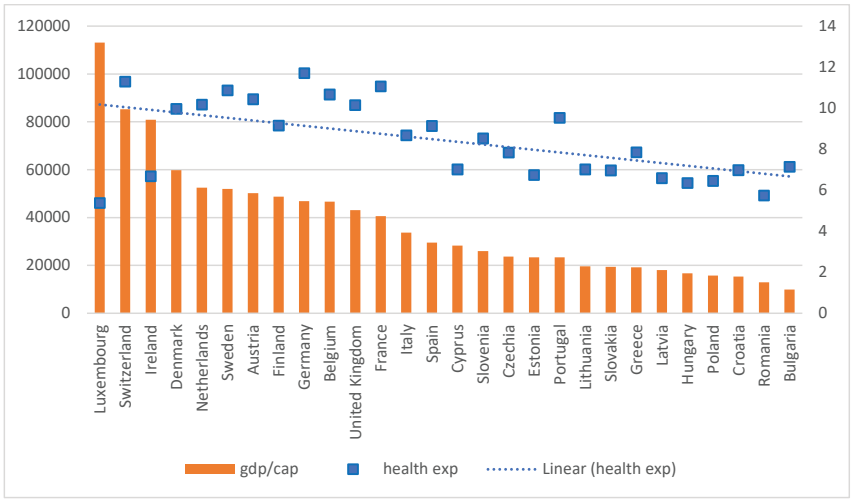


Chart 1. GDP/capita ($) and health expenditures as a share of GDP (%) at EU level in 2019

Source: Authors' calculation based on the Eurostat (2021a) data

The significant differences in the financing of the health sector are also observed depending on the level of economic development, the more developed the countries, the higher the share of investments in GDP compared to the less developed countries, except for Luxembourg (the share of expenditures is reduced due to the absolute value of GDP - $ 73.31

allocates less money on health expenditures compared to the European Union average (5 times less Euros/capita of around 3200, or 2 times less Euros PPP of 2572, in 2019- OECD / EU 2020), also observed in the number of medical staff per 100,000 inhabitants (1.7 times less than EU average - Chart 3) and migratory tendencies of the medical staff, which tends to work in other destination states, not only from the perspective of higher earnings, but also of precarious working conditions supported by the limited resources allocated.

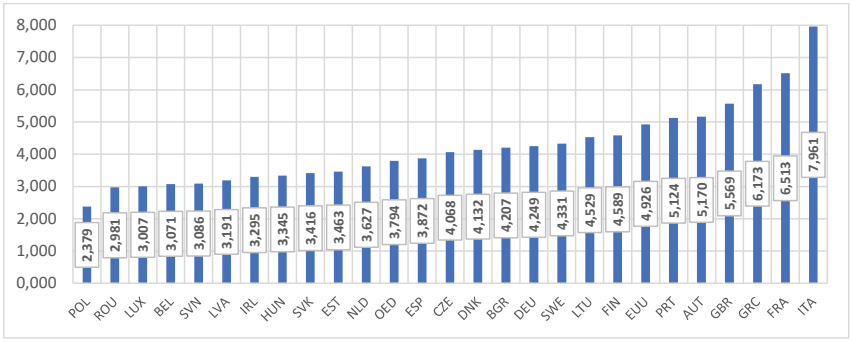


Chart 3 Physicians (per 1,000 people) - European Union 2017
Source: WB database, For Bulgaria, 2018,

In addition to the low level of health care in Romania, which does not allow the purchase of modern equipment and the use of the latest technologies and AI, another problem facing the Romanian medical system is the high share of elderly medical staff, whose level of digital Readiness does not allow them to use new technologies. At the same time, studies (Reßing et al., 2018) show that they are not adept at providing e-health services, preferring to offer traditional health services. Thus, in the last 10 years, there has been an increase in the number of medical staff over the age of 65 and staff aged between 50-64, and on the other hand, the number of young staff up to the age of 25 is declining.

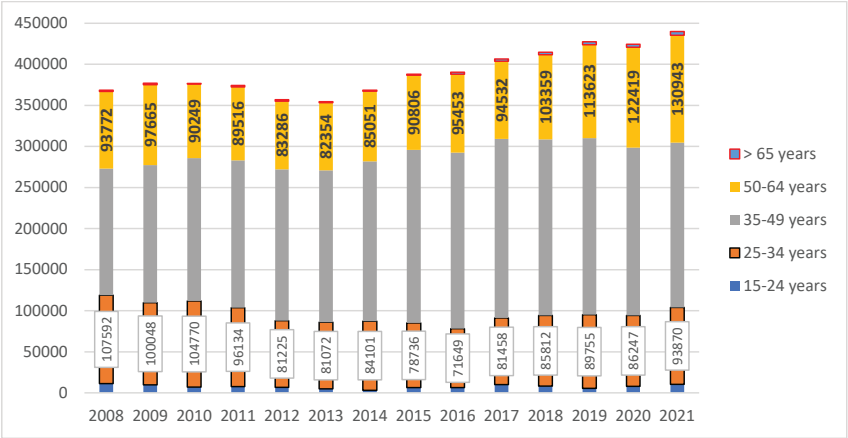


Chart 4. Evolution of the population employed in the health sector by age groups in Romania in 2008-2021, persons

Source: Authors' calculation based on the (National Institute of Statistics (2022) data

Exceptions are the years 2020 and 2021, when the authorities increased the number of residency places for young graduates (Ministry of Health, 2021), as a result of the COVID-19 pandemic, which keeps them in the short term for new graduates, but the problem is the uncertainty of the ability to retention and employment once they complete their residency studies. Even in these conditions of increasing number, forced by the pandemic, the number of employed persons in health sector of age group 25-34 fail to recover the level from 2008-2011.

According to the latest available data Eurostat (2021b; 2021c), we find that at the level of European Union, Romania is among the countries with the fewest doctors / 100,000 inhabitants (318), along with Latvia, Slovenia, France, and Belgium. In terms of the number of nurses, Romania is still on the last positions with 1,142 / 100,000 inhabitants, fewer nurses being only in Croatia (740 / 100,000 inhabitants).

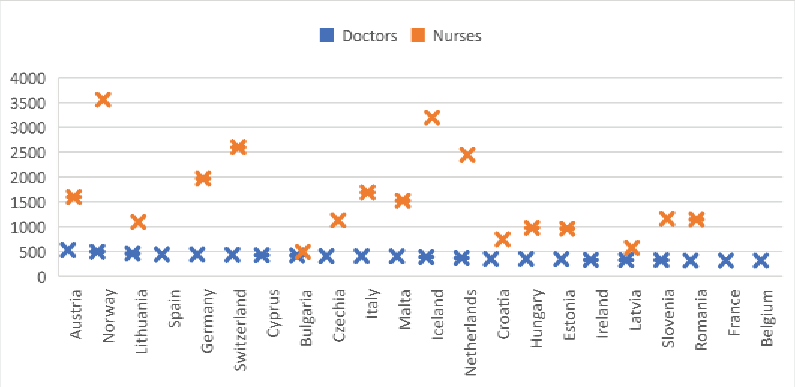


Chart 5. Medical doctors and nurses at EU level in 2019, staff/100,000 inhabitants
Source: Authors' calculation based on the Eurostat (2021b; 2021c) data
*This indicator is not available for: Spain, Cyprus, Ireland, France, and Belgium.

The reduced number of medical staff in Romania is not due to the decrease of the medical graduates, their number being increasing (Vasile et al., 2021) but rather to their migration to other destination countries. According to official data provided by the OECD (2020), over 30% of doctors born and educated in Romania have decided to migrate to other countries. Thus, Romania provides over 20,000 doctors (being on the 5th position in the top countries of origin) and 40,698 nurses (8th position) in the OECD area alone.

Table 2. Romanian medical doctors by destination country in OECD area, persons

Country / year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Austria	16	22	23	31	51	53	51	55	62	60	n/a
Belgium	566	744	866	975	1064	1172	1247	1300	1319	1371	1411
Chile	n/a	n/a	n/a	n/a	12	12	12	12	12	12	5
Czech Republic	-	1	2	1	1	6	8	8	8	11	11
Finland	38	44	46	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
France	n/a	2697	3099	3410	3705	3993	4272	4497	4697	4911	n/a
Germany	1269	1840	2559	3042	3363	3503	3661	3857	3978	4058	n/a
Hungary	1701	1652	1624	1623	1683	1630	1699	1819	1870	1963	n/a
Ireland	n/a	226	286	341	487	625	723	733	715	709	710
Israel	1206	1245	1252	1263	1308	1388	1445	1538	1635	1754	1913
Italy	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	266	335
Netherlands	43	47	43	45	46	57	60	64	57	n/a	n/a
New Zeeland	18	18	20	23	20	19	19	19	19	25	n/a
Norway	56	65	73	85	96	120	138	159	169	172	182
Poland	-	-	-	-	4	4	6	6	7	17	26
Slovenia	2	2	4	4	4	4	3	3	3	3	4

Sweden	421	485	564	628	735	569	729	814	921	n/a	n/a
Switzerland	148	156	178	204	230	248	276	305	334	365	395
Turkey	4	4	4	4	4	4					
Great Britain	435	582	639	764	852	872	949	1037	1129	1274	1347
US	2141	2324	2430	2457	2446	2457	2455	n/a	n/a	n/a	n/a
TOTAL	**8,064**	**12,154**	**13,712**	**14,900**	**16,111**	**16,736**	**17,753**	**16,226**	**16,935**	**16,971**	**6,339**

Source: Authors' calculation based on the OECD (2021) data
*Not all countries reported data for 2020

As we can see, the number of Romanian doctors (Table 2) and nurses (Annex 2) migrating to OECD countries has doubled in the last 10 years, with Romania being a significant provider of labor in the medical sector. Thus, the migration trend continues to increase for both doctors and nurses. However, Romania continues to export medical personnel, even during this global health crisis, with 231 medical staff leaving to Austria at the end of March (Euractiv, 2020), even if the pandemic have seriously affected the Romanian health sector and the shortage of doctors has increased further as a result of the large number of more than 1,000 diseases with the new virus among medical personnel (Romanian Government, 2020b).

Although statistics are predominant across the two categories of medical staff, we should also consider lower skilled workforces, such as stretchers or careers, who are also migrating, to similar positions or to other jobs, but unfortunately there are no official statistics available to the general public yet. This category of workers, although seemingly overlooked, are very important because it would change the pattern of health care retention on national labor markets, considering all workers, regardless of their skills and qualification level.

Therefore, developed countries are attractive to migrant physicians in other less developed countries, both because of the differential pay gap and better working conditions (Apostu et al., 2022; Ciuhu et al., 2018; Bazillier & Boboc, 2016) and because of the high level of digital development of health services provided to patients.

Table.3 The impact of health expenditures on foreign-trained doctors using OLS with random effect

	Coefficient	**Standard error**	**t test**	**p-value**
Intercept	1.083718	0.309187	3.5051	0.0004565 ***
Health exp	0.531229	0.032359	16.4169	< 2.2e-16 ***
p-value < 2.2e-16				
Hausman test – 0.2199				

The econometric analysis conducted by applying OLS with random effects supports the theory that states that invest a higher share of GDP are more attractive to migrant doctors from other less developed countries of origin. Thus, the 1% increase in the value of expenditures in

the health sector determines on average the 0.53% increase in the number of migrant doctors who choose to migrate to another destination state, thus H_2 is accepted.

Table.4 The impact of health expenditures and DESI Index on foreign-trained doctors using OLS with random effect

	Coefficient	**Standard error**	**t test**	**p-value**
Intercept	2.296355	1.030038	2.2294	0.02579 *
Health exp	0.577721	0.093673	6.1674	6.942e-10 ***
DESI	-0.049543	0.023461	-2.1117	0.03471 *
p-value 5.258e-09				
Hausman test – 0.832				

Developing the econometric model previously applied, by introducing a new independent variable, which measures the level of digitization in each country, we find that e-health services have a reverse impact on the migration of doctors. This phenomenon can be explained by the possibilities of providing e-health services without doctors being present in a hospital unit, allowing the practice of the profession remotely. E-health services can be provided with ITC devices, including for surgeries perform, and also patients can benefit from the consultations of the best doctors in any country, without having to travel long distances in other countries, so, some doctors no longer have to migrate for that higher salary differential from another destination country.

Although the data for 2020 are not yet available, certainly the „forced" digitization of public institutions, especially medical institutions have contributed to the growth of the importance of e-health services in all countries of the world, especially for services that do not require surgery, hospitalization in hospitals or ICU.

CONCLUSIONS

Looking beyond the pandemic, digitalization in healthcare is expected to improve a broad range of outcomes, from the prevention and treatment of disease to nursing care. It will allow national health systems to use resources more efficiently, making them more effective and sustainable as societies age (Mihai et al., 2020) .

In this context of digitalization and redefining the health sector and health jobs - as essential jobs, the mobility of medical staff is undergoing a major change. On the one hand, we have the continuation of the trend of the last decades of attracting foreign labor in deficient jobs in the more developed countries of the world (this trend continues for all categories of personnel working in the health sector - auxiliary lower skilled personnel for activity in hospitals and clinics - stretchers, etc.; medium skilled - nurses and similar staff; high skilled -

specialists). On the other hand, the development of the e-health segment based on remote jobs (with important development on 2 subsequent: -preventive medicine and counselling services, including 2 opinions for diagnosis, which can be done by experts based on digitized documents and online consultations with patients - the intervention of highly qualified experts and doctors and their participation in interventions through the remote system. Given this, the mobility / migration of the health workforce will be developed on the 2 channels: - work abroad by establishing the usual residence of migrant workers and - by employment in telework system without physical mobility or with physical mobility occasionally, without changing the usual residence.

For countries of origin, like Romania, the migration / mobility of medical staff involves several externalities, such as:

Negative externalities

- Human capital loss of educated and experienced medical staff, with negative spillover effects on those who remain.
- Increasing the shortage of medical staff, especially in rural or remote areas.
- Reducing the level of medical care and medical conditions.
- Increasing mortality.
- The inability to manage extreme situations such as the COVID-19 pandemic.
- Loss of public education expenditures and scholarships for migrant doctors.
- The cost of investing in training to health professionals.
- Encouraging population migration as a result of a poorly developed medical system.
- Reduction in tax revenues and economic growth.

Positive externalities

- Gaining skills to use new technologies and AI.
- Increasing the level of digitalization by investing capital in new clinics.
- Offering teleconsultations to patients in Romania, with knowledge from the destination countries.
- Remittance receiving by the family members.

The general effects of medical staff migration on source countries depend on the interaction of various factors. These involve modifications in the stock of human capital and

skilled personnel, the number of graduates, the remittances received, the impact on the labor market and changes in the requirement of health care and the health status of the population.

As future research, we intend to extend the analysis to the pandemic period, by completing data not yet available, and identifying the effects of graduate migration on jobs' digital transformation in the Romanian health system, compared to the performance of countries of preference for migration of doctors and nurses.

Acknowledgement: This paper received financial support through the project entitled „DECIDE - Development through entrepreneurial education and innovative doctoral and postdoctoral research, project code POCU / 380/6/13/125031, project co-financed from the European Social Fund through the Operational Program Human Capital 2014 – 2020".

REFERENCES

Apostu, S. A., Vasile, V., Marin, E., & Bunduchi, E. (2022). Factors Influencing Physicians Migration—A Case Study from Romania. *Mathematics* , *10*(3), 505. https://doi.org/10.3390/MATH10030505

Apostu, S. A., Vasile, V., & Veres, C. (2021). Externalities of Lean Implementation in Medical Laboratories. Process Optimization vs. Adaptation and Flexibility for the Future. *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 12309*, *18*(23), 12309. https://doi.org/10.3390/IJERPH182312309

Atasoy, H., Greenwood, B. N., & McCullough, J. S. (2019). The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. *Https://Doi.Org/10.1146/Annurev-Publhealth-040218-044206*, *40*(1), 487–500. https://doi.org/10.1146/ANNUREV-PUBLHEALTH-040218-044206

Bagchi, A. D., Melamed, B., Yeniyurt, S., Holzemer, W., & Reyes, D. (2018). Telemedicine delivery for urban seniors with low computer literacy: A pilot study. *Online Journal of Nursing Informatics*, *22*(2). https://doi.org/10.2/JQUERY.MIN.JS

Bazillier, R., & Boboc, C. (2016). Labour migration as a way to escape from employment vulnerability? Evidence from the European Union. *Applied Economics Letters*, *23*(16), 1149–1152. https://doi.org/10.1080/13504851.2016.1139670

Behringer, A. (2018). Could telemedicine cure Germany's health system? . Retrieved December 11, 2021, from Healthcare in Europe website: https://healthcare-in-europe.com/en/news/could-telemedicine-cure-germany-s-health-system.html

Berrouiguet, S., Baca-García, E., Brandt, S., Walter, M., & Courtet, P. (2016). Fundamentals

for Future Mobile-Health (mHealth): A Systematic Review of Mobile Phone and Web-Based Text Messaging in Mental Health. *Journal of Medical Internet Research*, *18*(6), e5066. https://doi.org/10.2196/JMIR.5066

Boboc, C., Vasile, V., & Ghiţă, S. (2011). Migration of physicians: Causes and effects in CEE countries. *Communications in Computer and Information Science*, *210 CCIS*(PART 3), 514–520. https://doi.org/10.1007/978-3-642-23065-3_74

Çela, A., Hysa, E., Voica, M. C., Panait, M., & Manta, O. (2021). Internationalization of Large Companies from Central and Eastern Europe or the Birth of New Stars. *Sustainability* , *14*(1), 261. https://doi.org/10.3390/SU14010261

Ciuhu, A.-M., Vasile, V., & Boboc, C. (2018). Occupations with Multiple Vulnerabilities in Romania. Retrieved February 8, 2022, from Romanian Statistical Review website: https://www.researchgate.net/publication/325902474_Occupations_with_Multiple_Vulnerabilities_in_Romania

Collins, K., Walters, S., & Bowns, I. (2004). Patient satisfaction with teledermatology: Quantitative and qualitative results from a randomized controlled trial. *Journal of Telemedicine and Telecare*, *10*(1), 29–33. https://doi.org/10.1258/135763304322764167

Euractiv. (2020). Austria imports workers from Bulgaria, Romania to plug gaps in COVID-19 care .

Eurostat. (2021a). Expenditure for selected health care functions by health care financing schemes. Retrieved December 14, 2021, from european Commission website: https://appsso.eurostat.ec.europa.eu/nui/setupDownloads.do

Eurostat. (2021b). Health personnel (excluding nursing and caring professionals). Retrieved December 21, 2021, from European Commission website: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_rs_prs1&lang=en

Eurostat. (2021c). Nursing and caring professionals. Retrieved December 21, 2021, from European Commission website: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_rs_prsns&lang=en

Farrell, M. (2016). Use of iPhones by Nurses in an Acute Care Setting to Improve Communication and Decision-Making Processes: Qualitative Analysis of Nurses' Perspectives on iPhone Use. *JMIR MHealth and UHealth*, *4*(2). https://doi.org/10.2196/MHEALTH.5071

Groyer, A., & Campbell, R. (2018). *Digital Health and Disability Claims*.

Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery*, *268*(1), 70–76.

https://doi.org/10.1097/SLA.0000000000002693

Herman, E. (2020). The Influence of ICT Sector on the Romanian Labour Market in the European Context. *Procedia Manufacturing*, *46*, 344–351. https://doi.org/10.1016/J.PROMFG.2020.03.050

Jayadev, C., & Shetty, R. (2020). Artificial intelligence in laser refractive surgery – Potential and promise! *Indian Journal of Ophthalmology*, *68*(12), 2650. https://doi.org/10.4103/IJO.IJO_3304_20

Mihai, M., Titan, E., Manea, D.-I., & Ionescu, C.-D. (2020). DIGITAL INNOVATION IN THE HEALTH SECTOR – A DETERMINANT OF HEALTH STATUS. RECORDS IN THE EU . *New Trends in Sustainable Business and Consumption*, 579–586. Retrieved from www.editura.ase.ro

Ministry of Health. (2021). Numărul locurilor de la Rezidenţiat, suplimentat . Retrieved February 5, 2022, from http://www.ms.ro/2021/11/26/numarul-locurilor-de-la-rezidentiat-suplimentat/

National Institute of Statistics. (2022). Populatia ocupata pe activitati, grupe de varsta si sexe. Retrieved February 5, 2022, from AMIGO website: https://insse.ro/cms/files/publicatii/Statistica teritoriala/Forta de munca_ind_JudLoc.htm

OECD. (2020). *Contribution of migrant doctors and nurses to tackling COVID-19 crisis in OECD countries*. Retrieved from https://www.oecd.org/coronavirus/policy-responses/contribution-of-migrant-doctors-and-nurses-to-tackling-covid-19-crisis-in-oecd-countries-2f7bace2/

OECD. (2021a). Health Workforce Migration : Foreign-trained doctors by country of origin - Stock. Retrieved April 21, 2021, from https://stats.oecd.org/index.aspx?queryid=68336

OECD. (2021b). Health Workforce Migration : Foreign-trained nurses by country of origin - Stock. Retrieved December 21, 2021, from OECD Statistics website: https://stats.oecd.org/Index.aspx?QueryId=68336

Olesch, A. (2021). Germany benefits from digital health infrastructure during COVID-19 pandemic . Retrieved December 11, 2021, from Healthcare IT News website: https://www.healthcareitnews.com/news/emea/germany-benefits-digital-health-infrastructure-during-covid-19-pandemic

Oliviera Hashiguchi, T. (2020). *Bringing health care to the patient : An overview of the use of telemedicine in OECD countries | OECD Health Working Papers | OECD iLibrary* (No. 116). Retrieved from https://www.oecd-ilibrary.org/social-issues-migration-health/bringing-health-care-to-the-patient_8e56ede7-en

Parker, S., Prince, A., Thomas, L., Song, H., Milosevic, D., & Harris, M. F. (2018). Electronic, mobile and telehealth tools for vulnerable patients with chronic disease: a systematic review and realist synthesis. *BMJ Open*, *8*(8), e019192. https://doi.org/10.1136/BMJOPEN-2017-019192

Perkins, Z. B., Yet, B., Sharrock, A., Rickard, R., Marsh, W., Rasmussen, T. E., & Tai, N. R. M. (2020). Predicting the Outcome of Limb Revascularization in Patients With Lower-extremity Arterial Trauma: Development and External Validation of a Supervised Machine-learning Algorithm to Support Surgical Decisions. *Annals of Surgery*, *272*(4), 564–572. https://doi.org/10.1097/SLA.0000000000004132

Player, M., O'bryan, E., Sederstrom, E., Pinckney, J., & Diaz, V. (2018). Electronic Visits For Common Acute Conditions: Evaluation Of A Recently Established Program. *Health Affairs*, *37*(12), 2024–2030. https://doi.org/10.1377/HLTHAFF.2018.05122

Qudah, B., & Luetsch, K. (2019). The influence of mobile health applications on patient - healthcare provider relationships: A systematic, narrative review. *Patient Education and Counseling*, *102*(6), 1080–1089. https://doi.org/10.1016/J.PEC.2019.01.021

Reßing, C., Mueller, M., Knop, M., & Niehaves, B. (2018). *Building Digital Bridges: Exploring the Digitized Collaboration of General Practitioners and Mobile Care in Rural Areas SenseVojta: Sensor-based diagnostics, therapy and aftercare following the Vojta principle View project ANTARES View project*. Retrieved from https://www.researchgate.net/publication/343693011

Romanian Government. (2020). COVID-19 știri oficiale.

Shuaib, A., Arian, H., & Shuaib, A. (2020). The Increasing Role of Artificial Intelligence in Health Care: Will Robots Replace Doctors in the Future?</p>. *International Journal of General Medicine*, *13*, 891–896. https://doi.org/10.2147/IJGM.S268093

Siemens Healthineers. (2021). *Insights Series Digitalizing Healthcare*. Retrieved from https://www.siemens-healthineers.com/insights/digitalizing-healthcare

Statista. (2021). Awareness and adoption of AI and automation in healthcare worldwide in 2019 and 2020. Retrieved December 12, 2021, from Statista website: https://www.statista.com/statistics/1223613/state-of-healthcare-automation-worldwide/

Sundberg, K., Eklöf, A. L., Blomberg, K., Isaksson, A. K., & Wengström, Y. (2015). Feasibility of an interactive ICT-platform for early assessment and management of patient-reported symptoms during radiotherapy for prostate cancer. *European Journal of Oncology Nursing*, *19*(5), 523–528. https://doi.org/10.1016/J.EJON.2015.02.013

Timus, M., Ciucan-Rusus, L., Stefan, D., & Popa, M.-A. (2020). Student Relationship

Management Optimization Using Organizational Process Automation Tools . *Acta Marisiensis, Seria Oeconomica*, *14*(1), 31–40. Retrieved from https://sciendo.com/downloadpdf/journals/amso/14/1/article-p31.xml

Toquero, C. (2021). Mobile Healthcare Technology for People with Disabilities amid the COVID-19 pandemic. *European Journal of Environment and Public Health*, *5*(1), em0060. Retrieved from https://www.ejeph.com/download/mobile-healthcare-technology-for-people-with-disabilities-amid-the-covid-19-pandemic-8551.pdf

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., … Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335. https://doi.org/10.1002/WPS.20883

van Dis, J. (2002). Where We Live: Health Care in Rural vs Urban America. *JAMA*, *287*(1), 108–108. https://doi.org/10.1001/JAMA.287.1.108-JMS0102-2-1

Vasile, V., Bunduchi, E., Boboc, C., & Vasile, R. (2021). GRADUATES AND THE LABOR MARKET DEFICIT IN THE ROMANIAN HEALTH SECTOR . *CKS*, 919–924. Retrieved from https://www.researchgate.net/publication/352019224

World Bank. (2021). Rural population (% of total population). Retrieved December 12, 2021, from World Bank Indicators website: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

Wu, C. F., Chang, T., Wang, C. M., Wu, T. P., Lin, M. C., & Huang, S. C. (2021). Measuring the Impact of Health on Economic Growth Using Pooling Data in Regions of Asia: Evidence From a Quantile-On-Quantile Analysis. *Frontiers in Public Health*, *9*, 999. https://doi.org/10.3389/FPUBH.2021.689610/BIBTEX

Zhou, X. Y., Guo, Y., Shen, M., & Yang, G. Z. (2020). Application of artificial intelligence in surgery. *Frontiers of Medicine*, *14*(4), 417–430. https://doi.org/10.1007/S11684-020-0770-0

Annexes

Annex 1. Variables used in econometric analysis

Variables	Abbreviation	Data source
Inflows of foreign-trained doctors	migr doct	OECD
GDP per capita	gdp_cap	World Bank
Health care expenditures	health exp	Eurostat
Digital Economy and Society Index	DESI	European Commission

Annex 2. Romanian nurses by destination country in OECD area, persons

Country / year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Belgium	298	421	690	888	1068	1224	1329	1431	1598	1713	1791
Canada	430	448	472	508	515	520	527	524	518	513	n/a
France	68	115	147	164	179	193	203	221	238	250	n/a
Greece	22	22	22	22	22	22					n/a
Hungary				431	530	576	591	603	626	621	n/a
Israel	98	92	86	82	79	77	71	63	59	55	50
Italy	10570	11215	11531	11731	11820	12159	11714	10969	10690	10635	11253
Netherlands	n/a	n/a	n/a	n/a	n/a	9	12	14	13	n/a	n/a
New Zeeland	n/a	20	17	16	13	13	14	12	14	14	12
Norway	34	34	42	46	57	76	88	99	81	88	92
Sweden	2	4	8	17	24	30	49	58	73	n/a	n/a
Turkey	2	3	3	3	3	3	n/a	n/a	n/a	n/a	n/a
United Kingdom	1272	1909	2254	2606	3739	5997	8115	7725	7542	7407	7421
Total	12,796	14,283	15,272	16,514	18,049	20,899	22,713	21,719	21,452	21,296	20,619

Source: Authors' calculation based on the OECD (2021b) data
*Not all countries reported data for 2020

Dealing with outliers generated by the COVID-19 pandemic in the process of seasonal adjustment of macroeconomic time series

Andreea MIRICĂ – Lecturer, PhD. (miricaandreea89@gmail.com)
Bucharest University of Economic Studies, Bucharest, Romania

Octavian CEBAN – PhD. Candidate (octavianceban1995@gmail.com)
Bucharest University of Economic Studies, Bucharest, Romania

Traian-Ovidiu CALOTĂ – Associate Professor, PhD. (traian.calota@infofisc.ro)
Titu Maiorescu University, Faculty of Finances, Banks, Accounting and Busineess Administration, Bucharest, Romania

Roxana-Violeta PARTAS-CIOLAN – PhD. Candidate (roxana.partas@gmail.com)
Bucharest University of Economic Studies, Bucharest, Romania

Liliana CATRINA – PhD. Candidate (liliana.catrina@gmail.com)
Bucharest University of Economic Studies, Bucharest, Romania

1. Introduction

Economic activities have been severely affected by the COVID-19 outbreak. This causes the presence of outliers at the end point of time series, which may affect the revision of the seasonally adjusted series each time new data becomes available (Eurostat, 2020a). Given the special circumstances, as it is impossible to predict the development of the crisis, Eurostat considers that modelling outliers based solely on statistic criteria using the automatic is an acceptable solution; however, using statistical criteria and economic information is recommended (Eurostat, 2020a). It should also be noted that revisions of the increase rates calculated based on seasonally adjusted data should be kept reasonable (Mirica et al. 2016).). The economy is prone to various shocks, especially on long term, such as war crises, pandemics, and a "good" statistical approach can be compromised by fitting to data ignoring some assumptions or a time-series characteristic (Hendry et al. 2011). Using outliers when modelling implies high risk of inflated errors (Osborne at al. 2004). This, along with the fact that only 8% of the researchers report checks for outliers in their work as outlined by an empirical study (Osborne et al 2001), leads to concerns. The main problem with outliers is that

they appear to be generated from a distribution that is different from the one that is modelled and based on which the assumptions are made. An illustration of this situation is presented by Hawkins D.M. who emphasizes that "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980, p.1). Modelling with outliers increase error variance, reduce the power of statistical test, and can decrease the normality (Osborne 2004). Besides the variance, outliers can add bias to the predictions or estimators (Zimmerman 1994). These uncontrolled or unexpected interventions leads to a "moderate to significant impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation and forecasting" (Chung and Lon-Mu 1993).

The outlier's detection, although it seems easy, in temporal data becomes very challenging and it has been addressed from different perspectives (Battaglia et al. 2005; Caroni and Karioti 2004. In some cases, the outlier may have an economic interpretation such as fraud detection (Takeuchi and Yamanishi, 2006). An example of relatively new approach with good results for detecting additive outliers is based on genetic algorithms (Cucina et al. 2014).

One solution for modelling with outliers is presented by Kourentzes et al. 2014 and is based on ensemble of neural networks operators. The idea is to use a mode ensemble operator based on kernel density estimation, since this technique is insensitive to outliers and deviations from normality. Martinez et al (2018) suggest the use of a multidimensional nearest neighbours algorithms for modelling seasonal time series affected by outliers.

2. Data and methods

The purpose of this paper is to present a procedure for dealing with outliers observed at the end of a raw series. In this respect, the quarterly data for the Romanian GDP will be used. The data was retrieved from the Tempo Online Database of the National Institute of Statistics Romania on October 9th, 2020 and comprises of raw figures from the first quarter of 2015 up to the second quarter of 2020 (22 observations). Subsequently, new data points presenting different possible scenarios for the third quarter of 2020 have been added, resulting in 10 series:

1. GDP_0.9 – in this scenario GDP in the third quarter of 2020 is 90% of the GDP in the third quarter of 2019.
2. GDP_0.91 – in this scenario GDP in the third quarter of 2020 is 91% of the GDP in the third quarter of 2019.
3. GDP_0.92 – in this scenario GDP in the third quarter of 2020 is 92% of the GDP in the third quarter of 2019.
4. GDP_0.93 – in this scenario GDP in the third quarter of 2020 is 93% of the GDP in the third quarter of 2019.
5. GDP_0.94 – in this scenario GDP in the third quarter of 2020 is 94% of the GDP in the third quarter of 2019.
6. GDP_0.95 – in this scenario GDP in the third quarter of 2020 is 95% of the GDP in the third quarter of 2019.
7. GDP_0.96 – in this scenario GDP in the third quarter of 2020 is 96% of the GDP in the third quarter of 2019.
8. GDP_0.97 – in this scenario GDP in the third quarter of 2020 is 97% of the GDP in the third quarter of 2019.
9. GDP_0.98 – in this scenario GDP in the third quarter of 2020 is 98% of the GDP in the third quarter of 2019;

10. GDP_0.99 – in this scenario GDP in the third quarter of 2020 is 99% of the GDP in the third quarter of 2019.

Next, the facilities of the open-source software JDemetra+ 2.2.3 will be used for series treatment. JDemetra+ is the software officially recommended by Eurostat for seasonal adjustment (Eurostat, 2020b). This software has multiple features that will be briefly presented below.

Firstly, according to Grudkowska (2021a), JDemetra+ 2.2.3 incorporates adapted versions of TRAMO-SEATS and X13ARIMA-SEATS that permit the use of plugins and extensions within a graphical user interface. One should note that "*both methods are officially recommended by Eurostat, ECB, OECD, IMF and others and produce similar but not identical results*". (UNECE, 2020, p. 38). According to the same source, both methods share common features, however, X-13 offers highly adaptable tools for the pre-adjustment part as well as several filters to choose from for the decomposition part. Also, JDemetra+ incorporates an automatic procedure for seasonal adjustment both within TRAMO-SEATS and X13, that proved to be effective and easy to use (Toma et al, 2018). Better results were obtained with X13 for series affected by calendar effects or outliers (Mirica, 2019).

Secondly, JDemetra+ 2.2.3 includes two useful tools for the automatic detection of outliers: Anomaly Detection and Check last (Grudkowska 2021b). According to the same source, both these tools are built upon the TRAMO app, TERROR. TERROR stands for "TRAMO for errors" (Caporello et al. 2000, p.20). According to the same source, TRROR computes an ARIMA model for the time series, that is further used to compare the actual values with the adjusted values and if any of the given difference is higher than a value specified by the user, the point in the time series is identified as error (outlier).

Thirdly, JDemetra+ 2.2.3 offers an easy-to-use instrument for defining calendar effects (Grudkowska 2021c). This is an important component in the seasonal adjustment process, as most series are influenced by the daily intensity of the economic activity (Ladiray, 2006). In this respect, JDemetra+ 2.2.3 offers the possibility to set a specific weight for each holiday, predefined by the user, to account for the impact of that specific day on the entire series (Grudkowska 2021d).

3. Results

Firstly, the 10 series were tested for the presence of outliers using the outliers detection tool in JDemetra+2.2.3. As can be observed from table 1, there are significant outliers within the series.

Table 1

Series	Outlier type	**Period**	**Value**	**StdErr**	**TStat**
GDP_0.9	Level Shift	II-2020	-6095.6923	570.2203	-10.6901
GDP_0.91	Level Shift	II-2020	-5804.4633	566.5768	-10.2448
GDP_0.92	Transitory Change	II-2020	-5846.1316	764.0029	-7.6520
GDP_0.93	Transitory Change	II-2020	-5674.6938	726.7415	-7.8084
GDP_0.94	Transitory Change	II-2020	-5571.0726	700.1872	-7.9565
GDP_0.95	Transitory Change	II-2020	-5483.7524	688.7626	-7.9617
GDP_0.96	Transitory Change	II-2020	-5391.5962	629.5402	-8.5643
GDP_0.97	Transitory Change	II-2020	-5216.2521	643.3317	-8.1082
GDP_0.98	Transitory Change	II-2020	-5094.2040	671.2414	-7.5892
GDP_0.99	Additive Outlier	II-2020	-4331.5818	537.4194	-8.0600

Secondly, in order to perform the seasonal adjustment of these time series, a calendar is defined comprising of all the legal holidays in Romania including the Julian Easter and related holidays. The calendar is then incorporated within the X13 RSA5c specification available in JDemetra+ 2.2.3. With regard to the calendar definition, one can observe that we chose not to define special days for the periods corresponding to the emergency state and alert state related to COVID-19 outbreak within the second quarter of 2020 in Romania. That is because "*the selection of relevant calendar effects used for calendar adjustment should be kept constant over appropriately long time periods*" and "*changes in the selection of calendar effects should be based on both empirical evidence and economic explanation*" (Eurostat, 2008 p.2). As nobody can anticipate how long an emergency state or alert state along with underlying restrictions due to pandemics can occurs, we chose to preserve the consistency of the calendar, rather than introduce additional special days that might not affect the series on the long run. Also, as nobody can state the exact effects of the mitigation efforts on the economic activity, it would be impossible to choose an appropriate weight for such special days.

Thirdly, the automatic procedure for choosing the ARIMA Model and the Henderson filter available in JDemetra+ 2.2.3 for seasonal adjustment is used. This procedure has a built-in outlier detection and correction tool. This initial specification has a default TC rate (rate of decay of the transitory change outlier) of 0.7 and it can be changed with a number between 0 and 1. It also uses Msr as a default seasonal filter that can be changed with several other options: S3x1, S3x3, S3x5, S3x9, S3x15, Stable, X11 Default. The default critical value for detecting outliers is 4 and it can be changed with a value chosen by the user.

One should note that according to the press release 264/October 9th, 2020 of the National Institute of Statistics, GDP – Seasonally adjusted decreased by 11.9% in the second quarter of 2020 compared to the first quarter of 2020 and remained the same in the first quarter of 2020 compared to the last quarter of 2019. In order to obtain acceptable revisions (more specifically, in order to preserve the increase rates that the National Institute of Statistics Romania announced), some modifications had to be performed to the initial specification. Table 2 presents these modifications as well as the results of the seasonal adjustment procedure. As one can observe, performing these modifications resulted in satisfactory results from the perspective of the quality of the seasonal adjustment process. Moreover, the revised increase rates for the first and second quarters of 2020 are reasonable.

Table 2

	Modifications of the initial specification				Seasonal adjustment quality outcome	Increase rates (%)	
Series	Series span	TC rate	Seasonal filter	Critical values for outliers		revised for the first quarter	Revised for the second quarter
GDP_0.9 (log transform)	From 2016, first quarter	-	S3x9	6	good	0.01	-12.27
GDP_0.91 (log transform)	From 2016, first quarter	0.9	S3x1	6	good	0.04	-12.26

GDP_0.92 (auto transformation)	-	0.9	S3x1	-	good	0.13	-12.48
GDP_0.93 (auto transformation)	-	-	-	-	good	0.23	-11.38
GDP_0.94 (auto transformation)	-	0.6	S3x5	-	good	0.38	-11.36
GDP_0.95 (auto transformation)	-	0.942	S3x5	-	good	0.19	-11.47
GDP_0.96 (auto transformation)	-	0.91	S3x3	-	good	0.02	-11.64
GDP_0.97 (auto transformation)	-	0.92	-	-	good	0.53	-11.62
GDP_0.98 (auto transformation)	-	0.85	S3x3	-	good	0.05	-11.77
GDP_0.99 (auto transformation)	From 2016, first quarter	0.85	S3x9	6	Good	0.09	-11.74

Next, further justifications for the modifications performed to the initial specification are provided. More specifically, we will explain our choices with regard to the critical value for the detection of outliers, seasonal filter, transitory change rate and series span. In this process, one should also bear in mind the principle stated in the beginning of this paper, namely keeping revisions reasonable.

Regarding the critical value for the detection of outliers, one should note that the relative frequency of detection of an outlier when none is present decreases as the critical value increases (Kaiser and Maravall, 1999). Table 3 illustrates the outliers detected with the default critical value (4) as well as a critical value equal to 6. All other elements of the specification are kept the same as in table 2For the first 2 series, if the default critical value is used, the automatic procedure must correct for 4 and 2 outliers, respectively. However, the initial evaluation detected only one Level Shift within the series, that is the value for the second quarter of 2020. This means that if the critical value is set to default, the automatic procedure commits a type 1 error detecting outliers. Moreover, it was noted that the quality of the seasonal adjustment process is severe. For the last series, the automatic procedure finds an additive outlier if the critical value is set to default and a transitory change for critical value equal to 6. There are two possible approaches for this case: set the critical value to 4 as the output matches the initial evaluation of the outlier detection tool or set the critical value to 6 if there is uncertainty about the trend-cycle component being left untouched by the crisis (see Eurostat 2020a, with regard to the Additive outlier). The second approach may be used at least until more data become available. Therefore, it is safer to increase the critical value for outlier

detection if there are too many detected outliers or if there is not enough data to support the existence of a certain type of outlier.
Table 3

Series	Outliers obtained and corrected with a critical value = 6	Outliers obtained and corrected with a critical value = 4
GDP_0.9; From 2016, first quarter	Level Shift (II, 2020); Probability = 0.0000	Level Shift (II, 2020); Probability = 0.0000 Level Shift (I, 2017); Probability = 0.0003 Transitory change (III, 2016); Probability = 0.0009 Additive outlier (I 2019); Probability = 0.0032
GDP_0.91; From 2016, first quarter	Level Shift (II, 2020); Probability = 0.0000	Level Shift (II, 2020); Probability = 0.0000 Level Shift (I, 2017); Probability = 0.0005
GDP_0.99; From 2016, first quarter	Transitory change (II, 2020); Probability = 0.0000	Additive Outliers (II, 2020); Probability = 0.0000

The Transitory change rate (TC rate) represents the rate of decay of a Transitory Change outlier, namely, parameter δ in the following equation, describing a time series y_t affected by such an outlier at time t = k (Galeano and Pena, 2013, p.248):

$$y_t = x_t + \frac{1}{1-\delta B}\omega I_t^{(k)}$$

$$\delta\ tends\ to\ 1, Transitory\ change = Level\ Shift$$

$$\delta\ tends\ to\ 0, Transitory\ change = Additive\ Outlier$$

The choice of the TC rate towards 1 or 0 should be based on economic insights as the entire modelling process of outliers (Eurostat 2020a) as well as ensuring low revision rates for the previous rates of increase announced (Mirica et al. 2016).

With regard to the series span, Mirica et al. (2016) concluded that a series span of 20 to 24 quarters is the best choice for obtaining high quality seasonally adjusted data as well as reasonable revisions while the minimum length of a series that is to be seasonally adjusted is 16 to 20 quarters. According to the same source, one can choose to log-transform the series if the variance fluctuates along with the trend, as in the case of the Romanian quarterly GDP. For the purpose of this paper, we used mainly the auto function embedded within JDemetra+ for the transformation process. However, in the case of the first two series, if the auto function had been used, the increase rate on the seasonally adjusted series for the first quarter of 2020 would have been revised to -3% and -3.4% respectively. As these rates were far from the estimates published by the National Institute of Statistics for the first quarter of 2020, it is clear that the model needed some adjustments. In this respect, a log transformation has been applied to these series. Table 4 presents the results for different time spans and transformation choices for GDP_09, GDP_091 and GDP_099. The TC rate, Seasonal filter and Critical values for outliers were the ones from table 2 in each case. As one can observe, the lowest AIC for GDP_09 and GDP_091 is obtained for the log transformation and series span from 2016q1. Moreover, the

lowest AIC for GDP_099 is obtained for the series span from 2016q1. This proves that shortening the time span as well as transforming the series is the best statistical choice also.

Table 4

Series	AIC value
GDP_09 log transformation, series span from 2016q1	239.4
GDP_09 log transformation, series span from 2015q1	278.6
GDP_09 no transformation, series span from 2016q1	263.0
GDP_09 no transformation, series span from 2015q1	327.4
GDP_091 log transformation, series span from 2016q1	240.0
GDP_091 log transformation, series span from 2015q1	279.5
GDP_091 no transformation, series span from 2016q1	263.5
GDP_091 no transformation, series span from 2015q1	327.9
GDP_099 series span from 2016q1	243.3
GDP_099 series span from 2015q1	317.7

In choosing the most suitable seasonal filter, the principle of minimum revision of the increase rate prevailed. In this respect, three rules emerged for each of our series: firstly, the increase rate for the first quarter must be positive; secondly, the increase rate for the second quarter must be as close to -11.9% as possible: thirdly, the increase rate for the first quarter must be as close to 0 as possible. Table 5 shows the increase rates for the different filters for each series.

Table 5

Series	Seasonal Filter	Increase rate q1 2020	Increase rate q2 2020
GDP_09	S3x1	0.52	-12.79
	S3x3	0.41	-12.67
	S3x5	0.16	-12.41
	S3x9	0.01	-12.27
	S3x15	0.02	-12.27
	Stable	0.02	-12.27
	X11 Default	0.16	-12.41
	Msr	0.16	-12.41
GDP_091	S3x1	0.04	-12.26
	S3x3	0.12	-12.30
	S3x5	-0.02	-12.15
	S3x9	-0.04	-12.11
	S3x15	-0.03	-12.11
	Stable	-0.02	-12.15
	X11 Default	-0.02	-12.15
	Msr	-0.02	-12.15
GDP_092	S3x1	0.13	-12.48
	S3x3	0.19	-12.47
	S3x5	-0.01	-12.3

	S3x9	-0.35	-12.02
	S3x15	-0.48	-11.82
	Stable	-0.48	-11.82
	X11 Default	-0.02	-12.26
	Msr	-0.02	-12.26
GDP_093	S3x1	0.68	-11.22
	S3x3	0.43	-11.31
	S3x5	0.18	-11.4
	S3x9	-0.13	-11.27
	S3x15	-0.21	-11.25
	Stable	-0.21	-11.25
	X11 Default	0.23	-11.38
	Msr	0.23	-11.38
GDP_094	S3x1	1.02	-10.97
	S3x3	0.66	-11.16
	S3x5	0.38	-11.36
	S3x9	0.02	-11.19
	S3x15	-0.05	-11.22
	Stable	-0.05	-11.22
	X11 Default	0.44	-11.36
	Msr	0.44	-11.36
GDP_095	S3x1	0.69	-11.33
	S3x3	0.45	-11.4
	S3x5	0.19	-11.47
	S3x9	-0.12	-11.33
	S3x15	-0.21	-11.32
	Stable	-0.21	-11.32
	X11 Default	0.23	-11.44
	Msr	0.23	-11.44
GDP_096	S3x1	-0.07	-11.65
	S3x3	0.02	-11.64
	S3x5	-0.25	-11.64
	S3x9	-0.44	-11.52
	S3x15	-0.58	-11.46
	Stable	-0.58	-11.46
	X11 Default	-0.19	-11.64
	Msr	-0.19	-11.64
GDP_097	S3x1	1.1	-11.33
	S3x3	0.75	-11.38
	S3x5	0.41	-11.46
	S3x9	0.06	-11.33
	S3x15	-0.02	-11.34
	Stable	-0.02	-11.34
	X11 Default	0.53	-11.62
	Msr	0.53	-11.62
GDP_098	S3x1	-0.02	-11.79
	S3x3	0.05	-11.77
	S3x5	-0.25	-11.73

	S3x9	-0.45	-11.6
	S3x15	-0.58	-11.53
	Stable	-0.58	-11.53
	X11 Default	-0.18	-11.74
	Msr	-0.18	-11.74
GDP_099	S3x1	0.05	-11.49
	S3x3	0.25	-11.99
	S3x5	0.1	-11.57
	S3x9	0.09	-11.74
	S3x15	0.08	-11.7
	Stable	0.08	-11.7
	X11 Default	0.1	-11.57
	Msr	0.1	-11.57

4. Conclusions

In this paper, a methodology for dealing with outliers when computing seasonally adjusted series was presented. This methodology is bases on two principles: (1) keeping the revisions in the increase rates of the seasonally adjusted series as low as possible and (2) using the automatic procedure in JDemetra+ as much as possible, with minimum modifications. Of course, a good quality of the seasonally adjusted data is crucial.

The automatic procedure in the X-13 package, RSA5c specification, in JDemetra+ 2.2.3 is very useful in dealing with outliers. However, some minor modifications should be performed to keep revisions reasonable. These modifications depend on the outlier type of the last data point and may consist of: transforming the series using the logarithm function; choosing a custom time span, TC rate or seasonal filter; setting a different critical value for the detection of the outliers. There is no specific order in performing these modifications, the process being iterative until the best possible combination is obtained.

5. References

1) Eurostat, 2020a METHODOLOGICAL NOTE GUIDANCE ON TREATMENT OF COVID-19-CRISIS EFFECTS ON DATA https://ec.europa.eu/eurostat/cros/system/files/treatment_of_covid19_in_seasonal_adjustment_methodological_note.pdf retrieved October 14th 2020
2) Eurostat, 2020b SEASONAL ADJUSTMENT https://ec.europa.eu/eurostat/web/research-methodology/seasonal-adjustment retrieved October 14th 2020
3) Grudkowska S. (2021a) A brief description of JDemetra+ https://jdemetradocumentation.github.io/JDemetra-documentation/ retrieved February 4th 2021
4) Grudkowska S. (2021b) Statistical methods https://jdemetradocumentation.github.io/JDemetra-documentation/pages/reference-manual/Statistical-methods.html retrieved February 4th 2021
5) Grudkowska S. (2021c) Calendars https://jdemetradocumentation.github.io/JDemetra-documentation/pages/reference-manual/calendars.html retrieved February 4th 2021
6) Grudkowska S. (2021d) National calendar https://jdemetradocumentation.github.io/JDemetra-documentation/pages/case-studies/calendars-national.html retrieved February 4th 2021

7) Caporello G., Maravall A., Sanchez F. (2000) Program TSW Reference Manual, https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/01/Fic/dt0112e.pdf retrieved February 4th 2021
8) Ladiray, D. 2006 Calendar Effects and Seasonal Adjustment: A Review https://ec.europa.eu/eurostat/documents/4578629/4579724/LADIRAY-AB.pdf retrieved February 4th 2021
9) MIRICĂ, A., Andrei, T., Dascălu, E. D., MINCU RĂDULESCU, G. I., & GLĂVAN, I. R. (2016). REVISION POLICY OF SEASONALLY ADJUSTED SERIES–CASE STUDY ON ROMANIAN QUARTERLY GDP. *Economic Computation & Economic Cybernetics Studies & Research*, *50*(3).
10) MIRICĂ Andreea, GLAVAN Ionela Roxana, TOMA Iulia Elena, PATRASCU Lucian, Understanding Patterns in the Consumption of Agro-Food Products in Romania – An Analysis at Regional Level, Romanian Statistical Review / Revista Romana de Statistica, Vol. , Nr. 4, 2019, pg. 81 - 96, ISSN:1844-7694 http://www.revistadestatistica.ro/2019/10/understanding-patterns-in-the-consumption-of-agro-food-products-inromania-an-analysis-at-regional-level/
11) TOMA Iulia Elena, MIRICĂ Andreea, PAUNICA Mihai, Seasonal Adjustment of the Industrial Production Index for Romania – An Innovative Approach Using JDemetra+ 2.1, Romanian Statistical Review / Revista Romana de Statistica, Vol. , Nr. 4, 2018, pg. 121 - 134, ISSN:1844-7694 http://www.revistadestatistica.ro/2018/12/seasonal-adjustment-of-theindustrial-production-index-for-romania-an-innovative-approach-using-jdemetra-2-1/
12) UNECE (2020) PRACTICAL GUIDE TO SEASONAL ADJUSTMENT WITH JDEMETRA+ FROM SOURCE SERIES TO USER COMMUNICATION https://unece.org/DAM/stats/publications/2020/ECECESSTAT20203.pdf accessed February 4th 2021
13) Eurostat. (2008). Guidelines on seasonal adjustment by Task force on Seasonal adjustment of QNA endorsed by the CMFB. Retrieved from http://ec.europa.eu/eurostat/ramon/statmanuals/files/sawd_recommendations.pdf retrieved February 4th 2021
14) Kaiser R., Maravall A. (1999) SEASONAL OUTLIERS IN TIME SERIES https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/99/Fic/dt9915e.pdf 8th February 2021
15) Galeano P and Pena, D. 2013 Finding Outliers in Linear and Nonlinear Time Series http://halweb.uc3m.es/esp/Personal/personas/dpena/publications/ingles/2013GatherBook_galeano.pdf 8th February 2021
16) Osborne, Jason & Overbay, Amy. (2004). The Power of Outliers (and Why Researchers Should Always Check for Them). Pract. Assess. Res. Eval.. 9.
17) Osborne, J. W., Christiansen, W. R. I., & Gunter, J. S. (2001). Educational psychology from a statistician's perspective: A review of the quantitative quality of our field. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.
18) Hawkins, D.M. (1980). Identification of outliers. London: Chapman and Hall.
19) Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. Journal of General Psychology, 121(4), 391-401
20) Takeuchi J.-i., Yamanishi K. A unifying framework for detecting outliers and change points from time series IEEE Trans. Knowl. Data Eng., 18 (4)

(2006), pp. 482-492
21) Battaglia F., Orfei L. Outlier detection and estimation in nonlinear time series J. Time Series Anal., 26 (1) (2005), pp. 107-121
22) Caroni C., Karioti V. Detecting an innovative outlier in a set of time series Comput. Statist. Data Anal., 46 (3) (2004), pp. 561-570
23) Tsay R.S. Outliers, level shifts, and variance changes in time series J. Forecast., 7 (1) (1988), pp. 1-20
24) Cucina D., Di Salvatore A., Protopapas M.K. Outliers detection in multivariate time series using genetic algorithms Chemometr. Intell. Lab. Syst., 132 (2014), pp. 103-110
25) Kourentzes, Nikolaos, et al. "Neural Network Ensemble Operators for Time Series Forecasting." Expert Systems With Applications, vol. 41, no. 9, 2014, pp. 4235–4244.
26) Kourentzes, Nikolaos, et al. "Neural Network Ensemble Operators for Time Series Forecasting." Expert Systems With Applications, vol. 41, no. 9, 2014, pp. 4235–4244.
27) Chen, Chung, and Lon-Mu Liu. "Joint Estimation of Model Parameters and Outlier Effects in Time Series." Journal of the American Statistical Association, vol. 88, no. 421, 1993, pp. 284–297.
28) Martínez, F., Frías, M. P., Pérez-Godoy, M. D., & Rivera, A. J. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with kNN. Expert systems with applications, 103, 38-48.

Statistical study on the stock of foreign direct investments in Bulgaria and Romania

Popescu Liviu (liviu.popescu@edu.ucv.ro; liviunew@yahoo.com)
University of Craiova, Faculty of Economics and Business Administration, Department of Statistics and Economic Informatics

Brostescu Simina (simina_brostescu@yahoo.com)
University of Craiova, Faculty of Economics and Business Administration, Department of Statistics and Economic Informatics

ABSTRACT

In this paper, a statistical research was done on the stock of foreign direct investments in the period 1995-2018 for Bulgaria and Romania. Several valid econometric models have been found that explain the determinants influencing the evolution of the stock of FDI inputs in the two countries. The stock of of foreign direct investments was used as a dependent variable, calculated as a percentage of gross domestic product, and as an independent variables, indicators of international trade, economic trends and components of the signaling indicator given by economic freedom indices. It was found that in the case of Bulgaria, the increase in exports, imports, trade balance and balance of payments lead to an increase in the stock of FDI. Also, a higher score of the components of economic freedom, which is characterized by a business environment as free as possible, by a level of taxation as low as possible, as well as by a diminished degree of corruption, lead to an increase in the stock of FDI inputs in the Bulgarian economy. In the case of Romania, it turned out that exports, trade balance, balance of payments have a positive influence on the stock of FDI, as well as a low level of taxation, both in the case of personal income and corporations, lead to growth stock of FDI. Contrary to Bulgaria, an increase in imports leads to a decrease in the stock of FDI in Romania. Finally, it is found that the indices of diversification of imports negatively influence the stock of FDI in both countries, which means that a greater divergence of imports from the world model leads to a decrease in the stock of FDI.

Keywords*: foreign direct investment stock, multiple linear regression, economic freedom index, international trade, quantitative methods.*

JEL classification: C12, C52, O24

1. Introduction

It is well known that direct investment means long-term investment relations between certain resident and non-resident entities that involve the exercise by investors of significant managerial influences in the companies in which they have invested. The main components of foreign direct investment (FDI) are given by the equity participations of non-resident investors holding at least 10% of the subscribed share capital of some resident enterprises, by the profit reinvested by them, as well as debt instruments (loans, trade credits, external receivables) between investors and the companies in which they have invested.
Given that more than 35% of EU assets are held by foreign companies, it is clear that it has one of the most open investment systems in the world. FDI stocks of non-EU investors reached the threshold of 6.295 billion euros at the end of 2017, which generated approximately 16 million new jobs in the European Union.
Statistical-econometric modeling of foreign direct investment has experienced a sharp diversification in recent decades, as the resulting models have included new factors that influence them and are important for their economic impact, in the literature there are a variety of econometric models related to FDI, each with a certain degree of originality and in order to anticipate economic developments.

Thus, in relation to a set of relevant statistical indicators, several types of statistical-econometric models can be identified in the literature. A first set of models quantifies the link between FDI and country risk assessment as the only exogenous variable or coupled with other social variables (Thomas (2006), Vijayakumar, Rasheed and Tondkar (2009), Lee and Rajan (2011), Săvoiu, Dinu and Ciucă (2013), Săvoiu and Țaicu (2014), Lacroix, Méon and Sekkat (2021)).
A second set of models describes the positive or negative correlation between corruption and FDI, increasing the impact or perception of corruption by increasing or decreasing the volume of investments (Barassi and Zhou (2012), Udenze (2014), Brada et all. (2019), Burlea-Schiopoiu, Brostescu and Popescu (2021)).
A third set of models is based on statistical indicators resulting from the notion of economic freedom generating an increase in FDI (Wells and Wint (2000), Bengoa and Sanchez–Robles (2003), the evolution of the value of the economic freedom index (Caetano and Caleiro (2009), Rožāns (2016)), export and import flows (Greenaway and Kneller (2007), Smits (1988)), local financial markets (Alfaro et all. (2004), Azman – Saini, Law and Ahmad, (2010)).
From the multitude of scientific papers studying FDI, we have selected mainly those that refer to the member countries of the European Union. Thus, Feldstein (1983) analyzed the relationship between domestic economies and international capital movement in 17 countries (1960 - 1979). Smits (1988) studied the correlation between FDI and export and import value in 30 countries. Culem (1988) analyzed the influence of FDI location in 6 European countries between 1969 and 1982. Fatehi and Safizadeh (1994) investigated the impact of social and political change on FDI in 15 of the least developed countries (1950-1982). Borensztein, De Gregorio and Lee (1998) researched the impact of FDI in 69 countries (1970-1989). Hejazi and Safarian (2001) established that trade and foreign direct investment (FDI) are complementary, using trade and FDI stock data on a bilateral basis between the U.S. and 51 other countries over the period 1982 to 1994. Noorbakhsk, Paloni and Youssef (2001) assessed the impact of FDI on human capital in 36 countries (1980-1994). Globerman and Shapiro (2002) assesses the impact of government policy, human capital and the environment on FDI in 144 countries (1995-1997). Bengoa and Sanzhez – Robles (2003) conducted an analysis of the interaction between FDI, economic growth and economic freedom in 18 Latin American countries (1970-1999). Bevan and Estrin (2004) studied the determinants of FDI in 11 European countries in transition (1994 - 2000). Bevan A., Estrin S., Meyer K. (2004) analyzed the impact of FDI on institutional development in 12 countries in transition (1994-1998). Alfaro, Chanda, Kalemli – Ozcan and Sayek (2004) investigated in their paper the impact of local financial markets on economic growth and FDI in 71 states (1975-1995). Durham (2004) measured the impact of FDI and foreign portfolio

71 states (1975-1995). Durham (2004) measured the impact of FDI and foreign portfolio investment on economic growth in 83 countries (1979-1998). Li and Liu (2005) assessed the impact of FDI on economic growth in 84 states (1970-1999). Agosin and Machado (2005) measured the impact of FDI on domestic investment in 12 countries (1971-2000). Schneider (2005) observed the interaction between international trade, economic growth and intellectual property rights in 47 countries (1970-1990). Vadlamannati and Tamazian (2009) measured the impact of FDI on economic growth in 80 countries (1980-2006). Kinda (2010) analyzed the correlation between FDI and the investment climate in 77 countries (2000-2006). Azman - Saini, Law and Ahmad (2010) analyzed the interaction between local financial markets and FDI in 91 countries (1975-2005). Doytch and Uctum (2011) investigated the impact of FDI on the growth of production and services in 60 countries (1990-2004). Fillat and Woerz (2011) assessed the impact of FDI on productivity growth in 35 countries (1987–2002). Barassi and Zhou (2012) analyzed the effect of corruption on the incentives of multinational enterprises to undertake FDI in a given country. Morrissey and Udomkerdmongkol (2012) assessed the interaction between private investment, FDI and government in 46 countries (1996-2009). Tintin (2013) analyzed the determinants of FDI in 6 European countries (1996-2009). Fereidouni (2013) studied the effect of the environment on FDI in 31 emerging economies (2000-2008). Thangavelu and Narjoko (2014) analyzed the relationship between free trade agreements and FDI in 39 countries (2000-2009). Imai, Gaiha, Ali and Kaicker (2014) measured the impact of FDI on economic growth in 24 countries (1980-2009). Goswami and Haider (2014) investigated the impact of political risk on FDI in 146 countries (1984–2009). Samargandi, Fidrmuc and Ghosh (2015) looked at the interaction between FDI and economic growth in 52 countries (1980 - 2008). Gui - Diby and Renard (2015) analyzed the relationship between industrialization and FDI in 49 countries (1980 - 2009). Broștescu (2018) studied a classic model of foreign direct investment for Romania and Bulgaria. Broștescu and Săvoiu (2019) investigated structural models based on the association of foreign direct investment flows and outflows (FDI) in some countries in ex-socialist, central and eastern Europe. Sujit, Kumar and Oberoi (2020) conducted an analysis on the impact of macroeconomic, governance and risk factors on FDI intensity. Sultana and Turkina (2020), studied the link between foreign direct investment, technological progress and absorption capacity. Burlea-Schiopoiu, Broștescu and Popescu (2021) analyzed the impact of foreign direct investment on the economic development of 10 emerging countries of the European Union in the period 2007-2017, including Romania and Bulgaria.

The stock of FDI is also studied. Thus, Kornecki and Raghavan (2011) estimate the impact of the FDI stock on economic growth in Central and Eastern Europe (CEE) during the post-communist era using a regression growth model based on the production function. Cardamone and Scoppola (2015) assessed the impact of tariffs on the European Union's external FDI stocks, using a sample of five EU countries and 24 partner countries in the period 1995-2008. Anghelache and Anghel (2015) performed an analysis of the dynamics of the FDI balance correlated with the evolution of GDP at European level. Dauti (2015) presented the main determinants of FDI stocks in 5 countries in south-eastern Europe and 10 new EU countries using an augmented gravity model in order to calculate potential levels of FDI stocks in Macedonia.

The main purpose of this paper is the statistical-econometric study of FDI input stocks in two member countries of the European Union, namely Bulgaria and Romania. Another goal of our research is to analyze a set of indicators from 1995 to 2018, which have the potential to predict the development of FDI stocks in terms of each of the economies of the countries analyzed. In the econometric modeling undertaken, export, import, import diversification index, trade balance, balance of payments, as well as some components of the economic freedom indicator for each country were considered as exogenous factors.

The novelty of the study consists in finding original and valid econometric models that explain the evolution of the stock of FDI inputs according to indicators of international trade, economic trends and components of the signaling indicator given by the index of economic freedom. Also, based on this research, we can identify measures that can be taken to increase the volume of FDI in Bulgaria and Romania.

The paper is structured as follows. The introductory part presents the notion of foreign direct investment and a classification of the factors that influence it. It also presents the current state of knowledge in the field through the review of the literature.
The second section of the paper presents the methods of research, the statistical indicators used in the study, as well as the statistical tests used in the validation of econometric models.

The third section contains the original results obtained by the authors, the interpretation of the results and possible measures that can be taken to increase the stock of foreign direct investment in the two countries analyzed. The paper ends with the section of conclusions, where a comparative analysis of the two countries is made in the light of the results obtained in the previous section, as well as possible further developments.

2. Research methodology

In this paper, multiple linear regression was used in order to find statistical-econometric relationships between the dependent variable and the independent variables, but also influenced between these variables (see Table 2.1). Our sample consists of two emerging markets in Bulgaria and Romania. The data analyzed cover the period 1995-2018 (see Annex 1).
In the econometric models from this paper, the stock of foreign direct investment is considered as a dependent variable, calculated as a percentage of gross domestic product, and as independent variables we have indicators of international trade, economic trends and components of the signaling indicator given by the economic freedom index (ILE). The indicators used in the study are presented in Table 2.1 and are calculated as a percentage when are used in the regression model for each country.

Table 2.1 Economic indicators used in the econometric modeling of stock FDI inflows.

Indicators	Unit	Abbreviations	Short description
Stock FDI inflows (% GDP)	%	$S_{FDI\_in(\%GDB)}$	The stock of FDI inflows represents the value of the share of capital and reserves (including retained profits) attributable to the parent company, plus the net debt of the affiliates to the parent companies. It is approximate the accumulated value of past FDI flows. This indicator is calculated as a percentage of GDP.
Exports (% global total)	%	$X_{\%G}$	Exports include all goods that leave the free movement of a country. This indicator is calculated as a percentage of the total globally.
Imports (% global total)	%	$M_{\%G}$	Imports include all goods entering the free zone of a country. This indicator is calculated as a percentage of the total globally.
Import diversification index	%	Id_M	This indicator is a modified Finger-Kreinin measure of similarity in trade, which takes values between 0 and 1. A value closer to 1 indicates a greater divergence from the global pattern of imports.
Trade balance (% imports)	%	$BC_{\%M}$	The trade balance is calculated as the difference between exports and imports, this indicator being expressed as a percentage of imports.
Balance of payments, current account balance, (% GDP)	%	$BP_{\%GDB}$	The balance of external payments is a system of accounts that includes the synthesis of economic and financial transactions of an economy with the rest of the world, over a period of time. The current account is part of the balance of payments and displays the flows of goods, services, primary and secondary income between residents and non-residents of an economy. The current account balance generally measures the difference between current receipts and expenses for internationally traded goods and services. At the same time, from a national perspective, the current account balance is the gap between domestic savings and investment
ILE - government integrity	%	ILE_{IG}	The ILE component that describes corruption that erodes economic freedom by introducing insecurity and coercion into economic relations. The score for each country is a number between 0 and 100, with the value 100 indicating the lowest level of corruption.

ILE - tax burden	%	ILE_{PF}	The ILE component that reflects marginal tax rates on both personal and corporate income and the general level of taxation (% of GDP), including direct and indirect taxes imposed by all levels of government. The score for each country is a number between 0 and 100, with the value 100 indicating the lowest level of taxation.
ILE – business freedom	%	ILE_{LA}	The ILE component that measures the degree to which regulatory environments and infrastructure constrain the efficient operation of business. The business freedom score for each country is a number between 0 and 100, with the value 100 indicating the freest business environment

Source*:* Authors' contribution based on the information available online at:
https://unctadstat.unctad.org/wds/ReportFolders/reportFolders.aspx
https://www.heritage.org/trade/report/2018-index-economic-freedom-freedom-trade-key-prosperity

It should be noted that other indicators were initially introduced in the study, such as the export diversification index, the import and export concentration index and other components of the economic freedom index, but no valid econometric models were found containing these indicators as exogenous variables..

The original econometric models presented in this paper rigorously went through the stages of specification, parameterization, testing and decision, with an emphasis on validation. Thus, the estimation of the parameters used the least squares method, aiming in the final models to obtain high values of the coefficient of determination (R-squared quantifying the percentage by which the influence of significant factors is explained, and adjusted R-squared representing a corrected value of R-squared, an increase possible and may sometimes be due to the number of variables in the model). The main tests used in the models were t-Student (with the null hypothesis H0: the coefficients are not significantly different from zero and the alternative hypothesis H1: the coefficients are significantly different from zero), the F test (checking if at least one coefficient is significantly different from zero, null hypothesis H0: all coefficients are not significantly different from zero, and H1: there is at least one non-zero coefficient), Durbin-Watson test to verify model error autocorrelation, Jarque-Bera test to prove whether model errors follow or not a normal distribution and the White test to verify the homoscedasticity or heteroskedasticity of econometric models.

By combining the indicators whose metadata were specified in Annex 1, a series of multifactor models were validated, with high values of the coefficient of determination, calculated F-statistic> tabulated F-statistic and Durbin-Watson test values in the range which errors are independent.

In order to validate the following proposed models, performed by means of the least squares method, the values of the tests on the significance of the independent variables, their influence on the evolution of the dependent variable, the verification of the asymmetry and kurtosis properties of the residual variable series, (their autocorrelation analysis or independent), as well as the verification of the homoskedasticity hypothesis, must satisfy the following conditions:

Test F:

- for models with two independent variables: F_calculated > 3.47 and associated probability less than 0.05;
- for models with three independent variables: F_calculated > 3.10 and associated probability less than 0.05;

Student *t* test:

- for models with two independent variables: |t_calculated|> 2.08 and associated probability less than 0.05;

- for models with three independent variables: |t_calculated|> 2,086 and an associated probability less than 0.05;

Durbin-Watson test:

- for models with two independent variables, the value must be between 1.55 - 2.45, the range in which the errors are independent;
- for models with three independent variables, the value must be between 1.66 - 2.34, the range in which the errors are independent;

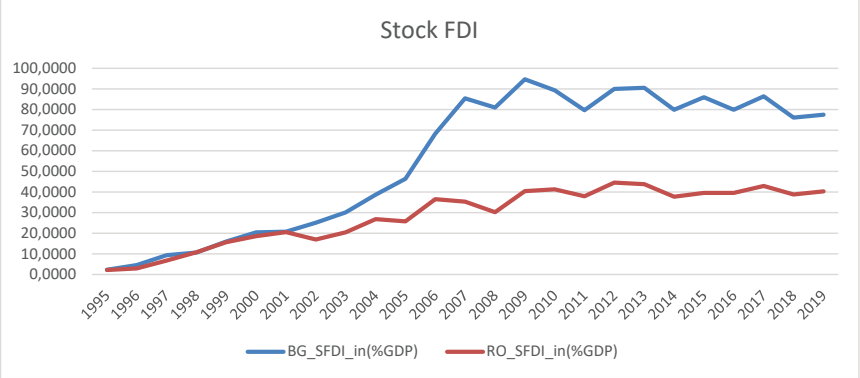
Probability associated with the Jarque-Bera test: greater than 0.05;
Probability associated with the White test: greater than 0.05.

3. Econometric models of the stock of FDI inflows in Bulgaria and Romania

Romania and Bulgaria had a similar history in terms of economic and political regime before 1989, and the transformations that took place after 1990 in the process of transition to a market economy and EU accession in 2007 support the comparative analysis between these countries. According to the study "Privatization and Restructuring in Central and Eastern Europe" conducted by the World Bank, in 1997, only 15% of companies in the Romanian manufacturing industry were privatized in 1995 and 8% of Bulgarian companies, compared to other Central European countries, such as Poland, Hungary and the Czech Republic where the percentage was over 60%, Privatization determines both costs and benefits. The costs consist of many restructurings and massive increases in unemployment, and the benefits are given by increasing employee productivity and attracting FDI. Consequently, emerging countries with massive and rapid privatizations have benefited from increased labor productivity and an increased volume of FDI. The slowdown in this process in Romania and Bulgaria is one of the causes of the reduced attraction of FDI.

Figure 1 shows the evolution of stocks of FDI inflows (% of GDP) in the economies of Bulgaria and Romania, in the period 1995 - 2019:

Figure 1. Evolution of stocks of FDI inflows (% of GDP) in the economies of Bulgaria and Romania in the period 1995 – 2019



Source: Authors' contribution based on the information available online at:
https://unctadstat.unctad.org/wds/TableViewer/tableView.aspx?ReportId=96740

Based on the data in Figure 1, it can be seen that Bulgaria had a very good evolution of the FDI stock, reaching 94.67% of GDP in 2009, being the highest value recorded in the analyzed period. Romania left the lowest FDI stock as a percentage of GDP of only 2.18%, registering higher values since 2006, the maximum value of 44.59% of GDP being reached in 2012 (see Annex 1).
Next, it was decided to build multifactorial econometric models for the two countries with stocks of FDI inflows, calculated as a percentage of GDP in the position of dependent variables, in an attempt to determine the factors with major influence on them, based on indicators of international trade, economic trends and components of the signaling indicator given by the economic freedom index. The database used in the construction of the econometric models proposed in this paper can be found in Annex 1 containing data from 1995-2018.

3.1. Econometric models of the FDI stock in Bulgaria – BG

Bulgaria has maintained its macroeconomic and financial stability amid a restrictive fiscal and monetary policy, but with a negative impact on the living standards of the population and sometimes the business environment.
The increase of the attractiveness of the business and institutional environment was achieved by practicing a profit rate of 10%, dividend tax of 5% and income tax from independent activities of 15%. According to the Bulgarian Investment Agency (www.investbg.government.bg), this country has undergone an accelerated privatization process in recent years, so that currently about 88% of the country's economy belongs to the private sector. The structure of Bulgarian exports includes raw materials 42.6%, consumer goods 28.1%, investment goods 16.4%, mineral products and energy 12.8%. As external partners, for export, in order of its volume are: Germany, Italy, Romania, Turkey, Greece. At import, the share is held by: raw materials 33.7%, investment goods 25.1%, consumer goods 20.8%, petroleum products and electricity 20.1%. According to the volume of imports, the main partners are the Russian Federation, Germany, Italy, Romania, Greece, Turkey. Bulgaria has improved its investment legislation in order to promote and facilitate priority and green field investments.

Table 3.1 Model of $S_{FDI\_in(\%GDP)}$ by $X_{\%G}$, Id_M and ILE_{LA} for BG

Variables	Coefficient	Std. Error	t-Statistic	Prob
C	-3.200829	29.74237	-0.107618	0.9154
$X_{\%G}$	333.4586	73.31514	4.548291	0.0002
Id_M	-244.2447	46.97849	-5.199075	0.0000
ILE_{LA}	1.549448	0.311065	4.981109	0.0001
R-squared	0.961736	Mean dependent var		54.63978
Adjusted R-squared	0.955997	S.D. dependent var		33.85309
S.E. of regression	7.101328	Akaike info criterion		6.909453
Sum squared resid	1008.577	Schwarz criterion		7.105795
Log likelihood	-78.91343	F-statistic		167.5638
Durbin-Watson stat	2.246595	Prob(F-statistic)		0.000000
Jarque-Bera	0.514481	Prob(J-B)		0.773182
Skewness	-0.357081	Kurtosis		2.933275
White Heteroskedasticity Test:				
F-statistic	0.574957	Probability		0.796543
Obs*R-squared	6.476829	Probability		0.691408

Source: Authors' contribution with software EViews

$S_{FDI\_in(\%GDP)}$=-3.200829+333.4586*$X_{\%G}$-244.2447*Id_M+1.549448*ILE_{LA}+ε_i [1]

Table 3.1 shows that the factor with the greatest influence on the stock of FDI inputs is exports, whose percentage increase with 1% in this case generates a 333% increase in the stock of FDI while the others remain constant. Similarly, the increasing variation of Id_M by 1% leads to a decrease in the stock of FDI inflows by 244%. At the same time, it is found that a better score of the ILE_{LA} component has as an effect an increase of the stock of the studied dependent variable. The model specified, parameterized and tested in the previous table explains the phenomenon studied in the Bulgarian economy in proportion of 95,5%.

Table 3.2 1 Model of $S_{FDI\_in(\%GDP)}$ by $M_{\%G}$, Id_M and $BC_{\%M}$ for BG

Variables	Coefficient	Std. Error	t-Statistic	Prob
C	110.9777	41.14887	2.696980	0.0139
$M_{\%G}$	464.2059	91.68981	5.062787	0.0001
Id_M	-336.8783	82.54470	-4.081163	0.0006
$BC_{\%M}$	0.737036	0.190484	3.869271	0.0010
R-squared	0.927339	Mean dependent var		54.63978
Adjusted R-squared	0.916440	S.D. dependent var		33.85309
S.E. of regression	9.785820	Akaike info criterion		7.550758
Sum squared resid	1915.245	Schwarz criterion		7.747100
Log likelihood	-86.60909	F-statistic		85.08390
Durbin-Watson stat	2.052333	Prob(F-statistic)		0.000000
Jarque-Bera Test	1.280355	Prob(J-B)		0.527199
Skewness	0.388964	Kurtosis		2.178305
White Heteroskedasticity Test:				
F-statistic	0.888640	Probability		0.558022
Obs*R-squared	8.725717	Probability		0.462969

Source: Authors' contribution with software EViews

$S_{FDI\_in(\%GDP)}$=110.9777+464.2059*$M_{\%G}$-336.8783*Id_M+0.737036*$BC_{\%M}$+ε_i [2]

For the model proposed in the previous table, it is found that in Bulgaria, the explanatory variables with the greatest impact on the stock of FDI inflows (% of GDP) are imports and their diversification index. If in the case of the increase of imports by a percentage, in the conditions in which Id_M and $BC_{\%M}$ remain constant, the stock of FDI has an increase of 464%, a pronounced divergence of imports compared to the global pattern leads to the decrease of the variable explained by 337%. The trade balance (% of imports) positively influences Bulgaria's FDI stock, which means that a higher value of exports automatically leads to an increase in the trade balance and $S_{FDI_}$in (% of GDP) by 0.74%, even if this increase is not significant. The adjusted coefficient of determination demonstrates that the proposed model for BG explains 91.64% of the investigated economic phenomenon.

Table 3.3 Model of $S_{FDI\_in(\%GDP)}$ by $M_{\%G}$, $BP_{\%GDP}$ and ILE_{LA} pentru BG

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	-131.6941	12.82042	-10.27222	0.0000
$M_{\%G}$	489.8016	65.62289	7.463884	0.0000
$BP_{\%GDP}$	0.547394	0.224921	2.433724	0.0244
ILE_{LA}	1.782159	0.300480	5.931047	0.0000
R-squared	0.958796	Mean dependent var		54.63978
Adjusted R-squared	0.952615	S.D. dependent var		33.85309
S.E. of regression	7.369173	Akaike info criterion		6.983500
Sum squared resid	1086.094	Schwarz criterion		7.179842
Log likelihood	-79.80200	F-statistic		155.1286
Durbin-Watson stat	1.840211	Prob(F-statistic)		0.000000
Jarque-Bera Test	1.789993	Prob(J-B)		0.408609
Skewness / Asimetrie	-0.661815	Kurtosis / Aplatizare		2.805068
White Heteroskedasticity Test:				
F-statistic	0.301528	Probability		0.961930
Obs*R-squared	3.896791	Probability		0.918073

Source: Authors' contribution with software EViews

$S_{FDI\_in(\%GDP)}$=-131.6941+489.8016*$M_{\%G}$+0.547394*$BP_{\%GDP}$+1.782159*ILE_{LA}+ε_i [3]

The variables of the model specified for the Bulgarian economy in Table 3.3 explain in proportion of 95.26% the variation of the stock of FDI inputs (% of GDP) based on the adjusted coefficient of determination. At the same time, it is found that all three have a positive influence on the studied dependent variable, so that an increase of a percentage of $M_{\%G}$ leads to an increase of the FDI stock by 489.80%, as in the previous model, if $BP_{\%GDP}$ % and ILE_{LA} remain constant.

Table 3.4 Model of $S_{FDI\_in(\%GDP)}$ by $M_{\%G}$, ILE_{PF} and ILE_{LA} for BG

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	-145.5968	9.467480	-15.37862	0.0000
$M_{\%G}$	217.2547	56.79942	3.824945	0.0011
ILE_{PF}	0.802021	0.155660	5.152398	0.0000
ILE_{LA}	1.653140	0.222919	7.415862	0.0000
R-squared	0.977053	Mean dependent var		54.63978
Adjusted R-squared	0.973610	S.D. dependent var		33.85309
S.E. of regression	5.499392	Akaike info criterion		6.398164
Sum squared resid	604.8662	Schwarz criterion		6.594506
Log likelihood	-72.77797	F-statistic		283.8519
Durbin-Watson stat	2.174807	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.392595	Prob(J-B)		0.821768
Skewness	-0.178951	Kurtosis		2.485704
White Heteroskedasticity Test:				
F-statistic	1.794266	Probability		0.160176
Obs*R-squared	9.305556	Probability		0.157109

Source: Authors' contribution with software EViews

$S_{FDI\_in(\%GDP)}$=- 145.5968+217.2547*$M_{\%G}$+0.802021*ILE_{PF}+1.653140*ILE_{LA}+ε_i [4]

The S_{FDI}_in model (% GDP) according to $M_{\%G}$, ILE_{PF} and ILE_{LA} for BG has a high coefficient of determination which explains about 97.36% of the variation of the investigated phenomenon. It is found, based on the data in Table 3.4, that an increased input of goods entering the territory of Bulgaria also leads to an increase in the stock of FDI ($M_{\%G}$ coefficient of 217.25%). Along with imports, it can be seen that an improvement in the score of the ILE component - business freedom by 1% leads to an increase in the dependent variable by 1.65%. However, the decrease of tax rates for both personal and corporate income does not significantly affect the stock of FDI inflows (% of GDP), a change of one percentage reflecting an increase of only 0.80% in the dependent variable.

Table 3.5 Model of $S_{FDI\_in(\%GDP)}$ by Id_M, ILE_{IG} and ILE_{LA} for BG

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	-104.1019	42.99991	-2.420981	0.0251
Id_M	-175.9396	58.55131	-3.004879	0.0070
ILE_{IG}	1.851335	0.481401	3.845720	0.0010
ILE_{LA}	2.363019	0.280941	8.411099	0.0000
R-squared	0.955250	Mean dependent var		54.63978
Adjusted R-squared	0.948538	S.D. dependent var		33.85309
S.E. of regression	7.679672	Akaike info criterion		7.066042
Sum squared resid	1179.547	Schwarz criterion		7.262385
Log likelihood	-80.79251	F-statistic		142.3099
Durbin-Watson stat	1.844073	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.427592	Prob(J-B		0.807513
Skewness	0.003639	Kurtosis		2.346135
White Heteroskedasticity Test:				
F-statistic	0.459235	Probability		0.878663
Obs*R-squared	5.470365	Probability		0.791531

Source: Authors' contribution with software EViews

$S_{FDI\_in(\%GDP)}$=- 104.1019-175.9396*ID_M+1.851335*ILE_{IG}+2.363019*ILE_{LA}+ε_i [5]

The model proposed and validated by the Bulgarian economy in Table 3.5 explains the change in the stock of FDI inputs (% of GDP) in the proportion of 94.85%. Improving the quality of each of the two components of the ILE - government integrity and business freedom by a percentage has a positive effect on the dependent variable, a low degree of corruption and an effective business regulatory environment reflected in a stock of FDI inflows. (% GDP) by 1.85% higher, respectively 2.36%. In contrast to ILE components, an increased divergence of imports from the global pattern has a significant negative effect on the volume of the dependent variable studied, leading to a decrease of 175.94% as found in all previously validated models.

From the data presented in Annex 1 it is found that exports (calculated as a percentage of the global total) have subunit values, so it is more realistic an increase of 0.01% per year which leads to an increase of 3.33% of the FDI stock. Similarly, a 0.01% increase in imports (calculated as a percentage of the global total) leads to an increase in FDI stocks with values between 2.17% and 4.48%, as shown by previous econometric models. Regarding the index of diversification of imports, it has subunit values and a decrease of 0.01% can generate an increase in the stock of FDI with values between 1.75% and 3.36%.

3.2. Econometric models of the FDI stock in Romania-RO

According to a report on FDI in Romania in 2017 made to the Council of Foreign Investors (FIC) together with the Academy of Economic Studies (ASE) in Bucharest (https://fic.ro/Documents/view/Raport-Investitiile-straine-directe-evolutia-si-importanta-lor-in-Romania) Foreign investments in Romania have significantly contributed to the modernization of the national economy and their integration into the European economy and international production chains. FDI companies employ a third of Romania's private workforce, approximately 1.2 million people. FDI companies have a labor productivity twice as high as those with Romanian capital and invest twice as much in each employee. FDI companies make on average 70% of Romania's exports, but also 60% of its imports. Although the perception is that the volume of FDI is high, Romania has the lowest stock of FDI per capita in the region (3,130 EUR / inhabitant). The period of preparation and accession to the EU overlaps the most favorable periods for attracting FDI in Romania. FDI flows increased more than 5 times between 2003-2008. The Netherlands, Austria and Germany are the most important economies investing in Romania, holding a share of over 50% of the total FDI stock. Almost half of total FDI has been directed to industry, which should lead to significant volume and long-term investigations. The share of gross value added of multinational companies exceeds 60% in industries such as automotive and IT&C, according to Eurostat (FATS) data. Romania is on the last positions in the region among some indicators such as infrastructure, tertiary education and vocational training, labor market efficiency. The single share of profit and income tax is 16% and the dividend tax is 5%. From the above, it is important to find some determinants that influence the stock of FDI.

Table 3.6 Model of $S_{FDI\_in(\%GDP)}$ by $M_{\%G}$, Id_M and ILE_{PF} for RO

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	42.53046	26.94568	1.578378	0.1302
$M_{\%G}$	-57.29563	18.59278	-3.081606	0.0059
Id_M	-140.8939	54.82854	-2.569718	0.0183
ILE_{PF}	0.672961	0.136791	4.919636	0.0001
R-squared	0.954397	Mean dependent var		28.16701
Adjusted R-squared	0.947557	S.D. dependent var		13.68928
S.E. of regression	3.134904	Akaike info criterion		5.274086
Sum squared resid	196.5525	Schwarz criterion		5.470428
Log likelihood	-59.28903	F-statistic		139.5238
Durbin-Watson stat	1.958780	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.112713	Prob(J-B)		0.945202
Skewness	-0.121135	Kurtosis		2.767583
White Heteroskedasticity Test:				
F-statistic	1.593099	Probability		0.209663
Obs*R-squared	12.14308	Probability		0.205363

Source: Authors' contribution with software EViews

$$S_{FDI\_in(\%GDP)} = 42.53046 - 57.29563 * M_{\%G} - -140.8939 * ID_M + 0.672961 * ILE_{PF} + \varepsilon_i \quad [6]$$

In the first model of the FDI stock (% of GDP) at the level of the Romanian economy, the following were used as explanatory variables: imports (global total%), the import diversification index and the ILE component (fiscal burden). The model specified in the table 3.6 explains in proportion of 94.76% the variation of the investigated economic phenomenon. The variable that produces the most significant changes in the $S_{FDI_}$in level (% of GDP) is the index of import diversification, thus, a higher value by a percentage of its value, which means a more pronounced divergence from the world import pattern, has the effect of a decrease with 140% of $S_{FDI_}$in (% GDP). The second variable with significant effect is represented by the level of goods entering the territory of Romania, i.e. imports, calculated as a percentage of the global total. The increase of this variable by one percentage, entails a decrease of the FDI stock (% of GDP) by 57.29563%. The third variable present in the model, the ILE component - the tax burden, does not have a significant impact on the target variable, an improved result with a percentage bringing an increase of $S_{FDI_}$in (% GDP) of only 0.672961%.

Table 3.7 Model of $S_{FDI\_in(\%GDP)}$ by $X_{\%G}$, Id_M for RO

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	67.32861	18.28979	3.681212	0.0014
$X_{\%G}$	50.58041	22.99035	2.200071	0.0391
Id_M	-173.6302	40.83534	-4.251960	0.0004
R-squared	0.910992	Mean dependent var		28.16702
Adjusted R-squared	0.902516	S.D. dependent var		13.68928
S.E. of regression	4.274136	Akaike info criterion		5.859509
Sum squared resid	383.6329	Schwarz criterion		6.006766
Log likelihood	-67.31411	F-statistic		107.4675
Durbin-Watson stat	1.800891	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.960553	Prob(J-B)		0.618612
Skewness	-0.483017	Kurtosis		3.165323
White Heteroskedasticity Test:				
F-statistic	2.272573	Probability		0.091021
Obs*R-squared	9.287540	Probability		0.098130

Source: Authors' contribution with software EViews

$$S_{FDI\_in(\%GDP)} = 67.32861 + 50.58041 * X_{\%G} - 173.6302 * ID_M + \varepsilon_i \quad [7]$$

Table 3.7 shows that the factor with a positive influence on the stock of FDI inputs is exports, the variation of which increases by one percent in this case generates an increase of 50.5%. Similarly, the 1% increase in Id_M leads to a 173% decrease in the stock of FDI inflows. The model specified, parameterized and tested in the previous table explains the phenomenon studied at the level of the Romanian economy in proportion of 90,2%.

Table 3.8 Model of $S_{FDI\_in(\%GDP)}$ by Id_M, $BC_{\%M}$ for RO

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	112.3385	4.759749	23.60176	0.0000
Id_M	-249.2065	14.58327	-17.08852	0.0000
$BC_{\%M}$	0.349923	0.085254	4.104453	0.0005
R-squared	0.939229	Mean dependent var		28.16702
Adjusted R-squared	0.933441	S.D. dependent var		13.68928
S.E. of regression	3.531703	Akaike info criterion		5.477906
Sum squared resid	261.9314	Schwarz criterion		5.625163
Log likelihood	-62.73487	F-statistic		162.2790
Durbin-Watson stat	2.367805	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.462224	Prob(J-B)		0.793650
Skewness	-0.159219	Kurtosis		2.399315
White Heteroskedasticity Test:				
F-statistic	2.493616	Probability		0.069634
Obs*R-squared	9.821226	Probability		0.080462

Source: Authors' contribution with software EViews

$$S_{FDI\_in(\%GDP)} = 112.3385 - 249.2065 * ID_M + 0.349923 * BC_{\%M} + \varepsilon_i \quad [8]$$

For the model proposed in the previous table, it is found that in Romania, the index of import diversification has a negative impact on the FDI stock, as seen in previous models. The trade balance (% of imports) influences Romania's FDI stock in a positive way, which means that a higher value of exports, automatically leads to the increase of the trade balance and $S_{FDI_}$in (% GDP) by 0.34%, even if this increase is not one significant. The adjusted coefficient of determination demonstrates that the proposed model for RO explains 93.3% of the investigated economic phenomenon.

Table 3.9 Model of $S_{FDI\_in(\%GDP)}$ by Id_M, $BP_{\%PGD}$ for RO

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	111.4965	4.925257	22.63770	0.0000
Id_M	-257.8297	15.17434	-16.99116	0.0000
$BP_{\%GDP}$	0.877708	0.236579	3.709996	0.0013
R-squared	0.933840	Mean dependent var		28.16702
Adjusted R-squared	0.927539	S.D. dependent var		13.68928
S.E. of regression	3.684952	Akaike info criterion		5.562861
Sum squared resid	285.1563	Schwarz criterion		5.710118
Log likelihood	-63.75433	F-statistic		148.2067
Durbin-Watson stat	2.374869	Prob(F-statistic)		0.000000
Jarque-Bera Test	0.735363	Prob(J-B)		0.692338
Skewness	-0.376583	Kurtosis		2.589997
White Heteroskedasticity Test:				
F-statistic	2.537132	Probability		0.073881
Obs*R-squared	8.355984	Probability		0.079375

Source: Authors' contribution with software EViews

$$S_{FDI\_in(\%GDP)} = 111.4965 - -257.8297 * ID_M + 0.877708 * BP_{\%GDP} + \varepsilon_i \quad [9]$$

In the previous model, the place of trade balance was taken by the balance of payments, which has a positive influence on the stock of FDI. Thus, a 1% increase of $BP_{\%GDP}$ generates a 0.87% increase in the FDI balance, in conditions where Id_M remains constant. The econometric model and tested and validated in the previous table explains the phenomenon studied at the level of the Romanian economy in proportion of 92,7%.

Analyzing the data from Annex 1, we find that exports (calculated as a percentage of the global total) have subunit values, which leads to the conclusion that a growth of 0.01% per year is more sustainable, which generates a growth of 0.5% of the stock of FDI. Similarly, an increase of 0.01% in imports (calculated as a percentage of the global total), generates a decrease in FDI stocks by 0.57% as shown by previous econometric models. Regarding the index of diversification of imports, it has subunit values and a decrease of 0.01% can generate an increase in the stock of FDI with values between 1.4% and 2.57%.

4. Conclusions and further developments

For a more visible influence of the indicators used in previous econometric models, we made a summary table.

Table 4.1 Influences of international trade indicators, economic trends and components of the ILE signaling indicator in Bulgaria and Romania

Cauntry	$X_{\%G}$	$M_{\%G}$	Id_M	$BC_{\%M}$	$BP_{\%GDP}$	ILE_{IG}	ILE_{PF}	ILE_{LA}
BG	+	+	-	+	+	+	+	+
RO	+	-	-	+	+		+	

Source: Authors' contribution

In the models previously presented for Bulgaria, all independent variables positively influence the change in the stock of FDI inputs (% of GDP), except for the import diversification index, in which case this country must align with the world import pattern to generate increases in the studied dependent variable. The change in the growth of imports, exports, trade balance and balance of payments leads to an increase in the dependent variable $S_{FDI_}$in (% of GDP) in Bulgaria. Also, a higher score of the components of economic freedom, which is characterized by a business environment as free as possible, by a level of taxation as low as possible, as well as by a diminished degree of corruption, lead to an increase in the stock of FDI inputs in the Bulgarian economy.

At the level of the Romanian economy, based on the models previously validated by this country, it is found that the increase in the stock of FDI inflows (% of GDP) can be achieved by increasing the volume of exports to the global total, trade balance and balance of payments and a level of taxation for both personal and corporate income and the general level of taxation (% of GDP), including direct and indirect taxes imposed by all levels of government as low as possible, based on an increased score of the ILE component - the tax burden.

Related to imports, it is found that an increase along with a pronounced divergence from them in the world pattern lead to a decrease in the stock of FDI inflows (% of GDP) in Romania. To counteract this effect, Romania must align its imports as close as possible to the global model and reduce as much as possible their volume, relative to the global total.

The analysis of the summary table shows that exports (% of total global) had a positive influence on the growth of the FDI stock (% of GDP) in both countries. Regarding the imports (% of the global total) they had different influences. Thus, if in Bulgaria a demand for imports generates an increase in the stock of FDI, in Romania the effect is completely opposite. This correlation between the decrease of imports and the increase of the stock of foreign direct investments is beneficial for Romania and in accordance with the economic policies in this country.

Another indicator that has the same effects in both countries is the index of diversification of imports, which negatively influences the increase of the FDI stock, a greater divergence of imports compared to the world model generating the decrease of the FDI stock.
The trade balance and the balance of payments positively influence the FDI stock in both countries, which means that a higher value of exports also generates the increase of the trade balance and implicitly of the FDI stock. Similarly, the increase in the current account balance, which generally measures the difference between current receipts and expenditures for internationally traded goods and services, generates an increase in the FDI balance in both countries.
A sixth indicator with influence in both countries is the index of economic freedom through its component given by the fiscal burden. Thus, a higher value of this index, which indicates the lowest level of taxation, has the positive effect of increasing the stock of FDI.
The other components of the economic freedom index, namely government integrity and business freedom, have positive effects on Bulgaria, in the sense that a diminished level of corruption and a freer business environment generate an increase in the stock of FDI.
As further developments, we believe that it is important to study other emerging economies in Central and South-Eastern Europe in terms of the stock of FDI inputs, as well as the source countries of these FDI. Also, other exogenous variables can be taken into account in finding econometric models, which explain the evolution of the FDI stock. A study of the influence of the COVID crisis on FDI may be a future direction of research.

References

Agosin M., Machado R. (2005). Foreign Investment in Developing Countries: Does it Crowd in Domestic Investment?, *Oxford Development Studies*, 33 (2): 149 – 162.

Alfaro L., Chanda A., Kalemli – Ozcan S., Sayek S. (2004). FDI and economic growth: The role of local financial markets, *Journal of International Economics*, 64 (1): 89-112.

Anghelache, C., Anghel, M. (2015). Model for analyzing the dynamics of FDI balance correlated with the evolution of GDP at European level, *Romanian Statistical Review - Supplement*, 10: 73-78.

Azman – Saini W.N.W., Law S.H., Ahmad A.H. (2010). FDI and economic growth: New evidence on the role of financial markets, *Economics Letters*, 107(2): 211 – 213.

Barassi, R. M., Zhou, Y., (2012). The effect of corruption on FDI: A parametric and non-parametric analysis*, European Journal of Political Economy*, 28(3): 302-312.

Bengoa M., Sanchez–Robles B. (2003). Foreign direct investment, economic freedom and growth: New evidence from Latin America, *European Journal of Political Economy*, 19 (3): 529 – 545.

Bevan A., Estrin S. (2004). The determinants of foreign direct investment into European transition economies, *Journal of Comparative Economics*, 32(4): 775 – 787.

Bevan A., Estrin S., Meyer K. (2004). Foreign investment location and institutional development in transition economies, *International Business Review*, 13(1): 43 – 64.

Borensztein E., De Gregorio J., Lee J. W. (1998). How does foreign direct investment affect economic growth? *Journal of International Economics*, 45 (1): 115 – 135.

Brada, J.C, Drabek, Z., Mendez, Perez, M..F. (2019). National levels of corruption and foreign direct investment, *Journal of Comparative Economics*, 47 (1):31-49,
https://doi.org/10.1016/j.jce.2018.10.005.

Broștescu**,** S.I., (2018) A classic model of foreign direct investment (FDI), parameterized and validated, similar for Romania and Bulgaria, *Romanian Statistical Review Supplement*, 66(3): 155-164.

Broștescu, S.I., Săvoiu, G. (2019), Structural Models Based on Associating The Foreign Direct Investment (FDI) Inflows and Outflows, in Some of The Ex-Socialist, Central and Eastern European Countries (CEEC -10), *Romanian Statistical Review*, 1, 57-73.

Burlea-Schiopoiu A. Broștescu**, S.,** Popescu, L. (2021), The impact of foreign direct investment on the economic development of emerging countries of the European Union**,** *International Journal of Finance & Economics,* pp. 1-30, https://doi.org/10.1002/ijfe.2530

Caetano, J., Caleiro A., (2009). *Economic freedom and foreign direct investment: How different are the MENA countries from the EU,* University of Évora (Portugal), Department of Economics, Economics Working Paper 02/2009.

Cardamone, P., Scoppola, M., (2015). Tariffs and EU countries foreign direct investment: Evidence from a dynamic panel model, *Journal of International Trade & Economic Development*, vol. 24 (1), pp. 1-23.

Culem C. (1988). The locational determinants of direct investments among industrialized countries, *European Economic Review*, 32 (4): 885 – 904.

Dauti, B., 2015. Determinants of foreign direct investment in transition economies, with special reference to Macedonia: evidence from gravity model, *South East European Journal of Economics and Business*, 10 (2): 7-28.

Doytch N., Uctum M. (2011). Does the worldwide shift of FDI from manufacturing to services accelerate economic growth? A GMM estimation study, *Journal of International Money and Finance,* 30 (3): 410 –427.

Durham J.B. (2004). Absorptive capacity and the effects of foreign direct investment and equity foreign portfolio investment on economic growth, *European Economic Review*, 48(2): 285 – 306.

Fatehi K., Safizadeh H. (1994). The effect of sociopolitical instability on the flow of different types of foreign direct investment, *Journal of Business Research*, 31 (1): 65 – 73.

Feldstein M. (1983). Domestic saving and international capital movements in the long run and the short run, *European Economic Review*, 21 (1-2): 129 – 151.

Fereidouni H.G. (2013). Foreign direct investments in real estate sector and CO2 emission: Evidence from emerging economies, *Management of Environmental Quality*, 24 (4): 463 – 476.

Fillat C., Woerz J. (2011). Good or bad? The influence of FDI on productivity growth. An industry level analysis, *Journal of International Trade and Economic Development*, 20 (3): 293 – 328.

Globerman S., Shapiro D. (2002). Global foreign direct investment flows: The role of governance infrastructure, *World Development,* 30 (11): 1899 – 1919.

Goswami G.G., Haider S. (2014). Does political risk deter FDI inflow?: An analytical approach using panel data and factor analysis, *Journal of Economic Studies*, 41 (2): 233 – 252.

Greenaway, D., Kneller, R., (2007). Firm heterogeneity, exporting and foreign direct investment. *The Economic Journal*, 117 (517): 134–161.

Gui–Diby S.L. Renard M.F. (2015). Foreign Direct Investment inflows and the industrialization of African countries, *World Development*, 74: 43-57.

Hejazi, W., Safarian, A.E. (2001), The complementarity between U.S. foreign direct investment stock and trade. *Atlantic Economic* Journal, 29, 420–437. https://doi.org/10.1007/BF02299331

Imai K., Gaiha R., Ali A., Kaicker N. (2014). Remittances, growth and poverty: New evidence from Asian countries, *Journal of Policy Modelling*, 36 (3): 524 –538.

Kornecki, L. Raghavan, V. (2011), Inward FDI Stock and Growth in Central and Eastern Europe, *The International Trade Journal*, 25(5): 539-557, DOI: 10.1080/08853908.2011.604297

Kinda, T. (2010). Investment Climate and FDI in Developing Countries: Firm-Level Evidence, *World Development*, 38 (4): 498 –513.

Lacroix, J., Méon, P-G., Sekkat, K., (2021). Democratic transitions can attract foreign direct investment: Effect, trajectories, and the role of political risk, *Journal of Comparative Economics*, 49 (2): 340-357, https://doi.org/10.1016/j.jce.2020.09.003.

Lee, H-H., Rajan, R.S. (2011). Foreign Direct Investment in the APEC Region: the Role of Country Risks, *Journal of Korea Trade*, 15 (3): 89-123.

Li X., Liu X. (2005). Foreign Direct Investment and economic growth: An increasingly endogenous relationship. *World Development*, 33 (3): 393- 407.

Morrissey, O., Udomkerdmongkol, M. (2012). Governance, Private Investment and Foreign Direct Investment in Developing Countries, *World Development*, 40 (3): 437 – 445.

Noorbakhsh, F., Paloni, A., Youssef, A. (2001). Human Capital and FDI Inflows to Developing Countries: New Empirical Evidence, *World Development*, 29(9): 1593 – 1610.

Rožāns, E., (2016). The Impact of Economic Freedom on the Attraction of Foreign Direct Investment in the Baltics, *Journal of Economics and Management Research,* 4 (5): 74-94.

Samargandi N., Fidrmuc J., Ghosh S. (2015). Is the Relationship Between Financial Development and Economic Growth Monotonic? Evidence from a Sample of Middle-Income Countries, *World Development*, 68: 66-81

Săvoiu, G., Dinu, V., Ciucă, S., (2013). Foreign direct investment based on country risk and other macroeconomic factors. Econometric models for Romanian Economy, *Romanian Journal of Economic Forecasting*, 16 (1): 39-61.

Săvoiu, G., Ţaicu, M., (2014). Foreign Direct Investment Models, Based on Country Risk for Some Post-Socialist Central and Eastern European Economies, *Procedia Economics and Finance*, 10: 249 – 260.

Schneider P.H. (2005). International trade, economic growth and intellectual property rights: A panel data study of developed and developing countries, *Journal of Development Economics*, 78 (2): 529 –547.

Smits, W.J.B. (1988). Foreign direct investment and export and import value – A cross – section study for thirty less developed countries, *De Economist*, 136 (1): 91 – 117.

Sujit, K.S., Kumar, B.R., Oberoi, S.S. (2020), Impact of Macroeconomic, Governance and Risk Factors on FDI Intensity—An Empirical Analysis. *Journal of Risk and Financial Management*. 13 (12): 304.

Smits, W.J.B. (1988). Foreign direct investment and export and import value – A cross – section study for thirty less developed countries, *De Economist*, 136 (1): 91 – 117.

Sultana, N., Turkina,E. (2020), Foreign direct investment, technological advancement, and absorptive capacity: A network analysis, *International Business Review*, 29(2): 101668, https://doi.org/10.1016/j.ibusrev.2020.101668.

Thangavelu, S.M., Narjoko, D. (2014). Human capital, FTAs and foreign direct investments flows into ASEAN, *Journal of Asian Economics*, 35: 65-76.

Thomas, R., (2006). Foreign direct investment: An activity to assess country risk, *Teaching Business & Economics*, 10 (3):13-16.

Tintin, C. (2013). The determinants of foreign direct investment inflows in the Central and Eastern European Countries: The importance of institutions, *Communist and Post – Communist Studies*, 46 (2): 287 – 298.

Vadlamannati, K.C., Tamazian, A. (2009). Growth effects of FDI in 80 developing economies: The role of policy reforms and institutional constraints. *Journal of Economic Policy Reform*, 12(4): 299 – 322.

Vijayakumar, J., Rasheed, A., Tondkar, R. (2009). Foreign Direct Investment and Evaluation of Country Risk: an Empirical Investigation, *Multinational Business Review*, 17(3): 181-204. https://doi.org/10.1108/1525383X200900023

Wells, L., Wint, A., (2000). *Marketing a Country: Promotion as a Tool for Attracting Foreign Investment*, Occasional Paper no. 13, World Bank. World Development Indicators, Washington, D.C.

ANNEX 1.

BG	$S_{FDI\_in(\%GDB)}$	$X_{\%G}$	$M_{\%G}$	Id_M	$BC_{\%M}$	$BP_{\%GDB}$	ILE_{IG}	ILE_{PF}	ILE_{LA}
1995	2.3456	0.1035	0.1081	0.4342	-5.3887	-0.1788	30.00	46.00	55.00
1996	4.5098	0.0904	0.0923	0.4356	-3.6453	0.1553	30.00	50.60	55.00
1997	9.3607	0.0882	0.0867	0.4367	0.2028	3.8131	30.00	48.90	55.00
1998	10.6218	0.0780	0.0879	0.3772	-13.1206	-0.4231	30.00	53.50	55.00
1999	16.0138	0.0695	0.0935	0.3445	-27.3576	-4.8289	30.00	68.00	55.00
2000	20.4114	0.0752	0.0983	0.3819	-25.8554	-5.3500	29.00	67.90	55.00
2001	20.7661	0.0826	0.1135	0.3465	-29.6952	-5.7167	33.00	58.10	55.00
2002	25.1040	0.0882	0.1194	0.3583	-27.9194	-1.9601	35.00	68.30	55.00
2003	30.1314	0.0994	0.1397	0.3293	-30.6154	-4.8715	39.00	72.40	55.00
2004	38.6417	0.1077	0.1525	0.3321	-31.2804	-6.4378	40.00	81.10	55.00
2005	46.4338	0.1118	0.1685	0.3335	-35.3686	-11.2937	39.00	80.30	55.00
2006	68.3021	0.1242	0.1883	0.3461	-35.2623	-17.1787	41.00	83.20	70.50
2007	85.4337	0.1321	0.2105	0.2806	-38.1925	-25.7533	40.00	82.40	70.30
2008	80.9333	0.1385	0.2241	0.2742	-39.4114	-21.8246	40.00	82.70	68.40
2009	94.6652	0.1300	0.1855	0.2912	-30.6738	-8.2034	41.00	86.20	73.50
2010	89.2909	0.1348	0.1654	0.2950	-19.1409	-1.9060	36.00	86.30	77.80
2011	79.6291	0.1538	0.1767	0.2865	-13.4232	0.4746	38.00	86.90	75.80
2012	90.0083	0.1441	0.1753	0.2868	-18.4178	-0.9807	36.00	93.60	72.70

2013	90.4832	0.1561	0.1809	0.2986	-13.7827	1.2215	33.00	94.00	73.60
2014	79.9199	0.1539	0.1818	0.2920	-15.5992	1.3152	35.20	91.20	73.50
2015	85.9176	0.1532	0.1746	0.3208	-13.1292	-0.0380	41.00	91.00	68.50
2016	79.8797	0.1656	0.1786	0.3121	-8.1587	2.6535	43.00	91.10	66.90
2017	86.4442	0.1772	0.1902	0.3142	-8.0342	3.1728	41.80	91.00	66.70
2018	76.1073	0.1727	0.1911	0.2999	-11.1950	4.5635	38.20	90.90	64.30
2019	77.4909								
RO									
1995	2.1802	0.1528	0.1963	0.3953	-23.0388	-4.7110	10.00	39.40	55.00
1996	2.9520	0.1494	0.2080	0.3797	-29.3022	-6.9172	30.00	42.40	55.00
1997	6.7431	0.1506	0.1984	0.3946	-25.2542	-5.8707	30.00	44.30	55.00
1998	10.7848	0.1506	0.2099	0.3570	-29.7894	-6.9489	50.00	43.90	55.00
1999	15.6812	0.1488	0.1778	0.3661	-18.2269	-3.5845	34.00	45.00	55.00
2000	18.5715	0.1614	0.1976	0.3480	-20.8083	-3.6192	30.00	58.30	55.00
2001	20.4808	0.1839	0.2428	0.3517	-26.8147	-5.4744	33.00	57.60	55.00
2002	16.9931	0.2135	0.2679	0.3466	-22.2764	-3.3027	29.00	64.40	55.00
2003	20.3824	0.2327	0.3083	0.3442	-26.3551	-5.5305	28.00	69.10	55.00
2004	26.8783	0.2553	0.3449	0.3122	-27.9511	-8.3734	26.00	69.90	55.00
2005	25.7820	0.2636	0.3759	0.2884	-31.6669	-8.5663	28.00	70.10	55.00
2006	36.4818	0.2676	0.4141	0.2789	-36.5567	-10.4505	29.00	87.50	74.60
2007	35.2885	0.2888	0.4941	0.2600	-42.4177	-13.5982	30.00	85.90	73.20
2008	30.2164	0.3067	0.5104	0.2425	-41.0671	-11.6971	31.00	85.60	74.90
2009	40.4017	0.3231	0.4281	0.2611	-25.3239	-4.7704	37.00	87.00	74.90
2010	41.3291	0.3240	0.4027	0.2630	-20.1736	-5.0873	38.00	85.80	72.50
2011	37.8933	0.3437	0.4148	0.2603	-17.5797	-5.0316	38.00	86.80	72.00
2012	44.5855	0.3124	0.3763	0.2557	-17.6136	-4.7768	37.00	87.40	70.50
2013	43.7965	0.3475	0.3875	0.2762	-10.3995	-1.0840	36.00	87.90	70.40
2014	37.7066	0.3668	0.4080	0.2715	-10.3194	-0.6931	37.70	87.00	71.00
2015	39.5730	0.3660	0.4175	0.2661	-13.2173	-1.2117	43.00	86.90	69.80
2016	39.5602	0.3960	0.4603	0.2604	-14.7883	-2.1088	43.00	87.50	66.10
2017	42.9709	0.3989	0.4756	0.2620	-17.2243	-3.1892	45.90	87.40	65.90
2018	38.7755	0.4092	0.4934	0.2690	-18.5043	-4.4754	40.00	87.30	65.20
2019	40.3192								

https://unctadstat.unctad.org/wds/ReportFolders/reportFolders.aspx

https://www.heritage.org/trade/report/2018-index-economic-freedom-freedom-trade-key-prosperity