Romanian Statistical Review
Revista Română de Statistică

THE JOURNAL OF NATIONAL INSTITUTE OF STATISTICS

1/2023

www.revistadestatistica.ro

CONTENTS 1/2023

Demonstration of an Exploratory Method for Categorical Data Imputing Inventories Zero or Non-zero Values

Anri Mutoh (anry_1225@fuji.waseda.jp)
Rissho University, Tokyo, Japan

Ichiro Murata (imurata@nstac.go.jp)
National Statistics Center, Tokyo, Japan

ABSTRACT

In Unincorporated Enterprise Survey of Japan, the problem of poor accuracy in the imputation of missing values for Inventories was caused by the inclusion of many zero values. To address this issue, we used a strategy that involves identifying categorical variables at first that could potentially indicate whether Inventories are zero or non-zero. This was done through summary statistics S1 which was obtained through Exploratory Data Analysis. S1 is simply a summary statistic and is not designed to accurately predict the values of Inventories. Therefore, we conducted an experiment using real data to test whether S1 could guarantee the same accuracy as other methods that pursue higher accuracy in variable selection. The results showed that S1 was as accurate as these other methods. Additionally, the variables selected by the other methods were difficult to interpret, but the variables chosen by S1 were easily understandable based on practical experience.

Keywords: *Unincorporated Enterprise Survey, categorical data analysis, inventories, Exploratory Data Analysis*

JEL Classification: *C14*

1. INTRODUCTION

The National Statistics Center of Japan has been working to improve the imputation methods for missing values in Unincorporated Enterprise Survey. One challenge has been dealing with missing values for items such as Initial and Final inventories, which often have zero values (known as zero-inflation). This can decrease the accuracy of imputation for missing values.

To address this challenge, we have been searching for a method that is not overly complex, can easily be put into existing imputation processes, does not require advanced technical expertise to understand, and can be explained from a practical perspective when applied.

The classification of Inventories as zero or non-zero is influenced by characteristics represented by categorical variables such as industrial classification. Therefore, we have decided to use a two-step method for simple imputation of Inventories: first, impute zero values based on categorical variables that contribute to whether Inventories are zero or non-zero, and then apply ordinary imputation methods for non-zero samples.

To select the categorical variables that contribute to whether Inventories are zero or non-zero in the first step imputation process, we used summary statistics (Mutoh and Shirakawa, 2023) based on Exploratory Data Analysis (EDA). However, the accuracy of this method has not been guaranteed. Therefore, we conducted an experiment to examine the usefulness of these summary statistics through some comparisons with existing methods that prioritize accuracy.

In the following sections, the background of this study will be discussed in more detail, followed by a review of previous research on zero-inflated data and its analysis in the context of statistical imputation. Next, the summary statistics used to select categorical variables that contribute to whether Inventories are zero or non-zero will be introduced, and the experiment to demonstrate their validity will be described.

2. BACKGROUND

The purpose of Unincorporated Enterprise Survey is to obtain background information for administrative policy by surveying around 40,000 samples of solo proprietorships, which includes questionnaire items related to business owners, employees, problems in business management, actual business conditions such as annual operating revenues and expenditures, and so on.

Ideally, we would obtain accurate responses from the survey target. However, it is expected that some missing values will arise. Therefore, for key items such as the amount of sales used in the System of National Accounts in Japan, we prioritize the imputation of missing values in order to maintain the reliability of the statistics. The items for imputation include Sales, Purchases, (Initial/Final) inventories, Total Expenses, and Salaries.

In the case of sales, its value is imputed using the time-adjusted last observed data. Meanwhile, the nearest neighbor hot-deck method is used for the imputation of purchases, total expenses, and salaries (calculating distances by variables other than the imputation target). However, using the same method for Inventories leads to lower imputation accuracy due to the presence of many zero values.

In some enterprises, such as drop shipping stores or consulting businesses, there is no need for inventory. In these cases, the amount of inventory is not easily affected by sales and purchases. This results in poor accuracy when imputing a large value for inventory where it should be zero. When it comes to imputation for zero-inflated data, it is common to use methods that take quantitative variables as a reference (as described later). However, our previous analysis has shown that zero values for inventory are dependent on some categorical variables, particularly industrial classification.

Therefore, we are investigating the contribution of categorical variables to whether the amount of inventory is zero or not for imputation.

3. RELATED WORKS

3.1. Zero-inflated data and imputation

In the review by Min and Agresti (2002), various modeling methods for zero-inflated data are discussed. The main approaches include the Tobit model, which is useful for data that must take a certain number of values to be zero, and the two-part model, which estimates two independent models depending on whether the data are zero or non-zero. Other commonly-used models are based on mixture distributions such as the Zero-Inflated Poisson regression (ZIP) model and the Negative Binomial distribution (ZINB) model. Overall, the review explains a comprehensive overview of the approaches to zero-inflated data.

In the context of statistical imputation, the use of Multiple Imputation (MI) is prevalent in the analysis of zero-inflated data. This leads to the thought that the implementation of MI through the commonly-used ZIP and ZINB models is natural (e.g. Pahel et al. (2011) in the use of ZIP model for MI in dental caries data). Kleinke and Reinecke (2013) proposed an algorithmic implementation of MI for comprehensive zero-inflated data using the ZIP and ZINB models, and Chen and Haziza (2017) proposed a robust MI algorithm for model selection. Lee et al. (2020) proposed a method to improve the accuracy of MI by focusing on covariance.

In addition, Mbarek et al. (2015) proposed the use of Zero-Inflated Ordered Probit (ZIOP) for imputation in quantitative variables that can be considered discrete such as count data. For zero-inflated data with large variance, the ZINB model is appropriate, but for non-unimodal distributions, the Tweedie distribution (Jorgensen, 1987) can also be considered (with a concise maximum likelihood estimation).

One study that has a high affinity with our imputation process is by MacNeil et al. (2016). In their evaluation of cost-effectiveness in randomized controlled trials, costs are often difficult to obtain and missing as outcomes. In the paper, they compare four strategies – complete case analysis, predictive mean matching, log predictive mean matching, and two-step MI – in a simulation of missing data. The results showed that log predictive mean matching was able to maintain stable accuracy even with an increased number of zeros. However, in the Unincorporated Enterprise Survey data that we are trying to impute, there are many cases where quantitative explanatory variables cannot explain the increased number of zeros for Inventories. Therefore, we used the two-step method instead of log predictive mean matching in MacNeil et al. (2016).

3.2. **Summary measures of association in contingency tables**

Agresti (2002) is an appropriate textbook that comprehensively covers the history and system of summary measures of association in contingency tables. In addition to classical summary measures of association, the Proportional Reduction in Error (PRE), AIC (and BIC) are also comprehensively implemented in the pw.assoc() function of the "StatMatch" package in R (D'Orazio, 2006). Furthermore, measures of association based on AIC are also implemented in the R package "CATDAP" (Sakamoto, et al.), which can also consider associations in contingency tables of up to four dimensions.

Kendall and Stuart (1979) offer a broader explanation of measures of association in contingency tables through the concept of PRE. PRE is determined by measuring the decrease in variance between the objective variable's marginal distribution and its conditional distribution given an explanatory variable. Examples of measures of association that utilize PRE include "Goodman-Kruscal λ", "Goodman-Kruscal τ", and "Theil's uncertainty coefficient", among others.

3.3. **S1 - Summary Statistics for Categorical Variables**

S1 is a summary statistic for categorical variables based on Exploratory Data Analysis (EDA) proposed by Mutoh and Shirakawa (2023) (in press). The fundamental idea and the measure of association in contingency tables' PRE index are similar to those mentioned earlier, but the calculation is simpler, and it emphasizes the unilateral contribution of the explanatory variables to the response variable.

The definition of summary statistics S1 is the following. Let Y represent a categorical objective variable with possible values $y_i\,(i = 1, \ldots, k)$, and let X represent a categorical explanatory variable with possible values $x_j\,(j = 1, \ldots, l)$. Then,

$$SI_{Y|X} = \frac{1}{l}\sum_{j=1}^{l}\sum_{i=1}^{k}\left|P(Y = y_i) - P(Y = y_i|X = x_j)\right|.$$

The S1 statistic expresses the difference between the empirical proportion of the objective variable and the proportion grouped by the explanatory variables, which is conditional probability.

A high value of S1 indicates that the explanatory variable has a significant impact on the objective variable, while a low value suggests that it has minimal effect. In situations where the proportion of levels of the objective variable in the data is affected by the explanatory variables, a high value of S1 will be observed. Conversely, a low value indicates the proportion of levels of the objective variable remains unchanged regardless of the condition by the explanatory variables. It is worth noting that the minimum possible value of S1 is 0.

In the case that multiple explanatory variables are present, we can compute the summary statistics for each variable.

It should be noted that the S1 statistic does not necessarily reflect accuracy, but rather serves as a means for facilitating exploratory data analysis.

4. PURPOSE

The mentioned methods for Zero-inflated data analysis for imputation employ quantitative variables to iteratively calculate estimates of the mixture distribution. However, due to their complexity, their implementation within our current imputation process is difficult and their comprehension by the responsible workers poses a challenge. Moreover, as stated earlier, categorical variables such as industrial classification play a significant role in Inventories.

In this study, we propose a two-step method for imputing the values of inventories. The initial phase entails identifying a categorical variable that signifies the presence or absence of zero inventories. This variable is subsequently utilized in conjunction with the current imputation process to estimate the value of the Inventories.

To make this approach easily integrated into existing imputation processes, we aim to use a simple mechanism that can be stably calculated. In addition, we prioritize interpretability in order to facilitate understanding by those responsible for the imputation and to fulfill our accountability to the general public.

There are several methods for identifying variables that contribute to the zero value of Inventories, such as a forward-backward stepwise method using AIC, Lasso regression, and various measures of association in contingency table. However, these methods require specialized knowledge for understanding the mechanism and have a large calculation amount, making it difficult to say that they have high interpretability. Therefore, we consider using the summary statistics S1. However, the summary statistics S1 is based on the idea of EDA and prioritizes interpretability over accuracy, so we need to confirm that it should not fall behind existing accuracy-seeking methods.

Therefore, the objective of this study is to assess the summary statistics S1 in the context of inventory imputation from the perspective of both accuracy and interpretability.

5. ACCURACY ASSESSMENT EXPERIMENT AND RESULT

To assess the summary statistics S1 from the perspective of accuracy, we conducted an experiment.

The process of the experiment was as follows: Using the data from the Unincorporated Enterprise Survey in Japan 2019, the Initial inventory and Final inventory were classified into two values based on whether they were zero or not, and were treated as categorical variables consisting of two levels. These Inventories were considered as the objective variables and other categorical variables were treated as explanatory variables.

To compare with S1, the methods and indicators (and R packages) used for variable selection are as follows:

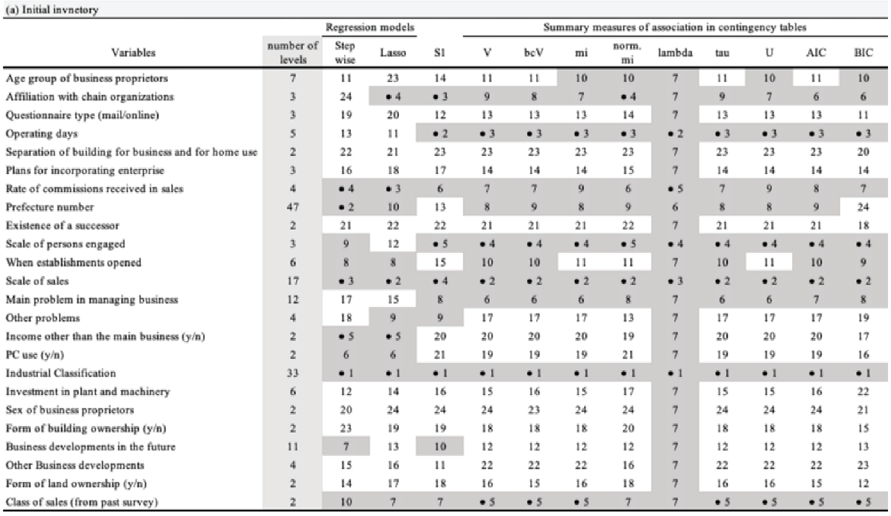
- ● Regression models
    - ♦ forward-backward stepwise method using AIC (Step Wise); “stats”
    - ♦ results of Lasso regression (Lasso); “glmnet”
- ● Summary measures of association in contingency tables
    - ♦ Cramer’s V (V); “StatMatch”
    - ♦ Bias-corrected Cramer’s V (bcV); “StatMatch”
    - ♦ mutual information (mi); “StatMatch”
    - ♦ normalized mutual information (norm.mi); “StatMatch”
    - ♦ Goodman-Kruscal λ (lambda); “StatMatch”
    - ♦ Goodman-Kruscal τ (tau); “StatMatch”
    - ♦ Theil’s uncertainty coefficient (U); “StatMatch”
    - ♦ AIC; “StatMatch”, “CATDAP”
    - ♦ BIC; “StatMatch”

When using categorical variables as explanatory variables in a regression model, they are treated as (number of levels - 1) dummy variables during estimation. However, in practice, it is preferred to treat the variable itself as a single unit rather than divided dummy variables. Therefore, in this experiment, the contribution of the selected dummy values was averaged to discuss variables itself as a single unit. The ranking of variables with high contributions was conducted by using each of the three methods for the objective variables representing whether each Inventories was zero or not. The accuracy on test data (10% of the whole sample) was calculated based on the top 5 variables with the highest contribution.

In Table 1, we present the categorical variables that we used as explanatory variables and their ranking based on the contribution to the objective variable according to the methods we applied. The results of our test data estimation using the variables with a high contribution to the objective variable by the Area Under the Curve (AUC) are provided in Table 2.

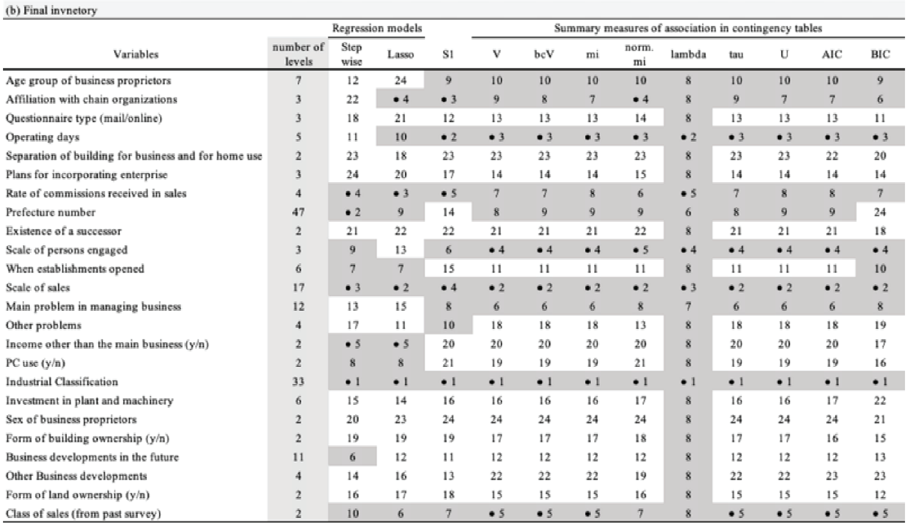
There are two types of inventories: Initial inventory and Final inventory. However, we obtained similar results in terms of variable selection and accuracy regardless of which of these variables we used as the objective variable. Therefore, unless otherwise stated, we will refer to Initial and Final inventories without distinguishing between them in the following discussion.

The ranking of the contribution of the categorical variable used as the explanatory variable, and the zero-non-zero value of (a)Initial inventory and (b)Final inventory by each method, is shown. The variables colored in gray are the top 10 variables in terms of contribution, and the black circles are the top 5 variables used as explanatory variables for testing data verification



(a) Initial invnetory

		Regression models			Summary measures of association in contingency tables									
Variables	number of levels	Step wise	Lasso	SI	V	bcV	mi	norm. mi	lambda	tau	U	AIC	BIC
Age group of business proprietors	7	11	23	14	11	11	10	10	7	11	10	11	10
Affiliation with chain organizations	3	24	● 4	● 3	9	8	7	● 4	7	9	7	6	6
Questionnaire type (mail/online)	3	19	20	12	13	13	13	14	7	13	13	13	11
Operating days	5	13	11	● 2	● 3	● 3	● 3	● 3	● 2	● 3	● 3	● 3	● 3
Separation of building for business and for home use	2	22	21	23	23	23	23	23	7	23	23	23	20
Plans for incorporating enterprise	3	16	18	17	14	14	14	15	7	14	14	14	14
Rate of commissions received in sales	4	● 4	● 3	6	7	7	9	6	● 5	7	9	8	7
Prefecture number	47	● 2	10	13	8	9	8	9	6	8	8	9	24
Existence of a successor	2	21	22	22	21	21	21	22	7	21	21	21	18
Scale of persons engaged	3	9	12	● 5	● 4	● 4	● 4	● 5	● 4	● 4	● 4	● 4	● 4
When establishments opened	6	8	8	15	10	10	11	11	7	10	11	10	9
Scale of sales	17	● 3	● 2	● 4	● 2	● 2	● 2	● 2	● 3	● 2	● 2	● 2	● 2
Main problem in managing business	12	17	15	8	6	6	6	8	7	6	6	7	8
Other problems	4	18	9	9	17	17	17	13	7	17	17	17	19
Income other than the main business (y/n)	2	● 5	● 5	20	20	20	20	19	7	20	20	20	17
PC use (y/n)	2	6	6	21	19	19	19	21	7	19	19	19	16
Industrial Classification	33	● 1	● 1	● 1	● 1	● 1	● 1	● 1	● 1	● 1	● 1	● 1	● 1
Investment in plant and machinery	6	12	14	16	15	16	15	17	7	15	15	16	22
Sex of business proprietors	2	20	24	24	24	23	24	24	7	24	24	24	21
Form of building ownership (y/n)	2	23	19	19	18	18	18	20	7	18	18	18	15
Business developments in the future	11	7	13	10	12	12	12	12	7	12	12	12	13
Other Business developments	4	15	16	11	22	22	22	16	7	22	22	22	23
Form of land ownership (y/n)	2	14	17	18	16	15	16	18	7	16	16	15	12
Class of sales (from past survey)	2	10	7	7	● 5	● 5	● 5	7	7	● 5	● 5	● 5	● 5



(b) Final invnetory

Variables	number of levels	Regression models			Summary measures of association in contingency tables								
		Step wise	Lasso	S1	V	bcV	mi	norm. mi	lambda	tau	U	AIC	BIC
Age group of business proprietors	7	12	24	9	10	10	10	10	8	10	10	10	9
Affiliation with chain organizations	3	22	• 4	• 3	9	8	7	• 4	8	9	7	7	6
Questionnaire type (mail/online)	3	18	21	12	13	13	13	14	8	13	13	13	11
Operating days	5	11	10	• 2	• 3	• 3	• 3	• 3	• 2	• 3	• 3	• 3	• 3
Separation of building for business and for home use	2	23	18	23	23	23	23	23	8	23	23	22	20
Plans for incorporating enterprise	3	24	20	17	14	14	14	15	8	14	14	14	14
Rate of commissions received in sales	4	• 4	• 3	• 5	7	7	8	6	• 5	7	8	8	7
Prefecture number	47	• 2	9	14	8	9	9	9	6	8	9	9	24
Existence of a successor	2	21	22	22	21	21	21	22	8	21	21	21	18
Scale of persons engaged	3	9	13	6	• 4	• 4	• 4	• 5	• 4	• 4	• 4	• 4	• 4
When establishments opened	6	7	7	15	11	11	11	11	8	11	11	11	10
Scale of sales	17	• 3	• 2	• 4	• 2	• 2	• 2	• 2	• 3	• 2	• 2	• 2	• 2
Main problem in managing business	12	13	15	8	6	6	6	8	7	6	6	6	8
Other problems	4	17	11	10	18	18	18	13	8	18	18	18	19
Income other than the main business (y/n)	2	• 5	• 5	20	20	20	20	20	8	20	20	20	17
PC use (y/n)	2	8	8	21	19	19	19	21	8	19	19	19	16
Industrial Classification	33	• 1	• 1	• 1	• 1	• 1	• 1	• 1	• 1	• 1	• 1	• 1	• 1
Investment in plant and machinery	6	15	14	16	16	16	16	17	8	16	16	17	22
Sex of business proprietors	2	20	23	24	24	24	24	24	8	24	24	24	21
Form of building ownership (y/n)	2	19	19	19	17	17	17	18	8	17	17	16	15
Business developments in the future	11	6	12	11	12	12	12	12	8	12	12	12	13
Other Business developments	4	14	16	13	22	22	22	19	8	22	22	23	23
Form of land ownership (y/n)	2	16	17	18	15	15	15	16	8	15	15	15	12
Class of sales (from past survey)	2	10	6	7	• 5	• 5	• 5	7	8	• 5	• 5	• 5	• 5

The table of Area Under the Curve (AUC) of the three methods

Table 2

	Regression models		Summary measures of association in contingency tables			
	StepWise	**Lasso**	**S1**	**norm.mi**	**lambda**	**V, bcV, mi, tau, U, AIC, BIC**
Initial	0.826	0.820	0.818	0.818	0.821	0.818
Final	0.829	0.823	0.823	0.821	0.823	0.821

From Table 2, it can be seen that the accuracy of the Stepwise method is slightly better than the other methods, but all methods are almost equivalent. Despite its simplicity, S1 provides equal or better accuracy than other measures of association.

6. DISCUSSION

The results of the experiment in the previous section demonstrate that the summary statistics S1 possibly achieves accuracy that is almost the same as the other methods, even though it may not necessarily reflect accuracy.

Based on Table 1, the summary measures of association in contingency tables indicate a negligible disparity in the explanatory variables' contribution order between the Initial Inventory and Final Inventory. Furthermore, S1 exhibits a comparable contribution order to these measures of association, particularly with norm.mi, revealing identical top 5 variables as in the Initial Inventory.

All methods select two variables in common that contribute to Inventories being zero or non-zero: “Industrial Classification” and “Scale of sales” (or “Rate of commissions received in sales”). These variables can be easily understood, as “Industrial Classification” refers to industries that provide services rather than goods, and “Rate of commissions received in sales” refers to cases where all sales are from commissioned work such as Original Equipment Manufacturing.

The variables that are suggested to contribute by method S1 and measures of association are “Operating days” and “Scale of persons engaged.” These variables can also be understood based on practical experience, as businesses with fewer operating days or fewer employees may not require Inventories.

The variables that are suggested to contribute by method S1 and norm.mi and Lasso are “Affiliation with chain organizations,” which makes sense as businesses affiliated with chain organizations in the wholesale and retail industries often need Inventories.

The variables that are suggested to contribute by StepWise and Lasso are “Prefecture number” and “Income other than the main business.” However, it is difficult to provide a practical explanation for how these variables are related to whether Inventories are zero or non-zero. One possible reason for the selection of these variables may be the existence of an unpredicted small group. For example, if the sample size of learning data corresponding to a certain industry in a certain prefecture is extremely small and Inventories are zero for all of them, the accuracy in the learning process will be 100%. While “Income other than the main business” is a variable with only two levels, “Yes” or “No,” it is also possible that an extremely small size will appear in the direct product set with other variables.

Many measures of association in contingency tables propose “Class of sales (from past survey).” Although this relationship is not difficult to interpret, there is a risk that the relationship may change if there are significant economic changes between the past survey and the present.”

Based on this analysis, it can be seen that the variable selection by summary statistics S1 is not only accurate but also practical and easy to explain in a real-world context.

7. CONCLUSION

In this paper, we demonstrated that summary statistics S1 is useful for obtaining categorical variables that contribute to whether Inventories in the Unincorporated Enterprise Survey are zero or non-zero in missing value imputation.

S1, which is based on exploratory data analysis and is easy to understand and compute, also has practical interpretability. However, we were not confident about the accuracy of the S1 in predicting outcomes. In order to address this issue, we conducted an experiment using real data to compare the S1 method to other methods, which prioritize accuracy. The aim of the experiment was to assess the ability of each method to accurately select variables.

The results of the experiment showed that S1 is as accurate as these other methods. Furthermore, the variables selected by the other methods were difficult to explain based on practical experience, but the variables chosen by S1 were easily understandable in this context.

Therefore, we have shown that S1 is a suitable summary statistic for practical purposes, as it not only guarantees accuracy but also has interpretability that is relevant to practical experience.

About the Data

In this paper, the microdata of Unincorporated Enterprise Survey conducted by Statistics Bureau, Ministry of Internal Affairs and Communications (MIC), was used for statistical research under the Statistics Act of Japan. All statistical results herein are produced by the authors while MIC doesn't necessarily disseminate these results.

References

1. **Agresti, A**. (2002), *"Categorical Data Analysis"*. Second edition. Wiley, New York.
2. **Chen, S., & Haziza, D.** (2017), *"Multiply robust imputation procedures for zero-inflated distributions in surveys"*, Metron, 75(3), 333-343.
3. **D'Orazio M., Di Zio M., Scanu M**. (2006) *"Statistical Matching, Theory and Practice"*. Wiley, Chichester.
4. **Jørgensen, B** (1987), *"Exponential dispersion models"*, Journal of the Royal Statistical Society, Series B. 49 (2): 127–162.
5. **Kendall, M., and Stuart, A**. (1979), "*The Advanced Theory of Statistics, Vol. 2; Inference and Relationship"*, 4th ed. New York: Macmillan.
6. **Kleinke, K., & Reinecke, J**. (2013), *"Multiple imputation of incomplete zero-inflated count data"*, Statistica Neerlandica, 67(3), 311-336.
7. **Lee, S. M., Lukusa, T. M., & Li, C. S**. (2020), *"Estimation of a zero-inflated Poisson regression model with missing covariates via nonparametric multiple imputation methods"*, Computational Statistics, 35(2), 725-754.

8. **MacNeil Vroomen, J., Eekhout, I., Dijkgraaf, M. G., van Hout, H., de Rooij, S. E., Heymans, M. W., & Bosmans, J. E.** (2016), "*Multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best?"*, The European Journal of Health Economics, 17(8), 939-950.
9. **Mbarek, M., Rousselière, D., Salanié, J.** (2015), *"Using multiple imputation for a zero-inflated contingent valuation with a potentially biased sampling"*, In Southern Economic Association Annual Meeting.
10. **Min, Y., Agresti, A.** (2002), *"Modeling nonnegative data with clumping at zero: A survey"*, Journal of The Iranian Statistical Society, 1, 7-33.
11. **Mutoh, A., and Shirakawa, K.** (2023), *"Summary Statistics Based on Level of Measurement for Exploratory Data Analysis"*, Data Science Research, 2. (Japanese) (in press)
12. **Pahel, B. T., Preisser, J. S., Stearns, S. C., & Rozier, R. G.** (2011), *"Multiple imputation of dental caries data using a zero-inflated Poisson regression model"*, Journal of public health dentistry, 71(1), 71-78.
13. **Sakamoto, Y., Ishiguro, M. and Kitagawa, G**. (1980), *"Computer Science Monograph*, No.14, CATDAP, A CATEGORICAL DATA ANALYSIS PROGRAM PACKAGE, DATA No.2". The Institute of Statistical Mathematics.

Some Approaches to Outliers' Detection in R

Marcello D'Orazio (madorazi@istat.it)
Italian National Institute of Statistics - Istat

ABSTRACT

Outlier detection is part of data editing phase for numerical variables. This work investigates outlier detection in the R environment by comparing "traditional" methods, popular in official statistics, with techniques developed in the field of data mining and statistical learning. The comparison is done considering longitudinal data where a set of quantitative non-negative variables are observed twice (or more) on the same sample of units. The work attempts to identify some "recent" outlier detection methods, already available in the R environment, that seem suitable for application in official statistics. This study takes stock of findings of a previous work investigating outlier detection in the univariate case that showed the goodness of some "recent" approaches; in this article we go a step further and investigate the behavior of "traditional" and recent methods also in the multivariate case. The first preliminary results are quite interesting and useful as guidance towards application of the chosen methods in the production of official statistics using the R facilities.

Keywords: *binary recursive partitioning, clustering, nearest neighbor distance, panel data.*

JEL classification*: C – Mathematical and Quantitative methods; C14 Semiparametric and Nonparametric Methods: General; C33 Panel Data Models; C83 Survey Methods*

1. INTRODUCTION

National Statistical Institutes (NSIs) spend non-negligible efforts in checking the incoming data to detect actual or potential errors. This *data editing* phase (or *statistical data editing*, sometimes also referred as to *input data validation*) can make use of a variety of statistical methods depending on the type of variables (continuous, categorical, or mixed-type) and the relationships existing between them; the causes of errors and how they can affect the final estimates; the data collection mode; etc. This note concentrates on the subset of data editing methods tailored to *outlier detection*; according to UNECE (2000) "an outlier is a data value that lies in the tail of the statistical distribution of a set of data values", whereas "outliers in the distribution of uncorrected (raw) data are more likely to be incorrect". Waal *et al* (2011, pp. 7-8) give a more general definition: "a value, or a record, is called an outlier if it is not fitted well by a model that is posited for the observed data"; it is said *univariate outlier* if it is a single value of the whole record; while, on the contrary, it is called *multivariate outlier* if it consists of "an entire record, or at least a subset consisting of several values, is an outlier when the values are considered simultaneously, that is if they do not fit the posited model well when considered simultaneously".

When observing a single continuous variable (household income; firm production; harvested area in a farm etc.) an outlier is often caused by a *measurement error*, i.e. an error occurred in data collection and such that the observed value is not equal to the true value (and the true value is not expected to be in the tail of the distribution), in this case the outlier should be removed and replaced (imputed). In some cases, an outlier can also be a non-erroneous "extreme" value which, however, may have a great influence on the final estimates (*influential value*) and therefore may deserve a "special" treatment in the analysis. Waal *et al* (2011, p. 8) note that an influential value "is often an outlier, and vice versa; however, an outlier may also be a noninfluential value and an influential value may also be a nonoutlying value". For this reason, detection of outlies often takes place in *selective-editing* sub-phase tailored to identify outliers and *influential errors*.

This work investigates traditional and "recent" approaches to outlier detection currently implemented in the R environment (R Core Team, 2022) by carrying out a small empirical comparison on panel survey data, where a continuous variable is observed on the same set of units (households, firms or agriculture holdings) in different time occasions. Next Section briefly describes well-known outlier detection methods based on fitting explicit statistical models; in addition, it summarizes the Hidiroglou and Berthelot (1986) method, developed for outlier detection with panel survey data. Section 3 introduces "traditional" and "recent" nonparametric outlier detection methods, mainly proposed in the field of data mining or machine learning, which seem suitable for application in official statistics. Since many of these recent methods rely on calculation of distance between observation, Section 4 discusses the major issues related to application of distance-based approaches in the multivariate case. Section 5 compares the results provided by the chosen methods with data related to farms and firms. Finally, Section 6 summarizes the main findings and future areas of work.

2. PARAMETRIC APPROACHES TO OUTLIER DETECTION

This Section is not intended to provide a comprehensive overview of parametric methods available for detecting outliers, rather it just mentions some popular approaches implemented in the R environment.

Formally, in the univariate case we consider a single quantitative non-negative variable *Y* observed on a random sample of *n* units ($y_i \geq 0;\ i = 1,2, \dots, n$). In a parametric framework, it is common to assume a Gaussian distribution (for raw data or log-transformed raw data) and then search for outlier in the tails of the estimated distribution; robust methods are applied to estimate the location and scale parameters; the median is a popular robust estimate of the location parameter, while several alternative robust estimators of scale parameter exist, ranging from simple ones considering the *inter quartile range* (IQR), to more complex ones using the *median absolute deviation* (MAD), S_n and Q_n estimators (see Rousseeuw and Croux, 1993) and τ estimator proposed by Maronna and Zamar (2002). In R all these robust estimators are made available by the package **robustbase** (Maechler *et al*., 2022). A wrapper to most of the methods is included in the package **univOutl** (D'Orazio, 2022).

A different approach tries to identify outliers by assuming that data are generated by a mixture of two Gaussian distributions sharing the same location but with different scale. The underlying idea is that the "contaminated" data points (outliers) are caused by additive measurement errors with zero mean but variance proportional to that of non-contaminated data. In this setting, estimation of parameters of the two distributions allows to identify the outliers as those observations that have higher probability of being generated by the contaminated Gaussian distribution (see e.g. Di Zio and Guarnera, 2013). This approach is implemented in the R package **SeleMix** (Guarnera and Buglielli, 2020) that allows including error-free predictors of *Y*.

When data do not follow the Gaussian distribution, it is possible to resort to approaches based on other models; for instance, the R the package **extremeValues** (van der Loo, 2010) covers Gaussian and Log-Normal as well as Weibull and Pareto distributions (van del Loo, 2010). Alfons *et al*. (2013) suggest to fit a semi-parametric Pareto tail model to detect outliers in complex sample surveys designed to estimate indicators on social exclusion and poverty (in EU traditionally produced by the European Union statistics on income and living conditions survey); this approach is implemented in the R package **laeken** (Alfons and Templ, 2013).

In the multivariate case we consider *p* continuous variables, $\boldsymbol{y}_i' = (y_{i1}, \dots, y_{ip})$ having non-negative values ($y_{ic} \geq 0;\ i = 1,2, \dots, n; c = 1, \dots, p$) observed on a sample of *n* units. When assuming a multivariate Gaussian distribution a wide set of methods identify potential outliers as observations having the largest Mahalanobis distance from the center of the data. In this setting the squared Mahalanobis distance follows a Chi-Square distribution with *p* degrees of freedom and it is common to identify as potential outliers those units having a squared distance greater than $\chi^2_{p,1-\alpha}$. Commonly adopted methods to achieve robust estimates of the parameters of the multivariate Gaussian distribution are MCD, MVE, OGK, etc. (see e.g. Todorov and Filzmoser, 2009). The robust estimation of location and variance-covariance parameters allows to compute a robust Mahalanobis distance and, accordingly, observations whose robust distance is greater than $\chi^2_{p,1-\alpha}$ are identified as potential outliers. This way of working basically identifies α percent of observations as outliers; for this reason, Filzmoser *et al.* (2005) suggest an "adaptive" approach that compares the theoretical distribution (χ^2_p) with the empirical distribution of the squared robust distances. All these features are implemented in the R package **mvoutlier** (Filzmoser and Gschwandtner, 2021; in particular see functions `arw()` and `dd.plot()` that are based on MCD estimator for the parameters of the multivariate Gaussian distribution). It is worth noting that the package **mvoutlier** includes various methods for multivariate outlier detection (see Filzmoser and Gschwandtner, 2021).

The package **SeleMix** (Guarnera and Buglielli, 2020) permits to apply the mixture-based approach also in the multivariate case.

A special multivariate case occurs when one or more continuous non-negative variables are observed repeatedly over time on the same set of units (panel surveys); in this case it is expected a high correlation between subsequent measurements and this feature becomes crucial when the objective is the estimation of the change over time of a population parameter related to one of more of the considered variables. In the bivariate case ($p = 2$) a very popular approach is suggested by Hidiroglou and Berthelot (1986).

2.1 Hidiroglou-Berthelot method for outlier detection with longitudinal data

Hidiroglou and Berthelot (1986) suggest to detect outliers by analyzing scores obtained by transforming the ratios $r_i = y_{t_2 i}/y_{t_1 i}$ $(i = 1,2, \dots, n)$; r_i denotes the "individual change" from time t_1 to time t_2 $(t_2 > t_1)$ for unit *i*. In practice, at first ratios are transformed in the following manner:

$$s_i = \begin{cases} 1 - \frac{r_M}{r_i}, & \text{if} \quad 0 < r_i < r_M \\ \\ \frac{r_i}{r_M} - 1, & \text{if} \qquad r_i \geq r_M \end{cases} \qquad [1]$$

where r_M is the median of the ratios, then to account for the magnitude of data and give more "importance" to units involving high values of *Y*, the following *scores* are derived:

$$E_i = s_i\left[\max\left(y_{t_1 i}, y_{t_2 i}\right)\right]^U \qquad [2]$$

where U can range from 0 to 1 ($0 \leq U \leq 1$) and controls the role of magnitude in determining importance associated to transformed ratios (a common choice consist in setting $U = 0.5$). Finally, assuming that *E*-scores follow a Gaussian distribution the potential outliers are the units whose scores fall outside the interval:

$$\left[E_M - C \times f_{Q1}, E_M + C \times f_{Q3}\right] \qquad [3]$$

Being E_M the median of the *E* scores, while f_{Q1} and f_{Q3} are functions of the quartiles of the *E*-scores with a correction factor that avoids drawbacks of distributions highly concentrated around the median; in addition, the bounds allow for a slight skewness in the distribution of the *E* scores. Recently, Hidiroglou and Emond (2018) suggest to replace the quartiles with the deciles ($P_{10,E}$ and $P_{90,E}$ instead of respectively $Q_{1,E}$ and $Q_{3,E}$) in cases where a large proportion of units (>1/4) share the same value of the ratio, since in this case the "standard" method would detect too many observations as potential outliers. The parameter *C* in expression [3] determines how far from the median the bounds should be; commonly suggested values are $C = 4$ or $C = 7$ but larger values can be considered, depending on the tails of the distribution of the *E* scores. Practically, the choice of the constants *U* and *C* is not straightforward and it is preferable to graphically investigate how the *E* scores are distributed.

It is worth noting that data editing literature suggests alternative methods to check whether the individual change (r_i) is too large or too low (see e.g. the theme "Editing for Longitudinal Data" in Eurostat, 2014).

In the R environment the Hidiroglou Berthelot (HB) procedure is implemented in the package **univOutl** (D'Orazio, 2022) that includes also graphical facilities for inspecting the scores, in line with Hidiroglou and Emond (2018) suggestion.

3. NONPARAMETRIC APPROACHES TO OUTLIER DETECTION

Nonparametric outlier detection methods avoid explicit assumption on the underlying distribution; in this group fall many outlier detection methods proposed in the domain of data mining and statistical learning. This Section summarizes the features of a subset of popular methods that have the advantage of being easily applicable in the production of official statistics by using the facilities of the R environment. D'Orazio (2023) carried out an empirical investigation of some of these methods but only in the univariate setting.

3.1 Boxplot

Detecting outliers by plotting a *boxplot* (*box-and-whisker* plot) is very popular in the univariate case; the units outside the *whiskers* are considered outliers. To account for skewness Hubert and Vandervieren (2008) suggested an "adjusted" boxplot taking into account a measure of the skewness, the *medcouple*, that works with moderate skewness. The R package **univOutl** (D'Orazio, 2022) includes a series of functions for outlier detection based on the standard or adjusted boxplot.

3.2 Outlier detection based on nonparametric density estimation

Literature on data mining and statistical learning provides many suggestions for outlier detection, often indicated as *unsupervised* outlier detection methods; the largest group of methods is the one "inspired" to nonparametric estimation of density. As noted by Zimek and Filzmoser (2018), several approaches under this umbrella simply consists in calculating distances between observations. Some of them are variants of Ramaswamy *et al.* (2000) proposal that suggest identifying the potential outliers by calculating the *k* nearest neighbor (*k*-NN) distance; in practice, if $d_{i,(k)}$ is the distance of the *i*th from its *k* nearest neighbor, then units showing largest values of $d_{i,(k)}$ are potential outliers. Angiulli and Pizzuti (2002) suggest to analyze the "weight" obtained by summing up all the distances from the corresponding *k* nearest neighbor observations:

$$\omega_{i,(k)} = \sum_{j=1}^{k} d_{i,(k)} \qquad [4]$$

Similarly, Hautamäki *et al* (2004) suggest using the average of distances ($\bar{\omega}_{i,(k)} = \omega_{i,(k)}/k$). Campos et al. (2016) note that the sum (or average) of distances reduces the variability of the scores and return scores less sensitive to the value of *k*.

In general, there's no rule-of-thumb for deciding *k*; D'Orazio (2023) investigated these approaches in the univariate case highlighting the difficulties in analyzing the final scores (distance or "weight") whereas the magnitude, indicating the chance of being an outlier, increases by increasing the value of *k*. In particular, it seems difficult to identify a threshold such that scores greater than it can be considered potential outliers. Hautamäki *et al.* (2004) suggest to derive the threshold as a fraction of the observed maximum first order difference calculated after sorting the scores in increasing order. Practically a graphical inspection can be more effective: after sorting the scores increasingly, good candidate thresholds are the values corresponding to “jumps” in the plot (abnormal increase in the score). In this setting, a good candidate threshold is the point of maximum curvature in the graph, as shown later (see Section 3.3).

The before mentioned approaches are labeled as “global” in data mining literature and for this reason not flexible enough to catch situations where sub-groups of observations may have different “local” densities; this latter case is better handled by approaches developed starting from the *local outlier factor* (LOF) (Breuning *et al*, 2000). The idea is very simple and consists in comparing the *local reachability density* of each observation with the average local reachability calculated on the q nearest neighbors (the *local reachability density* is obtained as the inverse of the average *reachability distance* between each unit and its q closest neighbors). An outlier is expected to have a local reachability distance smaller than the average on neighboring units and consequently a LOF score greater than one. The larger is the LOF score the higher is the chance of finding an outlier; unfortunately, also in this case there’s not a rule for setting a threshold.

The R package **DDoutlier** (Madsen, 2018) provides functions to apply many density and distance-based outlier detections approaches. k-NN distance is however calculated in many other R packages, the main drawback is that most of them permit to use just a limited set of popular distance functions (typically Euclidean and Manhattan distances).

3.3 Outlier detection with clustering-based algorithms

Density-based spatial clustering with noise (DBSCAN; Ester *et al*, 1996) separates the observations that do no not belong to any cluster, because not “reachable” by any other observations, as “noisy” observations, i.e. outliers. The “reachability” depends on a distance threshold ε; in practice two units i and j are *directly reachable* if their distance is less or equal than ε ($d_{i,j} \leq \varepsilon$), while they are only *reachable* if there is a path of three of more observations to go from i to j, where each couple of units in the path *is directly reachable*. In addition, the DBSCAN algorithm requires setting also the number g of “core” observations, i.e. observations that have at least $g - 1$ distinct units at a distance smaller or equal to ε. The literature suggests to set $g = 2p'$ (cf. Schubert *et al*, 2017) where p' ($p' \leq p$) is the number of variables used to calculate the distance, i.e. those for which we search for potential multivariate outliers. Empirical results in Schubert *et al* (2017) seem to support the conclusion that often g has a limited impact on the results; on the contrary ε is crucial. A common suggestion is to plot the ($g-1$)-NN distances in increasing order and set ε equal to the distance where the plot shows a “valley,” “knee,” or “elbow” (Schubert *et al*, 2017). Unfortunately, with many observations the graphical representation may show points rather close each other and difficult to investigate. A simpler criterion could be that of approximating ε with the point of maximum curvature, i.e. the point where the empirical curve has the maximum distance from the straight line connecting the smaller and the larger ($g-1$)-NN distances. This is basically the approach for identifying a “knee” in a series of discrete points showing an increasing concave-down curvature (plot of sorted ($g-1$)-NN distances usually shows an increasing concave-up curve); in particular, in this paper we opt for the *Kneedle algorithm* (Satopaa *et al.*, 2010) that is simple and can be easily implemented in R.

Although there are several proposals to improve the seminal DBSCAN idea, it still remains an approach that provides quite good results. In R the DBSCAN is made available by the package **dbscan** (Hahsler *et al.*, 2019; Hahsler and Piekenbrock 2022).

3.4 Outlier detection with recursive partitioning trees

The underlying idea is that outliers have a higher chance of being separated by the other ones in one branch of the partitioning tree with relatively few splits. In the univariate case an arbitrary threshold y_o is selected at random within the range of *Y* ($y_c^{(min)}, y_c^{(max)}$) and all the observations are divided into two groups according to whether they show higher or lower values than y_o. This randomized splitting process is applied recursively (i.e., divide the units into two groups then repeat the process in each group, and so on) until no further split is possible (or until meeting some other criteria). The final outcome is an *isolation tree* where the more observations show similar *Y* values, the longer (more splits) it will take to separate them in small groups (or alone) compared to less occurring *Y* values; for this reason, the *isolation depth* (number of splits needed to isolate a unit) can be considered as a tool for detecting outliers.

Since the isolation depth estimated in a single isolation tree would be characterized by a high variability, the common adopted device consists in building an ensemble of isolation trees – the *isolation forest* – and then derive the final score by averaging over the fitted trees (Liu *et al.*, 2008 and 2012). Like random forests, each single isolation tree can be fit on a bootstrap sample of *m* ($m < n$) observations randomly selected. In the Liu *et al* proposal (2008 and 2012) the partitioning stops when a node has only one observation or all units in a node have the same values (in some cases it is introduced a maximum value for the tree height). The isolation forest returns a score u_i ranging from 0 to 1 ($0 < u_i \leq 1$); scores close to 1 indicate observations with a very short average path length that tend to be isolated earlier than the other ones and therefore denote outlying observations. As a consequence, setting a threshold u_0 ($0 < u_0 < 1$), returns as outliers all the units having a score $u_i > u_0$; it is suggested to consider $u_0 = 0.5$ but D'Orazio (2023) shows that a graphical inspection of the ordered scores can be beneficial in deciding u_0.

The isolation forest is very efficient and requires setting just two tuning parameters, the subsample size *m* and the number of trees to fit. Liu *et al* (2008 and 2012) claim that even a small subsample size ($m = 256$) can work with very large data-sets, even though some implementations of the algorithm avoid subsampling if *n* is not very large. Concerning the number of trees to fit, it is suggested to start with at least 100, but this figure should be increased when the achieved scores are on average quite below 0.5, as this may point out a problem of unreliable estimation of the average path length.

In the multidimensional framework, the various trees are derived by picking at random at each iteration one of the available variables. Unfortunately, the standard way of working (branching and bounding) would consist in an ensemble of results related to the application of isolation forest independently variable by variable. To compensate this drawback, it is preferable to consider an *extended isolation forest* (Hariri et al 2018) that in the branching step considers jointly two or more variables; for instance, with $o = 2$ variables the algorithm partitions repeatedly the units according to a regression line whose intercept and slope are randomly generated each time (when $o > 2$ the branching considers randomly generated hyperplanes). Liu *et al* (2010) suggest to set $o = 2$ as it seems to work well even with many starting variables. Hariri *et al* (2018) suggestion is to set $o = 2$ or $o = 3$.

In R the standard isolation forest is implemented in the package **solitude** (Srikanth, 2021) while the package **isotree** (Cortes, 2022) permits to fit also the extended isolation forest.

4. ISSUES IN COMPUTING DISTANCES IN THE MULTIVARIATE FRAMEWORK

Many of the outlier detection methods presented in the previous sections rely on calculating distances between observations. This is often perceived as a very simple way of working, although in the multidimensional setting it involves taking additional decisions on: (i) the distance function, and (ii) whether the variables need to be scaled in advance. These choices are often understated and the practitioner accepts without criticism the default choices made by the developers of the used software packages, without worrying whether the default settings can work in their case studies. A common choice of many distance-based outlier detection is that of considering the Euclidean distance or the Manhattan distance, that are particular cases of the *L*-norm :

$$d_{i,j} = \sqrt[T]{\sum_{c=1}^{p} \left| y_{ic} - y_{jc} \right|^{T}} \qquad [5]$$

with respectively $T = 2$ and $T = 1$. This expression shows that a variable having larger values than another tends to dominate the overall distance and the dominance increases with increasing values of *T*. In multivariate outlier detection this would mean that an outlier in a distance-dominant variable influences greatly the detection of multivariate outliers; for this reason, before calculating the distances it may be necessary to scale the variables (preliminary scaling is not needed when applying the Mahalanobis distance).

Common choices for scaling the variables are the range or the standard deviation but, as in the case of outlier detection with Gaussian distributed data, the scaling should involve a robust estimate of the standard deviation or replacing the range with a function of inter-percentile range (e.g. difference between the whiskers of the boxplot, etc.).

Another common mistake in multivariate outlier detection is to jointly consider all or almost all the *p* available numerical variables despite how large is *p*. With a quite high number of continuous variables there is the risk of incurring in effects of the *curse of dimensionality*. In particular, in unsupervised outlier detection with a high-dimensional dataset, Zimek *et al* (2012) stress that the major problem of distance-based methods is the loss in discrimination ability, i.e. a reduction of the ability of the distance in discriminating between near and far neighbors; this is known as *concentration effect.* This problem affects the Mahalanobis distance too, where the efforts in estimating the variance-covariance matrix should also be taken into account. For these reasons, with many variables the detection of multivariate outliers should be done focusing just on the subset of relevant continuous variables p' $(p' < p)$.

5. APPLICATION OF THE CHOSEN METHODS TO SOME DATA FROM PANEL SURVEYS

This section investigates the performances of the methods presented in previous Sections when applied to a couple of datasets related to panel or pseudo-panel surveys that are described in Table 1.

Datasets used in the experiments

Table 1

Dataset/survey	Number of units	Type of units	Description
RDPerfComp	509	firms	R&D performing US manufacturing; yearly observations from 1982 to 1989 of the following variables: production, labor and capital[1]
RiceFarms	171	farms	Indonesian rice farm dataset, 171 farms producing rice observed 6 times. Many variables are available; the ones used in this study are: the total area cultivated with rice (in hectares); the total number of worked hours and the gross output of rice (in kg)[2].

In practice, in both the datasets the HB procedure is applied to each of the chosen variables in order to derive the corresponding scores ($E_{ci},\ i = 1,2,\dots,n;\ c = 1,\dots,p'$) provided by expression [2] with $U = 0.5$. These *E*-scores become the input of the following outlier detection techniques:

Md) robust Mahalanobis distance with adaptive distance threshold derived by comparing the theoretical distribution ($\chi^2_{p',0.975}$) with the empirical distribution of the squared robust distances (function `dd.plot()` in **mvoutlier** package);

SM) fit of a mixture of Gaussian distributions; outliers identified as "contaminated" units having a posterior probability greater than 0.5 (function `ml.est()` in **SeleMix** package with argument `tau=0.5`);

kNNw) *k*-NN weight (sum of *k*-NN distances; function `kNNdist()` in package **dbscan**) considering Euclidean distance and variables scaled in advance with a robust estimate of the standard deviation (based on inter-quartile range); approximate distance threshold set using the Kneedle algorithm;

LOF) Local Outlier Factor (`LOF()` function from **DDoutlier** package) with respectively $k = 5$ and $k = 10$; Euclidean distance is calculated on variables scaled in advance with a robust estimate of the standard deviation based on inter-quartile range; approximate distance threshold set using the Kneedle algorithm;

[1] https://www.nuffield.ox.ac.uk/users/bond/index.html. See also the R package pder https://CRAN.R-project.org/package=pder

[2] R package plm https://cran.r-project.org/package=plm

DBS) DBSCAN clustering algorithm with $g = 2p'$ and ε (distance threshold) set by using the Kneedle algorithm (function `dbscan()` in the package **dbscan**); Euclidean distance is calculated on the chosen variables scaled in advance with a robust estimate of the standard deviation based on inter-quartile range;

EIF) Extended isolation forest (with $o = 2$, no sub-sampling and 5000 trees in each forest; function `isolation.forest()` in the package **isotree**).

All the variables observed on the firms in the RDPerfComp dataset (production, labour and capital; $p' = p = 3$) are simultaneously considered; for each variable the *E*-scores are derived by comparing values observed in year 1986 vs. those in 1985.

Scores and outliers given by model-based, DBSCAN and Extend Isolation Forest outlier detection approaches with RDPerfComp dataset

Figure 1

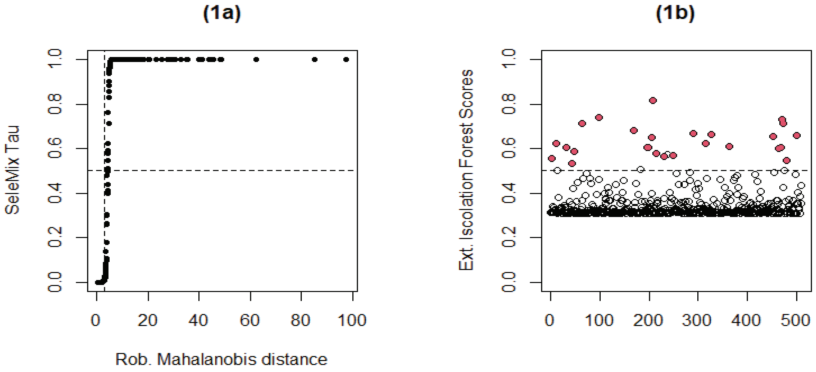


Figure 1a reports the scatterplot of the posterior probabilities estimated ("Tau") by SM vs. the robust Mahalanobis distance calculated by Md; the dashed vertical and horizontal lines indicate the corresponding thresholds and show that Md would return a higher number of outliers than SM.

Figure 1b shows the scores of EIF whereas the horizontal dashed line corresponds to the "standard" threshold $u_0 = 0.5$; the red-color filled dots indicate "noisy" observations (outliers) identified by DBS. Both the approaches return almost the same results, while the number of identified outliers is smaller if compared to that of SM or Md.

Table 2 compares result of approaches providing a direct identification of potential outliers (Md, SM, DBS) with the discretized scores of EIF. As already shown in Fig. 1b, DBS and EIF (when the rule-of-thumb $u_0 > 0.5$ is considered) point to almost the same relatively few outliers (27). The same units would be identified by Md and SM, that however, if jointly considered would return a nonnegligible number of additional potential outliers (97).

Outliers given by model-based, DBSCAN and Extend Isolation Forest outlier detection approaches with RDPerfComp dataset

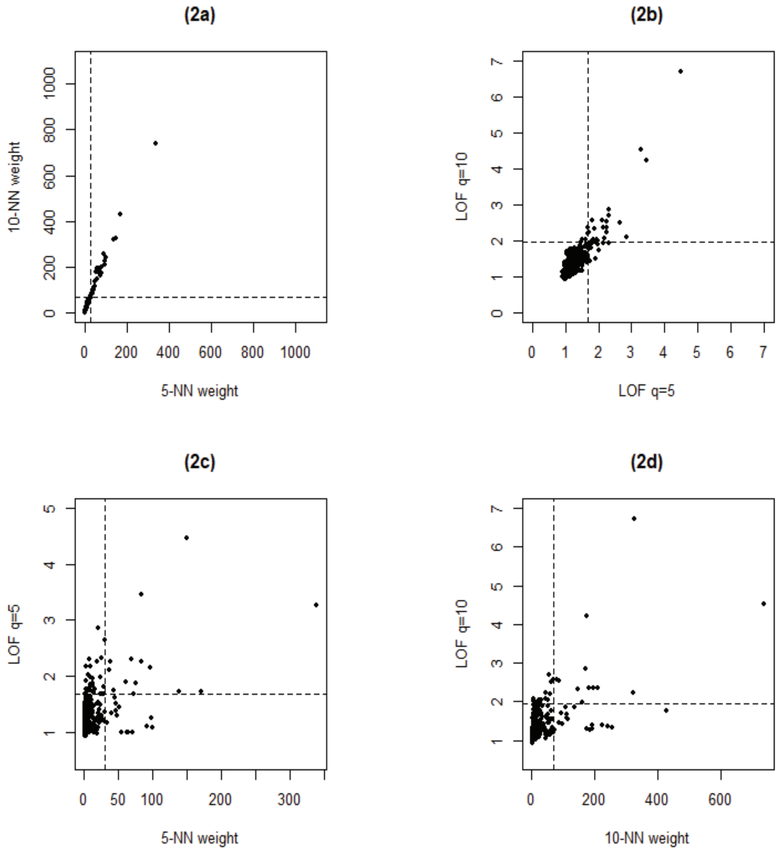
Table 2

			Scores EIF				
Md	SM	DBS	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 1.0]
Not-Outlier	Not-outlier	Not-outlier	351	0	0	0	h0
Outlier	Not-Outlier	Not-outlier	31	0	0	0	0
		Outlier	0	0	0	0	0
	Outlier	Not-outlier	58	39	3	0	0
		Outlier	0	0	8	14	5

Figure 2 summarizes results of kNNw and LOF. Figure 2a compares the scores obtained with $k = 5$ with those related to $k = 10$; the vertical and horizontal dashed lines denote the corresponding thresholds identified by applying the Kneedle algorithm; setting $k = 10$ gives a slightly higher number of potential outliers (5) than with $k = 5$; as in the expectations, the choice of k does not affect the results markedly. Similarly, Figure 2b compares the scores given by LOF with $q = 5$ and $q = 10$; the dashed lines indicate, as usual, the thresholds estimated with the Kneedle algorithm; the threshold identified with $q = 5$ returns a higher number of possible outliers compared to $q = 10$. Finally, Figures 2c and 2d compares scores of kNNw with those of LOF with respectively $k = q = 5$ and $k = q = 10$. In both the cases there is a high fraction of observations that are judged differently by the compared techniques.

Scores and outliers given by distance-based outlier detection approaches with RDPerfComp dataset

Figure 2



Figures 3 and 4 summarize the results obtained by applying outlier detection techniques to the panel of farms producing rice ("RiceFarms" dataset), when considering only three of the available continuous variables ($p' = 3 < 15 = p$): the total area cultivated with rice (in hectares); the total number of worked hours and the gross output of rice (in kg). In calculating the HB E-scores the 3rd and 4th observation occasions are considered.

Figure 3a shows the posterior probabilities estimated by SM vs. the robust Mahalanobis distance calculated by Md; as observed in the previous plots, the dashed lines indicate the corresponding thresholds. Also in this case study, Md returns a higher number of potential outliers if compared to SM.

Figure 3b shows that EIF scores are concentrated around the value 0.35 and there are relatively few potential outlying observations above the horizontal dashed line (corresponding to the $u_0 = 0.5$ threshold). In this case DBS with a distance threshold decided according to the Kneedle algorithm identifies just 5 “noisy” observations (outliers) (see also Table 3).

Scores and outliers given by model-based, DBSCAN and Extend Isolation Forest outlier detection approaches with RiceFarms dataset

Figure 3

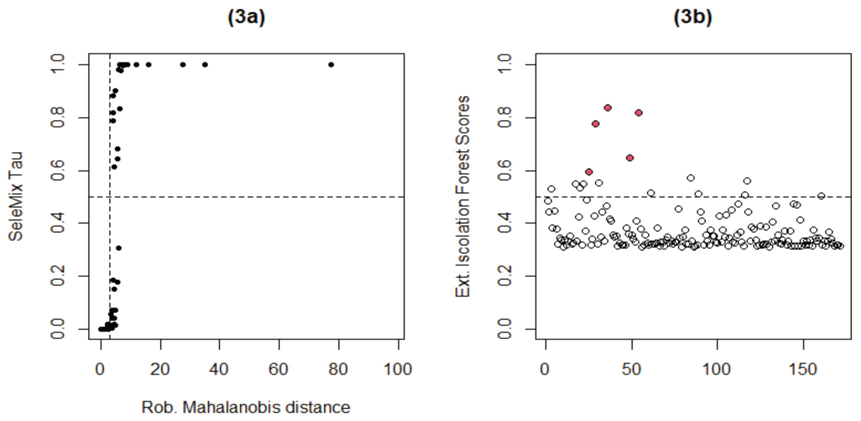


Table 3 counts the number of potential outliers obtained when crossing outcomes of Md, SM, DBS and the categorized scores of EIF. In this case study DBS identifies only 5 noisy observations out of 16 potential outliers given by EIF with the rule-of-thumb $u_0 > 0.5$. As in the previous example Md identifies a higher number of potential outliers if compared to the other procedures.

Outliers given by model-based, DBSCAN and Extend Isolation Forest outlier detection approaches with RiceFarm dataset

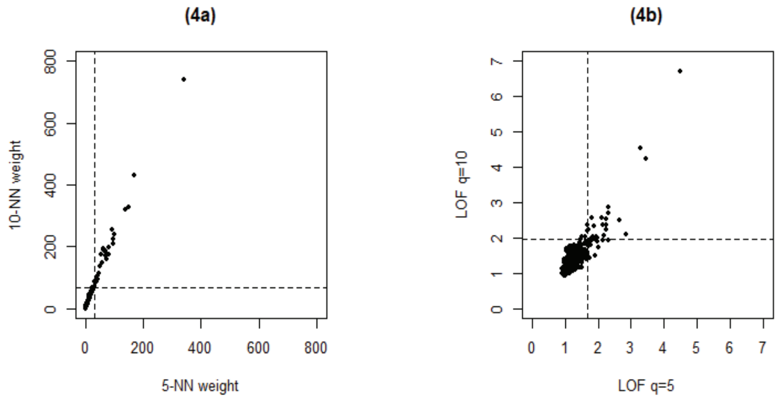
Table 3

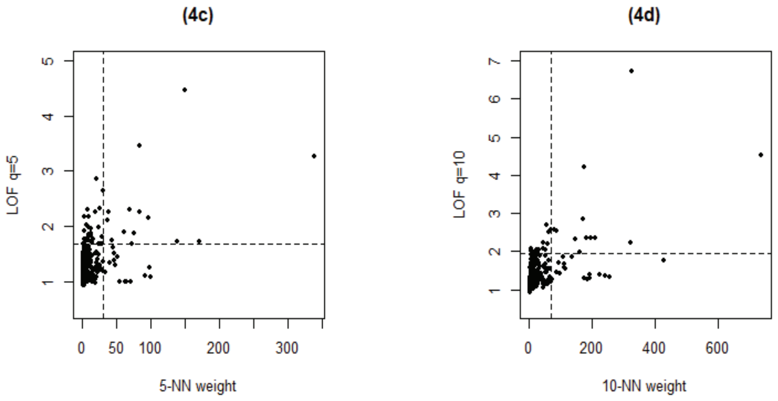
Md	SM	DBS	Scores EIF (0.3, 0.4]	 (0.4, 0.5]	 (0.5, 0.6]	 (0.6, 0.7]	 (0.7, 1.0]
Not-Outlier	Not-outlier	Not-outlier	127	1	0	0	0
Outlier	Not-Outlier	Not-outlier	4	14	2	0	0
		Outlier	0	0	0	0	0
	Outlier	Not-outlier	0	9	9	0	0
		Outlier	0	0	1	1	3

Figure 4 summarizes outcomes of kNNw and LOF. Figure 4a compares the scores of kNNw obtained with the chosen values of k (5 and 10); the vertical and horizontal dashed lines denote the corresponding thresholds identified by applying the Kneedle algorithm. Also in this case, increasing the value of k gives a slightly higher number of potential outliers. Figure 4b compares the scores given by LOF with $q = 5$ and $q = 10$, again the lower value of q returns a higher number of possible outliers compared to $q = 10$. Finally, Figures 2c and 2d compares scores of kNNw with those of LOF with respectively $k = q = 5$ and $k = q = 10$. Both the scatterplots show that a high fraction of observations is judged differently by the compared techniques, as in the case of firms' dataset.

Scores and outliers given by distance-based outlier detection approaches with RiceFarms dataset

Figure 3





6. CONCLUSIONS

This short note compares a series of approaches to detect multivariate outliers in particular when dealing with panel survey data and the objective consists also in measuring change over time. In particular, we search for multivariate outliers by analyzing the *E*-scores derived from the initial ratios of values observed in subsequent time occasions ($r_i = y_{t_2 i}/y_{t_1 i}$), as proposed by Hidiroglou-Berthelot; this way of working has the advantage of reducing the dimensionality of the data and tailor the outlier detection towards ratios involving large quantities that may have a higher influence on the final survey estimates.

The model-based outlier detection approaches considering a multivariate Gaussian distribution – robust Mahalanobis distance from the bulk of the data (Md) and fitting of mixture of Gaussian distributions (SM) – permit a direct identification of the potential outliers (units above thresholds) but in our small empirical study they tend to identify a larger number of potential outliers than the other approaches; in any case in both the approaches it is possible to use the resulting "score" (respectively robust Mahalanobis distance and posterior estimated probabilities of being outliers) to sort observations in decreasing order and start inspecting them with the support of subject matter experts. Both the approaches do not need to scale the variables in advance; they are quite easy to apply although fitting a mixture of distributions require setting a number of parameters related to the estimation process.

The nonparametric approaches based on calculating the distances are quite straightforward to apply; our study uses the Euclidean distance function (default in all the considered R functions) calculated on the *E*-scores, scaled in advance by a robust estimate of the standard deviation to account for different variability. In all the cases we decided to estimate approximately the threshold needed to identify potential outliers by applying the Kneedle algorithm, as sorted distance-based scores usually show an increasing concave-up curve. The algorithm is quite simple and easy to be coded in R.

In both the considered datasets, increasing the value of *k* in the *k*-NN weight, coupled with the chosen strategy for setting the threshold, tends to provide a higher number of potential outliers. Outcomes of LOF seem in contrast with those of *k*-NN weight; this seems due to the fact that the chosen datasets do not include sub-groups of observations with different local densities.

The DBSCAN clustering method uses also the Euclidean distances calculated on scaled *E*-scores. The key role is played by ε – the distance threshold – that in our empirical study is decided by applying the Kneedle algorithm; this latter decision led to identify relatively few units with a high potential of being outliers in both the considered datasets. The results of BDSCAN are quite aligned with those of the Extended Isolation Forest (that in our application partitions units according to randomly generated regression lines). The EIF approach has the advantage of avoiding to transform the variables in advance and has relatively few tuning parameters to set that, however, can be decided without resorting to additional algorithms/methods; in addition, the final scores range between 0 and 1 simplifying their analysis. It is worth noting that the rule-of-thumb consisting in identifying as potential outliers those units whose score is grater that 0.5 should be applied carefully; this limited empirical study seems to suggest that a slight inferior value, between 0.4 and 0.5 may work better; this finding need to be confirmed by additional investigation.

In summary, the limited empirical comparison carried out in this work, jointly to the results obtained by D'Orazio (2023) in the univariate case, show that the Isolation Forest and its extended version in the multivariate case (EIF) represent a promising approach for detecting potential multivariate outliers in official statistics with the great advantage that the results do not seem markedly affected the starting tuning parameters even if the rule-of-thumb of considering as potential outliers observation with a score greater than 0.5 should be applied carefully.

More in general, this study further confirms, if still needed, that the R environment is a fundamental software package for official statistics as it offers implementations of both "traditional" and recent statistical methods, as shown in this study limited to outlier detection procedures.

References

1. **Alfons A, Templ M** (2013). *"Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken."* Journal of Statistical Software, 54, pp. 1–25.
2. **Andreas Alfons, A., Templ, M., Filzmoser, P.** (2012) *"Robust estimation of economic indicators from survey samples based on Pareto tail modelling*", Journal of the Royal Statistical Society. Series C (Applied Statistics), 62, pp. 271-286
3. **Angiulli, F. Pizzuti, C.**, (2002) "*Fast Outlier Detection in High Dimensional Spaces"*. Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, pp. 15–26.
4. **Breunig, M.M., Kriegel, H.-P., Ng, R. T., Sander, J.** (2000). *"LOF: Identifying Density-Based Local Outliers"*. Proceedings of the International Conference On Management of Data. Dallas, TX. pp. 93-104
5. **Campos, G. O. Zimek, A., Sander, J., Campello, R.J.G.B., Micenková, B., Schubert, E., Assent, I., Houle, M. E.** (2016) *"On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study",* Data Mining and Knowledge Discovery, 30, pp. 891–927.
6. **Cortes, D.** (2022) *isotree: Isolation-Based Outlier Detection*. R package version 0.5.15, https://CRAN.R-project.org/package=isotree
7. **D'Orazio, M.** (2022). *univOutl: Detection of Univariate Outliers*. R package version 0.3. https://CRAN.R-project.org/package=univOutl
8. **D'Orazio, M.** (2023) *"An empirical comparison of some outlier detection methods with longitudinal data"*, To be published in Istat's Working Paper series.
9. **Di Zio, M., Guarnera, U.** (2013) "A Contamination Model for Selective Editing", Journal of Official Statistics, 29, pp. 539-555.
10. **Ester, M., Kriegel, H.-P., Sander, J., Xu, X.** (1996) *"A density-based algorithm for discovering clusters in large spatial databases with noise"*. Proceedings of the 2nd Int. Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, pp. 226–231
11. **Eurostat**, 2014. *Memobust Handbook on Methodology of Modern Business Statistics.* Luxembourg https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en
12. **Filzmoser, P., Gschwandtner, M.** (2021) *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. R package version 2.1.1, https://CRAN.R-project.org/package=mvoutlier
13. **Filzmoser, P., Garrett, R.G., Reimann, C.** (2005) *"Multivariate outlier detection in exploration geochemistry"*, Computers & Geosciences, 31, pp. 579–587.
14. **Guarnera, U., Buglielli, T**. (2020) *SeleMix: Selective Editing via Mixture Models*. R package version 1.0.2, https://CRAN.R-project.org/package=SeleMix.
15. **Hautamäki, V., Kärkkäinen, I., Fränti, P**. (2004) "*Outlier Detection Using k-Nearest Neighbour Graph*". International Conference on Pattern Recognition, pp. 430-433
16. **Hahsler, M., Piekenbrock, M.** (2022) d*bscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. R package version 1.1-10, https://CRAN.R-project.org/package=dbscan
17. **Hahsler, M., Piekenbrock, M., Doran, D.** (2019). *"dbscan: Fast Density-Based Clustering with R"*. Journal of Statistical Software, 91, pp. 1-30
18. **Hariri, S., Kind, M. C., Brunner, R. J.** (2018) "*Extended Isolation Forest"* arXiv preprint arXiv:1811.02141.
19. **Hidiroglou, M.A., Berthelot, J.-M**. (1986) "*Statistical editing and Imputation for Periodic Business Surveys"*, Survey Methodology, 12, pp. 73-83.
20. **Hidiroglou, M.A. Emond, N.** (2018) *"Modifying the Hidiroglou-Berthelot (HB) method"*. Unpublished note, Business Survey Methods Division, Statistics Canada, May 18, 2018.

21. **Hubert, M., Van der Veeken, S.** (2008) *"Outlier Detection for Skewed Data"*. Journal of Chemometrics, 22, pp. 235-246.
22. **Hubert, M., Vandervieren, E**. (2008) "*An Adjusted Boxplot for Skewed Distributions*" Computational Statistics & Data Analysis, 52, pp. 5186-5201
23. **Liu, F.T., Ting, K.M., Zhou, Z**. (2008) *"Isolation forest"*. In: Eighth IEEE International Conference on Data Mining, pp. 413–422
24. **Liu, F. T., Ting, K. M., Zhou, Z.** (2010) *"On detecting clustered anomalies using SCiForest"* Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2010.
25. **Liu, F. T., Ting, K. M., Zhou, Z.** (2012) *"Isolation-based anomaly detection"*. ACM Transactions on Knowledge Discovery from Data (TKDD), 6, pp. 1–39
26. **Madsen, J.H.** (2018) *DDoutlier: Distance & Density-Based Outlier Detection*. R package version 0.1.0. https://CRAN.R-project.org/package=DDoutlier
27. **Maechler. M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., di Palma, M.A**. (2022). *robustbase: Basic Robust Statistics* R package version 0.95-0. http://CRAN.R-project.org/package=robustbase
28. **Maronna, R.A., Zamar, R.H**. (2002) *"Robust estimates of location and dispersion of high-dimensional datasets"*, Technometrics, 44, pp. 307–317.
29. **R Core Team** (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
30. **Ramaswamy S., Rastogi, R., Shim, K.** (2000) *"Efficient Algorithms for Mining Outliers from Large Data Sets"*. Proceedings of the International Conference on Management of Data (SIGMOD '00), pp. 427-438.
31. **Rousseeuw, P.J., Croux, C**. (1993) *"Alternatives to the Median Absolute Deviation"*, Journal of the American Statistical Association, 88, pp. 1273–1283.
32. **Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B**. (2010) *"Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior"*, Proceedings of the 30th International Conference on Distributed Computing Systems SIMPLEX Workshop (ICDCS 2010), Genoa, Italy http://www.icsi.berkeley.edu/pubs/networking/findingakneedle10.pdf
33. **Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.** (2017) *"DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN"*. ACM Trans. Database Syst., 42, pp. 19:1-19:21.
34. **Srikanth, K.** (2021). *solitude: An Implementation of Isolation Forest*. R package version 1.1.3, https://CRAN.R-project.org/package=solitude
35. **Todorov, V., Filzmoser, P.** (2009) *"An Object-Oriented Framework for Robust Multivariate Analysis"*. Journal of Statistical Software, 32, pp. 1–47.
36. **UNECE** (2000) *Glossary of Terms on Statistical Data Editing*. Geneva https://ec.europa.eu/eurostat/ramon/statmanuals/files/UN_editing_glossary_2000.pdf
37. **Waal, T-de, Pannekoek, J., Scholtus, S**. (2011) *Handbook of statistical data editing and imputation*. John Wiley & Sons, Inc., Hoboken.
38. **van der Loo, M.P.J**., (2010) *"Distribution based outlier detection for univariate data"*, Discussion paper 10003, Statistics Netherlands, The Hague.
39. **Zimek, A. Filzmoser, P.** (2018) *"There and back again: Outlier detection between statistical reasoning and data mining algorithms"*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8, e1280.
40. **Zimek, A., Schubert, E., Kriegel, H.-P.** (2012) *"A survey on unsupervised outlier detection in high-dimensional numerical data"*. Statistical Analysis and Data Mining, 5, pp. 363–387.

Improvement of Model Construction based on Reliability Scores of Objects for Autocoding

Yukako Toko (ytoko@nstac.go.jp)
National Statistics Center, Tokyo 162-8668

Mika Sato-Ilic (mika@risk.tsukuba.ac.jp)
University of Tsukuba, Japan,

Takayuki Sasajima (tsasajima@nstac.go.jp)
National Statistics Center, Tokyo, Japan

ABSTRACT

This paper proposes a new method for autocoding, including the filtering task for constructing training data in a machine-learning language model. Misspecification of the training data causes biased outputs and exhibits undesirable behavior. Therefore, improving the training data for Natural Language Processing (NLP) tasks in supervised learning is essential for obtaining more accurate results and avoiding harmful outputs. This paper improves the training data in our proposed supervised machine-learning method for autocoding based on reliability scores over text descriptions. In the improvement task, which is a filtering task, we exploit data classified with high-reliability scores based on the idea that data classified with high-reliability scores are clearer data; adding the information of those data to the training dataset is performed to obtain a better classification accuracy. The numerical examples for the coding task of the National Survey of Family Income, Consumption and Wealth and the Family Income and Expenditure Survey show a better performance of the proposed method.

Keywords: *Coding, Reliability scores, Fuzzy logic, Text classification*
JEL Classification*: C38*

1. INTRODUCTION

In the governmental survey, assigning corresponding labels to text descriptions (coding) is an often performed activity for efficient data processing. Though coding was 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 a classification method which is the reliability score-based classification model (Toko et al., 2018, Toko et al., 2019, Toko and Sato-Ilic, 2020). In addition, as the state-of-the-art in machine learning, we have developed autocoding techniques using clustering, including the concept of Computational Intelligence (CI) such as fuzzy c-means clustering (Bezdek, 1981) to deal with more complex and unrobust features of data space (Toko and Sato-Ilic, 2021, Toko and Sato-Ilic, 2022). In these models, the main idea is based on capturing core words consisting of the same cluster in NLP, such as Anchors (Ribeiro et al., 2018, Guidotti et al., 2018) in explainable machine learning. In these systems, the reliability scores are defined as a probability measure. However, in our proposed method, the reliability scores have been defined considering both the probability measure and fuzzy measure. In addition, the target data of our proposed method is not only text descriptions provided by survey respondents to open-ended questions but also includes text descriptions obtained from images of shopping receipts. This method has been practically implemented for the coding task of *the Family Income and Expenditure Survey* in Japanese official statistics from January 2022.

However, generally, the machine learning language model requires fine-tuning tasks for the training data, and the literature indicates this process is an essential task for not causing the misspecification of the training data and avoiding generating harmful and biased outputs from the system. To overcome this problem, effective utilization of the large language model (LLM) such as GPT-3 (Solaiman and Dennison, 2021) and several filtering techniques for improving the training dataset by using similar datasets (Brown et al., 2020, Kenton et al., 2021) were proposed. However, in these methods, we need a large amount of training data or different kinds of datasets, which may have bias.

Therefore, in this paper, we propose an autocoding method including the implementation task of the training data by considering the highest score of reliability scores for each text description. This method can improve the training data by using its own dataset, so does not require a large amount of fine training data and other kinds of additional datasets.

In addition, as it is assumed that data classified with high-reliability scores are clearer data, adding the information of those data to the training dataset is performed to obtain a better classification accuracy. First, the proposed method performs the label assignment to evaluate data with the original training dataset. Then, we extract data whose reliability scores are relatively high evaluated data. After that, it adds the information of the extracted data to the original training dataset. Then, it performs the label re-assignment with the improved training dataset to the evaluation data.

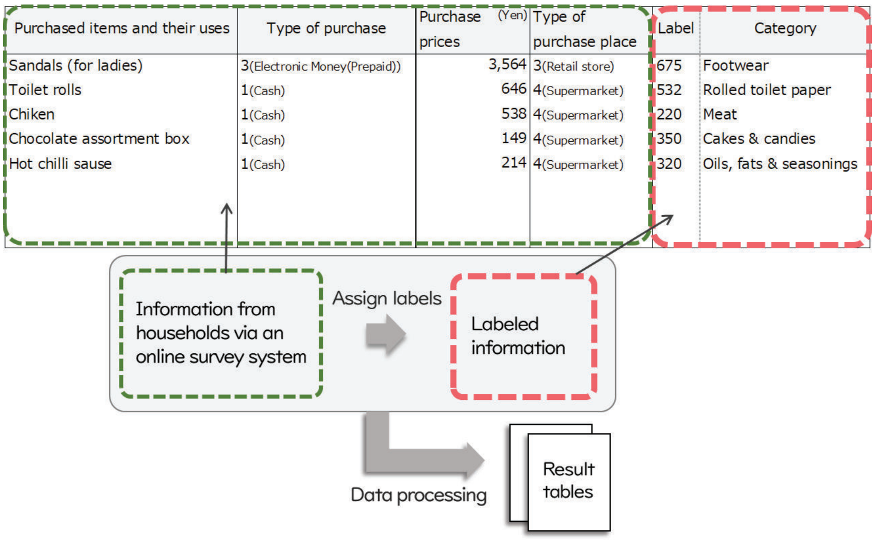
The rest of this paper is organized as follows: The classification method based on the reliability score is explained in section 2. The new method for the improvement of training data is proposed in section 3. The numerical examples are illustrated in section 4, and conclusions are described in section 5.

2. CLASSIFICATION METHOD BASED ON RELIABILITY SCORE

In the family income and consumption survey, households are asked to keep their daily incomes and expenditures, including images of shopping receipts. The responded information includes purchased items' names or receipt items' names in short text descriptions. Using that information, we assign a corresponding label for each text description. Then we use the labeled information for data processing to create result tables (see figure 1).

Coding task for data related to family income and consumption survey

Fig1



Purchased items and their uses	Type of purchase	Purchase prices (Yen)	Type of purchase place	Label	Category
Sandals (for ladies)	3(Electronic Money(Prepaid))	3,564	3(Retail store)	675	Footwear
Toilet rolls	1(Cash)	646	4(Supermarket)	532	Rolled toilet paper
Chiken	1(Cash)	538	4(Supermarket)	220	Meat
Chocolate assortment box	1(Cash)	149	4(Supermarket)	350	Cakes & candies
Hot chilli sause	1(Cash)	214	4(Supermarket)	320	Oils, fats & seasonings

In the classification method, as a pre-processing, we exclude symbols that are not related to the conceptual meaning of the text descriptions. Then, objects (or words) are extracted, and an object frequency table is created. We took word-level N-grams from the word sequences of text descriptions after tokenizing each description using MeCab (Kudo et al., 2004), which is a well-known dictionary-attached morphological Japanese text analyzer. Here, unigrams, bigrams, and entire sentences are considered objects. After objects are extracted, the system tabulates all extracted objects based on their given classification codes into an object frequency table. The classification method based on reliability score performs the extraction on objects and retrieval of candidate labels from the object frequency table provided by using the extracted objects. Then, it calculates the relative frequency of j-th object to a label k defined as

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

where n_{jk} is the number of occurrences of statuses in which an object j is assigned to a label k in the training dataset. J is the number of objects and K is the number of labels.

However, this classifier has difficulty correctly assigning labels to text descriptions for complex data, including uncertainty. To address the problem, we developed the overlapping classifier that assigns labels to each text description based on the reliability score. 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, $\{\tilde{\tilde{p}}_{j1}, \cdots, \tilde{\tilde{p}}_{j\tilde{K}_j}\}, \tilde{K}_j \leq K$ are created. That is, each object has a different number of labels. Then, the classifier calculates the reliability score for each label of each object. The reliability score of j-th object to a label k is defined as

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

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

These reliability scores were defined considering both probabilitv 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 labels. T shows

T-norm in statistical metric space (Menger, 1942, Mizumoto, 1989, Schweizer and Sklar, 2005). In machine learning, the robustness of the results obtained from a model is an important issue. *T*-norm is a kind of metric defined in a function family. Including such a function family, which has a multiple number of exact *T*-norms for the aggregation of probabilities, might affect sustainably obtaining the robustness of the classification model in machine learning. We generalize the reliability score by using the idea of *T*-norm, which is a binary operator in statistical metric space. *T*-norm satisfies the following four conditions:

· Boundary conditions

$$0 \leq T(a,b) \leq 1, \qquad T(a,0) = T(0,b) = 0, \qquad T(a,1) = T(1,a) = a$$

· Monotonicity

$$a \leq c\,, b \leq d \ \rightarrow T(a,b) \leq T(c,d)$$

· Symmetry

$$T(a,b) = T(b,a)$$

· Associativity

$$T(T(a,b),c) = T(a,T(b,c))$$

where $\forall a,b,c,d \in [0,1]$. Though there are numerous possible choices for *T*-norms, the following *T*-norms are employed in this study:

· Algebraic product

$$T(a,b) = ab$$

· Hamacher product

$$T(a,b) = \frac{ab}{p + (1-p)(a+b-ab)}, \quad p \geq 0$$

· Minimum

$$T(a,b) = min\{a,b\}$$

· Einstein product

$$T(a,b) = \frac{ab}{1 + (1-a)(1-b)}$$

Furthermore, to prevent an infrequent object from having significant influence, sigmoid functions $g(n_j)$ were introduced to the reliability score. By using equations (1) and (2), the reliability score considering the frequency of each object over the labels for each object in the training dataset is defined as follows:

$$\bar{\bar{p}}_{jk} = g(n_j) \times \bar{p}_{jk}. \tag{3}$$

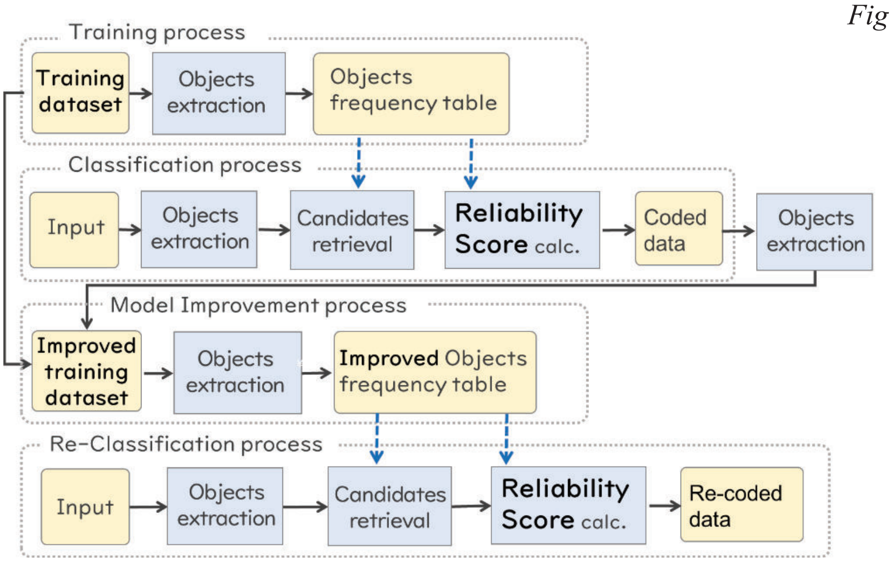
3. PROPOSED METHOD FOR THE IMPROVEMENT OF TRAINING DATA BASED ON RELIABILITY SCORES OF OBJECTS

The proposed method is developed for the improvement of training data based on reliability scores over objects for autocoding.

First, the proposed method performs label assignment, based on the reliability score, to the evaluation data with the original training dataset. Then, it extracts data whose reliability scores are relatively high evaluated data. After that, it adds the information of the extracted data to the original training dataset and creates the improved training dataset. Then, it re-assigns labels to the evaluation data with the improved training dataset. Figure 2 shows the flowchart of label re-assignment with the improved training dataset. It updates the object frequency table using the improved training dataset. Then, it performs label re-assignment based on the reliability score by using a newly calculated object frequency table.

Flowchart of label re-assignment with the improved training dataset

Fig 2



The detailed algorithm of the proposed method is the following:

Step 1. Label assignment by the autocoding method based on the reliability score to evaluation data.

For the reliability score, we determine the following:

· Fuzzy measure to be applied

· *T*-norm in statistical metric space to be applied

· Sigmoid function to be applied

Step 2. Extract data whose reliability scores are relatively high evaluated dataset.

In this study, we extract data whose reliability scores are over 0.99.

Step 3. Add the information of the extracted data to the original training dataset to improve the training dataset.

Step 4. Label re-assignment with the improved training dataset to evaluation data.

4. NUMERICAL EXAMPLE

For the numerical example, the proposed method is applied to the *National Survey of Family Income, Consumption and Wealth* dataset. The *National Survey of Family Income, Consumption and Wealth*, conducted every five years by the Statistics Bureau of Japan, is a sampling survey investigating family consumption, income, assets and liabilities. 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), including text descriptions obtained from shopping receipt images and their corresponding labels. The size of the targeted data from the *National Survey of Family Income, Consumption and Wealth* is approximately 5,750,000. 70% of the randomly extracted targeted data is used for training, and the remaining 30% is used for evaluation. In addition, the *Family Income and Expenditure Survey* dataset is also used for training as it is like the *National Survey of Family Income, Consumption and Wealth,* including the classification structure. The *family Income and Expenditure Survey*, monthly conducted by the Statistics Bureau of Japan, is a sampling survey related to household income and expenditure. For this numerical example, the corresponding label for each text description in the *Family Income and Expenditure Survey* dataset is rearranged according to the classification used in the *National Survey of Family Income, Consumption* and Wealth. In fact, around 350 different labels are available in the *National Survey of Family Income, Consumption, and Wealth*, whereas around 520 different labels are available in the *Family Income and Expenditure Survey*. The size of the targeted data from the Family Income and Expenditure Survey is approximately 19,000,000.

The proposed method was developed by combining VB.Net and R. As the practically implementing autocoding system based on reliability scores has been developed in VB.Net, we utilized it for training and label assignment. Meanwhile, the other parts were developed in R, as R has advantages in data wrangling and a rich set of packages for data handling.

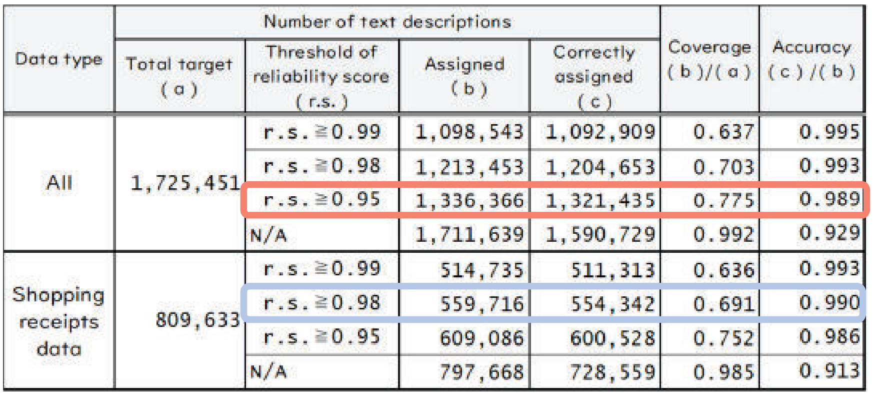
Table 1 shows the classification accuracy of the classification method with the original training dataset. In this numerical example, partition entropy was applied as a fuzzy measure. Hamacher product was applied as *T*-norm in statistical metric space and $n_j/\sqrt{1+n_j{}^2}$ is taken as a sigmoid function. That is, the reliability score shown in equation (3) is formulated as follows:

$$\bar{\bar{p}}_{jk} = \frac{n_j}{\sqrt{1+n_j{}^2}}\left(\frac{\tilde{\tilde{p}}_{jk}\left(1+\sum_{m=1}^{\tilde{K}_j}\tilde{\tilde{p}}_{jm}\log_K\tilde{\tilde{p}}_{jm}\right)}{1+\sum_{m=1}^{\tilde{K}_j}\tilde{\tilde{p}}_{jm}\log_K\tilde{\tilde{p}}_{jm}-\tilde{\tilde{p}}_{jk}\sum_{m=1}^{\tilde{K}_j}\tilde{\tilde{p}}_{jm}\log_K\tilde{\tilde{p}}_{jm}}\right). \qquad (4)$$

From table 1, it is found the classification method assigned labels to over 77.5% of target data with about 0.99 of classification accuracy with the original training dataset. When we focus on shopping receipts data, which contains more complex data, the classification method assigned labels to over 69.1% of target data with over 0.99 of classification accuracy. On the other hand, table 2 shows the classification accuracy of the proposed method with the improved training dataset. From table 1 and table 2, it is found that the proposed method increases the number of assigned data holding higher values of classification accuracy. For example, when the threshold of reliability score is set as 0.95, the number of assigned text descriptions increases by over 4,000, holding about 0.99 of classification accuracy. Also, for the shopping receipts data, when the threshold of reliability score is set as 0.98, the number of assigned text descriptions increases by over 2,000 holding 0.99 classification accuracy. In addition, by comparison between the classification method with the original training dataset and the proposed method, even if the classification accuracy is the same between two methods, however, the coverage is increased for the proposed method.

Classification accuracy of the classification method based on reliability score

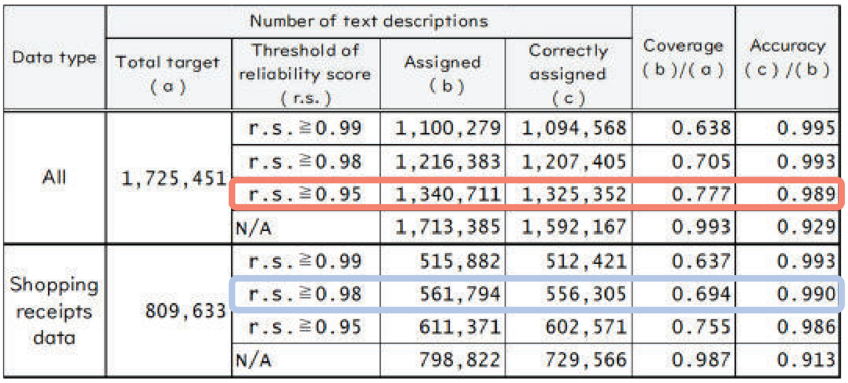
Table 1



Data type	Number of text descriptions				Coverage (b)/(a)	Accuracy (c) /(b)
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)		
All	1,725,451	r.s.≧0.99	1,098,543	1,092,909	0.637	0.995
		r.s.≧0.98	1,213,453	1,204,653	0.703	0.993
		r.s.≧0.95	1,336,366	1,321,435	0.775	0.989
		N/A	1,711,639	1,590,729	0.992	0.929
Shopping receipts data	809,633	r.s.≧0.99	514,735	511,313	0.636	0.993
		r.s.≧0.98	559,716	554,342	0.691	0.990
		r.s.≧0.95	609,086	600,528	0.752	0.986
		N/A	797,668	728,559	0.985	0.913

Classification accuracy of the proposed method with the improved training dataset

Table 2

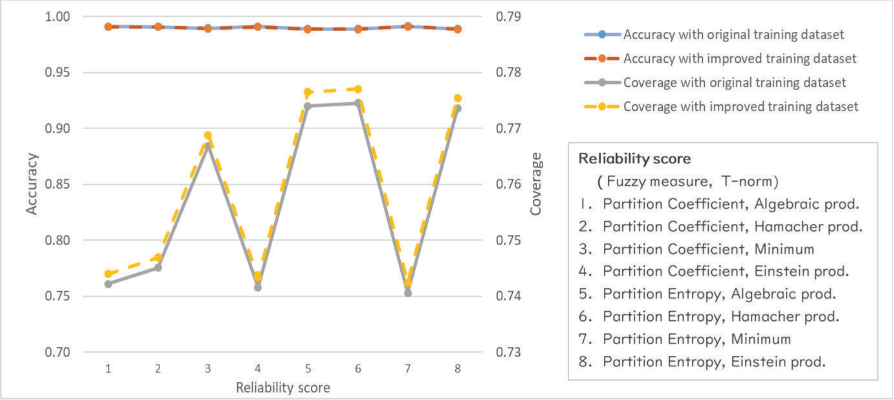


Data type	Number of text descriptions				Coverage (b)/(a)	Accuracy (c) /(b)
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)		
All	1,725,451	r.s.≧0.99	1,100,279	1,094,568	0.638	0.995
		r.s.≧0.98	1,216,383	1,207,405	0.705	0.993
		r.s.≧0.95	1,340,711	1,325,352	0.777	0.989
		N/A	1,713,385	1,592,167	0.993	0.929
Shopping receipts data	809,633	r.s.≧0.99	515,882	512,421	0.637	0.993
		r.s.≧0.98	561,794	556,305	0.694	0.990
		r.s.≧0.95	611,371	602,571	0.755	0.986
		N/A	798,822	729,566	0.987	0.913

Also, we compared the proposed method with the previous method for eight kinds of reliability scores. Figure 1 shows the proposed method increases the coverage holding higher values of classification accuracy for each reliability score. The blue line shows the classification accuracy of the classification method with the original training dataset, and the orange dotted line shows the classification accuracy of the classification method with the improved training dataset. These two lines seem to overlap. Meanwhile, the gray line shows coverage of the classification method with the original training dataset, and the yellow dotted line shows coverage of the classification method with the improved training dataset. This means that the high accuracy can keep for the proposed method, which is the same as the ordinary method; however, the proposed method can obtain better coverage compared with the ordinary method. That is, the proposed method increases the number of correctly assigned data. And this tendency has robustness over the difference in the reliability score.

Classification accuracy and coverage of both the previous method and the proposed method for each reliability score

Fig. 1



In addition, the proposed method is applied to the *Family Income and Expenditure Survey* dataset. In this numerical example, we used the following data from the *Family Income and Expenditure Survey* dataset: data from the 2018 to 2020 survey were used for training, and data from January to March 2021 survey were used for evaluation. The size of the training data from the *Family Income and Expenditure Survey* is approximately 9,500,000, whereas the size of the evaluation data for each month is approximately 700,000 on average. For the reliability score, the partition coefficient was applied as a fuzzy measure, and the Algebraic product was applied as *T*-norm in statistical metric space and $n_j/\sqrt{1+n_j{}^2}$ is taken as sigmoid function.

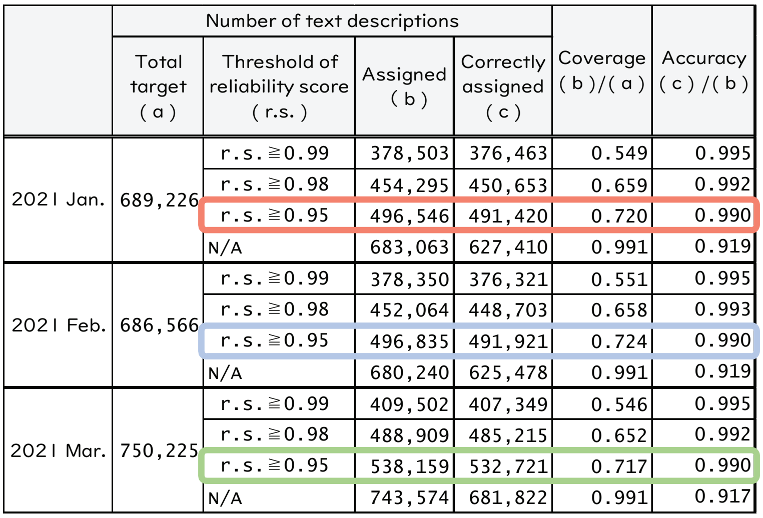
Table 3 shows the classification accuracy of the classification method based on reliability score, and table 4 shows the classification accuracy of the proposed method with the improved training dataset. From table 3 and table 4, it is found that the proposed method increases the number of assigned data holding higher values of classification accuracy for the *Family Income and Expenditure Survey* dataset too. For example, when we set the accuracy as

0.99, then for all months as January, February, and March 2021, coverage is higher scores for the proposed method shown in table 4 when compared with the scores of the previously proposed method shown in table 3. Moreover, as the *Family Income and Expenditure Survey* is a monthly survey, the model improvement process can be iteratively performed. Therefore, the degree of improvement might improve each month as the information of clearer data is iteratively added.

In the coding tasks of the *Family Income and Expenditure Survey,* we add the information from the previous month's survey data to the training dataset every month before we perform the label assignment. For considering a real application to this data, we extract data whose reliability scores are relatively high but not added in the improvement step from the previous month's survey data and additionally add the extracted information to the training dataset before we perform label assignment. In this numerical example, we extracted data whose reliability scores are 0.98 to 0.99 (a) or 0.95 to 0.99 (b) from the previous month's survey data. We started with the 2021 January survey data, and then the extended method was applied to the 2021 February survey data. Table 5 shows the classification accuracy when data from the previous month's survey data whose reliability scores are 0.98 to 0.99 were added, and table 6 shows the classification accuracy when data from the previous month's survey data whose reliability scores are 0.95 to 0.99 were added. From tables 5 and 6, from the cases when the score of the accuracy is 0.99, even only using the previous month's data, the proposed method can obtain almost similar accuracy to the result shown in table 4. That is, considering real data processing on the coding task for the *Family Income and Expenditure Survey*, the proposed method has certain validity for real applications.

Classification accuracy of the classification method based on reliability score

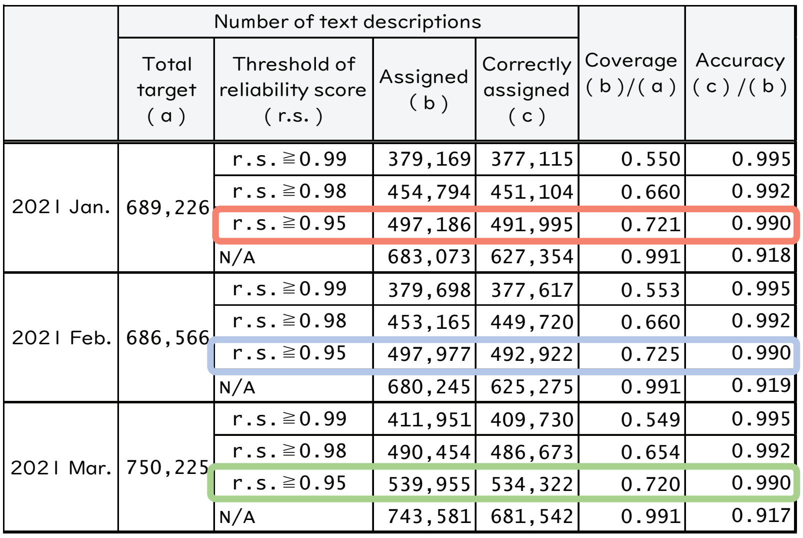
Table 3



	Number of text descriptions				Coverage	Accuracy
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)	(b)/(a)	(c) /(b)
2021 Jan.	689,226	r.s.≧0.99	378,503	376,463	0.549	0.995
		r.s.≧0.98	454,295	450,653	0.659	0.992
		r.s.≧0.95	496,546	491,420	0.720	0.990
		N/A	683,063	627,410	0.991	0.919
2021 Feb.	686,566	r.s.≧0.99	378,350	376,321	0.551	0.995
		r.s.≧0.98	452,064	448,703	0.658	0.993
		r.s.≧0.95	496,835	491,921	0.724	0.990
		N/A	680,240	625,478	0.991	0.919
2021 Mar.	750,225	r.s.≧0.99	409,502	407,349	0.546	0.995
		r.s.≧0.98	488,909	485,215	0.652	0.992
		r.s.≧0.95	538,159	532,721	0.717	0.990
		N/A	743,574	681,822	0.991	0.917

Classification accuracy of the proposed method with the improved training dataset

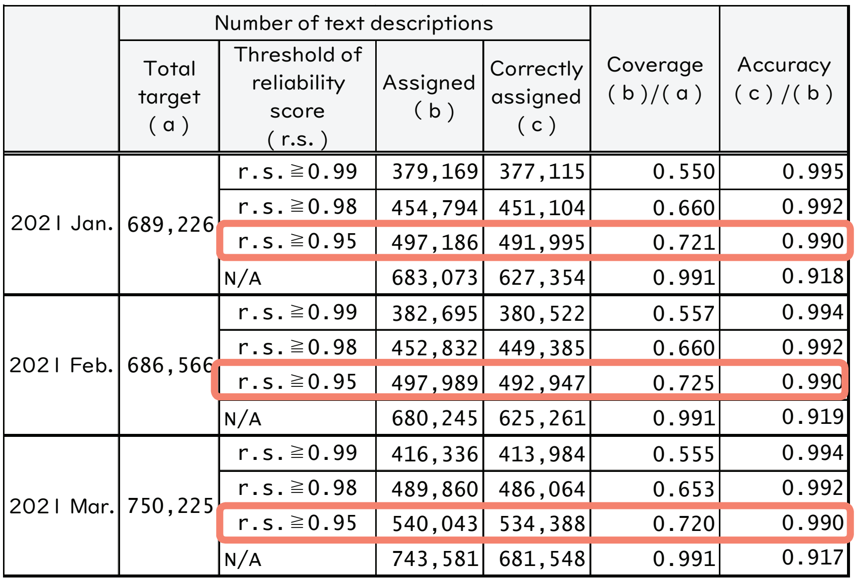
Table 4



	Number of text descriptions				Coverage	Accuracy
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)	(b)/(a)	(c) /(b)
2021 Jan.	689,226	r.s.≧0.99	379,169	377,115	0.550	0.995
		r.s.≧0.98	454,794	451,104	0.660	0.992
		r.s.≧0.95	497,186	491,995	0.721	0.990
		N/A	683,073	627,354	0.991	0.918
2021 Feb.	686,566	r.s.≧0.99	379,698	377,617	0.553	0.995
		r.s.≧0.98	453,165	449,720	0.660	0.992
		r.s.≧0.95	497,977	492,922	0.725	0.990
		N/A	680,245	625,275	0.991	0.919
2021 Mar.	750,225	r.s.≧0.99	411,951	409,730	0.549	0.995
		r.s.≧0.98	490,454	486,673	0.654	0.992
		r.s.≧0.95	539,955	534,322	0.720	0.990
		N/A	743,581	681,542	0.991	0.917

Classification accuracy of extended method (a)

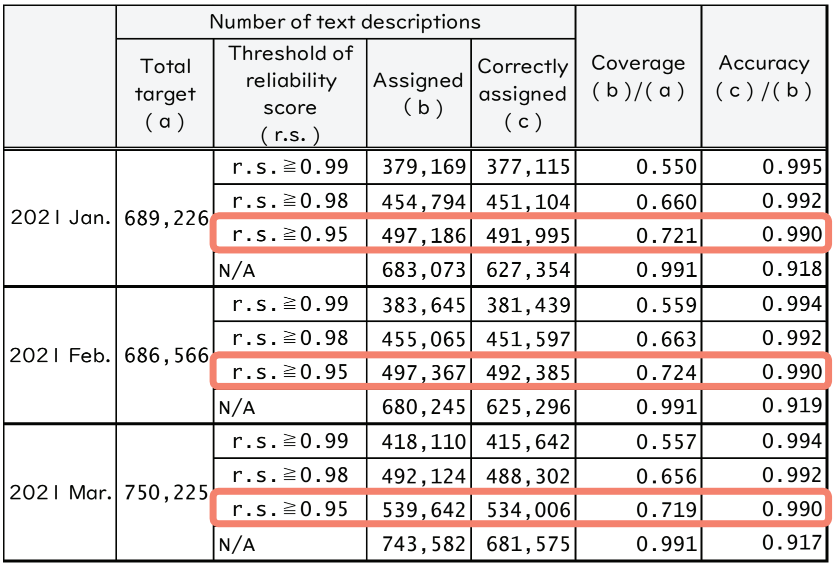
Table 5



	Number of text descriptions				Coverage (b)/(a)	Accuracy (c) /(b)
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)		
2021 Jan.	689,226	r.s.≧0.99	379,169	377,115	0.550	0.995
		r.s.≧0.98	454,794	451,104	0.660	0.992
		r.s.≧0.95	497,186	491,995	0.721	0.990
		N/A	683,073	627,354	0.991	0.918
2021 Feb.	686,566	r.s.≧0.99	382,695	380,522	0.557	0.994
		r.s.≧0.98	452,832	449,385	0.660	0.992
		r.s.≧0.95	497,989	492,947	0.725	0.990
		N/A	680,245	625,261	0.991	0.919
2021 Mar.	750,225	r.s.≧0.99	416,336	413,984	0.555	0.994
		r.s.≧0.98	489,860	486,064	0.653	0.992
		r.s.≧0.95	540,043	534,388	0.720	0.990
		N/A	743,581	681,548	0.991	0.917

Classification accuracy of extended method (b)

Table 6



	Number of text descriptions				Coverage (b)/(a)	Accuracy (c) /(b)
	Total target (a)	Threshold of reliability score (r.s.)	Assigned (b)	Correctly assigned (c)		
2021 Jan.	689,226	r.s.≧0.99	379,169	377,115	0.550	0.995
		r.s.≧0.98	454,794	451,104	0.660	0.992
		r.s.≧0.95	497,186	491,995	0.721	0.990
		N/A	683,073	627,354	0.991	0.918
2021 Feb.	686,566	r.s.≧0.99	383,645	381,439	0.559	0.994
		r.s.≧0.98	455,065	451,597	0.663	0.992
		r.s.≧0.95	497,367	492,385	0.724	0.990
		N/A	680,245	625,296	0.991	0.919
2021 Mar.	750,225	r.s.≧0.99	418,110	415,642	0.557	0.994
		r.s.≧0.98	492,124	488,302	0.656	0.992
		r.s.≧0.95	539,642	534,006	0.719	0.990
		N/A	743,582	681,575	0.991	0.917

5. CONCLUSION

This paper proposes a new method for the improvement of training data based on the reliability scores of objects for autocoding. This method includes the filtering task for the construction of training data in our previously proposed machine learning language model based on the idea that data classified with high-reliability scores are clearer data; adding the information of those data to the training dataset is performed to obtain a better classification accuracy. Generally, machine learning language model requires fine-tuning tasks for the training data, and this task requires a considerable amount of training data or different kinds of datasets which have essential difficulties in collecting and even we succeed in obtaining such a large amount of data; however, they may still have bias. To overcome this problem, the proposed method takes a different approach using only the initially obtained training data, so other kinds of additional datasets are not required, because the new approach is recursively improving using its own dataset.

The numerical examples show a better performance of the proposed method with both the *National Survey of Family Income, Consumption and Wealth* data and the *Family Income and Expenditure Survey* data.

References

1. **Bezdek, J.C.** (1981), *Pattern recognition with fuzzy objective function algorithms*, Plenum Press.
2. **Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R.** (1999), *Fuzzy models and algorithms for pattern recognition and image processing,* Kluwer Academic Publishers.
3. **Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan , J., Dhariwal, P., Neelakantan A., Shyam, P., Sastry , G., Askell, A., Agarwal , S., Herbert-Voss, A., Krueger, G., Henighan , T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.** (2020), *"Language Models are Few-Shot Learners"*, arXiv:2005.14165
4. **Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.** (2018), *"Local rule-based explanations of black box decision systems"* arXiv preprint arXiv:1805.10820.
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. **Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., Irving, G.** (2021), *"Alignment of Language Agents"*, DeepMind, arXiv:2103.14659.
8. **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, Barcelona, Spain, 25-26, Jul. 2004, pp. 230-237.

9. **Menger, K.** (1942), *"Statistical metrics"*, in Proceedings of the National Academy of Sciences of the United States of America, Vol. 28, pp. 535-537.
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. **Ribeiro, M.T., Singh, S., Guestrin, C.** (2018), *"Anchors: High-precision model-agnostic Explanations"*, The Thirty-Second AAAI Conference on Artificial Intelligence, pp. 1527-1535.
12. **Schweizer**, **S., Sklar, A.** (2005), *Probabilistic metric spaces*, Dover Publications.
13. **Solaiman, I., Dennison, C.** (2021), *"Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets"*, 35th Conference on Neural Information Processing Systems (NeurIPS 2021), arXiv:2106.10328.
14. **Toko, Y., Iijima, S., Sato-Ilic, M.** (2018), "*Overlapping classification for autocoding system"*, Romanian Statistical Review, Vol. 4, pp. 58-73.
15. **Toko, Y., Iijima, S., Sato-Ilic, M.** (2019), *"Generalization for improvement of the reliability score for autocoding"*, Romanian Statistical Review, Vol. 3, pp. 47–59.
16. **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.
17. **Toko, Y., Sato-Ilic, M**. (2021), "*Efficient autocoding method in high dimensional space"*, Journal of Romanian Statistical Review, Vol. 1, pp. 3-16.
18. **Toko, Y., Sato-Ilic, M**. (2022), *"Autocoding based multi-class support vector machine by fuzzy c-means"*, Romanian Statistical Review, Vol. 1, pp. 27-39.
19. **Statistics Bureau of Japan**: Outline of the National Survey of Family Income, Consumption, and Wealth. Available at: https://www.stat.go.jp/english/data/zenkokukakei/index.html, last accessed 2022/10/31

Impact of Main Skills Targeted by CVT Courses on Economic Activities of Enterprises

Nicolae Marius JULA (marius.jula@faa.unibuc.ro)
Faculty of Business and Administration, University of Bucharest,

Dorin JULA (dorinjula@yahoo.fr)
Institute of Economic Forecasting, Romanian Academy and Ecological University of Bucharest

ABSTRACT

In this paper, we analyse the impact of Continuing Vocational Training (CVT) courses on turnover, value-added and gross operating surplus, labour productivity (apparent and wage adjusted), and other special aggregates (dimensional and qualitative) of enterprises' economic activity. The analysis is developed (for 27 European countries, during the 2010-2020 period) starting with the main type of skills targeted by CVT courses (general and professional IT skills, management and office administration skills, team working, customer handling, problem-solving, and numeracy skills, foreign language, and communication skills, technical, practical, or job-specific and other skills and competences). As a methodology, we use R software to solve econometric models with panel data (time series and longitudinal data).

Keywords: *Continuing Vocational Training, employees' skills, panel data models, R software*

JEL Classification: *C33, I26, M53*

INTRODUCTION

It is generally accepted that firms invest in their employees by providing Continuing Vocational Training (CVT). In an economic context marked by globalization, population aging, and technological advancement, businesses engage in training to boost productivity and regularly update the abilities of their employees (automation and digitalization).

The goal of CVT (Continuing Vocational Training) courses is to raise employee skill levels, which will have a favourable effect on businesses' economic activity. In the following areas the impact should be most visible:

- ✓ Increased productivity: Improved skills lead to better performance and increased efficiency.
- ✓ Improved competitiveness: Up-to-date skills help companies remain competitive in their industry.

- ✓ Attraction and retention of skilled employees: CVT courses can make a company more attractive to job candidates and help retain current employees.
- ✓ Innovation: New skills and knowledge can drive innovation and creativity within a company.
- ✓ Improved financial performance: Companies with highly skilled employees are more likely to have strong financial performance.
- ✓ Better customer satisfaction: Improved skills can lead to better customer service and satisfaction.
- ✓ Adaptation to changing market conditions: CVT courses can help companies stay ahead of market changes and adapt to new technologies and trends.
- ✓ Improved risk management: Enhanced skills can help companies manage risks and make informed decisions.

As Brunello & Wruuck (2020) suggest, human capital investments are essential to the health and expansion of the economy. Making investments in human capital is crucial to maintaining high levels of competitiveness and employment at a time when tastes and technologies are changing quickly. It is challenging to fully benefit from technological advancements without a staff that is constantly learning new skills. A lack of appropriate abilities in the workforce can also make inequality worse.

Some papers intensively discuss the need for continuous training to decrease employee gaps, but also to provide useful skills for juniors or newly graduated employees.

In Gabor, Blaga, & Matis (2019) the authors suggest that the CVT can lead to:

- ✓ a decrease in employee turnover in the first few months of employment.
- ✓ a decrease in the possibility that workers will lose interest in their jobs and become at risk of long-term unemployment.
- ✓ assistance to employers in obtaining the maximum commitment and productivity from young workers.
- ✓ improvement in employees' capacity to adapt to and settle into subsequent jobs.
- ✓ creation of a stronger foundation for long-term employment.

According to the findings of Konings & Vanormelingen (2015), training-related productivity growth is greater than salary growth. More specifically, an increase of 10 percentage points in the proportion of workers who obtain training results in an increase in effective labor input of 1.7% to 3.2%, but only 1% to 1.7% in average wage growth. The authors suggest that their results are statistically significant and consistent, and the difference between the salary premium and the productivity premium holds across a wide variety of variables. The authors argue that training has a marginally greater influence in non-manufacturing sectors than in manufacturing ones.

Another aspect to be considered in the actual context of automatization and advances in AI is studied by Pouliakas (2018). In his study, he identifies factors that contribute to "automatability risk," or the likelihood that EU workers work in positions that are susceptible to being replaced by machines, robots, or other algorithmic processes, and examines how this risk affects labor market outcomes. The risk of automation is estimated to be very high for 14% of adult workers in the EU. It is also discovered that routine professions with minimal demand for transversal and social skills are more likely to have high automatability dispersion across industries and occupations. There is a minimal indication of polarization, and the risk of job displacement by robots is larger for men and lower-skilled individuals. It is common in occupations in the private sector that do not offer employees additional training, highlighting the vulnerability of at-risk workers and the demand for more robust lifelong learning programs at the EU level.

The effects of continuing vocational training on production and efficiency indicators were studied, among others, by Zwick (2005), Kuckulen (2007) and Morris, Steinmüller & Rohs (2022) for Germany, Mara (2018) for Catalonia, Makkonen & Lin (2012), European Commission (2022) Markowitsch & Bjørnåvold (2022) and Milmeister, Rastoder & Houssemand (2022) for Europe.

The available literature still offers insufficient support for the advantages of training for businesses in terms of its effects on production. In turn, a more methodical evaluation of the advantages, considering how they vary by nation, might help to explain the disparity in training expenditure and activity across the EU.

1. DATA AND METHODOLOGY

a. Methodology

We analysed the impact of Continuing Vocational Training (CVT) courses on the enterprises' economic activity, based on:

✓ Production indicators:

− *Turnover (*or *Gross Premiums Written),*

− *Production Value*

− *Value Added* at factor cost

✓ Efficiency indicators:

− *Turnover per Person Employed,*

– *Apparent Labour Productivity* (Gross Value Added per person employed)

– *Gross Operating Surplus/Turnover* (Gross Operating Rate)

✓ Development indicators:

– *Growth Rate of Employment* (percentage)

– *Investment per person employed*

The analysis is carried out for 27 European countries, over the period 2010-2020, starting from the main types of skills targeted by CVT courses.

We built an econometric model with panel data for each economic indicator. The model is as follows:

$$\mathrm{ind}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \mathrm{skills}_{k,it} + \{\text{control variables}\} + \mathrm{ISE}_i + e_{it}$$

where: i – index of cross-section data (European countries).

t – time index (2010, 2015, 2020)

ind_{it} – economic indicators {three production indicators, three efficiency indicators, two development indicators}.

skills_k – skills targeted by CVT courses {General IT skills, Professional IT skills, Management skills, Team working skills, Customer handling skills, Problem solving skills, Office administration skills, Foreign language skills, Technical, practical or job-specific skills, Oral or written communication skills, Numeracy and/or literacy skills, Other skills and competences}. In the models, we didn't do calculations for "Other skills and competences".

control variables: Gross domestic product /capita at market prices (index 2015 = 100), and Gross domestic product /capita, at UE27 level, percentage change on the previous period.

ISE_i – Individual specific effects (fixed or random)

e_{it} – Idiosyncratic errors

The eight econometric models are detailed in the Annex (one econometric model with panel data, for each indicator from the three groups).

Technically, we use **R** software to solve econometric models (Jula & Jula, 2019) with panel data – time series and cross-section data (Croissant & Millo, 2019), more specific, *plm* package. According to the description, the *plm* package in R is a tool for estimating linear panel data models. It can handle fixed effects, random effects, and first-difference models, as well as unbalanced panels with missing data. The package provides diagnostic tests and robust standard error options. The *plm* package is suitable for empirical research in various fields.

b. Data

For the models, we used the *main skills needed for the development of the enterprise by type of skill, % of all enterprises*. As the source of data, we used Eurostat, table *trng_cvt_10n2*, at https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=trng_cvt_10n2&lang=en (extracted on November 06, 2022).

The next variable was related to the *main skills targeted by CVT courses by type of skill, % of enterprises providing CVT courses,* and the source of data: Eurostat, table *trng_cvt_29n2*, at: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=trng_cvt_29n2&lang=en (extracted on November 06, 2022). The data are for the Eurostat Continuing Vocational Training Survey. The survey methodology is described in (Eurostat, 2022).

Data refer to 27 European countries (Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom) for three years (2010, 2015 and 2020). The countries were selected based on data availability.

The next two variables, *rGDP* and *iGDP2015* are defined as:

rGDP – Gross Domestic Product / capita (at market prices), chain-linked volumes, percentage change on the previous period,

iGDP2015 – Gross Domestic Product / capita (at market prices), chain-linked volumes, index 2015 = 100,

and we used as the source of data: Eurostat, *Main GDP aggregates per capita*, table nama_10_pc, https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nama_10_pc&lang=en (extracted on Nov. 20, 2022).

c. Main skills needed/targeted by Continuing Vocational Training (CVT) courses

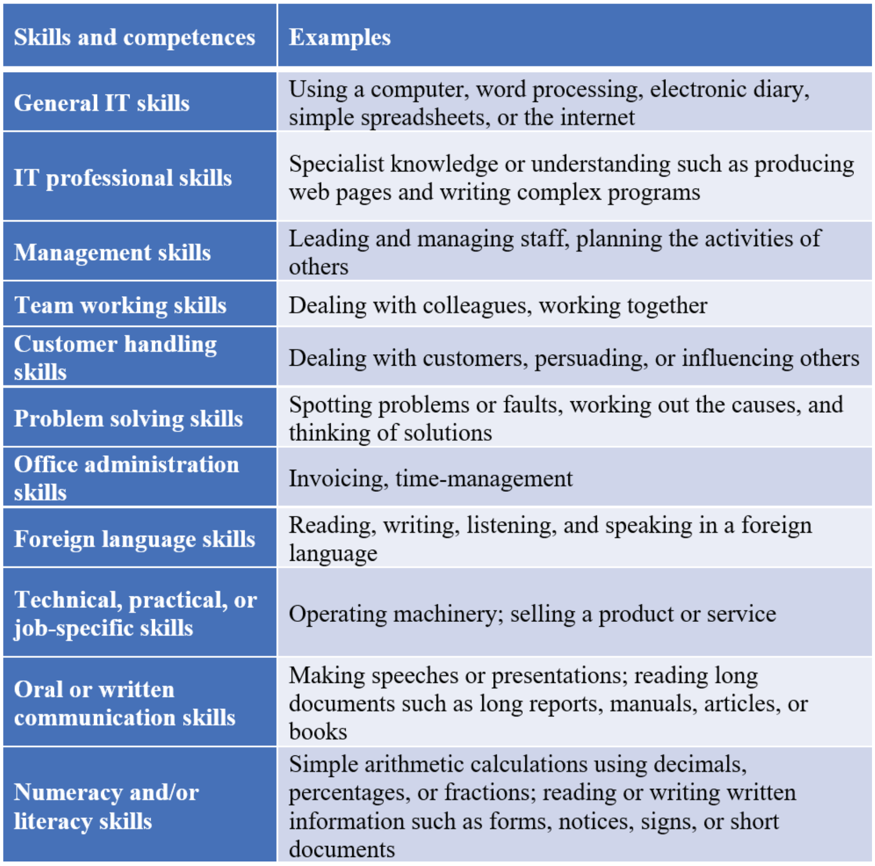
According to *Eurostat* (2022) definition:

> "*Continuing vocational training* (CVT) are training measures or activities which have as their primary objectives the acquisition of new competencies or the development and improvement of existing ones and which must be financed at least partly by the enterprises for their persons employed who either have a working contract or who benefit directly from their work for the enterprise such as unpaid family workers and casual workers" (Eurostat, 2022, p. Table 1) – retrieved on November 2022, from the following address: https://ec.europa.eu/eurostat/cache/metadata/en/trng_cvt_esms.htm).

The skills needed/targeted by Continuing Vocational Training courses, registered by Eurostat are detailed *Image 1.*

Skills needed/targeted by Continuing Vocational Training courses

Image 1



Skills and competences	Examples
General IT skills	Using a computer, word processing, electronic diary, simple spreadsheets, or the internet
IT professional skills	Specialist knowledge or understanding such as producing web pages and writing complex programs
Management skills	Leading and managing staff, planning the activities of others
Team working skills	Dealing with colleagues, working together
Customer handling skills	Dealing with customers, persuading, or influencing others
Problem solving skills	Spotting problems or faults, working out the causes, and thinking of solutions
Office administration skills	Invoicing, time-management
Foreign language skills	Reading, writing, listening, and speaking in a foreign language
Technical, practical, or job-specific skills	Operating machinery; selling a product or service
Oral or written communication skills	Making speeches or presentations; reading long documents such as long reports, manuals, articles, or books
Numeracy and/or literacy skills	Simple arithmetic calculations using decimals, percentages, or fractions; reading or writing written information such as forms, notices, signs, or short documents

Source: Eurostat, *CVTS 6 manual – Annexes*, Table A5.a: Types of skills and competences for the data collection in CVTS, p. 15, https://circabc.europa.eu/ui/group/d14c857a-601d-438a-b878-4b4cebd0e10f/library/b2d54858-dedb-4a5a-a12d-e08ea0198b2a/details

2. RESULTS

We have tested the impact of main skills needed for the development of the enterprise on economic activities, using 8 models.

a. Impact on production indicators

First, we analysed the impact on 3 production indicators:

✓ *Turnover* or *Gross Premiums Written* (Model 1)

✓ *Production value* (Model 2)

✓ *Value A dded at factor cost* (Model 3)

Model 1: **Turnover** or **Gross Premiums Written** (symbol: *tur*, million euro)

We used the following equation:

$$\mathrm{tur}_{\mathrm{it}} = \mathrm{a}_0 + \sum_{\mathrm{k}=1}^{11} \mathrm{a}_{\mathrm{k}} \cdot \mathrm{skills}_{\mathrm{k,it}} + \{\mathrm{control\ var iables}\} + \mathrm{ISE}_{\mathrm{i}} + \mathrm{e}_{\mathrm{it}}$$

where:

$\mathrm{skills}_{\mathrm{k}}$ ∈ {General IT skills, Professional IT skills, Management skills, Team working skills, Customer handling skills, Problem solving skills, Office administration skills, Foreign language skills, Technical, practical or job-specific skills, Oral or written communication skills, Numeracy and/or literacy skills, Other skills and competences}. In the models, we didn't do calculations for "Other skills and competences".

i – index of cross-section data (European countries).

t – time index (2010, 2015, 2020)

$\mathrm{ISE}_{\mathrm{i}}$ – Individual specific effects (fixed or random)

e_{it} – Idiosyncratic errors

The control variables are:

GDP_{it} – Gross domestic product /capita at market prices (index 2015 = 100), for country *i* and year = *t*.

$\mathrm{rGDP}_{\mathrm{EU27}}$ – Gross domestic product /capita (market prices), at UE27 level, percentage change on the previous period.

According to the results of the model, the effects of different skills needed for the development of the enterprise on the *Turnover* (or *Gross Premiums Written*) can be divided into *Positive Impact* and *No significant positive impact value*. Using this approach, we found out that there are 6 skills with a *positive impact*:

Professional IT skills	2194.190
Management skills	1110.353
Team working skills	740.706
Problem solving skills	1335.634
Office administration skills	3965.717
Foreign language skills	4681.331

(*Source*: authors' calculations)

All the coefficients of variables with positive impact are significantly different from zero at the standard threshold of 5%. With *no positive impact value*, we found: General IT skills, Customer handling skills, Technical, practical, or job-specific skills, and Oral or written communication skills and Numeracy and/or literacy skills.

Also, the model indicates positive individual-specific fixed effects for Belgium, France, Germany, Italy, Netherlands, Spain, and United Kingdom, and negative ones for Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, and Sweden. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}). The control variables have a significant, positive impact on *turnover* (or *Gross Premiums Written*). The econometric model is detailed in Annex.

Model 2: **Production value** (symbol: *prod*, million euro)

$$\mathrm{prod}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \mathrm{skills}_{k,it} + \{\mathrm{control\ var iables}\} + \mathrm{ISE}_i + e_{it}$$

Except for “prod” (i.e., *production value*), the other symbols are identical to those used in the first model. As control variable, only the *Gross domestic product /capita* (market prices), at UE27 level (percentage change on the previous period) has a significant effect.

According to the results of this second model, the effects of different skills needed for the development of the enterprise on the *production value* can be divided, likewise, into *positive impact* and *no significant positive impact values*. The variables with a positive impact on the *production value* are the following:

Positive impact:	
General IT skills	163.862
Professional IT skills	102.033
Management skills	1865.366
Team working skills	2045.978
Problem solving skills	330.2339
Numeracy and/or literacy skills	897.750

(Source: authors' calculations)

Except for *General* and *Professional IT skills*, the other coefficients are significantly different from zero at the 5% thresholds. With *no positive impact value*, we found: Customer handling skills, Office administration skills, Foreign language skills, Technical, practical, or job-specific skills, and Oral or written communication skills.

The second model indicates positive individual-specific fixed effects for Belgium, France, Germany, Italy, Netherlands, Spain, and Poland, and negative ones for Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Lithuania, Luxembourg, Malta, Norway, Portugal, Romania, Slovakia, Slovenia, and Sweden. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

Model 3: **Value added** at factor cost (symbol: *vad*, million euro)

The equation is the following:

$$\mathrm{vad}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \mathrm{skills}_{k,it} + \{\text{control var iables}\} + \mathrm{ISE}_i + e_{it}$$

Except for "vad" (i.e. *value added*), the other symbols are identical to those used in the previous models are the same. As control variable, we used only the *Gross domestic product /capita* at market prices (index 2015 = 100), for country *i* and year = *t*. The impact on *value added* is positive for the same set of needed skills as in the first model.

Positive impact:

Professional IT skills	930.169
Management skills	306.438
Team working skills	45.260
Problem solving skills	773.613
Office administration skills	235.422
Foreign language skills	290.923
Numeracy and/or literacy skills	1224.888

(Source: authors' calculations)

Except for *Team working skills*, the other coefficients are significantly different from zero at the standard thresholds. Also, the control variable induced a positive effect on value added. Likewise, the model indicates positive individual-specific fixed effects for Belgium, France, Germany, Italy, Netherlands, Spain, and Sweden . The negative individual specific fixed effects are for Poland and Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Lithuania, Luxembourg, Malta, Norway, Portugal, Romania, Slovakia, and Slovenia. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

b. Impact on efficiency indicators

Next, we tested the *impact of needed skills* on the following *efficiency indicators*:

✓ *Turnover per person employed* (Model 4)

✓ *Apparent labour productivity* (Gross Value Added per person employed) (Model 5)

✓ Gross *Operating Surplus / Turnover* (Gross Operating Rate) (Model 6)

Model 4: **Turnover per person employed** (symbol: *tpe*, thousand euro)

The equation that we used is:

$$\text{tpe}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \text{skills}_{k,it} + \text{ISE}_i + e_{it}$$

Compared to the production indicators, for the *efficiency indicators*, the set of necessary skills that had a positive impact change. They are no longer found in this set the Professional IT skills, Office administration skills, Numeracy and/or literacy skills, Team working skills, and they appear additionally the General IT skills, Customer handling skills, and Oral or written communication skills. As control variable, we used only the *Gross domestic product /capita* (market prices), at UE27 level, percentage change on the previous period. This variable has a significant positive effect on Turnover per person employed. The following required skills had, as well, a *positive impact* on turnover per employed person:

General IT skills	0.716125
Management skills	0.329466
Customer handling skills	0.724130
Problem solving skills	0.000506
Foreign language skills	0.134037
Oral or written communication skills	0.293339

(Source: authors' calculations)

Except for *Problem solving skills*, the other coefficients are significantly different from zero, at the standard thresholds. With *no positive impact value* are: Professional IT skills; Office administration skills, Technical, practical, or job-specific skills, Numeracy and/or literacy skills.

The set of countries for which the individual specific fixed effects are positive changes is: Austria Belgium, Denmark, Finland, France, Germany, Italy, Luxembourg, Netherlands, Norway, Sweden, and United Kingdom. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

Model 5: **Apparent labour productivity** (symbol: *alw*, gross value added per person employed, thousand euro)

The equation we used is:

$$\text{alw}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \text{skills}_{k,it} + \text{ISE}_i + e_{it},$$

where "alw" stands for *apparent labour productivity*.

The control variables did not significantly improve the performance of the econometric model. The skills needed from enterprises with a *positive impact* on *apparent labour productivity* are the following:

General IT skills	0.043272
Professional IT skills	0.106335
Management skills	0.060315
Team working skills	0.034526
Office administration skills	0.030136
Technical, practical, or job-specific skills	0.030672
Numeracy and/or literacy skills	0.147411

(Source: authors' calculations)

This set is different from the one in the previous model, both in structure and in the size of the impacts. The statistical performance of this model is lower than the previous one: 2 out of 7 positive coefficients are not significantly different from zero, at the standard threshold of 5% (i.e. *General IT* and *Office administration skills* coefficients). With *no positive impact value* are Customer handling skills, Problem solving skills, Foreign language skills, and Oral or written communication skills. Compared to the previous model, Italy is no longer among the countries for which a specific positive fixed individual effect is reported. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

Model 6: **Gross operating surplus/turnover** (symbol: *rgos*, gross operating rate, percentage)

We used the equation:

$$\text{rgos}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \text{skills}_{k,it} + \text{ISE}_i + e_{it}$$

where "rgos" is a symbol for stand for *gross operating surplus/turnover* (gross operating rate).

The control variables did not significantly improve the performance of the econometric model. The skills that had a positive impact on *gross operating surplus/turnover* (*gross operating rate*) are the following:

Positive impact:

Professional IT skills	0.040862
Management skills	0.005009
Team working skills	0.019344
Foreign language skills	0.025848
Technical, practical, or job-specific skills	0.018532
Numeracy and/or literacy skills	0.018087

(Source: authors' calculations)

Only the coefficient of the variable *Management skills* is not significant at the 5% level. With *no positive impact value* are General IT skills; Office administration skills, Customer handling skills, Problem solving skills, Oral or written communication skills.

Among the countries for which the specific individual fixed effect is positive, Romania appears, along with Austria, Croatia, Cyprus, Denmark, Italy, Malta, Netherlands, Norway, Poland, and the United Kingdom. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

c) Impact on development indicators

The last 2 models were designed to analyse the impact on development indicators:

✓ *Growth rate of employment* (percentage) (Model 7)

✓ *Investment per person employed* (Model 8)

Model 7: **Growth rate of employment** (symbol: *gre*, percentage)

We used the equation:

$$\mathrm{gre}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \mathrm{skills}_{k,it} + \{\text{control var iables}\} + \mathrm{ISE}_i + e_{it}$$

As a control variable, we used the *Gross domestic product /capita* (market prices), at the UE27 level, percentage change from previous period. The following analysed skills had *positive impact:*

General IT skills	0.027411
Professional IT skills	0.009031
Management skills	0.012052
Team working skills	0.001234
Customer handling skills	0.117043
Foreign language skills	0.073444
Oral or written communication skills	0.226335

(Source: authors' calculations)

Coefficients of the variables *Management* and *Team working skills* are not significant at the standard level.

With *no positive impact value* are: Problem solving skills, Office administration skills, Technical, practical, or job-specific skills, Numeracy and/or literacy skills.

Romania appears with a positive specific positive individual fixed effect, along with Belgium, Croatia, Estonia, France, Germany, Hungary, Italy, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovakia, Slovenia, and Spain. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

Model 8: **Investment per person employed -** (symbol: *inv*, thousands of euros)

The equation for this model is:

$$\mathrm{inv}_{it} = a_0 + \sum_{k=1}^{11} a_k \cdot \mathrm{skills}_{k,it} + \{\mathrm{control\ var iables}\} + \mathrm{ISE}_i + e_{it}$$

As control variables, we used *Gross domestic product /capita* (market prices), at the UE27 level, percentage change from previous period and *Gross domestic product /capita at market prices (index 2015 = 100)*, for country i and year = t.

Positive impact on *investment per person employed:*

General IT skills	0.032299
Professional IT skills	0.011686
Management skills	0.036787
Team working skills	0.010951
Office administration skills	0.045561
Technical, practical, or job-specific skills	0.010542
Numeracy and/or literacy skills	0.058969

(*Source*: authors' calculations)

Except for *Professional IT* and *Team working skills*, the other variables had significant coefficients at the 1% level. With *no positive impact value* are: Customer handling skills, Problem solving skills, Foreign language skills, Oral or written communication skills.

Most of the countries show specific negative fixed individual effects. Individual positive effects are reported only for Austria, Belgium, Denmark, Finland, France, Luxembourg, Netherlands, Norway, and Sweden. The LR test reject the hypothesis of redundant specific individual fixed effects (at a threshold < 10^{-4}).

CONCLUSIONS

We analysed the impact of the main skills needed for the development of the enterprise on economic activities using 8 indicators divided into 3 groups:

- ✓ three production indicators (*Turnover, Production Value, Value Added at factor cost*)
- ✓ three efficiency indicators (*Turnover per Person Employed, Apparent Labour Productivity*, *Gross Operating Surplus/Turnover*)
- ✓ two development indicators (*Growth Rate of Employment*, *Investment per person employed*)

Subject to the limited time dimension of the data series, we summarized the 8 models presented above and found that the impact of Continuing

Vocational Training (CVT) courses on the economic activity of enterprises was differentiated as follows (Table 1):

- ✓ Management skills was positively correlated with all economic indicators,
- ✓ Professional IT skills and Team working skills where correlated with almost all indicators
- ✓ General IT skills, Office administration skills, Foreign language skills, and Numeracy and/or literacy skills had a positive impact on most economic indicators, but not as much as the previous ones, while
- ✓ Customer handling skills, Technical, practical, or job-specific skills, and Oral or written communication skills have mainly been correlated with indicators of efficiency and development.

Impact of the skills needed for the development of enterprises economic activities

Table 1

Skills needed for the development of the enterprise	Impact to:		
	Production	Efficiency	Development
General IT skills	positive (1/3)	positive (2/3)	positive (2/2)
Professional IT skills	positive (3/3)	positive (2/3)	positive (2/2)
Management skills	positive (3/3)	positive (3/3)	positive (2/2)
Team working skills	positive (3/3)	positive (2/3)	positive (2/2)
Customer handling skills		positive (1/3)	positive (1/2)
Problem solving skills	positive (3/3)	positive (1/3)	
Office administration skills	positive (2/3)	positive (1/3)	positive (1/2)
Foreign language skills	positive (2/3)	positive (2/3)	positive (1/2)
Technical, practical, or job-specific skills		positive (2/3)	positive (1/2)
Oral or written communication skills		positive (1/3)	positive (1/2)
Numeracy and/or literacy skills	positive (2/3)	positive (2/3)	positive (1/2)

(*Source*: authors' calculations)

Note: (3/3) means that the mentioned skills had a positive impact for *3* among the *3* indicators analysed for the respective field (production, efficiency, development).

As median values (Table 2),

- ✓ the percentage of enterprises that needed some skills for development is greater than the percentage of enterprises providing the corresponding Continuing Vocational Training (CVT) courses, for almost all types of skills: *General IT skills* (+9.1 percentual points), *Management skills* (+3.6 pp)*, Team working skills* (+24.2 pp)*, Customer handling skills* (+18.1 pp)*, Problem-solving skills* (+9.7 pp)*, Foreign language skills* (+6.3 pp)*, Oral or written communication skills* (+3.2 pp) and *Numeracy and/or literacy skills* (+2.9 pp)
- ✓ only for *Professional IT skills* (-2 pp)*, Office administration skills* (-0.8)*, Technical, practical, or job-specific skills* (-12.95 pp), and *Other skills* (-13.8) the median value of provided courses was greater than the needed ones.

The coefficient of correlation between the percentage of enterprises that needed some skills for development and the share of enterprises that are providing the corresponding Continuing Vocational Training (CVT) courses is close to or greater than 90% for almost all types of skills (with 96% for *Oral or written communication skills* and 94.5% for *Numeracy and/or literacy skills*). The exceptions (with lower values of the coefficient of correlation) were *General IT skills* (78.6%) and *Technical, practical, or job-specific skills* (81.6%).

Correlation between main skills needed for the development of the company and main skills targeted by enterprises providing CVT courses

Table 2

	Median value of skills (as % of total enterprises):		Coefficient of correlation
	Needed	Targeted	
General IT skills	**25.95**	16.90	0.7862
Professional IT skills	12.55	**14.50**	0.8314
Management skills	**28.10**	24.50	0.9299
Team working skills	**47.00**	22.80	0.9269
Customer handling skills	**48.00**	29.90	0.8771
Problem solving skills	**29.30**	19.60	0.8931
Office administration skills	10.90	**11.70**	0.8953

	Median value of skills (as % of total enterprises):		Coefficient of correlation
	Needed	Targeted	
Foreign language skills	**14.70**	8.40	0.8853
Technical, practical, or job-specific skills	51.40	**64.35**	0.8160
Oral or written communication skills	**9.10**	5.90	0.9618
Numeracy and/or literacy skills	**4.30**	1.40	0.9446
Other skills	3.70	**17.50**	0.8904

(*Source*: authors' calculations)

In a study from 2017, Wiseman June and Emma Parry for the UK market (Wiseman & Parry, 2017), suggested that while some organizations experienced supply-side barriers (for instance, 19% of non-training organizations reported a lack of suitable courses), the main obstacles were the perceptions that the business's current skill levels were sufficient (89%) or that hiring could be a substitute for training (73%). Organizations that had trained had comparable excuses for not training more: While a significantly lesser percentage (21%) claimed they couldn't find the CVT courses they needed, 84% claimed they had completed all the training required or were able to find the requisite abilities. These figures on the barriers to training are consistent with both the 2010 CVTS survey results and the results of numerous other UK skills surveys (such as the national Employer Skills Surveys)

To achieve efficacy, support for training must be combined with assessment. Additionally, analysis can assist in identifying the most effective responses to widespread issues like population aging and digitalization as well as their implications for training policies, such as sustaining adequate levels of training investment in the face of an aging workforce, low participation rates among lower-skilled workers, and addressing the rapidly shifting skill requirements associated with digitalization.

To keep up with the rapid rate of technological change in the AI era, businesses must invest in CVT. For businesses to stay competitive, recruit and retain skilled workers, and get ready for the future of work, they must invest in employee training. Companies can guarantee that their staff has the skills required to compete in the future and take advantage of the potential given by AI by investing in CVT.

As policy implications, the paper can be useful for the targeted development of CVT activities.

Limitations of the analysis come from the short time series. Also, the data used refers to the enterprises that needed CVT courses, respectively that provided training, as a share (%) of all enterprises and not to the number of participants in CVT courses (the size of activities).

As a future development of the analysis, it could be envisaged to detail the study by NACE activities.

References

1. **Brunello, G., & Wruuck, P.** (2020). *Employer Provided Training in Europe: Determinants and Obstacles.* IZA Discussion Papers, No. 12981, Institute of Labor Economics (IZA), Bonn. Retrieved October 24, 2022, from https://www.iza.org/publications/dp/12981/employer-provided-training-in-europe-determinants-and-obstacles
2. **Croissant, Y., & Millo, G.** (2019). *Panel Data Econometrics with R.* West Sussex, UK: Wiley.
3. **European Commission.** (2022). *Vocational education and training. Skills for today and for the future.* Luxembourg: Publications Office of the European Union. doi:10.2767/811982
4. **Eurostat.** (2022, October 14). Continuing vocational training in enterprises. Luxembourg. Retrieved November 5, 2022, from Eurostat metadata: https://ec.europa.eu/eurostat/cache/metadata/en/trng_cvt_esms.htm
5. **Eurostat.** (2022). *Statistics Explained.* Retrieved November 5, 2022, from Continuing Vocational Training Survey (CVTS) methodology: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Continuing_Vocational_Training_Survey_(CVTS)_methodology
6. **Gabor, M. R., Blaga, P., & Matis, C.** (2019, June 17). Supporting Employability by a Skills Assessment Innovative Tool—Sustainable Transnational Insights from Employers. *Sustainability, 11*(12), 3360. doi:10.3390/su11123360
7. **Jula, D., & Jula, N.-M.** (2019). *Econometria seriilor de timp (Time series econometrics).* București: Mustang.
8. **Konings, J., & Vanormelingen, S.** (2015, May 01). The Impact of Training on Productivity and Wages: Firm-Level Evidence. *Review of Economics and Statistics, 97*(2), 485-497. doi:10.1162/rest_a_00460
9. **Kuckulenz, A**. (2007). *Studies on Continuing Vocational Training in Germany: An Empirical Assessment.* Heidelberg: Physica-Verlang.
10. **Makkonen, T., & Lin, B.** (2012). Continuing vocational training and innovation in Europe. *International Journal of Innovation and Learning, 11*(4), 325-338. doi:10.1504/IJIL.2012.047135
11. **Mara, L.-C.** (2018). *Innovation in the continuing vocational education and training (CVET) government-run programme in Catalonia.* Doctoral Thesis, Univerisitat Rovira i Virgili, Department of Business Management, Reus. Retrieved September 15, 2022, from https://www.tdx.cat/bitstream/handle/10803/662731/Thesis.pdf?sequence=3
12. **Markowitsch, J., & Bjørnåvold, J.** (2022, August 30). Scenarios for vocational education and training in Europe in the 21st century. *Hungarian Educational Research Journal, 12*(3), 235-247. doi:10.1556/063.2021.00116
13. **Milmeister, P., Rastoder, M., & Houssemand, C.** (2022, May 30). Mechanisms of Participation in Vocational Education and Training in Europe. *Frontiers in psychology, 13*, 12p. doi:10.3389/fpsyg.2022.842307

14. **Morris, T. H., & Rohs, M.** (2022, September 21). Examining barriers to participation in further and continuing education in Germany: Why a regional perspective is (still) important. *International Review of Education, 68*, 551-577. doi:10.1007/s11159-022-09968-4
15. **Pouliakas, K.** (2018). *Automation risk in the EU labour market. A skill-needs approach.* European Centre for the Development of Vocational Training (Cedefop). Retrieved September 15, 2022, from https://www.cedefop.europa.eu/files/automation_risk_in_the_eu_labour_market.pdf
16. **Wiseman, J., & Parry, E.** (2017). *Continuing vocational training survey: CVTS 5.* DfE research report, no. DFE-RR754A, UK Gouvernement, Department for Education, Manchester, England. Retrieved from https://dera.ioe.ac.uk/30307/1/Continuing_vocational_training_survey-main_report.pdf
17. **Zwick, T.** (2005). Continuing Vocational Training Forms and Establishment Productivity in Germany. *German Economic Review, 6*(2), 155-184.

Annex

1. Panel data estimations for production indicators

Variables	Coefficients in equation of:		
	Turnover	Production value	Value added
C	1005124.52 (40379.94)	658425.69 (19623.22)	233232.8 (10727.31)
General IT skills	-2494.5 (470.28)	163.86 (634.32)	-715.5 (162.88)
Professional IT skills	2194.19 (502.67)	102.03 (276.25)	930.17 (394.37)
Management skills	1110.35 (322.79)	1865.37 (217.97)	306.44 (117.42)
Team working skills	740.71 (359.49)	2045.98 (178.97)	45.26 (110.39)
Customer handling skills	-5291.28 (442.7)	-2878.83 (458.29)	-604.23 (142.12)
Problem solving skills	1335.63 (417.9)	330.23 (194.07)	773.61 (72.02)
Office administration skills	3965.72 (686.37)	-740.72 (442.63)	235.42 (143.73)
Foreign language skills	4681.33 (1185.5)	-347.99 (224.83)	290.92 (260.08)
Technical, practical, or job-specific skills	-5163.95 (534.3)	-845.86 (354.47)	-1090.9 (340.39)
Oral or written communication skills	-3483.74 (956.62)	-1719.44 (470.31)	-1914.16 (193.67)
Numeracy and/or literacy skills	-298.75 (347.73)	897.75 (467.3)	1224.89 (311.37)

GDP/capita at market prices (index 2015 = 100)		3266.22 (77.05)	–	593.21 (71.16)
GDP/capita at UE27 level, % change on the previous period		10208.69 (2476.44)	11517.81 (1986.24)	–
Fixed Effects (Cross)				
Belgium		148563.71	85524.76	16086.53
Bulgaria		-731514.95	-535874.01	-187197.62
Czechia		-508057.40	-249400.13	-117600.35
Denmark		-405595.00	-216815.40	-70550.78
Germany		5372012.29	3673524.59	1443916.32
Estonia		-994315.84	-604316.18	-227531.39
Greece		-715575.37	-429522.17	-173697.73
Spain		707222.39	537450.12	217949.19
France		2639672.91	1862365.20	705244.37
Croatia		-1029647.91	-602692.87	-215402.28
Italy		1712334.07	1486888.09	427500.98
Cyprus		-920706.00	-576745.69	-218437.18
Lithuania		-958191.36	-594089.75	-224869.86
Luxembourg		-796146.31	-525062.61	-182214.74
Hungary		-901703.51	-488986.47	-209958.37
Malta		-911355.52	-614440.55	-219732.76
Netherlands		581849.13	358767.24	139461.14
Austria		-237608.96	-146785.59	-23332.43
Poland		-43787.06	76827.76	-17740.97
Portugal		-760142.53	-453164.06	-168332.15
Romania		-692350.36	-475173.97	-172280.83
Slovenia		-917018.43	-567090.00	-212158.45
Slovakia		-817945.85	-439525.01	-179448.09
Finland		-549181.35	-357491.72	-132378.78
Sweden		-132758.38	-43281.90	7513.08
Norway		-187953.69	-160889.68	-4806.83
United Kingdom		3074851.924	–	–
R^2		0.9998	0.9987	0.9984
Redundant Fixed Effects Tests	Statistic	691.8582	458.7465	273.3240
	prob.	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$

(Under coefficients, in round bracket, is the standard error)

(*Source*: authors' own calculations)

2. Panel data estimations for efficiency indicators and for development indicators

Variables	Coefficients in equation of:				
	Turnover per person employed	Apparent labour productivity	Gross Operating Rate	Growth rate of employment	Investment per person employed
C	95.3084 (3.252)	51.3516 (1.262)	51.3516 (1.262)	-3.4618 (0.539)	10.7630 (0.697)
General IT skills	0.7161 (0.061)	0.0433 (0.046)	0.0433 (0.046)	0.0274 (0.013)	0.0323 (0.008)
Professional IT skills	-0.5042 (0.086)	0.1063 (0.067)	0.1063 (0.067)	0.0090 (0.002)	0.0117 (0.020)
Management skills	0.3295 (0.074)	0.0603 (0.021)	0.0603 (0.021)	0.0121 (0.013)	0.0368 (0.012)
Team working skills	-0.6488 (0.106)	0.0345 (0.022)	0.0345 (0.022)	0.0012 (0.028)	0.0110 (0.021)
Customer handling skills	0.7241 (0.126)	-0.1164 (0.029)	-0.1164 (0.029)	0.1170 (0.021)	-0.0543 (0.015)
Problem solving skills	0.0005 (0.189)	-0.1025 (0.029)	-0.1025 (0.029)	-0.1019 (0.026)	-0.0705 (0.014)
Office administration skills	-0.4973 (0.145)	0.0301 (0.058)	0.0301 (0.058)	-0.2191 (0.034)	0.0456 (0.018)
Foreign language skills	0.134 (0.036)	-0.1754 (0.029)	-0.1754 (0.029)	0.0734 (0.046)	-0.0486 (0.010)
Technical, practical, or job-specific skills	-0.0814 (0.047)	0.0307 (0.021)	0.0307 (0.021)	-0.0373 (0.019)	0.0105 (0.006)
Oral or written communication skills	0.2933 (0.100)	-0.1725 (0.033)	-0.1725 (0.033)	0.2263 (0.028)	-0.0432 (0.008)
Numeracy and/or literacy skills	-0.7619 (0.100)	0.1474 (0.026)	0.1474 (0.026)	-0.0559 (0.030)	0.0590 (0.008)
GDP/capita at UE27 level, % change on the previous period	–	–		1.7876 (0.076)	0.3555 (0.052)
GDP/capita % change on the previous period	0.7835 (0.020)	–		–	-0.1643 (0.030)
Fixed Effects (Cross)					
Belgium	158.1659	31.0849	-0.1066	0.4362	10.7522
Bulgaria	-103.6557	-33.9293	-0.8792	-2.3093	-4.9348
Czechia	-58.3889	-17.6106	-0.5074	-2.5754	-2.3378
Denmark	95.0410	38.9049	2.1261	-1.5048	4.6851
Germany	29.5643	11.7190	-0.2878	0.0088	-0.8844
Estonia	-51.3008	-17.6059	-2.2008	0.9399	-0.8407

Greece		-102.1451	-23.0486	-2.2283	-4.6101	-4.5725
Spain		-28.1544	-5.8969	-0.7721	2.1994	-3.5322
France		64.0134	9.9085	-4.6780	4.6765	1.7802
Croatia		-114.3600	-25.9567	1.3929	3.1278	-5.8065
Italy		6.0142	-1.1610	0.0058	0.7816	-3.3787
Cyprus		-74.5790	-8.3455	1.6201	-2.1976	-3.0940
Lithuania		-117.0091	-26.6578	-0.2701	0.1759	-3.3916
Luxembourg		375.2745	42.2079	-4.3327	2.7848	4.5622
Hungary		-64.8213	-24.2047	-0.7788	1.6482	-1.6465
Malta		-49.0325	-9.3949	4.3047	1.2495	-2.9998
Netherlands		67.8103	18.2893	0.6872	0.9937	0.8633
Austria		63.0646	23.6486	0.0967	-2.7985	4.4826
Poland		-81.9271	-24.2134	0.7356	2.0753	-4.0675
Portugal		-75.3497	-24.6944	-1.9456	1.1452	-4.6541
Romania		-107.1250	-30.0479	0.4521	0.8713	-0.9450
Slovenia		-49.7504	-9.2476	-0.9130	0.4411	-1.5063
Slovakia		-80.1295	-19.1693	-0.6435	0.8157	-1.5092
Finland		55.3033	17.6255	-1.0927	-0.6299	1.2985
Sweden		61.5632	23.9774	-0.5972	-1.6770	5.0485
Norway		157.9664	74.9980	7.3119	-6.0683	17.6459
United Kingdom		35.9209	13.2306	5.2512	–	-1.5253
R^2		0.998	0.998	0.965	0.976	0.998
Redundant Fixed Effects Tests	Stat.	387.019	247.463	35.396	11.741	334.526
	prob	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$

(Under coefficients, in round bracket, is the standard error)

(*Source*: authors' own calculations)

Geostatistical Coordinates Regarding the Suicidal Phenomenon in Romania between 2011-2020

Vicențiu-Robert GABOR[1] (Vicentiu.Gabor@iasi.insse.ro)
"Al. I. Cuza" University of Iasi, Iasi County Directorate of Statistics

Octavian GROZA (grozaoctavian@uaic.ro)
"Al. I. Cuza" University of Iasi

Ciprian IFTIMOAEI (Ciprian.Iftimoaei@iasi.insse.ro)
National Institute of Statistics, "Al. I. Cuza" University of Iasi

ABSTRACT

As an absolute first in the Romanian scientific space, being conducted on a local scale, the research that led to the publication of this article aims to provide expert assistance to public authorities that have to manage the dynamics and social impact of suicide and to psychologist and other social scientist who study the various other dimensions of the suicide phenomenon. The general hypothesis of the present research is the following: although suicide seems to be an a-spatial phenomenon, it can have resilient, persistent evolutions that lead over time to the formation of relatively stable geospatial (geographical) structures. To study the phenomenon of suicide in Romania, we have at our disposal data provided by the National Institute of Forensic Medicine (NIFM) and the National Institute of Statistics (NIS). The data from NIFM are publicly available only for the period 2003-2019, at the national level and in the county profile. In our research, we used data from the NIS, for the period 2011-2020, at a much finer scale (Local Area Unit2 - LAU2). The data collected is processed using geostatistical methods that are accessible, robust and capable of identifying and visually representing (cartographically) the sensitive spaces where the suicide phenomenon takes place. For the quantitative and structural highlighting of suicides, we used cross tabulation analysis, box-plot analysis, neighbours order smoothing autocorrelation analysis with PhilCarto. The suicide rate calculated for the period 2011-2020 was 10.75 suicides per 100,000 inhabitants; in absolute values, for the studied period, 21,252 suicides were registered. Regarding this situation, the legitimate question arises: how many suicide attempts are behind of total number of 21,252 suicides? The conclusions of this article can be used in the public policymaking process aimed at preventing suicidal acts.

Keywords*: suicide, suicide distribution, mental health, geostatistics, cartography*
JEL Classification*: C18, I12, J19*

[1] Corresponding author

1. INTRODUCTION

Suicide or death committed by the victims themselves has intrigued, fascinated or aroused people's interest throughout the ages. Philosophers, theologians, historians, psychiatrists, psychologists, sociologists, demographers, and geographers have investigated this phenomenon with their own theoretical and methodological tools, in an attempt to discover the causes of suicide, the factors influencing it and the consequences for communities and societies everywhere.

The decision to give up one's own life remains an elusive mystery. Sociodemographic profiles of those who have chosen to end their lives have been drawn up based on statistical data provided by various institutions responsible for official statistics and public health. The social, economic and cultural causes of suicide attempts and fatal suicide were also investigated. Using qualitative (interviews, focus groups, life stories) and quantitative (questionnaires applied to people with at least one suicide attempt) research methods, typologies of suicidal behaviour were developed and described. Demographers and geographers produced maps of the suicide phenomenon (incidence of suicide per 100,000 inhabitants) at global, regional, national and local levels, based on which comparisons were made. Sociologists and social workers are increasingly tempted to formulate public policy proposals to prevent and combat suicidal behaviour.

Political or religious leaders, public authorities, notables of the communities or ordinary people have dealt with the phenomenon of suicide differently throughout history. Suicide has been condemned, tolerated or glorified. Writers, poets and painters have each been sensitised in their own way by the contexts and life stories woven around prominent personalities who resorted to suicide as the ultimate solution to the existential problems they faced. Literature and art collections in museums bear witness to artists' preoccupation with the issue of suicide.

According to Georges Minois (2002), historian of suicide, the ancient world was quite tolerant of suicide (voluntary killing). In some ancient cities (Athens, Sparta, Thebes) there were certain penalties against the bodies of those who voluntarily gave up their lives. Greek and Roman antiquity records a number of high-profile suicides: Aristodemus (suicide for remorse), Cleomenes (suicide for honour), Pythagoras (suicide for a religious idea), Hippolytus (suicide for chastity), Zeno, Diogenes, Epicurus, (philosophical suicides), the death or 'questionable suicide' of Socrates, the suicide of the Romans Cato, Seneca and others.

Ancient history also records a famous case of honor suicide (rape) - Lucretia (6th century BC) during the period of Royal Rome (753-509 BC) -

resulting in mass indignation and a popular revolt with political consequences: the transition of the Roman state from monarchy to republic. The tragedy of this historical legend impressed painters and sculptors such as Titian, Damià Campeny and others who immortalised the suicide scene in their works.

The world of the Middle Ages, under the influence of Christian doctrine, harshly sanctioned suicide, albeit not in the same way for all. In the medieval period, there were differences in the way suicide was reported, depending on whether one belonged to one social class or another. Suicides of peasants and artisans were condemned, while those of the nobility were treated with some leniency or passed over in silence. The *Renaissance* and the *Age of Enlightenment* were periods of emancipation, including on attitudes towards suicide. Voluntary death was seen by some thinkers as a manifestation of individual freedom. In his *Essay on Suicide*, the philosopher David Hume (1783) believes that if suicide is not a crime, then we can accept those who courageously give up life when it becomes a burden.

Nowadays, there is more and more talk about assisted suicide or euthanasia (passive or active), practices that are legalized in some countries of the world (e.g. Switzerland, Belgium, Netherlands, Luxembourg, Canada, USA, South Korea, etc.). In Romania, such practices are forbidden and punishable by criminal law. Assisted suicide is most often requested by people suffering from serious, degenerative, incurable illnesses which cause suffering to the person concerned and their relatives. There are also situations where both members of the couple/family have resorted to suicide as a sign of affection and solidarity towards the physical and/or mental suffering of one of them.

Another problem of the contemporary world is suicidal terrorism, which is used by groups driven by fundamentalist ideologies and beliefs that propose the suppression of entire social groups that share other values, principles and norms of life. The attacks on the World Trade Center in the US on 11 September 2011 are a grim reminder of suicidal terrorism, which has reshaped international relations and the global security environment. For Mia Bloom (2002), the aim of terrorism is to demoralise, to induce panic and fear among the civilian population in order to change the course of domestic and international politics of some states. She adds that terrorist organisations also use suicide terrorist attacks to motivate and mobilise their supporters from whom future terrorists are selected. They are indoctrinated to give up their own lives and trained militarily to kill people and destroy critical infrastructure with the promise of a reward in the afterlife.

Thousands of books have been written on the subject of suicide without fully clarifying what makes people commit suicide or stay alive in similar circumstances. Suicide is present in all eras and societies, with certain variations in time and space, influenced by psychological, social, cultural, economic, political and environmental factors. Suicide theories and empirical research data are becoming more and more pervasive over time. Commenting on this, Cristina E. Brădățan of Texas Tech University says: „People, as well as societies, change over time, and what was seen as consistent, meaningful or true 100 years ago is probably seen in a very different light today. Can old theories be used for anything other than studying them over time in a sociology history course? Are they still informative? How useful are some 19th-century theories to explain variations in suicide now? Can the concepts and explanations used a century ago make any sense of the statistics of our time?" (Brădățan, 2007).

2. LITERATURE REVIEW

Scientists have reported differently on the phenomenon of suicide. Psychiatrists believe that suicide is caused by some form of mental alienation, insanity, anxiety or depression. Psychologists consider suicide to be self-destructive behaviour, a form of individual deviance, a disturbance of the self-preservation instinct of the human person. Sociologists focus on the social causes of suicide. For lawyers specialising in criminal law, suicide is of interest if it is caused (rape, assault) or facilitated by criminally sanctioned conduct (abuse, harassment, grievance). Demographers and geographers are particularly concerned with the incidence of suicide, the spatial and seasonal nature of suicide.

There is a certain controversy between psychiatrists and psychologists, on the one hand, who attribute suicide to causes related to the physical and psychological constitution of the individual, and sociologists, on the other hand, who believe that suicide is explained by social factors, independent of the biological and psychological characteristics of the person who resorts to this self-destructive act. In this respect, Jean Bachler (1975) considers that voluntary death should not be explained by statistics, but on the basis of individual cases of suicide (*apud* Minois 2002). The founder of psychoanalysis, Sigmund Freud (1905), considered that suicide is a turning of aggression against the self (*ibid*.).

Psychologists believe that suicide is a complex process that includes several stages: (1) the emergence of the thought of suicide (communicated verbally or non-verbally to family, group of friends, etc.), (2) suicide planning (context, means, impact-message), (3) suicidal behaviour (attempted or completed suicide). Although suicidal thoughts are the first concrete step towards self-destructive behaviour, at this stage the person may still be receptive to therapy/intervention/prevention measures for suicidal acts (Yip et al., 2012).

Psychologist G. Havârneanu (2014) includes suicide in the sphere of dysfunctional behaviours generated by the individual's mental health problems (emotional imbalances, anxiety, depression), advocating for psychological interventions (counselling, psychotherapy, social support) offered within individual or community assistance programmes. In addition to all these measures to prevent suicidal acts or suicide attempts, psychiatrists propose medication treatment, including hospitalisation for people at risk of suicide.

Without denying the influence of economic and socio-cultural factors, psychologists believe that the motivation of suicides must be sought in the personality structure of the individual and in the psychopathology of his life. Psychiatrists provide explanations within the sphere of neuropsychological pathologies (e.g. neurotransmitter depletion) associated with illnesses such as alienation and depression.

For the sociological literature on suicide, Émile Durkheim's *On Suicide* (*ed. fr.*1897) [rom. 2007] remains representative, topical and controversial among researchers. The French sociologist defines suicide as any case of death resulting directly or indirectly from a positive or negative act committed by the victim himself, who knows the result of his self-destructive act (Durkheim, 2007). In disagreement with psychologizing approaches, Durkheim states that suicide is not a result of individual problems (dysfunctions) related to the biological and/or psychological construction of the person, such as mental alienation. In his view, suicide is nothing more than the result of a 'collective tendency', with social rather than individual causes. In this sense, sociology aims to study the social and extra-social causes of suicide.

The epistemological and methodological imperative of Durkheim is to avoid a similarity between suicide and mental illness: "...each social group has a specific tendency towards suicide which neither the organic-psychic consciousness of individuals nor the nature of the physical environment can explain. It follows, by elimination, that this tendency depends on social causes and constitutes, by itself, a collective phenomenon. Even certain facts we have examined, such as geographical and seasonal variations in suicide, have led us to these conclusions" (*ibid*., p. 133)..

Using quantitative (statistical) methodology, Durkheim elaborates the following typology of the suicide phenomenon: (1) selfish suicide (action of religious, cultural, family, political factors); (2) altruistic suicide (when the individual is too integrated into society); anomic suicide (socio-economic causes). On the basis of statistical data, Durkheim notes that the incidence of suicide is higher among Protestants than among Catholics, and higher among Catholics than among Jews; men are more likely to commit suicide than women; in some (primitive) societies, suicide of old or sick men, suicide of women when their husbands die, suicide of servants when their masters die, etc.; social anomie caused by sudden (economic crises) or slower changes (industrialisation, urbanisation) leads to an increase in the incidence of suicide.

Periods of crisis are complementary to changes in socio-economic development and, by implication, to changes in society's standard of living and individual well-being. The incidence of suicide (number of suicides per 100 000 inhabitants) is also an indicator of the social well-being and health of the population in a region, country or locality. However, it should be noted that there is also empirical research focusing on the link between periods of economic crisis and deterioration in social well-being, including health, which shows that there is no strong statistical link between poverty (poverty or social exclusion rate) and suicide (suicide rate). Poverty may have a conjunctural influence on the incidence of suicide in a territorial-administrative unit, but it is not determinant.

There is an association between the health status of a population and its suicidal tendencies, with variations from region to region. This conjectural link can be tested by correlating mortality rates with suicide rates and their territoriality. Based on official statistics, a positive correlation between indicators of population health and suicide rates can be shown. Exceptions can be explained by other social, economic or cultural factors: employment, unemployment, poverty, alcohol and/or drug use, domestic violence, secularisation, etc. In general, it is appreciated that sudden social changes or periods of crisis make people more vulnerable, and some people are more prone to suicidal (self-destructive) behaviour than others.

Sociologist S.M. Rădulescu (2014a, 2014b, 2015) stresses that the suicidal phenomenon is generated by causes and factors that are interdependent, acting correlatively or successively on the individual's behaviour. However, periods of crisis (economic, pandemic, geopolitical) do not in themselves determine the incidence of suicide, but through the action of factors (intermediate variables) such as unemployment, falling income, alcohol or drug consumption, deterioration in the physical and/or mental health of individuals.

Regarding the self-destructive methods used by suicides in Romania, according to data provided by the National Institute of Forensic Medicine, more than 72% of suicides choose to end their lives by hanging, followed by voluntary poisoning and precipitation ("jumping from heights"), which are used by about 8-9% of suicides (Rădulescu, 2014). These methods are also used, but in a higher proportion than in Romania, by suicides in other European countries, such as England and the Nordic countries (who choose poisoning as the main method after hanging), Luxembourg, Malta and Spain (who prefer precipitation immediately after hanging) (Ajdacic-Gross, 2008; Rădulescu, 2014).

In a sociological research entitled "In Durkheim's footsteps" (2017), based on quantitative methods (data-supported analysis, online questionnaire) and qualitative methods (content analysis, life story interview), Emanuel Adrian Sârbu aims to draw up a "map of suicide in post-communist Romania". Although the title of the book creates some expectations of a geospatial analysis of the phenomenon of suicide in Romania, based on methods that are part of the arsenal of spatial statistics and modern cartography, the reader is presented with an update of the descriptive statistical analysis of suicide based on data provided by the National Institute of Legal Medicine and the National Institute of Public Health, complemented by the analysis of data collected through the online questionnaire to people who have at least one suicide attempt, as well as the qualitative analysis of the testimonies collected online on specialized platforms dedicated to preventing and combating suicide attempts in Romania. This work is particularly valuable for the qualitative information it provides for empirical research on the suicide phenomenon.

In a theoretical study dedicated to suicide in new interpretative contexts, specific to Romanian society, sociologist D. Stan (2021) of the University "Al.I. Cuza" of Iași provides the following conclusions: (1) the right of ownership of one's own body should not be metamorphosed into the individual's right to commit suicide; (2) even if suicide were a normal fact, this phenomenon belongs to the pathological aspects of everyday life; (3) social actions against suicide can only be ante factum. The sociologist believes that individuals, groups, communities and society as a whole must be informed, educated and made aware of the causes that lead to suicide and the psychosocial, cultural and economic factors that will influence the dynamics of this social phenomenon: "the act of suicide can only be traumatic and condemnable for society".

The phenomenon of suicide is increasingly perceived as a societal problem and is coming to the attention of public authorities with increasing intensity. The World Health Organization launched a global suicide prevention initiative in 1999 (UN-WHO, 2000). Also, at the United Nations, the World Health Assembly in 2013 adopted a first-ever action plan for the World Health Organization on mental health, which called for a 10% reduction in suicides by 2020 (UN-WHA, 2013). A comprehensive prevention policy has long been in place in the European Union and is closely followed by Member States. In 2009, the European Parliament adopted Resolution 2008 on mental health, which draws attention to the need for early detection of depression, a common cause of suicide. This resolution was the basis for various updates to the

Romanian law on mental health (no. 487/11.07.2002) and the introduction of suicide prevention among the objectives of the National Program of Mental Health and Prophylaxis in Psychiatric Pathology, adopted by Romanian Parliament in 2015. (Dumitru et al., 2019). Following the development of this legislation on suicide prevention, more and more legislation is emerging, not only for people in need of psychiatric care, but also for people in many other areas of work (education, military, law enforcement, social work, etc.).

This state of affairs explains the growing interest in the phenomenon of suicide shown by Romanian society, through the increase in the number of associations and NGOs, urban prevention centres, through the implementation of hotlines (Sârbu, 2017) or through the increased frequency of presentation of cases in traditional or online media (Rădulescu, 2015). The scientific community reacted accordingly. An analysis of the reports of the relevant institutions (National Institute of Criminology, National Institute of Forensic Medicine, National Institute of Statistics) and of the scientific literature in various fields shows that suicides are systematically and thoroughly analysed, providing a solid basis for action by the actors called upon to manage and control the dynamics of suicides.

The scientific community reacted accordingly. An analysis of the reports of the relevant institutions (National Institute of Criminology, National Institute of Forensic Medicine, National Institute of Statistics) and of the scientific literature in various fields shows that suicides are systematically and thoroughly analysed, providing a solid basis for action by the actors called upon to manage and control the dynamics of suicides.

Place matter… Where things happen is as important as *when*, *how* and *why* they happen. The general hypothesis that structures our research is that an apparently a-spatial phenomenon such as suicide can have resilient, perennial developments that lead over time to its territorialisation and, therefore, to the emergence of relatively stable geographical structures that, once revealed, can serve as a framework for effective public policy intervention.

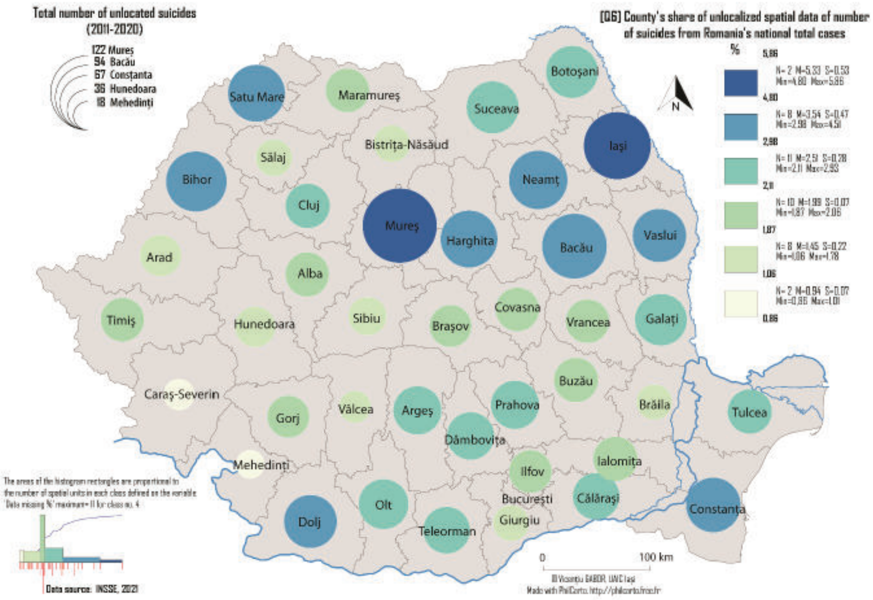
3. CRITICAL PRESENTATION OF DATA

In Romania, official statistics on suicide are provided by two institutions, namely the National Institute of Statistics (NIS) and the Institute of Forensic Medicine "Mina Minovici" from Bucharest. Although discrepancies between the datasets and some differences from international standards have been noted (Rădulescu, 2014a), the existing information can provide a fairly clear picture of the magnitude of the phenomenon and its structural features. Data from the Institute of Forensic Medicine "Mina Minovici" are publicly available for the period 2003-2019 and only at the Romanian county level. This article has used these data to extend backwards the interval that constitutes the researched period, i.e. 2011-2020. Such a perspective contributes to a more precise framing of the phenomenon and allows a more efficient approach to specific aspects, such as the seasonality of suicides.

The main dataset used in this study was provided by the National Institute of Statistics, for the period 2011-2020, at a finer scale - that of the 3,181 TAUs (Territorial Administrative Units)[1], called in the European taxonomy LAU2 (Local Area Unit), which corresponds to the communes and cities of Romania. The data were not provided for localities where fewer than 3 cases were recorded. For cartographic and geostatistical reasons, however, we assigned a value of 1 for localities with unavailable data but where suicides occurred. To summarise, only 9.8% of cases could not be geographically located at the local authority level; the situation by county is shown in Figure 1.

Spatial distribution of the statistical data deficit for the counties' LAU2

Figure 1


Total number of unlocated suicides (2011-2020)
122 Mureş
94 Bacău
67 Constanta
36 Hunedoara
18 Mehedinţi
[Q6] County's share of unlocalized spatial data of number of suicides from Romania's national total cases
%
5,86
N= 2 M=5.33 S=0.53 Min=4.80 Max=5.86
4,80
N= 8 M=3.54 S=0.47 Min=2.98 Max=4.51
2,98
N= 11 M=2.51 S=0.28 Min=2.11 Max=2.93
2,11
N= 10 M=1.99 S=0.07 Min=1.87 Max=2.06
1,87
N= 8 M=1.45 S=0.22 Min=1.06 Max=1.78
1,06
N= 2 M=0.94 S=0.07 Min=0.86 Max=1.01
0,86
N
Satu Mare
Maramureş
Botoşani
Suceava
Iaşi
Bistriţa-Năsăud
Sălaj
Bihor
Cluj
Neamţ
Mureş
Harghita
Bacău
Vaslui
Arad
Alba
Timiş
Hunedoara
Sibiu
Braşov
Covasna
Vrancea
Galaţi
Caraş-Severin
Gorj
Vâlcea
Argeş
Prahova
Buzău
Brăila
Tulcea
Dâmboviţa
Mehedinţi
Ilfov
Ialomiţa
Bucureşti
Călăraşi
Giurgiu
Constanta
Dolj
Olt
Teleorman
0
100 km
The areas of the histogram rectangles are proportional to the number of spatial units in each class defined on the variable 'Data missing %' maximum= 11 for class no. 4

Data source: INSSE, 2021

© Vicenţiu BABOR, UAIC Iaşi
Made with PhilCarto http://philcarto.free.fr

[1] In Romanian: UAT - Unităţi Teritoriale Administrative. Including the 6 sectors of the capital, the official number is 3187. In this article we have not analysed the sectors of Bucharest.

This apparent inaccuracy also concerns the data on the number of inhabitants on the basis of which the suicide rates were calculated, data which are the result of annual estimates made by INS specialists and which may be influenced by random processes, such as the presence of Moldovan citizens registered in some of the Moldovan counties. We say *apparent inaccuracy* because we can sustain the adjective with two observations. The first is that the greatest "losses" of cases are precisely in the counties with the highest intensities of the phenomenon, and this normal statistical effect cannot structurally affect the analysis. The second is based on the idea that this work aims to capture the spatial structures of the phenomenon, which are less sensitive to small oscillations in the available information. This is, in fact, also the reason why we will not analyse the purely statistical dimensions of the phenomenon except to the extent that this will bring more clarity to the logical approach. Therefore, classical statistical analyses will not be the direct basis for the hypotheses, statements and conclusions of our research. Thanks to the level of aggregation at the county level, it is possible to use the full information on the number of suicide cases.

In order to have the necessary global benchmarks to frame the suicide phenomenon at national level, we used statistical data provided by *Eurostat*, *Global Burden of Disease Collaborative Network*[1], with data available up to 2019, presented in detail in the paper *Suicide*, accessible on the interactive research platform *Our World in Data* of the University of Oxford (Ritchie et al., 2015).

4. RESEARCH METHODOLOGY

Given that the main purpose of the research is to demonstrate the existence of zonal structures of the phenomenon analysed, the first concern was to limit as much as possible the transfer of the inconsistency of the statistical series to the cartographic representation. The obvious choice was to use the classical *suicide rate* (an indicator that standardises the incidence of suicide and eliminates to some extent the effect of size[2]), calculated according to the formula:

$$SR = \frac{NS_i}{NI_i} * 100.000$$

where:

SR = suicide rate;

NS_i = suicides number in spatial unit i;

NIi = number of inhabitants in spatial unit i.

[1] https://vizhub.healthdata.org/gbd-results

[2] Effect that contributes significantly to increasing the likelihood that the incidence of a phenomenon is higher in areas with larger populations.

Also, in order to mitigate the effect of size, we have carried out cross-tabular analyses, allowing for the most accurate reading of the quantities and interacting structures.

The methodological architecture of the research combines a number of analytical methods, based on descriptive statistics (*box-plots*) and classical spatial analysis (neighbour's order smoothing, autocorrelation analysis). The *box-plot* representation is a simplified graphical method of visualising the distribution of a set of values, by presenting a summary consisting of five specific values of the distribution: minimum value, Q1 (first quartile), median (i.e. Q2), Q3 (third quartile), average, maximum value and outliers.

The box, described by the interquartile range, allows the rapid identification of the type of distribution of values over the range and allows the construction of hypotheses regarding the methods necessary to map as accurately as possible the phenomenon under analysis.

One of the most difficult problems was to (carto)graphically represent the phenomenon. Rossen and Khan (2016) summarize very well the problems raised by the spatial analysis of this phenomenon: it is characterized by rare events, and their frequency becomes lower when dealing with lower population densities. This leads to discontinuous, inconsistent and highly turbulent statistical series. Analysing the USA at the county scale, the authors propose several solutions for carrying out geographic research (e.g. using 5-year averages) and focus methodologically on a spatial information aggregation model based on multi-year averages and a Bayesian smoothing model.

The annual geostatistical analysis of suicide rates in Romania clearly demonstrates the randomness of the phenomenon, with the Moran and Geary autocorrelation coefficients stubbornly remaining very close to zero. Taking the 10-year multi-year average changes the situation radically, as the coefficients demonstrate the existence of autocorrelation, even if their values decrease abruptly beyond third-order neighbours (Table 1). The analysis of the dynamics of these coefficients justifies the use of a smoothing technique in order to outline the spatial structures more clearly.

Autocorrelation coefficients by neighbour's order

Table 1

Coefficients of spatial autocorrelation					
Average suicide rate 2011-2020			**Smoothed average suicide rate 2011-2020**		
Neighbor's order	MORAN coeff.	1-GEARY coeff.	Neighbor's order	MORAN coeff.	1-GEARY coeff.
1	0,25	0,29	1	0,94	0,95
2	0,18	0,21	2	0,86	0,87
3	0,15	0,17	3	0,74	0,75
4	0,11	0,12	4	0,58	0,58
5	0,08	0,08	5	0,44	0,41
6	0,05	0,04	6	0,30	0,26
7	0,02	0,01	7	0,18	0,12
8	0,01	-0,01	8	0,09	0,01
9	0	-0,04	9	0,01	-0,08
10	-0,01	-0,05	10	-0,04	-0,16

In contrast to the model proposed by the American researchers, in our study we chose a different data listing option: neighbourhood smoothing. (*lissage par voisinage*, cf. Waniez, 2023a). The choice is justified by the fine level at which we carry out the analysis, that of LAU2, which is also the level of primary information collection. The principle of this type of smoothing is based on the proximity of spatial units, with two units having at least one line segment as a common boundary being considered neighbours.

The immediate neighbours of unit *i* are neighbours of order 1, the immediate neighbours of those of order 1 are neighbours of order 2 for unit *i*, etc. The calculations are performed using the mapping software *PhilCarto*[1]. The calculation consists of identifying neighbours from order 1 to order 10 for all units *i* (in this case Romania's 3,181 TAU), then calculating a smoothed value for each of these units. The smoothed value is the weighted average of the unit in question and its neighbours of different orders. The weighting is 1 for unit *i*, 1/2 for its neighbours of order 1, 1/3 for those of order 2, etc. This gradual weighting - according to distance, after all - makes it possible to group the elementary units into larger spatial structures, which preserve the initial role of the elementary unit in the territorial whole, but illustrates it better by articulating it with that of its neighbours. The fact that the basic structure

[1] Free-ware, available at http://philcarto.free.fr

from which the smoothing was performed is preserved is also demonstrated by the dynamics of the Moran and Geary autocorrelation coefficients for the smoothed variable (Table 1), dynamics parallel to those calculated from the actual values.

The inconsistency of the data series collected at LAU2 level also poses serious problems for descriptive statistics, with obvious results on cartographic representations. In order to reduce as much as possible, the risk of misinterpretation, which is very possible with analytical mapping, we have used on the final synthesis map a series of elements highlighting the areas where suicide is really an exceptional phenomenon (statistically speaking!) and the areas where this phenomenon seems to be spatialized, to take root.

5. GEOSTATISTICAL ANALYSIS AND RESULTS

Suicide is a sensitive subject for society, although its study is essential in identifying its causes, spatial and temporal distribution trends, and social diffusion channels. The phenomenon of suicide is treated with great caution by most of the world's nations for various reasons. In the European Union, legislation on personal information is a serious obstacle to comprehensive research into the phenomenon. In other countries, such as Islamic countries, the phenomenon is stigmatised and declared illegal, which could cause serious distortions of statistical or geographical reality. (Mishara et al., 2016; Arafat et al., 2022).

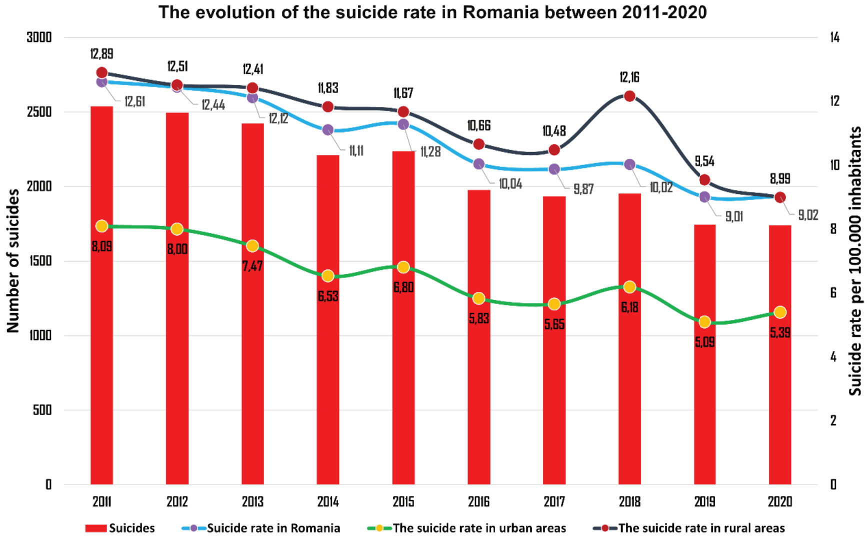
Despite the more or less justified obstacles, the subject should not be lost sight of, as its dimensions can become worrying. In 2019, 1.3% of deaths worldwide were caused by suicide; in absolute terms, this means that the world lost 759,028 people to suicide, which far exceeds Luxembourg's entire population. (Ritchie et al., 2015). Therefore, Romania's positioning in this global context becomes necessary, as it is thus possible to relativize the orders of magnitude of the phenomenon under investigation domestically.

At the undesirable pole, i.e. the countries with the highest suicide rates, are Lesotho (42.2 suicides/100,000 inhabitants), Guyana (32.6 suicides/100,000 inhabitants), Kiribati (26.6 suicides/100,000 inhabitants) and Ukraine (26.3 suicides/100,000 inhabitants). High values are also recorded in the republics that were part of the former Soviet Union. At the other end of the ranking were Syria (2.3 suicides/100,000 inhabitants), Jordan (2.5 suicides/100,000 inhabitants), Sao Tome and Principe (2.6 suicides/100,000 inhabitants), Oman (2.6 suicides/100,000 inhabitants) and Kuwait (2.8 suicides/100,000 inhabitants). However, these figures should be viewed with caution because of the strong interference of statistics with the religious dimension in the public life of Muslim societies which systematically report low suicide rates. (Ritchie et al., 2015).

According to the data provided by the National Institute of Statistics, during the period analysed (2011-2020) in Romania there was a downward trend in the incidence of the phenomenon, whose rate decreased from 12.61 suicides per 100,000 inhabitants in 2011 to a value of 9.02 in 2020 (Figure 2). The tables on the Eurostat[1] website, slightly different from those of the NSI, show that both the values and the trend place our country at the level of the European Union average.

The evolution of the suicide rate in Romania between 2011-2020,

Figure 2



Data source: INSSE 2021

The average rate calculated for the period analysed was 10.75 suicides per 100,000 inhabitants. (Figure 2). In absolute values, this translates into 2,538 suicides in 2011 and 1,739 cases in 2020. Overall, for the 10 years studied, a total of 21,252 suicides[2] were recorded, a value equivalent to the population of the municipality of Moinești. Thus, even if the national suicide rate places Romania in a reasonable position in the global hierarchy, the losses

[1] https://ec.europa.eu/eurostat/databrowser/

[2] The total from the spatial analysis performed at the commune level in this study differs from the actual total because communes with less than 3 suicides were not analysed with their officially recorded headcount.

(and their effects) are not insignificant. Extending the analysed interval with data taken for the period after 2003 from the Institute of Forensic Medicine “Mina Minovici” in Bucharest, we reach the impressive figure of 48,951 suicides, which means the population of a medium-sized Romanian city such as Turda.

Elementary statistical analysis applied to statistical information shows that, apart from a few particularities, Romania generally fits the accepted pattern of evolution and structuring of the suicidal phenomenon (Sârbu, 2017), a pattern described by a multitude of studies and researches, conducted in various fields (medical sciences, humanities and socio-human sciences, administrative sciences, etc.) and in different parts of the world. Thus, the Romanian gender behaviour in the face of suicide obeys the so-called “female paradox” Andrade Palma et al. (2021), i.e. the finding that, compared to men, the rate is 3 times higher for suicide attempts among women, but the mortality by suicide is 4 times higher in men than in women. As a result, also in Romania (Table 2), most suicide acts are committed by males (Sava and Papari, 2015), with male suicide mortality being 5 times higher than that of females (Rădulescu, 2014).

In order to highlight the social impact that this act has on the social body, perhaps it would be appropriate here to recall Cedric Mims’ (2020) statement that for every completed suicide there are around 20 failed attempts. If we extrapolate, exaggerating somewhat[1], it would result that in Romania, for the period studied, more than 425,040 people have at some time resorted to a suicide attempt.

Suicide rate by gender

Table 2

	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
European Union (general)	12,39	12,38	12,25	11,81	11,39	10,75	10,49	10,49	10,15
Romania (general)	***13,00***	***12,67***	***12,24***	***11,33***	***11,43***	***10,16***	***9,97***	***10,02***	***9,03***
European Union (females)	5,22	5,22	5,30	5,15	5,03	4,66	4,65	4,67	4,41
Romania (females)	***3,89***	***3,42***	***3,60***	***3,16***	***3,32***	***3,06***	***3,22***	***3,11***	***2,39***
European Union (males)	20,73	20,69	20,28	19,52	18,75	17,76	17,21	17,20	16,70
Romania (males)	***23,07***	***22,98***	***21,78***	***20,49***	***20,55***	***18,17***	***17,52***	***17,84***	***16,47***

Data source: Eurostat, 2023

The incidence by age group (Table 3) also confirms the known fact that the cohorts most affected by suicide are the mature and elderly, followed at a large distance by the young.

[1] Studiile statistice arată o medie mondială de 5-7 tentative eșuate la un act finalizat.

Suicide rate by age groups (samples)

Table 3

	2011	2012	2013	2014	2015	2016	2017	2018	2019
European Union (15-19)	4,85	4,82	4,76	4,90	4,35	4,21	4,39	4,20	4,30
Romania (15-19)	***6,60***	***5,84***	***5,42***	***4,52***	***5,64***	***3,98***	***5,17***	***4,88***	***4,17***
European Union (50-54)	19,32	19,21	18,71	18,06	16,97	15,95	15,37	15,16	14,51
Romania (50-54)	***22,62***	***20,30***	***20,52***	***18,48***	***18,44***	***16,69***	***16,71***	***17,34***	***12,33***
European Union (+85)	25,36	25,88	25,13	25,15	24,99	24,04	23,80	24,31	23,29
Romania (+85)	***17,13***	***14,62***	***17,35***	***15,60***	***21,97***	***14,60***	***12,04***	***14,98***	***12,14***

Data source: Eurostat, 2023

The situation by setting also does not differ from the general pattern outlined in most of the studies of the last two decades (Casant and Helbich, 2022). During the period analysed, most suicides occurred in rural areas which, with 10,980 suicides, cover 57% of the cases[1], the remaining 43%, i.e. 8,189 suicides, belong to urban localities. The greater intensity of the phenomenon in rural areas is also shown by the relative values, the suicide rate, with an average of 11.31 suicides per 100,000 inhabitants, being higher than that recorded in urban areas, with 6.5 suicides per 100,000 inhabitants (Figure 2).

The seasonality of suicide incidence shows that, contrary to common myths promoted by the national media, most suicides do not occur during the winter holidays, but during the spring and summer months (Rădulescu, 2014; Sava and Papari, 2015), peaking in July. This incidence of seasonality has long been confirmed by Dr. Nicolae Minovici, who, 150 years ago, in his work "Study on Hanging", noted a significant increase in the incidence of suicides in the spring period (Minovici, 2007).

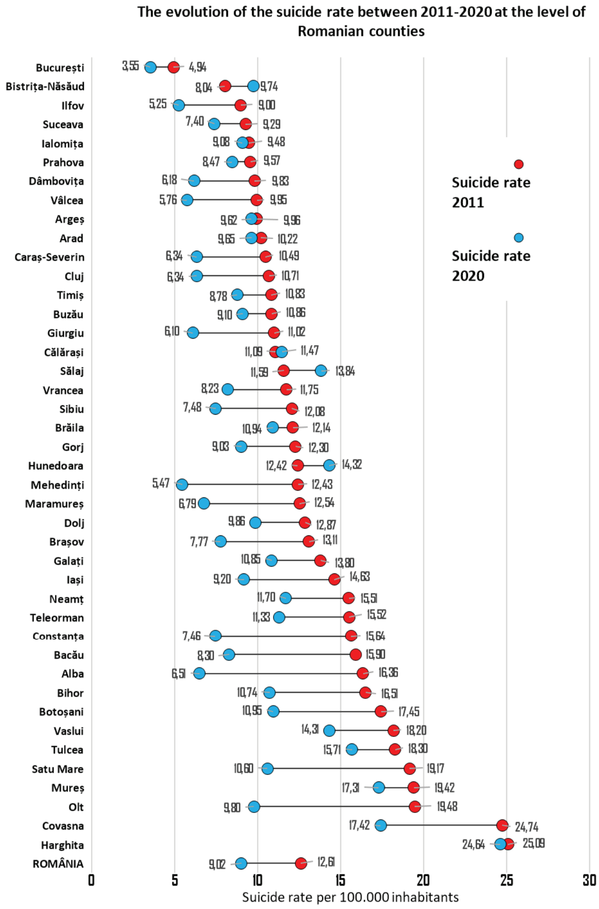
The general framing of the phenomenon at the NUTS 2 (development regions) and NUTS 3 (counties) administrative levels is also necessary to introduce the spatial analyses that follow. Among the development regions, the highest average suicide rate was recorded by the Centre region, with an average of 15.40 suicides per 100,000 inhabitants. Its leading position is due to a strong cultural dimension, which is more clearly evident when analysing the situation by counties and even more clearly when analysing the situation at the municipal level. The North East region, with an average of 12.58 suicides per 100,000 inhabitants, is a close second, and an analysis of the factors responsible would probably highlight the role of its eastern, peripheral position combined with that of an austere economic environment. The South-East region follows

[1] Values are approximate, taking into account that 9.8% of cases could not be located in the territorial profile

with a rate of 11.76 suicides per 100,000 inhabitants, followed by the North-West region with a rate of 11.53, the South-West-Oltenia region with a rate of 9.89 and the South-Muntenia region with a rate of 9.73 suicides per 100,000 inhabitants. The lowest values belong to the Bucharest-Ilfov and West regions, with average rates of 4.69 and 9.26 suicides per 100,000 inhabitants respectively. In general, the dynamics of the regional values over the period analysed was moderate, maintaining the general downward trend, in line with the national trend, which is also true for the counties, which, with small exceptions such as Bistrita, Calarasi, Hunedoara or Sălaj, followed the same trend (Figure 3).

The evolution of the suicide rate between 2011-2020 of Romanian counties

Figure 3

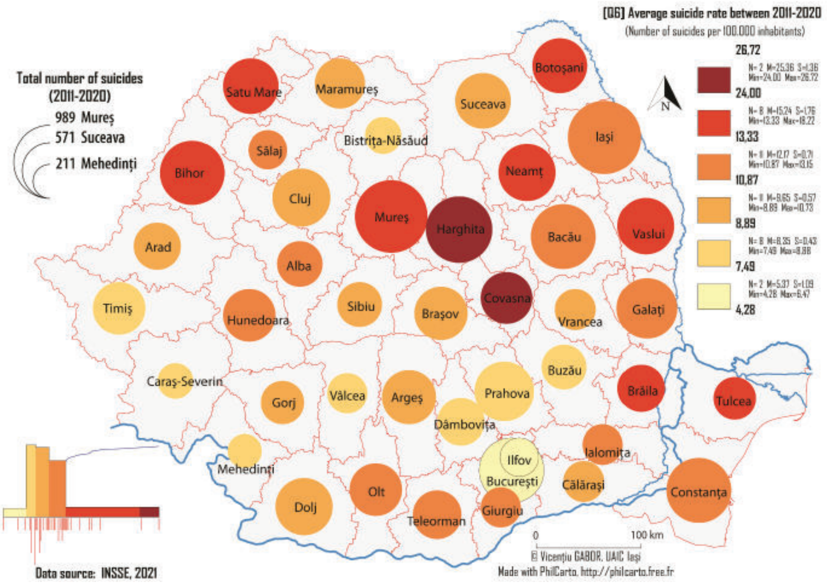


Data source: INSSE 2021

The highest average rates were recorded, unsurprisingly (see also Rădulescu, 2014), by Harghita (26.72), Covasna (23.99), Mures (18.21), compact spatial block (Figure 4), to which, as usual, Satu Mare (17.09) is added. The perennial association of the four counties at the top of the hierarchy is explained in particular by the geographical distribution of the Hungarian ethnic population.

Distribution of suicides by county

Figure 4



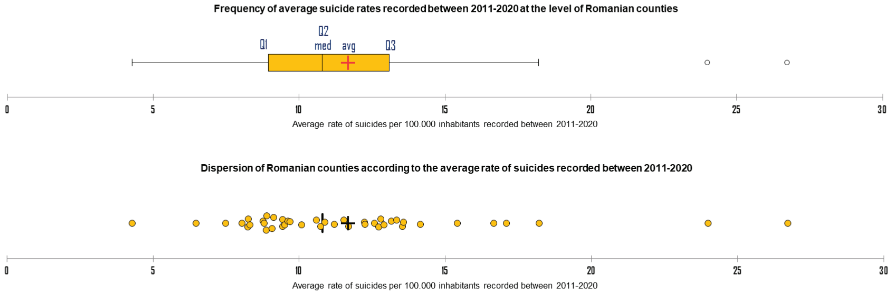
Data source: INSSE, 2021

The counties with the lowest average rates for the analysis interval ranked, in order, as follows: Bucharest (4.28), Ilfov (6.46), Timiș (7.4), Dâmbovița (8.04) and Vâlcea (8.23). The overall county dynamics shows a decreasing trend, with only Harghita county maintaining a certain stability of the rate values at the end of the ten-year interval (Figure 3).

The box-plot analysis (Figure 5) shows that at the county level the distribution is relatively homogeneous and allows mapping by discretizing the values into quantile (Q) based classes, slightly modified geographically (Waniez, 2023b). The modification involves initially removing the 5% lowest and 5% highest values, performing the discretization, re-inserting the removed values into the corresponding classes, and mapping the classes. This method leads not to statistical classes (with equal numbers) but to geostatistical classes (with slightly unequal numbers), favouring cartographic expression over statistical rules, which are sometimes too rigid about territorial realities.

Boxplots of the distribution of the suicide rate incidence at county level

Figure 5



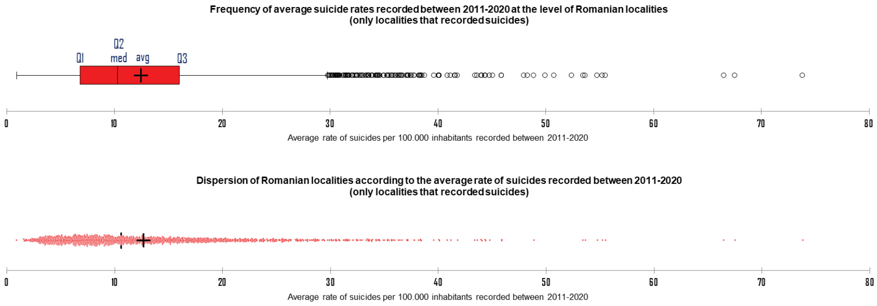
Data source: INSSE 2021

For the county values, the minimum is recorded by Bucharest, with 4.28 suicides per 100,000 inhabitants, and the maximum is reached by Harghita county with 26.72 suicides per 100,000 inhabitants. Most of the relevant county values (50%) oscillate between Q1 = 8.93 and Q3 = 13.09 suicides per 100,000 inhabitants, the median value being Q2 = 10.79 suicides per 100,000 inhabitants. The difference between average and median is much smaller than in the case of communes, i.e. only 0.87, a good sign for the normality of the distribution. Nevertheless, outliers are present, being represented by Harghita and Covasna counties, even if their values have decreased a lot compared to the situation presented for 1996-2012 by Rădulescu (2014).

At the communal level, things are more spectacular, but more difficult to map. Box-plot analysis at the scale of the UATs (Figure 6) shows that the asymmetry is exacerbated, making it difficult not so much to analyse the values statistically, but to represent them cartographically.

Boxplots of the distribution of the suicide rate incidence at local level (only TAUs with incidents)

Figure 6



Data source: INSSE 2021

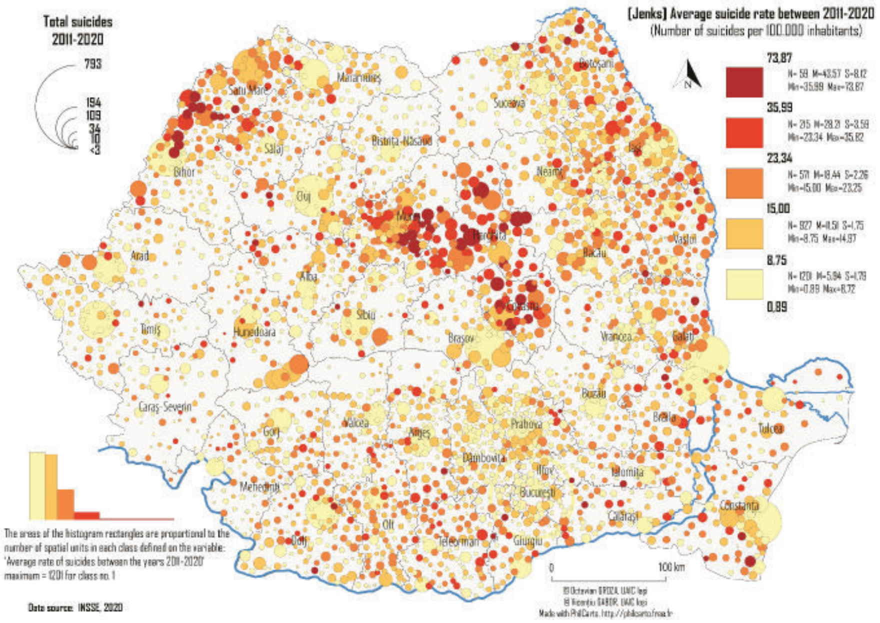
The chart refers only to the localities with incidents recorded over the decade studied, i.e. 2,973 units, totalling 93.46% of Romania's area. The situation seems catastrophic, but this is only one aspect of the effect of size: it is enough to separate the period into two intervals (2011-2015 and 2016-2020) to see that in the first a number of 548 TAUs did not register any case of suicide, and in the second the number increases to 679. As these are not the same TAUs, the total for the ten years may lead to unjustified, if not pernicious, conclusions. Returning to the box-plot, the minimum suicide rate for the decade analysed was recorded by the town of Măgurele in Ilfov county, i.e. 0.9 suicides per 100. 00 inhabitants. At the opposite end of the distribution, the commune of Brateş in Covasna county reached a value of 73.86 suicides per 100,000 inhabitants, thus leading the outliers and increasing the amplitude to 72.96. This obviously complicates the discretization into classes, with the interquartile range clustering 50% of the values on only 9.19 units (Q1 = 6.81 and Q3 = 16). The median (Q2 = 10.29) and even the mean = 12.43) confirm this. Therefore, a discretization close to the geometric one would be most appropriate. In this case we used the Jenks algorithm (minimization of intra-class variance and maximization of inter-class variance), available in the discretization modules of the mapping software PhilCarto. The result of the discretization led to the analytical cartogram in Figure 7.

Its analysis allows, on the one hand, to confirm some previous conclusions (notable differences in suicide rates between urban and rural areas, the outline of the areas in which the Hungarian population is concentrated) and, on the other hand, it opens up new questions, concerning, for example, the notable differences between the east and the south of the country or between the three compartments of the Dobrogea region, from north to south. What is important, however, is the overall visual impression created by the map, i.e. the alternation of empty and full spaces whose territorial arrangement is not random, at least to the geographer's trained eye.

For example, the "gaps" are not only due to difficult natural conditions (high mountain heights, delta), nor are they only due to the pitfalls that lurk for those unfamiliar with territorial analysis (mountain and sub-mountain municipalities with large areas and population centres located towards the periphery, in their lower parts). Many of these gaps are simply due to the lower incidence of suicide.

Distribution of suicides by TAU (Territorial Administrative Units)

Figure 7



Data source: INSSE, 2020

During the period 2011-2020, 208 administrative units (6.5% of the total) had no cases of suicide and 1,414 administrative units (44.5% of the total) had three or fewer cases. These two categories, totalling 1,622 administrative units (51% of all TAUs), accounted for a total of 2,912 cases during the decade under study, which represents only 15.2% of the total number of 19,169 suicides counted at municipal level. Their geographical distribution is quite wide, covering most of the territory. The low absolute values that characterise them (although sometimes accompanied by high rates due to the inconsistency of the statistical series) tend to be spatially associated. This means that the TAUs they describe form contiguous areas, such as the Banat, the Apuseni Mountains, southern and northern Transylvania, the extreme west of Oltenia, the low-density areas of the southern and eastern Carpathians or central Dobrogea (Figure 7).

Conversely, "full" areas are generated by the association of high absolute values and high rates, such as the eastern half of Moldavia, the contact area between Muntenia and Oltenia, southern and northern Dobrogea or the almost continuous area described by the location of ethnic Hungarians.

Further analysis would be difficult and could lead to hasty conclusions, even if we were to complement the visual analysis with that of data tables and various statistical parameters. A single example can support this statement. The map in Figure 7 shows both rural and urban TAUs. Cities, even though generally characterised by low suicide rates, by necessity have high absolute values, represented by circles, which cannot be enlarged very much without affecting the overall readability of the map. These circles "tase" the circles with the small values of rural communes (and small and very small towns), often masking relevant spatial configurations. Conversely, the high rates of some rural communes may hide rate variations within the urban system. For this reason, there is a need to extend the research methods carried out in the second part of this article.

The second part of our research, therefore, has two objectives. The first is to mitigate as far as possible the inaccuracies in the interpretation of territorial realities. Uncertainties can be caused in particular by over-generalising (geo)statistical analyses, influenced either by the coarse level of aggregation of statistical information (county, regional, national), or by the mixing of orders of magnitude that are too different (e.g. undifferentiated analysis of small, medium and large towns), or by the mixing of different structures (e.g. rural analysed together with urban). The second objective, which is already mentioned in the *Abstract*, focuses on the need to discover the spatial (geographical) structures of a phenomenon which seems to have nothing spatial in its emergence and development, and which may therefore give the false impression that it cannot be the subject of territorially differentiated public policies.

The analysis of Table 4 is optimistic, as the rows seem to indicate both a decrease in the intensity of the suicide phenomenon (gradual reduction in absolute values, rates and annual shares of the total for the decade) and a spatial restriction (fewer administrative units are affected as time goes by). But how consistent is this block analysis of the Romanian territory, even if it is carried out on the basis of the population's average living conditions? Analysis of such a table induces in the viewer's mind the idea of a frozen, inert spatial framework, in which things happen and over which time passes, as the only variable that would explain the dynamics of events. Reductionism is a natural and necessary reflex for understanding the surrounding complexity, but in territorial analysis it must not limit the analyst's thinking to the level of homogeneous categories with smooth and compact dynamics. If, for example, the suicide phenomenon was to disappear completely from Bucharest in the decade under analysis, the annual percentage shares of the total for the ten years would decrease in the table by an average of 1%, while in the other 319 cities the suicide dynamics would not change.

Spatial distribution of suicides in the urban* and rural TAU of Romania

Table 4

		2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Urban (320 TAU)	Number of cities with suicide cases	243	238	225	224	227	227	219	224	198	208
	% of cities	*75,9*	*74,4*	*70,3*	*70,0*	*70,9*	*70,9*	*68,4*	*70,0*	*61,9*	*65,0*
	Total suicide cases in urban	1028	1013	944	822	856	731	708	775	638	674
	% of number of suicides in urban (8.189 between 2011-2020)	*12,6*	*12,4*	*11,5*	*10,0*	*10,5*	*8,9*	*8,6*	*9,5*	*7,8*	*8,2*
Rural (2861 TAU)	Number of rural TAU with suicide cases	1045	1014	1023	975	983	896	904	886	815	803
	% of rural TAU	*36,5*	*35,4*	*35,8*	*34,1*	*34,4*	*31,3*	*31,6*	*31,0*	*28,5*	*28,1*
	Total suicide cases in rural	1255	1218	1207	1149	1133	1035	1016	1178	922	867
	% of number of suicides in rural (10.980 between 2011-2020)	*11,4*	*11,1*	*11,0*	*10,5*	*10,3*	*9,4*	*9,3*	*10,7*	*8,4*	*7,9*
National (3181 TAU)	Number of TAU with suicide cases	1288	1252	1248	1199	1210	1123	1123	1110	1013	1011
	% of TAU	*40,5*	*39,4*	*39,2*	*37,7*	*38,0*	*35,3*	*35,3*	*34,9*	*31,8*	*31,8*
	Total suicide cases	2283	2231	2151	1971	1989	1766	1724	1953	1560	1541
	% of total number of suicides (19.169 between 2011-2020)	*11,9*	*11,6*	*11,2*	*10,3*	*10,4*	*9,2*	*9,0*	*10,2*	*8,1*	*8,0*

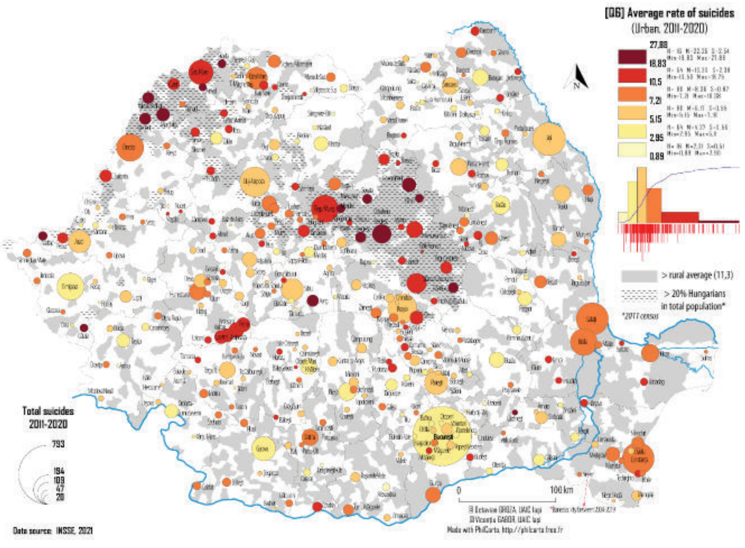
**Between 2004 and 2019, the TAU Băneasa, Constanța county, functioned as a city. In calculations, its values are included in urban.*

Data source: INSSE, 2021

Moreover, from year to year, the values do not belong to the same spatial units, as we might falsely assume, so the dynamics illustrated do not belong to the same collective of individuals, fixed in the same locations. For example, 471 TAUs, which between 2011-2015 totalled 870 suicides (8.2% of the 10,625 recorded in the first 5 years analysed), have not experienced any new cases since 2016. On the other hand, a number of 340 TAUs, which had not experienced this phenomenon, totalled 585 suicides between 2016-2020, representing 6.65% of the total of 8,544 cases of the last 5 years analysed. In the table, the decrease in values is evident, both at national level and at the level of the two categories, urban and rural. Much less obvious is whether the decrease is the result of a uniform behaviour or not of the spatial units. The major discontinuities in the series and the small number of observations make it unnecessary to use dispersion parameters, either absolute or relative. In any case, they would provide information about something quite different from where and how strongly things are evolving there. Cartographic analysis, with all its limitations, can add clarity to this type of analysis. Figure 8, which illustrates the spatiotemporal dynamics of suicide in the Romanian urban environment allows, on the one hand, to deepen the observations and, on the other hand, initiates a series of other questions, as possible hypotheses for future studies (e.g. explaining the similar behaviour of large and medium-sized cities in the south-east of Romania).

Distribution of suicides in the Romanian cities and their territorial context

Figure 8



First of all, it should be noted that, with the exception of the area with a high presence of ethnic Hungarians, urban areas are relatively independent of the behaviour of the rural areas in which they are located. Thus, in rural areas with sub-average suicide rates, cities can have low or high rates (Focșani and Tecuci, Gătaia and Bocșa, Vișeu de Sus and Borșa), while the opposite is also possible in rural areas with above-average rates (Zalău and Marghita, Sighișoara and Cristuru Secuiesc, Costești and Scornicești). Secondly, it is clear that the separate analysis of urban areas reduces the impression of homogeneity in the behaviour of cities that Figure 7 suggested, where, because of the high rates in rural areas, cities were mostly concentrated in only two classes, with the described limits of 0.89 and 15 suicides per 100,000 inhabitants. Thus, the high rates, visible in Figure 8, of some large and medium-sized cities, such as Brăila, Galați, Constanța, Tulcea, Slatina, or the mining towns in the Jiu coal basin (Uricani, Lupeni, Vulcan, Aninoasa, Petroșani, Petrila), were hidden on the general analytical map in Figure 7. Similarly, the mass of rural TAUs masked to a large extent the dynamics of small and very small towns, such as those in the Danube valley, in the Prahova Subcarpathians or in the east and south-east of the country. The Banat region no longer seems very homogenous either. Analysis of the map also gives legitimacy to the hypothesis of the role of town size on the incidence of suicide (Table 5).

Distribution of suicides by classes of size of urban TAUs

Table 5

	Suicides	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
under 10.000 inhab.	Number	81	87	73	70	82	75	58	89	44	55
	Rate	*10,0*	*10,8*	*9,1*	*8,8*	*10,3*	*9,5*	*7,4*	*11,3*	*5,6*	*7,1*
10.000-25.000 inhab.	Number	180	154	160	143	136	155	140	143	120	127
	Rate	*10,3*	*8,9*	*9,3*	*8,3*	*7,9*	*9,0*	*8,2*	*8,4*	*7,1*	*7,5*
25.000-50.000 inhab.	Number	188	143	139	104	117	93	96	135	88	93
	Rate	*11,9*	*9,1*	*8,8*	*6,6*	*7,5*	*5,9*	*6,1*	*8,6*	*5,6*	*6,0*
50.000-100.000 inhab.	Number	113	140	113	104	101	69	86	85	72	82
	Rate	*7,3*	*9,1*	*7,4*	*6,9*	*6,7*	*4,6*	*5,8*	*5,7*	*4,9*	*5,6*
over 100.000 inhab.	Number	466	489	459	401	420	339	328	323	314	317
	Rate	*6,6*	*7,0*	*6,6*	*5,7*	*6,0*	*4,9*	*4,7*	*4,6*	*4,5*	*4,5*

Data source: INSSE, 2021

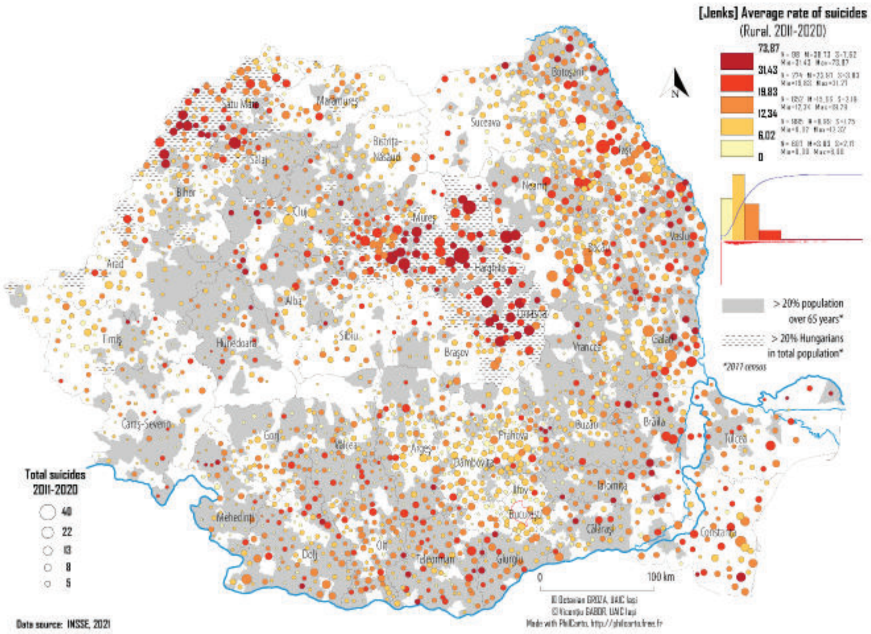
Figure 8 and Table 5 raise questions not so much about the high rates of small and very small cities, many of which have urban status only by force of law, but about the fluctuating rates of medium-sized cities and especially about the low rates of large cities, where the stresses and problems of modern life should work against them.

In addition to these elements, the map also allows us to detect a certain regionalisation of urban behaviour, as a result of the probable action of a number of factors that can be suspected: relatively homogeneous ethno-confessional areas, isolation and peripheral character (southern and eastern borders), former mining, oil or mono-industrial areas (Valea Jiului, Maramures, west of Bacău, Prahova)

In the same vein, a separate analysis of rural areas (Figure 9) allows a better understanding of the spatial aggregations of TAUs with the same behaviour: the area with the presence of Hungarian communities, the eastern half of Moldavia, eastern Oltenia and western Muntenia, eastern Bărăgan, northern and southern Dobrogea. Equally important on the map are the relative "gaps", formed by contiguous UATs, with low values (and sometimes rates). Both types of rural areas show how difficult it is to suggest the existence of explanatory factors: the shade of grey highlights ageing areas, which are usually also isolated and with a predominantly agricultural economy, relatively poor, but which may also be present within the "gaps" and in the case of "full" areas.

Distribution of suicides in the Romanian rural space

Figure 9



Data source: INSSE, 2021

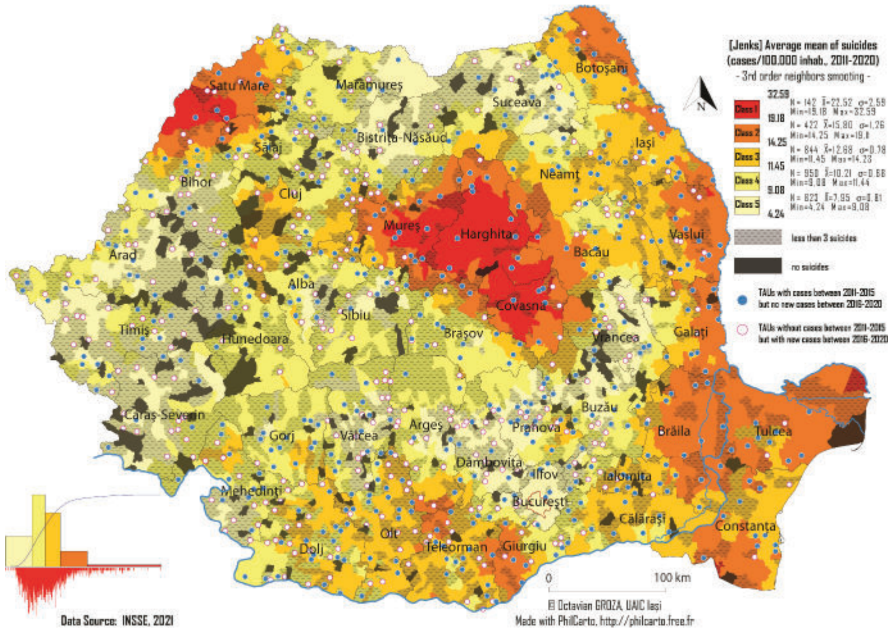
© Octavian GROZA, UAIC Iași
© Vicențiu GABOR, UAIC Iași
Made with PhilCarto, http://philcarto.free.fr

What is more important, however, for the main objective of our research, is the fact that both tend, through the coagulation of the TAUs, to form spatial structures, which justifies the last step, namely the attempt to discover the territorial architecture resulting from the territorial dynamics of the suicide phenomenon.

For this, according to the methodology, we used the listing of the values of neighbouring spatial units, which dilutes individual situations in favour of the expression of collective structures based on contiguity. Listing the values down to the level of 3rd order neighbours of each of the 3,181 TAUs gives a clearer picture of the distribution and geographical position of the voids and fills analysed above (Figure 10). The resulting map is not the result of a simple geostatistical game. In order to justify this statement, a synthetic presentation of the five classes of mapped values is necessary, which we will analyse in relation to the real values (Table 6). The real values are already present on the figure through the colours that individualise the 208 TAUs that did not experience any case of suicide during the ten years, as well as through the hatches that describe the 1,414 TAUs with less than three suicides during the whole period.

Spatial structures of the suicidal phenomenon in Romania

Figure 10



Data Source: INSSE, 2021

© Octavian GROZA, UAIC Iași
Made with PhilCarto, http://philcarto.free.fr

The presence of these graphical artifices has the gift, on the one hand, of underlining the contiguity and consistent behaviour of the areas less affected by the phenomenon analysed, and on the other hand, of reminding us that the areas highly affected by suicide are nevertheless the product of random, individual acts and do not necessarily characterise the entire population concerned.

Spatial declination of suicidal phenomenon in Romania

Table 6

		2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**Average**
Class 1 142 TAU (4,5%)	% of population	3,8	3,8	3,8	3,8	3,8	3,8	3,8	3,8	3,8	3,8	**3,8**
	% of suicides	7,4	9,1	8,5	8,6	9,1	9,9	8,9	9,5	10,7	10,1	**9,2**
	suicide rate	*20,08*	*23,90*	*21,68*	*20,07*	*21,52*	*20,73*	*18,27*	*22,25*	*20,04*	*18,79*	***20,7***
Class 2 422 TAU (13,3%)	% of population	10,8	10,8	10,8	10,8	10,7	10,7	10,7	10,6	10,6	10,6	**10,7**
	% of suicides	15,8	17,5	15,7	16,6	15,6	17,6	17,6	17,4	15,1	14,3	**16,3**
	suicide rate	*14,85*	*16,14*	*14,01*	*13,61*	*12,98*	*13,05*	*12,82*	*14,39*	*10,03*	*9,44*	***13,1***
Class 3 844 TAU (26,5%)	% of population	24,3	24,2	24,3	24,3	24,4	24,4	24,4	24,3	24,3	24,3	**24,3**
	% of suicides	28,7	26,9	27,1	29,0	28,6	25,9	26,1	25,0	27,0	27,1	**27,1**
	suicide rate	*12,05*	*11,05*	*10,74*	*10,52*	*10,46*	*8,43*	*8,31*	*9,04*	*7,80*	*7,77*	***9,6***
Class 4 950 TAU (29,9%)	% of population	26,4	26,4	26,4	26,4	26,3	26,3	26,3	26,2	26,2	26,1	**26,3**
	% of suicides	25,6	24,7	26,5	24,1	24,6	26,1	24,9	25,8	23,7	25,4	**25,1**
	suicide rate	*9,86*	*9,33*	*9,66*	*8,08*	*8,33*	*7,87*	*7,36*	*8,64*	*6,36*	*6,76*	***8,2***
Class 5 (823 TAU (25,9%)	% of population	34,7	34,8	34,8	34,7	34,8	34,8	34,9	35,0	35,2	35,3	**34,9**
	% of suicides	22,4	21,8	22,2	21,8	22,1	20,6	22,4	22,3	23,5	23,0	**22,2**
	suicide rate	*6,56*	*6,24*	*6,13*	*5,54*	*5,68*	*4,68*	*4,99*	*5,59*	*4,70*	*4,55*	***5,5***

Data source: INSSE, 2021

Class 1 is the most problematic, over the 10 years it has averaged 9.2% of suicides for only 3.8% of the country's population, which is a ratio of 2.44, well above, for example, class 5, which covers on average 22.2% of suicides but for only 34.9% of Romania's population, a ratio of 0.64. Since it is comprised, with extremely few exceptions, of administrative units where the Hungarian community exceeds 20% of the population, it is extremely tempting to put forward a culturally based explanation. Such an explanation has to be constructed very carefully because local situations can be quite complicated (as shown, for example, by the study carried out by Mitroi in 2007). The average rate for this class is double the national rate.

Class 2 is also characterised by a high ratio of 1.52 (16.3% of suicides for 10.7% of the population). High ratios (14.9 in 2011 and 13.1 in 2020) characterise a number of TAUs clustered in haloes around the compact blocks of the first class. Another series of units creates an almost continuous band on the eastern extremity of Moldova, with extensions into the Bărăgan and northern Dobrogea. Finally, a third series is grouped in nuclei, more extensive (on the border between Giurgiu and Teleorman counties or on the borders between Teleorman, Argeș and Olt) or more isolated, in Alba, Cluj and Mehedinți counties.

Class 3, with a ratio slightly above unity (1.12, i.e. 27.1% of suicides for 24.3% of the population) is similar in this respect to class 4 (a ratio of 0.95), but its spatial role is different. Forming haloes around the aggregates of the previous class, the TAUs in this group often form “bridges” between the aggregates with higher listed rates (Iasi-Neamt, Vaslui-Bacău, Ialomița-Călărași-Constanța, Mureș-Cluj, Giurgiu-Teleorman. The spatial structures of this class could be seen as transitional spaces, a feature also suggested by the fact that the exceptional events, represented by the blue and white circles on the map, have a higher spatial frequency than in the higher classes, but lower than in the lower classes. The unlikely but possible hypothesis that this spatial pattern might play some role in the spatial diffusion of suicide should be explored.

Classes 4 and 5, generally characterized by the lowest absolute values, group 1,773 of the TAUs (55.7%), which encompass 61.38% of Romania’s population (2020) and cover 54.83% of the national territory. The areas described by these two classes are characterized by an extremely fluid and random dynamics of the suicidal phenomenon. For example, they concentrate 65.6% (309) of the 471 TAUs with no suicide cases after 2015 and 69.1% (235) of the 340 TAUs unaffected until 2015 but with suicide cases after this year. If we were to add class 3, which rather outlines a transition area, then these exceptional values, which would cover 2,617 TAUs, would rise to 90.23% and 90.88% respectively, which would leave the idea of the structurality of the suicide phenomenon hanging over less than 10% of the national territorial ensemble. Also, in the areas circumscribed by classes 4 and 5 are almost all the TAUs which, at least during the period analysed, did not have any suicides.

The obvious conclusion is that the first two classes, to which we can add with reservations the third class, which we consider more of an interstitial space, tend to exhibit forms of structural, territorially rooted suicide. In the spaces outlined by these first two/three classes, absolute values and suicide rates remain high over time, and TAUs with turbulent behaviour (long-term disappearances of the phenomenon and unexpected reappearances) are rare or exceptionally rare. Public authorities and social scientists should focus their attention on these areas more vigorously, as it is unlikely that random factors, such as the health of the individual, are responsible for this territorialisation.

6. CONCLUSIONS

In the analysed interval (2011-2020), a downward trend in the incidence of suicides was recorded in Romania, the rate decreasing from 12.61 suicides per 100,000 inhabitants in 2011 to a value of 9.02 in 2020. In 2011 , a number of 2,538 suicides were recorded in 2011, and in 2020, a number of 1,739 suicides. The total number of suicides recorded during the analysed period was 21,252.

In a regional profile, the highest average suicide rate was recorded by the Center region, with an average of 15.40 suicides per 100,000 inhabitants. This region is followed in the order of severity of this phenomenon by the North East region with an average value of 12.58 suicides per 100,000 inhabitants. The lowest values belong to the Bucharest-Ilfov and West regions, with average rates of 4.69 and 9.26 suicides per 100,000 inhabitants, respectively.

At the county level, the highest average rates were recorded by the counties of Harghita (26.72), Covasna (23.99), Mureș (18.21), Satu Mare (17.09). The counties with the lowest average rates for the period 2011-2020 were ranked, in order, as follows: Bucharest (4.28), Ilfov (6.46), Timiș (7.4), Dâmbovița (8.04) and Vâlcea (8,23).

Regarding the phenomenon of suicides at the local level, the lowest value of the suicide rate for the analysed decade was recorded by the town of Măgurele in Ilfov county, respectively 0.9 suicides per 100,000 inhabitants. At the opposite pole of the distribution, Brateș commune in Covasna county reached a value of 73.86 suicides per 100,000 inhabitants.

The obtained results confirm the hypothesis that governs this study, demonstrating that suicidal phenomena may have a tendency to be territorially rooted and that the perpetuation of this behaviour among the components of a population may lead to the individualization of relatively homogeneous and relatively stable areas over time. While at aggregate levels of statistical information corresponding to higher-level administrative divisions (counties and regions) the phenomenon may appear to be random in nature, analysis at a finer scale identifies areas where the phenomenon can take on structural dimensions, even in relatively short time series.

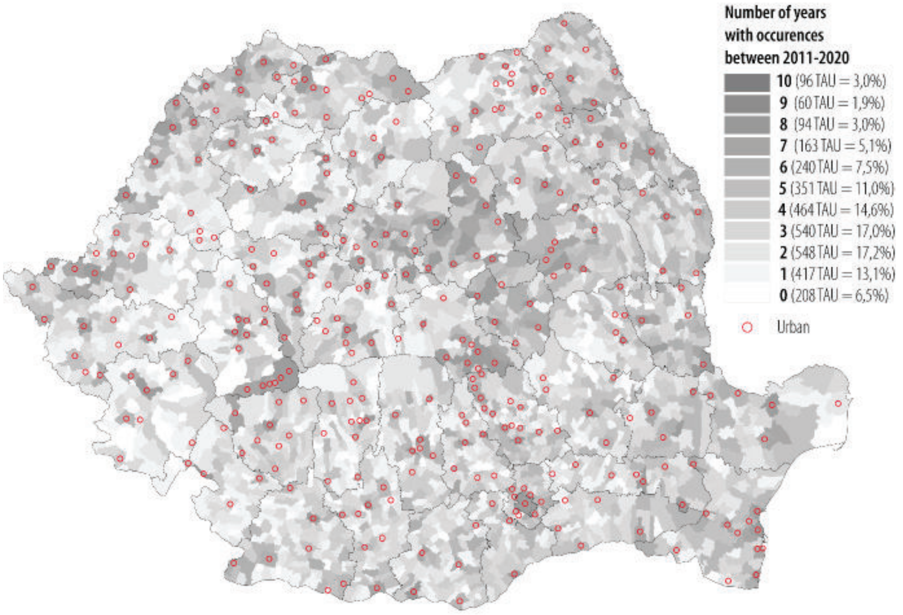
Multiple additional checks carried out (spatial autocorrelation analyses, annual analytical maps, annual and synthetic occupancy maps – **see Annex 1** etc.), which cannot be presented here due to editorial constraints, confirmed the research results. We can affirm that in some areas of Romania, other than the classic ones defined by the spread of the Hungarian ethnic group, suicide shows tendencies of territorialisation and that these areas must urgently come to the attention of the authorities called to prevent and reduce the incidence of the territorial phenomenon.

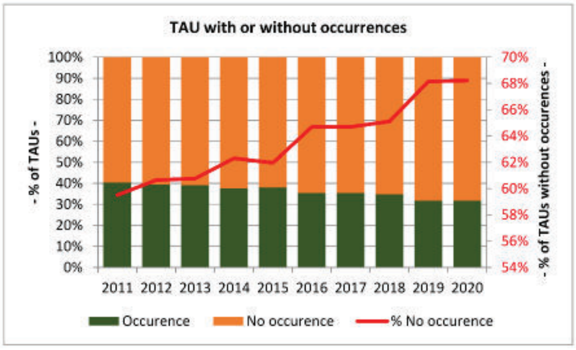
REFERENCES

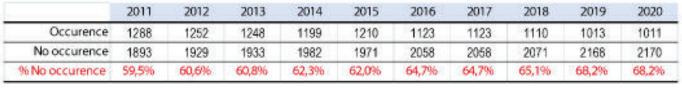
1. **Ajdacic-Gross, V., Weiss, M.G., Ring, M., Hepp, U., Bopp, M., Gutzwiller, F., Rössler, W.** (2008). Methods of Suicide: International Suicide Patterns derived from the WHO Mortality Database, *Bulletin of the World Health Organization*, 86, p. 726–732.
2. **Andrade Palma D. C., Ives de Oliveira B. F., Ignotti E.** (2021). *Suicide rates between men and women in Brazil, 2000-2017*, Cadernos de Saúde Pública 2021; 37(12), DOI: https://doi.org/10.1590/0102-311X00281020, available online at: https://www.scielo.br/j/csp/a/fQ6krtdS3Sstmc89zWCR VFq/?lang=en
3. **Arafat S. M. Y., Marthoenis M., Khan M. M., Rezaeian M.** (2022). *Association between Suicide Rate and Human Development Index, Income, and the Political System in 46 Muslim-Majority Countries: An Ecological Study*. European Journal of Investigation in Health, Psychology and Education, 2(7), 754-764, DOI: https://doi.org/10.3390/ejihpe12070055
4. **Bloom, M.** (2002). Rational Interpretations of Palestinian Suicide Bombing. Paper presented at the *Program on International Security Policy*, University of Chicago.
5. **Brădățan, C.** (2007). About Some 19th-Century Theories of Suicide. Interpreting Suicide in a East European Country, *International Journal of Comparative Sociology*, 48, 2007, p. 423, available online: https://doi.org/10.1177/0020715206070269.
6. **Brădățan, C.** (1999). Sinuciderea ca fenomen social: Suicide as Social Phenomenon. *Sociologie Românească*, *8*(2), 85-96. On-line:https://revistasociologieromaneasca.ro/sr/article/view/1324
7. **Casant J., Helbich M.** (2022). *Inequalities of Suicide Mortality across Urban and Rural Areas*: A Literature Review. International Journal of Environmental Research and Public Health, 19(5), 2669, DOI: https://doi.org/10.3390/ijerph19052669
8. **Durkheim, E.** (2007). *Despre Sinucidere*, Editura Institutul European: Iași.
9. **Havârneanu, G.** (2014). Sinuciderea: repere pentru strategii preventive eficiente. I. Dafinoiu & Ș. Boncu (Eds.), *Psihologie socială clinică* (pp. 120-132), Polirom: Iași
10. **Hume, D.** (1783). *Essay on Suicide*, Smith, London.
11. **Yip, P., Caine, E., Yousuf, S., Chang, S-S., Wu, K., & Chen, Y-Y**. (2012). Means restriction for suicide prevention. *Lancet*, 379, 2393-2399.
12. **Mimis, C.** (2020). *When We Die : The Science, Culture, and Rituals of Death*. 1st edition. St. Martin's Griffin: New York. ISBN: 978-0312264116. 384 p.
13. **Minois, G.** (2002). *Istoria Sinuciderii. Societatea occidentală în fața morții voluntare*. Humanitas: București.
14. **Minovici, N.** (2007). *Studiu asupra spânzurării*, Curtea Veche: București, ISBN: 978-973-669-396-0. 248 p.

15. **Mishara B.L., Weisstub D.N.** (2016). *The legal status of suicide: A global review*. Int. J. Law, Psychiatry, DOI: https://doi.org/10.1016/j.ijlp.2015.08.032, available online: https://www.sciencedirect.com/science/article/abs/pii/S0160252715001429?via%3Dihub
16. **Mitroi, O.** (2007). *Cercetare cu privire la fenomenul suicidar din România, în general şi din Judeţul Covasna*, *ANGVSTIA*, 11, 2007, Sociologie. 381-400, available online: https://biblioteca-digitala.ro/reviste/Angvstia/dl.asp?filename=11-Revista-Angvstia-11-2007-istorie-sociologie-38.pdf
17. **Rădulescu, S.M.** (2015). *Caracteristici ale sinuciderilor din România reflectate în mass-media şi mediul virtual. Evaluări cantitative şi clasificări calitative*. Revista română de sociologie, Serie nouă, XXVI (3-4), 223-251, available online: https://revistadesociologie.ro/pdf-uri/NR.3-4-2015/03-SRadulescu.pdf
18. **Rădulescu, S.M.** (2014a). *Constatări şi evaluări statistice cu privire la gradul de răspândire a sinuciderilor în România*. Revista română de sociologie, Serie nouă, XXV (5-6), 405-428, avaiable online: https://www.revistadesociologie.ro/pdf-uri/nr.5-6-2014/02-SRadulescu.pdf
19. **Rădulescu, S.M.** (2014b). *Evoluţii şi tendinţe ale fenomenului suicidar în România, în perioada 1996–2012.* Revista română de sociologie. serie nouă, XXV (3-4). 175–202, available online: https://www.revistadesociologie.ro/pdf-uri/nr.3-4-2014/02-Radulescu.pdf
20. **Ritchie, H., Roser, M., Ortiz-Ospina, E.** (2015). *Suicide*, Published online at OurWorldInData.org, Retrieved from: https://ourworldindata.org/suicide [Online Resource]
21. **Rossen, L. M., Khan, D.** (2016). *Mapping Suicide Death Rates: Geographic Aggregation Tools and Spatial Smoothing with Hierarchical Bayesian Models*, National Center for Health Statistics Geospatial Web Applications, Tools, and Data Workshop, US National Center for Health Statistics. Division of Vital Statistics, available online: https://nces.ed.gov/fcsm/pdf/GIG_Workshop_2016_Mapping_Suicide_Death_Rates_NCHS.pdf
22. **Sava, N. I., Papari A. C.** (2015). Comparative study on the phenomenon of suicide based on gender and season. *Procedia - Social and Behavioral Sciences*. 2015. 187. 532 – 535. DOI: 10.1016/j.sbspro.2015.03.099
23. **Sârbu, E.A**. (2017). *Pe urmele lui Durkheim. Harta sinuciderii în România post-comunistă*, 256 p., Tritonic: București.
24. **Sârbu, E.A.** (2015). The Lifeline for Suicide Prevention of Children and Teenagers in Bucharest, Romania. *Sociology Study*, 2015. 5 (5). 415-427. DOI: 10.17265/2159-5526/2015.05.009, available online: https://www.researchgate.net/publication/310779397_The_Lifeline_for_Suicide_Prevention_of_Children_and_Teenagers_in_Bucharest_Romania
25. **Stan, D.** (2021). Current Suicidal Situations Within the Romanian Area, *Scientific Annal of "Al. I. Cuza" University of Iasi, Sociology and Social Work,* Vol. XIV/ 1.
26. **UN-WHA** (2013). *Sixty-sixth World Health Assembly. Resolutions and decisions. Annexes*. Geneva, 20–27 may 2013. 171 p.
27. **UN-WHO** (2000). *Preventing suicide. A resource for general physicians*. Mental and Behavioural Disorders. Department of Mental Health World Health Organization. Geneva. 15 p.
28. **Waniez Ph.** (2023a). *Documentation complémentaire : lissages par voisinage* available at http://philcarto.free.fr/03_documentation/03_documentation.html
29. **Waniez Ph.** (2023b). *Cartographie thématique et Analyse des Données.* In *Documentation de base,* available at http://philcarto.free.fr/03_documentation/03_documentation.html

ANNEX 1







	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Occurence	1288	1252	1248	1199	1210	1123	1123	1110	1013	1011
No occurence	1893	1929	1933	1982	1971	2058	2058	2071	2168	2170
% No occurence	59,5%	60,6%	60,8%	62,3%	62,0%	64,7%	64,7%	65,1%	68,2%	68,2%

The Impact of the Covid-19 Crisis on the Labour Market. Implications Of Employment Policy

Ioana Manuela MINDRICAN (mindrican.ioanamanuela@gmail.com)
Doctoral School Economy I, Bucharest University of Economic Studies

Elena Florentina MATEI[1] (matei.elena96@gmail.com)
Doctoral School Economy I, Bucharest University of Economic Studies

ABSTRACT

The onset of the health crisis generated by the Covid-19 pandemic led to the transformation of the labor market by changing the dynamics between employers and employees, but also by introducing a new way of working, which although it was known until the crisis, it was not as wide-spread. The pandemic drew attention to the level of training of employers in terms of their ability to manage major problems in a short time. In this context, the health crisis has forced a change in the way of working not only in Romania, but globally. A large part of the employees were able to work from home, and the company quickly adapted to the new changes to come to the aid and protection of the employees. Thus, a series of structural changes occurred, which in the employer-employee relationship involved a new type of benefits and a much more efficient communication. As the impact of the pandemic was strong, the labor market was severely affected, which will result in a subsequent long-term recovery. The impact of the pandemic was felt by the temporary closure of some sectors of activity, the reduction of population movements, the loss of jobs, and implicitly the increase of the unemployment rate. All these things have a strong negative impact on the economy and the population by reducing tax revenues and increasing budget expenditures that will result in a strong increase in the country's indebtedness. Finally, the pandemic has led to an increase in poverty and social and regional disparities. An effective solution for improving labor market conditions has been to focus on digitization, which has made this period easier by accessing resources to the population and continuing professional and development activities.

Keywords*: COVID-19, unemployment rate, regional disparities, digitization*
JEL classification*: J24, J38, O15, G01*

[1] Corresponding author: Elena Florentina MATEI, Doctoral School Economy I, Bucharest University of Economic Studies

INTRODUCTION

The health crisis caused by the Covid-19 pandemic is a global challenge, as it has spread rapidly from one state to another regardless of the level of development. At European level, the authorities have responded quickly to support Member States, but have also shown solidarity through global efforts to reduce and combat the negative effects of the pandemic. In this context, the President of the European Council, Charles Michel, stated that "only a common spirit of global solidarity and responsibility will overcome the crisis caused by Covid-19", European Council (2022). The Covid-19 pandemic was not just a simple global health crisis, but it is much more complex because it has economic, social, and political implications. Member States have intervened by adopting unprecedented budgetary and political measures to improve the public health system and to help the population and economic sectors severely affected by the crisis. The economic impact of the current health crisis varies from one sector of activity to another depending on the level of exposure of the population, depending on several factors such as dependence on production processes, stocks, or the possibility of adapting to supply chain disruption. An unprecedented decision by the European Commission in the context of the new economic situation is to activate the general derogation clause of the Stability and Growth Pact on the possibility for Member States to deviate from budgetary requirements to be able to adequately address the new economic problems. In this context, President Ursula von der Leyen said, "Today we are proposing maximum flexibility in our rules, which will allow national governments to come to the aid of all health systems, staff, and people so severely affected by the crisis. I want to make sure that we respond as best we can to the human and socio-economic dimension of the coronavirus pandemic", European Commission (2022). These new measures have been an important source of support for the labor market, which has been severely affected by the new economic context, where a large part of the population has lost their jobs and companies have been forced to reduce their activities. For example, European states can provide financial assistance from the European Union of up to 100 billion euros in the form of money loans through which they can respond to the rapid increase in public spending on maintaining the population's jobs. Thus, this assistance is provided through the SURE instrument, which is a key element of the Union's strategy for mitigating socio-economic consequences and protecting European citizens. Moreover, this instrument acts as a second line of defense, supporting technical unemployment measures and others of this kind, to help Member States protect their jobs, the self-employed and employees, against unemployment and decrease in income. Romania has benefited from 3 billion Euros through this European support instrument.

The motivation for choosing this topic is represented by the topicality of the problem, having a significant impact on the economic situation of Romania, which is strongly affected by the negative effects of the pandemic that materialized on the one hand by declining jobs following the restriction of activity. Moreover, all these negative effects of the pandemic have determined both globally, but in this case, in Romania, the loss of previously made progress that had a positive impact on the possibility of adopting the single currency in the near future. In this context, the authorities need to focus on promoting policy measures that contribute to economic recovery, and once these things materialize, measures can be considered to support the process of adopting a single currency and achieving the goals assumed upon accession to the European Union.

The main objective of this paper is to identify the impact of the health crisis generated by the Covid-19 pandemic on the labor market that can have significant negative effects on the economy due to increased budget expenditures and declining tax revenues.

This paper is structured in several distinct parts, such as (i) the first part includes the introduction, (ii) the second review of the literature, (ii) the third research methodology, (iv) the fourth case study, which aims to analyze macroeconomic indicators on employment, identify the effects of Covid-19 on employment policy and identify the measures implemented by the authorities to reduce the negative effects of the pandemic (v) in the fifth part of future research directions, (vi) and in the last part are found the conclusions and recommendations.

LITERATURE REVIEW

The European Employment Strategy dates to 1997, and through it several common objectives have been set which will help to improve the situation of the Member States through related funding instruments and monitoring processes. Moreover, the main objective of the Europe 2020 Strategy is to create more and better jobs, and in addition, the Commission has proposed new and more ambitious goals in the field of employment, social protection, and skills to strengthen Europe by 2030. Even if the responsibility for employment and social protection lies with the Member States, European Union law also allows for the intervention of European authorities. Additionally, among the main activities, principles, and objectives found in the Treaty on the Functioning of the European Union is the promotion of a high level of employment through the development of a coordinated strategy, especially regarding the creation of a highly qualified, adaptable and prepared workforce, as well as the creation of labor markets that adapt and react to economic change. In this context, in accordance with the clause set out in Article 9 of the aforementioned Treaty, the objective of registering a high level of employment must be taken into account when defining and implementing the activities and policies of the Union. In a flexible and dynamic labor market, the workforce can change jobs in a short time, which stimulates the creation of more productive jobs. In this context, the dynamism of the labor market helps to improve employment, while the barriers found in formal employment contribute to the relocation of employees to informal sectors that are less productive or even inactive, Kuddo (2009). Moreover, public policies on the labor market must include the unemployed and the employed population, but also the reactivation of the labor force by promoting policies that contribute to stimulating reintegration into the labor market, Șerban (2013). The triggering of the health crisis generated by the Covid-19 pandemic had a negative impact on the labor market by restricting economic activity and limiting the movement of the population, Vasile et al. (2020). At European level, restrictions can have asymmetric effects on the labor market, affecting the most vulnerable workers, Perez et al. (2020). Regarding the impact of the Covid-19 crisis on the labor market, it was a strong one, with the European Commission highlighting several aspects, including the high risks to the forecast and the accentuation of employment imbalances.

Moreover, the European Commission draws attention to the risk of widening development and relaunch divergences between European states in the event that national policy responses at the national level are not sufficiently coordinated or if there is no common policy response at European level. Another risk identified by the Commission is the triggering of strong and long-lasting changes in global value chains and international cooperation, which can severely affect the labor market in open and underdeveloped economies, which are also dependent on trade international, as is the case of Romania, the European Parliament (2020). Additionally, in a recent research conducted by the IFO Institute (2020), it was found that the reduction of investment is the main problem with a strong impact on national economies. Moreover, at the time of the study, there was the problem of a long period of economic recovery in the labor market in terms of increasing income inequality and poverty, increasing unemployment, increasing population debt, and declining purchasing power. At the same time, such perspectives generate pressure on temporary migration for work abroad, a phenomenon that occurs in Romania and which attracts not only seasonal workers, but also the laid-off population or small farmers who are affected by the lack of demand or inability to sell products online that has grown since the Covid-19 pandemic broke out. The Covid-19 pandemic has led to a change in the global paradigm, with a significant impact on the economy and sustainable development scenarios. Thus, in the vision of the authors Contipelli and Picciau (2020), the pandemic accentuated and highlighted the weakness of the current government, the fragility of education and health systems, poverty and the low level of international cooperation. In terms of the impact on the labor market, the crisis has brought, in addition to well-known disadvantages, a number of advantages, such as the possibility to study and work much easier in any part of the world which has a positive impact on the labor market, through developing skills and learning new things that until now were harder to achieve.

The effects of the crisis were strong on both labor supply and demand, as in the acute phase of the pandemic it was decided to segment the labor force by profession and its role in reducing the negative effects of the pandemic. At the same time, the pandemic accentuated the need to carry out economic activities in telework, which was not widespread. As regards the preconditions for employment policy at European Union level, in 1989, the Member States of the European Union, except for the United Kingdom, adopted by means of a declaration, the Community Charter of the Fundamental Social Rights of Workers known in the literature and under the name of "Social Charter". Since its inception, the Charter has been a political instrument made up of "moral guarantees" and the aim of which is to ensure that people's social rights are respected and respected in the Member States. These rights mainly concern the labor market, equal opportunities in terms of obtaining employment in accordance with the studies and needs of individuals, as well as training and professional development. At the same time, this instrument includes a request to the European Commission to submit a proposal to transpose the content of the Charter into legislation, so that social rights on the labor market are truly correlated and connected with the economic context of each state. Subsequently, the Charter was followed by various social action programs. The Treaty of Amsterdam was developed in 1997 and included the "Social Policy Agreement", which had increased provisions in particular in the social chapter of the European Community. At the same time, this treaty plays a key role in the development of European employment policy, as it has created a legal basis for equal opportunities for women and men in the workplace and clarifies to measures against social exclusion. Finally, it can be argued that a reference to fundamental rights has created a new dimension to the objectives of social policy. Another aspect is that this treaty covers both social policy and employment policy, thus achieving the most important pact aimed at economic stability and growth, ensuring the necessary balance between economic integration and employment policy. In April 2005, the European Commission published a document on the implementation of the Lisbon Strategy in line with the changes made in 2003. The plans for this strategy focus on three key directions, depending on micro-economic, macro-economic, and employment priorities. Regarding employment, these plans envisage the replacement of the current National Employment Plans, which will be implemented through their integration into the Lisbon national program. Next, the integrated guidelines will be presented to achieve sustainable economic growth on the one hand, and increase employment for the citizens of the Member States of the European Union on the other. Thus, macroeconomic guidelines consist of (i) safeguarding economic sustainability through the efficient allocation of resources, (ii) security of economic stability, (iii) promoting high coherence between structural and macroeconomic policies, (iv) efficient development of wage income so that they contribute to macroeconomic stability and sustainable economic growth, (v) contributing to a functional and dynamic EMU. The microeconomic guidelines consist of (i) expanding internal markets, (ii) developing open and highly competitive markets, (iii) creating and developing a business-friendly environment, (iv) consistently promoting entrepreneurship among young people, (v) creating and developing a favorable environment for small and medium-sized enterprises, (vi) increasing investment in research, development and innovation, (vii) facilitating access to innovation and increasing ICT development, (viii) encouraging resource efficiency; and strengthening the synergy between economic growth and environmental protection, (ix) forming and implementing a solid industrial base. Regarding to the employment guidelines, they relate to (i) the implementation of employment policies that have been established by the European Commission to achieve a full employment rate, improve the quality and productivity of work, strengthen social and territorial cohesion, (ii) promoting and supporting lifelong learning, (iii) ensuring proper integration into the labor market and employment for jobseekers and the disadvantaged; , (iv) effectively combining labor market flexibility with job security and reducing market segmentation, (v) increasing volume and improving investment in human capital, (vi) streamlining the link between demand at labor market with the existing supply, (vii) adapting and improving the education system in line with the new requirements that are being registered at European level. The above-mentioned directions for employment will be covered by three broad guidelines, which were drawn up mainly in 2003, as part of the evaluation of the Lisbon Strategy. These consist of (a) improving the capacity of employees and enterprises to adapt to the requirements of the labor market and (b) attracting as many employees as possible to the labor market and improving the social protection system. In the current context, taking into account the effects of the 2008-2009 crisis and the COVID-19 pandemic, it is particularly important to achieve the target of total employment while reducing unemployment and inactivity by increasing demand and supply in the labor market. This goal must be strongly correlated with the increasing attractiveness of jobs and the quality and productivity of work. From a macroeconomic point of view, increasing employment is the most effective way to generate sustainable economic growth and at the same time promote social inclusion. In addition, the approach from a new perspective of work, one known in the specialized economic literature as “lifecycle approach to work”, as well as the modernization of social protection systems, will play a fundamental role in the future in the context of the declining population during the work. However, special attention must be paid on the one hand to the problem of employment differences between women and men, as well as to the declining employment rates among the elderly and the elderly. young people. Taking these aspects into account, one can state the idea that another important guideline is that of jobseekers. In this case, it is particularly important to facilitate people’s access to job vacancies, to have constant information on the labor market, through all these aspects to increase their employability.

RESEARCH METHODOLOGY

This part of the presentation of the research methodology aims to facilitate the completion of the following parts of the paper by presenting the methodology and the database used, but also the sources of information. This paper is conducted using a mixed research methodology, as it is based on qualitative and quantitative data. More precisely, through quantitative data, the macroeconomic analysis of the case study was performed, these data being introduced in the Microsoft Excel program to make graphs that allow easier identification of the evolution of macroeconomic indicators on employment, unemployment rate and migration. In terms of qualitative data, they consisted of the method of descriptive analysis used for the review of the literature, and in this stage were used several platforms including Enformation, which includes a diverse base of platforms such as Scopus, ProQuest, ScienceDirect and many others.

To carry out this work, the database of various institutions at European level was used, such as Eurostat and the Ministry of Labor and Social Protection on the macroeconomic indicators mentioned above in Romania for 2019 and 2020. The analysis includes this period to be able to identify the impact of the 2020 pandemic compared to 2019 in terms of labor market macroeconomic indicators.

CASE STUDY

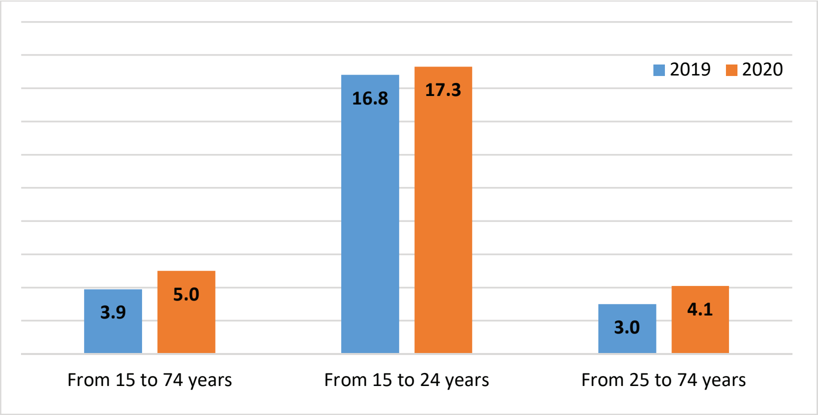
1. The evolution of the unemployment rate in the period 2019-2020 in Romania

The onset of the health crisis had a negative impact on the labor market by severely restricting economic activity, which led to job losses for a large part of employees, which led to an increase in the unemployment rate in 2020 compared to 2019. The emergence of the pandemic affected the population in Romania regardless of age group, according to the attached graphic representation. For example, the unemployment rate for the 15-24 age group increased by 0.5 percentage points, for the 15-74 age group by 1.1 percentage points, while for the 25-74 age group it also increased by 1.1 percentage points.

The authorities quickly intervened in the economy to support the population by ensuring technical unemployment, the possibility of employees entering the telework, and other types of aid, but nevertheless some of the employees went into unemployment. This situation does not have a favorable effect on the economy, as it leads to an increase in public spending and a reduction in tax revenues by lowering budget revenues from contributions and taxes. Moreover, this situation is not favorable for the population, because the incomes received from unemployment benefits are much lower compared to the incomes obtained from salaries, which determines the worsening of the living conditions of the population and the increase of the risk of poverty and social exclusion.

Evolution of the unemployment rate in the period 2019-2020 depending on the age class

Chart no. 1



Source: Own processing based on data taken from the Eurostat website

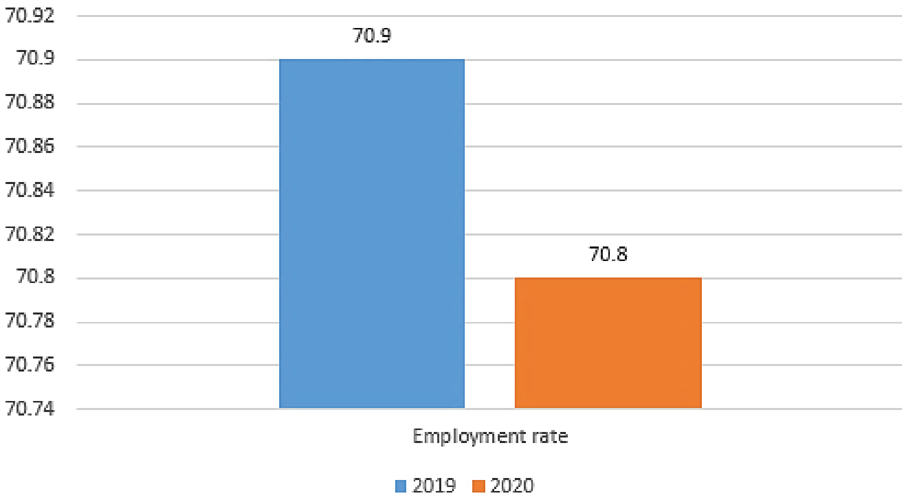
2. The evolution of employment in the period 2019-2020 in Romania

The tightening of economic activity, starting with March 2020, has brought with it a series of negative and high-intensity effects on both the demand and supply sides, which are transmitted through several channels, such as the decrease in final demand. and the demand for labor in the fields and activities dependent on and correlated with international flows of raw materials and consumables, which is influenced by the blockade of normal supply of highly complex supply chains established in the global economy. Another effect is the gradual decline in consumption and the blocking of investment, as the economic situation in the near future is quite predictable. Under these conditions, we can state the idea that Romania has faced a pandemic recession of the labor market, because the main shock could be identified in the labor market, but in the context of maintaining the technological and physical characteristics of the fixed capital stock.

According to the attached graph which highlights the evolution of the employment rate in 2019-2020 in Romania, we can see a decrease in the employment index, this being accompanied by a decrease in labor availability, because many areas have restricted activity. The notable effects of the Covid-19 pandemic could be observed on both the demand side and the labor supply side, depending on the industry. An example can be the shortage of labor in the medical sector, taking into account the emergence of an increasing number of patients infected with various complications, which has led to the management of the whole situation in optimal conditions, with considerable efforts by all medical staff. At the same time, it was necessary to develop and implement measures to improve the management of medical units and facilitate the process of hiring additional staff. At the same time, the analysis of the impact of the pandemic on employment in the context of the current crisis highlights not only the limitations and dysfunctions related to the employment model, but also the manifestation of opportunities that can gradually lead to change both at the organizational level and at the technological level.

Evolution of employment in the period 2019-2020

Chart no. 2

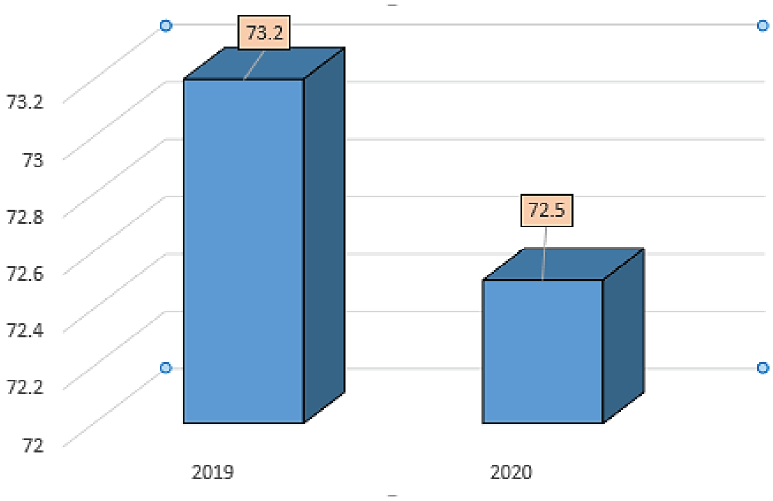


Source: Own processing based on data taken from the Eurostat website

Consistent with the data presented in the graph below, it can be seen that the European Union average in terms of employment rate decreased in 2020 compared to 2019 amid the onset of the COVID-19 pandemic. Prior to this crisis and the impact it generated, the recovery of the labor market led to an employment rate close to the target set for 2020, namely, 75%. As a result of previous experiences and lessons learned from other crises, one of the most important ways to recover the economy quickly and efficiently is to strengthen skills and keep workers in touch with the labor market. An important aspect is the fact that the level of employment has shown a very slow growth trend, especially in the case of jobs that are paid at an average level, this being achieved during the period of employment recession. Another change that occurred and could be observed at European level in 2019 compared to 2020, was the increase in the number of precarious jobs for certain groups, such as workers with “other types of contracts” or people who work “without contracts”. In view of the spread of the Covid-19 pandemic, the above-mentioned categories of workers have been most affected by the crisis and are therefore at very high risk of poverty. Additionally, the increase in various forms of atypical work has led to deep divisions in the European labor market between employees who receive adequate protection and those who have limited access to social protection, but also to the rights related to the employment process at work. The gradual increase in the number of precarious jobs mainly requires solutions that are materialized through policies that are closely in line with the needs of the market and those of employees. This is therefore becoming particularly important in the context of the emerging impact of the pandemic, which carries major risks for employees in precarious jobs, as well as for those who work independently.

The evolution of employment in the period 2019-2020 at the level of the European Union

Chart no. 3



Source: Own processing based on data taken from the Eurostat website

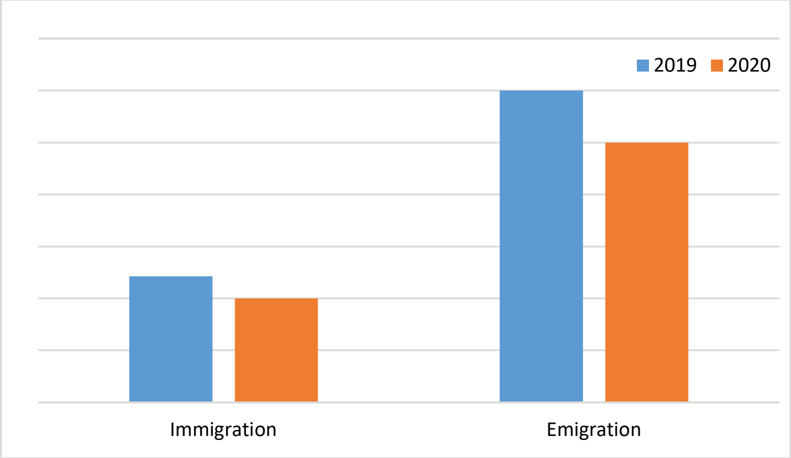
3. The evolution of migration in the period 2019-2020 in Romania

International migration is a phenomenon that has negative effects on the economies of all countries worldwide. As for Romania, it is not only a country of origin, but also one of destination and transit. The evolution of migration over the years has been an upward trend due to the free movement of the population and better working and living conditions in different countries compared to the country of origin, which drives the population to take this step. The term international migration is used by the United Nations (2019) to refer to the population living in a different region or country from the one in which they were born or to which they belong. The data for 2019 were taken from the Eurostat website, while the data for 2020 are estimates of the authorities, because at the time of writing this official data has not yet been published.

The outbreak of the Covid-19 pandemic severely affected all forms of human mobility, including international migration. The closure of national borders, together with the strong disruption of international travel by land, sea, or air, has forced the population to cancel or postpone their plans to move outside their country of origin. Moreover, a large part of the migrant population remained stranded, unable to return to their country of origin, while others were forced to return to their country of origin earlier than planned due to the loss of places of residence work or school closure. While the Covid-19 pandemic has caused major changes in migration flows by 2020, the number of international migrants has increased considerably over the years.

The evolution of international migration in the period 2019-2020

Chart no. 4



Source: Own processing based on data taken from the Eurostat website

4. The effects of the Covid-19 pandemic on the Romanian labor market

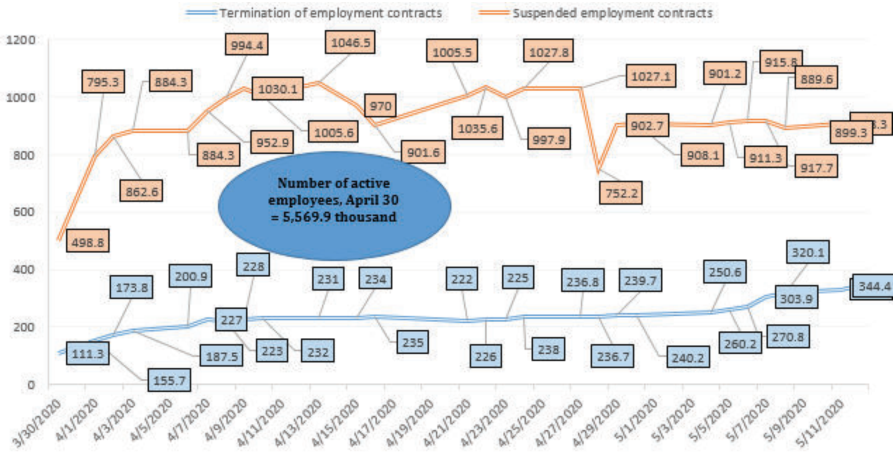
The Covid-19 pandemic proved to be a particularly large phenomenon with major implications on various sectors of the Romanian economy, especially on the labor market. In Romania, the level of the poverty rate has remained at a high level, including during periods of economic growth. An example can be represented by the fact that in the period 2013-2016, about a quarter of the total population had very low incomes, so that it was impossible for them to reach an acceptable standard of living, this situation ranking Romania among the last countries in Europe. According to the information published by Eurostat, Romania has the highest rate of persistent poverty, at about 19%, which highlights the fact that one in five people were in poverty in at least two of the previous three years. For these reasons, in Romania there are approximately 7 million people at risk of poverty, respectively, social exclusion. An important aspect is the fact that the periods of high economic growth and the increase of jobs have failed to reduce the number of poor and socially excluded people. This change was one million in four years, a difference recorded in the period 2013-2017.

Given the negative effects of the pandemic, it can be said that that 1 million people can return to poverty, especially because they are not economically active, without adequate and well-paid jobs full-time employment contracts for an indefinite period of time. At the same time, the most affected age group was between 25-49 years old, because approximately 2.5 million people were at risk of poverty, more precisely 35% of the total of 7 million poor people. This is due to the fact that very low labor intensity is considered a widespread phenomenon. In February 2020, the people who benefited from a guaranteed minimum income were 169,907 people, more precisely 3.5% of the total number of people in poverty. The monthly value borne by the budget for persons receiving social assistance is approximately 45 million lei. Hence, the idea that the main solution to poverty reduction is the creation and development of new jobs, which are permanent and based on employment contracts, a process that can be stimulated with the help of government programs in support of the private sector. However, given the current context, as a result of the impact of the Covid-19 crisis, it is quite difficult to implement such a measure, as many jobs have dwindled, the market has shrunk and focused only on key areas, and the incomes of disadvantaged categories have decreased considerably.

According to the information published by the Labor Inspectorate, in the first 3 months of 2020, the number of employees in Romania decreased by 904.4 thousand from the value of 6,474.3 thousand in December 2019 to 5569.9 thousand people in March 2020. Thus far, the information presented by the Ministry of Labor and Social Protection gives us the opportunity to have an overview of the impact that the Covid-19 pandemic has had on employees and the main areas of activity, in terms of the number of contract work that has been suspended or terminated. In contrast to the number of active employees on March 30, 2020, of 5,569.6 thousand people, between March 30 and May 12, 2020, as a result of the propagating effects of the pandemic, the trend of increasing the number of contracts can be seen according to the attached graph of the work that were terminated, more precisely their value increased from 111.3 thousand to 344.4 thousand.

Number of employment contracts terminated and suspended during 30.03.2020-12.05.2020

Chart no. 5

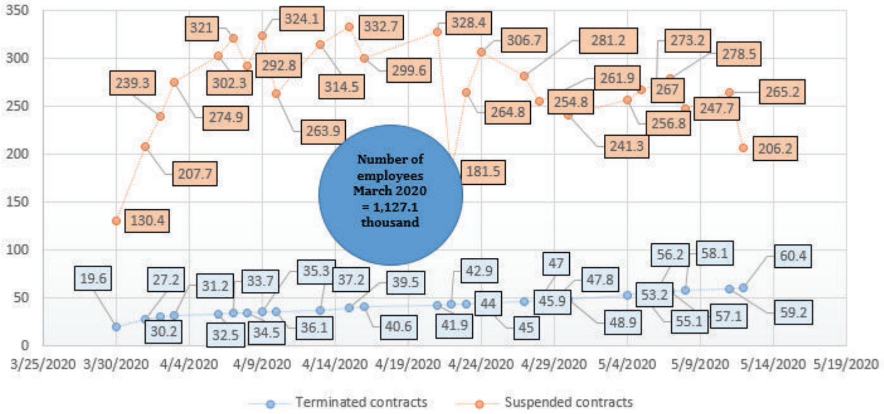


Source: Own processing based on data provided by the Ministry of Labor and Social Protection

Furthermore, the number of suspended contracts registered the highest value on April 10, 2020, were 1046.5 thousand. Additionally, during the analyzed period, the values fluctuated considerably, having a decreasing trend at the end of the presented interval, more precisely their number was 899.3 thousand. In terms of areas of activity, most suspended employment contracts could be identified, in particular in the manufacturing industry, wholesale trade, car repair, hotels, and restaurants, which account for more than 50% of all suspended contracts.

Number of employment contracts suspended and terminated in the manufacturing industry during 30.03.2020- 12.05.2020

Chart no.6



Source: Own processing based on data provided by the Ministry of Labor and Social Protection

Taking into account the situation of suspended employment contracts, respectively, those terminated as a result of the effects propagated by the COVID-19 pandemic, the effects have been and continue to be increasingly complex, leaving deep imprints on various sectors of the economy. According to a study carried out by Eurofound, the idea was reached that people who have become unemployed as a result of the COVID-19 crisis, unfortunately, suffer from financial insecurity, which is considered the most acute and profound, as it unquestionably highlights the “need for protection in the context of the existence of the concept of physical distance and the diminution of social contracts”. However, in Romania in the current context, there are significant categories of labor force that are much more vulnerable compared to those mentioned above, to the effects of the crisis. One of the most important categories of people vulnerable to the current crisis is the structure of the level of education of the employed population, because this category, at any decrease in income, can enter the risk of poverty. This category includes people with a low level of education and training, who are generally engaged in seasonal paid work or even without registered employment contracts. According to the information published by Eurostat in 2019 in Romania, the employed population in the aforementioned category was 1,423 thousand people, a very high number taking into account the fact that the other social categories are not in a positive outlook. The distribution of the employed population according to professional status shows the very high share of the two categories, which were not greatly affected by the propagating effects of the pandemic, especially those of social isolation and abolition of companies, more precisely self-employed and of course, unpaid family workers, which in 2019, represented almost 22.1% of the total employed population in Romania and amounted to approximately 1.9 million individuals, according to Eurostat. The Covid-19 pandemic, which broke out in the first months of 2020 globally, has a severe impact, including on the Romanian economy, with an emphasis on the labor market and the level of employment. Given the deterioration of the financial and macroeconomic framework, the proportions of which are quite difficult to anticipate and manage in terms of its size, timing, and depth, and despite the anticrisis measures that have been developed and implemented by the issuing authorities of this field of activity, in the short and medium term, the Romanian labor market will face increases in the unemployment rate. This change will be felt not only in 2020, but also in the coming years, as a result of the restriction of activity in several areas of activity, which clearly relaxes the tensions that are manifested in the labor market and which have become severe in many branches in the period preceding the outbreak of the external shock of the Covid-19 pandemic. However, in the context of the existence of a favorable situation from a macroeconomic point of view and the restart of the Romanian economy that would allow on the one hand a rapid recovery, and on the other hand a recovery as close as possible to the levels recorded in the precrisis period. It is also likely that in 2021, the specific problems of the labor market, as well as the tensions accumulated at its level, will decrease and have a smaller amplitude compared to 2020.

Taking into account the information presented above, one of the characteristics of Romania today, but also of other Eastern European states, members of the European Union refers to the existence of massive influxes of emigrants who returned to their country of origin due to the effects propagated by the Covid-19 pandemic. In the circumstances in which approximately one million people who emigrated back to Romania in the first months of 2020 and hoping that the decision-making authorities will manage this in optimal conditions to improve the conditions in the labor market, these people would remain on the territory of the Romanian state. One of the benefits of this situation is the mitigation of qualitative and quantitative deficits that have occurred in recent years in the labor market, especially in those sectors of the economy where staff do not require higher education, namely food and agriculture, and this contributes to maintaining the economy in functional parameters according to the criteria established by the European Union. Otherwise, in the context in which Romanians decide to emigrate back to the states where they were employed, this will amplify the existing tensions in Romania both economically and socially.

5. Measures adopted by the authorities in response to the Covid-19 pandemic on the Romanian labor market

The COVID-19 pandemic and the measures taken by policy makers to limit the spread of the virus have led to major contractions in gross domestic product, which will at some point lead to appropriate adjustments in the labor market. The elaboration and subsequent implementation of measures to support economic agents, more precisely their companies and employees affected by the new economic context, was one of the priorities of public policies that were adopted during the crisis. However, their long-term timeliness and sustainability remains a key issue given the level of uncertainty about the path back to the pre-pandemic situation, with a major risk of inefficient allocation of budgetary resources.

Taking into account the situation in Romania, the public authorities focused on measures to encourage the maintenance of labor relations, and in this regard offered financial support to companies in the economy. The initiative implemented by public authorities is similar to the general trend that has been registered at European level, more precisely to perform daily work with a reduced schedule, this being a measure belonging to the social security system through which employers are allowed for a short period of time to adjust its workforce in the context of the deepening recession. Specifically, it aims to reduce the number of hours worked or even the total, which will later lead to the suspension of the employment contract and the provision of technical unemployment, in both cases, promoting the adoption of the European model until overcoming the shock of the crisis. Additionally, workers' unemployment benefits are calculated as a percentage of the employment salary they obtain and are borne by the state. The main objective is to avoid redundancy, which is an adjustment by extensive margins and on the other hand to mitigate the loss of income of employees while adapting the work schedule to disadvantageous economic conditions and of course, with the benefit of retaining already qualified and familiar workforce in the operations that are carried out within the companies. Under these conditions, companies have the opportunity to retain human capital, avoiding the costly redundancy process that occurs during the recession, followed by reemployment and training for the return period. Initially, economic theory aimed at the preference of economic agents to adapt the production process to the short-term fluctuations of demand through the quantity achieved by the labor factor, in the context in which capital is fixed and more difficult to adjust. Over time, the situation has proved to be paradoxically different from economic theory, because when a shock occurs, economic operators tend to resort to an intensive margin, namely, to reduce the use of labor. Recourse to an extensive margin, it to adjust the workforce by reducing jobs, is done only after a certain period of time, more precisely when companies perceive that the changes in activities are permanent, and this strategy becomes much more expensive, because it involves expenses incurred by the dismissal process followed by the search, employment and finally, the training of new employees. Another strategy available to companies is salary adjustment, but this is difficult to implement, because in the context of rigidity in declining incomes, which is caused by institutional factors such as minimum wage policy or coverage through collective agreements, as well as economic ones that focus on the negative impact of wage adjustments on labor productivity. Although the labor market and its specific issues, such as unemployment and migration, remain a widely debated topic at European level, it is still a rather vague concept. A essential theory belongs to Pissarides and was presented in 1997, referring to the capacity and speed with which the labor market absorbs the shocks that occur in the economy, aspects most often assessed and analyzed based on the employment rate the work. An eloquent example can be represented by the gradual flexibility of the US labor market, which refers to looser regulations regarding the employment process, respectively, dismissal has gradually induced a change in the behavior of economic agents over the years, another measure being moreover, the adjustment of the labor factor to the occurrence of shocks with the help of the extensive margin thus becoming much faster Fernald (2015). Hence, the idea that the coordinates are different at European level from the perspective of the labor market, because this time the focus is on "flexicurity" measures, the latter focusing on combining the need for flexibility of employers with the security of the place which is indisputably important for employees. In the current context, in the short term, European companies have resorted to a major role to a strategy of reducing labor costs with the help of the intensive margin, more precisely the reduction of the work schedule.

An important aspect is the fact that the adjustment of the labor market in relation to the shocks that may occur in the economy can have significant implications on the duration and amplitude of the economic cycle, the latter influence being the factor that triggers the use of reduced work schemes. However, their erroneous calibration can lead to inefficient use of resources economically, thus limiting the transfer of labor from firms that are less productive to performing. Taking into account the situation in Romania, the Ministry of Labor decided that employees should benefit from technical unemployment in the context of the existence and development of the pandemic, and their employment relationships are maintained after the resumption of activities by employers. These persons will benefit for a period of approximately 3 months, with the help of the employer, from the payment of 41.5 percentage points of the total basic salary in accordance with the job they have, but not more than 41.5 of the total gross earnings provided in the

Law on the state social insurance budget issued in 2020, more precisely no. 6/2020. To apply this measure, companies have the obligation to maintain the employment relationships of employees until 31 December 2020, except for seasonal work and situations in which the termination of the individual employment contract arises from reasons not attributable to the employer. Moreover, the above measure applies in particular to persons who have had a period of suspension of the individual employment contract of at least 15 days related to the period of emergency or alert. At the same time, the executive decided to stimulate the employment of people over the age of 50 whose employment relationships ceased for various reasons not attributable during the state of emergency or alert and who are also registered as unemployed in the records of territorial agencies for employment. Another measure taken is to reduce the working hours from five days a week to four days, as mentioned in this article, of course, at the same time as lowering the salary, but this has been implemented only in the areas where it has been registered a temporary decline in activity, such as the tourism and hotel sector. In the context of very serious situations in terms of declining profits or turnover, this means the suspension of employment contracts, without the payment of any compensation to the affected staff. An important aspect to be mentioned is that to determine whether a company can suspend the employment contracts of the staff due to the propagating effects of the pandemic, a detailed analysis of the respective conditions was necessary, and subsequently the need was established for the adoption and implementation of such a measure. Additionally, the Covid-19 pandemic left deep imprints on various sectors of the economy, more than 875,000 Romanians lost their jobs, as the number of suspended employment contracts doubled from July 15, 2020 to June 1, 2020, as a result of the negative effects of this process.

CONCLUSIONS

As a result of the research carried out in this paper, the idea was reached that the crisis caused by the Covid-19 pandemic affected the entire workforce in Romania, but in different intensities, more precisely not only those active in terms of legally on the labor market or those who work in the “gray” area of the Romanian economy, but also important segments of the population that depend on them. Regarding to employees, both those belonging to the public sector and those belonging to the private sector have materialized through the obligation to comply with certain measures, such as social distancing and changing the work regime by introducing work at home and reducing working time. The decrease in the number of active employees by over 900 thousand people in April 2020 compared to the end of 2019, the suspension of a significant number of individual employment contracts, which reached a maximum of about 1 million during the period considered, as well as the gradual increase of the number of employment contracts concluded at a maximum of over 340 thousand in the first part of May 2020 represent some essential changes of the COVID-19 pandemic on the labor market, but which had a significant economic and social impact. In terms of the areas of activity, the most affected were the manufacturing industry, retail and wholesale trade, repair of motor vehicles, motorcycles, and the hotel and restaurant sector, which concentrates more than 50% of all contracts that were suspended economically in Romania.

At the same time, the Romanian labor market does not only imply the existence of processes specific to employee turnover, they have, in addition to the general characteristics of employment in the market economy and the particularities resulting from the structure of the economy, and its level of development, among which two are particularly important for setting up anticrisis measures in the context of the existence and spread of the COVID-19 pandemic, namely the high share of activities that are directly unpaid, in a contractual basis for family workers, and the high share of wage labor that carries out without the existence of written and registered employment contracts, namely, informal and, finally, unpaid salary work. An important aspect to mention is that these categories together with self-employed workers, are the most vulnerable groups to the current crisis. At the same time, it is about 2.2 million people who depend on other people or the state or who depend on other incomes as well as households. Regarding the particularities of these categories of people, they have a low level of education, the lack of an employment contract, the occasional nature of the activities performed, and the significant lack of forms of protection, and social assistance. The analysis presented in this paper reveals sufficient arguments that the Covid-19 pandemic has and continues to have a disproportionate impact, and the most exposed people are vulnerable groups, namely people in the informal area of the economy, individuals working in the most affected areas, as well as workers with a low level of professional qualification.

Additionally, the negative effects of the pandemic on the economic situation of Romania determined the reduction of the progress registered in the previous years that contributed to the fulfillment of the necessary conditions for the adoption of the single currency, an objective assumed with the accession to the European Union. All measures implemented following the outbreak of the pandemic have led to a sharp reduction in tax revenues and increased budget expenditures that have implicitly widened the fiscal gap, and in this context the authorities must focus on implementing measures to resume the process of economic growth and subsequently measures to support the process of joining the Eurozone.

REFERENCE

1. **Boumans, D., Link, S. & Sauer, S**. (2020), *Covid-19: The world economy needs a lifeline - but which one ?*, EconPol Policy Brief, Vol.4, https: // www .econpol.eu / sites / default / files / 2020-04 / EconPol_Policy_Brief_27_COVID19_Economy_Lifeline. pdf
2. **Chivu, L. & Georgescu, G**. (2020), *Labor market and employment vulnerabilities in the context of the COVID-19 pandemic. Possible solutions*, Romanian Academy, National Institute of Economic Research "Costin C. Kiritescu", https://acad.ro/SARS-CoV-2/doc/d04-Vulnerabilitati_ale_pietei_muncii.pdf
3. **Chivu, L. & Georgescu, G**. (2020), *Labor market vulnerabilities under the COVID-19 impact in Romania*, Munich Personal RePEc Archive, Paper no. 101676, https://mpra.ub.uni-muenchen.de/101676/1/MPRA_paper_101676.pdf
4. **Contipelli, E. & Picciau, S.** (2020), *Post Covid-19 rebuilding our paradigms through sustainable development goals and the sufficiency economy philosophy*, https://www.researchgate.net/publication/343079841_Post-COVID-19_Rebuilding_Our_Paradigms_Through_Sustainable_Deviciency
5. **European Commission** (2017), *Active Labor Market Policies*, https://ec.europa.eu/info/sites/default/files/file_import/european-semester_thematic-factsheet_active-labour-market-policies_en. pdf
6. **European Commission** (2022), *Coronavirus: Commission proposes to activate fiscal framework's general escape clause to respond to pandemic*, https://ec.europa.eu/commission/presscorner/detail/en/ip_20_499
7. **European Council** (2022), *The COVID-19 coronavirus pandemic: the EU response*, https://www.consilium.europa.eu/en/policies/coronavirus/
8. **European Parliament** (2020), *The economy and coronavirus: weekly picks*, https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/645745/IPOL_BRI(2020)645745_EN.pdf
9. **Eurofound** (2021), *Employment and Labor Markets*, https://www.eurofound.europa.eu/en/topic/employment
10. **European Parliament** (2021), https://www.europarl.europa.eu/factsheets/ro/sheet/54/politica-de-occupare-a-fortei-de-munca
11. **Government of Romania** (2005), *Labor Market and Employment Policy*, http://ier.gov.ro/wp-content/uploads/publicatii/Piata_muncii.pdf
12. **Gi-Group** (2020), *Impact of Covid-19 on the labor market*, https://ro.gigroup.com/wp-content/uploads/sites/12/2020/09/Impact-Covid-19-pe- labor market.pdf
13. **Kuddo, A**. (2009), *Employment services and active labor market programs in Eastern European and Central Asian Countries*, World Bank SP Discussion Paper No. 0918, https://web.worldbank.org/archive/website01507/WEB /IMAGES/0918.PDF
14. **Ministry of Labor**, *National Strategy for Employment, 2014-2020*, https://mmuncii.ro/j33/images/Documente/Munca/2018/SN_Ocupare_forta_munca_2018.pdf
15. **National Bank of Romania** (2020), *Inflation Report*, https://www.bnr.ro/DocumentInformation.aspx?idInfoClass=3922&idDocument=37812&directLink=1
16. **Perez, ST, Fana, M., Gonzalez-Vazquez, I. & Fernandez-Macias, E**. (2020), *The asymmetric impact of Covid-19 confinement measures on EU labor markets*, Center for Economic Policy Research, VoxEu , https://voxeu.org/article/covid-19-lockdown-and-eu-labour-markets

17. **Șerban, A. C.** (2013), *Public policies targeting labor market rigidities*, Journal of Theoretical and Applied Economics, Vol. XX, Nr. 2 (579), http://store.ectap.ro/articole/831_ro.pdf
18. **United Nations** (2019), *International Migration 2019,* Department of Economic and Social Affairs, https://www.un.org/en/development/desa/population/migration/publications/wallchart/docs/MigrationStock2019_Wallchart.pdf
19. **Vasile, V., Boboc, C., Ghiță, S., Apostu, S., Pavelescu, FM & Mazilescu, R**. (2020), *The effects of the Sars pandemic -Cov-2 on employment*, https://acad.ro/SARS-CoV-2/doc/d19-Efectele_pandemiei_asupra_ocuparii.pdf