Romanian Statistical Review
Revista Română de Statistică

THE JOURNAL OF NATIONAL INSTITUTE OF STATISTICS

2/2022

www.revistadestatistica.ro

CONTENTS 2/2022

Using R to determine the impact of an image on viewers

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

Diana Mihaela JULA, PhD student (diana.jula@yahoo.com)
The Romanian Academy,

ABSTRACT

In this paper, we are addressing some aspects regarding the likeability of an image. Besides the obvious answers as the quality, resolution, size, we suggest that there are also other properties that make people like and remember an image. This study answers the questions like: what is the best time frame to post on social media for the highest impact, is the dominant color, main subject, or the number of subjects having an impact on clients/followers?

We are answering these questions using AI packages like YOLO for classification and object detection, countcolors for findind pixels by color range, and imager for an array of functions for working with image data.

Keywords: *image processing, color detection, object detection*

1. INTRODUCTION

The main objective of this study is to identify main aspects that have an impact on the likability of an image. In order to obtain this, we intend to propose a model/validate a model to evaluate the impact. On the long run, we intend to create a tool (shiny app) to estimate the impact of an image on target audience.

The idea of identifying aspects that make an image memorable started from the fact that today a person is "flooded" with an extensive number of images. According to some studies (https://ppcprotect.com/blog/strategy/how-many-ads-do-we-see-a-day/), the average person is now estimated to encounter between 6,000 to 10,000 ads every single day. So, the need to understand what makes an image memorable to the viewers is more important than ever.

2. METHODOLOGY

To analyze the impact of different elements of the picture on viewers, we decided to start from the well-documented relation between photography and painting. The relationship between the two arts was long discussed, having a lot of people voting for one or another in terms of impressing people or creating a memorable feeling.

Van Gogh once said "I always think photographs abominable, and I don't like to have them around, particularly not those of persons I know and love... photographic portraits wither much sooner than we ourselves do, whereas the painted portrait is a thing which is felt, done with love or respect for the human being that is portrayed."

VINCENT VAN GOGH, Artist, b. 1853, Zundert, Netherlands, d. 1890, Auvers-sur-Oise, France - Letter to Wilhelmina van Gogh, 19 September 1889). In contrast, George Bernard Shaw boldly affirmed that "I would willingly exchange every single painting of Christ for one snapshot" (GEORGE BERNARD SHAW - Writer, critic, and dramatist, b. 1856, Dublin, d. 1950, Ayot St. Lawrence, Hertfordshire, England.)

The purpose of this study is not to argue with these opinions, but to see if some of the techniques used in painting are also valid for photography in terms of good practices for creating a memorable piece of art.

2.1. **What makes a good painting and a good photo?**

According to different experts, there are some aspects that are more or less generally accepted that are qualities of a good painting. Usually, these are grouped by technical aspects, emotional impact and novelty.

What makes a good painting?

Table 1

Technical Aspects	**Personal & Emotional Involvement**	**Novelty**
Composition	Freshness	Innovation
Choice of Colors	Inspiration	Originality
Focal Point	Readiness	Ingenuity
Patterns & Groups		

Source: https://feltmagnet.com/painting/What-Makes-A-Good-Painting

When discussing about good photos, there are also some aspects that pop out:

- Line
- Shape
- Form
- Texture
- Color

- Size
- Depth

Source:https://www.bhphotovideo.com/explora/photography/tips-and-solutions/elements-of-a-photograph-size

Besides these aspects, there are also elements related to picture quality (resolution, color space, size), but also elements that are specific to the platforms used to display images – for example, the time and date when uploading the image. We do not pursue this direction, because these aspects seem to be very platform dependent. We are trying to identify the general elements that make an image memorable. For this purpose, we selected the colors to be element of interest.

When analyzing the colors in photography, we started from the definition:

a: a phenomenon of light (such as red, brown, pink, or gray) or visual perception that enables one to differentiate otherwise identical objects

b: the aspect of the appearance of objects and light sources that may be described in terms of hue, lightness, and saturation for objects and hue, brightness, and saturation for light.

also*: a specific combination of hue, saturation, and lightness or brightness*

c: a color other than and as contrasted with black, white, or gray

Source: https://www.merriam-webster.com/dictionary/color

Hue is simply the description of the color (e.g., blue, red, yellow, etc.). Value is the relative brightness or darkness of a color. Saturation is the intensity or purity of a color. The purest color is a hue with no white, black, or gray added to it.

There are studies that suggest one can expect that:

- muted colors might provoke indifference or even melancholic feelings
- harmonic colors—colors that complement each other—serve to create distinct feelings in photographs
- highlighting the colors will enhance the subject and accentuate the mood of the image

We are not interested in a specific mood that a color can induce to a certain viewer, but the fact that it indeed creates a measurable reaction – we are counting the likes that the image is obtaining.

2.2. Working hypotheses

According to Curalate.com, on social media (Instagram):

- The lighter the image, the better impact (24 percent more likes).
- More background, less close-up (29 percent more likes).
- Bluish, or cold, images fare better than reddish, or warm, images (24 percent).
- Images with lower saturation get more likes than those with more intense colors (18 percent).
- Photos with higher levels of texture are 79% more likely to be appreciated than smoother photos

In a study from 2016 (Schultz C. D., 2016), the author explained the post interactions using variables like post vividness, content, timing, post length, fans and others. Schultz concluded that "the (ln) number of likes is significantly positively affected by picture, hashtag, voting, call to act, contest, coverage and product category, while events, competitions, and post length have a negative effect." There is also an interesting result about the number of fans, namely it has a negative effect on sharing habits.

Yiyi Li and Ying Xie in the paper "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement" (Li Y., Xie Y., 2020) analyzed three effects of image content: mere presence, image characteristics and image-text fit effect. The analysis was conducted on Twitter and Instagram. We are focusing on the results for Instagram (to be able to compare with our findings).

One interesting finding was that pictures with human faces tend to attract fewer likes and lower engagement. The colorfulness of the pictures may have an impact but depends on the product category (in their study, Li and Xie found that color pictures are driving more engagement on travel-related social posts and less on cars). Regarding the text, on Instagram, they concluded that the fit has no positive influence on likes.

When testing one or some of these hypotheses, we are facing a lot of challenges. The most obvious ones are:

- Meeting "caeteris paribus" conditions – same number of followers, same tags, similar audience, same posting time/day
- Continuous changes in social media platforms' algorithm of displaying posts
- Others (e.g., "hot" topics)

2.2.1. **First hypothesis**

Colors' distribution influences number of reactions (likes)

To test this, we used 2 R packages *colordistance* and *countcolors*, the test sample included 62 images form an Instagram account with 3300 followers. The average number of likes/image is 143. The used Instagram account is the author's and it was selected to have access to insight details about posts frequency, audience, promoted posts, invited friends and so on. These details (like locations, age range, gender distribution) are to be included in further analyses.

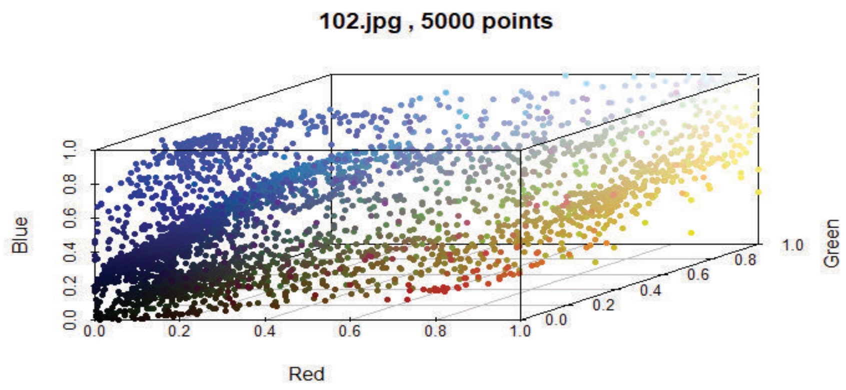
We decided to include only 62 images to keep the sample as consistent as possible (for example, less followers and hot topics variations).

First, we use colordistance. According to the description, this package contains functions for quantifying the differences between colorful objects. Colordistance had different uses and it was validated also in studies like improving the accuracy when quantifying the area of P. destructans infection in bat wing-membranes. (Hooper S., et.all., 2020)

We downloaded the images from an Instagram account and used a recursive method in R to extract the data for each image. We wanted to minimize the effect of different sizes and compression of the original images on our model.

Image 1. 102.jpg, test image

Image 2: *colorcount* result



Using the code *colordistance::plotPixels("102.jpg", lower = NULL, upper = NULL, n = 5000)* we obtain:

The initial linear model tested is
number_of_likes~ colordistance(red)+ colordistance(green)+ colordistance(blue)
but after econometric hypothesis testing, the model resulted was:

Image 3: Econometric model

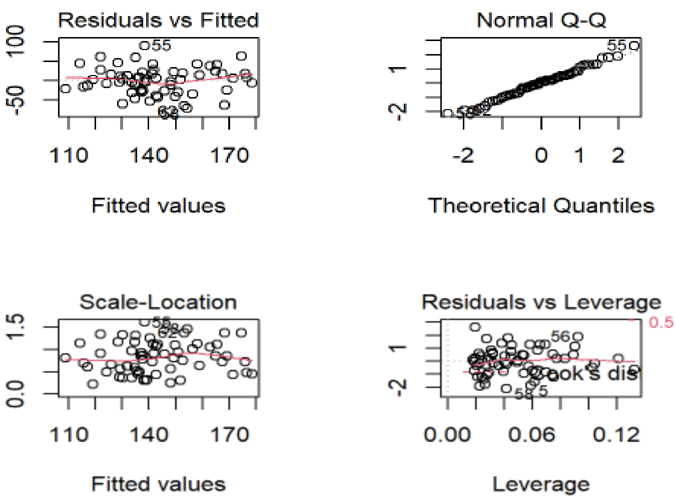
```
Call:
lm(formula = y ~ R + G)

Residuals:
    Min      1Q  Median      3Q     Max
-72.342 -24.763   2.467  18.923  89.597

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   148.76      15.78   9.425 2.26e-13 ***
R             137.54      43.17   3.186 0.002307 **
G            -174.94      47.65  -3.671 0.000521 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.65 on 59 degrees of freedom
Multiple R-squared:  0.1947,    Adjusted R-squared:  0.1674
F-statistic:  7.13 on 2 and 59 DF,  p-value: 0.001685
```

Image 4: Plot diagnostics



The model suggests that RED and GREEN might positively influence the number of likes of an image.

We can use the model to predict the number of likes based on the colors of the image. Also, we must take into consideration aspects like the number of followers. But we suggest to use the model to compare the results with the average number of likes.

Next model included the *countcolors* package, which finds pixels by color range in an image.

Using the code *image.spherical <- countcolors::sphericalRange(image, center = center.spherical, radius = 0.2, color.pixels = FALSE, plotting = FALSE);*

on the Image 1 (102.jpg), we obtain:

Image 5: *countcolors* result

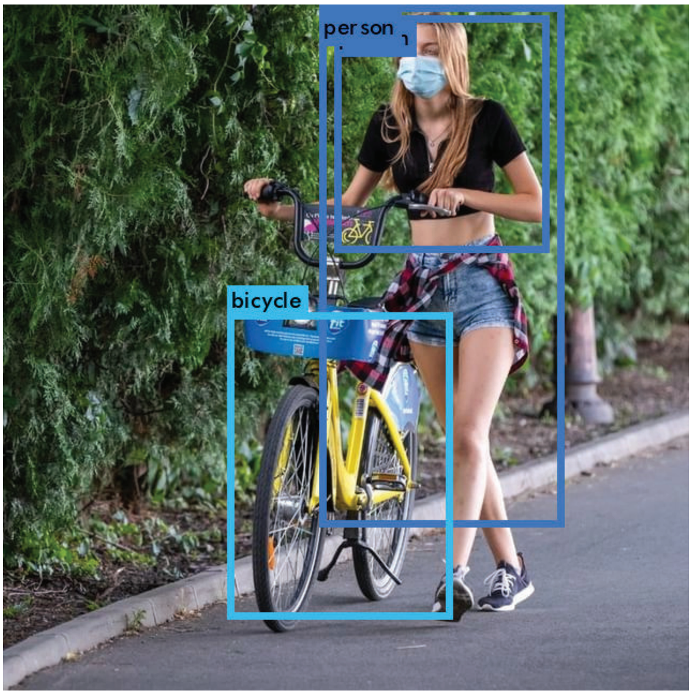
The issue that we were facing using this approach was the high difficulty in calibrating the parameters of the *countcolors* function (the sensibility in detecting color range), in order to create a valid econometric model.

Our next approach included the use of Tiny Yolo - R package *image.darknet*. The functions included in this package allow us to identify different types of subjects.

Image 6: test photos/228.jpg: Predicted in 1.818000 seconds.

Boxes: 845 of which 1 above the threshold.
person: 46%

Image 7: test photos/124.jpg: Predicted in 1.801000 seconds.



Boxes: 845 of which 3 above the threshold.
person: 22%
person: 78%
bicycle: 62%

Even if the model identified the objects with an acceptable level of confidence, it didn't lead to a valid model. We suspect that the main reason for this is the insufficient number of objects in the sample. A future model should be based on more consistent samples.

3. Results and next steps

These initial models are presenting encouraging results. The model based on *colordistance* function resulted in an econometric model able to predict the impact of colors on the number of likes (red and green colors were the ones that generated a boost in number of likes). This model will be further tested on larger samples (bigger accounts in terms of images and number of followers) and on different platforms. We suspect that the unique algorithm each platform uses to display images and the number of users that are able to see a post can highly affect the result.

We also wanted to test the hypothesis that pictures with human faces tend to attract fewer likes and lower engagement suggested by Li and Xi, but the results were inconclusive. More extensive samples should be used in further studies to test this approach.

We are aware of the limitations of this study, mainly based on the reduced number of test sample. So as a future step, we plan to test the model using more images and different accounts (number of followers, audience, posting habits, tags). Also, we intend to test other parameters – quantitative and qualitative and on different platforms – not only social media.

References:

1. **Benve R.**, 2021, *Elements of a Good Painting: Art Tips*, https://feltmagnet.com/painting/What-Makes-A-Good-Painting
2. **Carr S.**, 2021, *How Many Ads Do We See a Day In 2021?* https://ppcprotect.com/blog/strategy/how-many-ads-do-we-see-a-day
3. **Schultz C. D.**, 2016, *Driving likes, comments, and shares on social networking sites: how post characteristics affect brand interactions in apparel retailing*. In Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World (ICEC '16). Association for Computing Machinery, New York, NY, USA, Article 9, 1–9. DOI: https://doi.org/10.1145/2971603.2971612
4. **Hooper S.E., Weller H., AmelonS.K**.; Countcolors, 2020, *An R Package for Quantification of the Fluorescence Emitted by Pseudogymnoascus Destructans Lesions on the Wing Membranes of Hibernating Bats*. J Wildl Dis 1 October 2020; 56 (4): 759–767. doi: https://doi.org/10.7589/2019-09-231
5. https://cran.r-project.org/web/packages/countcolors/vignettes/Introduction.html, last accessed on 10.01.2022
6. https://rdrr.io/cran/imgpalr/man/image_quantmap.html , last accessed on 10.01.2022
7. https://www.merriam-webster.com/dictionary/color ,last accessed on 10.01.2022
8. **Li Y., Xie Y.**, 2020, *'Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement'*, Journal of Marketing Research, 57(1), pp. 1–19. doi: 10.1177/0022243719881113.
9. **Redmon J., Divvala S., Girshick R., Farhadi A.,** 2016, *You Only Look Once: Unified, Real-Time Object Detection*, University of Washington, https://arxiv.org/pdf/1506.02640.pdf
10. **Vorenkamp T.**, 2021, *Elements of a Photograph: Size* https://www.bhphotovideo.com/explora/photography/tips-and-solutions/elements-of-a-photograph-size
11. **Walker R.**, 2016, *Color Quantization in R*, https://www.r-bloggers.com/2016/01/color-quantization-in-r/

Economic Analysis between Gross Domestic Product and Foreign Trade in Romania

Costel POPESCU, PhD student,
The Romanian Academy / National Institute of Statistics Romania

ABSTRACT

Foreign trade is considered a major component of sustainable economic growth. The links between it and the gross domestic product (GDP) have been analyzed in numerous specialized economic works. The developed econometric models have demonstrated the strong connection, both in the short and long term, between these macroeconomic components.

This paper demonstrates once again the long-term and short-term links between these variables by using the VECM (vector error correction) econometric model on the annual GDP, Export and Import data with Romanian agricultural products from the period 1995-2020.

Keywords: *GDP, trade balance, trade efficiency, export, import*

JEL Classification: *F19*

1. INTRODUCTION

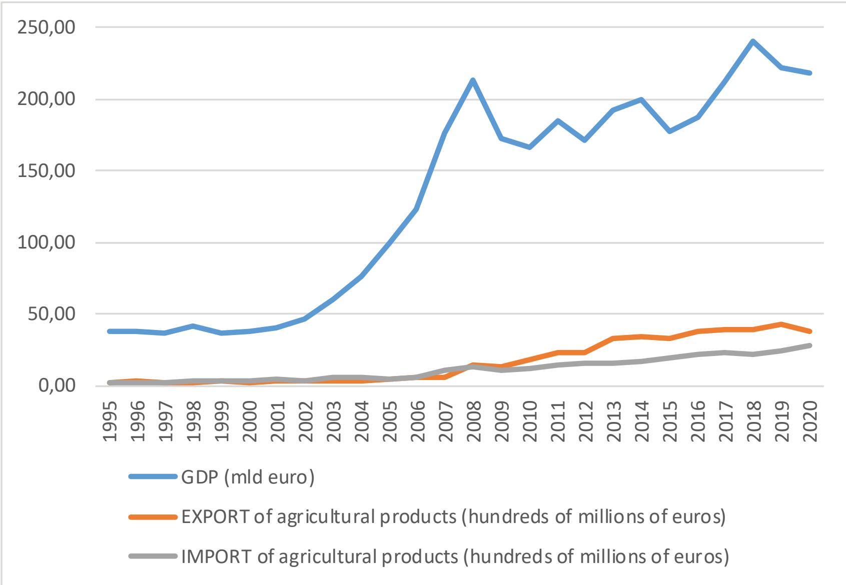
Economic growth is one of the most treated topics in the literature due to the major impact it has on the population of a country. Many studies have tested the dynamics of various factors (export, import) and the process of economic growth (GDP dynamics).

For the three data series, stationarity (unit root test - Dickey-Fuller Augmented), cointegration (Johansen cointegration test) and causality between variables (Granger test and Wald test) were tested.

For the regression equations, generated with the VECM model, the following residue tests were performed: residue normality test (Jarque-Bera test - distribution histogram), residue correlation test (Breusch-Godfrey test) and heteroskedasticity (Breusch-Godfrey test) Pagan-Godfrey and the ARCH test).

Evolution of GDP, exports and imports of agricultural products

Figure 1



Source: author's calculation based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

2. ECONOMIC ARGUMENTATION / LITERATURE REVIEW

The relationship between economic growth (GDP) and external trade (export and import) has long been one of the most debated aspects of international economic development, with a special focus on research. Fundamental economic theories show the contribution of exports to economic growth through the element called the multiplier effect of foreign trade (Tekin, 2012). At the same time, the increase of exports generates an increase of the degree of economic opening, because the respective economies will be able to absorb more quickly the technology of the more developed countries (Hart, 1983). Increasing global factor productivity will positively influence, in the long run, the rate of economic growth. At the same time, trade (import) allows faster access to high technologies, which is an important factor for sustainable economic growth.

Most experts deem that a country's persistent trade deficit is unfavorable, constituting a deterrent factor for sustainable growth of GDP. In the economic literature there are also more balanced opinions considering as dangerous only deficits caused:

- by loans used to finance current consumption to a greater extent than financing long-term investments
- by reducing labor force employment
- as outcome of intensified inflationist processes.

External trade is considered an important component of sustainable economic growth. The links between it and GDP have been analyzed in many specialized economic papers. The developed econometric models have demonstrated the strong connection, both in the short and long term, between these macroeconomic indicators.

3. RESEARCH METHODOLOGY

Selection of the research method

A VEC model was used to generate the regression equation, after testing the stationarity of the selected series (unit root test - Dickey-Fuller Augmented), cointegration (Johansen cointegration test) and causality between variables (Granger test and Wald test).

Selection of variables

In the study of the economic interdependence between GDP, export and import were used the annual data from 1995-2020 and were considered, in turn, GDP, export and import as dependent variables and the other variables, independent variables, as follows:

- dependent variable: GDP and independent variables: export, import
- dependent variable: export and independent variables: GDP, import
- dependent variable: import and independent variables: GDP, export

Description of variables

Gross domestic product (GDP) is a macroeconomic indicator that reflects the sum of the market value of all goods and services intended for final consumption, produced in all branches of the economy within a country within one year.

The Gross Domestic Product, according to the expenditure method, is:

$$GDP = CF + FBCF + \Delta S + (E - I)$$

where:
CF - final consumption,
GFCF - gross fixed capital formation,
ΔS - stock change,
I - import and
E - export.

Exports of goods and services consist of transactions in goods and services (sales, barter, donations, etc.) made by residents to non-residents.

Imports of goods and services consist of transactions in goods and services (purchases, barter, donations, etc.) made by non-residents to residents.

Theoretical presentation of the proposed analysis

For the three data series (GDP, export and import of agricultural goods) were tested:

- stationarity - unit root test - Dickey-Fuller Augmented,
- cointegration - Johansen cointegration test and
- causal link between variables (Granger test and Wald test).

For the regression equation, generated with the VECM model, were did the residue tests:

- residue normality test (Jaque-Bera test - Distribution histogram),
the residue correlation test (Breusch-Godfrey test) and
- heteroskedasticity (Breusch-Pagan-Godfrey test and ARCH test).

4. DATA ANALYSIS

4.1. Basic statistics

The table below shows the basic statistics (Summary Statistics: max, min, average, and SD values) for each variable in question (average, median, maximum, minimum, standard deviation, etc.):

Basic statistics

Table 1

	GDP	Export	Import
Mean	131.03	16.56	11.17
Median	169.16	9.57	10.42
Maximum	240.35	42.89	27.24
Minimum	35.84	2.12	1.83
Std.Dev.	75.17	15.24	8.03
Skewness	-0.16	0.50	0.45
Kurtosis	1.35	1.57	1.86
Jarque-Bera	3.05	3.33	2.30
Probability	0.22	0.19	0.32
Sum	3406.80	430.59	290.36
Sum Sq. Dev.	141255.5	5805.00	1613.93
Observations	26	26	26

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Based on these statistics, we can establish that the value of GDP is between 35.84 billion Euros in 1997 and 240.35 billion Euros in 2018. The average value of this indicator for the period 1995-2020 is 131.03 billion Euros. The values related to the Skewness and Kurtosis tests show us that the considered distribution is not a perfectly symmetrical one, predominating the values located between the average and the maximum of the data series (the median of the series is higher than the average of the series).

For exports, the values are between 2.12 in 1998 and 42.89 in 2019. The average value of this indicator for the period 1995 - 2020 is 16.56. The values related to the Skewness and Kurtosis tests allow us to state that the considered distribution is not a perfectly symmetrical one, predominating the values located between the minimum and the average of the data series (the median of the series is much lower than the average of the series).

Imports have values between 1.83 in 1995 and 27.24 in 2020. The average value of this indicator for the period 1995-2019 is 11.17. The values related to the Skewness and Kurtosis tests allow us to state that the considered distribution is almost symmetrical (the median of the series is close to the average of the series).

4.2. Stationary and integration series tests

Series Stationarity Test (Unit Root Test - Dickey-Fuller Augmented)

1. Unit root test for GDP series in level

a. Series in level, model without constant and without trend

Unit root test without constant and without trend

Table 2

Null Hypothesis: D(GDP) has a unit root
Exogenous: None
Lag Length: 3 (Automatic – based on SIC, maxlag=12

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.480309	0.0134
Test critical values:	1% level	-2.589531	
	5% level	-1.944248	
	10% level	-1.614510	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP,2)
Method: Least Squares
Date:01/10/22 Time: 10:38
Sample (adjusted): 1995 2020
Included observation: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob
D(GDP(-1))	-0.336736	0.135764	-2.480309	0.0150
D(GDP(-1),2)	-0.402145	0.134316	-2.994016	0.0035
D(GDP(-2),2)	-0.308602	0.122368	-2.521914	0.0134
D(GDP(-3),2)	-0.317283	0.097903	-3.240779	0.0017

R-squared	0.415182	Mean dependent var	-0.163158
Adjusted R-squared	0.395903	S.D. dependent var	647.9767
S.E. of regression	503.6314	Akaike info criterion	15.32276
Sum squared resid	23081657	Schwarz criterion	15.43029
Log likelihood	-723.8311	Hannan-Quinn criter	15.36621
Durbin-Watson stat	2.036466		

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

The value of the ADF test is -2.480309, higher than the critical threshold in the DF distribution for the 1% level (-2.589531). We reject the null hypothesis.

b. Series in level, the model with a constant:

Unit root test with a constant

Table 3

Null Hypothesis: GDP has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic – based on SIC, maxlag=12

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		1.764290	0.9997
Test critical values:	1% level	-3.497727	
	5% level	-2.890926	
	10% level	-2.582514	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)
Method: Least Squares
Date:01/10/22 Time: 10:50
Sample (adjusted): 1995 2020
Included observation: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob
GDP(-1)	0.012802	0.007256	1.764290	0.0808
C	-129.1668	214.1636	-0.603122	0.5478

R-squared	0.031092	Mean dependent var	237.9091
Adjusted R-squared	0.021103	S.D. dependent var	510.5735
S.E. of regression	505.1574	Akaike info criterion	15.30761
Sum squared resid	24752845	Schwarz criterion	15.36004
Log likelihood	-755.7268	Hannan-Quinn criter	15.32882
F-statistic	3.112720	Durbin-Watson stat	1.722480
Prob(F-statistic)	0.080831		

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

The value of the ADF test is 1.764290, higher than the critical threshold in the DF distribution for the level of 5% (-2.890926). We do not reject the null hypothesis: the GDP series in level has a unit root, in the model with a constant. The probability attached to the null hypothesis is 0.9997.

c. Series in level, trendy model:

Unit root test

Table 3

Null Hypothesis: GDP has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic – based on SIC, maxlag=12

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.381831	0.8605
Test critical values:	1% level	-4.053392	
	5% level	-3.455842	
	10% level	-3.153710	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP)
Method: Least Squares
Date: 01/11/22 Time: 12:52
Sample (adjusted): 1995 2020
Included observation: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDP(-1)	-0.038400	0.027790	-1.381831	0.1702
C	690.2619	478.8670	1.441448	0.1527
@TREND("1995")	12.97427	6.803923	1.906880	0.0595
R-squared	0.066452	Mean dependent var	237.9091	
Adjusted	0.047003	S.D. dependent var	510.5735	
S.E. of regression	498.4298	Akaike info criterion	15.29064	
Sum squared resid	23849498	Schwarz criterion	15.36928	
Log likelihood	-753.8865	Hannan-Quinn criter.	15.32245	
F-statistic	3.416754	Durbin-Watson stat	1.699051	
Prob(F-statistic)	0.036860			

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

The value of the ADF test is -1.381831, higher than the critical threshold in the DF distribution for the level of 5% (-3.455842). We do not reject the null hypothesis: the GDP series in level has a unit root, in the trend model. The probability attached to the null hypothesis (unit root) is 0.8605.

2. The series in the first difference

$$d(GDP_t) = GDP_t - GDP_{t-1}$$

a) the series in the first difference, the model without constant and without trend

Unit root test

Table 4

Null Hypothesis: D(GDP) has a unit root
Exogenous: None
Lag Length: 3 (Automatic – based on SIC, maxlag=12

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.480309	0.0134
Test critical values:	1% level	-2.589531	
	5% level	-1.944248	
	10% level	-1.614510	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(GDP,2)
Method: Least Squares
Date: 01/11/22 Time: 14:52
Sample (adjusted): 1995 2020
Included observation: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(GDP(-1))	-0.336736	0.135764	-2.480309	0.0150
D(GDP(-1),2)	-0.402145	0.134316	-2.994016	0.0035
D(GDP(-2),2)	-0.308602	0.122368	-2.521914	0.0134
D(GDP(-3),2)	-0.317283	0.097903	-3.240779	0.0017

R-squared	0.415182	Mean dependent var	-0.163158
Adjusted R-squared	0.395903	S.D. dependent var	647.9767
S.E. of regression	503.6314	Akaike info criterion	15.32276
Sum squared resid	23081657	Schwarz criterion	15.43029
Log likelihood	-723.8311	Hannan-Quinn criter.	15.36621
Durbin-Watson stat	2.036466		

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

For the series d(GDP) calculated by simply differentiating the series in level, the ADF test does not argue in favor of the unit root hypothesis. The value of the ADF test is -2.480309, lower than the critical threshold in the DF distribution for the level of 5% (-1.944248). In fact, the probability attached to the hypothesis (unit root) is 0.0134, less than 5%.

The ADF test rejects the unit root hypothesis for the level series, but finds no arguments in favor of the unit root hypothesis for the series calculated in the first difference. We say that the GDP series is non-stationary, integrated by order 1, symbolic I (1).

ADF test

Table 5

Variable	Exogenous					
	Non		Constant		Trend	
	ADF test	Critical value 1%	ADF test	Critical value 5%	ADF test	Critical value 5%
GDP	-2.480309	-2.589531	1.764290	-2.890926	-1.381831	-3.455842
d(GDP)	-2.480309	-1.944248	it's not necessary			

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Tests for residue diagnosis

- Autocorrelation of errors

The residue correlation test (Breusch-Godfrey test) shows that the errors are not correlated (does not reject the null hypothesis: there is no error correlation - Prob. Chi-Square(1)=0.06>0.05):

Breusch-Godfrey Serial Correlation LM Test:

Table 6

F-statistic	2.515305	Prob. F(2,68)	0.0883
Obs*R-squared	5.510686	Prob. Chi-Square(2)	0.0636

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

- Homoscedasticity of random errors

Breusch-Pagan-Godfrey and ARCH tests show that errors are not heteroscedastic (do not reject the null hypothesis: errors are homoscedastic):

- Breusch-Pagan-Godfrey test (Prob. Chi-Square(4)=0.07>0.05)

Heteroskedasticity Test: Breusch-Pagan-Godfrey

Table 7

F-statistic	1.830721	Prob. F(12,67)	0.0605
Obs*R-squared	19.75406	Prob. Chi-Square(12)	0.0719
Scaled explained SS	13.93166	Prob. Chi-Square(12)	0.3051

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

- ARCH test (Prob. Chi-Square(1)=0.97>0.05)

Heteroskedasticity Test: ARCH

Table 8

F-statistic	0.032735	Prob. F(2,75)	0.9678
Obs*R-squared	0.068029	Prob. Chi-Square(2)	0.9666

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

- Random errors have normal distribution

5. CONCLUSIONS

The series stationarity tests (Unit root test - Dickey-Fuller Augmented) indicated that the series are not stationary and become stationary after the first differentiation (export, import), respectively after the second differentiation (GDP). The series cointegration test (Johansen test) indicates the presence of 4 cointegration equations. The Granger short-term causality test shows that there are two-way influence relations between imports and GDP and imports. Unidirectional relations exist between: export and import (export influences import and not vice versa).

Even if it is not the best indicator of human well-being, GDP remains the most complex and accurate tool for judging a nation's well-being. In Romania, the consumption of expensive imported finished products will continue and the cheap basic products will be exported.

Bibliography

1. **Andrei T., Bourbonnais R.**, *2008*, *Econometrie*, Ed. Economică, Bucureşti
2. **Baltagi B.H.,** *2008*, *Econometrics*, Springer
3. **Dușa A., Oancea B., Caragea N., Alexandru C., Jula N.M., Dobre A.M.**, *2015*, *R cu aplicații în statistică*, Editura Universității din București, Decembrie 2015, ISBN 978-606-16-0643-6
4. **Greene W.H.**, 2003, *Econometric Analysis*, Prentice Hall
5. **Hayashi F**., 2000, *Econometrics*, Princeton University Press

6. **Johnston, J., DiNardo J.E.**, 1997, *Econometric Methods*, McGraw-Hill
7. **Jula D.**, 2003, *Introducere în econometrie*, Ed. Professional Consulting, Bucureşti
8. **Jula D., Jula N.M.**, 2016, *Prognoza economică, Aplicaţii financiare*, Editura Mustang, Bucureşti
9. **Jula N., Jula D.**, 2016, *Modelare economica, Modele econometrice şi de optimizare*, Editura Mustang, Bucureşti
10. **Maddala G.S.**, 2001, *Introduction to Econometrics*, Wiley
11. **Pecican E.S.,** 2006, *Econometrie*, Ed. All Beck, Bucureşti
12. **Pfaff, B.**, 2008a, *VAR, SVAR and SVEC Models: Implementation Within R Package vars*. Journal of Statistical Software 27(4). URL http://www.jstatsoft.org/v27/i04/;
13. **Pfaff, B**., 2008b, *Analysis of Integrated and Cointegrated Time Series with R*. Second Edition. Springer, New York. ISBN 0-387-27960-1
14. **R Core Team**, 2017, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
15. **Ramanathan R.**, 1992, *Introductory Econometrics*, Dryden Press
16. **Tekin, R. F.**, 2012, *Economic growth, exports and foreign direct investment in Least Developed Countries: A panel Granger causality analysis*, Economic Modelling, 29, (3), 868-878
17. **Verbeek M.**, 2005, *A Guide to Modern Econometrics*, Wiley
18. **Hart, O.**, 1983, *Imperfect Competition in General Equilibrium: An Overview of Recent Work*, STICERD - Theoretical Economics Paper Series 64, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.
19. **Eurostat database**, http://ec.europa.eu/eurostat/data/ database.

ANNEXES

Annex 1

Dickey-Fuller unit root test Augmented for the GDP data series

a) GDP gross data series

- autoregressive process with non-zero mean

Unit root test

Table 9

Null Hypothesis: GDP has a unit root Exogenous: Constant Lag Length: 8(Automatic – based on SIC, maxlag=10)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.566452	0.4757
Test critical values:	1% level	-3.920350	
	5% level	-3.065585	
	10% level	-2.673460	
*MacKinnon (1996) one-sided p-values.			

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

- autoregressive process with non-zero mean and trend

Unit root test

Table 10

Null Hypothesis: GDP has a unit root Exogenous: Constant, Linear Trend Lag Length: 10(Automatic – based on SIC, maxlag=10)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-10.04755	0.0000
Test critical values:	1% level	-4.800080	
	5% level	-3.791172	
	10% level	-3.342253	
*MacKinnon (1996) one-sided p-values.			

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Annex 2

VAR Lag Order Selection

Table 11

Criteria
Endogenous variables: DGDP DFBCF DEXPBS DIMPBS
Exogenous variables: C
Date: 01/11/22 Time: 16:17
Sample: 1995 2020
Included observations: 25

Lag	LogL	LR	FPE	AIC	SC	HQ
	-2718.963	NA	1.86e+24	67.23366	67.35190	67.28110
	-2649.069	131.1584	4.92e+23	65.90295	66.49417	66.14016
	-2552.480	171.7153*	6.74e+22*	63.91308*	64.97728*	64.34005*

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Annex 3

VECM model: 1st order differentiated data, 1st order delay and a cointegration equation.

Table 12

Vector Autoregression Estimates
Data: 01/12/22 Time: 11:56
Sample (adjusted): 1995 2020
Included observation: 25 after adjustments
Standard errors in () & t-statistics in []

	GDP	Import	Export
GDP(-1)	1.058294	0.015738	0.012658
	(0.28763)	(0.01978)	(0.03959)
	[3.67933]	[0.79574]	[0.31971]
GDP(-2)	-0.195332	0.016963	0.027321
	(0.26412)	(0.01816)	(0.03635)
	[-0.73956]	[0.93402]	[0.75151]
IMPORT(-1)	3.913094	0.627381	0.421049
	(3.91909)	(0.26948)	(0.53944)
	[0.99847]	[2.32809]	[0.78053]
IMPORT(-2)	-0.799562	-.458253	-0.168640
	(3.92644)	(0.26999)	(0.54045)
	[-0.20364]	[-1.69731]	[-0.31203]
EXPORT(-1)	-2.202919	0.022384	0.325252
	(2.05379)	(0.14122)	(0.28269)
	[-1.07261]	[0.15850]	[1.15054]
EXPORT(-2)	1.114250	0.276888	0.409588
	(2.10240)	(0.14456)	(0.28938)
	[0.52999]	[1.91533]	[1.41538]
C	9.031726	1.102625	-1.520316
	(10.9752)	(0.75467)	(1.51068)
	[0.82292]	[1.46106]	[-1.00638]
R-squared	0.934848	0.973696	0.972044
Adj. R-squared	0.911853	0.964412	0.962177
Sum sq. resids	7965.917	37.66391	150.9232
S.E. equation	21.64679	1.488464	2.979569
F-statistic	40.65450	104.8800	98.51559
Log likelihood	-103.7130	-39.46231	-56.11913
Akaike AIC	9.226084	3.871859	5.259928
Schwarz SC	9.569683	4.215458	5.603527
Mean dependent	138.8321	11.93140	17.70145
S.D. dependent	72.91032	7.890138	15.32057
Determinant resid covariance (dof adj.)		4827.195	
Determinant resid covariance		1715.568	
Log likelihood		-191.5336	
Akaike information criterion		17.71113	
Schwarz criterion		18.74193	
Number of coefficients		21	

Annex 4

Granger short-term causality test

Table 13

Pairwise Granger Causality Tests
Date: 01/12/22 Time 12:02
Sample: 1995 2022
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
IMPORT does not Granger Cause EXPORT	25	3.81487	0.0416
EXPORT does not Granger Cause IMPORT		1.07221	0.3631
GDP does not Granger Cause EXPORT	25	5.86659	0.0109
EXPORT does not Granger Cause GDP		0.41409	0.6671
GDP does not Granger Cause IMPORT	25	0.97483	0.3963
IMPORT does not Granger Cause GDP		0.36831	0.6970

Author's calculation in EViews, based on http://unctadstat.unctad.org/wds/TableViewer/tableView.aspx and http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Occupational Gender Disparities – Before and After Covid-19

Ana Maria Ciuhu (dobre.anamaria@hotmail.com)
Institute of National Economy, Romanian Academy & National Institute of Statistics, Romania

Simona Ioana Ghita (simo_ghita@yahoo.com)
Bucharest University of Economic Studies

Valentina Vasile (valentinavasile2009@gmail.com)
Institute of National Economy, Romanian Academy

Cristina Boboc (cristina.boboc@yahoo.fr)
Bucharest University of Economic Studies

ABSTRACT

Women are facing deep and lasting inequalities on the labor market. The effect of the pandemic on women's employment is evident after two years of pandemic, it has significantly affected the level of employment and increased the risk of women at work, inequalities have deepened and the recovery period will be longer than for men. In this paper we present an econometric analysis of occupational gender disparities at European level, considering as predictors the educational level and the countries' development level.

We inspected the influence of pandemic on employment gender gap focusing on the gender occupational gap, gender disparities in education and whether the gender disparities in employment significantly differs depending on the countries' development level. Main findings refer to the statistically proved influence of the pandemic factors on employment gender gap, more significant for employees with tertiary education and in the countries with a higher level of development. The data used in the analysis is provided by EU official statistics.

Keywords: *employment, gender disparities, women's employment, covid-19*
JEL codes: *J01, J16, J21*

1. INTRODUCTION

Even if, demographically, women are the majority, they face deep and lasting inequalities on the labor market (Rembeza, J., Radlińska K.. 2021). Several studies show that favorable developments in the field of gender equality have a positive effect on the entire economic and social situation of a country (Maceira, 2017; Bertay et al, 2020). Prior to the COVID-19 pandemic, gender inequality had a number of advantages and registered a positive trend. In 2019 the ILO reiterated the need for a human-centred approach to the future of work by achieving gender equality at work through “the effective realization of gender equality in opportunities and treatment” (ILO, 2019, p 6), pointing out: to ensure equality principle in opportunities, participation, remuneration and treatment on labour market; to promote work environment and working time arrangements able to achieved a more balanced sharing of family responsibilities and better work–life equilibrium; to invest in the care economy.

During the pandemic women suffered “disproportionate job and income losses” (ILO, 2021). The effect of the pandemic on women’s employment is evident after two years of pandemic, it has significantly affected the level of employment and increased the risk of women at work, inequalities have deepened and the recovery period will be longer ILO (2021). Several studies point to the negative impact of the COVID-19 pandemic on the situation of women, exacerbating the already existing inequalities between the two genders in economic, social and political life, and hampering progress in recent years in this direction (EC, 2021; Fisher & Ryan, 2021).

Women’s jobs are 1.8 times more vulnerable to this crisis than men’s jobs, and although women account for about 39% of total employment, they are responsible for 54% of total job losses (Madgavkar et al, 2020). One possible explanation for this disproportionate effect of the pandemic crisis on female employment is the growing burden of unpaid aid supported to a large extent by women (in the form of childcare or relatives / family members with poor health due to the pandemic). This also meant a reduction in working time and, implicitly, in women’s income. As such, it has been observed that female employment has declined during the pandemic period at a slower pace than the average employment, considering gender segregation in employment, with men and women predominantly employed in different sectors and occupations (Madgavkar A. et al, 2020). Also, a huge number of women is placed in the fight against the pandemic: 76% of the health and social care workers and 86% of the personal care workers in the health services are women, their situation being even more difficult, women have to deal with a high intensity of work, risks of infections and work-life balance (European Commission, 2021a)

In this paper we analyzed the disparities in occupational gender before and after the COVID-19 pandemic, in three scenarios: gender occupational gap, the influence of the education level and by development level of the country.

2. LITERATURE REVIEW

Women's jobs and quality of life are more vulnerable to economic and societal conjunctural phenomena, such as crises. Eagly and Wood (1999) consider that during the crises there is a sharpening of gender inequality, a phenomenon underlying the characteristics of traditional, conventional roles of the two genders: the richer human nature of women, who show a greater concern for the family, relatives or own welfare, but, at the same time, who assumes an inferior status in the society; the strength and stoicism of the man, who becomes the defender of the family outside the home, responsible for ensuring its well-being but who also has a higher status in society (Eagly & Wood, 1999). The two roles evolved differently from the classic expectations, the male remained constant, going in parallel with the expectations, while the role of the woman changed in relation to the traditional expectations, with the extension of her role outside the dwelling, with the more active participation in the labor market, borrowing some of the attributes that once belonged only to men (United Nations, 2020). There is a rapprochement between the two roles, which also, theoretically, leads to a reduction in gender inequalities, but, in practice, there are still significant gaps, determined by both objective and subjective factors. However, it seems that during crises of various types (economic, health) there is a return to the conventional, classic roles of the two genders, women reporting a higher level of stress than men in these circumstances, due to old social expectations to prioritize family obligations at the expense of professional career. The role of women in society depended more on providing opportunities through the legal and institutional framework, good practices and effective implementation of support policies, than on changing the old perception of potential differences, depending on gender. People often argue for gender inequality during times of crisis through stereotypical thinking that women are more inclined to care and domestic work than men (Brescoll et al., 2013). Other studies explain COVID-19 pandemic response policies through gender differences in leadership and risk perception. Thus, women are perceived as less willing to take risks than men (Bem, 1974). During the COVID-19 crisis, female decision-makers were more inclined to take large-scale lockdown action than male decision-makers, as the former put human life and human protection at the forefront, rather than the economic consequences. Meanwhile for male decision makers these priorities are exactly the opposite (Garikipati & Kambhampati, 2020). In normal times, women leaders are often more criticized for abdicating the classic role of empathizer than men (Rudman, 1998). Implicitly, such criticism can affect the decision-making process of female leaders. The current health crisis has given women leaders the opportunity to assert themselves precisely through these traditional qualities, of empathy and concern for others, behavior perceived as positive. Precisely because of the desire to protect and help those in a stalemate, women are more willing to hold leadership positions (Ryan & Haslam, 2007).

Actually, the effects of the pandemic on women's employment are divergent, from increasing the vulnerability of employment and deepening inequalities already present in the labor market to maximum involvement, associated with additional risks in the workplace, in key areas - health, education, social services.

More than ever, women's participation in reducing the effects of the pandemic on the labor market has been far more important than the expectations of gender experts and theorists.

For example, contractual hours worked have remained broadly unchanged for women and decreased for men in euro countries in 2020 comparing with 2019, and is related to composition effects (which, in fact, derive from gender gaps prior to the pandemic) such as the decline in part -time employment, which has affected women to a greater extent than men. (EU quarterly labor force survey, Q4 2020 comparing to Q4 / 2019). According to the same source, more men than women lost their jobs and became inactive during the COVID-19 crisis, therefore, women remained more present in the labor market during the pandemic than men. Moreover, the asymmetric sectoral impact of the COVID-19 crisis (Anderton et al, 2020, Botelho V., Neves P., 2021) has further affected the female labor force, being particularly affected the areas of activity with predominantly famines, the effect being in both directions: a) significant reduction of the activity - personal services, hospitality industry, etc. b) increase in activity associated with overtime demand - health services sector.

3. METHODOLOGY AND DATABASE

Beyond the bibliometric research that highlighted the importance of the female labor force and some implications of the pandemic on employment, we further develop a quantitative research, to emphasize whether gender disparities are a cause or an effect of the pandemic crisis.

In this sense, we performed a statistical analysis of the information on employment by gender available at Eurostat and then we addressed the gender dynamics in employment by gender occupational profile, level of education and countries' development level on which the female labor market activate. We used data for 2019 and 2020, as the data for 2021 is not yet available – this represent one limitation of the performed research.

We applied the following methods of analysis: testing hypothesis by using Paired Samples Statistics, Correlations and Test (for H1 - There is a statistically significant difference between the pre- and post-pandemic gender occupational discrepancy.), testing normality with Kolmogorov-Smirnov and Shapiro-Wilk tests (for H2 - For employees with education level k, there was a statistically significant differentiation of the gender occupational discrepancy in the year of the pandemic manifestation compared to the normal period) and the non-parametric Kruskal-Wallis test and Post-Hoc test (for H3 - Gender Employment Gap differs significantly depending on the level of development of European countries).

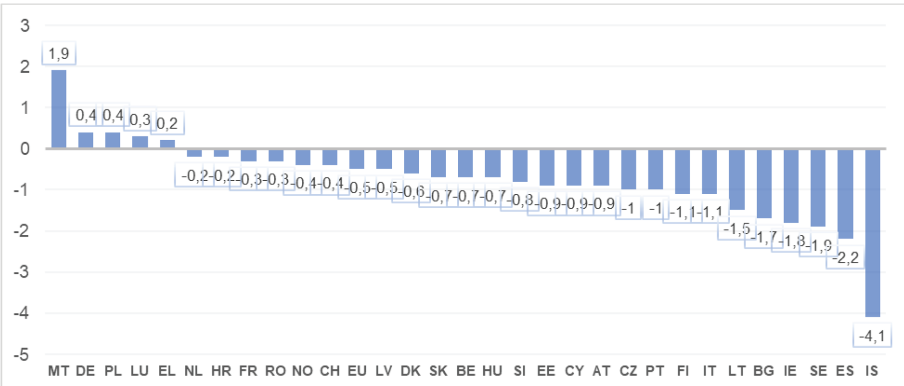
4. RESULTS AND DISCUSSIONS

4.1. Feminine employment in E.U. – a pre and during COVID-19 pandemic statistical analysis

The employment rate for women in most European countries was negative, decreasing by 0.5 percentage points (EU average) in 2020 compared to the pre-pandemic year 2019. The largest decrease was recorded by Iceland (over 4 percentage points), while Malta stood out with an increase in the level of the indicator by almost 2 percentage points, given the rate of female employment in this country close to the EU average. The decrease in the employment rate was double for men (almost 1 percentage point - EU average) (Figure 1)

Change in the female employment rate in 2020 compared to 2019 (%), by European countries

Figure 1



Source: Authors' calculation, based on Eurostat data

By age groups, at the level of the European Union in 2020 there was a sharper decrease in the employment rate among young people, both women and men (by about 2 percentage points), the decrease attenuating with the age increase. There is a greater stability of the employment rate in older age groups, so that in the case of women in 2020 it was only 0.4 percentage points lower than in the previous year, and in the case of men it remained constant.

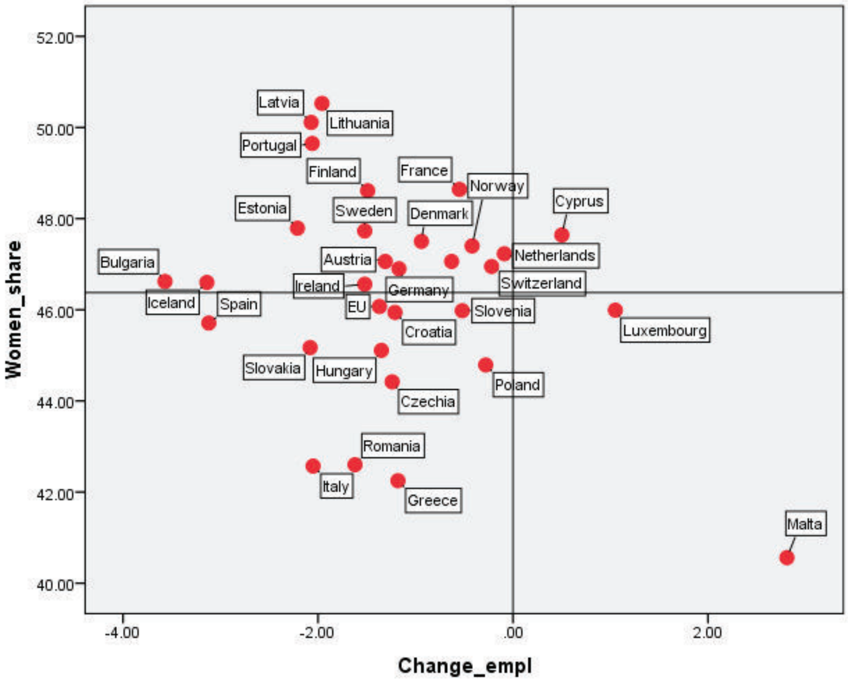
If we analyze European countries from two perspectives: one of the share of women in the employed population and the second of employment changes in the first year of the pandemic (2020), compared to the previous one, we can identify two basic patterns of behavior:

- In most European countries (Nordic countries, France, Portugal) the share of women in the employed population is above average, but there has been - under the impact of the pandemic - a reduction in the employed population in 2020 compared to the previous year;
- In the former socialist, developing countries or in some southern European countries, the share of women in the labor force is lower than the European average, given that even in these countries the pandemic has led to a reduction in employment.

Some isolated situations could be identified, such as that of Cyprus (a country where the share of the female population in the total employed population in 2019 was slightly above the average level - almost 48%) or Malta (where women have the lowest share in the labor force i.e. slightly more than 40%), but which do not seem to have been so much influenced by employment in terms of employment, as the pandemic did not result in a decrease by 2% (Figure 2). Luxembourg, on the other hand, is a country where women have a share of total employment below the European average, and it has been less affected than other European countries by the COVID-19 crisis. Thus, although the real GDP of this country (the largest of all European countries) decreased by 2.8% in 2020 compared to 2019 (Eurostat), this reduction was among the smallest compared to other countries. EU financial sector has shown resilience during the health crisis, government-initiated economy support programs have proven effective, and the country's employed population has seen a slight increase (around 1%) in 2020 compared to the previous year. (https://ec.europa.eu/economy_finance/forecasts/2021/winter/ecfin_forecast_winter_2021_lu_en.pdf)

Scatterplot between Relative change in total employment (2020/2019, %) and share of women in total employment (2019, %) by European country

Figure 2



Source: Authors' calculation, based on Eurostat data

By characterizing the economic sectors on all European countries from the same double perspective: that of the share of women in the employed population and that of the dynamics of total employment in 2020 compared to 2019, several patterns of behaviour of these sectors can be identified:

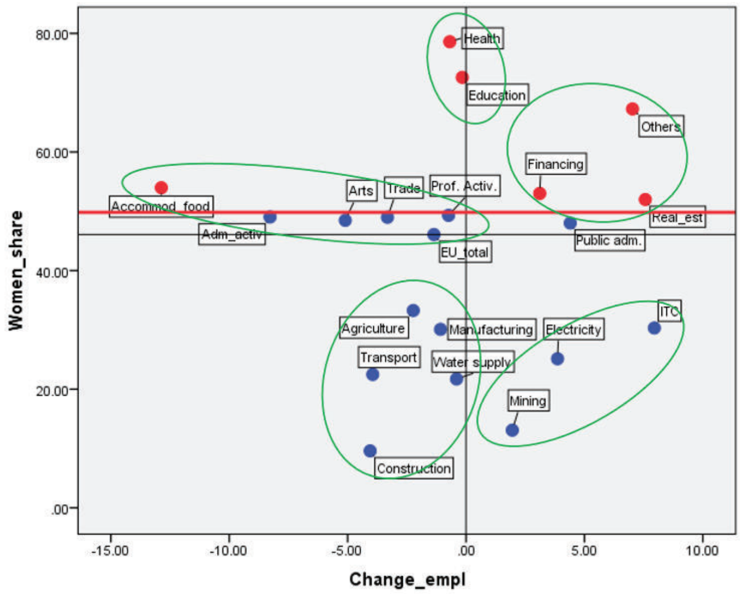
- Predominantly female employment sectors, in which the pandemic did not lead to a contraction of the employed population, to a restriction of activity, but on the contrary, to an increase of no more than 8% (Finance, Real-estate);
- Predominantly female employment sectors, whose total employed population did not change significantly in the year of the pandemic compared to the previous one, as they are sectors with essential activities that required continuity (Health, Education);
- Sectors of activity in which there is a balance between the two genders in terms of the employed population, whose activity was strongly negatively affected by the pandemic, suffering a decrease in

total employment in 2020 compared to 2019 (Accommodation and Food, Administrative activities, Arts, Trade, Professional activities);

- Sectors of activity in which there is a balance between the two genders in terms of the employed population but whose activity is essential, as such in the conditions of the pandemic it was necessary to expand the employed population by about 4% (Public administration);
- Predominantly male sectors of employment that have suffered a restriction of activity and the total employed population, under the rule of the restrictions imposed by the pandemic (Construction, Transport, Agriculture, Manufacturing, Water Supply);
- Sectors of activity with predominantly male employment, but essentials, whose continuity had to be ensured even during the pandemic, and therefore required an increase of employment (Mining, Electricity, ICT - registering the largest increase of the employed population).

Scatterplot between Relative change in total employment (2020/2019, %) and share of women in total employment (2019, %) by activity sector – EU level

Figure 3



Source: Authors' calculation, based on Eurostat data

Legend: 🔴 - Female-dominated sector; 🔵 - Male-dominated sector

4.2. The Covid-19 impact on gender disparities dynamics, on influencing factors

Beyond the results presented above, the research consists of an analysis of the pandemic impact on female labour force, focusing on (a) the gender occupational gap, (b) gender disparities in education and, (c) whether the gender disparities in employment significantly differs depending on the countries' development level.

a) ***Gender occupational discrepancy before and after the pandemic – regardless educational level***

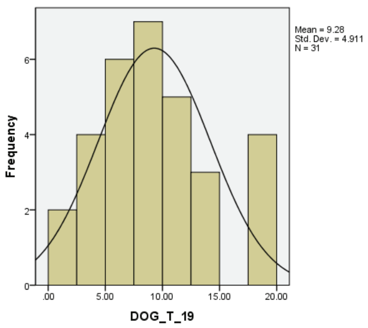
The changes in the employment rate of the two genders have led to statistically significant changes in the occupational gender gap. The variable tested is the disparity between the employment rates of men and women (%), calculated separately for 2019 and 2020, to capture the situation before and during the pandemic, as well as the changes between the two moments in time. The testing was performed by applying the Student's test for paired samples, after, previously, the hypotheses regarding the independence of the observations, the normality of the differences of the observations were checked.

In this regard, the following working hypothesis was formulated:

Hypothesis 1 (H1). *There is a statistically significant difference between the pre- and post-pandemic gender occupational discrepancy***.**
Following the analysis, it can be stated that across the European Union there is a statistically significant difference between the gender gap in employment before and after the pandemic (significance level 0.05).

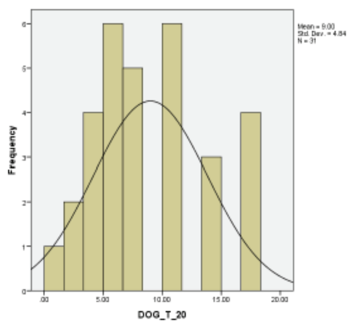
Histogram of the occupational gender disparities, 2019, at EU-level

Figure 4



Histogram of the occupational gender disparities, 2020, at EU-level

Figure 5



Source: Authors' calculation

Although there is a slight decrease of the gender gap in employment across the EU in 2020 compared to 2019, within the EU there are different patterns of change in this gap, under the impact of the pandemic, which can be stated with a high level of confidence (95%) that the pandemic led to a statistically significant differentiation of the occupational gender gap.

Output on Paired Samples Statistics, Correlations and Test for H1

Table 1

a. Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	DOG_T_19	9.2839	31	4.91101	.88204
	DOG_T_20	8.9968	31	4.83952	.86920

b. Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	DOG_T_19 & DOG_T_20	31	.987	.000

c. Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Sifference				
					Lower	Upper			
Pair 1	DOG_T_19 - DOG_T_20	.28710	.77921	.13995	.00128	.57291	2.051	30	.049

Source : Authors' calculation

b) *Gender occupational disparities before and after the pandemic by educational level*

The following hypothesis was formulated:

Hypothesis 2 (H2): *For employees with education level k, there was a statistically significant differentiation of the gender occupational discrepancy in the year of the pandemic manifestation compared to the normal period.* (where k = low education level (0-2 ISCED level); average education level (3-4 ISCED level); high education level (5-8 ISCED level)

Following the application of the t test for the educational groups less than primary, primary and lower secondary education (levels 0-2) and upper secondary and post-secondary non-tertiary education (levels 3 and 4) it turns out that for the low level educated labour force, as well as for the educated one, we cannot claim the existence of a significant difference between the gender discrepancy of employment before and after the pandemic. It seems that the pandemic did not generate a significant change (neither in the sense of attenuation nor in the sense of significant amplification) of the occupational gender gap in these education groups.

However, analyzing the educational group of tertiary education is observed by applying the same analysis shows the gender occupational discrepancy occurred in the group of employees with higher education, in the sense of a statistically significant attenuation of this difference (Table 2). The analysis reveals that in 2020 the gender occupational discrepancy suffered a statistically significant reduction compared to the year before the pandemic. Some possible causes would be the dynamics of pre / post pandemic activities, especially of the predominantly female sectors. Women have been involved in the front line of the fight against the coronavirus pandemic, to a greater extent than men. The sectors of activity that have been more involved in the fight against the pandemic are those with predominantly female occupational (health, social assistance), sectors in which the activity did not decrease during the pandemic, but on the contrary. Across EU, 76% out of 49 million employees in the care and assistance sector, who have been most exposed to the virus, is represented by females (Eurostat database). The widest range: in Latvia 88% of the health care labour force is female, and in Malta only 53%. Other areas where women are over-represented - and which have remained open, active during the pandemic - are essential services from sales to childcare. At EU level, 82% of the cashier and 95% of the cleaning and domestic care workers are women.

Testing the hypothesis of normality of populations according to the two variables is presented in the following.

Testing the differences between 2019 and 2020 for the population with Tertiary Education

Table 2

a. **Tests of Normality for 2019**

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
DOG_ter_19	.119	31	.200*	.941	31	.091

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

b. **Tests of Normality for 2020**

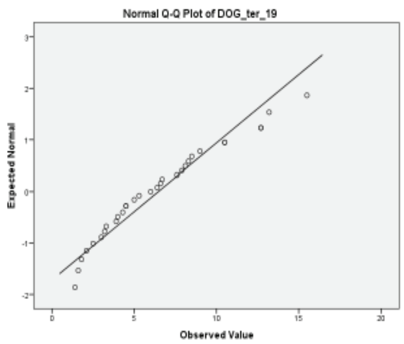
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
DOG_ter_20	.136	31	.153	.927	31	.036

a. Lilliefors Significance Correction

Source : Authors' calculation

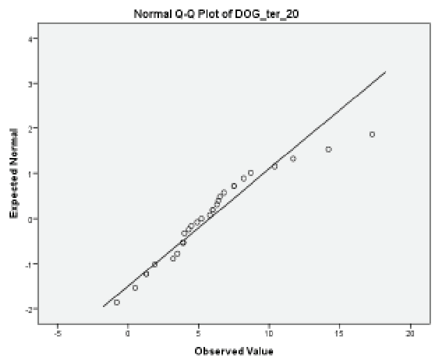
Normal Q-Q plot of employees with tertiary education, 2019

Figure 6



Normal Q-Q plot of employees with tertiary education, 2020

Figure 7



Source: Authors' calculation

The normality hypothesis is validated and the test results are presented in Table 3.

Output on Paired Samples Statistics, Correlations and Test for H2

Table 3

a. Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	DOG_ter_19	6.4710	31	3.76237	.67574
	DOG_ter_20	5.7645	31	3.84305	.69023

b. Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	DOG_ter_19 & DOG_ter_20	31	.923	.000

c. Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	DOG_ter_19 - DOG_ter_20	.70645	1.49397	.26833	.15846	1.25445	2.633	30	.013

Source: Authors' calculation

In conclusion, hypothesis H2 is not validated for low and average education level, but is validated for high education level.

c) ***Gender occupational disparities by development level of the country***

By introducing in the analysis the level of economic development of European countries, figured out by the Gross National Income per capita, we aim to check if the gender employment disparity differs significantly depending on this level. European countries were grouped by Gross National Income per capita. Thus, group 1 includes low-income countries - below $ 12,000 per capita, i.e. countries with economies in transition, non-EU members (North Macedonia, Serbia, Montenegro) and Bulgaria. The second group includes countries with an average income level - between $ 12000 and 35000 per capita, represented by newer EU member states, some also developing (mainly former socialist countries). In the third group are high-income countries - over $ 35,000per capita, these are countries with developed economies, EU members (such as France, Germany, Austria, Belgium, Finland, Denmark, Netherlands, Sweden) or developed countries that are not EU members (Iceland, Switzerland, Norway, United Kingdom).

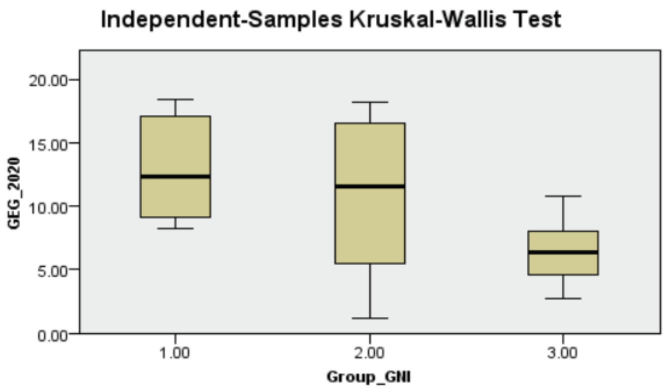
In this context, the following hypothesis was formulated:

Hypothesis 3 (H3): *Gender Employment Gap (GEG) differs significantly depending on the level of development of European countries***.**

Since the hypothesis of normality of the dependent variable is not verified and the three distributions belong to different populations (Figure 8), the non-parametric Kruskal-Wallis test was applied, which checks whether the median occupational gender disparity of the groups of countries formed by level of development differs significantly.

Box-plot diagram for distributions of gender occupational disparities by groups of countries, by development level

Figure 8



Source: Authors' calculation

Following the application of the Kruskal Wallis test, the hypothesis of the similarity of the distributions of the variable explained for the three categories of countries is rejected, at a significance level of 5% and therefore, the previously formulated hypothesis (Hypothesis 2) is validated. (Table 4)

Output on Kruskal Wallis test for H3

Table 4

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of GEG_2020 is the same across categories of Group_GNI.	Independent-Samples Kruskal-Wallis Test	.011	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Source: Authors' calculation

Ranks

Table 5

	Group_GNI	N	Mean Rank
GEG_2020	1 (low_income)	4	25.88
	2 (medium_income)	16	20.47
	3 (high_income)	14	11.71
	Total	34	

Source: Authors' calculation

Test Statistics[a,b]

Table 6

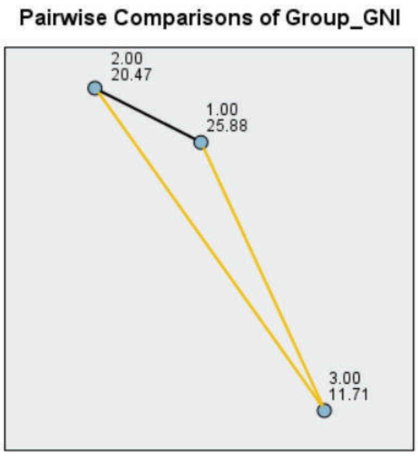
	GEG_2020
Chi-Square	8.983
df	2
Asymp. Sig.	.011
a. Kruskal Wallis Test	
b. Grouping Variable: Group_GNI	

Source: Authors' calculation

Following the above conclusion, Post-Hoc tests were performed to compare the median values of the response variable on the occupational gender gap for the three categories of countries. Thus, it is observed that the groups of countries with medium and low level of Gross National Income are not significantly differentiated by gender discrepancy, but - from this point of view - the differentiation occurs from countries with high level of explanatory variable, respectively countries with a high level of development, which is characterized by a higher similarity between the male and female employment (Figure 9, Table 7).

Graphic comparison of the three categories of countries in terms of gender occupational disparity

Figure 9



Source: Authors' calculation

Output on Post-Hoc test

Table 7

Each node shows the sample average rank of Group_GNI.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
3.00-2.00	8.754	3.643	2.403	.016	.049
3.00-1.00	14.161	5.644	2.509	.012	.036
2.00-1.00	5.406	5.565	.971	.331	.994

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Source: Authors' calculation

Across all countries included in the study, there is a statistically significant inverse correlation of average intensity between Gross National Income and Gender Employment Gap (Table 8, Figure 10).

Correlations

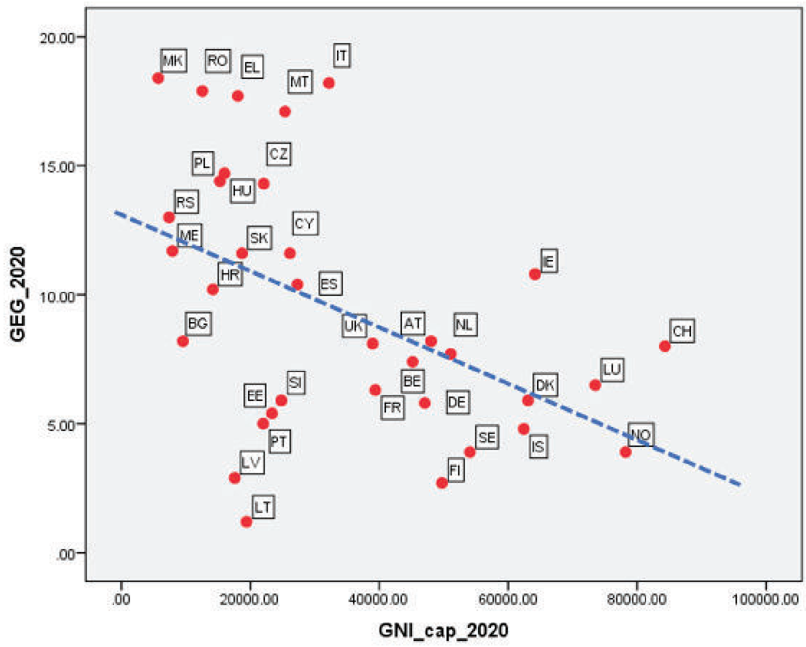
Table 8

			GNI_ cap_2020	GEG_2020
Spearman's rho	GNI_cap_2020	Correlation Coefficient	1.000	-.498**
		Sig. (2-tailed)	.	.003
		N	34	34
	GEG_2020	Correlation Coefficient	-.498**	1.000
		Sig. (2-tailed)	.003	.
		N	34	34
**. Correlation is significant at the 0.01 level (2-tailed).				

Source: Authors' calculation

Correlation between GNI per capita and Gender Employment Gap, 2020, EU countries

Figure 10



Source: Authors' calculation, based on Eurostat data

The figure above shows that the group of former socialist countries, with developing economies, is characterized by low incomes and a high gender disparity in employment, with the male employment rate being significantly higher than the female. At the other end of the spectrum are developed countries in northern Europe (such as Denmark, Norway, Iceland, Sweden) or other European areas (Luxembourg, Switzerland), which - amid very high incomes - have made real progress in closing the gender gap. on the labor market. However, there are also countries (e.g. the Baltic States) which - although low in income by 2020, also have a small gender gap.

CONCLUSIONS

According to the results of the analysis made in this paper, gender disparities have both a cause and an effect of the pandemic crisis, and during the crisis the gap worsened. The impact of the crisis is different in magnitude and is strongly influenced by the particularities of women's employment and the restrictions imposed to limit the incidence among the population. Among the common features of female employment, whose significance was reconsidered during the pandemic, we mention: a) Predominance of employment in lower paid activities with higher vulnerability for layoff; b) women's predisposition to adapt their work and rest time according to the needs of their family and children, which also implies the acceptance of atypical forms of employment, mainly vulnerable - part time or fixed-term labor contracts.

But for the pandemic period, other aspects have been added that accentuate the gender gap, namely: a) women participation in activities closely related to Covid-19 incidence is higher than for men- health sector, education, personal services, in commerce, social services, hospitality sector, just to name several activities with a significant higher presence of female labor force; b) outsourcing of functional services within companies, with the reduction of staff - accounting, financial, cleaning services in office spaces c) increasing the employment of women in services associated with current consumption of goods, which substantially increased the demand for products that meet the requirements of hygiene, usually jobs that require low / minimum qualification - packaging activities in the food processing industry, take a way services from trade and public catering

Women's participation in the labor market during the Covid crisis registered a strongly differentiated dynamic by fields of activity, professions and level of qualification, from activities in which the changes were insignificant (education, social services) or even the presence of women in employment increased (health services, digitized jobs - telework), up to significant temporary reductions (technical unemployment) or even permanent - in personal services, trade, etc.

The main findings of this research refer to the statistically proved influence of pandemic on employment gender gap for employees with tertiary education and across the countries with a high level of development. We identified three categories of EU countries, depending on their development level. Over the medium term, policy makers must seek to attenuate the negative impact of the pandemic on specific groups, including women. Thus, considering the results of the whole analysis, we formulate below some proposals of social and economic policies measures:

- Women-oriented policies for digitalization. Pandemic has spotted in light the digitalization in all aspects of life, and women shall benefit from special specialization in this field, as in general they are not so involved as men are in such digital skills. Those most digitally skilled have been able to cope with the crisis better. The use of digital technologies has been essential to democratic participation; access to public services and public transfers; employment opportunities; access to health; finance; social capital and networks; and even preventing or escaping from gender-based violence. Closing the gender digital divide can then contribute to maintaining of livelihoods, ensuring economic safety nets and even saving lives.
- Measures that attenuate the effects of educational loss. The temporary closure of schools will have important long-term negative effects on shaping the human capital. The spendings shall support remote learning, encourage reenrollment (by prioritizing those at higher risk of dropping out, especially girls), and offset learning losses.
- Introducing in the fiscal policy some gender social norms, such as nontransferable mandatory paternity leaves or tax incentives for companies meeting gender-equality targets.

In conclusion, starting from the results of the analysis in the paper and taking into account the preliminary assessments of experts and officials for 2021, the effect of the pandemic on women's employment is obvious after 2 years, respectively pandemic has more affected the level of employment and risk of women at work, and the recovery period will be longer ILO (2021). As future research directions we can mention the increase of the analyzed time period, with the inclusion of the year 2021, in order to verify if the conclusions are reconfirmed, but also the extension of the analysis, with the inclusion of other influencing factors, such as digitization aspects. employment of women or jobs' disruption as a result of digitization, the incidence of telework, or "women in digital score" (European Commission, 2021b).

Acknowledgement:

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

References:

1. **Anderton, R.; Botelho, V.; Consolo, A.; Da Silva, A.D.; Foroni, C.; Mohr, M.; Vivian, L.**, 2020, *The impact of the COVID-19 pandemic on the euro area labour market*. ECB Economic Bulletin, Issue 8/2020.
2. **Bertay A.C., Dordevic L., Sever C.**, 2020, *Gender Inequality and Economic Growth: Evidence from Industry-Level Data*. IMF Working Paper No. 20/119, ISSN 9781513546278/1018-5941, https://www.imf.org/en/Publications/WP/Issues/2020/07/03/Gender-Inequality-and-Economic-Growth-Evidence-from-Industry-Level-Data-49478
3. **Botelho V., Neves P.**, 2021, *The impact of the COVID-19 crisis on the euro area labour market for men and women*, Economic Bulletin, 4/2021, European Central Bank.
4. **Brescoll, V. L., Uhlmann, E. L., & Newman, G. E.**, 2013, *The effects of system-justifying motives on endorsement of essentialist explanations for gender differences*. Journal of Personality and Social Psychology, 105, 891–908. https://doi. org/10.1037/a0034701
5. **Bem, S. L.**, 1974, *The measurement of psychological androgyny*. Journal of Consulting and Clinical Psychology, 42, 155–162. https://doi.org/10.1037/h0036215
6. **Eagly, A. H., & Wood, W.**, 1999, *The origins of sex differences in human behavior: Evolved dispositions versus social roles*. American Psychologist, 54, 408–423. https://doi.org/10.1037/0003- 066X.54.6.408
7. **Garikipati, S., & Kambhampati, U.**, 2020, *Leading the fight against the pandemic: Does gender "really" matter?* SSRN. https://doi.org/10.2139/ssrn.3617953, in Fisher AN, Ryan MK. Gender inequalities during COVID-19. *Group Processes & Intergroup Relations*. 2021;24(2):237-245. doi:10.1177/1368430220984248. https://journals.sagepub.com/doi/full/10.1177/1368430220984248
8. **ILO**, 2021, *Building Forward Fairer: Women's rights to work and at work at the core of the COVID-19 recovery*, Policy Brief, July 2021, https://www.ilo.org/wcmsp5/groups/public/---dgreports/---gender/documents/publication/wcms_814499.pdf
9. **ILO**, 2019, *ILO Centenary Declaration for the Future of Work*, https://www.ilo.org/wcmsp5/groups/public/@ed_norm/@relconf/documents/meetingdocument/wcms_711674.pdf
10. **Maceira, H.M.**, 2017, *Economic Benefits of Gender Equality in the EU*. Intereconomics, 52, 178-183. DOI: 10.1007/s10272-017-0669-4
11. **Madgavkar A., White O., Krishnan M., Mahajan D., Azcue X.**, 2020, *COVID-19 and gender equality: Countering the regressive effects*. McKinsey Global Institute, https://www.mckinsey.com/featured-insights/future-of-work/covid-19-and-gender-equality-countering-the-regressive-effects.
12. **Rembeza, J., Radlińska K.**, 2021, *"Labor Market Discrimination – Are Women Still More Secondary Workers?"*. Journal of Business Economics and Management 22 (1):77-97. https://doi.org/10.3846/jbem.2020.13648.

13. **Rudman, L. A.**, 1998, *Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management*. Journal of Personality and Social Psychology, 74, 629–645. https:// doi.org/10.1037/0022-3514.74.3.629, in Fisher AN, Ryan MK. Gender inequalities during COVID-19. *Group Processes & Intergroup Relations*. 2021;24(2):237-245. doi:10.1177/1368430220984248
14. **Ryan, M. K., & Haslam, S. A.**, 2007, *The glass cliff: Exploring the dynamics surrounding the appointment of women to precarious leadership positions*. The Academy of Management Review, 32, 549–572. https://doi.org/10.2307/20159315
15. **European Commission,** 2021a, 2021 report on gender equality in the EU, https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/annual_report_ge_2021_printable_en_0.pdf
16. **European Commission**, 2021b, https://digital-strategy.ec.europa.eu/en/news/women-digital-scoreboard-2021
17. **European Institute for Gender Equality**, 2017, *Gender, skills and precarious work in the EU* (https://eige.europa.eu/resources/ti_pubpdf_mh0217250enn_pdfweb_20170503163908.pdf)
18. **United Nations**, 2020, *Policy brief: The impact of COVID-19 on women*. https://www.unwomen. org/-/media/headquarters/attachments/ sections/library/publications/2020/policybrief-the-impact-of-covid-19-on-women-en.pdf?la=en&vs=1406
19. https://ec.europa.eu/eurostat/data/database
20. https://ec.europa.eu/economy_finance/forecasts/2021/winter/ecfin_forecast_winter_2021_lu_en.pdf

Visualising the results of the life-events surveys on satisfaction with public administration in Germany

Daniel Kühnhenrich (daniel.kuehnhenrich@destatis.de)
Federal Statistical Office (Destatis)

Sylvana Walprecht (sylvana.walprecht@destatis.de)
Federal Statistical Office (Destatis)

ABSTRACT

Making data understandable via visualisation is an objective of many types of publications. While R offers a vast amount of packages for the task, new R coders may find themselves overwhelmed by the possibilities. This article presents three easy applications. As part of the Federal Government's Better Regulation programme, the Federal Statistical Office of Germany (Destatis) conducts surveys on the satisfaction of citizens and companies with public authorities every two years since 2015. In 2019, Destatis modernised data visualisation in its print publications. A Sankey diagram informs about the usage of digital devices, a word cloud displays reasons of dissatisfaction and a network diagram highlights essential and secondary agencies. They were generated with the R packages networkD3, wordcloud2, and igraph. During the 9th International Conference on the Use of R in Official Statistics (uRos2021), the authors presented these diagrams as part of the "Scientific Session Dissemination and visualization".

Keywords: *public administration, better regulation, Sankey diagram, word cloud, network diagram, life-events surveys, satisfaction*

JEL classification: *H (Public Economics) + Y (Miscellaneous Categories)*

Introduction

Since 2015, the Federal Statistical Office of Germany (Destatis) has conducted life-events surveys every two years as part of the Federal Government's Better Regulation programme. They examine the satisfaction of citizens and companies with public authorities they had contact with over the last two years. In 2019, for example, a random sample of 6,016 citizens and 2,697 companies was interviewed to this end. Their interactions with public authorities are grouped into life events such as childbirth or starting a business. Besides satisfaction, the questionnaire also deals with digitalisation, complexity of procedures and comprehensibility of forms and applications. In 2019, with the objective to modernise the visualisation of the results, Destatis introduced several new types of graphics as part of its print publications. On 25 January 2021 during the 9th International Conference on the Use of R in Official Statistics (uRos2021), the authors presented these graphics as part of the "Scientific Session Dissemination and visualization" (Universitatea Ecologica Bucuresti, 2021; Walprecht and Kühnhenrich, 2021). This article describes their application in more detail and offers new R coders a place to start in the vast array of available packages. Meanwhile, Destatis (2021) has published the results of the latest surveys.

Methodology

Three types of graphics were introduced into the publications: a Sankey diagram (Kennedy and Sankey, 1898) on the usage of digital devices, a static word cloud (also called tag cloud) (Khusro et al., 2021) for reasons of dissatisfaction and a network diagram (Fleischer and Hirsch, 2001) to highlight essential and secondary agencies in a life event (Destatis, 2019). All three were generated with R, whereas in the surveys before graphics were solely created with SAS and Microsoft Excel.

Sankey diagram

A major focus of the surveys is the digitalisation of public administration. Two questions deal with the citizens' usage of digital devices such as laptops, smartphones or tablets. The first one asked which digital devices respondents employed predominately for private online activities e.g. communication, information, entertainment and shopping, while the other one inquired the medium used for interactions with public authorities. The Sankey diagram in figure 1 illustrates the results in the 2019 survey and compares the usage during these two kinds of activities. The diagram highlights that with 94% the large majority of citizens uses digital devices such as PCs, laptops, smartphones and tablets. However, most of these people filled out paper forms when interacting with public administration. Almost no one employed smartphones and tablets to this end (Destatis, 2019). The Sankey diagram is an appropriate tool to emphasis this discrepancy and to depict the situation for each of the devices in detail.

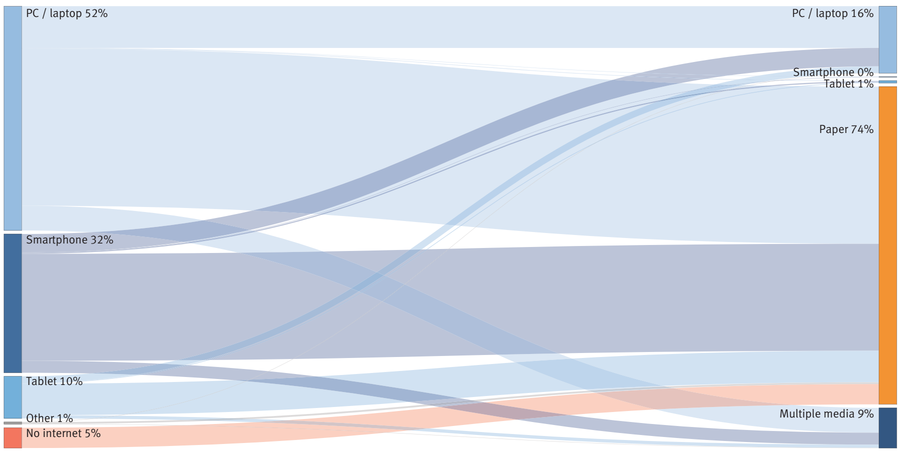


Figure 1: Usage of media for ordinary private activities (left side) and for communicating with public authorities (right side)
Adapted from: Destatis (2019, p. 24).

The **networkD3** package (Allaire et al., 2017) was employed creating the Sankey diagram. It requires two data frames as input:

1. Nodes: it contains the names of the nodes e.g. “PC / laptop” for the left side and “Paper” on the right side.
2. Links: this includes the source numbers of nodes on the left side, target numbers of the nodes on the right side and the respective frequency value e.g. 25% for the link between smartphone users for private online activities and filling out paper forms when in contact with public administration. Furthermore, colour groups were defined.

The `sankeyNetwork` function creates the diagram. It employs the two input data frames i.e. arguments `Links` and `Nodes`, the variables containing the specification of the nodes’ names (`Source`, `Target`), the respective frequency values (`Value`) and IDs (`NodeID`). Furthermore, design elements such as the font (`fontFamily`), its size (`fontSize`) and the colour of the nodes and links (`colourScale`) are defined. The arguments `LinkGroup` and `NodeGroup` are necessary to connect the respective nodes with the colour of the data frame in `colourScale`. Finally, the argument `iterations` enables to change the order of the nodes, but is not used in this case.

```
sankeyNetwork(Links = links,
              Nodes = nodes,
              Source = "source",
              Target = "target",
              Value = "value",
              NodeID = "name",
              fontSize= 32,
              fontFamily = "MetaNormalLF-Roman",
              nodeWidth = 50,
              iterations = 0,
              colourScale = my_color,
              LinkGroup = "group",
              NodeGroup = "group")
```

The function produces a JavaScript graphic in an HTML file. Since Destatis published the results as a printed booklet, the HTML file was converted into a vector graphic and slightly adapted prior to publication.

Word cloud

In an open question, respondents who indicated that they had not been satisfied with the respective public authority were asked for the causes of their dissatisfaction. Destatis categorised and coded the 2,895 answer texts manually. The word cloud in figure 2 displays the most important reasons for dissatisfaction with larger terms indicating a higher frequency. It shows that citizens repeatedly criticised the long processing time of their applications. Furthermore, complicated procedures, waiting time at the agency, rejections and unfriendly or seemingly incompetent staff were mentioned often (Destatis, 2019).

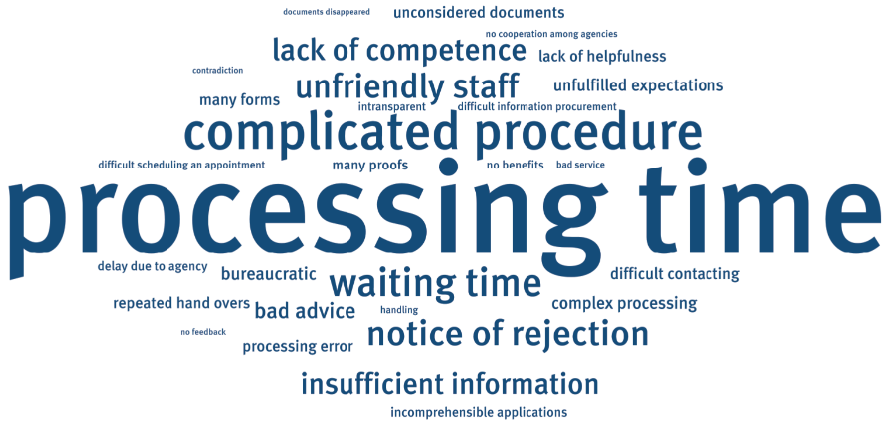


Figure 2: Stated reasons in case of dissatisfaction with public authorities
Adapted from: Destatis (2019, p. 18).

The **wordcloud2** package and its eponymous function were employed to generate the word cloud (Lang and tin Chien, 2018). It requires a data frame containing the terms that are to be displayed and their frequency. A vector was used to define the colours of the words, in this case different shades of blue, while the background colour is white. Since no rotation was needed, the argument `rotateRatio` was set to 0. Furthermore, the font and size of the diagram were specified according to the design guidelines and dimensions required for the print publication.

```
wordcloud2(complains,
           color = colour,
           backgroundColor = "white",
           rotateRatio = 0
           fontFamily = "MetaNormalLF-Roman",
           size = 5, gridSize = 100, widgetsize = c(9600, 5385))
```

Then, the resulting list was exported into an HTML file with the packages **webshot** (Chang, 2019) and **htmlwidgets** (Vaidyanathan et al., 2021). Finally, it was converted into a high-definition PNG file for printing.

Network diagram

At the beginning of the survey, respondents indicated which individual public authorities they had interacted with during the previous two years. With this information, a frequency distribution of contacts was calculated and displayed for the respective life event using a network diagram as shown in figure 3 for the life event "birth of a child". The graphic illustrates that combinations between the registry, parental allowance and family benefits offices, and the statutory health insurance occurred in the majority of cases. Hence, these four authorities play a key role during this life event. New parents interact with the three other agencies – statutory pension insurance, job centre and youth welfare office – less frequently, since they are only important under certain circumstance, e.g. the youth welfare office is responsible for the acknowledgement of paternity of unmarried fathers.

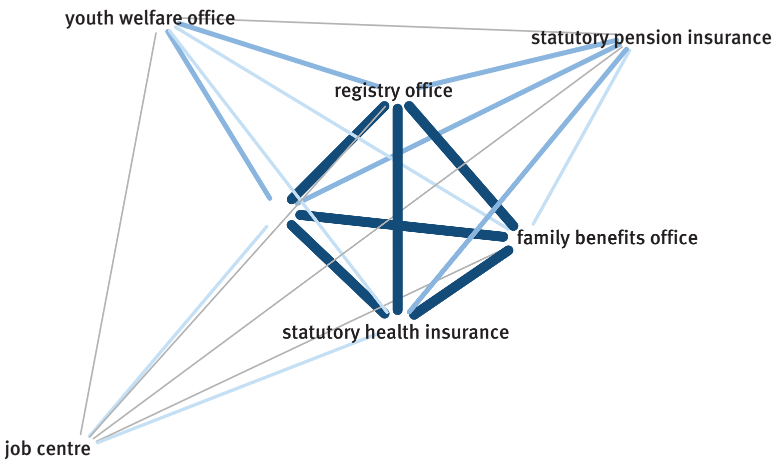


Figure 3: Frequency of combinations of contacts with public authorities in the life event "birth of a child"
Adapted from: Destatis (2019, p. 83).

This diagram was generated using the **igraph** package (Csardi and Nepusz, 2006). It requires a data frame containing two variables describing every possible combination – in the case of 7 public authorities 21 combinations ($= \binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{7!}{(7-2)!2!}$). A third variable includes the frequency of these contacts occurring together.

The `graph.edgelist` function creates a graph with the public authorities as vertexes i.e. nodes and uses no arrows at the end of the edges i.e. simple lines (`directed = FALSE`) for their connections. The command `E(g)$weight` refers to the vector with the frequencies of each combination. Based on that the thickness and colour of each line are allocated in the two new vectors `E(g)$edge.width` and `E(g)$edge.color`. Finally, the `plot` function creates the network diagram by specifying all needed arguments. Vertexes i.e. nodes are set invisible in this example. Only their label is printed. Since the network diagram is part of the printed booklet, a PDF file was generated.

```
network_mat <- as.matrix(data);
g <- graph.edgelist(network_mat [,1:2],directed = FALSE);
E(g)$weight = as.numeric(network_mat [,3]);
E(g)[weight <= 10 & weight > 0]$edge.color <- "#b3b3b3";
E(g)[weight <= 10 & weight > 0]$edge.width <- 1;

. . .

plot(g, vertex.shape = "none",
        vertex.frame.color = "#ffffff",
        vertex.label.color = "black",
        edge.width = E(g)$edge.width,
        edge.color = E(g)$edge.color,
        vertex.label.family = "mtnolfro"
        asp = 0
);
```

Conclusion

This article described three easy uses to illustrate survey results about the satisfaction with public authorities in Germany. Employing modern and innovative diagrams helps to explain complex relationships between variables or entities. In 2021, Destatis moved away from print to solely digital publications of the life-events survey results (Destatis, 2021). Due to the technical specifications of the website, all graphics were programmed in Highcharts (Highsoft, 2022) – a JavaScript library. Nevertheless, these first applications in R were decisive, since the current interactive implementations were built on those experiences. Furthermore, in R data management, data analysis and graphical illustration stem from one source which is a major advantage especially in explorative data analysis. Finally, the R packages presented allow users to create similar graphics for other purposes and to employ further specifications.

Acknowledgements

The authors wish to thank Carsten Hornbach and Susanne Michalik (both Destatis) for their valuable remarks. Furthermore, the authors wish to thank the uRos2021 organising committee for the invitation to the conference.

Bibliography

J. Allaire, C. Gandrud, K. Russell, and C. Yetman. *networkD3: D3 JavaScript Network Graphs from R*, 2017. URL `https://CRAN.R-project.org/package=networkD3`. R package version 0.4. [p2]

W. Chang. *webshot: Take Screenshots of Web Pages*, 2019. URL `https://CRAN.R-project.org/package=webshot`. R package version 0.5.2. [p3]

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`. [p4]

Destatis. *Zufriedenheit der Bürgerinnen und Bürger mit behördlichen Dienstleistungen: Ausgewählte Ergebnisse der Lebenslagenbefragung 2019*, 2019. URL `https://www.amtlich-einfach.de/SharedDocs/Downloads/Ergebnisse_Buerger_2019.pdf?__blob=publicationFile&v=2`. Accessed on 5 May 2022. [p1, 2, 3, 4]

Destatis. *Zufriedenheit der Bürgerinnen und Bürger mit behördlichen Dienstleistungen: Überblick und weitere themenbezogene Detailauswertungen*, 2021. URL `https://www.amtlich-einfach.de/DE/Ergebnisse/Buerger2021/Ueberblick_node.html`. Accessed on 5 May 2022. [p1, 4]

R. Fleischer and C. Hirsch. Graph drawing and its applications. In M. Kaufmann and D. Wagner, editors, *Drawing Graphs: Methods and Models*, pages 1–22. Springer, Berlin, Heidelberg, 2001. [p1]

Highsoft. *Highcharts Documentation*, 2022. URL `https://www.highcharts.com/docs/index`. Accessed of 5 May 2022. [p5]

A. B. W. Kennedy and H. R. Sankey. The thermal efficiency of steam engines. report of the committe appointed to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam engines: with an introductory note. *Minutes of the Proceedings of the Institution of Civil Engineers*, 134(1898):278–312, 1898. [p1]

S. Khusro, F. Jabeen, and A. Khan. Tag clouds: Past, present and future. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 91(2):369–381, 2021. [p1]

D. Lang and G. tin Chien. *wordcloud2: Create Word Cloud by 'htmlwidget'*, 2018. URL https://CRAN.R-project.org/package=wordcloud2. R package version 0.2.1. [p3]

Universitatea Ecologica Bucuresti. *uRos2021 - Scientific Session Dissemination and visualization (Day 2 - session 2)*, 2021. URL https://youtu.be/kjUtApjK6XE?t=2100. Accessed on 5 May 2022. [p1]

R. Vaidyanathan, Y. Xie, J. Allaire, J. Cheng, C. Sievert, and K. Russell. *htmlwidgets: HTML Widgets for R*, 2021. URL https://CRAN.R-project.org/package=htmlwidgets. R package version 1.5.4. [p3]

S. Walprecht and D. Kühnhenrich. *Modern graphics for presenting the results of life-events surveys on satisfaction with public administration*, 2021. URL https://r-project.ro/conference2021/book-of-abstracts-uros2021.pdf#page=21. Accessed on 5 May 2022. [p1]

COVID-19 epidemic in Spain in the first wave: Estimation of the epidemic curve inferred from seroprevalence data and simulation of scenarios based on SEIR model

Francisco-Javier Prado-Galbarro, PhD
PhD Orphan Drug Laboratory, Biologic System Department- Metropolitan Autonomous University, Mexico City, Mexico

Copytzy Cruz-Cruz, DSc
DSc Orphan Drug Laboratory, Biologic System Department- Metropolitan Autonomous University, Mexico City, Mexico

Ana-Estela Gamiño-Arroyo, MD,
MD Hospital Infantil de México Federico Gómez. Mexico City, Mexico.

Carlos Sanchez-Piedra, PhD (csanchez.job@gmail.com)
Research Unit, Health Technology Assessment Agency of Carlos III Institute of Health (AETS), Madrid, Spain

ABSTRACT

The COVID-19 pandemic represents one of the most severe challenges in the recent history of public health. The aim of this study is to ***e****stimate the transmission rate parameter (β) and to predict the epidemic progression in Spain. We integrated data from Our World in Data. Our model considered a mean time from infection to death to be 24 days and the results of the seroprevalence survey in Spain. We calculated β using a SEIR model estimated by least squares. We also used a SEIR model to evaluate four scenarios: 1) model 1: no containment measures, 2) model 2: containment measures from the beginning of the epidemic, 3) model 3: mild measures since the 20th day, 4) model 4: strict containment measures since the 20th day. The estimated β parameter was 1.087. We calculated 41,210,330 infected people and 725,302 deaths in model 1; 165,036 infected people and 2,905 deaths in model 2;*

4,640,400 infected people and 81,671 deaths in model 3; and, 62.012 infected people and 1,091 deaths in model 4. Peak of the epidemic varied from 69th day in model 1 to 216th day in model 4. Containment measures prevented a scenario with a significant increase in deaths and infected people. Our findings showed that, by stricter interventions such as quarantine and isolation could lead to reduce the potential peak number of COVID-19 cases and delay the time of peak infection.

Keywords: *public health, COVID-19, epidemiology, health policy*

1. INTRODUCTION

On December 31st, 2019, the Municipal Health Commission in Wuhan (Hubei province, China) reported 27 cases of pneumonia of unknown etiology that included seven serious cases, with common exposure to a seafood, fish and live animal wholesale market located in Wuhan city [1]. Onset of symptoms of the first case occurred on December 8th, 2019. On January 7th, 2020, Chinese authorities identified a new type of virus of the Coronaviridae family as the agent causing of the outbreak. Labelled afterwards as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), it is responsible for a disease called COVID-19 (Coronavirus Disease 2019).

Coronaviruses is a virus family causing infection in humans and in a variety of animals including birds and mammals, i.e. camels, cats, and bats. It is a zoonotic disease, which means that can be transmitted from animals to humans [2]. Human coronavirus infection (HCoV) can lead to clinical manifestations that range from the common cold with Winter seasonal pattern to other more serious ones such as those produced by Severe Acute Respiratory Syndrome (SARS) and the Middle East respiratory syndrome coronavirus (MERS-CoV) [3].

On March 11th, the WHO declared a global pandemic for COVID-19. From the beginning of this epidemic to the elaboration of this document, more than thirty million COVID-19 cases have been detected in the world. Out of this total, 640,040 cases have been detected in Spain according to data published by the Coordination Centre for Health Alerts and Emergencies of the Spanish Health Ministry [4]. So far, the gross fatality rate in Spain is 4.8% (30,495) [4] of reported cases, although it must be taken into account that these figures are subject to numerous reporting biases and different policies applied with diagnostic testing in each country [5].

Based on data accumulated so far in the European Union and United Kingdom, among confirmed cases 30% of people infected with COVID-19 required hospitalization and 4% were considered in critical condition, the latter defined as needing mechanical ventilation or any other criteria for intensive care unit (ICU) admission [5]. Along the same lines, in Spain, out of the first 18,609 cases with complete information, 43% required hospitalization and 3.9% were admitted to ICU [6]. An overall fatality rate of 14% (CI95% 3.9-32%) among hospitalized cases [7,8], and between 0.3% and 1% in the general population has been estimated through modelling [9].

Different public health strategies have been developed during this pandemic [10,11]. On Saturday 14 March 2020, the declaration of the state of alarm was published in the Spanish Official Gazette as a response to the public health crisis caused by COVID-19. The state of alarm was prolonged through the Royal Decree 476/2020 of 27 March, extending the state of alert declared by RD 463/2020 of 14 March. Assessing the impact of intervention strategies on the spread of COVID-19 and their consequences give valuable information to a better knowledge of the epidemiology of the infection and planning measures relating to future waves of COVID-19. This study aimed at estimating the transmission rate parameter (β) and to predict the epidemic progression taking into account four scenarios and incorporating the most recent COVID-19 epidemiological data from Spain.

2. DATA AND METHODS

Data sources

The information from the number of deaths per day from COVID-19, reported by Our World in Data from February 1st, 2020 to May 17th, 2020 was used.

Estimation of model parameters

Taking into account the mean incubation period of the COVID-19 infection to be 7 days [11] and the mean duration from onset of symptoms to death to be 17 days [12], the mean time between infection and death to be 24 days was considered. Therefore, the new theoretical cases for each day have been estimated moving back daily deaths 24 days and using the Spain seroprevalence survey results, which estimated that 5% of population was affected by SARS-CoV2 [13].

Once estimated the daily theoretical cases of COVID-19 in Spain, the β value was calculated using a SEIR model estimated by least squares. The β parameter value with the smallest root mean squared error was choosen. This parameter represents the transmission rate, which means the probability of transmitting disease between a susceptible and an infectious individual. This model assumes that Exposed [E] population is asymptomatic but infectious, and [I] refers to the symptomatic and infectious population. The incubation rate, σ was described as the rate by which the exposed individual develops symptoms.

After calculating the value for β, different SEIR models were applied taking into account various scenarios:

- Model 1: SEIR model without containment measures.
- Model 2: SEIR model with containment measures from the beginning of the epidemic (since the confirmation of the first case in Spain).
- Model 3: SEIR model with some containment measures (stay at home, wash the hands well and often and keep the distance) since the 20th day of the epidemic.
- Model 4: SEIR model with strict containment measures (additional restrictions to mobility and work activity) since the 20th day of the epidemic.

We summarized our parameters in Table 1.

Data sources for our SEIR model

Table 1

Parameters	Notation	Values	References
Initial population size	N	47,100,396	Instituto Nacional de Estadística (www.ine.es)
Initial susceptible population	S	31	Estimated based on WorldData data and seroprevalence survey
Transmission rate	β_0	1.087	Estimated based on WorldData data and seroprevalence survey
Average duration of infection	γ^{-1}	5 days	Lin et al [14]
Average incubation period	σ^{-1}	7 days	Prem et al [11]
Public sense of risk based on known critical cases and deaths	D(t)	0.05*I(t)	Wu et al [24]
Intensity of responds	k	100	Lin et al [14]
Governmental action strength	α	(0, 0.5, 0.7)	Lin et al [14]

The SEIR model is a compartmental model for modeling how a disease spreads through a population. It is an acronym for Susceptible [S], Exposed or latent [E], Infected [I], Recovered or dead [R]. When a disease is introduced to a population, the people move from one of these classes (or compartments) to the next. We assume that latent [E] population is asymptomatic but infectious, and [I] refers to the symptomatic and infectious population.

SARS-CoV-2 was first confirmed to have spread to Spain on 31 January 2020 and we simulated until September 27, 2020 (in particular, t = (0, 240)). In the first model, the transmission rate remained constant in the period (in particular, β = 1.087). Thus, the model is formulated as follows:

$$\begin{cases} S'(t) = -\dfrac{\beta S(t)I(t)}{N}, \\ E'(t) = \dfrac{\beta S(t)I(t)}{N} - \sigma E(t), \\ I'(t) = \sigma E(t) - \gamma I(t), \\ R'(t) = \gamma I(t), \end{cases}$$

In the rest of the models, the transmission rate function formulated in Qianying Lin et al. (2020) was adopted, considering the effect of governmental action and individual actions (stay at home, wash the hands well and often and keep the distance). The containment measures applied from the 20th day of the epidemic for models 3 and 4 were simulated. Thus, our SEIR model is formulated as follows:

$$\begin{cases} S'(t) = -\dfrac{\beta(t)S(t)I(t)}{N}, \\ E'(t) = \dfrac{\beta(t)S(t)I(t)}{N} - \sigma E(t), \\ I'(t) = \sigma E(t) - \gamma I(t), \\ R'(t) = \gamma I(t), \end{cases}$$

where

$$\beta(t) = \beta_0 \left(1 - \alpha(t)\right) \left(1 - \frac{D}{N}\right)^k,$$

for model with containment measures from the beginning of the epidemic, and

$$\begin{cases} \beta_0, & si\ t < 20 \\ \beta(t) = \beta_0 \left(1 - \alpha(t)\right) \left(1 - \dfrac{D}{N}\right)^k, & si\ t \geq 20 \end{cases},$$

for models with mild containment measures and strict containment measures, respectively; where β_0 is the transmission rate without containment measures, α(t) (with values in the interval [0, 1]) represents the government actions, D(t) is the public feeling of risk as a consequence of the known critical cases and deaths, and k measures the intensity of the individual reactions.

Under the naive scenario (model 1), the governmental action strength α = 0 and intensity of individual reactions k = 0 was assumed. The model with containment measures from the beginning of the epidemic and mild containment measures, α = 0.5 and α = 0.4 respectively, and k = 100. The model with strict containment measures was considering stricter government containment measures with α = 0.7 and K = 100. Lin et al. used K = 1117.3, but a higher value was considered because the political reality and social environment in China are different than in the Western world [14].

The estimated number of deaths from COVID-19 in Spain during the first 21 weeks of 2020 was 43,945 deaths [15]. Considering that in Spain approximately 2,500,000 inhabitants were infected [13], the case fatality rate (CFR) estimated was 1.76%.

3. RESULTS

The study estimated the basic reproduction number to be $R_0 = \frac{\beta_0}{\gamma} = 5.435$. The basic reproductive number is derived by assuming that the mean infectious period is 5 days, and the transmission rate is 1.087, which was estimated with a SEIR model. If the average R0 in the population was greater than 1, the infection will spread exponentially. If R0 is less than 1, the infection will spread only slowly, and it will eventually die out. The highest value of R0, the fastest an epidemic will progress. The basic reproduction number may vary across locations because contact rates among people may differ due to differences in population density and cultural differences [16].

Summary of predictions from four scenarios of containment measures against COVID-19 disease in Spain

Table 2

T= 82 days	No containment measures	Containment measures from the beginning of the epidemic (α=0.5)	Mild containment measures (α=0.4)	Strict containment measures (α=0.7)	Scenario based on reality
Case fatality rate	1.8%	1.8%	1.8%	1.8%	1.8%
Population without recovery	6035730	57863	1668610	12124	485498
Total infected population	41210330	165036	4640400	62012	2508658

Total population recovered or death	35174600	107173	2971790	49888	2023160
Total death population	725302	2905	81671	1091	44152
Population not affected	813362	46810900	39047300	47017500	43794700
Total recovered population	34449298	104268	2890119	48796	1979008
Max. Simultaneous infected population	9352370	57863	1668610	12124	485498
Max. New daily infections	2160730	16862	468500	2915	112030
Max. New daily recoveries	1872300	10973	318600	2370	95440
R0 (basic reproduction number)	5,435	5,435	5,435	5,435	5,435

Four different models according to control measures for containing the pandemic in Spain were considered. Table 2 summarizes these models. Model 1 implied no containment measures. In this scenario 41,210,330 infected people (with an estimated peak of 2,160,730 people infected per day) and 725,302 people could have died were estimated. In the second model, with containment measures from the beginning of the epidemic (since the confirmation of the first case in Spain), there was an estimated 165,036 infected people, with a peak of 16,862 infected patients per day. In model 3 and 4, the impact of containment measures to control the epidemic were studied on day 25 after the first reported COVID-19-related case. The third model considered a scenario with mild containment measures (stay at home, wash the hands well and often and keep the distance). This model estimated 4,640,400 infected people and 81,671 deaths. The estimated peak of new cases was 468,500 people per day. Finally, 62,012 infected people and 1,091 deaths were expected in the model 4 (strict containment measures including restrictions to mobility and work activity since the 20th day of the epidemic). The report of cumulative active infections up to May 17 for Spain was examined onto our predicted model and the peak of new daily infections reached 112,030 cases was found.

Estimated theoretical cases

Figure 1

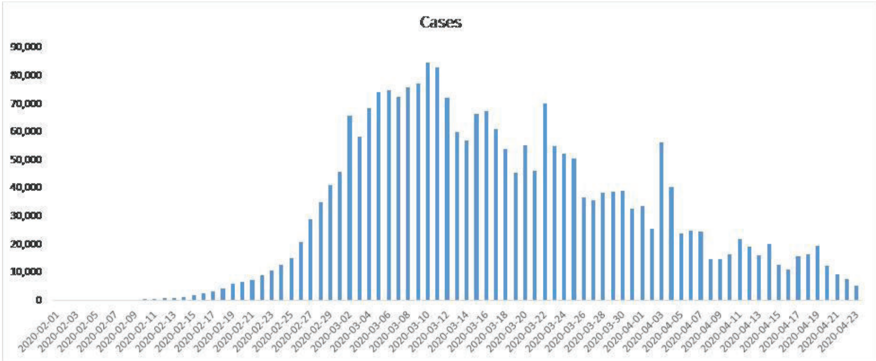


Figure 1 shows the data on the number of daily reported theoretical cases, which indicate that the peak of the epidemic was reached on March 10, 2020.

Results of four SEIR models depeding on containment measures for 240 days

Figure 2

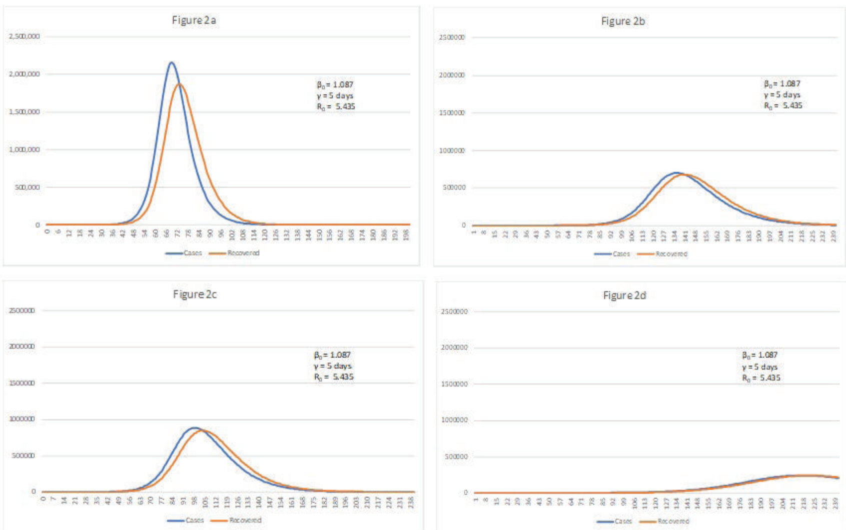


Figure 2a shows the first scenario where the epidemic would have peaked by 69th day (in this scenario using a period of 240 days,), although this scenario would imply a peak of 2,160,730 new daily infections. Figure 2b presents results from the second model. The maximum new daily infections would have been 702,050 and the epidemic would have peaked after 135 days. In this scenario, measure would have been implemented from the beginning of the epidemic (first infection detected in Spain). Figure 2c shows a third model with mild restrictions from day 82 after first case detected. The epidemic in this model would have reached by 99th day and the peak of new daily infections would have been 887,250. A scenario based on strict control measures would have reached the peak of the epidemic on day 216, but would have involved a maximum number of daily infections of 244,140 (Figure 2d).

Cases accumulated by scenarios according to SEIR model

Table 3

Time	Date	No containment measures	Containment measures from the beginning of the epidemic	Mild containment measures	Strict containment measures	Scenario based on reality
1	2020-02-01	4	4	4	4	31
2	2020-02-02	8	8	8	8	72
3	2020-02-03	13	12	13	13	128
4	2020-02-04	19	16	19	19	202
5	2020-02-05	25	20	25	25	302
6	2020-02-06	34	25	34	34	436
7	2020-02-07	44	30	44	44	615
8	2020-02-08	57	35	57	57	854
9	2020-02-09	72	42	72	72	1,176
10	2020-02-10	92	48	92	92	1,606
11	2020-02-11	117	56	117	117	2,183
12	2020-02-12	148	65	148	148	2,957
13	2020-02-13	187	74	187	187	3,993
14	2020-02-14	236	85	236	236	5,381
15	2020-02-15	297	97	297	297	7,242
16	2020-02-16	373	110	373	373	9,735
17	2020-02-17	469	124	469	469	13,075
18	2020-02-18	588	141	588	588	17,552
19	2020-02-19	737	159	737	737	23,550
20	2020-02-20	923	179	923	923	30,269
21	2020-02-21	1,157	202	1,141	1,129	37,793
22	2020-02-22	1,448	227	1,381	1,333	46,748
23	2020-02-23	1,813	255	1,652	1,538	57,403
24	2020-02-24	2,269	287	1,958	1,748	70,083
25	2020-02-25	2,840	322	2,307	1,965	85,172
26	2020-02-26	3,553	361	2,705	2,189	106,155
27	2020-02-27	4,446	404	3,159	2,422	135,083
28	2020-02-28	5,562	453	3,679	2,665	170,262
29	2020-02-29	6,957	507	4,274	2,919	211,512
30	2020-03-01	8,702	567	4,955	3,184	257,405
31	2020-03-02	10,885	634	5,734	3,461	323,297
32	2020-03-03	13,614	709	6,625	3,752	381,780
33	2020-03-04	17,026	793	7,646	4,056	450,440
34	2020-03-05	21,293	886	8,814	4,374	524,726
35	2020-03-06	26,628	989	10,151	4,707	599,547
36	2020-03-07	33,298	1,105	11,681	5,056	672,047

37	2020-03-08	41,637	1,234	13,432	5,421	747,851
38	2020-03-09	52,059	1,378	15,436	5,804	824,994
39	2020-03-10	65,086	1,538	17,729	6,205	909,815
40	2020-03-11	81,364	1,716	20,354	6,624	993,030
41	2020-03-12	101,701	1,915	23,357	7,064	1,065,262
42	2020-03-13	127,103	2,137	26,794	7,524	1,125,440
43	2020-03-14	158,820	2,385	30,727	8,007	1,182,315
44	2020-03-15	198,406	2,661	35,227	8,512	1,248,655
45	2020-03-16	247,788	2,968	40,375	9,041	1,316,244
46	2020-03-17	309,353	3,311	46,265	9,595	1,377,226
47	2020-03-18	386,045	3,693	53,002	10,175	1,431,244
48	2020-03-19	481,485	4,119	60,708	10,782	1,476,780
49	2020-03-20	600,116	4,594	69,521	11,419	1,532,047
50	2020-03-21	747,346	5,123	79,599	12,085	1,578,208
51	2020-03-22	929,729	5,713	91,122	12,783	1,648,476
52	2020-03-23	1,155,136	6,371	104,295	13,514	1,703,565
53	2020-03-24	1,432,922	7,104	119,351	14,279	1,755,797
54	2020-03-25	1,774,059	7,921	136,555	15,081	1,806,244
55	2020-03-26	2,191,210	8,832	156,209	15,921	1,842,851
56	2020-03-27	2,698,640	9,848	178,657	16,800	1,878,476
57	2020-03-28	3,312,040	10,980	204,284	17,721	1,916,869
58	2020-03-29	4,047,950	12,241	233,533	18,686	1,955,708
59	2020-03-30	4,922,980	13,648	266,900	19,696	1,994,994
60	2020-03-31	5,952,500	15,215	304,945	20,753	2,027,762
61	2020-04-01	7,149,140	16,962	348,305	21,861	2,061,512
62	2020-04-02	8,520,800	18,910	397,687	23,021	2,087,226
63	2020-04-03	10,068,690	21,080	453,890	24,236	2,087,226
64	2020-04-04	11,785,770	23,499	517,805	25,508	2,143,655
65	2020-04-05	13,655,660	26,196	590,424	26,841	2,184,101
66	2020-04-06	15,652,820	29,200	672,849	28,236	2,208,030
67	2020-04-07	17,743,890	32,549	766,294	29,697	2,233,119
68	2020-04-08	19,890,110	36,280	872,099	31,227	2,257,762
69	2020-04-09	22,050,840	40,438	991,721	32,829	2,272,405
70	2020-04-10	24,186,610	45,071	1,126,747	34,507	2,287,047
71	2020-04-11	26,261,850	50,234	1,278,883	36,264	2,303,565
72	2020-04-12	28,247,670	55,985	1,449,950	38,104	2,325,351
73	2020-04-13	30,121,970	62,393	1,641,869	40,031	2,344,369
74	2020-04-14	31,870,000	69,531	1,856,661	42,048	2,360,530
75	2020-04-15	33,483,600	77,482	2,096,382	44,161	2,380,797
76	2020-04-16	34,960,170	86,338	2,363,132	46,373	2,393,565
77	2020-04-17	36,301,410	96,200	2,659,000	48,689	2,404,547
78	2020-04-18	37,512,390	107,183	2,986,000	51,114	2,420,262
79	2020-04-19	38,600,070	119,410	3,346,030	53,654	2,436,690
80	2020-04-20	39,572,850	133,023	3,740,830	56,313	2,456,065
81	2020-04-21	40,439,840	148,175	4,171,900	56,314	2,468,387
82	2020-04-22	41,210,330	165,036	4,640,400	56,316	2,477,672
83	2020-04-23	41,893,330	183,798	4,640,465	56,318	2,485,440
84	2020-04-24	42,497,620	204,668	4,640,532	56,320	2,490,708

Table 3 shows the accumulated cases estimated to four scenarios, and the theoretical accumulated cases that were estimated with the mortality data provided by Our World in Data and the seroprevalence survey. The results show the sensitivity analysis of the SEIR models, and they indicate that on day 82 of the epidemic Spain reaches 2,500,000 cases, which coincides with the seroprevalence survey.

4. DISCUSSION

A modified SEIR compartmental model accounting for infection from undiagnosed individuals and four scenarios of control measures was implemented. Also, the hypothetical evolution of the COVID-19 disease in Spain under these four scenarios was evaluated. The most important findings of this work can be summarized as follows: 1) our SEIR model estimated that implementing strict containment measures 82 days after the first case in Spain could have had a relevant impact on the disease spread, 2) The real situation of infection in Spain seems to be in an intermediate scenario between our models with moderate measures and strict measures against the disease. The main purpose of this study was to provide an evaluation of the lessons learned from the first wave of the pandemic and to understand new tools able to support the policymakers decision about the action to minimize the impact of the disease in the future.

The simulation scenarios were adjusted according to the population of Spain for a 240-day epidemic evolution. Our models simulated the conditions where COVID-19 is spreading in a closed population of 47,100,396 people, with and without the effect of governmental action and individual actions (stay at home, wash the hands well and often and keep the distance). Every model suggest that Spain would have reached the peak of the epidemic at different times, delaying the peak with serious and early measures. The findings are comparable with other previous studies as the Wuhan study, where analyzed the epidemics trend of COVID-19 in China and they found that under strong suppression of "lockdown Hubei" the epidemic of COVID-19 in China would achieve peak by late February and decline by the ed of April 2020.REF [17,18]. A recent study using a mathematical model named SEMCR estimated that the optimal timing of interventions differs between suppression and mitigation strategies, as well as depending on the definition of optimal. The study showed that immediate suppression taken in Wuhan significantly reduced the total exposed and infectious populations, and its success required efficient government initiatives and effective collaborative governance for mobilizing of corporate resources to provide essential goods. Also, in London, it was possible to take a hybrid intervention of suppression and mitigation for every 2 or 3 weeks over a longer period to balance the total infections and economic loss [19].

This study found that the R_0 value is equal to 5.44, which means that, in Spain, on average, each infected person is capable of infecting 5.44 people. According to a study published in Emerging Infectious Diseases [20], the median R_0 value was estimated to be 5.7 during the coronavirus disease outbreak in China. Therefore, when a virus is more infectious, a higher percentage of people would need to have immunity to stop the spread.

The SEIR models simulated with different scenarios showed that serious measures taken before March 14 contributed to a considerable reduction of COVID-19 cases. The growth of cases could have been substantially greater if the Government of Spain had not decreed a state of alarm or implemented any containment measure. Given the rapid rate of spread as seen in outbreaks in Europe and the United States, the entire human population is potentially susceptible to SARS-COV-2 infection. Therefore, containment measures as strict lockdown or identification of cases by testing can help prevent further transmission of the virus. Countries, such as the United States, Brazil or Mexico, did not adopt effective mitigation measures (introducing control measures early, strict lockdown measures or identification of cases) and have not managed to stop the transmission. Countries are adopting different ways to contain the SARS-COV-2 spread but there is no one-size-fits-all approach.

This kind of analysis can be useful to face a second wave of the SARS-COV-2 infection [21,22]. In Spain, some researchers called for an independent and impartial evaluation as part of process to be prepared for the next steps [23]. The key point in the interpretation of the findings is to define clearly a model with strict containment measures. The study has limitations, such as: there is not enough information on the immunity of recovered persons, so the model did not consider the transition from the recovered category to the exposed category; the model did not consider the number of tests performed or hospitalizations; and the model did not distinguish between asymptomatic and symptomatic people. However, these models allow a better estimation of the parameters once that the epidemiological curve has decreased, and mortality data have been used to estimate the theoretical cases in Spain. Since the underestimation in the count of cases is greater than that of deaths. It would also be convenient to analyze the epidemic by regions of Spain and to control where the outbreaks are detected so to adapt containment measures to specific regions.

In conclusion, the well-mixed SEIR compartmental model was developed to know the transmission patterns, estimating the effect of control measures in the COVID-19 epidemic, and to evaluate the model´s parameters. The findings showed that, by stricter interventions such as quarantine and isolation could lead to reduce the potential peak number of COVID-19 cases and delay the time of peak infection. Also, the results can facilitate the interpretation of intervention strategies to certain countries in light of its multiple natures and capabilities. And the findings could be useful to help to future analyses and supporting the policymakers' decisions on the minimization of the COVID-19 disease impact.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Author Contributions:

Formal analysis: FJP-G.

Conceptualization and methodology: FJP-G and CS-P.

Writing – original draft: FJP-G and CS-P.

Writing – review & editing: FJP-G, CC-C, A-EG-A, CS-P.

REFERENCES

1. Isolate WSMPV. Wuhan-Hu-1, Complete Genome; 2020.
2. Organization WH. Novel Coronavirus (‎2019-nCoV)‎: situation report, 3. 2020.
3. **Paules CI, Marston HD, Fauci AS**., 2020, *Coronavirus infections—more than just the common cold*. *Jama.* 2020;323(8):707-708.
4. (CCAES) CdCdAyES. Enfermedad por nuevo coronavirus, COVID-19. Situación actual. 2020; https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Actualizacion_210_COVID-19.pdf. Accessed 21/09/2020, 2020.
5. Prevention ECfDCa. Coronavirus disease 2019 (COVID-19) pandemic: increased transmission in the EU/EEA and the UK – seventh update [Internet]. . 2020; https://www.ecdc.europa.eu/sites/default/files/documents/RRA-seventh-update-Outbreak-of-coronavirus-disease-COVID-19.pdf. Accessed 10/07/2020.
6. [Internet]. RNdVEIslsdC-eE. Madrid Mar 2020.
7. **Wu P, Hao X, Lau EH, et al.**, 2020, *Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China*, as at 22 January 2020. *Eurosurveillance.* 2020;25(3):2000044.
8. **Ryu S, Chun BC**, **of Epidemiology KS**., 2020, *An interim review of the epidemiological characteristics of 2019 novel coronavirus*. Epidemiology and health. 2020;42.
9. **Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD.**, 2020, *How will country-based mitigation measures influence the course of the COVID-19 epidemic? The Lancet.* 2020;395(10228):931-934.
10. **Phelan AL, Katz R, Gostin LO.**, 2020, *The novel coronavirus originating in Wuhan, China: challenges for global health governance*. *Jama.* 2020;323(8):709-710.
11. **Prem K, Liu Y, Russell TW, et al.**, 2020, *The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study*. Lancet Public Health. 2020;5(5):e261-e270.
12. **Wu JT, Leung K, Bushman M, et al.**, 2020, *Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China*. Nat Med. 2020;26(4):506-510.
13. **Pollán M, Pérez-Gómez B, Pastor-Barriuso R, et al.**, 2020, *Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study*. The Lancet. 2020.
14. **Lin Q, Zhao S, Gao D, et al.**, 2020, *A conceptual model for the outbreak of Coronavirus disease 2019 (COVID-19) in Wuhan, China with individual reaction and governmental action*., International journal of infectious diseases. 2020.

15. **(INE) INdE**. *Estimación del número de defunciones semanales durante el brote de covid-19*. 2020; https://www.ine.es/experimental/defunciones/experimental_defunciones.htm. Accessed 10/08/2020.
16. **You C, Deng Y, Hu W, et al.**, 2022, *Estimation of the time-varying reproduction number of COVID-19 outbreak in China*. International Journal of Hygiene and Environmental Health. 2020:113555.
17. **Degli Atti MLC, Merler S, Rizzo C, et al.**, 2008, *Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios*. PloS one. 2008;3(3):e1790.
18. **Yang Z, Zeng Z, Wang K, et al.**, 2020, *Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions*. Journal of Thoracic Disease. 2020;12(3):165.
19. **Yang P, Qi J, Zhang S, et al**., 2020, *Feasibility study of mitigation and suppression strategies for controlling COVID-19 outbreaks in London and Wuhan*. PloS one. 2020;15(8):e0236857.
20. **Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R.**, 2020, *Early Release-High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2*. 2020.
21. **He S, Peng Y, Sun K.**, 2020, *SEIR modeling of the COVID-19 and its dynamics*. Nonlinear Dynamics. 2020:1-14.
22. *Leung K, Wu JT, Liu D, Leung GM.*, 2020, *First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment*. The Lancet. 2020.
23. **García-Basteiro A, Alvarez-Dardet C, Arenas A, et al.**, 2020, *The need for an independent evaluation of the COVID-19 response in Spain*. The Lancet. 2020.
24. **Wu Z, McGoogan JM.**, 2020, *Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention*. Jama. 2020;323(13):1239-1242.

The Concentration of Tourism Resources versus Accommodation Statistics in Localities in Romania

Cristi FRENȚ, Senior Researcher, PhD (email: cristi.frent@incdt.ro)
National Institute of Research Development in Tourism, Bucharest, Romania

ABSTRACT

Purpose

In 2008 the Romanian government adopted the National Plan for Territorial Planning (rom. Planul de Amenajare a Teritoriului Național – PATN) section VIII – Areas with tourism resources. In this document a categorization of localities/municipalities that have tourism resources is undertaken based on tourism potential and other supply-side elements (e.g. tourism specific infrastructure and general infrastructure). The purpose of this paper is to see to what extent the categorization from PATN is reflected in the evolution of tourism flows (as part of demand) in the period 2001-2019. In other words, the aim is to see how the categorization from PATN is seen from the perspective of demand-side, a component that was not included in the methodology pertaining to PATN.

Design/methodology/approach

Data from National Institute of Statistics (INS) at municipality level have been used for deriving annual average growth rates for the period 2001-2019. These data refer to arrivals of tourists in accommodation establishments (herein called accommodation statistics) for a number of 786 municipalities for which data is available.

Main findings

It has been determined that one third of the localities listed in PATN as having high and very high concentration of tourism resources are not registering any flows of tourists in accommodation establishments. More, a superior performance of tourism demand was observed for localities not listed in PATN as compared with localities that were listed in PATN (an annual average growth rate of 8.6% versus 5.6%). This is quite surprising and may question the methodology underpinning PATN.

Originality/value

This paper is the first one that wants to challenge the official categorizations of localities in Romania from PATN in terms of tourism resources.

Conclusions/Recommendations

A revision of the methodology pertaining to PATN categorization of municipalities is recommended in order to include also variables of existing tourism demand at local level.

***Keywords**: tourism resources, accommodation statistics, territorial planning, locality, Romania*
***JEL classification**: Z32, O21*

1. INTRODUCTION

In 2008 Romania adopted the National Plan for Territorial Planning (rom. Planul de Amenajare a Teritoriului Național – PATN) section VIII – Areas with tourism resources officially stated in the Governmental Emergency Ordinance (OUG) 142 from 28 October 2008, subsequently updated in the Law no. 190/2009. It is envisaged that for the localities listed in these official documents "tourism is considered a priority economic activity and investments for developing this activity will be mainly oriented to these areas" (A/N localities). In PATN, an exhaustive identification of all administrative territorial units (localities) classified upon the concentration of tourism resources is performed. Also each locality is classified upon the type of dominant tourism resources: natural resources, man-made resources and mixed.

Almost 1,200 localities are listed in PATN, these representing little over 37% of the total localities in Romania (at 15th of July 2021 in Romania there were 3,181 localities defined as territorial administrative units according with Ministry of Development, Public Works and Administration (2021)). Most of these localities identified in PATN are classified as having a high concentration of tourism resources (989) while almost 200 localities are classified as having a very high concentration of tourism resources. The distinction between these two types of classification (high concentration versus very high concentration) is given by the score obtained, over 25 from maximum 50 for localities with very high concentration of tourism resources and a score between 14 and 24.99 for localities with high concentration of tourism resources according to the methodology for the analysis of tourism potential of territory approved by the Order no 549/518/2016 of vice-prime minister, minister of regional development and public administration and of vice-prime minister, minister of economy, commerce and relations with business community. It had to be mentioned that not all localities in Romania are listed in PATN (it can be deducted that those localities having a score under 14 are not listed in PATN). Therefore this categorization is strictly based on the tourism potential of the territory which resulted from the tourism resources each locality has. Tourism potential is considered an essential condition of tourism development in a certain territory (Minciu, 2004).

The purpose of this paper is to see to what extent the categorization from PATN is reflected in the evolution of tourism flows (as part of tourism demand) in the period 2001-2019. In other words, the aim is to see how the categorization from PATN is seen from the perspective of demand-side (a component that was not included in the methodology pertaining to PATN).

It is important to mention that tourism is among the six sectoral plans of PATN that represents the core component of national spatial development policy in Romania. These are: transport network, water management, protected areas, settlement network, natural risks and tourism (Benedek, 2013). Therefore, the importance of tourism from a spatial development policy perspective is clearly acknowledged.

2. METHODOLOGY

In a first step, both the number of localities listed in PATN and those localities that have data regarding number of tourists in accommodation establishments (herein named accommodation statistics) will be identified. This has been done by a cross-tabulation of localities from both accommodation statistics and PATN. Data regarding the number of tourists in accommodation establishments are provided by National Institute of Statistics (INS) with an annual periodicity in the period 2001-2019 (INS, 2021). 2001 is the first year when data is publicly available while 2020 has not been included in this paper due to strong impact of COVID-19 pandemics upon tourism sector in 2020 that certainly distort data series evolution in the last years. Subsequently, for each locality an analysis will be performed regarding the trend in number of tourists by types of dominant tourism resources of locality (natural, man-made and mixed) and by types of classification of localities (i.e. localities with high concentration of tourism resources and localities with very high concentration of tourism resources).

Then for each locality an annual average growth rate (AAGR) is calculated for the period 2001-2019 and AAGR should be seen in relation with the national level. AAGR also known as compound annual growth rate shows an average value for the annual rate of change over a period of time (Eurostat, 2021a). This rate is known in the Romanian statistical literature as being as average growth rate (rom. ritmul mediu de creștere $\overline{R}$) and is calculated as a difference between the dynamics average index ($\overline{I}$) expressed in percentage and 100% (Anghelache and Manole, 2012):

$$\overline{R} = (\overline{I} \text{ x } 100) - 100 \qquad [1]$$

where

$$\overline{I} = \sqrt[n-1]{\frac{Yn}{Y1}} \qquad [2]$$

for data series noted from 1 to n, where n =19 (period 2001-2019) but it might be different from locality to locality depending on data availability.

It is assumed that the usage of AAGR facilities the comparability between different periods more precisely "periods of different lengths, for example, comparing annual, five-yearly and ten-yearly rates" (Eurostat, 2021a).

A special treatment is given to localities that are not listed in PATN where the trend in accommodation statistics has been differentiated into two main periods: 2001-2008 and 2009-2019. These two periods have been considered taking 2008 as the benchmark year (of breaking the two periods) when the regulation (OUG 142/2008) has entered into force. An overall comparative analysis will be made between the trend of localities that are listed in PATN and localities that are not part of PATN in these two periods; this will be made also by using AAGR. In addition, taking 2001 as a base year, an illustration of dynamics index cu fixed base will be made between localities that are listed in PATN and localities that are not part of PATN. This index is calculated as a ratio between the level of each year and the level of the base year (Anghelache and Manole, 2012).

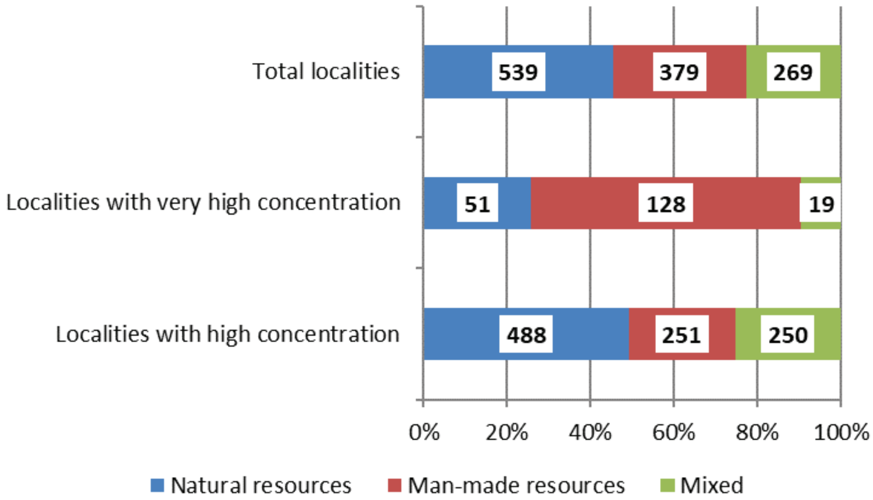
3. RESULTS

3.1. A general overview of the municipalities within PATN in Romania

First of all, it is important to have a general overview of the distribution of localities in PATN based on the dominant tourism resources: natural resources, man-made resources and mixed resources as well as the degree of concentration of tourism resources: very high concentration and high concentration (see figure 1). One can see that over 45% of localities (539 in absolute values) identified in PATN have natural resources as dominant tourism resources while man-made resources accounts for almost 32% of the total localities listed in PATN (379 in absolute values). However, if we consider strictly the localities with very high concentration of tourism resources, man-made resources are dominant (almost two-thirds of localities). Instead, if we look at the localities with high concentration of tourism resources almost half have natural resources as being the dominant tourism resources. It is important to add that there are in total 198 localities with very high concentration of tourism resources and 989 localities with high concentration of tourism resources; in other words, 16.6% localities from PATN have very high concentration of tourism resources while the rest have high concentration of tourism resources.

The typology of localities listed in PATN in Romania

Figure 1



Source: author's calculations based on PATN section VIII – Areas with tourism resources

At the same time, if we consider strictly the localities with man-made tourism resources (379), most of them (251) are found in the localities having a high concentration of tourism resources. It has to be reminded that a number of 379 localities have been identified as having man-made tourism resource as dominant tourism resources, representing 12% of all localities in Romania. As regards strictly localities with natural resources, a much higher predominance (almost 90%) is to be found in the localities with high concentration on tourism resources.

Also it is important to present the number of localities in Romania in a cross classification between their inclusion in PATN and in the accommodation statistics provided by INS (data on number of arrivals of guests in accommodation establishments). Once again, it has to be reminded that in total there are 3,181 localities in Romania (see table 1).

Number of localities in Romania by their inclusion in PATN and in accommodation statistics in the period 2001-2019

Table 1

Localities » »	Found in accommodation statistics (INS)	Not found in ccommodation statistics (INS)	**Total**
Listed in PATN, total, of which with	786	401	**1.187**
Very high concentration of tourism resources	*182*	*16*	*198*
High concentration of tourism resources	*604*	*385*	*989*
Not listed in PATN	484	1.510	**1.994**
Total	**1.270**	**1.911**	**3.181**

Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

From the above table one can see that almost two thirds (66.2%) of the localities listed in PATN were also found in accommodation statistics which corresponds to a number of 786 localities in absolute terms. At the same time, one third of localities listed in PATN (in fact 401 localities) listed as having high (385) or very high (16) tourism resources do not register any tourists flows according with accommodation statistics in Romania. For these localities there is a need to have a detailed analysis within PATN in order to see if they reported to have or not accommodation establishments. It has to be mentioned that according with the methodology pertaining to PATN elaboration (Order no 549/518/2016) the existence of accommodation establishments in a locality is scored with maximum 7 points (out of total 100). The author considers that this is a very low score considering accommodation is a basic product which determines to a greater extent the volume of tourism activity in a locality. In fact, in the tourism literature accommodation services is one of the four specific tourism services together with transportation, food and beverage and entertainment (Minciu, 2004).

At the same time, considering only the number of localities found in accommodation statistics, one can say that almost 62% of the total localities that register tourism flows in Romania are localities listed also in PATN (786 localities); though an important share (38%) is represented by localities that are not listed in PATN (484 localities were not listed in PATN but they registered tourism flows).

If one looks only at the number of localities classified upon the concentration of tourism resources, only 16 out of 198 localities with very high concentration of tourism resources (representing 8%) were not found in accommodation statistics. But in the case of localities with high concentration of tourism resources almost 40% of them (385 localities) did not reported any tourism flows. This seems to be a very important share which might challenge the PATN classifications.

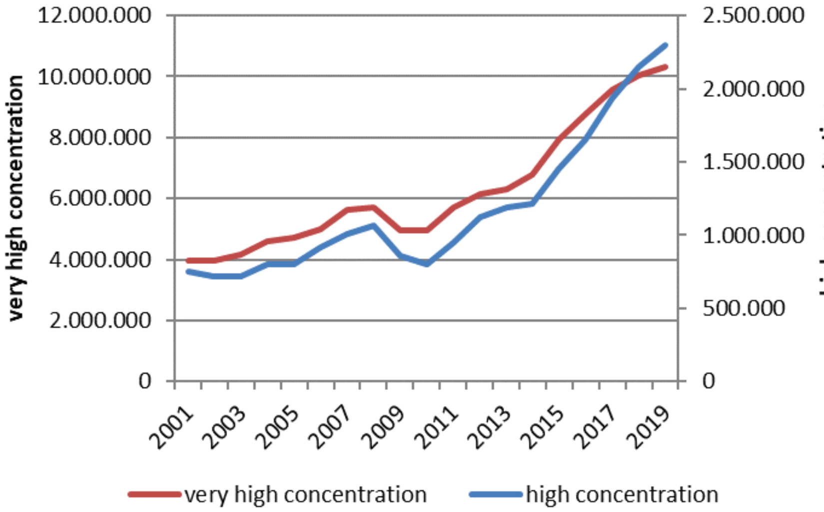
Not the least, if the overall situation is envisaged (considering all localities in Romania) one can say that only 40% of these (1,270 localities) registered tourists in accommodation establishments at least one year in the period 2001-2019. From an opposite perspective, one can say that 60% of localities in Romania are not involved in any tourism activity which shows a certain territorial concentration of tourism activity in Romania as well as limited spatial distribution. From this perspective, tourism sector in our country can increase territorial disparities rather than contribute to their reduction (Cehan et al, 2019).

3.2. Tourism flows in the muncipalities listed in PATN

If we consider strictly the number of tourists only in localities for which data is available (i.e. the amounts of tourism flows), the situation looks different than in the case where only the localities were counted. As showed in table 1 there are 786 localities in PATN that have data regarding the number of tourists in accommodation establishments, at least one year in the period 2001-2019. But it is important to recognize that the localities listed in PATN accounts for in average 95% from the total number of tourists registered in Romania. Most of the tourists have been registered in localities in PATN for which there is a very high concentration of tourism resources (in average over 80% from total localities listed in PATN) while in localities with high concentration of tourism resources fewer tourists have been registered (20% of all localities listed in PATN). One can see that the trend of these two components is similar: after the fall in 2009 and 2010 (a direct consequence of 2009-10 economic crisis) in the period 2011-2019 a robust growth of number of tourists is can be noticed starting 2011 (see figure 2).

Number of tourists in accommodation establishments in the localities listed in PATN classified upon the concentration of tourism resources, 2001-2019

Figure 2

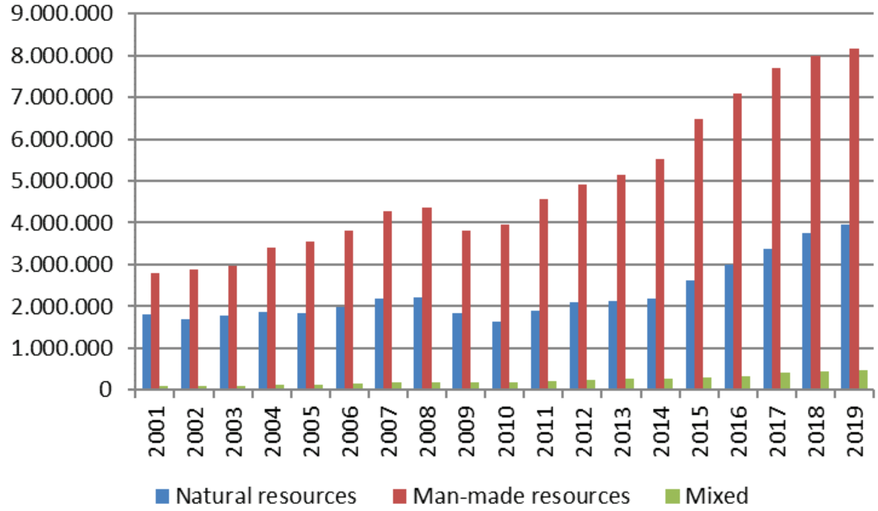


Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

As regards types of dominant tourism resources of localities listed in PATN, in terms of number of tourists, by far man-made tourism resources are best represented. In absolute terms, the PATN listed localities with man-made tourism resources as dominant tourism resources register flows of tourists two times higher than localities in PATN with natural resources as dominant resources (see figure 3). Localities listed in PATN with mixed resources as dominant resources occupy more than a modest place in terms of number of tourists, growing in absolute terms from almost 96 thousands in 2001 to over 479 thousands in 2019.

Number of tourists in accommodation establishments in localities listed in PATN classified upon the dominant tourism resources, 2001-2019

Figure 3

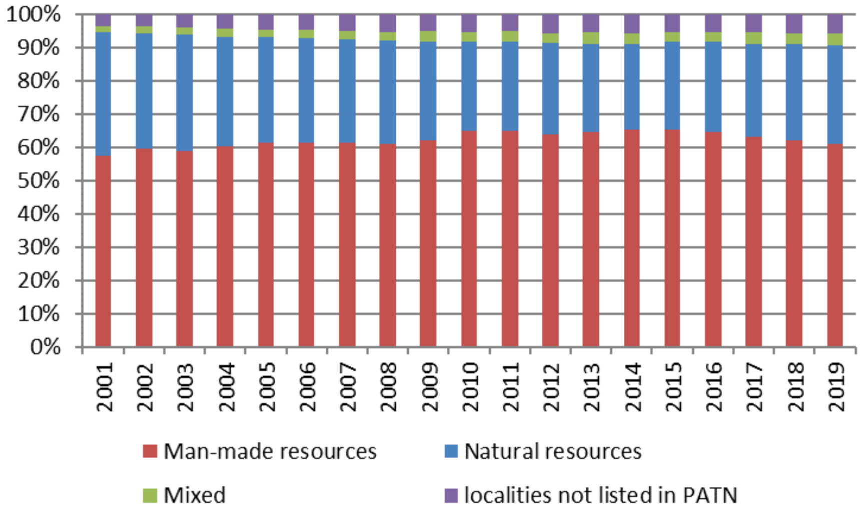


Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

In average, in the period 2001-2019, over 62% of the number of tourists in accommodation establishments in Romania were in localities listed in PATN as having man-made tourism resources as dominant resources. Of course, this share fluctuates from little over 57% in 2001 to 65% in 2015. At the same time, in average almost 30% of the number of tourists in accommodation establishments in Romania were in localities listed in PATN as having natural tourism resources. It has to be remarked that localities not listed in PATN have increased their share from 3.6% in 2001 to 5.8% in 2019 (see figure 4).

The distribution of tourists in Romania upon dominant tourism resources of localities listed in PATN and localities that are not listed in PATN, 2001-2019

Figure 4



Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

At the same time these figures have to be put in relation with the national level in terms of annual average growth rate for number of tourists in accommodation establishments (5.8%). It has to be observed the higher rate for localities listed in PATN with man-made tourism resources as dominant resources and very high concentration of tourism resources (6.2%) while for localities listed in PATN with natural resources as dominant resources and very high concentration of tourism resources this rate is much lower (3.5%) (see table 2).

Annual average growth rate for number of tourists in the period 2001-2019

Table 2

	Value (as %)
Total Romania	5.8
Total localities listed in PATN	5.6
Total localities with natural resources as dominant tourism resources, of which	4.5
Very high concentration	3.5
High concentration	6.5
Total localities with man-made resources as dominant tourism resources, of which	6.1
Very high concentration	6.2
High concentration	5.3
Total localities with mixed resources as dominant tourism resources, of which	9.6
Very high concentration	12.6
High concentration	8.2

Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

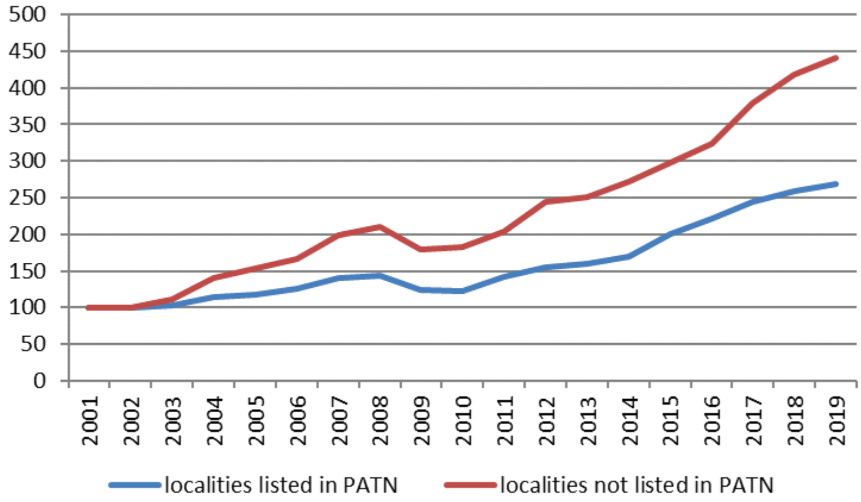
3.3. A different evolution of tourism in localities that are not listed in PATN?

As mentioned before, out of 1,270 localities from which INS provides data for at least one year in the period 2001-2009 within accommodation statistics, a total number of 484 localities (38.1%) were not listed in PATN. From an analytic point of view it is important to analyse the evolution of tourism in these localities. However, it has to be reminded once again that the number of tourists in localities that are not listed in PATN represents a low share in the total number of tourists in Romania but it is remarkable that this share has increased from 3.6% in 2001 to 5.8% in 2019.

Although the same general pattern of increasing the number of tourists in the period 2001-2019 is seen (see figure 5), one can observe interestingly a superior dynamics of localities that were not listed in PATN compared with localities that are listed in PATN. Also, it can be added that the annual average growth rate of localities that are not listed in PATN is much higher compared both with localities that are part of PATN and general national level, namely country level (8.6% compared with 5.6% and 5.8%).

Dynamics of number of tourists in localities listed in PATN versus in localities not listed din PATN, 2001-2019 (2001=100)

Figure 5



Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

Also, it is important to illustrate this evolution of tourism broken down into two main periods: 2001-2008 and 2019-2019. Again, one can see that localities that are not listed in PATN have performed superior both to the general country level (national level) and to localities that are listed in PATN (see table 3). However, there is a slowdown in the growth in the period 2009-2019 compared with the period 2001-2008. It is not clear yet whether this evolution is due to the fact the localities were not included in PATN, Section VIII, Areas with tourism resources. At the same time, one can see that localities with high concentration of tourism resources have an annual average growth rate in the period 2009-2019 almost double compared to the previous period 2001-2008 but once again in is not clear whether this is due to their inclusion in the list within PATN. Meanwhile, it has to be mentioned that 2009 was the year when tourism was strongly affected by the economic crisis, registering a decline compared with the previous years; therefore the basic level (the one referring to comparability) in the year 2009 was very low and this fact has somehow an influence in the analysis of the two analyzed periods 2001-2008 and respectively 2009-2019.

Annual average growth rate for number of tourists with differentiation for the period 2001-2008 and 2009-2019

Table 3

	Annual average growth rate for (%)		Difference in percentage points (B) - (A)
	2001-2008 (A)	2009-2019 (B)	
Total Romania	5.6%	8.1%	+2.5
Localities that are **not** listed in PATN	**11.3%**	**9.4%**	**-1.9**
Localities that are listed in PATN, of which with	5.3%	8.0%	+2.7
Very high concentration	5.4%	7.6%	+2.2
High concentration	5.0%	10.3%	+5.3
Man-made resources	6.5%	7.9%	+1.4
Natural resources	2.9%	8.0%	+5.1%
Mixed resources	11.4%	9.7%	-1.7%

Source: author's calculations based on PATN section VIII – Areas with tourism resources and INS (2021) data on accommodation statistics

4. CONCLUSIONS

This paper wants to draw attention about the need to discuss and possible revise the official methodology underpinning PATN Section VIII Areas with tourism resources. It is not clear, from the perspective of tourism demand how the categorization from PATN helped or not the development of tourism at municipality level in Romania. This is an issue of great importance since public investments in tourism (including the allocation of EU funds) take as a reference PATN. For instance, in the period 2007-2016 according to a study commissioned by Ministry of Tourism, within Regional Operational Program, major intervention area Urban poles for Growth (where funds were very significant) only municipalities having high and very high concentration of tourism resources benefited from EU funds (INCDT, 2017). The same study (p. 61) shows that localities having high and very high concentration of tourism resources cumulated 96% from the total value of "investment projects with tourism relevance" while 4% were allocated in localities that are not listed in PATN. However, despite these tourism investments, the evolution of tourism demand in localities listed in PATN has not experienced a higher dynamics than the national level and this may question the very official methodology by which localities are categorized in PATN.

One can consider that there is need to consider also the trend in tourism demand (in accommodation statistics we have number of arrivals of tourists in accommodation establishments, data available for each locality (INS, 2021)) as an element of analysis of tourism in each locality. The approach from PATN was based strictly on tourism supply-side elements (tourism resources, infrastructure) and has its limitations that can condition somehow unjustifiably the tourism development at municipality level. By including an analysis of tourism demand in the methodology pertaining to PATN, Section VIII Areas with tourism resources, a better connection with tourism market and the trend in tourism demand at local level will be provided. In other words, a revision of the methodology from PATN is recommended by including in the score of each locality also an index of tourism demand trend, data available in accommodation statistics. Another possibility is to include an indicator of tourism density defined either as a ratio between number of overnight stays to total area of a locality or as a ratio between number of overnight stays to resident population.

Of course, there is no guarantee that the trend in tourism demand from a past period will also continue in the future but one can consider that such index will allow a better connection with the trend of tourism market in each locality. Finally, one can consider that a better evaluation of tourism resources at municipality level can be made in PATN, section VIII Areas with tourism resources if one can consider also elements of tourism demand and/or including also the indicator of tourism density – see Eurostat (2021b).

However, it has to be admitted that the major limitation is given by the nature of accommodation statistics. It comprises only establishments licensed by the ministry in charge of tourism affairs neglecting accommodations in other forms of accommodation (such as staying at friends or relatives, in own vacation homes, or in non-licensed accommodation units). Moreover, the so called border effect can be envisaged meaning that at local level, a municipality is affected by the fact that guest are staying in a bordering municipality (Bohlin et al., 2016), thus the accommodation statistics fails to capture the entire guest flows in a municipality. Nevertheless, in order to minimize the border effect issue a detailed analysis is required especially in localities where there are no accommodation establishments (and thus no tourist flows registered). When analysing tourism demand in such localities (it has to be remember that there are more than 400 localities that are found in PATN but not in accommodation statistics) a sort of correction factor can be envisaged based on distance to the nearest accommodation establishments, for instance in a threshold of maximum 30 km.

Another extension of the research can be made by considering also the length of stay of tourists thus using the indictor of number of overnight stays. This can be made either in combination with the indicator of arrivals or by totally replacing the indictor of arrivals in accommodation units with overnight stays in accommodation units. Indeed in the statistical practice, in counting the tourism flows a higher relevance is given to number of overnight stays since this indicator avoids any double-counting risk which might occur during multi-destination trips undertaken by tourists. But one can consider that at the municipal level this is not the case since it is implicitly presumed that a guest staying in an accommodation establishment is a unique tourist for a locality.

5. REFERENCES

1. **Anghelache C.** and **Manole A**., 2012, *"Serii dinamice / cronologice (de timp) – prezentare, structură, relații dintre indici (en. Dynamic/chronologic (time) data series – presentation, structure and relation between indices)"*, *Romanian Statistical Review*, nr. 10/2012, pp. 68-77.
2. **Benedek J.**, 2013, *"The spatial planning system in Romania"*, Romanian Review of Regional Studies, volume IX, number 2, pp. 23-30.
3. **Bohlin M., Brandt D.** and **Elbe J.**, 2016, *"Tourism as a vehicle for regional development in peripheral areas – myth or reality? A longitudinal case study of Swedish regions"*, European Planning Studies, 24:10, pp. 1788-1805.
4. **Cehan A., Iațu C., Eva M et al**., 2019, *"Territorial dynamics of tourism in Romania: a long term-perspective (1990-2016)"*, Human Geographies – Journal of Studies and Research in Human Geography, vol. 13, no. 1, pp. 23-44.
5. **Eurostat**, 2021a, *"Glossary: Annual average growth rate (AAGR)"*, available at https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Annual_average_growth_rate_(AAGR) (accessed 20 September 2021).
6. **Eurostat**, 2021b, *"Tourism statistics at the regional level"*, available at https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Tourism_statistics_at_regional_level&oldid=515132#Tourism_pressures (accessed 21 September 2021).
7. **INCDT Institutul Național de Cercetare Dezvoltare în Turism (en. National Institute of Research Development in Tourism)**, 2017, *"Studiu de cercetare privind dezvoltarea investitiilor publice în infrastructura specific turistică din România pentru perioada 2007-2016 și direcțiile principale de dezvoltare a acestora (en. Research study regarding the development of public investments in tourism infrastructure for the period 2007-2016 and the main dirrections for developing of these investments)"*, București.
8. **INS - Institutul Național de Statistică (en. National Institute of Statistics)**, 2021, *"Tempo database"*, available at http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table (accessed 25 February 2021).
9. **Minciu, R.**, 2004, *Economia Turismului (en. Tourism economy)*, Editura Uranus, București.
10. **Ministry of Development, Public Works and Administration**, 2021, *"Numărul și clasificarea Unităților Administrativ Teritoriale și Sectoarelor din România (en. Number and classification of Territorial Administrative Units and Sectors in Romania)"*, available at http://www.dpfbl.mdrap.ro/nr_uat-uri.html (accessed 22 July 2021).