Romanian Statistical Review

Revista Română de Statistică

THE JOURNAL OF NATIONAL INSTITUTE OF STATISTICS

DATA-DRIVEN VISUALISATIONS FOR OFFICIAL STATISTICS – A CASE STUDY IMPLEMENTING CORPORATE DESIGN IN R

DIFFERENCE ON EVALUATION SCORES CONSIDERING IMAGE
DESCRIPTIONS FOR AUTOCODING

IMPROVING THE QUALITY OF CIVIL SERVANTS' PROFESSIONAL TRAINING PROCESS THROUGH DIGITALIZATION - AN EXPLORATORY RESEARCH FROM EMPLOYEES' PERSPECTIVE

IS GOOGLE TRENDS USEFUL IN NOWCASTING UNEMPLOYMENT RATE DURING THE PANDEMIC AT REGIONAL AND NATIONAL LEVEL IN ROMANIA?

1 /2024

www.revistadestatistica.ro

CONTENTS 1/2024

Data-driven visualisations for official statistics – A case study implementing corporate design in R

Hendrik Christian Doll[1] (hendrik.doll@bundesbank.de)
Deutsche Bundesbank, DG Data and Statistics

Daniel Ollech[2]
Deutsche Bundesbank, DG Data and Statistics

ABSTRACT

We introduce the {bbkplot} package, an easy-to-use R package to create graphics in the Deutsche Bundesbank corporate design. Thereby, we facilitate user-friendly and reliable production of data-driven graphics in exact accordance with the design guidelines. Lessons learned apply to many statistical offices and providers of data that aim to communicate the reliability of their published statistics through a unified corporate design.

The presented package allows the conversion of figures and tables into graphics precisely adhering to the corporate design standards. We provide an implementation for a complex corporate design standard that specifies formatting requirements for graphics in great detail. {bbkplot} automates the exact application of these requirements and enables a quick and easy production of high-quality graphics. Therefore, we aspire for {bbkplot} to serve as a blueprint for statistical offices to implement user-friendly pipelines using R to create high-quality output.

We see multiple value propositions from the package. (1) By advancing from pre-specified graphic types to flexible plot creation, {bbkplot} efficiently allows the creation of a wide range of graphics in corporate design. (2) By facilitating graphics creation for a wide range of publications, we support a more comprehensive external brand appearance for providers of official statistics. (3) By demonstrating an implementation of a complex corporate design in R, we hope to enable providers of official statistics to implement similar solutions.

[1] All views expressed in this paper are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

[2] The authors would like to thank Frederik Hering, Siddharth Marya, and Robin Kollmann for valuable support in developing the package and Carolin Schaaff for revisions.

Keywords: *R-package, data dissemination, statistics communication*
JEL classification: *C80, C88, Y10.*

1. INTRODUCTION

The past decade has witnessed immense growth in the amount of collected data and its use for analysis. Amidst the wealth of information produced, disseminated, and consumed, a large demand for reliable information remains. To answer this need, official statistical offices have recognised that efficient branding is paramount to convey reliability of the information they provide (UNECE, 2019).

Deutsche Bundesbank, as the Central Bank of Germany and part of the system of official statistics, regularly publishes a large number of reports, papers, and presentations. Many of these publications disseminate results of data-driven analyses. All publications have a unified brand appearance using a common corporate design. The common design aspires to enable readers to recognise easily the reliability of the source of information and to guide readers through publications in a structured way.

To achieve a common design, detailed design guidelines are available internally to enable all staff to streamline their external communication accordingly. The need for a unified branding is prevalent among many official statistical offices (UNECE, 2019). In our case study, design guidelines describe requirements on more than 150 pages, specifying everything from typography, colours, to line-by-line requirements for all chart elements (Deutsche Bundesbank, 2011).

To date, a dedicated team using commercial software, based on data and suggestions from the respective experts, compiles figures and graphics for publication manually. While this process ensures a high quality and coherent design of output visualisations for flagship publications, for a large and increasing number of standard visualisations for usage in less visible publications, this process can be time-consuming. This may lead to a situation, where analysts build their own graphics, which only partially adhere to the corporate design standard and thereby reduces overall coherency of the brand.

We introduce the {bbkplot} package, an easy-to-use open source software in R that produces high-quality graphics exactly adhering to Bundesbank corporate design (Doll et al., 2020 present a previous version). {bbkplot} is a user-friendly internal R package.[1] The package allows fast and independent creation of graphics by all staff. {bbkplot} allows to create a wide range of plots, going beyond what was representable in corporate design

[1] The name of the package derives from the abbreviation “bbk”, an often-used acronym for Bundesbank, in combination with “plot”, which refers to figures and charts in R.

to date. This provides interesting applications as novel visualisations such as interactive graphics become more popular in modern visualisations of official statistics (Forbes et al., 2011; Ten Bosch and de Jonge, 2008).

We see multiple value propositions from the package. First, by advancing from pre-specified graphic types that traditional graphics suites implement to flexible plot creation, {bbkplot} efficiently allows the creation of a wider range of graphics in corporate design. Second, we aspire to improve the external brand appearance by facilitating a low-threshold user-creation of graphics for presentations and thereby a wider usage of corporate design. Third, the application of our lessons learned applies to institutions beyond our case study. By highlighting a detailed user-friendly implementation of a complex corporate design in R, we hope to inspire official statistics providers to implement similar solutions.

2. CORPORATE DESIGN

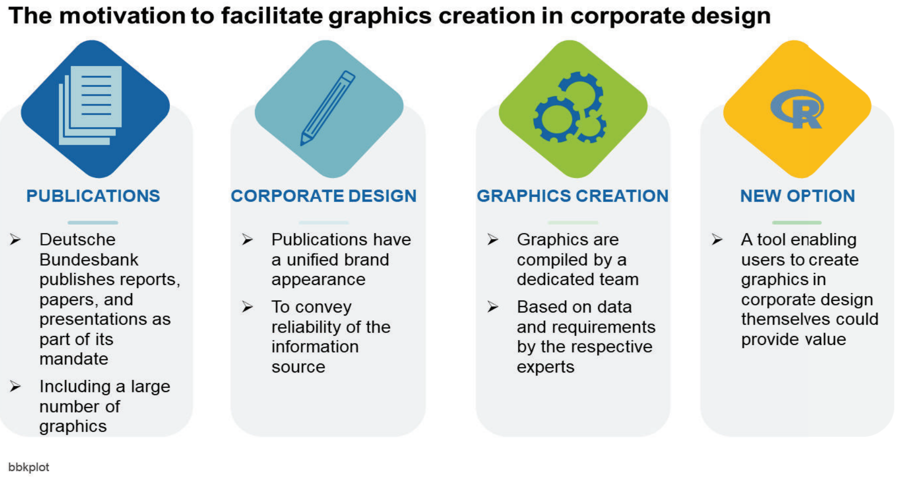
Efficient communication for official statistics

Since the communication of data-driven results becomes increasingly important, it is crucial to mandate a corporate design that ensures a consistent public appearance in published visualisations. Corporate design is an effective tool for both publishers and consumers of information as it enables publishers to establish a visual identity and supports reputation (Van den Bosch et al, 2005). At the same time, it facilitates users to verify the source of information and its trustworthiness. The benefit of using such guides is not limited to generating consistency; it also leads to efficiency gains by saving time in generating documents and creating a professional look in documents (Allen, 1996).

While trustworthiness refers to many issues, when producing the actual data, one important aspect is communication. In this regard, UNECE (2019) raises the need for statistical offices to enhance and promote statistical organisations' reputation, with branding being at the core. Other statistical offices also focus in part on building their brand through innovative forms of visualisations (Destatis, 2020).

The motivation for Deutsche Bundesbank and Statistical Offices in general to facilitate graphics creation in corporate design

Figure 1



The motivation to facilitate graphics creation in corporate design

PUBLICATIONS

- Deutsche Bundesbank publishes reports, papers, and presentations as part of its mandate
- Including a large number of graphics

CORPORATE DESIGN

- Publications have a unified brand appearance
- To convey reliability of the information source

GRAPHICS CREATION

- Graphics are compiled by a dedicated team
- Based on data and requirements by the respective experts

NEW OPTION

- A tool enabling users to create graphics in corporate design themselves could provide value

bbkplot

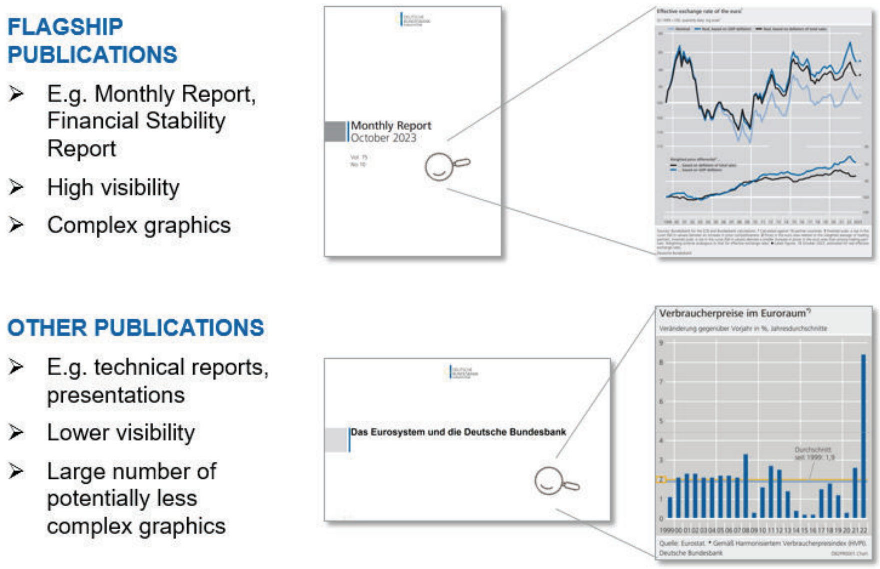
Source: Own depiction. R logo © 2016 The R Foundation.

Providers of official statistics in general and Deutsche Bundesbank in particular therefore realise the need to streamline their external communications in a coherent corporate design in all communications. This ensures a unified brand appearance and conveys reliability of the source. Until recently, designed, and formatted output in terms of publications is created in a rather manual process by dedicated communications departments. For textual documents, recent advances have been made to create publication-ready documents using Quarto[1] (e.g. Gomolka, 2018). For graphics, a tool to enable users to produce their own publication-ready visualisations, would fit right into such pipelines leveraging on capabilities tools such as R and Quarto documents offer (compare Figure 1 for a graphical depiction of the motivation).

[1] Quarto is a formatting syntax, similar to R Markdown, allowing creating, among others html and pdf output from within RStudio, a popular compiler for R. Using Quarto enables to combine R, Python and Julia Code into an easily reproducible output document.

Examples of publications from Deutsche Bundesbank exemplary for the data-driven communication that takes place in Central Banks and Official Statistical Offices

Figure 2



FLAGSHIP PUBLICATIONS

- E.g. Monthly Report, Financial Stability Report
- High visibility
- Complex graphics

OTHER PUBLICATIONS

- E.g. technical reports, presentations
- Lower visibility
- Large number of potentially less complex graphics

Monthly Report
October 2023
Effective exchange rate of the euro
Das Eurosystem und die Deutsche Bundesbank
Verbraucherpreise im Euroraum*)
Veränderung gegenüber Vorjahr in %, Jahresdurchschnitte
9
8
7
6
5
4
3
2
1
0
Durchschnitt seit 1999: 1,9
1999 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22
Quelle: Eurostat. * Gemäß Harmonisiertem Verbraucherpreisindex (HVPI).
Deutsche Bundesbank

Note the structure with a header and footer, both with different aligned text elements and with the body of the graphic including multiple diagrams and non-standard axes. These diverse elements necessitate a tailor-made solution: Own depiction based on Deutsche Bundesbank (2023a) and Deutsche Bundesbank (2023b)

In the context of our present case study, the corporate design guide mandates that all figures have the same structure. It contains three parts, namely (i) the header, (ii) the plotting area, and (iii) the footer. The header consists of the title, subtitle and number. The size of the header must be at least two times the format size and can be increased if more space is required. The plotting area contains at least one figure and optionally the legend. The footer contains a label with ”Deutsche Bundesbank” and an optional mention of the source. It is further possible to move the legend into the footer. The three distinct parts of a figure are separated by a white line of width 2 points (cf. Figure 2 for examples of published graphics that contain the mentioned elements).

The title of the graphic is set in the size of the continuous text. This can range from 9 to 12 pt. All other text element sizes are relative to the title. The size of supplementary text is scaled to 70 per cent of the title size. However, this should be at least 6 pt. The font style of the title is Frutiger Next Medium. The font style for the supplementary text is Frutiger Next Light. The style guide further requires 2 mm space on the borders of every element, certain margins between text lines, and specific line-by-line requirements for all plot elements, such as lines, dots, bars, areas.

Further extended guidelines exist for exact colour codes and their mandated shadings, axis formatting, grid lines, legend placement and various special types of charts (such as box plots, bar charts, maps, etc.). As corporate design guidelines such as the ones outlined in this section exist in many large organisations, it is perhaps unsurprising that a number of solutions for facilitated corporate implementation in open-source software exist. In the next section, we outline available solutions.

Existing Corporate Design Packages

Given the thrust towards open source technologies, it is not surprising that several authors have developed software and libraries that accommodate their respective corporate designs using open source software R. Given that users, such as analysts and journalists, become increasingly data-savvy, there is a large efficiency gain in giving users the proper tools to convert their figures the right away.

We draw inspiration from {bbplot} (Stylianou et al., 2018), which is an open source R package built by the British Broadcasting Company (BBC) for the production of figures. The BBC team went on to publish their charts using their package in 2018. A starting point for the package was the implementation of corporate colour palettes as described by Jackson (2018). Bion, Chang, and Goodman (2017) outline how Airbnb uses R for data visualisation to create branded graphs in addition to other stages of the analysis pipeline.

There even exists a dedicated {ggthemes} package that allows plotting in the style of *The Economist, FiveThirtyEight, Wall Street Journal*, and others (Arnold et al., 2020). Another more general example is the {ggCorpIdent} package (Klein and Wagner, 2018) which allows implementing any basic corporate design for {ggplot2} graphics. Neth and Gradwohl (2023) developed the {unikn} package that implements corporate design colours and text functions of the University of Konstanz and allows to adapt these functions to the needs of other institutions.

As we have discussed above, the Bundesbank use-case however incorporates rules that go beyond a simple {ggplot2} implementation using themes or any of the above-mentioned packages. These requirements drive our approach of relying on {grid} and {gridExtra} that provide the primitive functions, which {ggplot2} builds on. We create separate graphical objects ("grobs") for all parts of the graphic (header, plotting area, footer, legend, lines) and combine them.

Non-standard axis ticks (especially for time series, compare Figure 2 for an example) drive us to take special consideration for the axes. While {ggplot2} is surprisingly flexible, the challenge lies in identifying the right major and minor tick position for any temporal data, according to the corporate design specifications needed in such a case. We resort to nested for-loops that draw each single tick iteratively.

A further challenge in designing the package is handling the non-standard default font (Frutiger) that is not present on all machines, where the package is used. As a mitigation, we build on the {sysfonts} package, attempting to load fonts and enable fallback options, if not possible.

3. THE {BBKPLOT} PACKAGE

Design

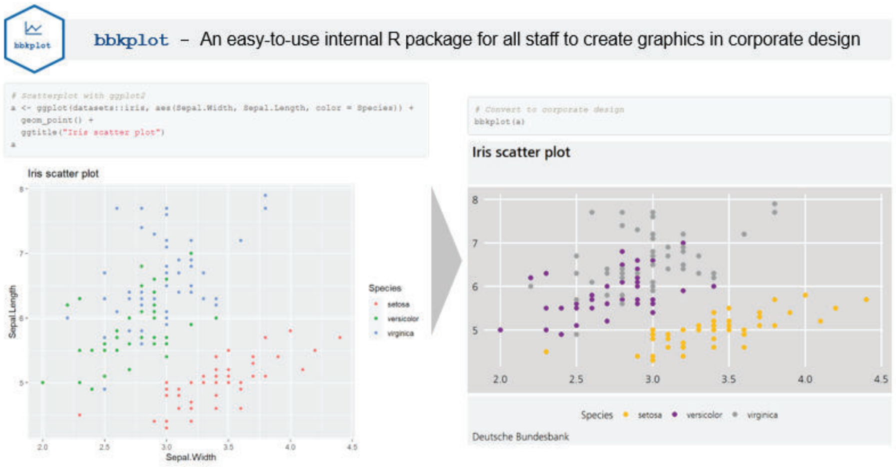
The primary goal of the {bbkplot} package is to develop an open source solution to visualise data adhering to the Corporate Design Manual of the Bundesbank.[1] It should fulfil the various corporate design requirements outlined above and enable Bundesbank employees to easily produce high quality graphics and include them in documents and presentations automatically. The complexity of our mandated corporate design prevents us from using the pre-existing R packages discussed before.

We developed the {bbkplot} package with a set of design objectives in mind. First, the functions shall be easy to use, even for inexperienced R users. Figure 3 exemplarily shows how we achieve this in practice with a one-line command. Second, the package should integrate well in the working process of R users, i.e. the notation and structure should be in line with the customs in R. Third, the package should strike a good balance between including as many plot types and relying on as few functions as possible.

[1] Due to concerns of unauthorized publication of graphics in corporate design, we are unable to publish the package source code, hence in the following, we describe the package and showcase its functionality and the applicability of lessons learned. In case of interest from providers from official statistics, we are happy to share code and lessons learned bilaterally.

Introducing {bbkplot} capabilities and easy handling

Figure 3



bbkplot – An easy-to-use internal R package for all staff to create graphics in corporate design

```
# Scatterplot with ggplot2
a <- ggplot(datasets::iris, aes(Sepal.Width, Sepal.Length, color = Species)) +
  geom_point() +
  ggtitle("Iris scatter plot")
a
```

```
# Convert to corporate design
bbkplot(a)
```

Source: Own depiction based on Iris dataset

With this in mind, a fruitful starting point is the {ggplot2} package developed by Wickham et al. (2019). The package is probably the most widely used and reputable visualisation package in R and allows far-reaching customisation of visualisations, which include all important plot types. Additionally, a large number of packages have been developed that extend {ggplot2} graphical capacities even further, including network, radar, geographic, and interactive graphics.

In principal, the {bbkplot} package is an add-on to {ggplot2}. Users can create graphics using the {ggplot2} infrastructure and translate the layout to the Bundesbank design using designated functions of the {bbkplot} package. However, we advance from {ggplot2}, by allowing for {grid} graphics and, to a limited extent, even to some extent {base} graphics. Furthermore, for time series, a function is available to plot the data directly.

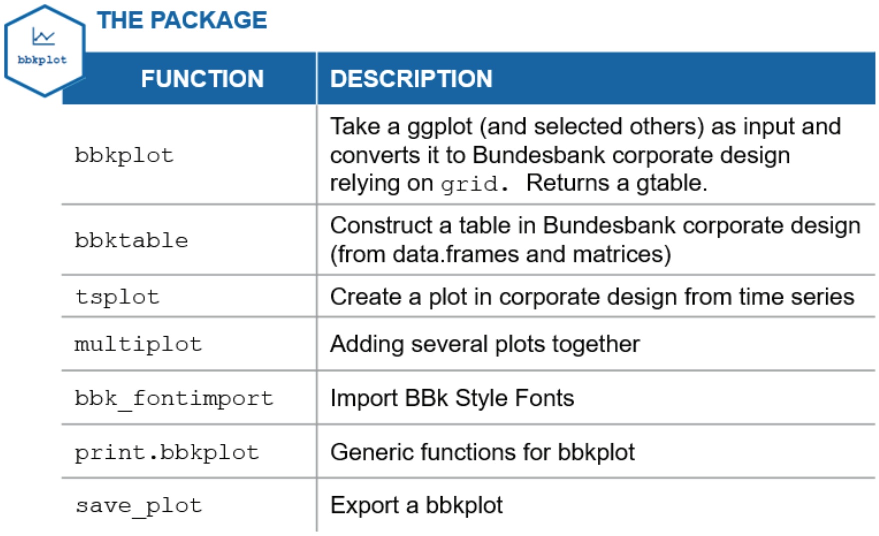
Functions

Currently, the {bbkplot} package includes functions for creating Bundesbank tables and graphics, changing the layout of a plot to the Bundesbank corporate design and helper functions, that import the Bundesbank font, define Bundesbank colour scheme and save plots and tables. Figure 4 shows the functions available to the user, several of which we will introduce in this section (please find a description of each functions' capabilities in Appendix 1)

The function `bbkplot()` is the workhorse of the {bbkplot} package. It converts `ggplot` objects into Bundesbank style graphics.[1] The function primarily uses the {ggplot2}, {grid} and {gridExtra} package and arranges the three distinct areas of a graph as mentioned above namely, the header, the plotting area and the footer.

Main functionalities of the {bbkplot} package

Figure 4



THE PACKAGE

FUNCTION	DESCRIPTION
`bbkplot`	Take a ggplot (and selected others) as input and converts it to Bundesbank corporate design relying on `grid.` Returns a gtable.
`bbktable`	Construct a table in Bundesbank corporate design (from data.frames and matrices)
`tsplot`	Create a plot in corporate design from time series
`multiplot`	Adding several plots together
`bbk_fontimport`	Import BBk Style Fonts
`print.bbkplot`	Generic functions for bbkplot
`save_plot`	Export a bbkplot

Source: Own depiction

There is a large number of optional arguments to the function call and these can be used to personalise the graphical object, e.g. by adding a title and subtitle to the plot, by placing the legend within the plot area. Appendix 2 summarises the whole range of optional arguments to personalise plots, while adhering precisely to corporate design using `bbkplot()`. Figure 3 shows the original `ggplot` (left) and the output in the form of a converted `bbkplot` object (right) in the default scenario, highlights the user-friendly design, with sensible settings for all optional parameters by default (Appendix 2 lists all available parameters).

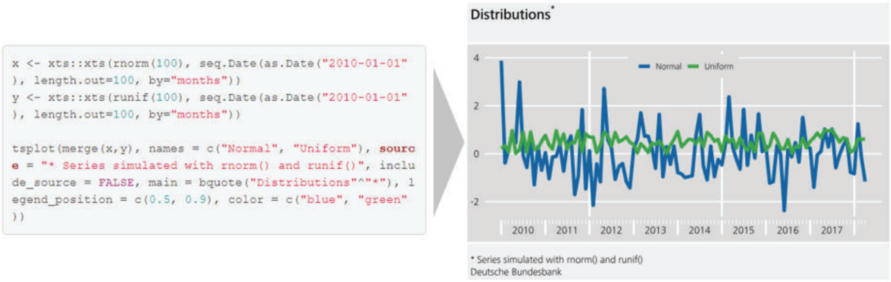
[1], The function also adds Bundesbank style header and footer to objects of class grid. These are often multiple plots generated by {ggplot2}. These grid objects need to be in the Bundesbank style for the whole plot generated by {bbkplot} to be in the style. This can usually be achieved by adding theme_bbk() to parts of the grid object.

A special function dedicated to plotting time series is included in the {bbkplot} package in order to facilitate an efficient visualisation of this data type, `tsplot()`. The additional benefit is twofold. First, because of the peculiarities of the temporal axes, reformatting time series is often quite cumbersome. Therefore, `tsplot()` accepts time series objects as input (ideally of class `xts` from the {xts} package) instead of a `ggplot` object as in `bbkplot()`. Second, `tsplot()` adds several functionalities for transforming the input and the output.

This is motivated by the detailed formatting requirements for time-series axis in our case, depending on the major axis ticks (e.g. yearly, monthly, weekly, daily) and the interval among them (minor ticks, e.g. quarter, months, weeks, days). As date axis ticks and labels follow very detailed style guidelines, special care is taken to ensure easy and user-friendly implementation of these (compare the x-axis in Figure 5).

Creating time series plots using the tsplot() function from {bbkplot} that offers a wide range of complex corporate design-adhering temporal x-axes that go beyond capabilities that can be implemented using {ggplot2}

Figure 5



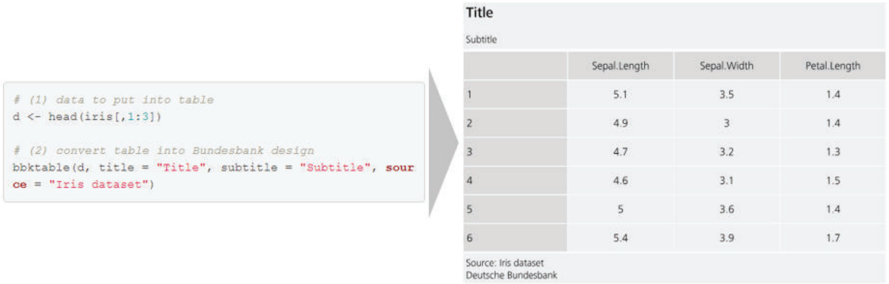
```
x <- xts::xts(rnorm(100), seq.Date(as.Date("2010-01-01"
), length.out=100, by="months"))
y <- xts::xts(runif(100), seq.Date(as.Date("2010-01-01"
), length.out=100, by="months"))

tsplot(merge(x,y), names = c("Normal", "Uniform"), sourc
e = "* Series simulated with rnorm() and runif()", inclu
de_source = FALSE, main = bquote("Distributions"^"*"), l
egend_position = c(0.5, 0.9), color = c("blue", "green"
))
```

Source: Own depiction.

The `tsplot()` function implements these options through easy to choose arguments for the user and otherwise relies heavily on the `bbkplot()` function. It is possible to directly load specified time series data from the Bundesbank database into R and then plot them in Bundesbank design, using further packages within the Bundesbank environment, e.g. the internal {rZISDB2}, which is used by analysts to load time series from the database into R (also using {xts}).

Using the bbktable() function to enable internal users to create publication-ready tables directly from within R

Figure 6



```
# (1) data to put into table
d <- head(iris[,1:3])

# (2) convert table into Bundesbank design
bbktable(d, title = "Title", subtitle = "Subtitle", sour
ce = "Iris dataset")
```

Title

Subtitle

	Sepal.Length	Sepal.Width	Petal.Length
1	5.1	3.5	1.4
2	4.9	3	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5	3.6	1.4
6	5.4	3.9	1.7

Source: Iris dataset
Deutsche Bundesbank

Source: Own depiction.

The function `bbktable()` can be used to create tables in the Bundesbank design. It accepts data.frames and matrices as input. Though the output looks like a table, it has a plot format. While the table therefore cannot be changed after its creation, such plots can be flexibly included in any publications and documents subsequently, i.e. LaTex, Excel and PowerPoint (compare Figure 6 for the resulting output). We outline examples, how to leverage on these functions in the next section.

Usage

This section provides examples for using the functions included in {bbkplot}. These examples illustrate how several types of plots can be created using the `ggplot()` function of the {ggplot2} package and then be converted to Bundesbank style using {bbkplot}. More and constantly updated examples can be found in the vignette of {bbkplot}. For all examples, the {ggplot2} and {bbkplot} package need to be loaded.

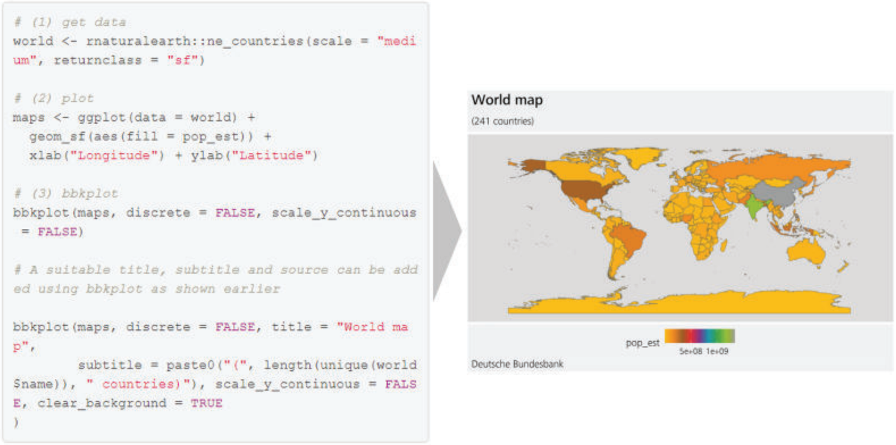
Generally, plots are created using {ggplot2}. Afterwards, the style is transformed according to the Corporate Design Manual using a call to functions in {bbkplot}. While these functions try to preserve all other settings defined in the original plot, some settings can be defined by the user both in {ggplot2} and in {bbkplot}. If for example, a title is defined in the original call to {ggplot} and in {bbkplot}, the title defined using {bbkplot} will have precedence.[1] Supported formats include, scatterplots (compare Figure 3 for the default plot produced), histograms, bar charts, boxplots. From the default setting, a wide range of customisation options within the scope of corporate design is possible.

[1] This is also the case for subtitles, where a subtitle defined using {bbkplot} will have precedence.

Boxplots are worth a short mention in this context, as they pose a challenge in our context. Generally, boxplots are an option to represent the conditional or unconditional distribution of the data. They have been developed by Tukey (1977) as a simple descriptive representation of the data that even can be drawn by hand. Traditionally, a boxplot comprises the median as a line, a box stretching from the lower to the upper quartile, two whiskers that depict extremes and single outliers outside of the extremes.

Using the bbkplot() function from {bbkplot} to create traditionally less common graphics in corporate design easily within R

Figure 7



```
# (1) get data
world <- rnaturalearth::ne_countries(scale = "medi
um", returnclass = "sf")

# (2) plot
maps <- ggplot(data = world) +
  geom_sf(aes(fill = pop_est)) +
  xlab("Longitude") + ylab("Latitude")

# (3) bbkplot
bbkplot(maps, discrete = FALSE, scale_y_continuous
 = FALSE)

# A suitable title, subtitle and source can be add
ed using bbkplot as shown earlier

bbkplot(maps, discrete = FALSE, title = "World ma
p",
          subtitle = paste0("(", length(unique(world
$name)), " countries)"), scale_y_continuous = FALS
E, clear_background = TRUE
)
```

Source: Own depiction.

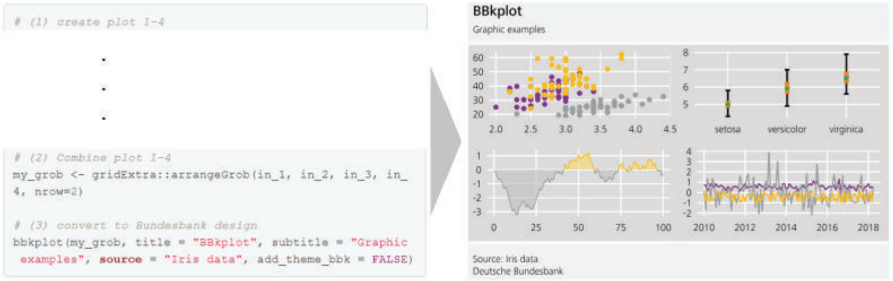
There exist numerous variations of this standard configuration, especially with respect to the whiskers (cf. Wickham and Stryjewski, 2011). In the Bundesbank design, the mean is included alongside the median in the boxplot. The mean is shown as a line and the median as a filled green circle. The outliers are usually not depicted. As requirements go beyond {ggplot2}, the necessary transformations of the boxplot are computed inside of {bbkplot}

Leveraging on the {ggplot2} package, a wide range of specific plots can be converted to corporate design, going beyond classical points, lines, and bars. For example, map graphic can be created relatively easily. Figure 7 shows an example of a heat map in corporate design, as a way to represent data pertaining to geographical locations. We use the Global Administrative Areas (GADM) maps and spatial data for our purpose. This data can be downloaded using the {GADMTools} package in R and is treated as “simple features” object as implemented by the {sf} package in R. The {ggplot2} package allows the use of “simple features” as graphical layers.

As a peculiarity of corporate design usage in our use-case, often multiple plots are combined in one graphic. {bbkplot} therefore allows combining multiple plots into one unified corporate design graphic. The procedure for this is to create the single (usually of `ggplot`) objects, combine them using `gridExtra::arrangeGrob()` and pass the resulting `grob` object to `bbkplot()` (compare Figure 8). While this precise capability may not be needed in most institutions, this is to show that relying on primitive graphic functions in R, virtually any requirement can be represented.

An example, how particularities of specific corporate designs can be implemented in R, in our use-case the propensity to combine multiple figures into one plot

Figure 8



```
# (1) create plot 1-4

        .

        .

        .

# (2) Combine plot 1-4
my_grob <- gridExtra::arrangeGrob(in_1, in_2, in_3, in_
4, nrow=2)

# (3) convert to Bundesbank design
bbkplot(my_grob, title = "BBkplot", subtitle = "Graphic
 examples", source = "Iris data", add_theme_bbk = FALSE)
```

Source: Own depiction.

As plotting capabilities according to precise corporate design requirements are supported for virtual any conceivable visualisation type for standard graphics, this offers potential for application in many producers of official statistics.

4. KEY TAKE-AWAYS AND APPLICABILITY FOR OFFICIAL STATISTICS

As a result of the technical capabilities of {bbkplot} outlined above, we note that {bbkplot} is an R package allowing efficient branding by empowering all staff to create high-quality graphics quickly and to their needs. While the package provides value on its own, {bbkplot} can be most valuable when integrated in a suitable environment, where multiple components can support each other.

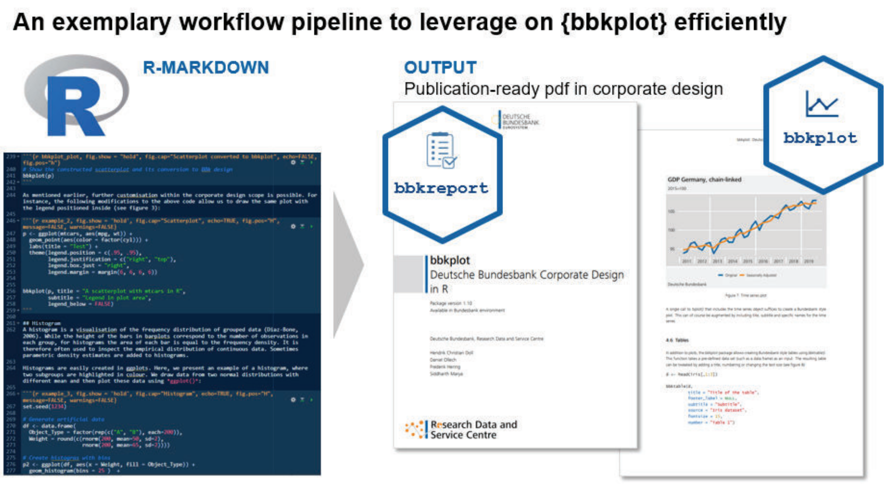
For maximum impact, we argue a tool such as {bbkplot} benefits from an ecosystem allowing data production, analysis and dissemination in the same environment; be it in R or any other tool. [1] While data production and analysis often take place in the same technical environment, to date, data dissemination is often thought of as a separate building block.

In our case study, the implementation of {bbkplot} is supported by the prior existence of an integrated pipeline, where {bbkplot} fits right in. This is the case if a skilled staff exists, that is empowered to use technical capabilities provided to them. From a technical side, a related tool exists in our case study, allowing staff to create documents for publication adhering to corporate design standards using Quarto.

This {bbkreport} package (cf. Gomolka, 2018) creates synergies, by providing a further building block to an integrated pipeline from data production, analysis to dissemination in the same environment. Using {bbkreport} and {bbkplot} or any other similar tools generally allows creating publication-ready documents from within R (cf. Figure 9).

An exemplary workflow pipeline, how {bbkplot} can be leveraged on to produce publication-ready pdf documents using Quarto

Figure 9



An exemplary workflow pipeline to leverage on {bbkplot} efficiently
R-MARKDOWN
OUTPUT
Publication-ready pdf in corporate design
bbkreport
bbkplot
DEUTSCHE BUNDESBANK
bbkplot
Deutsche Bundesbank Corporate Design in R
Research Data and Service Centre
GDP Germany, chain-linked
Original
Seasonally Adjusted

~~~
```{r bbkplot_plot, fig.show = "hold", fig.cap="Scatterplot converted to bbkplot", echo=FALSE, fig.pos="h"}
# Show the constructed scatterplot and its conversion to BBk design
bbkplot(p)
~~~

As mentioned earlier, further customisation within the corporate design scope is possible. For instance, the following modifications to the above code allow us to draw the same plot with the legend positioned inside (see figure 3):

```{r example_2, fig.show = 'hold', fig.cap="Scatterplot", echo=TRUE, fig.pos="H", message=FALSE, warnings=FALSE}
p <- ggplot(mtcars, aes(mpg, wt)) +
  geom_point(aes(color = factor(cyl))) +
  labs(title = "Test") +
  theme(legend.position = c(.95, .95),
        legend.justification = c("right", "top"),
        legend.box.just = "right",
        legend.margin = margin(6, 6, 6, 6))


bbkplot(p, title = "A scatterplot with mtcars in R",
        subtitle = "Legend in plot area",
        legend_below = FALSE)
```

Histogram
A histogram is a visualisation of the frequency distribution of grouped data (Diaz-Bone, 2006). While the height of the bars in barplots correspond to the number of observations in each group, for histograms the area of each bar is equal to the frequency density. It is therefore often used to inspect the empirical distribution of continuous data. Sometimes parametric density estimates are added to histograms.

Histograms are easily created in ggplots. Here, we present an example of a histogram, where two subgroups are highlighted in colour. We draw data from two normal distributions with different mean and then plot these data using *ggplot()*:

```{r example_3, fig.show = 'hold', fig.cap="Histogram", echo=TRUE, fig.pos="H", message=FALSE, warnings=FALSE}
set.seed(1234)

# Generate artificial data
df <- data.frame(
  Object_Type = factor(rep(c("A", "B"), each=200)),
  Weight = round(c(rnorm(200, mean=50, sd=2),
                   rnorm(200, mean=65, sd=2))))

# Create histogram with bins
p2 <- ggplot(df, aes(x = Weight, fill = Object_Type)) +
  geom_histogram(bins = 25 )  +
```

Source: Own depiction based on package by Gomolka (2018) and screenshot from Doll et al. (2020). R logo © 2016 The R Foundation.

[1] such as SAS traditionally or Jupyter Notebooks conceivably

While in our case study, the working environment for introducing the {bbkplot} package includes data-savvy and technologically adept staff, facilitating the introduction of the package in the organisation remains key in our view. For this purpose, we (i) create a detailed vignette with many usage examples and (ii) offer an introductory 2-day course for staff on usage.

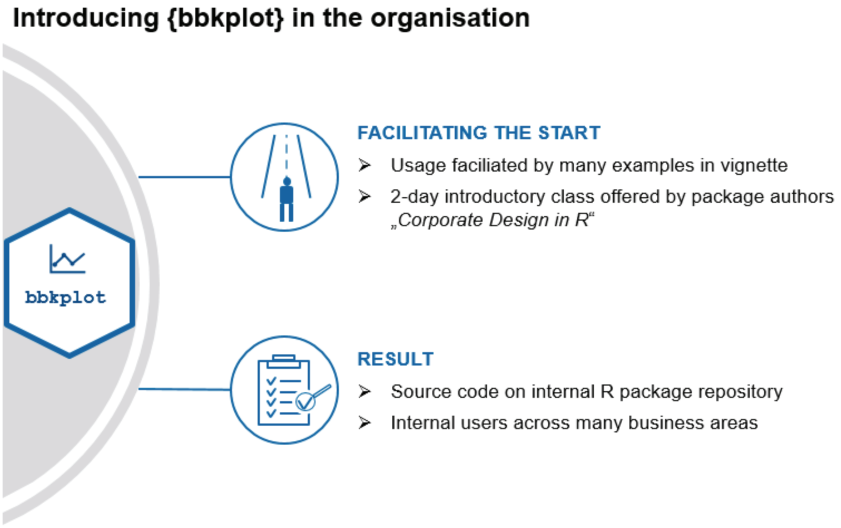
In order to provide native support directly with the package, we provide a detailed vignette. A vignette in the context of R packages is a guide explaining the usage of the package. The vignette is complementary to the R help, where users can get explanations on functions and parameters. The vignette provides exemplary code on how to create a wide range of graphics Thus, users get a head-start for the creation of their own graphics. This provides a type of self-learning guide to use {bbkplot}. However, this necessitates a knowledge basis of how to work with R packages, where to find them and how to find package-internal help.

For a target audience beyond those able to upskill themselves by exploring the package autonomously, we offer a 2-day internal class "*Corporate Design in R*". Contents are the basis of the corporate design, how to apply this efficiently and how to build pipelines to create graphics and documents according to high quality design standards. As a result, the {bbkplot} package is now available on the internal R package directory and users from many business areas leverage on {bbkplot} (see Figure 10).

As official statistics face the challenge to build a strategic approach to protect and promote the organisation's reputation and brand (UNECE, 2019), we aspire for our case study to provide lessons learned in order to advance these goals. Enabling reproducible pipelines in integrated environments evidently supports reliability of the statistics production process. Tools, such as {bbkplot} are building blocks to support such pipelines.

Introducing the package in the organisation

Figure 10



Introducing {bbkplot} in the organisation

FACILITATING THE START

- Usage faciliated by many examples in vignette
- 2-day introductory class offered by package authors *„Corporate Design in R“*

RESULT

- Source code on internal R package repository
- Internal users across many business areas

Source: Own depiction.

In the process of developing {bbkplot}, it becomes apparent, that most likely any complex design guidelines can be implemented in R to streamline corporate design graphic production. Upon implementation, efficiency gains for branding appear by enabling users can efficiently and reliably produce high-quality output. Quality gains for branding appear by a more widespread usage of precise corporate design, where beforehand coordination can be prohibitive in terms of effort for a large number of simple graphics.

Providers of official statistics can benefit from integrating the data dissemination process into the data production and analysis pipeline through tools such as {bbkplot} in several key ways. First, this enables flexibility. By building on the powerful and ever increasing visualisation capabilities that R and state-of-the-art packages such as {ggplot2} bring, novel visualization types can be accommodated quickly, even including making them interactive. Additionally, as R is a free tool and some barriers to the creation of graphics get removed, this may reasonably reduce the cost of data visualisations.

Of course, using R is by no means the only solution. Conceivably, using Python and Jupyter Notebooks can work just as well and institutions with a SAS legacy may build on this. The choice of the tool depends on the existing infrastructure, skillsets of staff and corporate design specifications. Providers of official statistics nowadays face a choice of high-quality software environments allowing integrated pipelines from data production all the way to dissemination of insights.

5. CONCLUSION

The past decade has witnessed a growth in the amount of available data and its use for analysis. As the digitalisation trend continues, there will be more data-driven analysis, decision- making, and dissemination of results. Along with this comes an increasing number of data-savvy employees carrying out data analysis and using programming languages such as R. Amidst the wealth of information produced, disseminated, and consumed, there is a large demand for reliable information. Official statistics can be more valuable than ever by providing high-quality information that serves as the factual basis for societal conversations.

Efficient communication of the reliability via corporate design lends the authenticity and trustworthiness of providers of official statistics to their communication with the public. To add to this, the spread of information on digital platforms has surpassed the traditional platforms and this means reduced time from production to publication. This adds a motivation to produce and disseminate reliable information from administrative sources in order for such information to be impactful.

We develop {bbkplot} to meet this contingency keeping in mind the tech- and data-savvy employees who can use it at scale to produce graphics for their presentations and communication. {bbkplot} further enables staff to produce a wide range of graphics including interactive, network and map graphics for digital presentations, an accomplishment that was not possible in the past.

In a modern, collaborative paradigm, it makes sense to develop a package like {bbkplot} on an open source platform. This enables a large number of analysts to produce graphics that complement their analysis and are fit to be included in reports and presentations on the go. The benefit is access to means of producing quality graphs directly by the responsible expert carrying out the analysis at an extra minimal cost to the organisation. As it makes sense to stay in the same computational ecosystem as used in data production and analysis, R is the natural tool of choice in our case study (in addition to its very powerful and adaptable graphics capabilities).

For official statistical offices, implementing comprehensive pipelines in one environment bring quality and efficiency gains by enabling easier reproducibility and updating of documents. The potential for typos or errors through incomplete copy pasting between tools is reduced. Furthermore, the speed from data production to publication can be increased because of existing recurring publications, the pipeline is only set up once and any updates generally require few changes. We consider {bbkplot} as one building block of a unified pipeline, facilitating data dissemination. By helping to consequently including corporate design in all communication, we support brand communication and spread of messages from a trusted institution.

References

1. Allen, P. (1996). User attitudes toward corporate style guides: a survey. Technical Communication, 43 (3), 237-243.
2. Akers, W. (2015). Visual resource monitoring for complex multi-project environments. Int. J. System of Systems Engineering, 6 (1/2), 112–126.
3. Arnold, J.B, Daroczi, G., Werth, B., Weitzner, B., Kunst, J., Auguie, B. Rudis, B., Wickham, H., Talbot, J., and J. London (2020). ggthemes: extra themes, scales and geoms for ggplot2. Retrieved on 01-01-2024 from https://CRAN.R-project.org/package=ggthemes.
4. Bion, R., Chang, R., and J. Goodman (2017). How R helps Airbnb make the most of its data. The American Statistician, 72 (1), 46-52.
5. Deutsche Bundesbank (2011). Corporate Design Manual. October 2011.
6. Deutsche Bundesbank (2023a). Monthly report, 75 (10). Retrieved on 01-01-2024 from https://www.bundesbank.de/resource/blob/913854/f463fc6fed22e587630a17cd11e469f1/mL/2023-10-monatsbericht-data.pdf.
7. Deutsche Bundesbank (2023b). Das Eurosystem und die Deutsche Bundesbank. Presentation slides. Retrieved 01-01-2024 from https://www.bundesbank.de/de/service/schule-und-bildung/unterrichtsmaterialien/sekundarstufe-ii/praesentation-fuer-den-unterricht-849308.
8. Doll, H. C., Ollech, D., Hering, F. & S. Marya (2020). bbkplot. Deutsche Bundesbank Corporate Design in R, Technical Report 2020-03, Deutsche Bundesbank, Frankfurt a.M. Retrieved on 01-01-2024 from https://www.bundesbank.de/resource/blob/831408/20861a1d419d93a1b2eba25ee829eae1/mL/2020-03-bbkplot-data.pdf.
9. Diaz-Bone, R. (2018). Statistik für Soziologen. Konstanz. UTBVerlag. Global Administrative Areas (2012). GADM database of Global Administrative Areas, version 2.0.
10. Forbes, S., Ralphs, M., Goodyear, R., and Pihama, N. (2011). Visualising official statistics. Statistics New Zealand Working Paper, (11-02).
11. Gomolka, M. (2018). DataReportR. uRos 2018, Conference proceedings. Retrieved on 01-01-2024 from http://r-project.ro/conference2018/presentations/Mattias_Gomolka_DataReportR.pdf.
12. Jackson, S. (2018). Creating corporate colour palettes for ggplot2. Retrieved on 01-01-2024 from https://drsimonj.svbtle.com/creating-corporate-colour-palettes-for-ggplot2.
13. Klein M. and S. Wagner (2018). ggCorpIdent: INWTlab wrapper for ggplot2 to create plots matching a corporate identity. INWTStatistics GmbH. Retrieved on 01-01-2024 from https://github.com/INWTlab/ggCorpIdent.
14. Neth, H. and N. Gradwdohl (2023). unikn: Graphical elements of the University of Konstanz's corporate design. Retrieved on 01-01-2024 from https://CRAN.R-project.org/package=unikn.
15. Destatis (2020). The Federal Statistical Office's 2020 Communication Strategy. Retrieved on 01-01-2024 from https://www.destatis.de/EN/About-Us/Goals-Strategy/communication-strategy-download.pdf?__blob=publicationFile.

16. Stylianou, N., Guibourg, C., Dahlgreen, W., Cuffe, R., Calver, T., and R. Mpini (2018). bbplot: Making ggplot graphics in BBC news style. British Broadcasting Company.
17. Ten Bosch, O., and De Jonge, E. (2008). Visualising official statistics. Statistical Journal of the IAOS, 25.3 (4), 103-116.
18. Tuckey, J.W. (1977). Exploratory data analysis. Addison-Wesley.
19. UNECE (2019), Strategic communications framework for statistical institutions. Note by the high-level group for the modernisation of official statistics project team on strategic communications framework. Retrieved 01-01-2024 from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2019/7_Strategic_commmunication_framework_for_consultation.pdf.
20. Van den Bosch, A. L., De Jong, M. D., and Elving, W. J. (2005). How corporate visual identity supports reputation. Corporate Communications: An international journal, 10 (2), 108-116.
21. Wickham, H, Chang, W., Henry, L., Takahashi, K., Wilke, C., Woo, K., and H. Yutani (2019). gg- plot2: Create elegant data visualisations using the grammar of graphics. Retrieved 01-01-2024 from https://CRAN.R-project.org/package=ggplot2
22. Wickham, H, and L. Stryjewski (2011). 40 years of boxplots. American Statistician.

Appendix

Appendix 1: {bbkplot} functions

bbkplot

The function `bbkplot()` is designed to convert `ggplot` objects into Bundesbank style graphics.[1] The function primarily uses the {ggplot2}, {grid} and {gridExtra} packages and arranges the three distinct areas of a graph as mentioned above namely, the header, the plotting area and the footer.

Adding the header and footer to the plotting area is achieved using the {gridExtra} package, the class of the resulting object changes from {ggplot2} to {grid}. Consequently, after applying the `bbkplot()` function, it is not possible to change the design of the plot using {ggplot2} functions. Put differently, all changes to the non-bbk specific layout and plot elements should be added before calling the `bbkplot()` function. Still, some additional changes can also be achieved using the `bbkplot()` function itself.

There are several arguments to the function call and these can be used to personalise the graphical object, e.g. by adding title and subtitle to the plot. Table 1 (see Appendix 2) summarises the arguments of `bbkplot()`.

theme_bbk

In {ggplot2} and other packages, there exist a wide array of `theme_x()` type functions that define the general layout of a plot. The above-mentioned {ggthemes} package incorporates various corporate styles via specifying themes. Similarly, the {bbkplot} package includes the `theme_bbk()` function, which can be used to change the style of the plotting area in accordance with Bundesbank's corporate design. The aforementioned `bbkplot()` function is the workhorse for adapting the plotting area.

In contrast to calling `bbkplot()` directly, using `theme_bbk()` enables extra calls to `ggplot2::theme()` to further personalise the plot. To add the Bundesbank corporate design specific footer and header, it is then necessary to call `bbkplot()` afterwards with the parameter setting `add_theme_bbk = FALSE`. This setting suppresses another call to `theme_bbk()`, which would potentially replace the plot adjustments made.

This is useful for instance, when one may need very specific settings, which usually would be overriden by `theme_bbk()`. Such examples could potentially include particular customised grid lines, axis labels or legend

[1] The function also adds Bundesbank style header and footer to objects of class grid. These are often multiple plots generated by {ggplot2}. These grid objects need to be in the Bundesbank style for the whole plot generated by {bbkplot} to be in the style. This can usually be achieved by adding theme_bbk() to parts of the grid object.

positioning. Also, it is often preferable to use `theme_bbk()` when working with other packages extending the possibilities of {ggplot2}.[1]

tsplot

Due to the relevance of time series to economists, this function is dedicated to displaying and manipulating time series. For visualising the data the `tsplot()` function relies heavily on the `bbkplot()` function. The additional benefit is twofold. First, because of the peculiarities of the time axis, reformatting time series is often quite cumbersome. Therefore, `tsplot()` accepts time series as input instead of a `ggplot` object as in `bbkplot()`. Second, `tsplot()` adds several functionalities for transforming the input and the output.

By default, the time series is depicted as a line plot and the function uses sensible decision rules to identify a meaningful date format for the x-axis. Using the parameters transform and type, it is easy to produce a figure of the period-to-period rate of change as a bar plot instead. As date axis ticks and labels follow very detailed style guidelines, special care is taken to ensure easy and user-friendly implementation of these.

bbktable

The function `bbktable()` can be used to create tables in the Bundesbank design. It accepts data.frames and matrices as input. Though the output is a table, the format is a plot. This has the disadvantage that the table cannot be changed after its creation. However such plots can be flexibly included in any publications and documents, i.e. LaTex, Excel and PowerPoint.

bbkcolor

In general, there are 13 specific colours defined in the Corporate Design Manual.[2] The `bbkcolor()` function defines the colours using the hexadecimal system that is used by {ggplot2} and {bbkplot}. The user can chose colours and, if needed, the shade. Each colour has up to five shades. If less than five shades are defined in the Corporate Design Manual, the shades are repeated, e.g. in `bbkcolor()` the fourth and fifth shade of green are

[1] Examples include the packages {ggiraph} and {plotly} that allow interactive plotting. Up-to-date examples of creating interactive graphics can be found in the vignette of the {bbkplot} package.

[2] Colours are Bundesbank-blue, Bundesbank-grey as so-called brand colours, and dark grey, yellow, orange, brown, red, pink, violet, light blue, petrol, dark green and light green as design colours for visualisations. The Corporate Design Manual specifies approved shades. For the shades, corporate design defines the colours as RGB and CMYK colours. We translate these into hexadecimals also known as HTML colour codes.

identical. The function allows the specification of multiple colours and shades and understands American and British spelling of the colours, e.g. grey and gray.

save_plot

To export the plot, the `save_plot()` function can be used. The recommended output format are PNG and PDF (`device = “png”` or `device = “pdf”`). By default, the plot is exported as a PNG image to the current working directory. A PDF can be created by setting `device = “pdf”`. Internally, the {grDevices} function `cairo_pdf()` is used.

For all formats but PDF, `ggsave()` of the {ggplot2} package is used. Currently, it allows exporting plots as EPS, PS, TEX, JPEG, TIFF, BMP, SVG and WMF. The function `save_plot()` allows to specify a path argument, to save the plot in any specified folder and a file name argument.

Appendix 2: Arguments for the bbkplot() function

	Argument	Definition
1	Object	A ggplot object to be put into bbk design. Must be specified
2	title	The title of the plot, a string. Default=NULL
3	main	Alternative argument name to pass the title of the plot, a string. Default=NULL. When both title and main are not NULL, title will be used
4	subtitle	The subtitle of the plot, a string. Default=NULL
5	submain	Subtitle of the plot, a string. Default=NULL
6	number	An optional number for the plot (as a string), in case multiple plots are produced and should be numbered (in the upper right corner). Default=NULL
7	footer_label	Label of the plot, a string, e.g. useful for adding datetime or the name of the file itself to the plot. Default=NULL
8	color	The colour of the plot, a string (one of c(”blue”, ”grey”, ”darkgrey”, ”yellow”, ”orange”, ”brown”, ”red”, ”violet”, ”petrol”, ”green”, ”darkgreen”, ”lightgreen”, ”black”), i.e. the bbk corporate style guide colours). Default=NULL
9	fill	The colour to fill areas
10	source	The source of the plot, a string
11	include_source	Whether the text component ”Source:” should be included
12	legend_below	Boolean, whether the legend should be extracted and plotted below the plot area itself. If FALSE, the user has to adjust legend position within the plotting area. Both are valid according to the Bundesbank style guide. Default=TRUE
13	add_theme_bbk	Whether to add the bbk theme to the plotting surface. Default=TRUE. Useful to be set to FALSE in non-standard plots, if different theme settings were manually set before, this option may override them)
14	fontsize	The size of the font to be used for the title. Everything else is scaled to 0.7*font size according to the style guide. Default=18
15	discrete boolean	Whether colour values should be used for plotting be discrete (or inter- polated). Default=TRUE
16	palette	The colour palette from the range of Bundesbank design colours (see details). Default=”main”. One of c(”main”, ”main_raster_1”, ”main_raster_2”, ”main_raster_3”, ”main_raster_4”, mixed”, ”mixed_raster_1”, ”mixed_raster_2”, hot”, ”cold”, ”bichromatic”, ”blackwhite”, ”random”)
17	axis_title	Whether to draw the axis titles. Default=TRUE
18	legend_title	Whether to draw the legend titles. Default=TRUE
19	distance_subtitle	The size of the space between title and subtitle. Default=1. May be useful to increase in case of multiline titles
20	plain_background	Whether the plot has a plain background without lines

21	y_title_direction	Defines the position of the y-axis title. Default=”vertical”. Other options are ”top”, to align the title on the top of the plotting area (horizontal) and ”subtitle” to align the title into the header
22	log_y	What kind of logarithmic scaling the y axis should have. log_y_scale takes the value of the base of the logarithm. It will usually be a value of c(1/2, 2, 10,exp(1)) but can take other numeric values. default=NULL (no logarithmic scale)
23	n_breaks	Number of grid labels, numeric value, default=5
24	scale_y_position	Position of the y axis scale, default = ”left”, can be set to ”right” e.g. to arrange multiple plots
25	horizontal_line	Whether to add a highlighted horizontal line, can be any numeric (usually at zero), default=0
26	expand_x_axis	Whether the x -axis should be expanded (the largest value by 10 percent) in order to open the grid, default=FALSE
27	legend_height	Height of the legend (if plotted below the plot area), default = 1, increase legend_height to increase vertical space for legend
28	scale_y_continuous	Whether the y_scale is continuous (TRUE) or discrete (FALSE), default = TRUE
29	secondary_axis	Creates a secondary axis based on a one to one transformation of the primary axis
30	secondary_axis_name	Name of the secondary axis
31	...	Additional arguments passed to ggplot2::theme()

Difference on Evaluation Scores Considering Image Descriptions for Autocoding

Yukako Toko (ytoko@nstac.go.jp)
National Statistics Center, Japan

Mika Sato-Ilic (mika@risk.tsukuba.ac.jp)
Institute of Systems and Information Engineering, University of Tsukuba, Japan

ABSTRACT

Autocoding plays an essential role in editing official statistics data, and here we have proposed a classification method which is fundamental to a hybrid autocoding system. This system has the essential feature of combining a rule-based classification method and a machine learning based classification method, in order to lead the coding task of the Family Income and Expenditure Survey. It is known that shopping receipt image data causes difficulty when using only the rule-based part of the proposed system, due to the complexity of the given data. Therefore, including a variety of criteria in the evaluation of the classification results for the shopping receipt images is essential for obtaining correct data. For this reason, this paper presents surveyed results of the various criteria for the shopping receipt image data based on the hybrid autocoding system. As a result, we found that the machine learning based classification element of the autocoding system chiefly works for dealing with the shopping receipt image data. Additionally, due to the recent increase in quantity of shopping receipt image data, the importance of the machine learning based classification method grows. Moreover, based on the results of various criteria, the existence of a dynamical sensitivity over the times was found. This may direct us to future developments in the machine learning-based classification method, such as fuzzy clustering based support vector machine currently in development (Toko and Sato-Ilic, 2021, Toko and Sato-Ilic, 2022).

Keywords: *Coding, Evaluation metrics, Fuzzy logic, Text classification*
JEL Classification: *C38*

1. INTRODUCTION

Official statistics often require coding and translating text descriptions into standardized labels for data processing. Traditionally, this labor-intensive task was performed manually by many experts over a long period of time. To improve efficiency, automated coding has become a focus of recent research. For example, Hacking and Willenborg (2012) described a methodology for autocoding. Gweon et al. (2017) illustrated methods for automated occupation coding based on statistical learning. In addition, Benedikt et al. (2020) introduced machine learning classification of shopping receipts data for the house budget survey. For the coding task of the Family Income and Expenditure Survey in Japan, we have developed a hybrid autocoding system (Toko et al., 2018, Toko et al., 2019, Toko and Sato-Ilic, 2020, Toko et al., 2023) and as an extension for more complex data coding, we have developed advanced machine learning based autocoding methods (Toko and Sato-Ilic, 2021, Toko and Sato-Ilic, 2022). The hybrid autocoding system uses a combined method with a rule-based method and a method based on machine learning and it has been practically implemented since Jan. 2022. The rule-based method was developed based on expert's knowledge and it assigns labels using human-crafted rules. The method based on machine learning is a classification method which is the reliability score-based classification model. The reliability scores have been defined considering two kinds of uncertain measures. One is probability measure which works for directly measuring data frequency and the other is fuzzy measure which works for measuring uncertainty of classification status for each data over labels. For evaluating the performance of this hybrid autocoding system, we have investigated only classification accuracy and coverage in our previous studies. Therefore, we have only evaluated true positive and true negative scores. That is, we did not consider false scores. To address this issue, this study evaluates the performance of this system for precision, recall, and f1 score including performance of false scores. In addition, in the Family Income and Expenditure Survey, the amount of target data has been increasing month by month. This data can be roughly divided into data obtained from shopping receipts images and manually inputted data. Although the numbers of both kinds of data have been increasing, the number of shopping receipts data has especially increased sharply compared with the other. However, there are difficulties for coding data obtained from shopping receipts images because of problems specific to receipt data such as the variety of product names. Therefore, this study evaluates different measures for the hybrid autocoding system with respect to the trends of different forms of data collection.

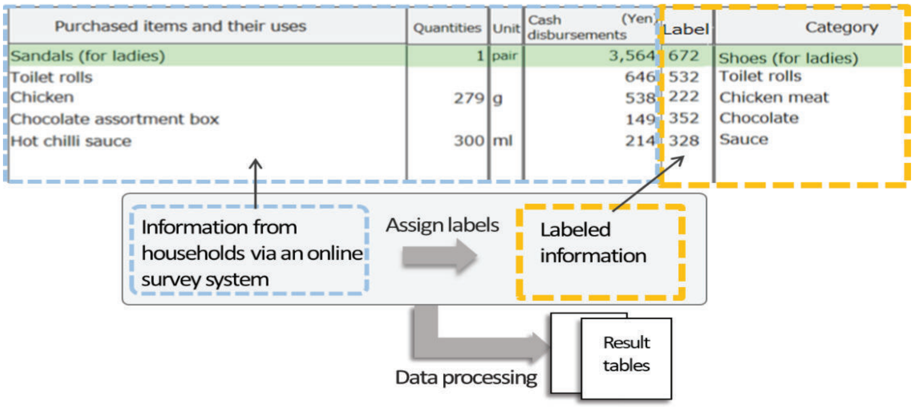
The rest of this paper is organized as follows: The hybrid autocoding system for the Family Income and Expenditure Survey is introduced in Section 2. The classification method based on machine learning is explained in section 3. The numerical examples are illustrated in section 4. Conclusions are described in section 5.

2. HYBRID AUTOCODING SYSTEM FOR THE FAMILY INCOME AND EXPENDITURE SURVEY

In the Family Income and Expenditure Survey, households are asked to keep their daily income and expenditures. Extract the following information from shopping receipts, whether provided as images or text descriptions: individual product names, item prices, and the total receipt amount. Using the collected information, each text description is assigned a corresponding 3-digit label. For example, if the target text description is “sandals for ladies”, then the corresponding category label is “672” shown in figure 1. Then the labeled information is used for data processing to create result tables.

Coding task for the Family Income and Expenditure Survey

Fig 1

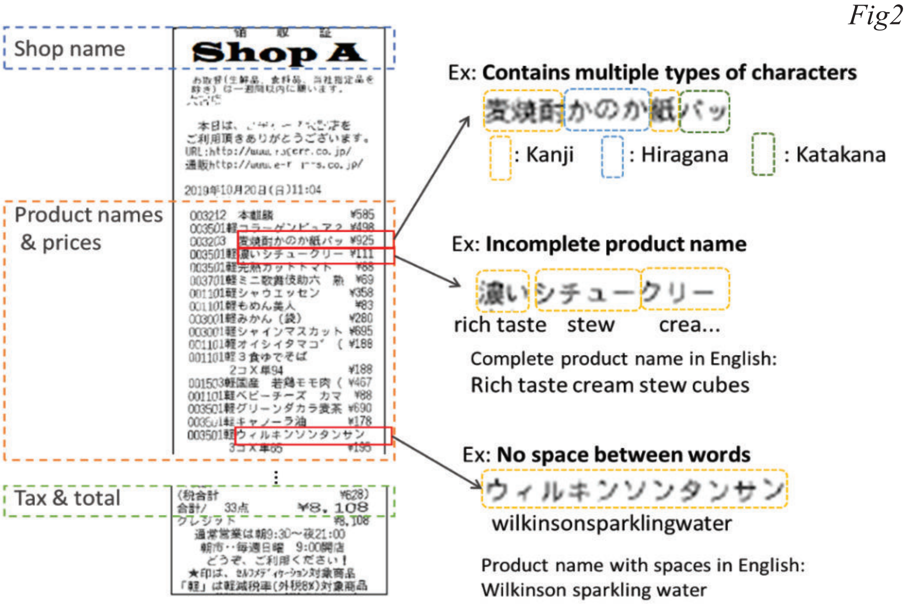


Purchased items and their uses	Quantities	Unit	Cash (Yen) disbursements	Label	Category
Sandals (for ladies)	1	pair	3,564	672	Shoes (for ladies)
Toilet rolls			646	532	Toilet rolls
Chicken	279	g	538	222	Chicken meat
Chocolate assortment box			149	352	Chocolate
Hot chilli sauce	300	ml	214	328	Sauce

As described above, the target data includes product names obtained from shopping receipt images. Figure 2 shows an example of a shopping receipt. For assigning labels to data obtained from images of Japanese shopping receipts, there are the following difficulties: First, product name sometimes contains multiple types of characters since we combine multiple types of characters for writing in Japanese. For example, product names contain 3 types of characters, “Kanji”, “Hiragana”, and “Katakana”. “Kanji” is the writing system originally adopted from China. “Hiragana” is phonetic-based lettering system developed in Japan along with “Katakana” which is also a phonetic-based Japanese alphabet used almost exclusively for writing foreign words. The second difficulty is the regularity of incomplete product names that periodically appear in Japanese shopping receipts. Here the example product name is “KOI STEW CREA” in Japanese. It means “rich taste stew crea…”, with “rich taste cream stew cubes” being the complete name in English. Experts assign a code for this type of product name complementing the missing word. Finally, in Japanese shopping receipts, there are no spaces between words. Therefore, it is difficult to split a product name into words. Again, in this example, the product name is “Wilkinson sparkling water”. It is easy to identify the product when we hear it, but it is much more difficult to identify it when we see the product name written like ”Wilkinsonsparklingwarter”.

Image of Japanese shopping receipt

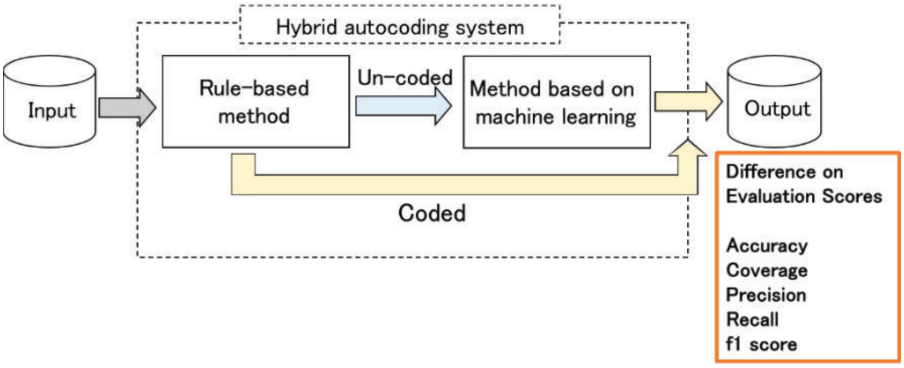
Fig2


Shop name
Shop A
Product names & prices
Tax & total
Ex: Contains multiple types of characters
麦焼酎かのか紙パッ
: Kanji
: Hiragana
: Katakana
Ex: Incomplete product name
濃いシチュークリー
rich taste
stew
crea...
Complete product name in English:
Rich taste cream stew cubes
Ex: No space between words
ウィルキンソンタンサン
wilkinsonsparklingwater
Product name with spaces in English:
Wilkinson sparkling water

We have developed an innovative hybrid autocoding system specifically designed for the Family Income and Expenditure Survey’s coding task (Toko et al., 2018, Toko et al., 2019, Toko and Sato-Ilic, 2020, Toko et al., 2023). The hybrid autocoding system uses a combined method with a rule-based classification method and a classification method based on machine learning. Figure 3 shows an image of the flowchart of the hybrid autocoding system. The hybrid autocoding system is executed as follows: First, the rule-based method is applied to the target dataset of text descriptions to classify clear text descriptions that are unambiguous and so clearly classifiable to specific labels. Then, the method based on machine learning is applied to the remaining text descriptions that the rule-based method could not assign any labels. That is the machine learning method is utilized to assign labels to more difficult data, such as text descriptions obtained from shopping receipt images, as compared with the straightforward data treated by the rule-based method. For evaluating the performance of this system, we have investigated only accuracy and coverage. Therefore, we only evaluated true-positive and true-negative scores. That is, we did not consider false scores. To overcome this issue, we evaluate the performance of this system for precision, recall, and f1 score, including performance of false scores. This paper shows several results related to these obtained evaluation scores.

Image of flowchart of the hybrid autocoding system

Fig 3



3. CLASSIFICATION METHOD BASED ON MACHINE LEARNING

In the method based on machine learning, first, symbols that are not related to the conceptual meaning of the text descriptions are excluded as a pre-processing. Then, objects (or words) are extracted. For the extraction of objects, word-level N-grams from the word sequences of text descriptions are taken after tokenizing each description using MeCab (Kudo et al., 2004), which is a morphological analyzer for Japanese text. Here, we consider unigrams, bigrams, and entire sentences as objects. After that, an object frequency table is created, that is all extracted objects are tabulated based on their given labels into an object frequency table. Then, the classification process in the machine learning method extracts objects in the same manner as described above and retrieves candidate labels from the object frequency table provided by using the extracted objects. Then, it calculates the relative frequency of j-th object to a label k defined as

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

where n_{jk} is frequency under a status that an object j is classified into a label k in the training dataset. J is the number of objects and K is the number of labels.

However, only using the relative frequency for this classification cannot obtain the satisfactory accuracy of the autocoding result. Therefore, we define the reliability score of j-th object to a label k as follows:

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

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

In (1) and (2), the status of classification structure for each object which are shown as $1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}$ or $\sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2$ are considered. Here, $\left\{\tilde{\tilde{p}}_{j1}, \cdots, \tilde{\tilde{p}}_{j\tilde{K}_j}\right\}, \tilde{K}_j \le K$ are putted such as $\tilde{\tilde{p}}_{j1} \ge \cdots \ge \tilde{\tilde{p}}_{j\tilde{K}_j} \ge \cdots \ge \tilde{\tilde{p}}_{jK}$, $j = 1, \cdots, J$, and $\tilde{K}_j$ shows the number of labels for an object j. These reliability scores were defined using two kinds of uncertainty measures. One is probability measure and the other is fuzzy measure (Bezdek, 1981, Bezdek et al., 1999). That is, $\tilde{\tilde{p}}_{jk}$ shows the uncertainty from the training dataset which is probability measure and $1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}$ or $\sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2$ shows the uncertainty from the latent classification structure in data which is fuzzy measure. These values of the uncertainty from the latent classification structure can show the classification status of each object; that is, how each object is classified to the candidate labels. *T* shows *T*-norm in statistical metric space (Menger, 1942, Mizumoto, 1989, Schweizer and Sklar, 2005). In machine learning, the robustness of the results obtained from a model is an important issue. Therefore, to address this issue, we generalized the reliability score by using the idea of *T*-norm in statistical metric space. *T*-norm is a measure based on uncertainty of dissimilarity defined in a function family which satisfies the following four conditions:

Boundary conditions: $0 \leq T(a,b) \leq 1,\ T(a,0) = T(0,b) = 0,\ T(a,1) = T(1,a) = a$

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

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

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

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

$$\text{Algebraic product: } T(a,b) = ab \tag{3}$$

$$\text{Hamacher product: } T(a,b) = \frac{ab}{p+(1-p)(a+b-ab)},\quad p \geq 0 \tag{4}$$

$$\text{Minimum: } T(a,b) = min\{a,b\} \tag{5}$$

$$\text{Einstein product: } T(a,b) = \frac{ab}{1+(1-a)(1-b)} \tag{6}$$

By considering a bias of infrequent words, we redefine the reliability score as follows:

$$\bar{\bar{p}}_{jk} = g(n_j) \times \bar{p}_{jk}, \tag{7}$$

where $g(\cdot)$ is sigmoid function.

4. NUMERICAL EXAMPLES

For the numerical examples, we applied the hybrid autocoding system to the Family Income and Expenditure Survey dataset. The Family Income and Expenditure Survey is a monthly sample survey conducted by the Statistics Bureau of Japan to understand the actual income and expenditure patterns of households. This dataset includes short text descriptions related to a household's daily incomes and expenditures such as receipt item names and purchased item names in Japanese including product names obtained from shopping receipt images and their corresponding labels. For labels in this survey, around 520 different labels are available. Figure 4 shows an increasing situation of the number of text descriptions in target dataset from Oct. 2019 to Jun. 2023. The blue line shows the number of target text descriptions obtained from shopping receipt images ("receipt data") whereas the red line shows the number of target text descriptions obtained from other than shopping receipt images, that is mainly manually inputted data ("not receipt data"). From figure 4, it is found that both types of data have been increasing month by month, but especially the number of receipt data has been increasing sharply compared with the number of non-receipt data. In this figure, the black dashed line shows the time (Jan. 2022) when the hybrid autocoding system was started to use. Before it, the rule-based system had been used.

Figure 5 shows the coverage of the hybrid autocoding system from Jan. 2022 to Jun. 2023. The blue line shows the coverage for receipt data whereas the red line shows the coverage for not receipt data. The coverage for receipt data has increased gradually whereas the coverage for not receipt data has stayed almost flat. However, the coverage for not receipt data is better than the coverage for receipt data so far as there is difficulty in coding receipts data because of problems specific to receipt data such as variety of product name described in section 2. In addition, figure 6 and figure 7 show the coverage of the rule-based method and the coverage of the machine learning method respectively. As mentioned in section 2, the machine learning method is applied to data which the rule-based method could not assign any labels, therefore the machine learning method has the role of treating difficult data for assigning correct labels. Therefore, the value of coverage of the machine learning method is lower than the value of coverage of the rule-based method. From figure 6 and figure 7, it can be seen that the coverage of the rule-based method has stayed almost flat for both receipt data and non-receipt data whereas the coverage of the machine learning method has increased especially for receipt data, so the method based on machine learning assigns labels for the increase of receipt data. In addition, the coverage for receipt data is lower than the coverage for non-receipt data in the rule-based method, while the coverage for receipt data is higher than the coverage for non-receipt data in the machine learning method. This indicates that the increase of the number of receipt data has been covered by the machine learning method.

For this numerical example, we used approximately 30 million text descriptions for training whereas approximately 0.99 million text descriptions per month for evaluation.

For evaluation measure for classification, we used the following criteria:

$$accuracy = \frac{TP + TN}{M}$$

$$macro\, precision = \frac{1}{K}\sum_{l=1}^{K}\frac{TP_l}{TP_l + FP_l} \tag{8}$$

$$macro\, recall = \frac{1}{K}\sum_{l=1}^{K}\frac{TP_l}{TP_l + FN_l} \tag{9}$$

$$macro\, f1\, score = \frac{1}{K}\sum_{l=1}^{K}\left(2 * \frac{precision_l * recall_l}{precision_l + recall_l}\right)$$

where K is the number of labels, M is the number of text descriptions, TP is number of true positive text descriptions, TN is the number of true negative text descriptions. TP_l is the number of true positive text descriptions at l-th label. $FP_l FP_l$ is the number of false positive text descriptions at l-th label, and FN_l FN_l is the number of false negative text descriptions at l-th label. "$precision_l$" is the value of macro precision shown in (8) at ll-th label and "$recall_l$" is the value of macro recall shown in (9) at ll-th label.

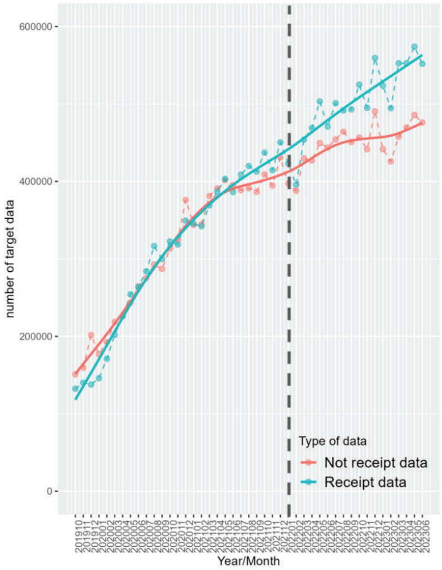
Figure 8 shows scores of evaluation measures and scores of coverage of the hybrid autocoding system. For this result, in the reliability score in the machine learning method shown in (7), we use equation (1) as $\bar{p}_{jk}$ where T-norm is the Einstein product shown in (6) and $g\left(n_j\right) = n_j/\sqrt{1 + n_j{}^2}$. That is, the reliability score shown in equation (7) is formulated as follows:

$$\bar{\bar{p}}_{jk} = \frac{n_j}{\sqrt{1 + n_j{}^2}}\left(\frac{\tilde{\tilde{p}}_{jk}(1 + \sum_{m=1}^{\tilde{K}_j}\tilde{\tilde{p}}_{jm}\log_K \tilde{\tilde{p}}_{jm})}{1 + \left(1 - \tilde{\tilde{p}}_{jk}\right)\left(-\sum_{m=1}^{\tilde{K}_j}\tilde{\tilde{p}}_{jm}\log_K \tilde{\tilde{p}}_{jm}\right)}\right). \tag{10}$$

In figure 8, dashed lines show the results of receipt data and solid lines show the results of not receipt data. From figure 8, it can be seen that values of accuracy are stable for both kinds of data over the targeted period. In the meanwhile, values of f1 score, precision, and recall are not stable, especially for significant difference between two kinds of data which are receipt and non-receipt data. This means even if the values of accuracy are almost consistent over the two kinds of data and the times, we can see variance of the change for other evaluation measures. The main difference between accuracy measure and other measures is the use of the false part of result. Therefore, by including consideration of the false part, we could see new tendency of data classification in the hybrid autocoding system. As previously mentioned, considerable difference exists between the two kinds of data for the coverage, however they are almost stable over the time period.

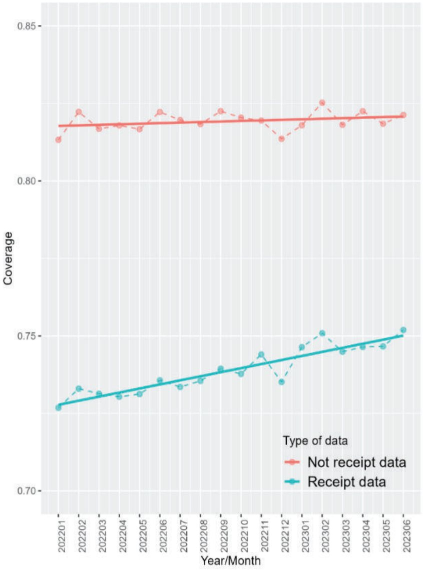
Number of text descriptions in target Dataset

Fig 4.



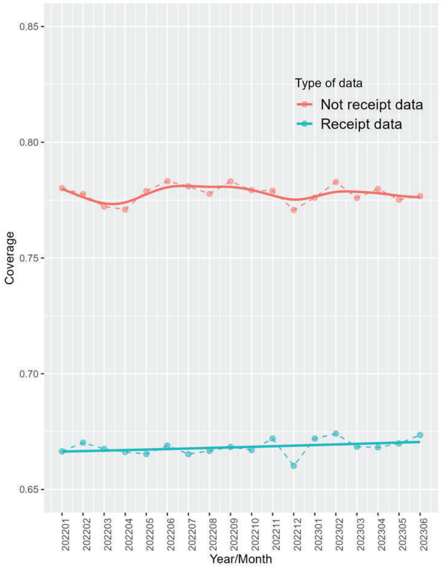
Comparison of coverage of the hybrid autocoding system for two types of data

Fig 5



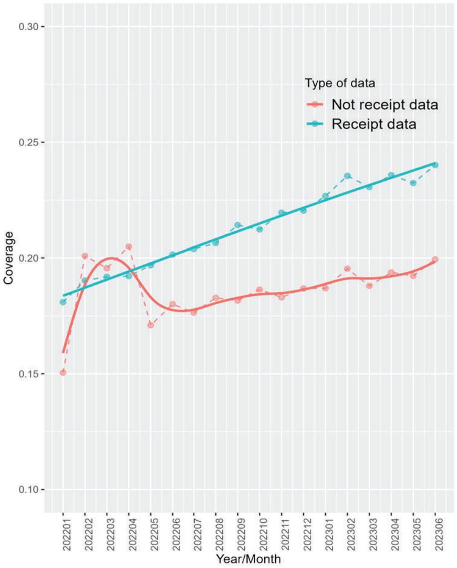
Coverage of the rule-based method

Fig 6



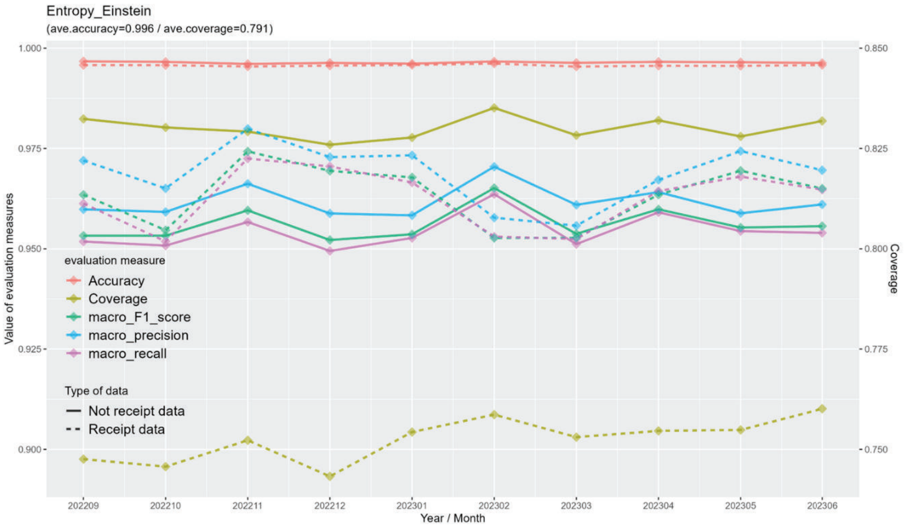
Coverage of the machine learning method

Fig 7



Scores of evaluation measures and scores of coverage of hybrid autocoding system by using Entropy based Einstein product shown in (10)

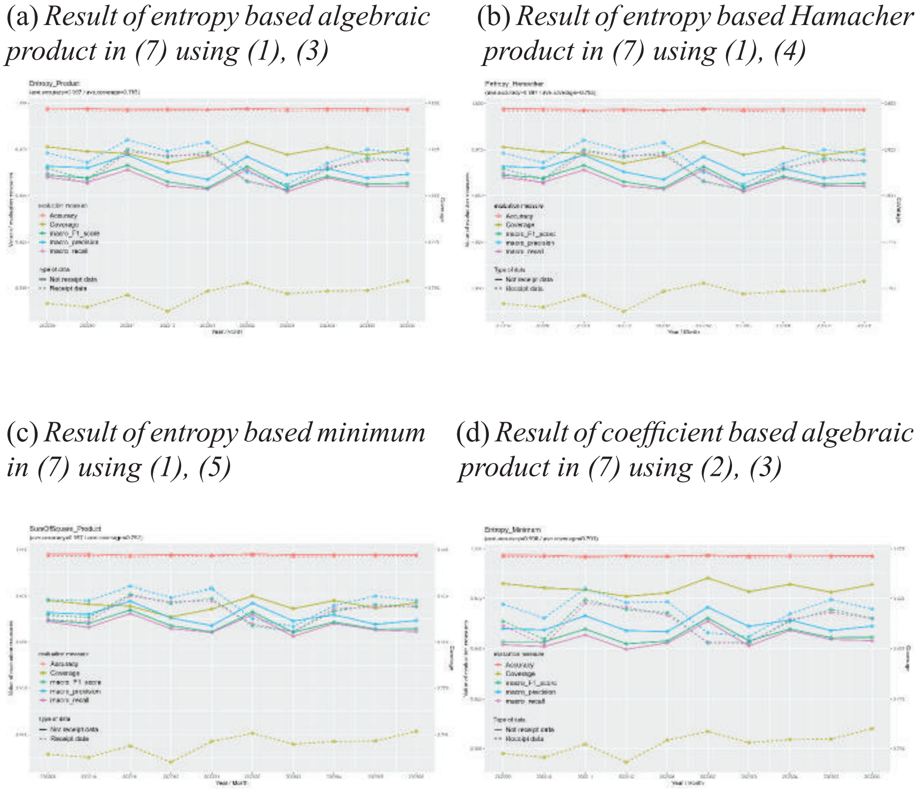
Fig 8



Although the best scores for average accuracy and average coverage were obtained by entropy based Einstein product, we obtained almost similar tendency for the scores of various evaluation measures by other aggregation operators shown in figure 9.

Scores of evaluation measures and scores of coverage of hybrid autocoding system by using other aggregation operators

Fig 9



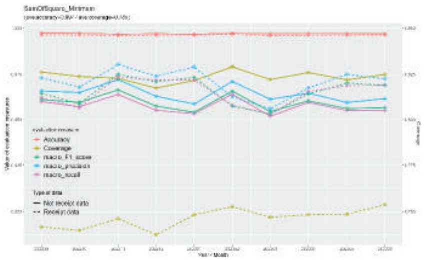
(a) *Result of entropy based algebraic product in (7) using (1), (3)*

(b) *Result of entropy based Hamacher product in (7) using (1), (4)*

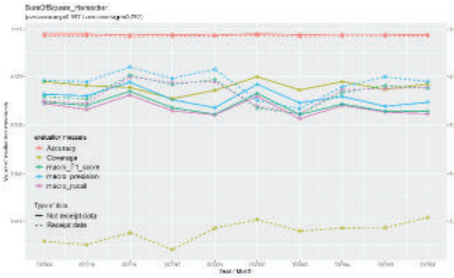
(c) *Result of entropy based minimum in (7) using (1), (5)*

(d) *Result of coefficient based algebraic product in (7) using (2), (3)*

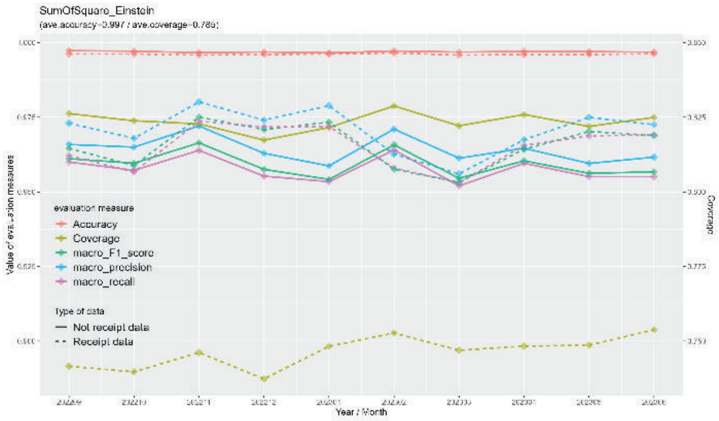
e) *Result of coefficient based Hamacher product in (7) using (2), (4)*



(f) *Result of coefficient based minimum in (7) using (2), (5)*



(g) *Result of coefficient based Einstein product in (7) using (2), (6)*



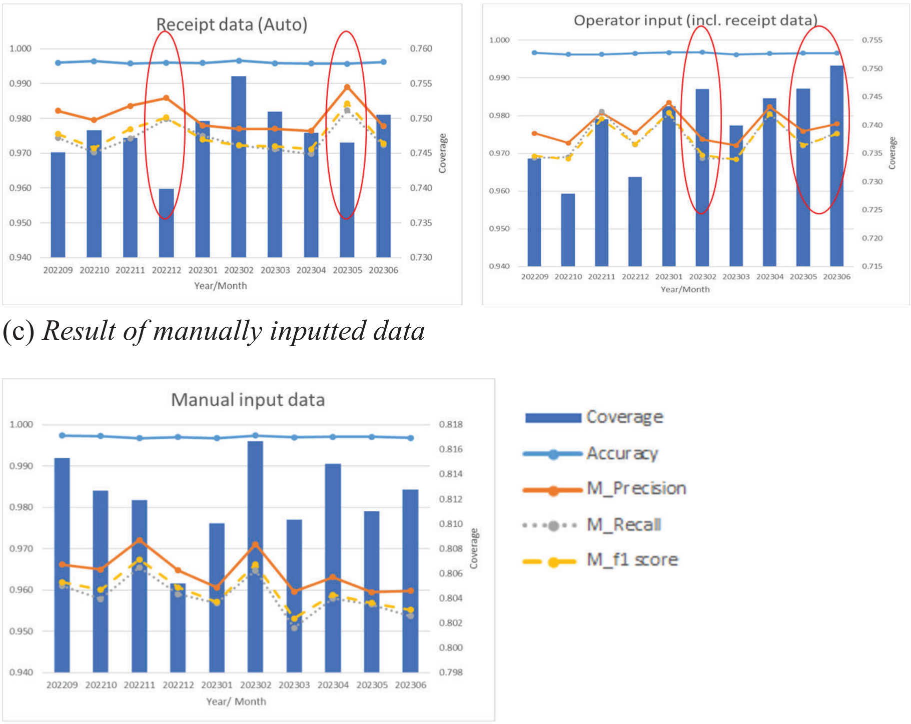
We also investigated the result of evaluation measures with respect to kinds of input data. Receipt data has two kinds of character recognition approaches. One is automatically recognized characters and the other is to recognize characters through operators. Figure 10(a) shows the result of automatically recognized receipt data, figure 10(b) shows the result of operator based receipt data, and figure 10(c) shows the result of manually inputted data. From these figures, we can see the same tendency over the scores of evaluation measures as mentioned above, which is accuracy are stable over the three kinds of data and time period, whereas other evaluation measures are not stable over the three kinds of data and the time period. However, we can see special features, in the case of automatically recognized receipt data. That is, even if coverage is lower but values of precision, recall, and f1 score, they have the special feature of inclusion of the false part of the evaluation and have higher increasing tendency. On the other hand, for another technique of receipt data which is operator based receipt data, we could see the opposite tendency that is even coverage is going to higher, however evaluation scores of precision, recall, and f1 scores are going in the decreasing direction. For the case of manually inputted data, tendency of increase and decrease change are almost simultaneously moving that is coverage scores and scores of other evaluation measures based on inclusion of the false part are moving a similar way. Moreover, even if coverage is higher, other evaluation scores including the false part are lower in the case of manually inputted data. In addition, for all evaluation measures, receipt data's evaluation scores are higher than evaluation scores of manually inputted data.

Result of evaluation measures with respect to kinds of input data

Fig 10

(a) *Result of automatically recognized receipt data*

(b) *Result of operator based receipt data*



(c) *Result of manually inputted data*

Results of the investigation were obtained using VB.Net and R. We implemented our hybrid autocoding system in VB.Net, as both rule-based and machine learning components were already developed within this language. Meanwhile, most of the other parts were implemented by R since R is convenient for data wrangling and it provides a rich set of packages for both data handling and data visualization such as “data.table” and “ggplot2”.

5. CONCLUSION

This paper evaluates different measures for the hybrid autocoding system under the assumptions of several kinds of data. In fact, we previously used only accuracy and coverage as evaluation measures for the hybrid autocoding system. In this study, we use various kinds of evaluation measures including the false part based evaluation measures such as precision, recall, and f1 score. From this, we capture various features based on different evaluation measures especially the false part based evaluation measures which have not been investigated in the previous study. In particular, we could see the machine learning method is covered for increase of receipt data. That is, we identified that the increase in amounts of data depends on the increase of receipt data, and increase of autocoding receipt data is treated by the machine learning method. Moreover, for all evaluation measures, evaluation scores of receipt data are higher than evaluation scores of manually inputted data. From these facts, we can see the stability of the machine learning method for receipt data.

As a further increase in the number of shopping receipt data is expected, method based on machine learning in the hybrid autocoding system will be expected to become more important. For further study, based on the investigation of the details of evaluation scores highlight differences of features among autocoding methods, therefore we believe this obtained knowledge can utilize further development of new methods or adoptable local application of methods for autocoding system for efficient and effective autocoding.

References

1. Benedikt, L., Joshi, C., De Wolf, N., Schouten, B. (2020), “Optical character recognition and machine learning classification of shopping receipts”, Available at: https://ec.europa.eu/eurostat/documents/54431/11489222/6+Receipt+scan+analysis.pdf (accessed Dec. 2023)
2. Bezdek, J.C. (1981), Pattern recognition with fuzzy objective function algorithms, Plenum Press.
3. Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R. (1999), Fuzzy models and algorithms for pattern recognition and image processing, Kluwer Academic Publishers.
4. Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., Steiner, S. (2017), “Three methods for occupation coding based on statistical learning”, Journal of Official Statistics, Vol. 33, No. 1, pp. 101-122.

5. Hacking, W., Willenborg, L. (2012), “Coding; interpreting short descriptions using a classification”, Statistics Methods, Statistics Netherlands, The Hague, Netherlands, Available at: https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/throughput/throughput/coding (accessed Dec. 2023).
6. Kudo, T., Yamamoto, K., Matsumoto, Y. (2004), “Applying conditional random fields to Japanese morphological analysis”, in the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25-26, Jul. 2004, pp. 230-237.
7. Menger, K. (1942), “Statistical metrics”, in Proceedings of the National Academy of Sciences of the United States of America, Vol. 28, pp. 535-537.
8. Mizumoto, M. (1989), “Pictorical representation of fuzzy connectives, Part I: Cases of T-norms, t-Conorms and Averaging Operators”, Fuzzy Sets and Systems, Vol. 31, pp. 217-242.
9. Schweizer, S., Sklar, A. (2005), Probabilistic metric spaces, Dover Publications.
10. Toko, Y., Iijima, S., Sato-Ilic, M. (2018), “Overlapping classification for autocoding system”, Journal of Romanian Statistical Review, Vol. 4, pp. 58-73.
11. Toko, Y., Iijima, S., Sato-Ilic, M. (2019), “Generalization for improvement of the reliability score for autocoding”, Journal of Romanian Statistical Review, Vol. 3, pp. 47-59.
12. Toko, Y., Sato-Ilic, M. (2020), “Improvement of the training dataset for supervised multiclass classification”, Czarnowski, I., Howlett, R.J., Jain, L. C. (Eds.), Intelligent Decision Technologies, Smart Innovation, Systems and Technologies, Springer, Singapore, Vol. 193, pp. 291-302.
13. Toko, Y., Sato-Ilic, M. (2021), “Efficient autocoding method in high dimensional space”, Journal of Romanian Statistical Review, Vol. 1, pp. 3-16.
14. Toko, Y., Sato-Ilic, M. (2022), “Autocoding based multi-class support vector machine by fuzzy c-means”, Journal of Romanian Statistical Review, Vol. 1, pp. 27-39.
15. Toko, Y., Sato-Ilic, M., Sasajima T. (2023), “Improvement of Model Construction based on Reliability Scores of Objects for Autocoding”, Journal of Romanian Statistical Review, Vol. 1, pp. 34-48.
16. Statistics Bureau of Japan: Outline of the Family Income and Expenditure Survey. Available at: https://www.stat.go.jp/english/data/kakei/1560.html, (accessed Dec. 2023)

Improving the quality of civil servants' professional training process through digitalization - an exploratory research from employees' perspective

Assoc. prof. Laura Mina-Raiu (laura.minaraiu@amp.ase.ro)
Bucharest University of Economic Studies, Associate professor

Lecturer Cătălin- Valentin Raiu (catalin.raiu@faa.unibuc.ro)
Bucharest University

Mihaela Comăniciu (mihaela comaniciumanuela21@stud.ase.ro**)**
Bucharest University of Economic Studies, MA student

ABSTRACT

Digitalization initiatives in the public sector require not only advanced ICT infrastructure and technologies, but also a good strategy in terms of stakeholders' management, especially citizens-clients and human resources. Because quality management plays a key role in improving the organizational performance and the quality of public services, this research is concerned with how quality management can enhance digitalization efforts and support public organizations adapt to the challenges of the technological revolution. The paper focuses on the particular case of National Institute of Administration (NIA) from Romania, the main provider of professional training courses and specialized training for public sector employees and also a very dinamic and proactive public organization regarding the quality management approach and digital transformation. Our findings suggest that implementing digital transformation is perceived by NIA employees to bring a lot of benefits, but at the same time participants to the digitalization process (public sector human resources and citizens) face several obstacles in this respect, such as lack of skills, competences and resources. We argue that a quality management approach may build a strong foundation for overcoming such obstacles, as TQM practices such as top management support, customer focus, continuous improvement, training and education enhance users' readiness for change.

Keywords*: digitalization, public sector organization, quality of training process, quality management*

INTRODUCTION

In the context of the COVID-19 pandemic, the low level of digitization has been strongly experienced by public sector institutions, being essential at that time for them to be able to continue their activities. Public institutions have been forced to adapt to working „online" and to seek solutions to provide citizens with quality services. However, there is a constant evolution of technology and modern working practices, which public institutions must take into account and improve their services offer, ultimately aiming to meet the needs of citizens. Education and training systems need to respond to the opportunities and challenges of today's digital transformation, as the COVID-19 pandemic has brought about unique changes in the educational landscape worldwide (Popescu et al. 2023).

This research explores the role of digitalization in public services and how it influences and enhances the quality of professional training process of public servants in an organizational setting where the quality management approach is part of the top management commitment and management. The paper is structured in four parts. The first one presents the concept of quality management and its particularities in public institutions, as well as the role of digitalization in enhacing the quality of public services in general and the quality of professional training for public servants in particular. The second part reveals the methology, aims, objectives, and hypotheses of the research, as well as the employed research methods. The case study on the role of digitalization in providing quality professional training services to public servants at the National Institute of Administration (NIA) is to be found in third part, while the last section focuses on conclusions and limitations.

1. THEORETICAL ASPECTS REGARDING DIGITALIZATION IN THE FIELD OF PUBLIC SERVANTS' PROFESSIONAL TRAINING

1.1 The concept of quality and total quality management in the public sector

Over time, various approaches and different viewpoints have been expressed regarding quality and its measurement, leading to a rich body of literature. The European Organization for Quality Control (EOQC) and the American Society for Quality (ASQ) define quality as the set of characteristics of a service or product that refers to its ability to satisfy given needs (Morgan and Murgatroyd, 1994, p. 8). Juran (1993) refers to quality as the ability of a product to satisfy a consumer's needs or the measure in which it successfully meets its expectations, or what the customers are willing to pay in exchange of what they receive.

The first concerns related to quality assurance are to be found in the business environment in the 1960s, with the goal of guaranteeing the quality of products and services, in the form of measures used for formalizing procedures of technical quality control and training and motivation procedures for personnel (Ebrahimi and Sadeghi, 2013, pp. 5625-5643.). Later, in the 1980s, the concept of quality is theorized, launching new ideas that develop total quality control and its advantages in the business environment. Total quality represents the totality of principles and methods aimed at satisfying the customer at the lowest costs. This encompasses the whole organization and its activities and focuses on several aspects, such as (Dobrin, 2005):

- Orientation towards customer needs, which implies surpassing the classical view that only quality departments are responsible for quality assurance. Thus, other departments in the organization also have a greater or lesser effect on the outcome;

- Extending the concept of the customer starting from the premise that the organization is a system that includes both suppliers and customers. Consequently, applying quality also involves satisfying the needs of internal customers;

- Leadership in terms of prices referring to the fact that customers' needs and expectations will be better met if costs transferred to the customer are reduced. Moreover, cutting costs increases the real chances of competing in the market;

- Management based on prevention, which refers to doing activities right the first time. In this way, the application of control actions is reduced, and costs are minimized;

improving human resources, which considers the quality achieved by the entire organization, making a well-established human resources management essential, which emphasizes motivation for quality and participation;

- Continuous improvement of quality means that services should be done even better and always adapted to the needs and expectations of the customer.

There are some differences between the classical concept of quality and total quality. In this respect, the objective of quality is conformity with standards, while the objective of total quality is customer satisfaction. Also, specialists define quality, and the focus is on products and services, while in the case of total quality, the customers define it, and all activities of the organization are involved. The set of methods through which total quality is achieved represents total quality management (TQM) (Vinni, 2007). This concept involves all functions and processes in an organization, aiming at the continuous improvement of the quality of services and goods, respectively, customer satisfaction (Stringham, 2004). Initially, Total Quality Management (TQM) emerged in the business environment in the USA during the 1980s-1989, and later in the central public administration at the state and federal level, to be adopted after 2000 by numerous units of public administration (Drăgulescu, 2013, p. 6). The first expert to address issues related to quality management was Juran, who considers that quality management can be divided into 3 stages: quality planning, quality control, and quality improvement (Carlos et al., 2014).

Another approach defines TQM as a management philosophy that develops all management principles and practices from the belief that success results from continuous quality improvement (Talib et al., 2012, p. 264; Petersen, 1999). Through his perspective, quality must be achieved in all corresponding stages of manufacturing a product, and improving market coverage and increasing long-term competitiveness leads to quality improvement (Ilieș and Crișan, 2011, p. 34).

1.2. Measuring quality in public organizations

Unlike private organizations, where performance measurement is achieved through economic benefit gained and the number of customers, public institutions aim to satisfy citizens, society, or individuals who use social or public interest services (Mina-Raiu et al., 2021). As providers of public services, both local and central public authorities are interested in evaluating and improving services, while their purpose and operations are reviewed by politicians and consumers. Currently, reducing bureaucracy represents the consensual approach for the evolution and modernization of public administration (Bellodi et al., 2024), but nevertheless many other intruments may contribute to the improvent of public service (Haruța, and Radu, 2010, pp. 62-70).

By contrast with the private sector, where there are many private companies competing for customers, in the public sector there is a single government that provides public services to citizens. Also, since the success of a business depends on the quality of the services offered, private companies carefully select the staff they display in front of customers, which includes careful recruitment and internal training. In the case of public services, this is difficult to achieve because you can’t always be sure that all employees working with customers have the appropriate training, characteristics, and skills (Tyasti and Caraka, 2017, p. 3286).

Public services differ from private services in several characteristics, which are essential for developing a culture for the successful implementation of Total Quality Management. Public services consisting of functions, responsibilities, and authority structures (Zehira, C., Ertosunb, O., Zehirc, S. and Müceldillid, 2012). Moreover, most public services are resource-rich, and Total Quality Management aims to provide better value for the resource used.

Measuring quality in public institutions is often performed by quantifying the current level of performance within a local public administration in accordance with performance standards (Zehira, et al., 2012), or by using indicators, through which the organization establishes its purpose and the ways in which it can achieve it (Titu and Vlad, 2014, p. 132). These are grouped into 5 categories, namely: (1) The quality of services offered to citizens by human resources in public administration; (2) The degree of professionalism shown in the relationship with citizens - individuals employed in the legal field; (3) The level of quality established at the level of coordination between branches of the public institution; (4) The coordination between objectives set by the public institution and the quality of services provided; and (5) The reduced response time to requests as an indicator in the quality management of services.

In assessing the quality of services, public institutions use the ISO 9000 class quality standards, aiming therefore to develop a set of principles, criteria, and procedures used by public institutions to ensure quality, influencing TQM practices, as well as the level of competitiveness and customer satisfaction. The ISO 9000 standards share the idea that a number of defining characteristics for a quality management system can be standardized in order to become efficient in improving various aspects of quality.

The spread of the European standard for quality management in the private sector at the end of the 1980s resulted in the launch of the Common Assessment Framework (CAF), which represents a standard of total quality management adapted for the public sector (Dediu et al., 2017, p. 27). Among quality standards, the ISO 9001 is the first used in public administration, and together with the Common Assessment Framework (CAF), they are the most used tools of total quality management in European public administration (López-Lemus, 2023, pp. 1143-1164). The EFQM model developed by the European Union for the public sector is the starting point for the CAF model. Thus, the logic and structure of EFQM were adopted by the CAF model, trying to adapt to the requirements of the public sector (Raboca, 2013, pp. 42-43).

1.3 The role of digitalization in provinding better quality public services

The low level of digitalization and online services was strongly challenged during the pandemic period worldwide (Hudrea et al., 2023). Public institutions were forced to find solutions to continue their activities and to still offer citizens quality services. At present, they need to take into account the constant evolution and adapt to the online work mode. At the same time, digitalization represents a key aspect in the professional training process of public officials, who can improve their knowledge with the help of online courses and training materials.

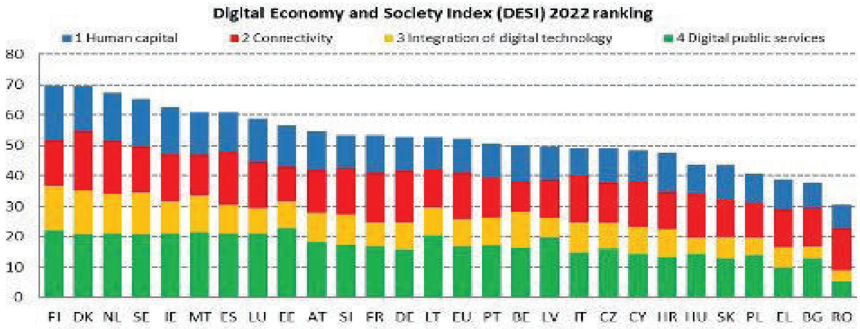
The European Commission uses the DESI index (Digital Economy and Society Index), elaborated in Romanian by the National Institute for Statistics, which summarizes relevant indicators in measuring the digital performance of the European Union and tracks the evolution of member states regarding digital competitiveness. It includes 5 dimensions, namely: connectivity, human capital, internet usage services, digital technology integration, and digital public services.

Human capital is divided into: „basic skills and usage", which includes indicators suggesting people's internet usage and the existence of their basic digital skills, and „advanced skills and development", which includes indicators on the employment of IT specialists and STEM graduates inscience, technology, and mathematics (European Commission, 2022).

The „digital public services" dimension consists of four indicators: the percentage that measures the number of internet users who have sent completed forms via the internet to an administration; the level of sophistication of existing e-government services in a country (Roztocki, N., Strzelczyk, W. and Weistroffer, 2021, pp. 3-6); how complete is the range of e-government services of a country, and the government's commitment to open data (European Commission, 2021). A general analysis of the DESI index shows that Romania has made slow progress in terms of digitalization compared to other EU countries, ranking last in 2023 with a general DESI index of 48, compared to the EU average of 77 (European Commission, 2023).

Digital Economy and Society Index (DESI) 2022 ranking

Figure 1



Source: European Commission, 2023

Similarly, in 2022, Romania's DESI index scored 30.6, compared to the EU average of 52.3, still positioning it at the last place in the ranking. Additionally, the lack of human resources and their skills regarding digitalization is a problem faced by public institutions. To all these is added the approach to digitalization from the citizens' perspective, as many are not familiar with working online or do not have electronic devices to benefit from digital public services.

In Romania, both the COVID-19 pandemic, which continued into 2021, and the periodic change of government represented a challenge (Raiu and Mina-Raiu, 2023). Digitalization remains a priority for the government, which coordinates the digital transformation of the public sector with the help of public policy in the field of e-government for the period 2021-2030, adopted on June 3, 2021, and which is in the initial phase of the application process (European Commission, 2021). In the economic field, e-government involves communication channels with lower costs with citizens, improved quality of services, and reduced delivery time. The managerial reasons behind the adoption of e-government are represented by the reform of the public sector, leading to increased efficiency of governmental management and increased accountability and transparency (Colesca and Dobrică, 2008, p. 205, Horák et al., 2021, pp. 87-89).

1.4 Digitalization as lever for improving the quality of civil servants training

Professional training or development of employees in the public sector is part of the human resources development in public administration. Every manager of a public institution should be concerned with the professional guidance of their employees since the performance and results of the organization largely depend on the quality of work and how public officials fulfill their tasks in achieving the established objectives (Ebrahimi, Mehran, and Mehran Sadeghi., 2013, pp. 5625-5643). Among the objectives pursued by the professional training of employees in the public sector are: the development of public officials' knowledge and skills, meeting the public institution's needs in the personnel and human resources domain, and increasing performance within the respective organization (Raboca, 2010, pp. 68-69).

In the context of society's digitalization and considering the challenges generated by the COVID-19 pandemic, e-learning and e-governance for public officials become increasingly important both in the case of traditional delivery of digital services and in finding new innovative methods (di Giulio and Vecchi, 2023). Information and communication technologies can enable tasks involving the collection, processing, storage, and distribution of graphical, textual, digital, audio, and video information, using electronic tools.

Moreover, a key role in the development of public administration is played by the development of e-governance (Profiroiu et al., 2023). The content and means of public officials' activities are continuously changing and evolving, so e-learning can provide a more flexible approach to meet the new challenges of public administration. As most public officials study and work at the same time, e-learning allows for combining these two activities, in line with the current situation (Khrykov et al., 2022, p. 126).

The development of e-learning systems in public administration offers many benefits to both individuals (public officials) and organizations. Employees do not need to physically attend courses in a single location, having the opportunity to study and learn in the comfort of their homes or workplaces. Another advantage is cost savings, as some training courses are far from the individual's locality, which means minimizing costs on the organization's part regarding accommodation, meals, and transportation for the beneficiary.

E-learning can also have disadvantages. Developing e-learning content involves a long time and technological resources. The unavailability of adequate IT infrastructure and poor internet connection could be an impediment for some organizations. Reluctance towards new technology, older age, and lack of motivation from both employees and trainers could also hinder the development of the e-learning process (Msoni et al., 2016, pp. 42-43).

According to Romanian national legislation, the professional training of public servants is perceived as a right but also as an obligation, and the expenses of institutions regarding the professional training of public officials must be included in the annual budget. Professional training of human resources at the level of public administration is currently provided by the National

Institute of Administration (NIA), which is responsible for training most of the staff of public administration (Simion et al., 2023).

Adapting to the present digital era and wishing to offer public officials quality professional training services, starting with 2021, the National Institute of Administration updated the existing online training platform to access the latest available version of the Learning Management System Moodle, configuring a new theme optimized for platform use on mobile devices. Moreover, the necessary connections between the NIA training platform and the portal „ina.gov.ro" were created to present the professional training programs organized by the institution and the automatic update of information in real-time. Thus, participant registrations for programs, course delivery, participant evaluation, and issuing of graduation certificates are carried out using this platform (NIA, 2023).

2. METHODOLOGY

The paper aims to analyze the impact of digitalization on the quality of the professional training process of public servants provided by the National Institute of Administration (NIA), in the post-pandemic context, from the perspective of NIA employees. In order to reach this goal our first research objective was to identify NIA᾽s employees perception related to the effects (both positive and negative) of digitalization on the professional training process.

The positive effects that were investigated looked at aspects such as time economy, quality of digital training materials, quality of teaching, while the negative effects focused on the difficulties for human resources to adapt to the changes in the process of providing online training, namely resistance to change, inadequate level of employees᾽ digital skills in order to work with an online training platform, loss of contact with learners and partner institutions representatives.

Second, we investigated ways to improve the quality of professional training of public officials using digital tools, such as: training materials used for courses (course textbooks, PowerPoint presentations, exercises and case studies, work plan, work agenda, and evaluation tests translated into online format through the tools offered by the NIA platform); Online feedback forms for learners; NIA training platform, updated to access the latest available version for the Learning Management System Moodle and configuring a new theme optimized for platform use on mobile devices; trainers ᾽ teaching methods.

The research objectives were reached using both a qualitative and quantitative methodology. In an early stage of the research we wanted to gain as many explorative insights as possible so we used secondary data analysis to investigate the impact of digitalization on the professional training process for public officials provided by NIA, from the perspective of participants to online professional training courses. By examining the pre-pandemic and post-pandemic activity reports of NIA, the evolution of the participation rate of the trainees was analyzed, as well as the extent to which their satisfaction increased or decreased as result of the digitalization of the training process.

Later on, in April 2023 we designed and conducted an online semi-structured survey, administered to NIA employees. The sample consisted of 36 employees from several NIA units such as: Directorate of Professional Development Programs (32%), Directorate of Specialized Training Programs (32%) and Continuous Education Territorial Centers for Public Administration from Constanța, Craiova, Iași, Sibiu, and Timișoara (36%). The questionnaire containing 17 close-ended questions and 3 open-ended questions, was administered online and has a 96% response rate.

The socio-demographic profile of the respondents is characterized by 68% female and 32% male, 91% hold non-management positions and 9% hold management positions. Most respondents have between 42 and 50 years of age and over 10 years of work experience.

3. CASE STUDY: THE ROLE OF DIGITALIZATION IN PROVIDING QUALITY PROFESSIONAL TRAINING FOR PUBLIC SERVANTS WITHIN THE NATIONAL INSTITUTE OF ADMINISTRATION (NIA)

3.1. Case selection

We limited our case selection to the National Institute of Administration (NIA) because it is a public organization that is very proactive and dinamic both in applying quality management principles, as well as digital transformation initiatives. Thus, NIA implemented two ISO quality standards, managing to establish an integrated management system, in accordance with the standards' requirements, harmonizing the documents and forms of the management system with the aim of increasing the quality and efficiency of its activities (NIA, 2023).

Regarding digitalization, NIA has impremented a Moodle platform for blended learning since 2018, but it took two years for it to become operational, because the pandemic created the appropiate context and pressure. Thus, in 2020 NIA digitalized its training activities from registration of participants, course delivery, skills assessment to the issuance of graduation certificates and digital supplements, including the selection process of course trainers (NIA, 2023).

The choice for NIA as a case study was also driven by the fact that there is a large body of literature that focuses on the practice of quality management and the antecedents of digital transformation from the perspective of private organizations, and only rarely from a public sector perspective (Wahi and Berenyi, 2023).

3.2. Case description

3.2.1 TQM initiatives within The National Institute of Administration

The integrated management system of the National Institute of Administration includes a quality management component (ISO 9001:2015) and an anti-bribery management component (ISO 37001:2017). It was certified in compliance with the requirements of the ISO 9001:2015 standard in the field of professional training in public administration on May 24, 2021. When the requirements of the quality management standard were implemented, 6 specific system procedures were designed and the already existing system or operational procedures at NIA were updated. Considering the creation and implementation of an integrated management system, the team responsible for implementing the Methodology for managing corruption risks and assessing integrity incidents at the level of INA worked together with the quality management component implementation team (NIA, 2023).

As part of the project „Quality, Standards, Performance - the premises of efficient management at the Ministry of Development, Public Works and Administration", the CAF tool was implemented in 2020. In 2021, the CAF Self-Assessment Group, composed of representatives from each functional structure of INA, went through the CAF evaluation stage, which aimed to analyze the level of achievement of the criteria in the INA improvement plan as a result of CAF implementation, the level of accomplishment of the objectives specified in the CAF plan, and the extent to which achieving the objectives contributed to improving the institution's activities. Furthermore, the National Institute of Administration was registered in the European database of CAF users, managed by the European Institute of Public Administration (EIPA) in Maastricht (NIA, 2023).

3.2.2 Digitalization initiatives within The National Institute of Administration

Before the COVID-19 pandemic, professional training courses for public officials were conducted face-to-face, although there was the possibility for participants to register online. The Moodle platform was used by NIA since 2018, through the implementation of a modern blended learning system, which later on, during the pandemic, allowed a quick transition to the online environment. The professional training process within NIA was digitized starting with 2020. Thus, participant registration, course delivery, competence evaluation, and issuing of graduation certificates and descriptive supplements in digital format were carried out using the online training platform and IT applications developed using internal resources.

In 2021, the digitization process continued and focused on two essential aspects: NIA training platform and the development of digital training materials. Additionally, NIA training platform was updated to access the latest version available for the Learning Management System Moodle and a new theme optimized for platform use on mobile devices was configured. Furthermore, the necessary connections between NIA training platform and the portal „ina.gov.ro" were created to present the professional training programs organized by the institution and the automatic update of information in real-time.

The digitization process was also extended to the recruitment, selection and management of specialized staff. Also, the materials used by specialized training programs and professional development were converted to digital format, using dedicated software applications, such as Articulate and Livresq Authoring Tools, and multimedia libraries. The obtained learning materials were designed in a standard format (SCORM) to facilitate navigation on the NIA training platform, contributing to ensuring the interactivity of the online training process and the self-assessment of participants' knowledge (NIA, 2023).

Within this institute, there are fee-based online professional development programs available from NIA's offering or at the request of institutions, beneficiaries, and public authorities, as well as specialized training programs. Through them, NIA aims for professional training to contribute to acquiring the necessary competencies to improve the quality of public services. In this respect, human resources in public administration can improve their specialized knowledge, exchange best practices and ideas, finding support in solving common problems.

Moreover, in the context of the pandemic, NIA implemented the project „Online Professional Development Programs", through which free professional development programs were offered to human resources in local and central public administration. Within the project 22 professional development programs (addressing topics such as internal public audit, project management, internal managerial control system, human resources management in public administration, emergency management, social media in administration, finance and accounting, urban planning, internal communication, public procurement, etc.) were developed and offered for free (NIA, 2023).

Enrollments in training programs organized by NIA, asynchronous learning modules, participants' knowledge evaluation, and the issuance of digitally signed graduation certificates are carried out through the Moodle platform. Participants just have to create an account on the platform, select the course they wish to take, check the date and time it is held and upload the requested documents on the platform. The program manager checks the enrollees' documents, creating a folder for each with these documents. Individuals who have uploaded all the requested documents and whose paperwork is in order are declared participants.

Following the completion of professional development programs, participants receive digital graduation certificates, accompanied by supplements describing the acquired competencies. The certificates are automatically generated from the online training platform, in digital format, based on the final grades from the trainer's register who conducted the program and have a QR code for each participant, which they can scan to view and verify the content. Thus, each participant can download their certificate from their account.

Additionally, course beneficiaries can provide online feedback to the trainers, which will be included in the NIA training platform's database. Measures will be taken to improve the quality of the institute especially if the feedback is negative.

To identify the training needs in public administration and the ways in which training programs can contribute to the professionalization of personnel, NIA organized 6 webinars – consultations with public institutions, associative structures, professional organizations, non-governmental organizations and public providers of training for public administration at the national level, during the period of March to April 2021.

The impact of all these digitalization initiatives are to found in NIA's activity reports, that highlight an increase in the number of participants in professional training courses and specialized training. Thus, in 2019, before the pandemic, a number of 103 participants in specialized training programs and 1063 participants in professional development programs were recorded, by comparison with the following years, when these numbers increased, reaching in 2022 a total of 305 individuals who underwent specialized training programs and 1412 individuals who underwent professional development programs (NIA, 2023).

4. RESULTS AND DISCUSSIONS

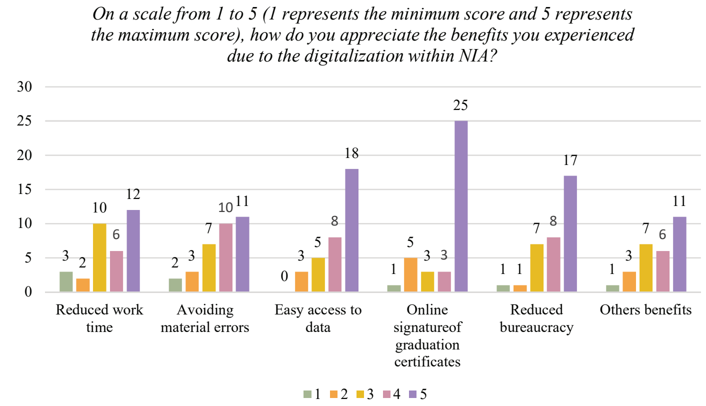
As previously mentioned, this study employed the opinion survey method, based on an online questionnaire, whose main findings are synthetized below, in relation with the three hypothesis that were formulated.

The first hypothesis tested whether the level of digitalization of the public officials' training process is perceived by NIA's employees to have a very positive impact on quality of their work.

Most respondents appreciate to a hight and very high extent (scores 4 and 5) benefits that derive from digitalization, such as: the implementation of the online signature of graduation certificates (82%), easy access to data (76%), reduced bureaucracy (74%), avoidance of material errors (62%), reduced work time (53%), and other benefits (50%).

Benefits perceived by NIA human resources after post pandemic digitalization

Figure 2



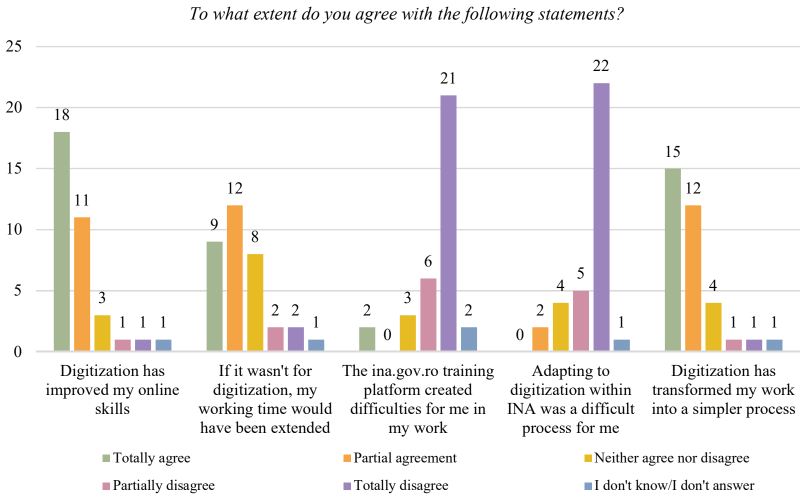
Source: questionnaire results, 2023

When asked to evaluate NIA's online training process on a scale from 1 to 5, 65% of the respondents awarded high scores (4 and 5), 21% gave a medium score of 3, and 14% gave a score of 2. Calculating the average of the scores, it results that respondents rated the current online professional training process with 4.11 out of 5, indicating that NIA employees are satisfied with the current level of digitalization of the professional training process.

Moreover, most respondents totally and partially agreed with the following statements „Digitalization has improved my online professional skills" (82%), „Digitalisation has transformed my work in a simpler process" (79%), „Without digitalization my work time would increase" (62%).

The efects of digitalization on NIA' employee

Figure 3

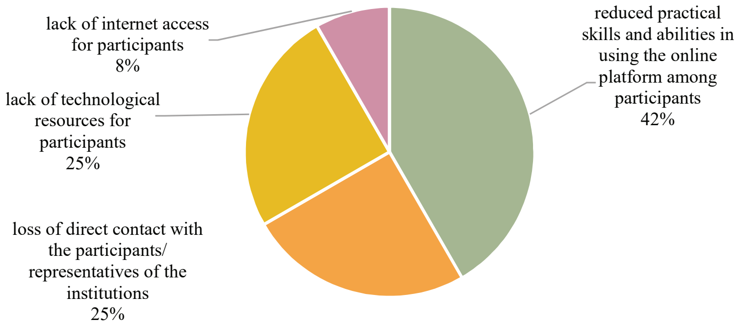


Source: questionnaire results, 2023

Although digitalization has brought many benefits to course participants, as well as employees, they have also encountered obstacles (Figure 4). By far the most important barrier seems to be linked with reduced practical skills and abilities in using the online platform among participants (42%).

Obstacles for the post-pandemic digitalization process

Figure 4



Source: questionnaire results, 2023

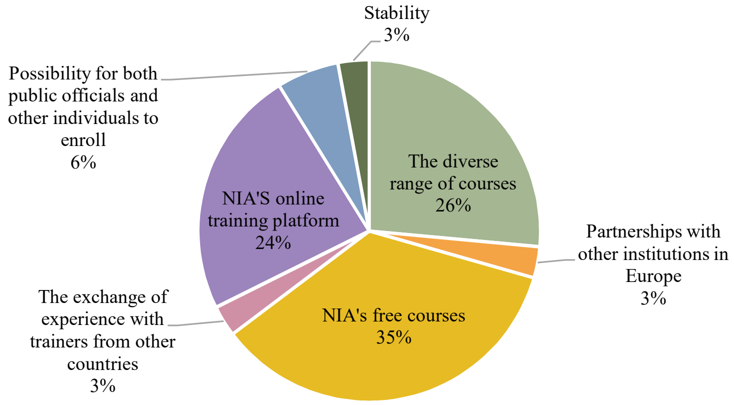
The second hypothesis tested whether the level of digitalization of NIA's training offer has a positive impact on NIA's image, expressed as perception of best provider of training courses for public servants and as increase in the number of course participants.

Out of 34 valid responses to the question „To what extent do you agree that the digitalization of the stages of conducting professional development and specialized training programs after the pandemic has contributed to an increase in the number of participants?", 67% of the respondents expressed their total or partial agreement.

The majority of respondents consider that the offer of free courses has contributed most to improving NIA's image (35%), followed by the diverse range of courses (26%), and the online training platform (24%) (Figure 5). Given that the 3 most important aspects for respondents are all linked to the digitalization of NIA's training offer (online courses and online platform) we can conclude that NIA's digital transformation improved significantly the reputation of the institution.

Aspects that contribute to the improvement of NIA's image

Figure 5



Source: questionnaire results, 2023

The third hypothesis tested whether the digitalization of the professional training process leads to increased levels of satisfaction and participation of beneficiaries (learners) in professional training courses.

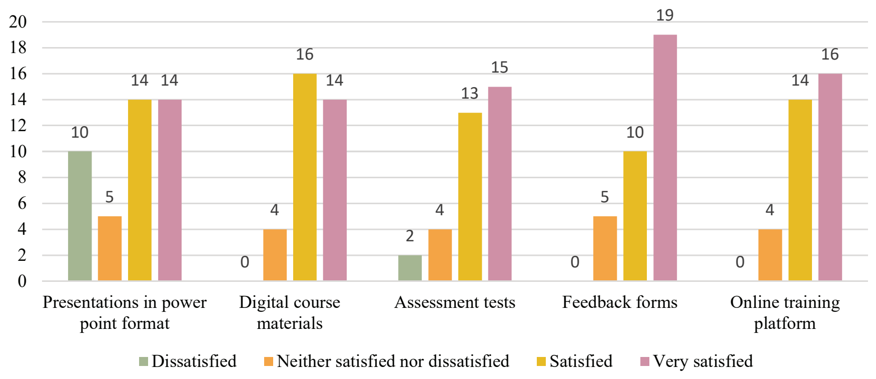
The majority of NIA staff (59%) consider that learners' satisfaction has increased following the digitalization of the stages of conducting professional training courses, while only 6% of respondents think that digitalization has contributed to a small extent to increasing beneficiary satisfaction.

In analyzing the impact of digitalization on the professional training process we also considered the digital tools used by trainers to observe the degree of participant satisfaction (Figure 6). The analysis of participant satisfaction concerning the digital tools used by trainers is conducted using as a reference point the traditional tools used before the pandemic. The statistical distribution of the results shows that participants are please with all online tools used during courses, as they were evaluated with „satisfied” and „very satisfied” by the majority of respondents.

Learners˒ satisfaction with NIA's digital tools

Figure 6

How do you rate the degree of satisfaction of the participants related to the following digital tools?



Source: questionnaire results, 2023

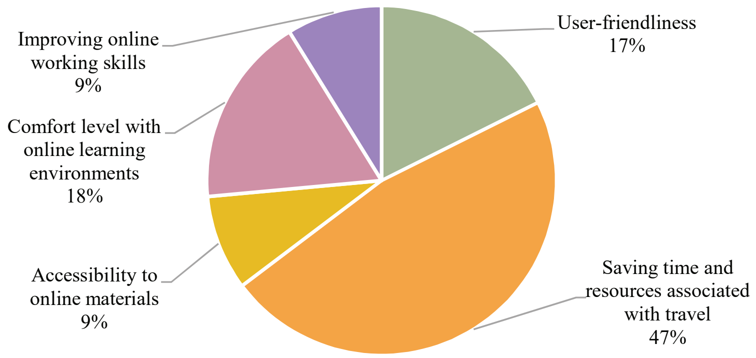
The results highlight a significant trend in the preferences of NIA's training participants towards digital learning. The reduction in travel time as a primary benefit, as acknowledged by 47% of respondents, underscores the value placed on the convenience and accessibility that digital platforms offer. This preference for online participation over physical attendance is reflective of the broader acceptance and appreciation for digitalization within professional training contexts.

The relatively lower emphasis on the importance of improving online working skills and easy access to online materials, as perceived by only 9% of respondents, may suggest that while these aspects are recognized, they are not the primary drivers for the preference for online learning. Instead, the convenience factor, particularly in terms of saving time and resources associated with travel, stands out as a key advantage of digitalized training offerings.

This trend points towards an increased comfort level with online learning environments among participants, highlighting the effectiveness of digital platforms in facilitating professional development. It also suggests that as digitalization continues to evolve, training providers like NIA may need to focus on enhancing the accessibility and user-friendliness of their online offerings to meet the growing demand for remote learning opportunities.

The benefits for participants after digitalization of NIA's professional training

Figure 7



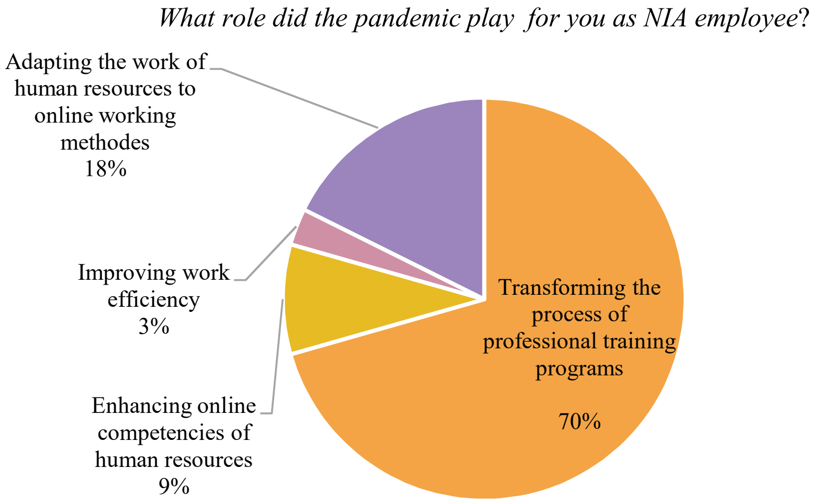
Source: questionnaire results, 2023

From the responses provided, it is obvious that the majority of the surveyed individuals (70%) consider that the pandemic has transformed the course delivery process. 18% consider that the role of the pandemic was to adapt their work to online working methods, while 9% think that improving online skills represents an effect of the pandemic. Efficiency in work (3%) ranks last, being considered less important in analyzing the effects of the pandemic on NIA.

This insight reflects a significant shift towards digital transformation within NIA, triggered by the pandemic's challenges. The emphasis on adapting to online working methods and enhancing online competencies underscores the accelerated transition to digital platforms and the necessity for both trainers and participants to acquire new skills in this changed environment. The relatively low importance attributed to work efficiency improvements may suggest that the primary focus during the pandemic was on continuity and adaptation rather than on optimizing productivity. This scenario highlights the urgent need for adaptation faced by professional education institutions during unprecedented times, leading to a rapid embrace of digital tools and platforms to ensure the uninterrupted delivery of professional training programs.

The role of pandemic in NIA

Figure 8



Source: questionnaire results, 2023

5. CONLUSIONS AND LIMITATIONS

Unlike many public institutions that were forced to carry their activity online during the pandemic, but then returned to their usual face-to-face operations in the post-pandemic period, NIA continued its' digital delivery mode. This change was brought about by the many benefits of digitalization that contribute to a better quality service, but also to increased levels of satisfaction among the users of digital training services, learners and employees.

Thus, participants prefer online courses over the physical ones because their travel time is reduced, the comfort of participation from home is increased, and participation is simplified due to the digitalization of course delivery stages. On the other side, NIA's human resources perceived the following benefits as a result of digitalization: reduced working time, avoidance of material errors, easy access to data, online signing of graduation certificates, and bureaucracy reduction.

Nonetheless, there are several negative effects were highlighted by the respondents: both employees and learners faced difficulties regarding online work and using the training platform. The online conduct of programs, using tools from the NIA training platform, accessing information, and uploading documents required for registration created difficulties for some individuals not familiar with technology. Additionally, not all those who enroll in NIA programs have the necessary means to participate well in online programs (video cameras, microphones) or face a poor internet connection. The loss of direct contact with representatives of institutions, public authorities, and participants seems to be a negative effect of the pandemic as well.

Considering NIA˒s particular organisational history and profile our assumption is that the digitalization process in this organization continued to grow even after the pandemic ended because of the quality management approach, based on practices such as top management support, customer focus, continuous improvement and continuous training and education for the human capital. However, this assumption needs more in depth research and may be considered as a future research direction of this study.

The main limitation of the research is the fact that the survey was limited to NIA employees, who may have been subjective when asked to assess benefits and challanges of digitalization process or the quality of their work as a result of digital transformation.

Bibliography

1. Bellodi, L., Morelli, M. and Vannoni, M. (2024) „A costly commitment: Populism, economic performance, and the quality of bureaucracy." *American Journal of Political Science* 68, no. 1 (2024): 193-209.
2. Carlos, V., Mendes, L. and Lourenco, L. (2014). The Influence of TQM on Organizational Commitment, Organizational Citizenship Behaviours, and Individual Performance. *Transylvanian Review Of Administrative Sciences*, 10(SI), 111-130. Available at https://rtsa.ro/tras/index.php/tras/article/view/416
3. Colesca, S. and Dobrică, L. (2008) "Adoption and use of e-government services: The case of Romania", *Journal of Applied Research and Technology*, Vol. 6, No. 3.
4. Dediu, B. , Moga, L. and Neculiță, M. (2017) *Calitatea în managementul administraţiei publice*, Calitatea: Acces la Success, Vol. 18,No. 159**.**
5. Dobrin, C. (2005), *Calitatea în sectorul public*, București: Editura ASE.
6. Drăgulescu, N. (2013) Calitatea serviciilor în administrația publică, *Calitatea: Acces la Success*; Vol. 14, No. 133.
7. Ebrahimi, Mehran, and Mehran Sadeghi. „Quality management and performance: An annotated review." *International Journal of Production Research* 51, no. 18 (2013): 5625-5643.
8. European Commission (2021), *Digital Economy and Society Index (DESI),* available at https://digital-strategy.ec.europa.eu/en/policies/desi-digital-public-services

9. European Commission (2022), *Digital Economy and Society Index (DESI) 2022*, available at https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2022
10. European Commission (2023), *Report on the State of Digital Decade 2023. Annex Romania*, available at https://digital-strategy.ec.europa.eu/en/library/2023-report-state-digital-decade
11. di Giulio, M. and Vecchi, G. (2023) „Implementing digitalization in the public sector. Technologies, agency, and governance." *Public Policy and Administration* 38, no. 2 (2023): 133-158.
12. Haruța, C. and Radu, B. (2010) „The Invisible Hand or What Makes the Bureaucracy Indispensable? A Short Theoretical Inquiry Into the Bureaucracy's Role in the Policy Making Process." *Transylvanian Review of Administrative Sciences* 6, no. 29 (2010): 62-70.
13. Horák, J., Bokšová, J. and Bokša, M. (2021) „Implementation of eGovernment from the Perspective of Public Administration." *International Advances in Economic Research* 27 (2021): 87-89.
14. Hudrea, A., Spoaller, D. and Urs, N. (2023). "Digital Tools in Romanian Higher Education: The Influence of the COVID-19 Pandemic on the Digitalization of Universities". *Transylvanian Review Of Administrative Sciences*, 19(69), 44-63. doi:http://dx.doi.org/10.24193/tras.69E.3
15. Ilieș, L. and Crișan, E. (2011) *Managementul calității totale*, Cluj-Napoca: Risoprint.
16. Khrykov, Y., Ptakhina, O., Sych, T., Dzvinchuk, D. and Kormer, M. (2022), Trends in the Development of e-Learning for Civil Servants, *Proceedings of the 1st Symposium on Advances in Educational Technology (AET 2020),* Vol.2.
17. López-Lemus, J. A. (2023) „ISO 9001 and the public service: an investigation of the effect of the QMS on the quality of public service organizations." *International Journal of Organizational Analysis* 31, no. 4.
18. Mina-Raiu, L., Bucura, I. A. and Raiu, C. V. (2021) „Transposing good practices in the Field Of Quality Management in Japan, within Romanian Public Administration." *Romanian Statistical Review* 2.
19. Morgan, C. and Murgatroyd, S. (1994) *Total Quality Management in the Public Sector*, Philadelphia: Open University Press.
20. Msoni, A.P., Munapo, E. and Choga, I. (2016) The conceptualisation of e-Learning at the public sector, *Problems and Perspectives in Management*, Volume 14, no.4.
21. NIA (National Institute of Administration/ Institutul Național de Administrație) (2021) *Raport de activitate al Institutului Național de Administrație 2021* [online], available at https://ina.gov.ro/wp-content/uploads/2022/02/Raport-de-activitate-NIA-anul-2021.pdf
22. Petersen, P. B. (1999) „Total quality management and the Deming approach to quality management." *Journal of management History* 5, no. 8.
23. Popescu, R. I., Sabie, O. M., Trușcă, M. I. (2023) "The Contribution of Artificial Intelligence to Stimulating the Innovation of Educational Services and University Programs in Public Administration", *Transylvanian Review of Administrative Sciences*, [S.l.], pp. 85-108, Oct. 2023.
24. Profiroiu, C., Negoiță, C. I. and Costea, A. V. (2023) „Digitalization of public administration in EU member states in times of crisis: the contributions of the national recovery and resilience plans." *International Review of Administrative Sciences* (2023): 00208523231177554.
25. Raboca, H. (2010) Formarea profesională a funcționarilor publici din cadrul instituțiilor publice din regiunea de nord-vest, *Revista Transilvană de Ştiinţe Administrative*, 2(26), 68-69.
26. Raboca, H. (2013) *Managementul calității în administrația publică*, București: Tritonic Books.

27. Raiu, C. and Mina-Raiu, L. (2023) „Who Runs Public Administration? A Longitudinal Study of Technocratic Ministerial Appointments in Post-Communist Romania (1991–2021)." *Transylvanian Review of Administrative Sciences* 19, no. 70 (2023): 109-127.
28. Roztocki, N., Strzelczyk, W. and Weistroffer, H. (2021) *Impact of Pandemics on e Government Services: A Pilot Study*, AIS Electronic Library (AISeL), pp.3-6
29. Simion, C .E., Nastacă, C. C., Drăguț, M. V. and Ștefănescu, M. S. (2023) „European models of professional training in public administration: a comparative approach" *Management Research & Practice* 15, no. 1.
30. Stringham, S. H. (2004) „Does quality management work in the public sector." *Public Administration and Management: An Interactive Journal* 9, no. 3 (2004): 182-211.
31. Talib, F. , Rahman, Z. and Qureshi, M.N. (2012) Total quality management in service sector: a literature review, *Int. J. Business Innovation and Research*, Vol. 6, No. 3.
32. Titu, A. and Vlad, A. (2014) *Quality Indicators in Reference to the Evaluation of the Quality Management of Services in Local Public Administration*, Elsevier B.V.
33. Titu, M. and Bucur, A. (2015) *Models for quality analysis of services in the local public Administration*, Springer Science and Business Media Dordrecht.
34. Tyasti, A. and Caraka, R. (2017) *A Preview of Total Quality Management (TQM) in Public Services*, E-Jurnal Ekonomi dan Bisnis Universitas Udayana 6.9.
34. Vinni, R. (2007) "Total quality management and paradigms of public administration." *International Public Management Review* 8, no. 1.
35. Zehira, C., Ertosunb, O., Zehirc, S. and Müceldillid, B. (2012) *Total Quality Management Practices' Effects on Quality Performance and Innovative Performance*, Elsevier Ltd. Selection.
36. Wahi, N. S. A. and Berényi, L. (2023) "Soft TQM elements for digital transformation in the public sector", *PRO PUBLICO BONO – Public Administration*, Vol. 3, pp. 29–48.

Is Google Trends Useful in Nowcasting Unemployment Rate During the Pandemic at Regional and National Level in Romania?

Mihaela Simionescu, PhD Professor (Full)
(mihaela.simionescu@ipe.ro, mihaela_mb1@yahoo.com)
Institute for Economic Forecasting, Romanian Academy

ABSTRACT

Given the role of Covid-19 pandemic in accelerating the digital transformation in Europe, the main aim of this paper is to explain the unemployment rate in Romania based on Internet searches for jobs during the epidemic at national and county level. Google Trends indexes for certain keywords related to jobs in Romanian language (”locuri de muncă” and ”joburi”) and the most famous websites with job announcements (”eJobs” and ”Hipo”) are considered. At national level, the unemployment rate in the period February 2020- December 2022 in Romania is explained using an autoregressive distributed lag model (ARDL) based on Google Trends index for ”locuri de munca”, while for youth unemployment ”joburi” and ”eJobs” are relevant. At county level, a spatial error model based on searches for ” locuri de munca” performs better than OLS regression. The results support the recommendations to improve governmental making-decision process.

Key-words*: unemployment; Google Trends; OLS; spatial model; ARDL model*

1. INTRODUCTION

The Covid-19 pandemic has accelerated digital transformation at global level, but it has also enhanced the concerns related to health and unemployment. Mangono et al. (2021) concluded that Google searches on unemployment were among the most popular searches during the pandemic in the US together with searches related to health issues. The same conclusion is suggested by Sotis (2021) for all US countries, for the District of Columbia and three US states (Nevada, Mississippi, and Utah). Using a Lotka–Volterra model, the author has proven that during the pandemic the relationship between Google searches on unemployment, news and symptoms has become strong.

Google Trends accessed through https://trends.google.com/trends/ provides information related to requests made to the Google search engine in all the countries. This tool is based on a random sample representative of all queries which are handled by Google daily (Caperna et al., 2020). Normalization of searches is made based on location of the request and time. Search results are normalized to the time and location of a query. Each data point for a specific time and location is divided by the sum of all searches to get the relative popularity that is scaled on a range of 0 to 100 considering one query's proportion to total searches on total queries. The index is named Google Trends Index and is denoted by GTI. Two main advantages are explained by Caperna et al. (2022): provision of data for indicators that are released with low frequency from official sources and the low sensitiveness to the small sample bias.

Even if it is a powerful tool, Google Trends presents more limitations. It depends on the Internet penetration, social and economic status of the individuals, and category of age. Young people are more likely to use Google to search for jobs. On the other hand, the old people are likely to be retired and are not interested in using Internet to find out job opportunities. People with high degree of poverty and/or illiterate that are looking for a job have less chances to use the benefits of technological progress to integrate on labour market.

The research question is related to the utility of Google in explaining the unemployment in real time in Romania during the pandemic when digital transformation has accelerated. The hypothesis that will be checked is that Internet searches for jobs explained unemployment rate in pandemic at national and regional level. Considering this hypothesis, the main objective of this paper is the evaluation of the capacity of Google searches for jobs to explain and predict unemployment rate. Starting from this general objective, two specific objectives are analyzed based on a regional approach: the hypothesis is checked at national level and at county level. Therefore, the time series approach at national level (OLS regressions and ARDL model) is combined to spatial approach at county level (OLS regressions and spatial autoregressive model). The results show common conclusions as well as differences: Google searches for the same key-words (”locuri de munca” and ”joburi”) explain unemployment at both levels and other key-words are relevant only at national level (”eJobs” and ”Hipo”).

These results are a step forward in nowcasting unemployment rate in Romania, a country where official statistics are still released late. The paper continues with a short presentation of previous achievements in different countries on the unemployment- Google searches nexus. The next sections present methodology, results with discussion and conclusions.

2. LITERATURE REVIEW

There are two major international contexts that have intensified the research around the utility of Google Trends in nowcasting and forecasting unemployment rate. In the first stage, the use of Google Trends is related to the necessity to anticipate the evolution on labour market in the context of global economic crisis. Therefore, econometric models were built mostly on time series to explain the evolution of unemployment in real time. Besides the econometric models on time series, some studies checked for Granger causality between unemployment and Google searches for jobs (Askitas, Zimmermann 2009a, Su 2014) and the causality was validated in one way or another. The Granger causality test on stationary time series makes also the subject of our paper. Two seminal papers of Askitas and Zimmermann (2009 a,b) established a significant relationship between unemployment and Google searches for jobs in Germany. The key-words used for searches refer to unemployment state (*unemployment rate*), job opportunities (*job search, short-term work most popular search engines in Germany*) and various institutions that manage the unemployed issues (*labour office, unemployment office or agency, Personnel Consultant*). In our paper, we employed as key-words those expressions referring to job opportunities (locuri de munca, joburi- words used for jobs in Romanian language) and the most important websites used for job searches (eJobs and Hipo). A previous study for Romania of Simionescu (2020) used two of our key-words (locuri de munca and joburi) and another similar word (angajari). However, the paper of Simionescu (2020) did not take into account the websites from Romania where job offers are posted. Two arguments are provided for these types of key-words in our paper. First, job opportunities are more appropriate than unemployment, since someone who is no unemployed might search for unemployment to understand better the concept. Second, websites used for job searches are more relevant than labour offices during the pandemic when people try to limit the physical contact as much as possible. Moreover, the efforts for digital transformation during the epidemic made online search a more simple and affordable way to search for jobs.

In general, the digital transformation has been made fast in developed countries even before pandemic. Therefore, in pre-pandemic period most of the studies discussing the role of Google Trends in anticipating the unemployment were made for developed countries like Norway, UK, Italy, France, Spain, Canada, US, Israel (Anvik, Gjelstad 2010, McLaren, Shanbhogue 2011, D'Amuri 2009, Naccarato et al. 2018; Fondeur, Karamé 2013, Vicente et al. 2015, Dilmaghani 2019, Choi, Varian 2012, Suhoy 2009). However, Google Trends was not a successful tool in improving the unemployment rate forecasts in all countries. For example, Barreira et al. (2013) indicated that only the predictions for France, Italy, and Portugal improved by including Google Trends indexes, while in Spain this tool was not efficient.

Only few studies addressed this topic in developing countries like Brazil, Turkey, Ukraine, V4 countries (Hungary, Poland, Czech Republic, Slovakia) and Romania (Lasso, Snijders 2016, Chadwick, Sengül 2015, Pavlicek, Kristoufek 2015, Oleksandr 2010, Simionescu 2020), because of low Internet penetration rate. Only the predictions made for Hungary and Czechia on the horizon January 2004 - December 2013 improved by taking into account Google queries for jobs, while in Poland and Slovakia the correlation between unemployment and Google Trends indexes was not significant (Pavlicek, Kristoufek 2015). For Romania, in pre-pandemic period, Simionescu (2020) showed that quarterly unemployment forecasts at county level based on dynamic panel data model and Google queries related to jobs are better than predictions based on models that do not include Internet data.

In the second stage, many papers analyzing the connection between unemployment and Google Trends have appeared in the context of Covid-19 pandemic and our paper belongs to this strand. Therefore, a deeper presentation is made to the research direction related to Google Trends and unemployment during the Covid-19 epidemic. Doerr and Gambacorta (2020) showed that the US regions that were more affected by Covid-19 pandemic reported more Google searches related to unemployment and pandemic.

Most of the papers implement a time series approach or panel data models and analyzed one specific country or a sample of countries. Few recent studies analyzed only one country using time series models. For example, for India the authors Fajar et al. (2020) predicted unemployment rate in Indonesia in the first months of pandemic (March-June 2020) using an ARIMAX model and "phk" (work termination) as key-word. Yurevich and Akhmadeev (2021) predicted unemployment rate in Russia during the pandemic using autoregressive models and a hybrid model that include Google Trends indexes related to job search. The NEET unemployment in Italy was nowcasted and forecasted for few years (2019-2021) by Fenga and Son-Turan (2020) using a feed-forward artificial neural network that includes Google Trend index related to job searches. Yi et al. (2021) proposed a semiparametric method called Penalized Regression with Inferred Seasonality Module based on Google Trends data to predict unemployment during the pandemic in the US.

A SARIMA model (Seasonal Autoregressive Integrated Moving Average) was used by Roopnarine and Spencer (2021) to forecast unemployment in Trinidad and Tobago.

There are other studies made for a sample of countries. A complex procedure is described by Caperna et al. (2022) for the EU-27 countries. The authors retrieved more than 400 queries connected to unemployment in each national language. The selection of the best queries is based on machine learning techniques that allow the combination of the queries and the creation of specific indicators related to these searches. Borup and Schütte (2022) showed that a panel data approach based on Google Trends information outperformed a time series approach to predict unemployment in US countries. Moreover, Brave et al. (2020) showed a positive connection between searches on unemployment and the rate of unemployment insurance during the pandemic in the US, which is explained by the variation across time for metro areas and less by variation in space. There are significant differences in this correlation during the last global economic crisis and during epidemic because of the federal Pandemic Unemployment Assistance (PUA) program.

The cross-sectional data are rarely used in papers related to Google Trends and pandemic. For example, Larson and Sinclair (2022) showed that cross-sectional data based on Google searches for jobs in the US states include relevant information that could improve forecasts of unemployment claims in spring 2020, but on longer periods a time series approach based on autoregressive models is better.

In this paper, the approach based on time series at national level has been combined with that based on cross-sectional data at county level to explain the monthly and the annual unemployment, respectively during the pandemic. The aim is not to make predictions, but to understand better if Google searches for specific key-words explained the unemployment rate during the pandemic. In this context, there are high chances to nowcast with enough accuracy the evolution of the unemployment rate in a dynamic framework based on digital transformation.

3. METHODOLOGY AND DATA

The methodology corresponds to the two specific objectives of the paper that responds to the research question at two levels: national level and county level. The analysis from a national perspective is based on time series and employs an ARDL model and OLS regressions. The research at county level is based on OLS regressions and spatial autoregressive models developed on cross-sectional data for 2020 and 2021, respectively, years corresponding to Covid-19 pandemic.

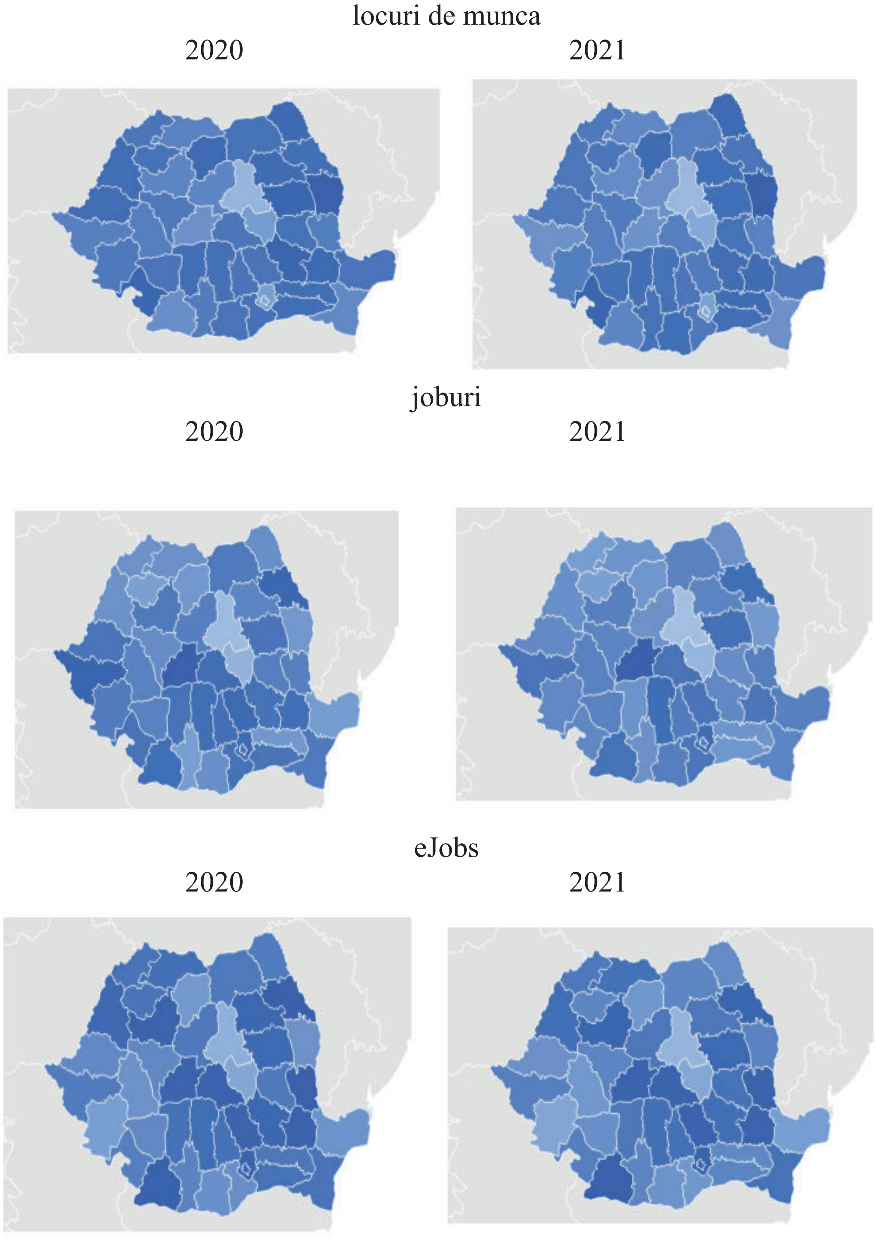
At national level, the data frequency is monthly starting with the first month of Covid-19 pandemic (February 2020) until December 2022. The unemployment rate (15-74 years) provided by Tempo online (official database of the National Institute of Statistics in Romania) and youth unemployment rate (15-24 years) from Eurostat are used as dependent variables in the models. The explanatory variables are represented by inflation rate based on index of consumer prices (monthly average value) denoted by inflation, real average wage (the nominal value is adjusted for inflation) denoted by wage, and Google Trend indexes for key-words related to job searching: ”locuri de munca” and ”joburi” that mean jobs and ”eJobs” and ”Hipo” that are two websites with job offers. The matrix of correlation for explanatory variables indicates no strong coefficients of correlation (values under 0.35).

At county level, the analysis is conducted for two different years covering the pandemic: 2020 and 2021. The annual data series for unemployment rate used as dependent variable is taken from Tempo online. The explanatory variables with annual data series correspond to more indicators with no significant correlations between them in each year: number of emigrants in each county (permanent and temporary migrants) provided by Tempo online (underestimated values, but more plausible compared to pre-pandemic values), real annual average wage denoted by wage, and monthly Google Trend indexes for key-words related to job searching: ”locuri de munca”, ”joburi”, ”eJobs” and ”Hipo”.

There are significant differences between the distributions of the counties according to searches for various key-words. According to Figure 1, the highest searches for ”locuri de munca” in 2020 and 2021 were made in Vaslui county (maximum value in 2020 and 2021) located in the Eastern of Romania and Mehedinti county in the South-West of the state. Sibiu (center of the country) was the county with maximum searches after ”joburi” in 2020 and 2021. The most of the searches after ”eJobs” were made in Iasi in 2020 (Eastern region) and in Dolj (South-Oltenia region) in 2021. In many counties the queries for ”Hipo” were not significant.

Google Trends indexes associated to searches for various key-words

Figure 1



locuri de munca

2020 2021

joburi

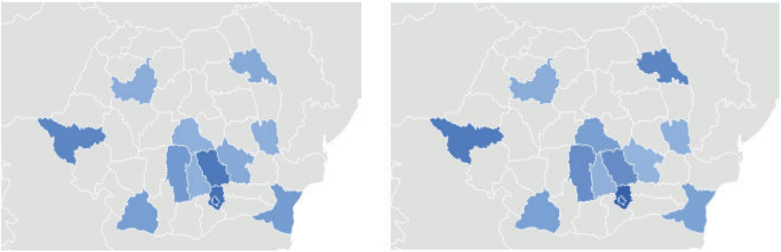
2020 2021

eJobs

2020 2021

Hipo

2020 2021



Source: Google Trends

The time series approach starts from a basic model that is extended by adding control variables:

$$unemployment_t = \alpha_{01} + \alpha_{11} \cdot GTI_t + \varepsilon_{1t}$$

$$unemployment_t = \alpha_{02} + \alpha_{12} \cdot GTI_t + \alpha_{22} \cdot wage_t + \alpha_{32} \cdot inflation_t + \varepsilon_{2t}$$

t- index for month

unemployment- dependent variable represented by total unemployment rate or youth unemployment rate

GTI- Google Trends index for a specific key-word ”locuri de munca”/ ”joburi”/”eJobs”/”Hipo”.

wage- monthly average real wage

inflation- inflation rate based on index of consumer prices

$\alpha_{01}, \alpha_{11}, \alpha_{02}, \alpha_{12}, \alpha_{22}, \alpha_{32}$- parameters to be estimated using the data series and ordinary least squares as estimation method

$\varepsilon_{1t}, \varepsilon_{2t}$- error terms

An autoregressive distributed lag model (ARDL) is also built. This type of model includes among regressors the dependent variable with a certain lag and lagged explanatory variables. Prior to estimations, a unit rot test should be applied to check for the order of cointegration. The results in the next section will indicate that GTI associated to all key-words are stationary in level, while the rest of the data series are integrated of order one (unemployment rate, youth unemployment rate, inflation rate, and wage) at 1% significance level. Given the fact that some data series are I(0) and others are I(1), an ARDL model could be constructed.

The ARDL models present more advantages: the description of short-run and long-run relationship when the time series are integrated of different orders, but without an order higher than 1 and superior estimations for small samples. Multicollinearity and serial correlation of errors might be potential issues when OLS is used. If Y is the endogenous variable and X_1, X_2,..., X_k represent the k explanatory variables, the general form of the ARDL (p, q_1, q_2,..., q_k) model is:

$$Y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{k} \sum_{l_j=0}^{q_j} \gamma_{j,l_j} X_{j,t-l_j} + \varepsilon_t \quad (1)$$

t- trend

α_0- intercept

α_1- parameter associated to trend

β_i- parameters associated to lagged values of Y

γ_{j,l_j}- parameters associated to lagged values of the k explanatory variables for j=1, 2,..., k

ε_t- innovations

If L is the lag operator, the polynomials in L are denoted by $\beta\,(L)$ and $\gamma_j\,(L)$:

$$\beta(L) = 1 - \sum_{i=1}^{p} \beta_i L^i$$

$$\gamma_j(L) = 1 - \sum_{l_j=1}^{q_j} \gamma_{j,l_j} L^{l_j}$$

Considering the polynomials in L defined above the equation (1) becomes:

$$\beta(L)Y_t = \alpha_0 + \alpha_1 t + \sum_{j=1}^{k} \gamma_j(L) X_{j,t} + \varepsilon_t \quad (2)$$

The analysis at county level is based on OLS regressions and spatial regressions to account for any spatial autocorrelation. The basic models and the extended one are represented below:

$$unemployment_i = \beta_{01} + \beta_{11} \cdot GTI_i + u_{1i}$$

$$unemployment_i = \beta_{02} + \beta_{12} \cdot GTI_i + \beta_{22} \cdot wage_i + \beta_{32} \cdot emigrants_i + u_{2i}$$

i-index for county

$\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \beta_{22}, \beta_{32}$- parameters to be estimated using the data series and OLS

u_{1i}, u_{2i} - error terms

Spatial autocorrelation should be taken into account since neighbouring counties might have similar values for unemployment. Moran's I statistic is used to check for spatial autocorrelation. If spatial dependence is accepted, then a spatial model could be constructed. One type of model is the spatial autoregressive model (SAR) with a spatial lag associated to dependent variable:

$$unemployment_t = \rho W unemployment_i + \gamma_{11} \cdot GTI_t + \gamma_{21} \cdot wage_t + \gamma_{31} \cdot emigrants_i + v_{1i}$$

The spatial error model (SEM) is another option that considers spatial dependence in the error:

$$v_{1i} = \lambda W v_{1i} + w_i$$

$$unemployment_t = \gamma_{02} + \gamma_{12} \cdot GTI_t + \gamma_{22} \cdot wage_t + \gamma_{32} \cdot emigrants_i + \lambda W v_{1i} + w_i$$

Lagrange multiplier test for spatial error and lag is used to select the best model. If the null hypothesis (spatial randomness) is not rejected, the OLS is better than the spatial model.

4. RESULTS AND DISCUSSION

At national level, a time series approach is employed based on ARDL model and OLS regression for the period February 2020-December 2022. All the data series for the variables mentioned in the previous section are seasonally adjusted using Tramo/Seats method. The Augmented Dickey-Fuller (ADF) is applied to check for unit root in the seasonally adjusted time series. The ADF test to check for unit root was considered following the descending strategy of Dickey and Pantula. In all the cases, the most suitable model for ADF equation included no trend and no intercept. The results in Table 1 indicated that unemployment rate, youth unemployment rate, inflation rate and wage are integrated of order 1, while the data series corresponding to different Google queries are stationary in level at 1% significance level.

The results of ADF test

Table 1

Variable	Type of data series	ADF stat.	p-value
unemployment rate	Data series in the second difference	10.015	<0.01
	Data series in the first difference	-5.190	<0.01
	Data series in level	0.266	0.757
youth unemployment rate	Data series in the second difference	-9.646	<0.01
	Data series in the first difference	-5.654	<0.01
	Data series in level	0.397	0.792
inflation rate	Data series in the second difference	-11.092	<0.01
	Data series in the first difference	-7.570	<0.01
	Data series in level	0.231	0.747
wage	Data series in the second difference	-9.777	<0.01
	Data series in the first difference	-5.320	<0.01
	Data series in level	3.342	0.999
locuri de munca	Data series in the second difference	-5.782	<0.01
	Data series in the first difference	-7.572	<0.01
	Data series in level	-6.608	<0.01
Joburi	Data series in the second difference	-5.399	<0.01
	Data series in the first difference	-5.109	<0.01
	Data series in level	-5.620	<0.01
eJobs	Data series in the second difference	-5.436	<0.01
	Data series in the first difference	-5.151	<0.01
	Data series in level	-5.655	<0.01
Hipo	Data series in the second difference	-6.215	<0.01
	Data series in the first difference	-7.033	<0.01
	Data series in level	-6.442	<0.01

Source: own calculations in EViews.

Granger causality could be checked only on stationary data series. According to Table 2, the variation in unemployment rate is cause for all the key-words excepting ”locuri de munca”, at different significance levels (1% for ”joburi”, 5% for ”eJobs” and 10% for ”Hipo”), but the reciprocal causality is not checked. On the other hand, the variation in youth unemployment rate is Granger cause only for GTI related to ”joburi” and ”eJobs” and none of the Google Trends indexes are causes for variation in youth unemployment.

The results of Granger causality test

Table 2

Cause	Effect	Stat.	p-value
Δunemployment	locuri de munca	1.790	0.102
Δunemployment	Joburi	5.496	0.0097
Δunemployment	eJobs	4.043	0.0287
Δunemployment	Hipo	2.681	0.0860
locuri de munca	Δunemployment	2.503	0.1000
Joburi	Δunemployment	1.229	0.3077
eJobs	Δunemployment	1.034	0.3686
Hipo	Δunemployment	0.955	0.3967
Δ youth unemployment	locuri de munca	0.014	0.9857
Δ youth unemployment	Joburi	5.802	0.0078
Δ youth unemployment	eJobs	4.125	0.0225
Δ youth unemployment	Hipo	0.187	0.8298
locuri de munca	Δ youth unemployment	0.109	0.8966
Joburi	Δ youth unemployment	0.178	0.8375
eJobs	Δ youth unemployment	0.868	0.4305
Hipo	Δ youth unemployment	1.706	0.1998

*Source: own calculations in EViews. Note: * for p-value<0.1, ** for p-value<0.05, *** for p-value<0.01*

Since unemployment series is integrated of order one (I(1)) and is not cause for ”locuri de munca”, inflation and wage, an ARDL model might be considered in this case. An ARDL(1,2,2,1) model was built to explain registered unemployment rate and only searches in the previous one and two months had impact on unemployment. The diagnostic tests suggest that the errors are independent, homoskedatic and normally distributed, while the model is correctly specified.

The Table 3 indicates that only after two months of searches for jobs, the unemployment at national level begins to reduce. Wage in the previous two months had a negative effect on unemployment rate which supports the idea that the increase in wage in the long run motivates unemployed to get hire faster. Inflation has a small and negative impact on registered unemployment rate. The rest of the key-words were not relevant in explaining unemployment at national level in the period February 2020-December 2022.

ARDL(1,2,2,1) model to explain registered unemployment rate in the actual month in Romania based on Google searches for ”locuri de munca” (jobs in Romanian language)

Table 3

Variable	Coefficient	p-value
unemployment(t-1)	0.836***	0.0000
locuri de munca (t)	-0.004	0.7547
locuri de munca (t-1)	0.05***	0.0062
locuri de munca (t-2)	-0.029**	0.0480
wage(t)	0.0001	0.3197
wage(t-1)	0.0003	0.1537
wage(t-2)	-0.0004**	0.0279
inflation(t)	-0.041*	0.0943
inflation(t-1)	-0.046*	0.0676
constant	9.037***	0.0063
Diagnostic tests		
R-square	0.8836	-
DW test (stat.)	1.931	-
Breusch-Godfrey Serial Correlation LM Test (autocorrelation of order 2) (stat. & p-value)	0.190	0.909
Breusch-Pagan-Godfrey test (stat. & p-value)	9.398	0.4013
Jarque-Bera test (stat. & p-value)	0.388	0.823
Ramsey RESET Test (stat. & p-value)	0.870	0.393

*Source: own calculations in EViews. Note: * for p-value<0.1, ** for p-value<0.05, *** for p-value<0.01*

In the case of variation in monthly youth unemployment rate, only the GTI for ”joburi” and ”eJobs” had a positive and significant impact on the dependent variables (see Table 4). ”Joburi” is the English takeover of the expression ”locuri de munca” that is more popular among young people. Moreover, ”eJobs” is the largest online recruitment platform in Romania, with an average of two million daily users. Variation in inflation had no effect on changes in youth unemployment rate, while the increase of wage from one month to another reduced the unemployment, but the influence is very small. This supports the hypothesis that young people are more motivated to get a job due to salary (Buheji, 2019). Actually, the salary is a target for young people because it offers to young people a financial independence from their parents (Bea and Yi, 2019).

OLS regressions to explain the variation in monthly youth unemployment rate in Romania using Internet searches for jobs

Table 4

Variable	Coefficient (p-value in brackets)			
Δwage(t)	-0.002*** (0.0006)	-0.002*** (0.0006)	-0.002*** (0.0005)	-0.0025*** (0.0003)
Δinflation(t)	0.097 (0.677)	0.124 (0.594)	0.132 (0.570)	0.117 (0.611)
locuri de munca (t)	0.0067 (0.413)	-	-	-
joburi (t)	-	0.013** (0.076)	-	-
eJobs (t)	-	-	0.018** (0.073)	-
Hipo(t)	-	-	-	0.0031 (0.625)
constant	2.361 (0.9162)	-0.873 (0.969)	-1.440 (0.948)	-0.608 (0.978)
Diagnostic tests				
R-square	0.426	0.423	0.425	0.434
DW test (stat.)	1.894	1.891	1.896	1.902
Breusch-Godfrey Serial Correlation LM Test (autocorrelation of order 2) (stat. & p-value)	0.224 (0.899)	0.187 (0.903)	0.229 (0.900)	0.195 (0.901)
Breusch-Pagan-Godfrey test (stat. & p-value)	5.275 (0.152)	3.194 (0.389)	3.679 (0.298)	1.846 (0.605)
Jarque-Bera test (stat. & p-value)	1.201 (0.548)	3.472 (0.180)	2.829 (0.242)	3.461 (0.177)
Ramsey RESET Test (stat. & p-value)	1.339 (0.191)	1.398 (0.172)	1.947 (0.173)	0.152 (0.699)

*Source: own calculations in MATLAB. Note: * for p-value<0.1, ** for p-value<0.05, *** for p-value<0.01*

The approach based on cross-sectional data for 2020 and 2021 (years of pandemic) is based on OLS regressions and spatial models. There is a negative, but very small influence of wage and number of emigrants on unemployment in 2020 and 2021 as Table 5 indicates. Google searches for ”locuri de munca” has a positive, but very low impact on unemployment rate. On the other hand, OLS suggested a positive effect of searches for ”joburi” on unemployment. The SAR and SEM models based on GTI for ”joburi” indicated that the coefficient for this variable is not statistically significant. More searches for ”locuri de munca” indicate higher unemployment rate at county level.

Moran’s I value is 2.933 with p-value less than 0.01, which indicates spatial correlation for unemployment rate. Consequently, the estimations for the entire country could not properly explain the unemployment in any county. The existence of spatial dependency implies the utilisation of spatial models (SAR or SEM) as Anselin (2005) indicated. The parameters of these models were estimated using maximum likelihood method. The results of likelihood ratio test suggest that OLS model is better than spatial lag model (p-value higher than 0.05). On the other hand, the parameters for spatial lag variables are not significant at 5% level. Having the spatial lag model rejected, the conclusion is that unemployment in a county was not correlated with the unemployment from neighbouring counties.

According to likelihood ratio test, the spatial error model performs better than OLS model. Moreover, the parameter of lambda is statistically significant at 5% level. Therefore, we can conclude that the spatial error model explains better the unemployment rate in Romania at county level in 2020 and 2021. Therefore, the spatial dependence in unemployment is present among Romanian counties, but there are other variables that are not included in this model that explain the correlation between neighbouring counties.

OLS regressions and spatial models to explain monthly unemployment rate in Romania using Internet searches for jobs during the Covid-19 pandemic (in 2020 and 2021)

Table 5

Variable	Coefficient (p-value in brackets)								
	2020	2020	2020	2020	2020	2021	2021	2021	2021
W unemplyoment (spatial lag)	-	-	-	0.225 (0.112)	-	-	0.187 (0.128)	-	-
λ	-	-	-	-	0.432** (0.016)	-	-	-	0.371** (0.022)
wage(t)		-0.003*** (0.0055)	-	-0.002*** (0.005)	-0.001*** (0.006)	-0.001***(0.007)	-	-	-0.001*** (0.007)
emigrants(t)	-	0.006** (0.012)	-	0.004** (0.022)	0.0043 (0.021)	0.003** (0.024)	-	-	0.0033** (0.024)
locuri de munca (t)	0.038** (0.014)	0.056*** (0.0004)	-	0.044*** (0.0009)	0.045*** (0.0009)	0.054*** (0.0009)	0.055*** (0.0009)	-	0.061*** (0.0006)
joburi (t)	-	-	-0.029** (0.036)	-	-	-	-	-0.030* (0.052)	-
constant	1.329 (0.212)	-1.932 (0.1479)	5.752*** (0.0001)	4.765*** (0.001)	2.233** (0.011)	0.199 (0.843)	0.225 (0.769)	5.369*** (0.000)	2.445** (0.023)
Diagnostic tests									
R-square	0.439	0.576	0.487	0.606	0.678	0.475	0.511	0.433	0.536
DW test (stat.)	2.053	2.315	1.957	-	-	1.776	-	1.899	-
Breusch-Godfrey Serial Correlation LM Test (autocorrelation of order 2) (stat. & p-value)	0.719 (0.697)	1.115 (0.5725)	0.623 (0.424)	-	-	0.603 (0.739)	-	2.191 (0.334)	-
Breusch-Pagan test (stat. & p-value)	4.327 (0.114)	7.963 (0.537)	2.580 (0.275)	3.199 (0.196)	4.022 (0.181)	3.341 (0.188)	3.556 (0.136)	1.080 (0.582)	3.2 (0.228)
Jarque-Bera test (stat. & p-value)	3.759 (0.152)	3.118 (0.210)	1.372 (0.503)	-	-	4.585 (0.1001)	-	4.570 (0.101)	-
Ramsey RESET Test (stat. & p-value)	0.655 (0.485)	0.584 (0.562)	3.632 (0.162)	-	-	3.332 (0.202)	-	0.652 (0.424)	-

Likelihood Ratio test (stat. & p-value)	-	-	-	2.553 (0.117)	4.022 (0.024)	-	2.422 (0.123)	-	3.96 (0.03)
Log likelihood	-355.27	-356.67	-352.49	-340.18	-338.41	-390.48	-370.36	-379.34	-369.46

*Source: own calculations in MATLAB. Note: * for p-value<0.1, ** for p-value<0.05, *** for p-value<0.01*

These findings are subject to discussion. At county level, higher average wage motivates the unemployed to search more for a job, while the emigrants enhance the tensions on labour market even if from theoretical point of view emigration should make available more jobs for the unemployed that remain in the county. However, the influence is very small. The results are in line with Škuflić, and Vučković (2018) who show that emigration increased unemployment in nine EU Member States, including Romania, in the period 2004-2015. These findings might be explained by structural issues of the labour market determined by high emigration, including a significant supply and demand mismatch on the labour market. Remittances from relatives working in a foreign country might encourage unemployed to remain in this state. The positive correlation between emigration and unemployment is explained by endogenous growth theories that consider low substitutability and human capital externalities for well skilled and less skilled workers. On the other hand, the skilled emigrants can play the role the citizens that support the better quality of institutions.

5. CONCLUSIONS

Google Trends represents a powerful tool in a world where digital transformation is a priority for the EU countries. The Covid-19 pandemic has enhanced the digital transformation, but also increase the concerns related to unemployment. In this context, this paper proposes a deep analysis of the capacity of Internet searches for jobs to nowcast unemployment in Romania, a country that has made progress in digitization, but still encounters difficulties in releasing official statistics in time. This research came up with a national and also with a regional approach to make a comparative analysis. At national level, only the keyword ”locuri de munca” explained the registered unemployment rate since the beginning of the pandemic until the end of 2022 showing that it is required more time to search for jobs until unemployment decreases. On contrary, the youth unemployment rate is explained by searches after ”joburi” and ”eJobs” platform. These findings can improve decision-making process at governmental level and can ensure the best policies to reduce unemployment and create job opportunities.

At county level, ”locuri de munca” also explains the unemployment rate in 2020 and 2021 according to a spatial error model, but the spatial dependence is attributed to other factors that are not included in the model.

Besides the value of these results for Romanian labour market, this study is subject to more limitations. A few number of control variables was included in the model because of little data availability. Other types of econometric models should be considered for robustness test and the study does not make a comparative analysis with other countries in the Eastern Europe. Therefore, future directions of research might refer to the use of panel data models at county level and a comparative analysis with other New EU Member States.

Acknowledgement

A part of this study related to data processing in MATLAB belongs to the research entitled “Econometric models in MATLAB”, which is part of the research plan of the Institute for Economic Forecasting of the Romanian Academy for 2024.

References

1. Anselin L (2005) Exploring spatial data with GeoDaTM: a workbook. *Center for spatially integrated social science* 165-223
2. Anvik C, Gjelstad K (2010) Just Google It!. BI Norwegian Business School
3. Askitas N, Zimmermann KF (2009a) Google econometrics and unemployment forecasting. *Applied Economics Quarterly* 55(2): 107-120. https://doi.org/10.2139/ssrn.1415585
4. Askitas N, Zimmermann KF (2009b**)** Googlemetrie und Arbeitsmarkt. *Wirtschaftsdienst* 89(7): 489-496. https://doi.org/10.1007/s10273-009-0957-0
5. Barreira N, Godinho P, Melo P (2013) Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking* 14(3):129-165. https://doi.org/10.1007/s11066-013-9082-8
6. Bea MD, Yi Y (2019) Leaving the financial nest: Connecting young adults’ financial independence to financial security. *Journal of Marriage and Family* 81(2): 397-414. https://doi.org/10.31235/osf.io/t6ar5
7. Borup D, Schütte ECM (2022) In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics* 40(1): 186-200. https://doi.org/10.2139/ssrn.3423124
8. Brave SA, Butters RA, Fogarty M (2020) Another Look at the Correlation Between Google Trends and Initial Unemployment Insurance Claims.
9. Buheji M (2019) Youth unemployment mitigation labs-an empathetic approach for complex socio-economic problem. *American Journal of Economics* 9(3): 93-105
10. Caperna G, Colagrossi M, Geraci A, Mazzarella G (2020) Googling unemployment during the pandemic: Inference and nowcast using search data. *JRC Working Papers in Economics and Finance* 2020/04:1-17 https://doi.org/10.2139/ssrn.3627754

11. Caperna G, Colagrossi M, Geraci A, Mazzarella G (2022) A babel of web-searches: Googling unemployment during the pandemic. *Labour Economics* 74:1-29 https://doi.org/10.1016/j.labeco.2021.102097
12. Chadwick MG, Sengül G (2015) Nowcasting the Unemployment Rate in Turkey: Let's Ask Google. *Central Bank Review* 15(3): 1-23.
13. Choi H, Varian H (2012) Predicting the present with Google Trends. *Economic Record 88*(s1): 2-9 https://doi.org/10.1111/j.1475-4932.2012.00809.x
14. D'Amuri F (2009) *Predicting unemployment in short samples with internet job search query data*. University Library of Munich, Germany.
15. Dilmaghani M (2019) Workopolis or The Pirate Bay: What Does Google Trends Say about the Unemployment Rate?. *Journal of Economic Studies* 46(2):422-445 https://doi.org/10.1108/jes-11-2017-0346
16. Doerr S, Gambacorta L (2020) *Identifying regions at risk with Google Trends: the impact of Covid-19 on US labour markets*. Bank for International Settlements 8.
17. Dumitrescu EI, Hurlin C (2012) Testing for Granger non-causality in heterogeneous panels. *Economic modelling* 29(4): 1450-1460 https://doi.org/10.1016/j.econmod.2012.02.014
18. Fajar M, Prasetyo OR, Nonalisa S, Wahyudi W (2020) Forecasting unemployment rate in the time of COVID-19 pandemic using Google trends data (case of Indonesia).
19. Fenga L, Son-Turan S (2020) Forecasting youth unemployment in the aftermath of the COVID-19 pandemic: the Italian case 1:1-28. https://doi.org/10.21203/rs.3.rs-74374/v1
20. Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment?. *Economic Modelling* 30: 117-125 https://doi.org/10.1016/j.econmod.2012.07.017
21. Larson WD, Sinclair TM (2022) Nowcasting unemployment insurance claims in the time of COVID-19. *International Journal of Forecasting* 38(2): 635-647 https://doi.org/10.1016/j.ijforecast.2021.01.001
22. Lasso F, Snijders S (2016) The power of Google search data; an alternative approach to the measurement of unemployment in Brazil. *Student Undergraduate Research E-journal! 2*: 1-9.
23. McLaren N, Shanbhogue R (2011) Using internet search data as economic indicators. *Bank of England Quarterly Bulletin* 2: 134-140 https://doi.org/10.2139/ssrn.1865276
24. Oleksandr B (2010) *Can Google's search engine be used to forecast unemployment in Ukraine* (Doctoral dissertation, Kyiv School of Economics).
25. Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: Evidence from the Visegrad group countries. *PloS one* 10(5): 1-11https://doi.org/10.1371/journal.pone.0127084
26. Roopnarine KA, Spencer JD (2021) *Exploring the Use of Internet Searches to Predict Unemployment in Trinidad and Tobago*. Central Bank of Trinidad & Tobago.
27. Škuflić L, Vučković V (2018) The effect of emigration on unemployment rates: the case of EU emigrant countries. *Economic research-Ekonomska istraživanja* 31(1): 1826-1836 https://doi.org/10.1080/1331677x.2018.1516154
28. Sotis C (2021) How do Google searches for symptoms, news and unemployment interact during COVID-19? A Lotka–Volterra analysis of Google Trends data. *Quality & quantity* 55(6): 2001-2016 https://doi.org/10.1007/s11135-020-01089-0
29. Stewart K (2005) Dimensions of well-being in EU regions: Do GDP and unemployment tell us all we need to know?. *Social Indicators Research* 73(2): 221-246 https://doi.org/10.1007/s11205-005-2922-7

30. Su Z (2014) Chinese online unemployment-related searches and macroeconomic indicators. *Frontiers of Economics in China* 9(4): 573-605.
31. Suhoy T (2009) *Query indices and a 2008 downturn: Israeli data*. Bank of Israel.
32. Vicente MR, López-Menéndez AJ, Pérez R (2015) Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?. *Technological Forecasting and Social Change* 92: 132-139 https://doi.org/10.1016/j.techfore.2014.12.005
33. Yi D, Ning S, Chang CJ, Kou SC (2021) Forecasting unemployment using Internet search data via PRISM. *Journal of the American Statistical Association* 116(536): 1662-1673 https://doi.org/10.1080/01621459.2021.1883436
34. Yurevich MA, Akhmadeev DR (2021) Predicting the unemployment rate: Analyzing statistics on search engine query. *Terra Economicus* 19(3): 53-64 https://doi.org/10.18522/2073-6606-2021-19-3-53-64