National Bank of Serbia

# WORKING PAPERS BULLETIN

March
2024

06

National Bank of Serbia

# WORKING PAPERS BULLETIN

March
2024

06

**NATIONAL BANK OF SERBIA**
**Belgrade, Kralja Petra 12**
**Tel: 011/3027-100**
**Tel: 011/333-8000**
**www.nbs.rs**

**Number of copies: 45**
**ISSN 2787-3315**

**Editor-in-Chief:**

Dr Jorgovanka Tabaković, Governor

**Deputy Editor-in-Chief:**

Dr Željko Jović, Vice Governor

**Editorial Board:**

Nikola Dragašević, Vice Governor
Savo Jakovljević, General Manager, Economic Research and Statistics Department
Darko Kovačević, General Manager, Financial Stability Department
Darko Stamenković, General Manager, Banking Supervision Department
Mirjana Miletić, Deputy General Manager, Economic Research and Statistics Department
Mirko Đukić, Head of Division for Nowcasting Using Big Data, Economic Research and Statistics Department

**Technical Processing:**

Sofija Vučetić, Researcher, Economic Research and Statistics Department

**Issued by:**

National Bank of Serbia

**Printed by:**

Institute for Manufacturing Banknotes and Coins – Topčider

# *Foreword by the Governor*

*The period since the last Working Papers Bulletin was marked by further alleviation of inflationary pressures, the anticipation of the onset of monetary policy accommodation by leading central banks, as well as new turbulences caused by intertwined political and economic developments at the global level.*

*The globalisation of economic flows in the decades before the global economic crisis brought numerous benefits, but also heightened the sensitivity of national economies to global trends. This is evidenced also by the recent disruptions in the overseas commodity transport and increase in transport costs for companies because of the attacks on cargo ships in the Red Sea. Central banks must adapt to the more unstable environment by using a wider range of relevant information in their decision-making (compared to the traditional sources) and techniques which provide sound assessments of the current situation before official data.*

*Along these lines, the first paper in the Bulletin introduces machine and deep learning methods (LSTM neural network) and mixed data sampling (MIDAS) regression models for forecasting current movements of Serbia's GDP. Apart from official indicators, models' assessment relies also on high frequency alternative indicators, such as Google Trends and daily electricity consumption. The deep learning model showed a smaller error in GDP forecast, but it turned out that the MIDAS regression model can also help decision makers, given that its structure (as opposed to the LSTM) enables insight into current developments underlying GDP dynamics.*

*As another alternative indicator of economic trends, the frequency of occurrence of certain topics in newspaper articles can be used, which is dealt with in the following paper in the Bulletin. The Latent Dirichlet Allocation model was used to classify articles into topics in this paper. The application of thematic modelling in the Serbian language is complicated by the fact that our language is highly inflected, i.e. words have a large number of forms, which the model recognises as words with different meanings. In this paper, this aggravating circumstance was turned into an advantage by reducing only economic terms to their base form. Thus, when classifying articles into topics, they were given greater importance compared to non-economic terms. The paper showed that the frequency of writing about certain topics is a good indicator of household inflation expectations, as well as that it explains relatively well the earlier episodes of elevated inflationary pressures in Serbia.*

*The last paper in the Bulletin focuses on payment system migration from the current ISO15022 to the new ISO20022 standard, which was initiated as a project by the National Bank of Serbia and which should end in November 2025. One of the main goals of the project, not only in Serbia but globally, is to facilitate cross-border payments,*

*which are still largely characterised by high costs, low speed and insufficient transparency. In addition, when it comes to the National Bank of Serbia, the motive is to achieve compatibility with the requirements of potential connection with other payment systems and joining the SEPA region. The new electronic messages are up to three times more voluminous and structured to offer greater flexibility and adaptability to the economic conjuncture and complex regulatory requirements.*

*We hope that the papers in this Bulletin will help the reader better understand the aspects of analyses and research that are part of the decision-making process at the National Bank of Serbia.*

*In the coming period, the National Bank of Serbia will continue to carefully monitor global economic flows, as well as trends in economic analysis and further development of the tools used. We will strive to apply the best world practice in our analyses in order to provide additional information for the decision-making process and the discharge of the functions that we are responsible for, according to the Law on the National Bank of Serbia, and above all in order to achieve our main goals – price and financial stability, but also to support the economic policy of the Government of the Republic of Serbia.*

*Dr Jorgovanka Tabaković, Governor*

# WORKING PAPERS:

Working Papers describe research in progress by the author(s) and are published to encourage discussion and suggestions for future work.

National Bank of Serbia

# A COMPARISON OF USING MIDAS AND LSTM MODELS FOR GDP NOWCASTING

Iva Glišić

© National Bank of Serbia, March 2024

Available at www.nbs.rs

The views expressed in the papers constituting this series are those of the author(s), and do not necessarily represent the official view of the National Bank of Serbia.

Economic Research and Statistics Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

## A comparison of using MIDAS and LSTM models for GDP nowcasting

Iva Glišić

**Abstract:** The paper elaborates on machine and deep learning methods, as well as mixed data sampling regression models, used for GDP nowcasting. The aim is to select an adequate model that shows better performance on the data used. The paper provides an answer to the question of whether the use of deep learning methods can improve GDP nowcasting compared to traditional econometric methods, as well as whether the use of specific high-frequency indicators improves the quality of the models used. The paper examines the selection of adequate indicators – both official and those from alternative sources, presents the framework of mixed data sampling regression models and deep learning models used for nowcasting, and gives an assessment of two such models on the example of Serbian GDP. Serbia's GDP was modelled for the period Q1 2016 – Q2 2023 and the end of the observed period (six quarters) was used for the forecast. Finally, two assessed models were compared – the mixed data sampling regression model and the LSTM neural network. A special focus is placed on ways to improve both models. The LSTM recurrent neural network model had a smaller forecast error, with the use of a combination of official and alternative (high-frequency) indicators, but the mixed data sampling regression model also proved to be a good tool for decision-makers, since its structure allows insight into the ongoing movements impacting GDP dynamics. The use of alternative indicators in nowcasting improved the projections through both presented models.

**Keywords:** GDP, nowcasting, MIDAS, neural networks, high-frequency indicators.
**[JEL Code]:** C32, C45, C53

## Non-technical summary

The models used to obtain flash estimates of macroeconomic variables or the overall state of economic activity have gained in importance in recent years, especially in periods of major crises, when reliable information on the intensity of the effect of a certain phenomenon or measure is very important for economic activity. This type of modelling is related to the concept of nowcasting, used to obtain the flash estimate of an economic variable when official data are not yet available. GDP is a particularly useful variable for nowcasting. Although not without its flaws, GDP provides the most complete picture of the state of economic activity at a given time. One of the shortcomings is certainly the time lag in the publication of data for the current period, given that the official flash GDP estimate is available one month after the end of the reference quarter, while the official GDP estimate during that quarter does not exist. The nowcasting models aim to overcome these shortcomings.

One of the main questions in the assessment of such models concerns the relevant indicators that can adequately approximate the movement of GDP. Indicators suitable for use in these models must be highly correlated with the GDP indicator, but also more frequent and regular in publication. Given the need for the simultaneous use of mixed frequency data, mixed data sampling (MIDAS) regression models were developed. In addition to them, dynamic factor models (DFM), as well as their combinations, are often used in econometric analyses.

Given the need for a more precise estimate of economic activity, in addition to official indicators, alternative indicators are increasingly being used in literature and practice. Alternative indicators are usually big data and, as such, are differently structured and usually much more frequent and timely than the official ones. Nevertheless, working with such data introduces new complexities into the model, and one of the answers to the growing computational complexity of such models is the increasingly frequent use of machine learning, especially deep learning for the purpose of nowcasting.

The paper contains the assessment of two nowcasting models – MIDAS and LSTM, of which the former is a traditional econometric model, and the latter is a deep learning model. The models were assessed using a combination of official and alternative indicators as regressors, with the target projection variable being GDP growth. Both models adequately model GDP dynamics in the short run, with the LSTM neural network model proving to be more accurate, but also more difficult for evaluation and interpretation than the MIDAS model. In addition, since the LSTM model does not provide insight into the individual contributions of the variables used, the MIDAS model is much more useful in identifying triggers for changes in GDP. Also, by looking at the structure of the MIDAS model, it can be concluded that alternative indicators are statistically significant for GDP nowcasts. Due to the use of alternative indicators and the nature of macroeconomic data, both models have the problem of small samples and, consequently, overfitting to the data. As the database grows, it is expected that the performance of both assessed models, especially the deep learning model, will improve.

## Contents:

# 1 Introduction

*An economist is an expert who will explain tomorrow why the things he predicted yesterday did not happen today.*

Laurence J. Peter

Economic forecasting is not an exact science, nor are its research results always the most precise, as the quote from the beginning of this paper suggests. Uncertainty and various risks affect the probability of realisation of the projected movement of a certain macroeconomic variable. The longer the prediction horizon, the more unreliable the projections are, and the same is true when uncertainty is high, as is the case in crisis periods, when, as a rule, macroeconomic projections are most closely observed and monitored. In crisis periods, the monitoring of official indicators of economic activity and their projections indicates the likelihood of deepening of or exiting from the current crisis.

The crisis caused by the covid pandemic, directly followed by the crisis triggered by the Ukraine war, showed that a large number of economic indicators are not timely, which further complicates the process of decision-making about adequate measures aimed at preserving a stable economic environment. This is especially true for the economic indicator such as GDP, as the first economic consequences of the covid pandemic were felt already in January (transportation issues), while the GDP indicator confirmed this only in late April or early May in most countries. Therefore, a focus in this period is placed on GDP nowcasting models. These models have been analysed in economic literature for a number of years, but they are especially useful in times of sudden changes in economic activity, as they provide a quick estimate of the current effect on GDP.

When assessing the GDP nowcasting models, the explanatory variables must be adequately selected. Already in that stage, two problems arise, the first relating to the fact that a large number of official indicators whose movement is correlated with the movement of GDP also have a publication lag. Another problem is that most of these indicators are expressed in a frequency (most often monthly) that differs from quarterly GDP data.

To model actual economic activity in real time, nowcasts often rely on high-frequency data, such as daily or weekly indicators. The first chapter of this paper deals with the selection of adequate indicators – both official and those from alternative sources, which almost as a rule have a shorter lag between the end of the reference period and the publication of the flash estimate of the results in the given period. The second chapter concerns traditional econometric methods that focus precisely on solving the problem of different frequencies of dependent and explanatory variables. In the second chapter, the framework of mixed data sampling regression models (MIDAS) will be presented, and one such model on the example of Serbian GDP will be assessed.

In a significant part of the literature cited in the paper, the authors refer to the growing number of available economic indicators that have not been published by the official statistical office of a country. The exponential growth in the availability of a huge amount of data is related to the term big data – those are large, complex and diverse datasets generated from

different sources. These data are generated at high speed, are large (large data quantity) and diverse (different formats and types of data). Such data can be included in traditional econometric models and are primarily used in various machine learning models, especially deep learning. The third chapter pertains to the development and application of machine learning in econometrics, with a focus on deep learning models used in nowcasting. This chapter also assesses a deep learning model, specifically a type of a recurrent neural network. A neural network consists of connected nodes, i.e. neurons. These nodes make up layers, including an input layer, one or more hidden layers, and an output layer. Data flow through the layers, activating functions within each neuron whose outputs provide input on further use of the data. Recurrent networks allow communication that is not exclusively one-way between layers and nodes of the network. The paper uses the neural network LSTM (long short-term memory), which is particularly suitable for modelling time series for nowcasting Serbia's GDP.

In the last, fourth chapter, two assessed GDP nowcasting models are compared – MIDAS and LSTM, with a focus on methods for improving the used models.

The aim of the paper is to select an adequate set of indicators – both from official and alternative sources, for the purposes of finding Serbia's optimal GDP nowcasting model.

## 2 Selection of adequate variables for GDP nowcasting

The Serbian Statistical Office, as the main producer and disseminator of statistical data, publishes a large number of announcements, indicators, bulletins and similar reviews. Over a number of years, those were the only available data for economic analysis and macroeconomic projections. In time, various sentiment measures and analyses of the subjective expectations of economic entities (such as the inflation expectations indicator and ESI) were added.

In addition to indicators based on surveys, with the emergence of the internet, a large number of indicators became available through user interaction with webpages. An example are the search indicators on the Google engine, used in this paper. During the covid pandemic, it was very important to monitor the Google mobility index, which shows trends and changes in movements in certain regions. A topic of particular interest for economic analysis was the impact of restrictions aimed at preventing the covid fallout on the economy. Ilin et al. (2021) demonstrated that mobility indicators (such as Google's) can be used to assess the effectiveness of non-pharmaceutical intervention and to predict the spread of covid. They found that mobility data alone are sufficient to predict coronavirus infections at all geographic scales – from counties and cities, through states and provinces, to countries and eventually the entire world. In addition, models that exclude mobility data perform significantly worse, indicating the important role of mobility data as a high-frequency and alternative data source for forecasting.

In this paper, four official monthly indicators and five alternative, high-frequency indicators are proposed for the purpose of short-term GDP projections and nowcasting.

## 2.1 Statistical Office's official monthly indicators

The goal of this paper is to find the most adequate model for the purpose of short-term GDP projections and nowcasting. The most commonly used economic indicator is GDP as it represents the total value of all goods and services produced in a country in a certain period (usually a year).

In Serbia, according to the production approach, the services sector accounted for the largest portion of GDP (51% on average in 2016–2022), followed by the industry sector, with around 21%. In terms of expenditure, personal consumption accounted for the largest portion of GDP – around 70% on average in 2016–2022, while private investment made up around 16%, and government investment around 5% on average. Net exports are under a strong impact of external factors, primarily the prices of energy and cereals, as well as external demand of Serbia's largest trading partner – the EU countries, primarily Germany. In the last seven years, with the exception of 2016, net exports negatively contributed to total GDP: -1.2 pp on average.

GDP is the key indicator used by governments, policy makers, businesses and economists to monitor and analyse a country's economic performance and progress. As such, it has a great weight in deciding on the adequacy of current and direction of future economic policies, and the construction of economic sentiment. The Statistical Office publishes preliminary data on quarterly GDP outturn (the so-called flash estimate) one month after the end of the reference quarter (e.g. the flash estimate of GDP in the first quarter is published on the last day of April). This means that the first estimate of the state of the economy is only available at the end of the fourth month. Given the importance of this indicator, for decades already, great effort is invested in economic literature to assess models that will adequately predict the movement of GDP in the short run, but also estimate the current situation (nowcasting).

The Statistical Office collects data on industrial production, retail trade and goods exports and imports on a monthly basis. The first data on the movement of these indicators are available on the last day of the month following the reference month (at the end of February for the January performance). The fact that these data are more frequent and up-to-date than economic activity indicators makes them suitable for assessing the movement of total GDP, if it turns out that they have predictive power. A part of this paper deals with the selection of the optimal set of indicators for forecasting GDP dynamics.

The connection between industrial production and the overall trend of economic activity has long been the subject of research. This relationship was particularly strong in periods when the largest part of added value was that produced in the industrial sector, before the services sector took over primacy. Nevertheless, since a large number of domestic and external shocks affecting industrial production also influence economic activity as a whole, the dynamics of this indicator can still provide insight into the movement of entire economic activity.

Back in the 1980s, Stock & Watson (1988) assessed the dynamic factor model in order to obtain an index of the movement of total economic activity. For the purpose of constructing the index, data on industrial production, real personal income, manufacturing output, trade and employment were used.

Particularly interesting is the research conducted by the ECB in 2007, relating to the impact of the lag between the end of the reference period and the publication of the indicator flash estimate. Ignoring the differences in publication lag, the authors concluded that data on actual activity (especially industrial production) are the most important source of information. However, when their less timely publication is taken into account, data on actual activities become much less relevant, and survey data take their place. In very late GDP forecasts, made in the last month of the corresponding quarter and thereafter, the industrial production data still contain significant information (Bańbura & Rünstler, 2011).

The paper produced by the Fed of Dallas is also interesting – when assessing the model, they do not take into account the flash estimate or final data on economic indicators, but rather their combination, as well as the difference created in the revision of these data. This is illustrated by the model using monthly industrial production, employment and retail sales (Koenig et al., 2003).

In the papers by Stock & Watson (1988) and Koenig et al. (2003), in addition to the industrial production indicator, the retail trade indicator was also used for the purpose of assessing the overall economic activity.

The paper of the US Treasury is also interesting – one of the most important economic indicators is the dynamics of the retail sales index. Retail sales data are released mid-month, and the retail sales component goes directly into the GDP calculation for that quarter. If only one month of data is available (for forecasts at the beginning of the reference quarter), it is often assumed that the data for the remaining two months of the quarter will be unchanged or the trend observed will continue (Kitchen & Monaco, 2003).

Also, retail trade proved to be a statistically significant indicator for euro area GDP nowcasts using the Kalman filter in bridge models (Angelini et al., 2011), as well as for estimates of Canada's monthly GDP (Mourougan, 2006).

Many papers have explored the connection between foreign trade and economic activity as a whole. A widely known paper was published by Balassa (1985), who states that the purpose of including exports (together with capital and labour force) in the production function is to test the hypothesis that export orientation increases total factor productivity through beneficial effects on more efficient resource allocation, capacity utilisation, economies of scale, and technological change. The results obtained show that exports significantly contributed to the economic growth rate, but also much improved the explanatory power of the equations.

The Bank of France used the bridge model for the purpose of estimating monthly GDP, which includes both export and import variables, bearing in mind that the condition for including the variable in the model is at least the monthly frequency and a publication lag of less than two months after the end of the reference period. Apart from official data concerning imports and exports, European Commission survey data were also used. When it comes to the link between official import and export indicators and survey expectation indicators, the link with exports is quite direct, as this variable refers to the activity of France's economic partners and, as such, represents a substitute for external demand directed to France. The connection with imports is less obvious, but it is clear that higher imports from France will stimulate the

activity of trading partners, with the economic cycles of European countries being quite aligned. When assessing the model, the authors concluded that foreign trade variables are extremely difficult to project due to their high volatility (Barhoumi et al., 2012).

For the purpose of Spain's GDP nowcasting and short-term forecast, a dynamic factor model was used, containing ten monthly indicators, eight of which relate to real activity, including goods imports and exports, as well as two survey-based data (Arencibia Pareja et al., 2020).

## 2.2 High-frequency indicators

The paper already touched upon the importance of timely data about the movement of economic activity for policy makers and economic agents, this being the most important input in making personal and business decisions. The purpose of a great body of research estimating GDP level before the official publication of the actual result is to find economic activity indicators which approximate this measure well, and with a smaller lag between the closing of the accounting period and data publication. These indicators gain particular importance in periods of departure from projected or long-term growth, during crises or sudden expansions. It was during the coronavirus-induced crisis that adequate business cycle indicators came to the fore as one of the main research issues. The potential candidates are the abovementioned Google trends and Google mobility index, as well as electricity consumption and production measures, air pollution measures, etc. Using some of the mentioned indicators, the models estimating real GDP growth at a given moment are assessed. In this paper, for the purpose of a short-term GDP projection, we used Google trends and electricity consumption indices, among other things.

### *2.2.1 Google trends indicator*

One of the indicators increasingly used in the past years as an explanatory variable in the forecast and assessment of economic activity is the Google Trends indicator. Google Trends (*trends.google.com*) allow us to access the search dynamics of a specific term or a category in a specific territory at any given time, and the database itself stores data from 2004 onwards. The index calculation method is specific, as the database values do not represent the scale of the searches, but are normalised and range from 1 to 100, where 100 represents the point with the highest interest recorded in the reference period. One of the biggest advantages of using Google Trends is that the data are available at any time during the calculation period for the past performance, that is, in real time.

In his paper, Wołoszko (2020) used over 200 categories in developing weekly GDP estimates for OECD countries. Owing to the nature of neural networks, which are good at modelling various non-linear relationships, there was no need to make any assumption about the type of correlation between GDP growth rates and Google Trends. This paper focused particularly on the fact that the emergence of high-frequency indicators (though they may not have been created with macroeconomic modelling in mind) enabled the use of deep learning techniques that require a large database for optimal functioning. The model from the abovementioned paper successfully signalled the effect that the outbreak of the pandemic

would have on economic activity of 45 out of 46 OECD countries and, overall, it had a lower quarterly GDP prediction error than the benchmark autoregression model. The results from the paper about the expected relationship between searches and GDP movements are in line with intuition (e.g. higher searches around *investment* category may signal higher GDP growth, while an increase in searches involving *crisis* and *recession* categories indicates lower growth).

Schmidt & Vosen (2009) were among the first to present in their paper a new indicator for private consumption as a component of GDP, based on Google Trends. Given that private consumption is the single most important component of GDP, timely and precise determination of its dynamic provides an insight into the current state of economic activity. Wu & Brynjolfsson (2015) used Google Trends as an explanatory variable in a model that predicts real estate prices, and claimed that economic predictions from this type of internet search based data can be applied to any market where internet search precedes a transaction, even when the transaction itself does not occur in virtual space. Choi & Varian (2012) state that the assertion made in their paper is not related to predicting future using Google Trends, but to estimating the present, thus entering the field of nowcasting. Google Trends have also been used in forecasting exchange rate trends (Markiewicz et al., 2018) and overcoming information gaps in developing countries (Narita & Yin, 2018).

A particularly interesting methodology of selection and processing of Google Trends data can be found in Wołoszko (2020), including the use of categories and not indices of searches created based on key terms, an approach applied in this paper as well. The paper also undertakes their transformation to address a few problems relating to the manner of data collection and presentation. The key problem with these data is that they show a relative significance of searches of a certain category relative to the total number of searches, which is then multiplied by a constant, so that at any time the index for the period when the search for that term was the most significant is 100, i.e.:

$$SVI_{ct} = \frac{SV_{ct}}{SVT_t} * C_c, \tag{1}$$

where *SVIct* is a relative share of searches by category *c* in period *t*. In this equation, only the denominator, i.e. the total number of searches, depends exclusively on the period *t*. The problem with this manner of index calculation is the fact that as the total number of searches increases, the relative share of relevant categories inevitably goes down, thus introducing bias in modelling. By transformation (1), we get:

$$svi_{ct} = log(SVI_{ct}) = sv_{ct} - svt_t + c_c. \tag{2}$$

Keeping in mind that *SVTt* remains unchanged for all categories in period *t*, we may isolate it by using the principal component analysis over the logarithmic series of *SVI*, which isolate the trend by means of the *HP* filter. The resulting first component is then recalculated to have the same mean and standard deviation as the average of the logarithmized *SVIs* and subtracted from the logarithmized *SVIs*. The transformation was carried out by using the y-o-y growth rates to overcome the seasonality problem, and indicators were taken from the source in the form of monthly indices.

In this paper, categories used as potential regressors are *business and industry, finance, real estate and travel.*

### *2.2.2 Electricity consumption*

Apart from Google Trends, one of the high-frequency indicators proposed as a potential regressor is the variable pertaining to electricity consumption. To collect data on electricity consumption in the territory of Serbia we used the ENTSO-E Transparency Platform (European Network of Transmission System Operators for Electricity), a joint venture of 39 official European transmission system operators (TSO), including Serbia's Elektromreža Srbije. TSOs are agents working independently from other participants in the electricity market, responsible for electricity transmission through the main high voltage electrical grids. TSOs enable network access to electricity market participants and guarantee safe system operation and maintenance. Among other things, through membership in ENTSO-E, TSOs are obliged to submit certain data which are made publicly available on the website. Among these publicly available data one can also find the data about the actual total load in the area of TSO operation within one hour. These data must be submitted no later than one hour after the end of the accounting period, which makes them timely. The data may be downloaded directly from the platform or using *API*, as a more efficient option for download and processing of large databases such as this one. In this paper data were downloaded by software *R* and converted from hourly to monthly amounts.

A question arises as to the relevance of including this piece of data in models for the projection of GDP growth rate, and/or justifiability of using electricity consumption as an indicator of economic activity. Not many papers have been published on this topic and one of them looked into the relationship between industrial output and electricity consumption, among other things. Using Johansen's cointegration methodology, Sun & Anwar (2015) concluded that there is a statistically significant long-term relationship when observing electricity consumption, industrial production and entrepreneurship, on the example of Singapore.

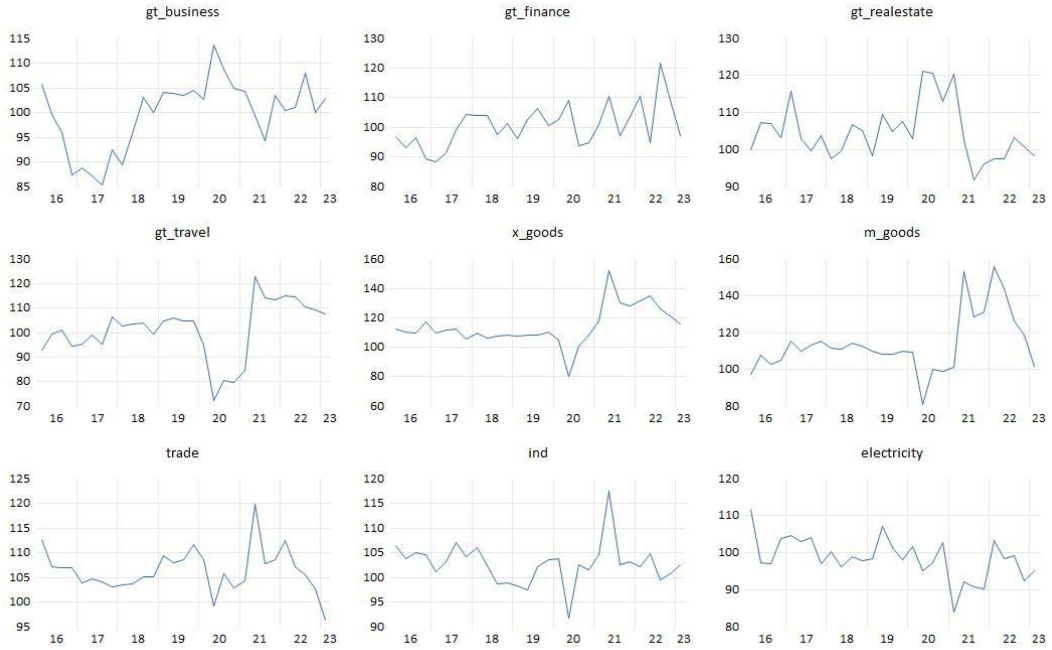
In empirical terms, the Serbian industry consumes over 30% of total available electricity (source: energy balances for 2021). It may for example be assumed that volatility in energy consumption can be attributed to shocks affecting industrial production, because the household sector is made up of a large number of small units whose consumption should be relatively stable. However, given that this paper deals with the projection of overall economic activity, there is no need to disaggregate total energy consumption, but only determine whether it represents a statistically significant explanatory variable in nowcasting and short-term GDP projections.

## 2.3 Transformation of used variables and checking of stationarity and multicollinearity

The data used in this paper were transformed in indices describing y-o-y dynamics, when it comes to quarterly and monthly data. As stated above, hourly data on electricity consumption

were aggregated into monthly amounts and then transformed to y-o-y indices. Plots of series indicate potential stationarity around a constant.

Figure 1 **Graphical overview of independent variables**




A formal application of the ADF test without added lags, which tests the stationarity around a non-zero constant, with the significance level of 5%, indicates that the series are indeed stationary around a constant.

Further, it needs to be checked whether there is a problem of multicollinearity within the database. A certain degree of correlation between the indicators is certainly expected, considering that there is a great number of factors affecting the dynamics of all the mentioned indicators, mainly from the external environment (prices of products established in the global market, such as cereals and oil, the degree of external demand, imported inflation, etc.), but also from the domestic one (such as inflation rate). Based on the multicollinearity analysis, we can conclude there is a high degree of correlation between exports and imports of goods, which is somewhat expected, as well as between these variables and the Google Trend index concerning travel. Attention should be paid to these variables and their interaction when constructing the model.

## 3 Econometric model used in nowcasting

The approach of reducing a large number of variables or time series to just a few factors was first mentioned in 1946 (Burns & Mitchell, 1946), and concerns the business cycles analysis. Namely, even at that time it was defined that while there is a great number of indicators which can move procyclically or countercyclically relative to the phase the economy is in, there is still a certain general tendency toward recession or expansion, not necessarily

noticeable in the same way as the number of the unemployed or the share of non-collectible loans. Further, these indicators of overall economic activity may be simultaneous (indicating the current state of the economy), leading (indicating the direction of future movement) or lagging (confirming tendencies that are underway or have occurred). As for the current projection of any variable, a successful forecasting requires us to determine a set of explanatory variables that best describe the movement of the forecasted variable. Ideal variables for these models would be those that are simultaneous and leading, and that predict current or future movements with great precision. Simultaneous indicators are particularly important if the lag period in publishing indicator value is shorter than in the variable that is being modelled.

The selection of potential regressors from the previous chapter is based on the principle of simultaneous indicators with the shorter lag period and leading indicators which indicate future changes in economic activity. Papers using the retail sale index have labelled it as a leading indicator, in the sense that it indicates emerging changes in the economy. We can add Google indices and electricity consumption, apart from the retail index, as leading indicators, while industrial production and export and import of goods are simultaneous indicators which have a relatively short lag period relative to GDP data.

## 3.1 Literature review

One of the most common models used for GDP nowcasting is a MIDAS model. This approach to data modelling was introduced for the first time in Ghysels et al. (2002), which presented the advantages of this approach for mixed data frequencies. According to authors, the typical regression models up to that time assumed the use of time series collected in regular and uniform time intervals. However, it is often the case that relevant information is contained in high-frequency indicators, while series that need to be modelled are of lower frequency, the typical example being macroeconomic data and the potential for modelling quarterly GDP data based on monthly indicators.

In their paper, Armesto et al. (2010) raised the question of the most efficient approach to mixed data modelling. They concluded that in models relying on a large number of explanatory variables, due to the problem of a large number of potential parameters, more parsimonious models may perform better, that is, aggregation models and MIDAS models, while the advantage of MIDAS models is that they also allow projections within the period.

Ghysels et al. also looked into the justifiability of incorporating financial data (which have a much higher frequency than macroeconomic variables) into macroeconomic projections. The premise is that a great number of daily financial time series contain information relevant for the further movement of economic activity that is lost through aggregation. The models used in the paper show that this hypothesis is accurate, i.e. that by using variations of MIDAS models with regressors from the financial sector, we arrive at a more reliable projection.

Frale & Monteforte (2010) combined two approaches in their work and proposed a mixed-frequency factor model, where high-frequency indicators are treated with a mixed data sampling regression model approach (FaMIDAS). They concluded that such models provide good assessments of monthly GDP and short-term quarterly projections.

## 3.2 Theoretical framework of MIDAS models

The theoretical framework was adopted from the papers by Ghysels et al. (2004), Ghysels et al. (2016) and Sinko (2008).

MIDAS models cannot be autoregression models due to the nature of AR models requiring equality of frequencies, but are most similar to distributed lag models.

A simple MIDAS model has the following structure:

$$Y_t = \beta_0 + B\left(L^{1/m}, \Theta\right)X_t^{(m)} + \varepsilon_t^{(m)}, \tag{3}$$

where $m$ indicates how much the frequency of the explanatory variable is higher than the dependent variable, so in the case of modelling annual data with quarterly explanatory variables, m would be 4. Further,

$$B\left(L^{1/m}, \Theta\right) = \Sigma_{k=0}^{K} B(k, \Theta)L^{k/m} \tag{4}$$

$$L^{1/m}X_t^{(m)} = X_{t-1/m}^{(m)} \tag{5}$$

mathematically, there would be equality between $L^{k/m}$ coefficient multiplied with $X_t^{(m)}$ and the value of $X_t^{(m)}$with *k/m* lags. In the example with annual and quarterly frequencies, this means that this year's value $Y_t$ is projected to the database with quarterly data on the value $X_t^{(m)}$ up to the $K^{th}$ quarter back. Theoretically, the number of parameters for assessment from the polynomial $B\left(L^{1/m}\right)$ is final for the purpose of simplification, keeping in mind that one $b_k$ needs to be assessed for each period, and in case of high-frequency data (daily or hourly), that is an extremely large set of parameters to be assessed. Empirically, this problem was initially treated by aggregating high-frequency data so that their frequency would be of the same order as in a low-frequency, dependent variable, in the absence of models which allow for different frequencies. The aggregation approach would lead to neglecting some potentially important information through series transformation. In order to overcome this problem, vector $\Theta$ is introduced in MIDAS models in the function from which the model parameters are assessed. There are several potential specifications $B\left(L^{1/m}, \Theta\right)$ which may be used for the assessment of model parameters. Besides, the use of parametric function $B\left(L^{1/m}, \Theta\right)$ also enables a more efficient selection of lags that need to be included in the model. The selection of an adequate form of vector $\Theta$ actually makes it possible for the selection of the number of lags to be imposed by the structure of data themselves (Sinko, 2008).

In the matrix representation, an example of a regression model of data of different frequencies in which the dependent variable $y_t$ is quarterly and whose dynamics is modelled using the lags of the dependent variable itself and the monthly variable $x_\tau$ and its lags, has the following form:

$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} \alpha_1 + \begin{bmatrix} x_6 & \cdots & x_1 \\ \vdots & \vdots & \vdots \\ x_{3n} & \cdots & x_{3n-5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{6}$$

Thus, in every quarter $t$, considering that $m = 3$, the dependent variable $y_t$ is a linear combination of variables $x_{3t}, x_{3t-1}, x_{3t-2}$ from the current quarter $m$, and $y_{t-1}$ and $x_{3(t-1)}, x_{3(t-1)-1}, x_{3(t-1)-2}$ from the previous quarter $t-1$.

In this way, a kind of frequency matching is performed, since the high-frequency variable $x_\tau$ is transformed into a lower-frequency vector $(x_{3t}, \ldots, x_{3t-5})^T$. It should be noted that for this type of transformation, the number of observations in variable $x_\tau$ should be exactly $3n$. In cases that are common in practice and which concern the so-called missing edges, i.e. lags and different times of publication of official data, it is necessary to either use balanced official data or estimate the missing values by using a model or applying the Kalman filter.

On the other hand, if we wanted to add to the model (6) another explanatory, high-frequency variable *z*, which is published on a weekly level, the model would need to be expanded. The use of weekly (as well as daily and business day data) reveals another limitation of multi-frequency regression models, i.e. the fact that these models do not allow variations in the number of observations within one calculation period *m* that refers to the dependent variable. This means that each month must strictly have four weeks, i.e. in this illustration each quarter must have 12 weeks. In the case of variable $z_\tau$, for the purpose of frequency matching, the value of *m* equals 12. Apart from the abovementioned variables and lags, we are adding to the model observations $z_{12t}, z_{12t-1}, \ldots, z_{12t-11}$ from the current quarter $m$ and $z_{12(t-1)}, z_{12(t-1)-1}, \ldots, z_{12(t-1)-11}$ from the previous quarter $t-1$. In matrix form:

$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} \alpha_1 + \begin{bmatrix} x_6 & \cdots & x_1 \\ \vdots & \vdots & \vdots \\ x_{3n} & \cdots & x_{3n-5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_5 \end{bmatrix} + \begin{bmatrix} z_{24} & \cdots & z_1 \\ \vdots & \vdots & \vdots \\ z_{12n} & \cdots & z_{12n-23} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \vdots \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \quad (7)$$

The frequencies of some high-frequency variable $x_\tau$ are generally matched through transformation into a low-frequency vector $(x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l}^{(i)})^T$, whereby we obtain a matrix representation of the general model in the following form:

$$\begin{bmatrix} y_l \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_{l-1} & \cdots & y_{l-p} \\ \vdots & \vdots & \vdots \\ y_{n-1} & \cdots & y_{n-p} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} + \sum_{i=0}^{k} X^{(i)} \begin{bmatrix} \beta_0^{(i)} \\ \vdots \\ \beta_l^{(i)} \end{bmatrix} + \begin{bmatrix} \epsilon_l \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (8)$$

$$X^{(i)} := \begin{bmatrix} x_{um_i}^{(i)} & x_{um_i-1}^{(i)} & \cdots & x_{um_i-l}^{(i)} \\ x_{(u+1)m_i}^{(i)} & x_{(u+1)m_i-1}^{(i)} & \cdots & x_{(u+1)m_i-l}^{(i)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{tm_i}^{(i)} & x_{tm_i-1}^{(i)} & \cdots & x_{tm_i-l}^{(i)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{(n-1)m_i}^{(i)} & x_{(n-1)m_i-1}^{(i)} & \cdots & x_{(n-1)m_i-l}^{(i)} \\ x_{nm_i}^{(i)} & x_{nm_i-1}^{(i)} & \cdots & x_{nm_i-l}^{(i)} \end{bmatrix}, \quad (9)$$

where $n$ is the number of observations of the dependent variable, $m$ is frequency, $p$ is the number of lags of the dependent variable included in the model and $u$ is the smallest integer to which $um_i - l > 0 \; and \; u > p$ applies.

When defining the MIDAS models, the notion of parametric functions was introduced. The selection of an adequate parametric constraint function leads to more desirable outcomes when it comes to model performances. The parametric constraint function and frequency matching are interdependent concepts which help to properly define the relationship between high-frequency and low-frequency variables. The goal is to approximate, as best as possible, the relationship between these variables despite their different frequencies.

The parametric constraint function plays the key role in frequency matching. It determines a functional form of the relationship between high-frequency and low-frequency variables. By imposing appropriate limitations, the function ensures that high-frequency information is appropriately mapped in the low-frequency domain.

The parametric constraint function provides a framework for defining how a low-frequency variable depends on the value or lags of the high-frequency variable. The function establishes a connection between two different frequency domains and allows for a meaningful interpretation and analysis of the relationship between variables observed at different frequencies.

There are several potential parametric constraint functions which may be used in modelling, including, inter alia, the exponential Almon polynomial, beta function (analogous to the probability function), Gompertz function, logarithmic Cauchy function. For the purposes of GDP forecast, considered in the remainder of this paper, Almon's exponential lag function is used (Almon, 1965).

This parametric function rests on the Weierstrass theorem. According to the Weierstrass theorem about extreme value (known as the extreme value theorem), used in mathematical analysis, if function *f(x)* is continuous on the closed and bounded interval [*a*, *b*], then *f(x)* must attain a maximum and a minimum on that interval, each at least once. Further, that function may be approximated by a polynomial *p(x)* of some order *P*. Take the following model for example:

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_n x_{t-n} + u_t, \quad t = 1, 2, \ldots, T. \tag{10}$$

By applying the Weierstrass theorem, Shirley Almon concluded that the parameters $\beta_0, \beta_1, \ldots, \beta_n$ can be approximated by the following polynomial of order *P*:

$$\beta_i = a_0 + a_1 i + a_2 i^2 + \cdots + a_p i^P. \tag{11}$$

That is, by replacing (11) in (10),

$$y_t = a_0 x_t + (a_0 + a_1 + a_2 + \cdots + a_p) x_{t-1} + (a_0 + 2a_1 + 4a_2 + \cdots + 2^P a_p) x_{t-2} + \ldots + (a_0 + n a_1 + n^2 a_2 + \cdots + n^P a_p) x_{t-n} + u_t \tag{12}$$

where it applies:

$$
\begin{aligned}
y_t &= a_0 z_{0t} + a_1 z_{1t} + a_2 z_{2t} + \cdots + a_P z_{Pt} + u_t \\
z_{0t} &= (x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-n}) \\
&\vdots \\
z_{pt} &= \left(x_{t-1} + 2^P x_{t-2} + \cdots + n^P x_{t-n}\right)
\end{aligned}
\quad . \tag{13}
$$

The use of Shirley Almon's polynomial when assessing the MIDAS model reduces the initial model with a large number of high-frequency parameters and their lags to a model with the smaller number of parameters for assessment.

### 3.3 Assessment of the MIDAS model

Standard statistical tests for establishing model validity and adequacy are used for assessing MIDAS models. It has already been determined that both the dependant variable (y-o-y GDP growth index) and all potential regressors (goods export and import, retail trade and industrial production, as well as indices of Google trends and electricity consumption) are stationary and that there is no significant problem of multicollinearity. Still, due to the relatively small sample (29 quarterly data and 85 monthly data for each potential regressor), variables adequate for the model needed to be selected. Also, the model was assessed based on the period Q1 2016 – Q4 2021, with the last six quarters being used for checking the validity of the forecast. Initially, the included variables were the Statistical Office's official indicators, as well as the electricity consumption index. The iterative process, monitoring the Akaike Criterion, showed that the best complement to official data is Google's index that pertains to business topics (gt_business), as well as the index for the topic of finance (gt_finance), in addition to the already present electricity consumption index.

A specificity of MIDAS models (as indeed of all models for GDP nowcasting) is that the primary indicator of the model's successfulness is the root mean forecast error (RMFE), even at the cost of including regressors that are not statistically significant by their p-value. Also, by monitoring the RMFE and the Akaike Criterion, we needed to decide whether a larger number of variables yields a greater contribution than a larger number of lags. As presented in the theoretical overview, the number of lags included in the model does not equal the number of model parameters as the impact of individual variables and their lags is assessed using a parameter function, in this case the Almon polynomial function. Given that we used monthly data, and that the quarterly GDP datum is being modelled, we chose the third order polynomial. In addition, the very procedure has the option of automatic selection of significant lags given that the model specification does not follow an already published model and is therefore not limited by the previously selected number of lags; rather, only the lags that turn out to be statistically significant in the given sample are included in the model. The initial model included all of the given potential regressors, and the principle of monitoring the RMFE together with the statistical significance of the given regressor iteratively excluded regressors, until the final model specification was selected. In regard to this, the model that performed the best on the base of proposed indicators is the following:

Figure 2 **The assessed MIDAS model**

Dependent Variable: GDP
Method: MIDAS
Date: 06/27/23 Time: 16:39
Sample (adjusted): 2016Q3 2021Q4
Included observations: 22 after adjustments
Method: PDL/Almon (polynomial degree: 3)
Automatic lag selection, max lags: 6
Chosen selection: 4 5 4

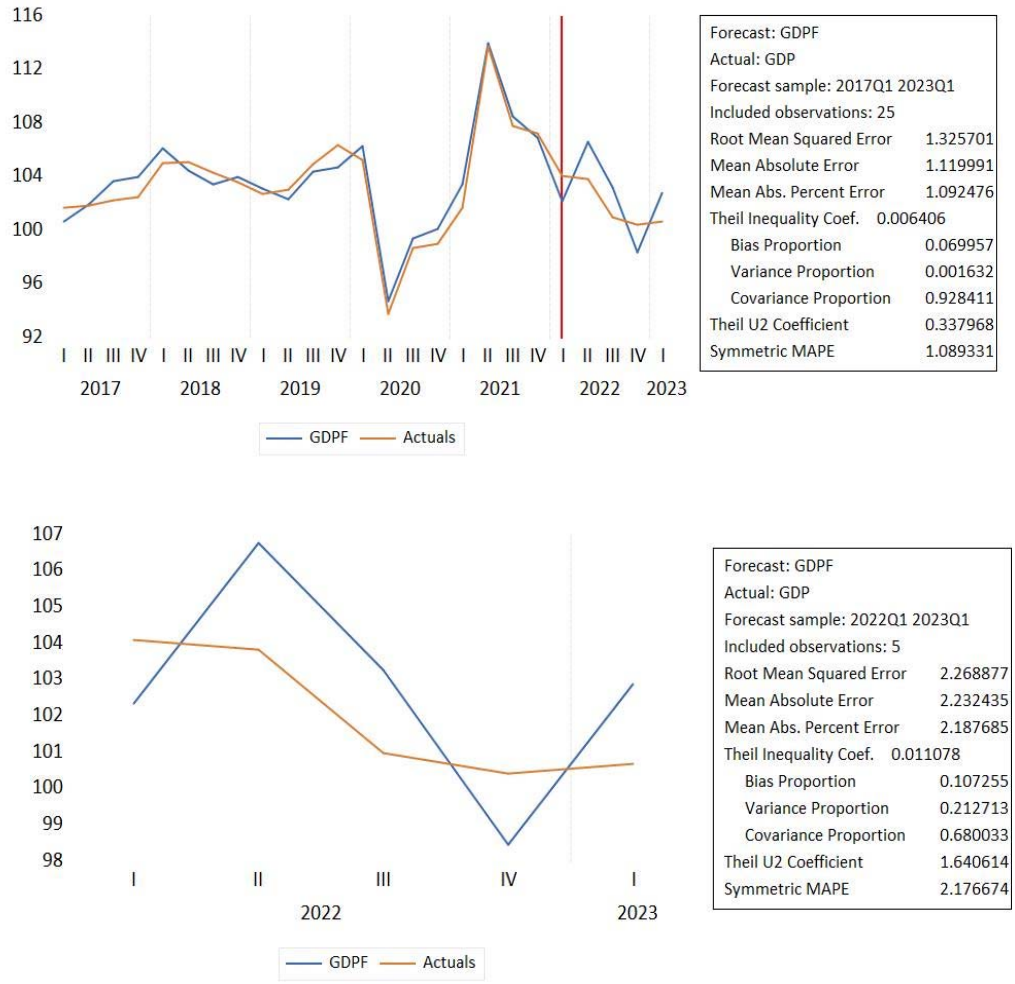
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| GDP(-1) | 0.701371 | 0.080774 | 8.683141 | 0.0000 |
| Page: GT_BUSINESS Series: GT_BUSINESS Lags: 4 | | | | |
| PDL01 | 0.127017 | 0.224591 | 0.565549 | 0.5821 |
| PDL02 | 0.075517 | 0.203254 | 0.371540 | 0.7167 |
| PDL03 | -0.041929 | 0.040506 | -1.035135 | 0.3210 |
| Page: IND Series: IND Lags: 5 | | | | |
| PDL01 | -0.700033 | 0.173006 | -4.046296 | 0.0016 |
| PDL02 | 0.845103 | 0.144796 | 5.836520 | 0.0001 |
| PDL03 | -0.161418 | 0.025421 | -6.349781 | 0.0000 |
| Page: ELECTRICITY Series: ELECTRICITY Lags: 4 | | | | |
| PDL01 | -0.242987 | 0.166338 | -1.460801 | 0.1698 |
| PDL02 | 0.301689 | 0.162097 | 1.861158 | 0.0874 |
| PDL03 | -0.068253 | 0.033511 | -2.036719 | 0.0644 |

| | | | |
|---|---|---|---|
| R-squared | 0.939454 | Mean dependent var | 103.5311 |
| Adjusted R-squared | 0.894044 | S.D. dependent var | 3.829498 |
| S.E. of regression | 1.246536 | Akaike info criterion | 3.581570 |
| Sum squared resid | 18.64624 | Schwarz criterion | 4.077498 |
| Log likelihood | -29.39727 | Hannan-Quinn criter. | 3.698396 |
| Durbin-Watson stat | 2.074969 | | |

The final model includes one dependent variable lag (GDP(-1)), Google's business trend index (GT_BUSINESS) with four selected lags, the industrial output index with five selected lags (IND), as well as the electricity consumption index (ELECTRICITY) with four selected lags. All of the explanatory variables are monthly frequencies, while the regressand is quarterly. Thus, a total of 13 parameters were assessed using nine Almon coefficients.

The selected model has the following performances in terms of both in-sample and out-of-sample error:

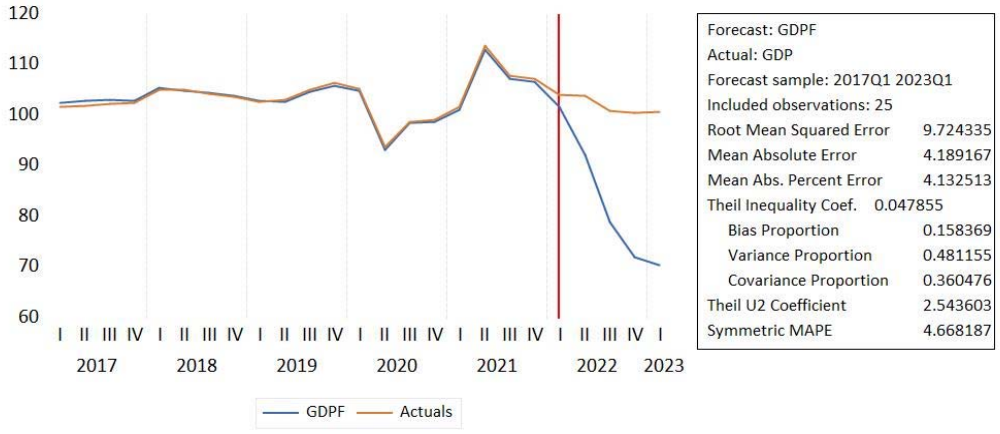
Figure 3 **Projections of the assessed MIDAS model and out-of-sample model projection**




Individually, the Google trends indicator used in the model is not statistically significant, however, its exclusion has shown to be unfavourable to the performances of out-of-sample forecast. The potential reasons for the lack of statistical significance are reflected in the fact that this is a relatively small sample for a model of this size, in the way in which data about Google trends are being collected (though this was partly neutralised by the use of growth rates), as well as in the fact that the correlation rate between movements of this index and the industrial forecast index, though insufficiently high on its own to be problematic, is not negligible (-0.4). Moreover, MIDAS models assessed on such a small sample generally imply the issue of overfitting, i.e. the tendency to "remember" data instead of abstracting the key relations, which is reflected in a significantly smaller error within the sample, with a poorer forecast and bigger out-of-sample forecast errors. Exclusion of the business trends index leads to a significantly poorer out-of-sample forecast and the bigger presence of the problem of overfitting given that the model's main goal is to achieve a forecast as accurate as possible, it is justified to include a variable that contributes to the reduction of the RMFE, though in itself

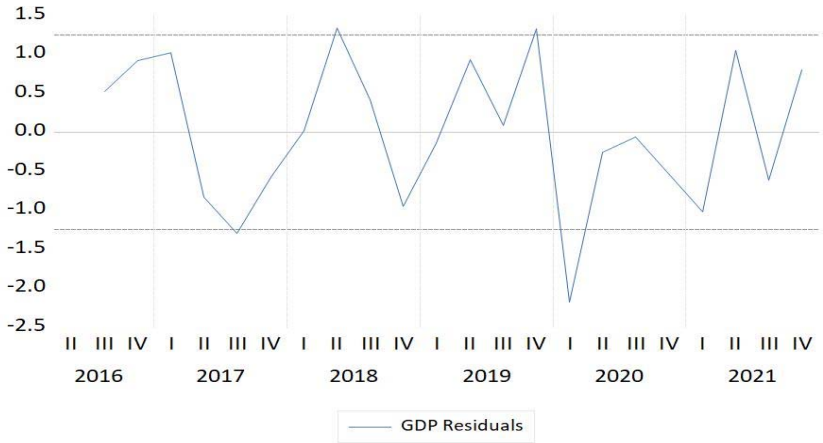
it is not necessarily significant as an explanation of the dynamics of the dependent variable. Besides, if we compare the selected model containing two variables that are not official monthly indicators of economic activity with the model containing only the indicators published by the statistical office, we can see that the inclusion of Google's trends and electricity consumption in the model led to a better situation as regards this issue, but did not reduce the forecast error. Below, we can see the performance of the MIDAS model which contains regressors such as goods export and import, retail trade and industrial production, and where the issue of overfitting is evident.

Figure 4 **Projections of an alternative MIDAS model**



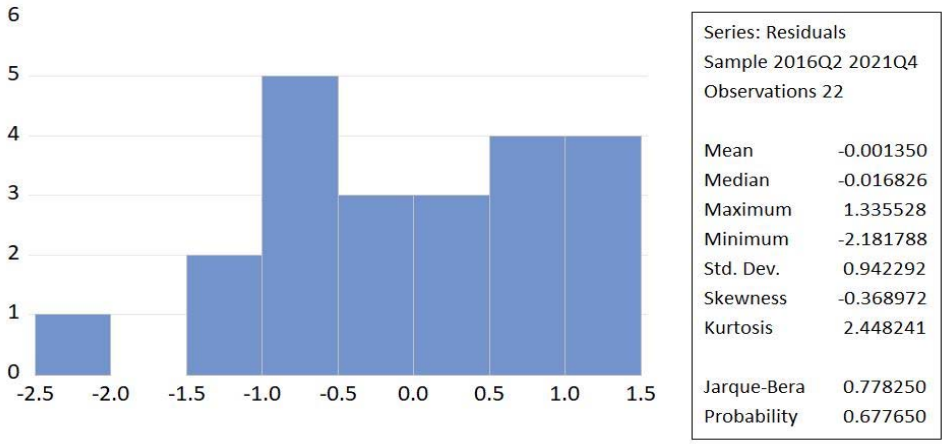

A further test of the model's adequacy is reflected in the testing of model residuals for the selected model, whose specification was previously shown in Figure 2 in this part.

Figure 5 **Residuals of the assessed MIDAS model**



Based on the chart, we can assume that the residuals are normally distributed, which can be verified using the Jarque-Bera normality test. The test results for the chosen MIDAS model are given below.

Figure 6 **Jarque-Bera normality test of residuals for the assessed MIDAS model**




The p-value for the Jarque-Bera normality test indicates the non-rejection of the zero hypothesis of the normality of residuals.

## 4 Machine learning models in econometrics

A specificity of the field of machine learning is partly reflected in the fact that there is no concrete overview of the history and development of this discipline (Plasek, 2016). The author stated an interesting assertion regarding the nature of the discipline, namely the fact that machine learning systems usually have the status of an experiment because they are used both as a confirmation and a counterweight to what we know. Still, the beginning of this discipline is usually associated with psychologist Frank Rosenblatt of the Cornell University, who invented a letter-recognition machine based on the ideas about the way the nervous system works. Called the Perceptron, the machine is considered the prototype of modern neural networks (Fradkov, 2020).

According to Murphy (2012), the goal of machine learning is the automatisation of the pattern-recognition process in data so that the patterns are relatively good at predicting future movements of the relevant variables. The need for this type of modelling arose with the growing amount of available data, i.e. it goes hand in hand with the big data concept, which is believed to be the first turning point in machine learning. The second turning point is associated with the costs of processing a large amount of data in increasingly complex systems, without adequate processors. Google and HeadsUp found a solution to this. The third turning point is the booming growth of deep learning.

Regardless of the lack of systematisation in this field, which is due to its explosive growth, Alzubi et al. (2018) underlined several key events that defined the development of this

discipline, some of them being the formulation of the Turing test, presentation of the artificial intelligence concept and then of the algorithm for pattern presentation, triumph of the IBM's computer of Garry Kasparov in chess and the creation of neural networks able to recognise patterns and faces.

As the complexity of econometric models increases, which is a natural consequence of the growing complexity of interactions between various economic entities due to the emergence of the internet, cashless and online trade and similar features of the time we live in, including the increasing availability of databases, a need arose to improve existing models and make them more efficient. One way to achieve this is to introduce machine learning in econometrics, which we can define as the ability of the system to learn from data in a way that will ensure the automatisation of the model construction process and more efficiently resolve the problems that might occur (Janiesch et al., 2021).

Machine learning implies the application of algorithms that learn in iterations from problem-specific data, thus enabling an insight into the relations and patterns for which they are not explicitly programmed. There are several types of machine learning: supervised learning, unsupervised learning and reinforcement learning (Sarker, 2021). Supervised learning implies providing inputs and outputs so that the algorithm can identify the existing relations and create an adequate model describing those relations. Unsupervised learning requires the algorithm to find structure in data sets that are not marked in any way. Reinforcement learning involves a dynamic environment where the algorithm receives a feedback reaction from the system after performing an action, and the goal is to maximise the positive feedback reaction. Deep learning, which is in the focus of this paper, lies within machine learning.

Generally, we need three things to create a machine learning model:

- input, i.e. data to be analysed (e.g. if we are analysing a sentiment, we need textual data to assess the sentiment present);
- examples of expected output, i.e. the results we expect to get from data (textual data based on which we analyse the sentiment must have associated sentiment assessments);
- the manner of measuring the successfulness of the algorithm, i.e. the difference between the assessed and actual value based on which the manner of algorithm functioning will be adapted (this is actually the "learning" part of machine learning).

In fact, all these elements are also present in classic modelling, but this does not necessarily mean that a machine learning method can be applied to any econometrics problem. When an adequate model is applied, we certainly get satisfactory results when it comes to modelling in-sample data, as the model will "learn" to recognise the patterns present. However, it is questionable whether the model will be efficient in predicting values outside of the sample on which it was trained. Namely, machine learning models (and deep learning in particular) are highly prone to overfitting to data (Chollet et al., 2022), which we have already covered when assessing the MIDAS model.

When making a good machine learning model, a sufficiently large database is very helpful. In their paper, Zhang & Ling (2018) called machine learning a data-driven approach and emphasized that the key ingredient behind the success of these models is the recognition of

patterns based on a sufficiently large number of learning inputs. Neural networks go beyond classical machine learning methods only when databases exceed a certain amount of data, and the model's performances begin to improve at a higher rate compared to classical algorithms.

Until a few years ago, the only data relevant for econometric analyses could be obtained from official statistical offices and relevant institutions. Today there is a multitude of high-frequency, relevant and available data that can be used to explain movements and predict economic variables, and machine learning allows us to assess the model that will most efficiently process such amount of input data.

Varian (2014) stated that the purpose of econometric analysis is in one of these four categories: prediction, summarising, assessment and hypothesis testing. Machine learning is primarily about predictions, therefore this paper assesses the quality of the machine learning model through its successfulness in predicting future movements of relevant variables.

From the very start, machine learning models were shown to be extremely useful in classification tasks. Some of the most famous machine learning algorithms are the following:

- Naïve Bayes, a classification algorithm based on the Bayes theorem, whose assumption is the mutual independence of characteristics of input data (hence the "naïve").

- Kernel methods, which include the support vector machines, a method that gained its current form in a paper from 1995 (Cortes & Vapnik, 1995), where it was defined as a new methodology for classification of two-class problems. The idea is that input vectors are mapped in a non-linear manner into a multidimensional space with characteristics, where the output is again a linear decision on input classification. The high level of generalisation with this approach is given special significance, and its essence is in finding optimal limits for decision-making as to the attribution to a specific class. In a multidimensional space, the question of setting optimal decision-making limits is not associated with determining the coordinates, but rather distances between the dots that represent inputs with the help of the kernel method. These models are often considered shallow representations of neural networks.

- Decision trees and tree-like methods, such as random forest and gradient boosting. The random forest algorithm involves many decision trees based on which the final output is created. Gradient boosting algorithms function in a similar manner and use the amplification of gradients to improve performance in dots where the previous model iterations turned out to be the weakest.

This paper deals specifically with the subsegment of machine learning that pertains to deep learning, more precisely neural networks, which has been increasingly present in time series modelling over the past decade. Deep learning is an approach to learning patterns from data which emphasizes the learning of sequential layers with growing efficiency. "Deep" in deep learning does not pertain to a better or deeper understanding that is reached by this approach, but rather it refers to the idea of the existence of sequential layers in the model. The depth of the model is defined by the number of layers contributing to the final outcome. Other appropriate names for this area could be *layered learning* and *hierarchical learning* (Chollet & Allaire, 2018). In theory, there is no upper limit for the number of layers through which learning can take place, but the main difference relative to other methods of machine learning

is that classic machine learning implies one or maximum two layers of processed data, which is why such methods are sometimes called shallow learning.

## 4.1 Neural networks

Initially, the rise of neural networks unfolded as this method was improved for the purpose of better classification of images. The abrupt surge in usage happened when the networks turned out to be extremely useful for a series of problems, notably because in the process of creating an adequate model, they automated one of the key steps – the so-called feature engineering. Before neural networks, feature engineering implied processing data in a manner that would provide best performance outputs depending on the model used. Neural networks process input data alone, in the first layer, after which data are transformed in all subsequent layers until the last, output layer.

The explanation as to how neural networks function has been borrowed from Chollet et al. (2022). The basis of the neural network is the block unit called the perceptron, and several perceptrons placed in layers create a neural network. The perceptron is composed of four units:

- input layer,
- weights (values with which input data are weighted),
- weighted sum,
- activation function (which may or may not be activated upon the receipt of the weighted sum).

At the end of the process, an output layer is generated, whose dimensions are defined by the type of problem being resolved – if it is a regression or a binary classification, there will be only one output layer. In case of multifold classification, where the probability of belonging to a specific class is being modelled, the output layer will have as many nodes as there are classes. The type of the problem being modelled will determine the type of the activation function to be used. Some of the possible activation functions are listed below:

- Rectified Linear Unit (ReLu) – negative values are output as zero, and positive values as the maximum in the set;
- Leaky Rectified Linear Unit (Leaky ReLu) – accepts negative values, but with low weights which depend on the $a$ coefficient;
- The Tanh activation function is non-linear, but has a narrow scope, hence it more often rejects important inputs and has activation values in the range [–1, 1];
- The sigmoid function is similar to the previous one, with the exception of activation values being in the [0, 1] range, therefore it is particularly suitable for modelling the probability problem;
- Softmax – activation values are added to 1, therefore it is suitable for multifold classification.

The minimum number of layers in a neural network is two, given that the input layer does not count, therefore the simplest, shallow neural network would be made up of an output layer and one hidden layer.

The network takes data from the input layer in batches, and the number of nodes is equal to the number of explanatory variables, which must be numerical. Several types of layers can be found in the hidden part. With dense layers, each node in the input layer is connected to each node in the output layer (fully connected layers). Convolutional layers are made up of filters which process a portion of input nodes, but always in such a manner than each input node is connected to at least one node in the hidden part of the network. This paper applied the model of the recurrent neural network (nodes can take their previous output values as input ones), more specifically the LSTM network (long short-term memory), which are the most suitable for working with time series, as we will present in the remainder of the chapter.

Each network must have a defined loss function, which is used to assess the quality of the model's prediction (successfulness). Depending on the number of output variables, we can have several loss functions in the model (one for each variable). The already known mean squared error (MSE) and mean absolute error (MAE) are applied in neural networks which are used for regression problems, while the binary cross entropy and hinge loss (a protocol similar to the support vector machines) are used for the problems of binary classification.

With the loss function, the optimisation protocol rounds up one neural network. With the help of the optimiser, weights are updated in each iteration so that the value of the loss function when generating a new output layer decreases. The most frequently used optimiser is the gradient descent. In fact, this is a simple principle of looking for the global minimum of the loss function by iteratively changing weights based on the previously realised output layers. As with all problems of this type, there is the issue of stopping at a local minimum, which is solved by the convex nature of the loss function and the learning rate. A high learning rate will move quickly down the convex loss function, and can thus "miss" the global minimum. A too low learning rate requires a large number of iterations to find the minimum, which means it is time-consuming. Finding the optimum learning rate has resulted in the improvement of the gradient descent optimiser and the creation of the ADAM optimiser (Adaptive Moment Estimation), which takes into account the previous values when calculating a new gradient.

Goulet Coulombe et al. (2022) compared different kinds of machine learning models based on data from the FRED base concerning industrial production, unemployment rates, inflation, the difference between the ten-year maturity rate of Treasury bills and the key policy rate, as well as real estate prices. They drew several conclusions, two of which are particularly relevant for this paper, namely the fact that the non-linearity of machine learning models is especially useful in situations of macroeconomic uncertainty, tight financial conditions and when the real estate bubble bursts, and the other conclusion concerns the amount of data in the model. The first conclusion is applicable in case of data-rich models, while smaller models have better performances if non-linear machine learning models are combined with the classic factor analysis.

Hopp (2022) said that of all machine learning models, the ones with neural networks recorded the largest growth. Still, of all areas in which neural network models are applied, the slowest progress has probably been recorded in current econometric projections, which is most

likely the consequence of the same problems with which traditional econometric models are faced, such as multicollinearity, missing data, mixed frequencies and lags in data publication. In addition, a large amount of data and the addition of explanatory variables to models that are already experiencing some of these issues can further exacerbate the situation. Hopp concluded that the LSTM model has better performances that the DFM model because it does not rely on the invertibility of the matrix, and can thus process any data set or any combination of frequencies. One of the key advantages of using the LSTM model is the fact that it does not rely on the invertibility of the matrix, therefore its input can also be the singular matrices or non-squared matrices, which is often the case when it comes to real data. Also, as the given models use large databases, the invertibility of the matrix is often a very demanding condition technically and timewise, which is not necessarily possible to meet. In contrast, the shortcomings compared to the DFM model include the stochastic nature of the LSTM model, the lack of interpretability in coefficients and the fact that this is a "black box" of contributions of individual explanatory variables.

This shortcoming is referred to by Medeiros et al. (2005), placing neural networks in the group of models that leave a functional form unspecified. Though these models contain parameters (and a large number of those), they are not globally identified or assessed, and therefore cannot be interpreted.

Elsayed et al. (2021) asked whether we need deep learning models at all for projecting time series, and concluded that simpler models of classic machine learning are sometimes quite sufficient for modelling dynamics of univariant and multivariant time series.

### *4.1.1 LSTM neural networks*

Siami-Namini et al. (2019) gave a brief overview of the LSTM model functioning. When the neural network is tasked with modelling interdependent data, i.e. data where the previous values affect the current ones (as is the case with time series data), it is desirable to use the models of recurrent neural networks (RNN). The RNNs differ from other neural networks by the presence of connections between layers that are not solely in one direction – from the input to the output layer. More specifically, the recurrent network "learns" by back propagation in the process of minimising loss function. Due to the issue of short memory of the previous state of inputs in classical RNN networks (the information is lost after only a few periods), the LSTM (long short-term memory) network model was developed. This model essentially functions the same as other RNN networks, with the addition of a special structure of the node itself, where there are several gates:

- input gate, that pertains to data entering an active cell;
- output gate, that processes data exiting an active cell;
- forget gate, which filters the relevant information received through the backward propagation method.

Though the RNNs have shown to be adequate for time series modelling, especially in the form of the LSTM model, a certain number of papers used other deep learning models for their current forecasts, and an overview of those can be found in Zheng et al. (2023). One of them

is the model presented by Loermann and Maas (2019) in their paper, for the needs of assessing US GDP levels in the current quarter and several quarters ahead using the ANN model. The Artificial Neural Network (ANN) models are the simplest neural network models containing one hidden layer in their structure. A special advantage of using the ANN lies in the possibilities for choosing the relevant variables and lags using the grid structure by way of iterative exclusion of variables and lags that did not turn out to be statistically relevant for forecast improvement. Given that the ANN models have a pronounced issue with the local minimum in the model error minimisation process (which, in this case, is the mean squared error), 100 random ANN models were assessed and then included in one kernel model. This approach is called the ensemble operator approach. The model was then compared to a dynamic factorial model, and the ANN model had a much lower forecasting error.

The success of different kinds of neural networks in modelling time series has induced a number of authors to create new deep learning models suitable for this type of data. Thus, the Temporal Fusion Transformer (TFT) model was created. The transformer-type deep learning models are characterised by the adoption of the self-attention concept, i.e. the possibility to access inputs along the entire sequence in any moment, as well as the possibility of weighting them based on the learned material, which is what sets them apart from the RNNs (Zeng et al. 2022). This is the model that should solve the biggest issue with the LSTM – the relatively short network memory (several periods back). Still, longer-term memory is not always necessary, especially if the goal is modelling only a few periods ahead or estimating current levels (nowcasting).

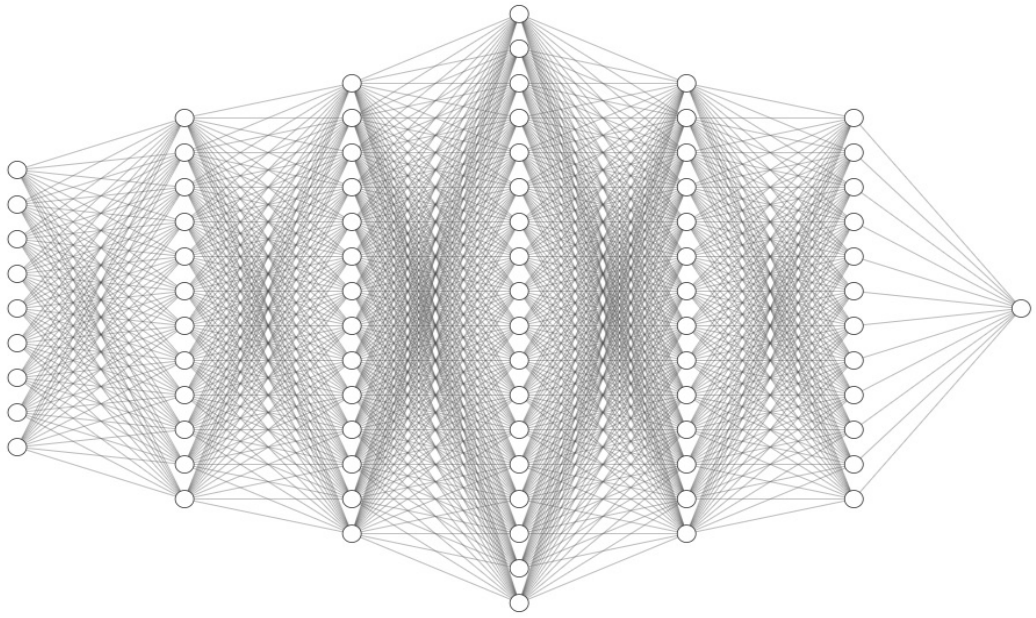
The above mentioned ensemble operator approach is a type of machine learning development relating to economic modelling by using a machine learning model to unify a larger number of already existing standard econometric models and/or machine learning models. The idea that a more reliable forecast can be obtained by averaging a larger number of models is not new in economics, but machine learning enables the weighting and selection of the most relevant models in a manner that minimises forecasting errors without *a priori* imposing the structure.

## 4.2 Assessing the LSTM model

The process of assessing LSTM models partly follows the methodology presented in Medeiros et al. (2005), though the model type is not the same. The approach boils down to the idea that the first assessed model should be as simple as possible, and then the performances of each ensuing, expanded model should be monitored. In this paper, the performances of each model iteration are compared using the mean squared error.

We assessed the LSTM RNN model with one input layer with nine regressors, and five hidden layers with 64, 128, 256, 128 and 64 nodes, and one node in the output layer. The structure of the assessed model's approximation is as follows.

Figure 8 **Approximation of the structure of the assessed LSTM model**



According to the already mentioned paper by Varian (2014), the problem of overfitting that often occurs with machine learning models is addressed by dividing the database into training data, validation data and assessment data. The model's efficiency is measured based on the reduction of the mean squared error, and the model was assessed in 1,000 epochs. The final value of the root mean squared error was 2.98, while the value for the validation set was 4.90.

The obtained error of out-of-sample forecast can be compared to the one that was obtained by assessing the MIDAS model. Namely, the exit from the LSTM model gives us information about three different mean squared errors – the first two are listed and pertain to the assessment part of the sample and the validation part of the sample, while an error obtained in the testing part of the sample in the LSTM model would correspond to out-of-sample error in the MIDAS model. The MIDAS model has an out-of-sample error of 2.27 (Figure 3 in part 3.3), while the comparable value with the LSTM model is 1.85.

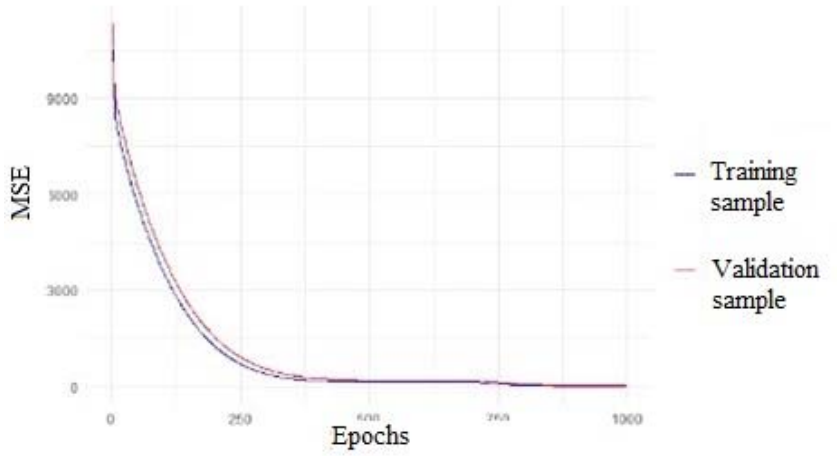
The fact that the LSTM model has a larger validation base error compared to the training base confirms that the issue of overfitting to data is still present, and in case of this model it is almost impossible to avoid. The reason for this is the relatively small database which is a necessary precondition for successfully setting a machine learning model, in particular a deep learning one. It is additionally worrying that the root mean squared testing-based forecasting error has an extremely low value compared to the training and validation bases, although it shows greater precision of the LSTM model compared to the MIDAS. The reasons for such inconsistency between models are several:

- Besides the already mentioned problem of base size, an additional issue is the fact that the validation set is much smaller than the training set, while the testing set contains only ten monthly regressor data for each series;

- Variance in the testing base is much smaller than the variance in the testing and validation bases;
- This type of problem occurs sometimes when the neural network is trained using the Keras/TensorFlow. When using the batch gradient descent, the value of error in each epoch is the average of all batches processed until that point. Hence, the high error value at the beginning shall affect the entire error through the average, regardless of how much it decreased during the model training. On the other hand, this problem is not present when testing model successfulness as in that case only the final error value is taken into account.

Below are model performances on the training and validation sets.

Figure 9 **Performances of the neural network model on the training and validation parts of the sample**




## 5 Conclusion

In their papers, Varian (2014) and Murphy (2012) stated that the basic measure of adequacy of a machine learning model is forecasting success. This is certainly not the only criterion when it comes to traditional econometric models, but it is the goal of the models used for nowcasting. The goal of assessing the MIDAS model and the neural network model is the assessment of GDP current movements, therefore it is justifiable to compare these two models using the root of the mean squared forecasting error.

Based on the out-of-sample forecasting error, we can conclude that the LSTM neural network model is more successful at assessing current GDP movements and the near-term GDP dynamics (in this case, up to five quarters ahead). Still, we stated the reasons why the results of the LSTM model should be taken with reservations because of the small sample, low variance, and the applied software. The root mean squared error on the database for LSTM validation is significantly higher than the forecasting error of the MIDAS model.

The contribution of the MIDAS model is significant when it comes to the differentiation between regressors that are statistically significant and the ones that are not, as the LSTM model has unspecified parameters, i.e. insight into its structure is not possible. A specificity of

the nowcasting model is also reflected in the manner of assessing the statistical significance of the explanatory variable, given that the justifiability of including regressors whose test of statistical significance has a high p-value is explained by the improvement of the forecast. The final form of the MIDAS model contains three variables, one of which is an official indicator of the statistical office, industrial production index, while two variables are alternative indicators of economic activity, Google's business trends index related to corporate-related searches, as well as the measure of electricity consumption. Interestingly, the industrial production index can be associated with the industrial sector, while the business trends index can intuitively be associated with the services sector which together make up almost three-quarters of total GDP, on the production side. Electricity consumption as an indicator can be associated with all branches of GDP on the production side, and can thus include the fluctuations not included in industrial production or Google's business trends index in the industry and service sectors, though it can contain shocks that affect the construction and agriculture sectors which, together with the above mentioned branches and net taxes, comprise the entire GDP.

This paper made several references to sources that underscore sufficiently large databases as a necessary condition for a successful machine learning model. The phrase "a sufficiently large base" means nothing on its own if it is not accompanied by a specific number. Sadly, there is not an accurate amount of data that would guarantee that a machine learning model will function in the best possible manner and give the best results. It is a fact that deep learning models were partly made precisely in order to process very large amounts of inputs generated each second, primarily through internet use. Therefore the bases used for the training of deep learning models have up to several million inputs, which is impossible to achieve with macroeconomic data. Nevertheless, deep learning models will record an improvement in their performances by adding new and good quality inputs (there has been talk about counterproductivity if bases are expanded with data characterised with issues such as missing segments, collinearity, autoregression, etc). This assertion is true not only for machine or deep learning models, but also for traditional econometric models, with which it is easier to come by meaningful representations of reality if there is a "sufficient" amount of data. There is no specific number, but it is almost certain that the database used in this paper does not meet the size criterion that would justify the transition from traditional econometric to deep learning models. In contrast, the complexity and non-linearity present in this base, reflected in a large number of potential regressors, their varying frequency and complex dynamics and problematics of GDP forecasting, partly justify the attempt to model alternative approaches relative to classic econometric models.

One of the reasons for a small amount of data is the inclusion of alternative indicators since their series are significantly shorter than the official macroeconomic indicators of economic activity. Still, these indicators have been shown to be statistically significant (electricity consumption) and better results were achieved when they were included in the model (Google trends) than when a comparable model was used based solely on official statistics. In addition, the complexity in the manner of collecting, processing, publishing and the frequency of these indicators also speaks in favour of using a machine learning model. Hence, if the number of available indicators and the length of the available series continue to

grow, we can also expect an increase in performances of both traditional and machine learning models.

In short, the assessed MIDAS model and the neural network model adequately model GDP dynamics in the short term, with the LSTM model turning out more efficient if the root mean squared forecasting error is used for comparison. Both models displayed a tendency towards overfitting to available data, which is mostly a consequence of the relatively small database for assessing the models. Still, though the LSTM model was shown to be better, if we bear in mind the computational demands and the inability to analyse the significance of individual parameters, the MIDAS model gives us more than an adequate short-term projection of Serbia's GDP dynamics. As one of the hypotheses of this paper is to verify the statistical significance of alternative parameters in combination with official macroeconomic indicators, after an insight into the MIDAS model we can conclude that Google's business trend indicator and the electricity consumption indicator from the ENTSO-E platform contribute to the improvement of the short-term projection. Besides, the alternative indicators used in combination with the industrial production indicator, in a simplified manner, model the individual branches within the production approach to the assessment of total GDP. Essentially, even though a deep learning model has turned out to be better at projecting Serbia's short-term GDP dynamics, because of its structure and transparency, the MIDAS model is a valuable input for economic policy makers, corporates and individuals who benefit from the timely and adequate economic activity assessment as well. As the database grows, the performances of both assessed models are expected to be increasingly better, especially the deep learning model.

## References


Almon, S. (1965). The Distributed Lag Between Capital Appropriations and Expenditures. *Econometrica*, 33 (1), Article 1. https://doi.org/10.2307/1911894.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012.

Anesti, N., Galvão, A. B. & Miranda-Agrippino, S. (2018). Uncertain Kingdom: Nowcasting GDP and its Revisions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3278618.

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. & Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 14 (1), C25–C44. https://doi.org/10.1111/j.1368-423X.2010.00328.x.

Arencibia Pareja, A., Gomez-Loscos, A., Luis López, M. D. & Perez-Quiros, G. (2020). A Short Term Forecasting Model for the Spanish GDP and its Demand Components. *Economía*, 43 (85), 1–30. https://doi.org/10.18800/economia.202001.001.

Armesto, M. T., Engemann, K. M. & Owyang, M. T. (2010). Forecasting with Mixed Frequencies. *Federal Reserve Bank of St. Louis Review.*

Balassa, B. (1985). Exports, policy choices, and economic growth in developing countries after the 1973 oil shock. *Journal of Development Economics*, 18 (1), 23–35. https://doi.org/10.1016/0304-3878(85)90004-5.

Bańbura, M. & Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, 27 (2), 333–346. https://doi.org/10.1016/j.ijforecast.2010.01.011.

Barbaglia, L., Consoli, S. & Manzan, S. (2021). Forecasting GDP in Europe with Textual Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3898680.

Barhoumi, K., Darné, O., Ferrara, L. & Pluyaud, B. (2012). Monthly GDP Forecasting Using Bridge Models: Application for the French Economy. *Bulletin of Economic Research*, 64, s53–s70. https://doi.org/10.1111/j.1467-8586.2010.00359.x.

Burns, A. & Mitchell, W. (1946). *Measuring Business Cycles*. NBER.

Chernis, T. & Sekkel, R. (2017). A dynamic factor model for nowcasting Canadian GDP growth. *Empirical Economics*, 53 (1), 217–234. https://doi.org/10.1007/s00181-017-1254-1.

Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Chollet, F. & Allaire, J. J. (2018). *Deep learning with R*. Manning.

Chollet, F., Kalinowski, T. & Allaire, J. J. (2022). *Deep learning with R* (Second edition). Manning.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S. & Schmidt-Thieme, L. (2021). *Do We Really Need Deep Learning Models for Time Series Forecasting?* (arXiv:2101.02118). arXiv. http://arxiv.org/abs/2101.02118.

Fradkov, A. (2020). Early History of Machine Learning. *IFAC-PapersOnLine*, 53 (2), 1385–1390. https://doi.org/10.1016/j.ifacol.2020.12.1888.

Frale, C. & Monteforte, L. (2010). *FaMIDAS: A Mixed Frequency Factor Model with MIDAS structure*.

Ghysels, E., Kvedaras, V. & Zemlys, V. (2016). Mixed Frequency Data Sampling Regression Models: The *R* Package *MIDASR*. *Journal of Statistical Software*, 72 (4), Article 4. https://doi.org/10.18637/jss.v072.i04.

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2002). *The MIDAS Touch: Mixed Data Sampling Regression Models*.

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2004). *Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies* (No. w10914; Issue w10914, p. w10914). National Bureau of Economic Research. https://doi.org/10.3386/w10914.

Giannone, D., Agrippino, S. M. & Modugno, M. (2013). *Nowcasting China Real GDP*.

Giannone, D., Reichlin, L. & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55 (4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010.

Goulet Coulombe, P., Leroux, M., Stevanovic, D. & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37 (5), Article 5.

Hopp, D. (2022). Economic nowcasting with long short-term memory artificial neural networks. *Journal of Official Statistics*, 38 (3), 847–873.

Ilin, C., Annan-Phan, S., Tai, X. H., Mehra, S., Hsiang, S. & Blumenstock, J. E. (2021). Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific Reports*, *11*(1), 13531. https://doi.org/10.1038/s41598-021-92892-8.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31 (3), 685–695. https://doi.org/10.1007/s12525-021-00475-2.

Kitchen, J. & Monaco, R. (2003). Real-time Forecasting in Practice: The U.S. Treasury Staff's Real-time GDP Forecast System. *U.S. Department of the Treasury*.

Koenig, E. F., Dolmas, S. & Piger, J. (2003). The Use and Abuse of Real-Time Data in Economic Forecasting. *Review of Economics and Statistics*, 85 (3), 618–628. https://doi.org/10.1162/003465303322369768.

Loermann, J. & Maas, B. (2019). *Nowcasting US GDP with artificial neural networks*.

Markiewicz, A., Verhoeks, R. C., Verschoor, W. F. C. & Zwinkels, R. C. J. (2018). *The Winner Takes it All: Measuring Attention and Predicting Exchange Rates with Google Trends*.

Medeiros, M. C., Terasvirta, T. & Rech, G. (2005). Building Neural Network Models for Time Series: A Statistical Approach. *Journal of Forecasting*, 25 (1), Article 1.

Moody, J., Levin, U. & Rehfuss, S. (1993). *Predicting the U.S. Index of Industrial Production (Extended Abstract)*.

Mourougane, A. (2006). *Forecasting Monthly GDP for Canada*. OECD. https://doi.org/10.1787/421416670553.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Narita, F. & Yin, R. (2018). *In Search of Information: Use of Google Trends' Data to Narrow Information Gaps for Low-income Developing Countries* [IMF Working Papers].

Plasek, A. (2016). On the Cruelty of Really Writing a History of Machine Learning. *IEEE Annals of the History of Computing*, *38*(4), 6–8. https://doi.org/10.1109/MAHC.2016.43.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2 (3), 160. https://doi.org/10.1007/s42979-021-00592-x.

Schmidt, T., & Vosen, S. (2009). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. *Ruhr Economic Papers*.

Siami-Namini, S., Tavakoli, N. & Namin, A. S. (2019). The Performance of LSTM and BiLSTM in Forecasting Time Series. *2019 IEEE International Conference on Big Data (Big Data)*, 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997.

Sinko, A. (2008). *Some Applications of Mixed Data Sampling Regression Models*.

Stock, J. H. & Watson, M. W. (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In *Handbook of Macroeconomics* (Vol. 2, pp. 415–525). Elsevier. https://doi.org/10.1016/bs.hesmac.2016.04.002.

Stock, J., & Watson, M. (1988). *A Probability Model of The Coincident Economic Indicators* (No. w2772; p. w2772). National Bureau of Economic Research. https://doi.org/10.3386/w2772.

Sun, S. & Anwar, S. (2015). Electricity consumption, industrial production, and entrepreneurship in Singapore. *Energy Policy*, 77, 70–78. https://doi.org/10.1016/j.enpol.2014.11.036.

Turnovsky, S. J. (2000). *Methods of macroeconomic dynamics* (2. ed). MIT Press.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28 (2), 3–28. https://doi.org/10.1257/jep.28.2.3.

Woloszko, N. (2020). *Tracking activity in real time with Google Trends* (OECD Economics Department Working Papers No. 1634; OECD Economics Department Working Papers, Vol. 1634). https://doi.org/10.1787/6b9c7518-en.

Wu, L. & Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2022293.

Zeng, A., Chen, M., Zhang, L. & Xu, Q. (2022). *Are Transformers Effective for Time Series Forecasting?* (arXiv:2205.13504). arXiv. http://arxiv.org/abs/2205.13504.

Zhang, Y. & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, *4*(1), 25. https://doi.org/10.1038/s41524-018-0081-z.

Zheng, Y., Xu, Z. & Xiao, A. (2023). Deep learning in economics: A systematic and critical review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10272-8.

# TOPIC CLASSIFICATION OF ECONOMIC NEWSPAPER ARTICLES IN A HIGHLY INFLECTIONAL LANGUAGE – THE CASE OF SERBIA

Mirko Đukić

© National Bank of Serbia, March 2024

Available at www.nbs.rs

The views expressed in the papers constituting this series are those of the author(s), and do not necessarily represent the official view of the National Bank of Serbia..

Economic Research and Statistics Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

**Topic classification of economic newspaper articles in a highly inflectional language – the case of Serbia**

Mirko Đukić

**Abstract:** The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. The application of topic modelling in the Serbian language, using the LDA model, is hampered by the fact that Serbian is a highly inflectional language, where words have a large number of forms which the model recognises as words with a different meaning. In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicability in further economic analyses.

**Keywords:** textual analysis, topic modelling, Latent Dirichlet Allocation, LASSO model
[**JEL Code]:** C13, C55, E31, E37, E52

## Non-technical summary

Newspaper articles are an important source of information about economic developments. They can cover a wide spectrum of economic topics – from the analysis of individual companies to the global economy. The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. A precondition for this type of analysis is that a large number of articles are classified based on the topic covered.

Topic modelling is a textual analysis technique that uncovers the patterns of common occurrence of certain words in a set of documents, interpreted as hidden topics in that set. In this paper, we used the Latent Dirichlet Allocation (LDA) method which is based on the assumption that each topic is a combination of different words, and each article a combination of different topics.

The application of topic modelling in the Serbian language, as well as indeed any other textual analysis method, is hampered by the fact that Serbian is a highly inflectional language, in which words have a large number of forms that the model recognises as words with a different meaning.

In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance (when applying the LDA model for classification into topics) than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicability in further economic analyses.

The analysis was applied to 25,248 articles in the economy section of the Politika daily in the period 2006–2023. The LDA model extracted 40 topics from these articles, and they were later named based on the most frequent words in each topic. In the majority of cases, the topic covered by the article was unambiguously determined, and only in two of the 40 cases were we unable to determine the content of the topic. Some topics in the sample cover broad areas (trade, corporates, economy…), while others are narrow and specific (fuel, petroleum, copper, steel, electricity…).

Lastly, the obtained series of the shares of topics over time were regressed to household inflation expectations using the LASSO model. The estimated model was well able to catch inflation cycles with a high determination coefficient. Of the 40 topics, the model kept 17 as relevant, some of which expectedly so, while for some there was no clear economic interpretation.

## Contents:

# 1 Introduction

Newspaper articles are an important source of information about economic developments. They can cover a wide spectrum of economic topics – from the analysis of individual companies to the global economy. The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. A precondition for this type of analysis is that a large number of articles are classified based on the topic covered.

Topic modelling is a textual analysis technique that uncovers the patterns of common occurrence of certain words in a set of documents, interpreted as hidden topics in that set. The first model for topic modelling was the Latent Semantic Analysis (Dearwester et al. 1990), which grouped documents based on words with a similar semantic structure.

In this paper, we used the Latent Dirichlet Allocation (LDA) model developed by Blei et al. (2003), which is based on the assumption that each topic is a combination of different words, and each article a combination of different topics. The LDA uncovers topics based on the patterns of the common occurrence of words (for each word, there is a certain likelihood that it would occur under a certain topic), and topics occur as the likelihood of occurrence in each document. This method can be used for analysis in different fields, such as law, humanities, classification of scientific papers, etc.

For the purpose of economic analysis, the LDA model can analyse different types of texts and we will list several examples here. Combining the LDA with a dictionary approach, Angelico et al. (2021) created a measure of inflation expectations based on Twitter (now the X platform) and established that it is highly correlated with the usual inflation expectation measures. Gonzales et al. (2018) created a measure of volatility of policies for ten Latin American countries based on the change in the topic of their presidents' speeches and concluded that greater changes in policies lead to smaller economic growth. Yono et al. (2020) measured news-based macroeconomic uncertainty for the purpose of investment decisions. An expanded topic model they proposed attributes a numerical value to each individual text, and the indices obtained in this manner correlate well with market volatility indices. The paper most similar to ours is by Larsen et al. (2021), who concluded that the media coverage of certain topics, some of which initially seemed unrelated to inflation, can play a significant role in predicting consumers' inflation expectations.

The application of topic modelling in the Serbian language, as indeed is the case with any textual analysis model, is hampered by the fact that Serbian is a highly inflectional language, in which words have many forms that the model recognises as words with a different meaning. In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance (when applying the LDA model for classification into topics) than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicable value in further economic analyses.

The use of textual analysis in economic research is not novel in the National Bank of Serbia. Đukić (2022) created an indicator of inflationary pressures based on counting

expressions related to price changes in newspapers, which was determined to precede inflation movements. In this paper, we analysed newspaper articles from a different, topic standpoint.

Below we will first present a theoretical description of the LDA model. Afterwards, we will explain the way in which we prepared the articles for topic analysis, and then present the results of the LDA model application. Lastly, we will conclude the paper by assessing the impact of the movement of topics on inflation expectations by applying the LASSO model.

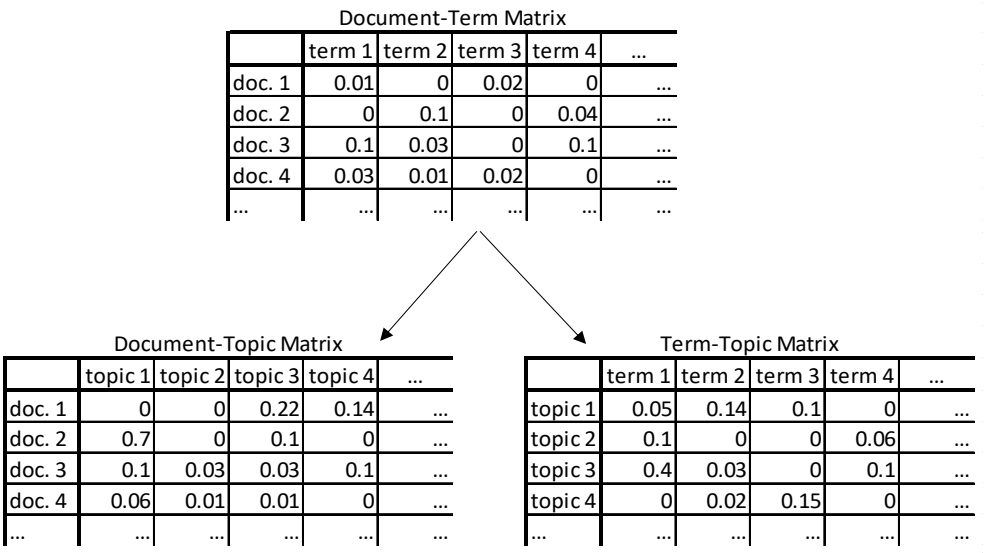
## 2 LDA model for classification of articles into topics

In this paper we use the Latent Dirichlet Allocation (LDA) model, developed by Blei et al. (2003). It is a widely used tool for detecting hidden topics within large sets of documents (newspaper articles in our analysis). The model assumes that documents are combinations of different topics and that topics are combinations of different words.

The LDA model treats documents as "bags of words", where their order and grammatical meaning play no role. Different forms of words with the same basic meaning are treated as entirely independent terms. The main aim of applying the LDA model is to reveal topics as sets of words that occur together, where some will have greater importance (probability) than others, and to present each document as a combination of different topics, where some topics will also be more important than others. The number of topics into which we wish to classify documents must be defined in advance.

The starting basis for the analysis is the Document Term Matrix (DTM), whose rows represent documents and columns all unique words in all documents. The element $i, j$ in the matrix shows the share of the occurrence of word $j$ in document $i$. As rows represent the shares of all words in the document, the sum of elements by rows is 1. Thus, the LDA from the DTM assesses the elements of the matrix of the share of topics in each document and the share of words in each topic (Figure 1).

Figure 1 **Matrices connecting words, topics and documents in the LDA model**



The LDA model algorithm is iterative. In the first iteration, one topic is randomly assigned to each word in each document, which is in the following iterations gradually adjusted based on specific criteria until optimal distribution is achieved.

The process is carried out for each "current" word by treating the topics assigned to other words as accurate. For instance, topic *t* in document *d* is assigned to current word *w*. The following probabilities are calculated:

- $p_1$: proportion of other words in document *d* that are assigned to the same topic *t*.
- $p_2$: proportion of documents assigned to topic *t* which stem from word *w*.

If a larger number of words from a given document belong to the same topic *t* (high $p_1$), it is more probable that the current word belongs to that topic. If the current word has a high probability of belonging to topic *t*, all documents containing *w* will be more strongly associated with topic *t* (high $p_2$). The higher the product of probabilities $p_1 \cdot p_2$, the more probable it is that the current word *w* belongs to topic *t*.

The LDA is run in a large number of iterations by assigning a new topic to word *w* based on the product of probabilities $p_1 \cdot p_2$ until equilibrium is achieved. The outcome are similar words grouped by topics, with the probabilities of each of them to stem from the given topics, and the distribution of the probabilities of shares of individual topics in each individual document. For a detailed mathematical overview of this method see Blei et al. (2003).

## 3 Text preparation for analysis

Before running the LDA model, texts should be abbreviated and adjusted, rendering them more conducive to faster machine processing. This includes reducing words to their base form, changing Latin letters specific for Serbian, eliminating frequent non-important words and punctuation, and turning uppercase letters into lowercase.

The Serbian language has a large number of forms of the same word, i.e. linguistically speaking it is a *highly inflectional* language, which generally aggravates any type of textual analysis. For instance, the algorithm recognises the words *nafta*, *nafte*, *naftu*, *naftni*, *naftna* (*oil* in different case forms) and similar words as different, although they all relate to the forms of the same base word. In our analysis, this would reduce the importance of the given term for classification of articles into topics, as its different forms will have a smaller frequency of occurrence.

One of the methods of reducing words to the base form is stemming, where suffixes are eliminated, keeping common beginning for all forms of the given word. Another method is lemmatisation, which reduces words to their roots as they would appear in the dictionary. The latter method is much more complex as it requires the existence of a dictionary with all forms of all words in a language. On the other hand, the stemmer is based on the rules whose number does not exceed several hundred, and can be developed much more easily.

Several automatic, programme stemmers have been developed for Serbian. The most famous was developed by Krovetz and Šipka (2008), according to whom the stemmer for Serbian must have eight times more rules for suffix stripping than for English, which is a good

illustration of how highly inflective Serbian language is. The above example contains several forms of the word *nafta* (*oil*), while *oil* is the only form of this word in English.

Instead of using the general stemmer that would treat all words in texts, we decided to reduce to the base form only those words that are relevant for our analysis, i.e. economic terms, whereby we favour their role in classifying texts into topics. Compared to non-economic terms, which remain in many different forms, economic terms reduced to the base form will have a larger share and be more relevant for classification. In this way, the aggravating circumstance of working with a language with a high number of word modifications was turned into an advantage for this type of analysis.

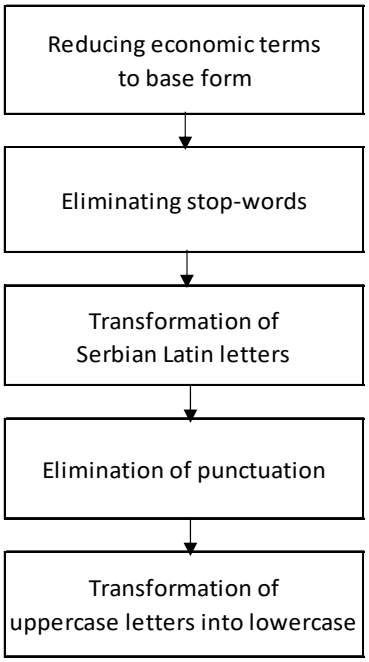
We interpreted the base form in the broadest possible way – unlike the standard stemmers, it includes all adjectival, verb and noun forms of a word. While stemmers for Serbian would treat the words *izvoz*, *izvozni* and *izvoziti* (*export* as a noun, an adjective, and a verb) as different base forms, in our analysis we reduced all forms of these words into the word *izvoz* (in order to attach greater importance to economic terms). Moreover, in some cases we grouped similar words with different prefixes as well, as is the case with *skupo (expensive)* and *poskupljenje (price increase)*.

Words were reduced to the base form in two ways. Where possible, words with the same base forms were identified based on the common beginning of all word forms, the so-called "stem", and were then replaced with the base form. For instance, the words *nafta, nafte, naftu, naftni, naftna*, etc. start with the letters *naft*, which is why we replaced them with the word *nafta*. Under the same principle, we replaced the words starting with *kamat* with the word *kamata* (*interest*), and the words starting with *inflaci* or *inflator* with *inflacija* (*inflation*) etc. This rule could be applied only if the stem for different forms of the base term is not at the same time the stem of another non-similar term.

Contrary to this, for instance, the stem for different forms of the word *cena* (*cenu, cene, cenama, cenovni* …) (*price* in different forms) – *cen* – is at the same time the stem for forms of some other non-similar terms (*centar*, *ceniti*, *cenzura*) (*centre, assess, censorship*). Applying the above rule in this case would unjustifiably transform the other words into the word *cena*, thus changing their meaning. Therefore, in this and similar cases, it was necessary to explicitly define all concrete forms of words that we wish to transform into the base form.

The first method (recognition based on the beginning of the word) is obviously simpler to apply, so we used it in certain cases where there are several terms for the same word beginnings. For example, words beginning with *bank* can be forms of the words *banka* and *bankina* (*bank* and *road barrier fence*), but as the latter term is very rare, if not non-existent in economic texts, there was no reason for concern that its unjustified transformation could disrupt the further analysis.

Figure 2 **Text preparation for the LDA model**




There was a certain degree of arbitrariness in choosing economic words to be reduced to the base form. Certain words, such as *saobraćaj, država, vlada* or *grejanje* (*traffic, state, government* or *heating*), strictly speaking, are not economic, but as we analysed economic texts only, we assumed that these words have an economic connotation, so we included them in the transformation list. In this list, we also included the names of institutions that are often mentioned in newspaper articles (IMF, EPS, Telekom).

While preparing the text for further analysis, we eliminated the stop-words, which are frequent in Serbian but are not essential, such as: *i*, *ili*, *ali*, *koji*, *to*, *od*, *gde* (*and, or, but, who/which, that, from, where*). Due to their frequency, the algorithm may recognise them as crucial in classifying texts into topics, which is not useful for economic or any other type of analysis.

As our text analysis program does not recognise Serbian Latin letters (*č, ć, ž, đ, š*), we transformed them into forms suitable for processing. We combined the base letters (without diacritics) with the letters not used in Serbian (*x*, *y*) as follows: *č→cx, ć→cy, ž→zx, đ→dx, š→sx.*

Punctuation was also eliminated, and uppercase letters were converted into lowercase letters. All these interventions – the elimination of frequent words and punctuation, and reducing frequent economic terms to the common base form make the text more conducive to our analysis, but also reduce the time for text processing, which is not insignificant given that the process may take several hours.

The above transformation rules can be illustrated with the following text:

*Vlada ograničila cene osnovnih životnih namirnica*
*Vlada Srbije na današnjoj sednici donela je odluku da ograniči visinu cena osnovnih životnih namirnica: šećer, brašno tip T-400, suncokretovo ulje, svinjsko meso i dugotrajno mleko sa 2,8 procenata mlečne masti, tako da one ne prelaze nivo cena na dan 15. novembar 2021.*
*Ograničenje cena utvrđeno je kako bi se otklonile štetne posledice i sprečili poremećaji na tržištu i neće se odnosit i na snižene cene, kao što su rasprodaje, sezonska sniženja ili akcijske prodaje, ukoliko su bila na snazi 15. novem bra, već na redovne, odnosno cene pre sniženja, saopšteno je iz vlade, preneo je Tanjug.*
*Odlukom, koja će biti primenjivana u trajanju od 60 dana, predviđeno je da proizvođači ove proizvode ne smeju isporučivati u količinama manjim od prosečnih u poslednjih 12 meseci.*
*Za kršenje navedenih odredaba, predviđene su i novčane kazne u iznosu od 100.000 do dva miliona dinara, kao i zabrana vršenja delatnosti u trajanju od šest meseci do jedne godine.*

and its transformed version:

*vlada ogranicxila cena osnovnih zxivotnih namirnica vlada srbije danasxnjoj sednici donela odluku ogranicxi visinu cena osnovnih zxivotnih namirnica sxecyer brasxno tip t400 suncokretovo ulje svinjsko meso dugotrajno mleko 28 procenata mlecxne masti one ne prelaze nivo cena dan 15 novembar 2021 ogranicxenje cena utvrdjeno bi otklonile sxtetne posledice sprecxili poremecyaji trzxisxte necye odnositi snizxene cena sx su rasprodaje sezonska snizxenja akcijske prodaje ukoliko su bila snazi 15 novembra redovne odnosno cena snizxenja saopsxteno vlada preneo tanjug odlukom primenjivana trajanju 60 dana predvidjeno proizvod proizvod ne smeju isporucxivati kolicxinama manjim prosecxnih poslednjih 12 meseci krsxenje navedenih odredaba predvidjene su novac kazne iznosu 100000 dva milion dinar zabrana vrsxenja delatnosti trajanju sxest meseci jedne godine*

Though for a Serbian reader the first article is certainly easier to understand, the transformed article is more conducive to the LDA method processing for the reasons we have outlined above. The transformed article is also shorter (by around 30%), which is important when processing large quantities of data.

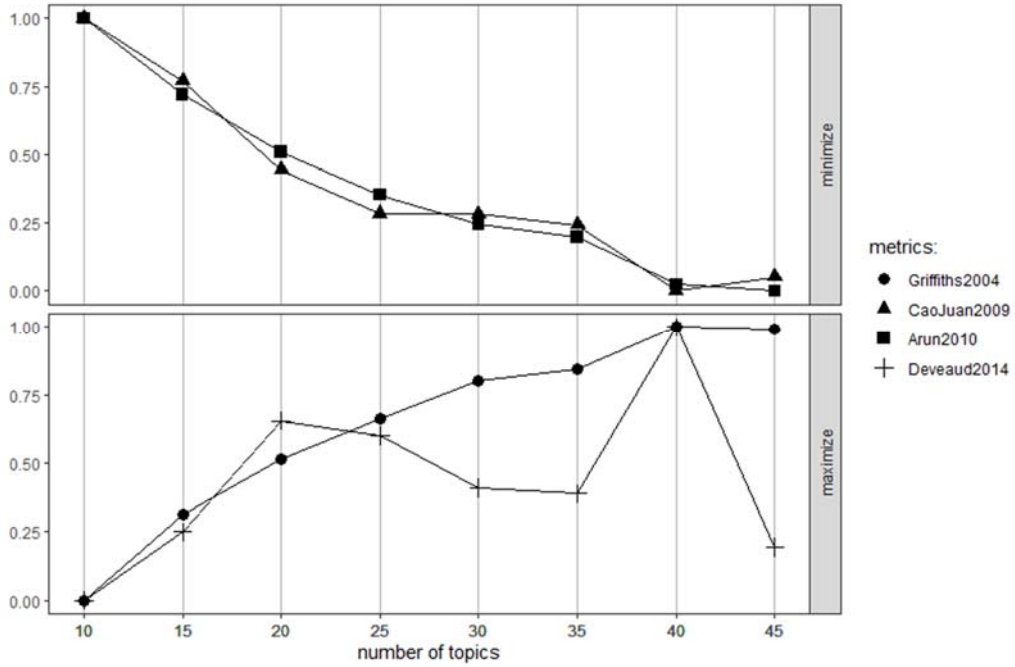
The choice of economic words for transformation and the elimination of frequent superfluous words was not entirely pre-defined, but was supplemented in several steps in parallel with the application of the LDA model. In the first step, the model was run with the initially shorter list of transformed and eliminated words, based on which the texts were classified into topics. Then, by analysing the most frequent words in topics, we extended the list of stop-words and economic terms to be reduced to their base forms. We ran the LDA model again with the new extended lists for word elimination/correction. We repeated the entire process several times, filtering the dominant words in topics. The final list of words for transformation is contained in Table A1 of the Appendix.

## 4 Text classification into topics with the LDA model

The starting point for the topic classification of texts with the LDA model is the Document-Term Matrix that contains the number of term occurrences per document. In our analysis, the Matrix contains 25.248 · 193.142 elements, where the first dimension represents the number of topics, and the second the number of unique words.

When running the LDA model, it is necessary to specify the number of topics into which we wish to classify the articles from the sample. To determine the optimal number of topics for classification, we used the four criteria developed by Griffiths, et al. (2004), Cao et al. (2009), Arun et al. (2020), and Deveaud (2014). According to three of these four criteria, the optimal number of topics for our sample is 40 (Chart 1).

Chart 1 **Criteria for the selection of optimal number of topics for the LDA model**




As already stated, we estimated the model in several steps, supplementing in each step the list of words for transformation or elimination, based on the analysis of the most frequent words in topics.

Chart A1 in the Appendix shows the final division into topics and the most frequent words. Technically, those are words with the highest estimated $\beta$ coefficient, which measures the probability of a concrete word stemming from a concrete topic. Note that the model did not specify topics titles on its own (as it only marked them with numbers), i.e. we have done this based on the dominant words per topic.

In most cases, it was easy to determine the topic based on the most frequent words. For instance, the topic where words such as *zaposlenost, radnik, rad, posao* (*employment, employee, labour, work*) were dominant is obviously related to the labour market; the topic with the words *proizvod, poljoprivreda, tržište, tona, pšenice, voća* (*product, agriculture, market, tonne, wheat, fruit*) concerns the agricultural commodities market; the topic with the words *energetika, EPS, elektrika, struja, uglja* (*energy, EPS (Electric Power Industry of Serbia) electrics, electricity, coal*) pertains to electrical energy, etc. Particularly relevant for the central bank and monetary policy are topics related to the banking sector (*kredit, banka, kamata, dinar…*) (*loan, bank, interest, dinar…*) and inflation (*inflacija, odsto, rast, cena…*) (*inflation, percent, growth, price…*). The titles of two topics could not be determined given

that non-economic and unrelated words were dominant in them. We named these topics *UNTITLED*[1]. In some cases, we had several topics dealing with the same area (*ECONOMY, AGRICULTURE, GOVERNMENT, PENSIONS, TOURISM, WORKERS*).

While each topic is a combination of different words, each article is a combination of different topics. Coefficient $\gamma$ represents the share of individual topics in each document. Articles can be dominantly represented with one topic or as a combination of several topics.

For instance, *GAS* ($\gamma = 0{,}98$) is the dominant topic in the following short article:

***The Balkan Stream gas pipeline connected at the border with Hungary*** *(4 June 2021)*

*Horgoš – The public enterprise Srbijagas and the Hungarian company FGSZ connected today the Balkan Stream gas pipeline at the border with Hungary, through which gas will in future be delivered from Turkey through Bulgaria and Serbia to Central Europe.*

*Dušan Bajatović, general director of Srbijagas, said that in this way, an end was put to the long-term great work by which our country finally solved the issue of gas supply from another direction, not only through Ukraine.*

*"Nobody in Serbia will be freezing anymore, and the price of gas for households will not be changed in autumn. The increase in the price of oil, which is followed by the price of gas, will spill over to Srbijagas", Bajatović explained.*

On the other hand, the article below is a combination of several topics, where none has a majority share (*FUEL 0.42; STOCK EXCHANGE 0.33; OIL 0.24*):

***European stock indices mostly down, gold price at a historic high*** *(4 December 2023)*

*NEW YORK/FRANKFURT/MOSCOW – European stock indices are mostly down at the beginning of the week, while the price of gold is at an all-time high today. The DAX index of the Frankfurt Stock Exchange rose to 16,421.45 points at 10:00 a.m. today, while the French CAC 40 fell to 7,333.14 points, as well as the London FTSE 100 – to 7,508.57 points and the Moscow MOEX – to 3,113.05. The value of the American stock index Dow Jones rose to 36,245.50 before today's opening of the stock market in America, as well as the value of the S&P 500 index – to 4,594.63 points and the value of the Nasdaq index – to 14,305.03.*

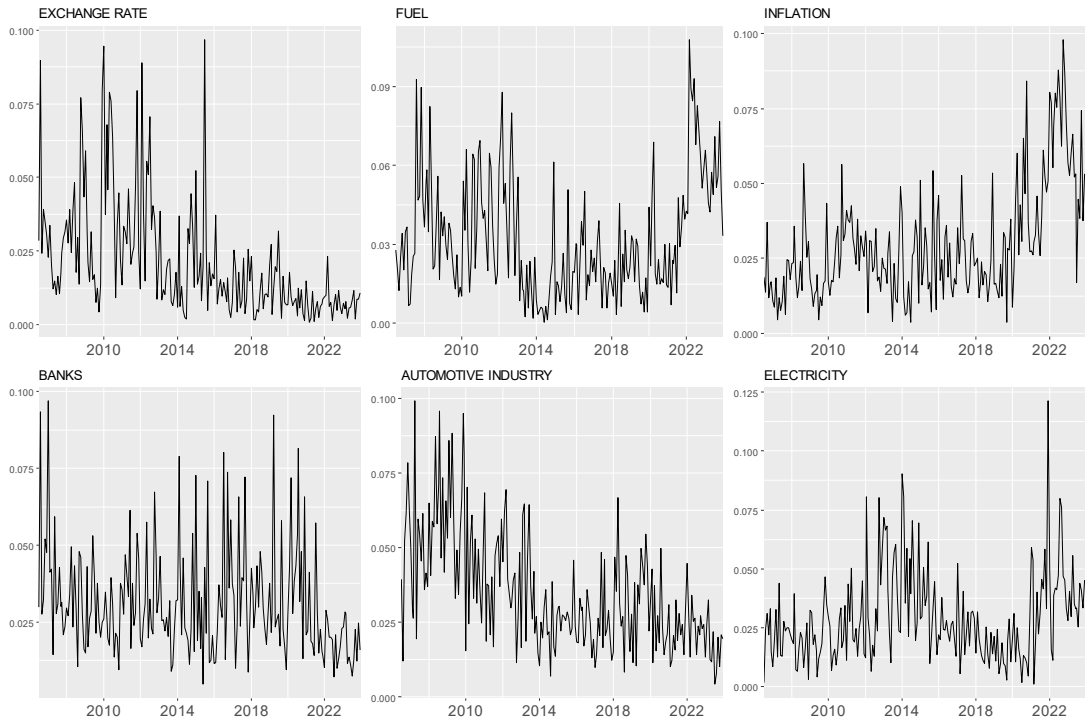
*According to stock exchange data, the price of crude oil fell to $72,874 per barrel, as well as the price of Brent oil – to $77,615. European gas futures were sold today at the opening of the TTF stock exchange at a price of EUR 42,750 per megawatt-hour.*

*The price of gold reached an all-time high of $2,110.8 an ounce early this morning, and by 10:00 a.m. it fell slightly to $2,069.16 an ounce (an ounce equals 28.35 grams). Wheat also rose to $5.7936 a bushel (a bushel equals 27,216 kg). The value of the euro against the dollar was 1.08692, which is approximately the same as on Friday, reports Tanjug.*

The time series of the share of topics display exceptionally high volatility (Chart 2). Although the topic classification sample contains a large number of articles (over 25,000) – 120 articles divided into 40 themes, on average, per month, implies a small monthly number of articles per topic.

[1] The titles of topics are written in uppercase letters.

Chart 2 **Frequency of selected topics in newspaper articles** (monthly averages of the γ coefficient)




Despite high volatility, the frequency of topics well reflects developments in the economy: the interest in the exchange rate grew during the periods of its high volatility, and subsided as it stabilised in the past several years; the automobile industry was a frequent topic at the time when Fiat arrived in Kragujevac; fuel and electricity are frequent topics during the periods of their price hikes; the interest in inflation is the highest during inflation cycles.

## 5 Estimate of the link between inflation expectations and topics

Households form inflation expectations based on the information they obtain from various sources, including the media. A feedback effect is also possible – media reporting can reflect inflation expectations. In any case, writing about specific topics is a potential indicator of inflation expectations.

In this paper, we estimated the link between inflation expectations and topics by the LASSO regression (Tibshirani, R. (1996)), which is suitable in cases when the number of variables in the regression is large, of which some are irrelevant. The key model characteristic is adding the penalty term $\lambda\sum_{j=1}^{p}|\beta_j|$ in the function of the ordinary least squares objective function, which encourages the model to reduce to zero the $\beta_j$ coefficients of less important variables:

$$min: \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

The λ parameter determines the degree to which we wish to penalize the coefficient retention in the model (higher λ – less non-zero coefficients). In our analysis, we used the $\lambda$ parameter, which results in the lowest out-of-sample mean forecasting error (Table A3 in the Appendix).

In our case, we regressed inflation expectations $\pi_t^{exp}$ to 40 variables of the movement of the share of individual topics $T_t^i$ in the period January 2009 – December 2023. To avoid the potential problem of endogeneity (simultaneous impact), we took one-month arrears as independent variables.

$$\pi_t^{exp} = \alpha + \sum_{i=1}^{40} \beta_i \, T_{t-1}^i + \lambda \sum_{i=1}^{40} |\beta_i|$$

The estimated model retained 17 variables as relevant and reduced the coefficients of the other ones to zero (Table 1), with the determination coefficient of $R^2 = 0.67$ (Table A3).

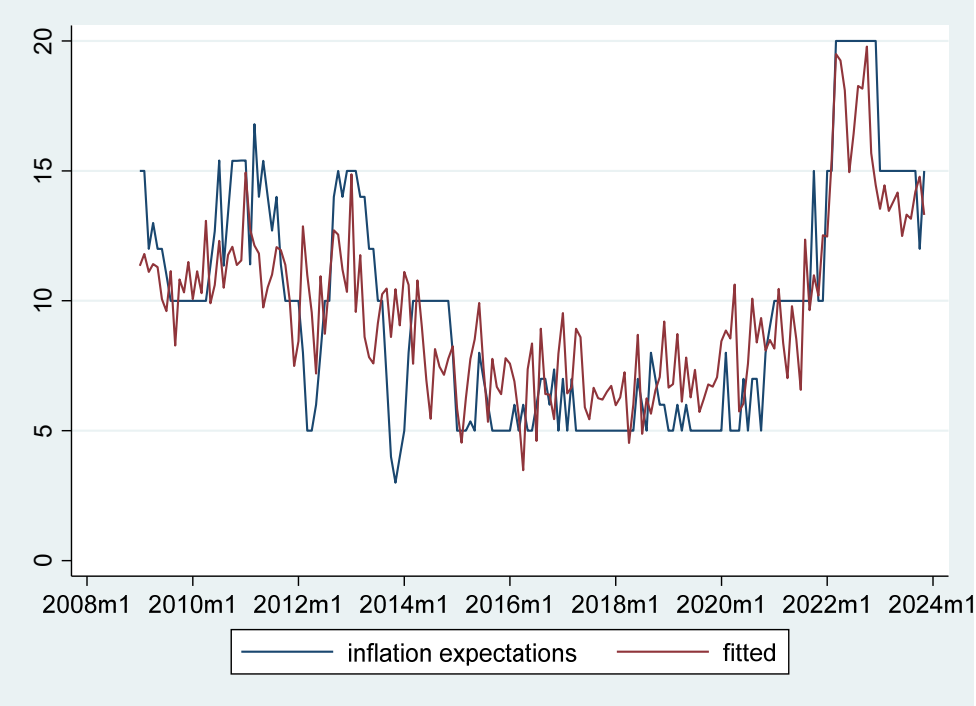
The variables with a positive coefficient include some expected topics: *INFLATION, FUEL, OIL, ELECTRICITY, WAGES-CONSUMPTION.* In these topics on energy products, the movement of their prices is often mentioned. In case of individual variables, however, there is no clear economic interpretation of the coefficient (e.g. a negative coefficient in topics *COPPER, EMPLOYMENT, ROAD CONSTRUCTION…*).

Table 1 **Estimated coefficients for topics from the LASSO model for inflation expectations**

| Topic | Coefficient | Topic | Coefficient |
|---|---|---|---|
| *1. ECONOMY* | 0 | *21. AUTOMOTIVE IND.* | 9.1 |
| *2. FX FLOWS* | 0 | *22. LAW-ECONOMY* | -16.6 |
| *3. COPPER* | -35.1 | *23. INTERNATIONAL* | 0 |
| *4. TOURISM* | 0 | *24. PENSIONS* | 0 |
| *5. TELEPHONY* | 0 | *25. OIL* | 70.3 |
| *6. FUEL* | 36.6 | *26. WAGES-CONSUMPTION* | 83.0 |
| *7. WORKERS* | 0 | *27. EMPLOYMENT* | -30.6 |
| *8. EXCHANGE RATE* | 0 | *28. AGGRICLUTURE* | 0 |
| *9. AGGRICLUTURE* | 0 | *29. LOCAL* | 27.2 |
| *10. TOURISM* | 0 | *30. ECONOMY* | 0 |
| *11. TAX* | 0 | *31. TRADE* | 0 |
| *12. BUDGET* | 0 | *32. AIR TRANSPORT* | -12.0 |
| *13. INFRASTRUCTURE* | 0 | *33. GAS* | 0 |
| *14. GOVERNMENT* | 0 | *34. STOCK EXCHANGE* | 0 |
| *15. ELECTRICITY* | 33.3 | *35. FLATS* | 50.1 |
| *16. INFLATION* | 40.0 | *36. GOVERNMENT* | 60.0 |
| *17. BANKS* | 0 | *37. FIRMS-BANKS* | 43.7 |
| *18. PENSION-INSURANCE* | -18.5 | *38. ROAD CONSTRUCTION* | -10.3 |
| *19. WORKERS* | 0 | *39. UNTITLED* | 0 |
| *20. UNTITLED* | 0 | *40. STEEL* | -23.6 |

Chart 3 shows that the model is good at predicting the inflation expectations cycles, with the estimated series being much more volatile than expectations. This is a natural consequence of high volatility in the movement of individual topics, on the one hand, and relatively stable inflation expectations from the survey, on the other.

Chart 3 **Inflation expectations and their fitted values based on the LASSO model with the movement of topics (in %)**




## 6 Conclusion

This paper shows the topic classification of 25,248 articles from the economic section of the Politika daily in the 2006–2023 period, by applying the LDA model.

The specificity of our approach lies in the fact that during the text preparation stage, we selectively chose only economic words to be reduced to their base form. By doing so, we provided them with an advantage in topic classification over non-economic words, which remained in a large number of forms with lower frequencies of occurrence. This advantage is particularly pronounced in languages with high inflection, such as Serbian, where words have numerous forms. The selection of words for transformation was not entirely pre-defined, but was supplemented in several steps in parallel with the application of the LDA model, based on the analysis of the most frequent words in topics.

The LDA model classified the modified newspaper articles into economic topics in a satisfactory way. In most cases, the content of the topic was unambiguous, while the content of two topics, with dominantly non-economic terms, was not evident. Some topics from the sample have a relatively large scope (*ECONOMY, INTERNATIONAL*…), while some are specific (*TELEKOM, OIL, COPPER, STEEL, ELECTRICITY*).

The monthly movement of topics displays high volatility, which can be explained by the insufficiently large sample for such a large number of topics. By using the LASSO model, we regressed these series to household inflation expectations. The inflation cycles were well

captured by the estimated model with a high coefficient of determination. Of the 40 topics, the model kept 17 as relevant, some of them expectedly so, while for others there is no clear economic interpretation. To have more reliable econometric analyses, it is probably desirable to expand the sample of documents, which will be one of our objectives going forward.

# Appendix

Table A1 **Replacement of the word with its base form based on the beginning of the word**

| Beginning of word | Replacement |
|---|---|
| *inflaci* | *inflacija* |
| *inflator* | *inflacija* |
| *deflaci* | *deflacija* |
| *deflator* | *deflacija* |
| *poskup* | *poskupljenje* |
| *skuplj* | *poskupljenje* |
| *pojeft* | *pojeftinjenje* |
| *jeft* | *pojeftinjenje* |
| *kurs* | *kurs* |
| *dinar* | *dinar* |
| *dolar* | *dolar* |
| *deviz* | *deviza* |
| *novc* | *novac* |
| *kredit* | *kredit* |
| *kamat* | *kamata* |
| *bank* | *banka* |
| *banc* | *banka* |
| *banaka* | *banka* |
| *finansi* | *finansije* |
| *monetar* | *monetarna* |
| *naft* | *nafta* |
| *barel* | *barel* |
| *goriv* | *gorivo* |
| *benzin* | *benzin* |
| *dizel* | *dizel* |
| *derivat* | *derivat* |
| *energ* | *energija* |
| *elektri* | *elektrika* |
| *struj* | *struja* |
| *grejanj* | *grejanje* |
| *proizvod* | *proizvod* |
| *preduzec* | *preduzeće* |
| *kompanij* | *kompanija* |
| *fabri* | *fabrika* |
| *penzi* | *penzija* |
| *zarade* | *zarada* |
| *zarada* | *zarada* |

| Beginning of word | Replacement |
|---|---|
| *zaposlen* | *zaposlenost* |
| *zapošlj* | *zaposlenost* |
| *radn* | *radnik* |
| *porez* | *porez* |
| *pores* | *porez* |
| *budžet* | *budžet* |
| *akciz* | *akciza* |
| *drzav* | *drzava* |
| *guverner* | *guverner* |
| *minist* | *ministar* |
| *ekonom* | *ekonomija* |
| *privred* | *privreda* |
| *makroekon* | *makroekonomija* |
| *uvoz* | *uvoz* |
| *uvezen* | *uvoz* |
| *izvoz* | *izvoz* |
| *izvezen* | *izvoz* |
| *trži* | *tržište* |
| *bdp* | *bdp* |
| *trgov* | *trgovina* |
| *kriz* | *kriza* |
| *recesi* | *recesija* |
| *investi* | *investicija* |
| *poljoprivr* | *poljoprivreda* |
| *potroš* | *potrošnja* |
| *milion* | *milion* |
| *milijard* | *milijarda* |
| *infrastrukt* | *infrastruktura* |
| *turis* | *turizam* |
| *turiz* | *turizam* |
| *transakci* | *transakcija* |
| *osiguran* | *osiguranje* |
| *eps* | *eps* |
| *mmf* | *mmf* |
| *telekom* | *telekom* |
| *berz* | *berza* |

Table A2 **Replacement of the word with its base form based on the entire word**

| Word | Replacement |
|---|---|
| *cene* | *cena* |
| *cenu* | *cena* |
| *cenama* | *cena* |
| *cenovni* | *cena* |
| *cenovna* | *cena* |
| *cenovne* | *cena* |
| *evra* | *evro* |
| *evri* | *evro* |
| *evru* | *evro* |
| *evrima* | *evro* |
| *plate* | *plata* |
| *platu* | *plata* |
| *plati* | *plata* |
| *platama* | *plata* |
| *gasa* | *gas* |
| *gasu* | *gas* |
| *gasni* | *gas* |
| *gasna* | *gas* |
| *gasovod* | *gas* |
| *gasovoda* | *gas* |
| *gasovodu* | *gas* |
| *vlade* | *vlada* |
| *vladu* | *vlada* |
| *vladi* | *vlada* |
| *firmi* | *firma* |
| *firme* | *firma* |
| *firmu* | *firma* |
| *firmama* | *firma* |
| *rada* | *rad* |
| *rade* | *rad* |
| *radu* | *rad* |

Figure A1-a **The most frequent words in topics** ($\beta$ coefficients)

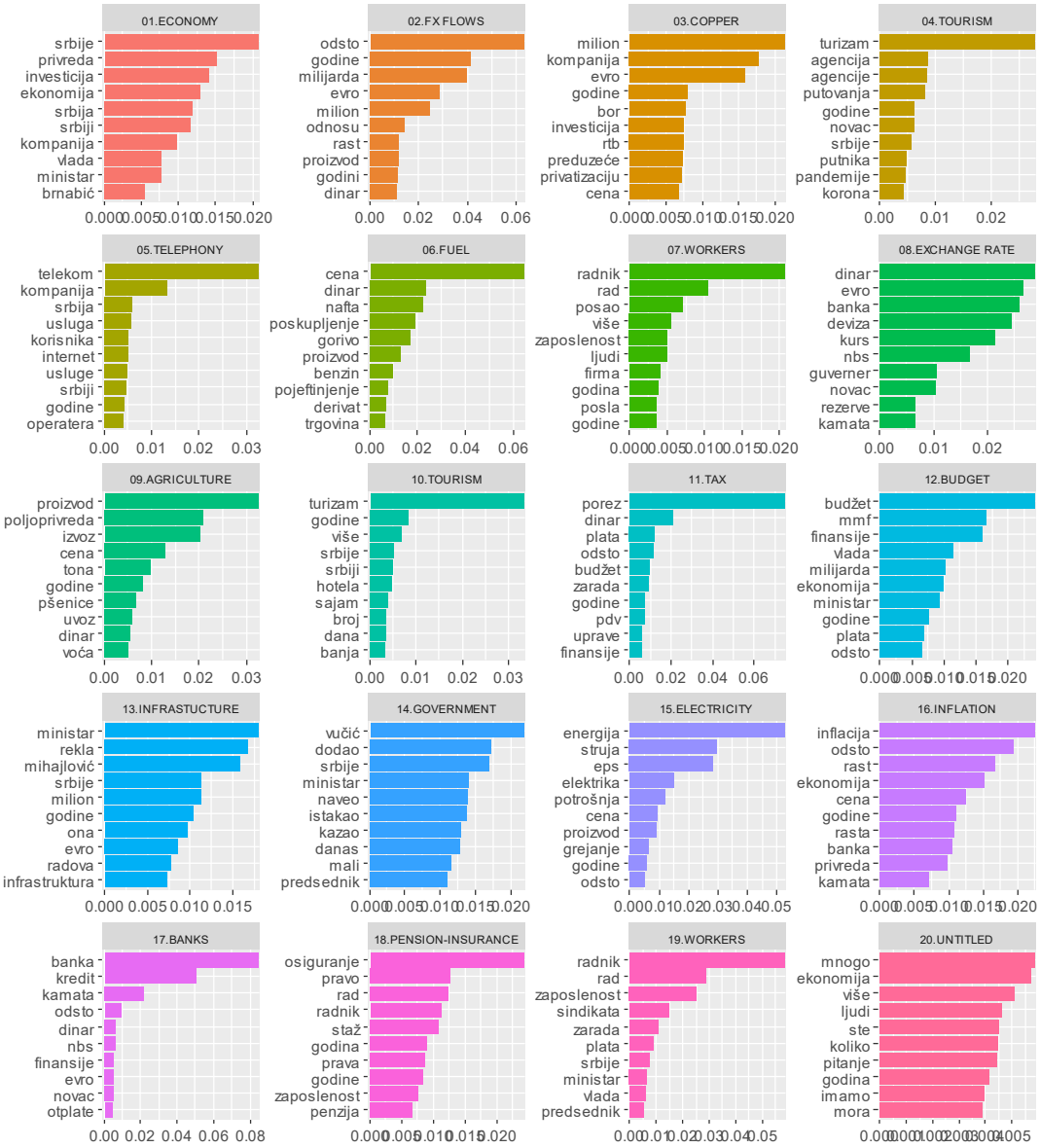


Figure A1-b **The most frequent words in topics** (*β* coefficients)

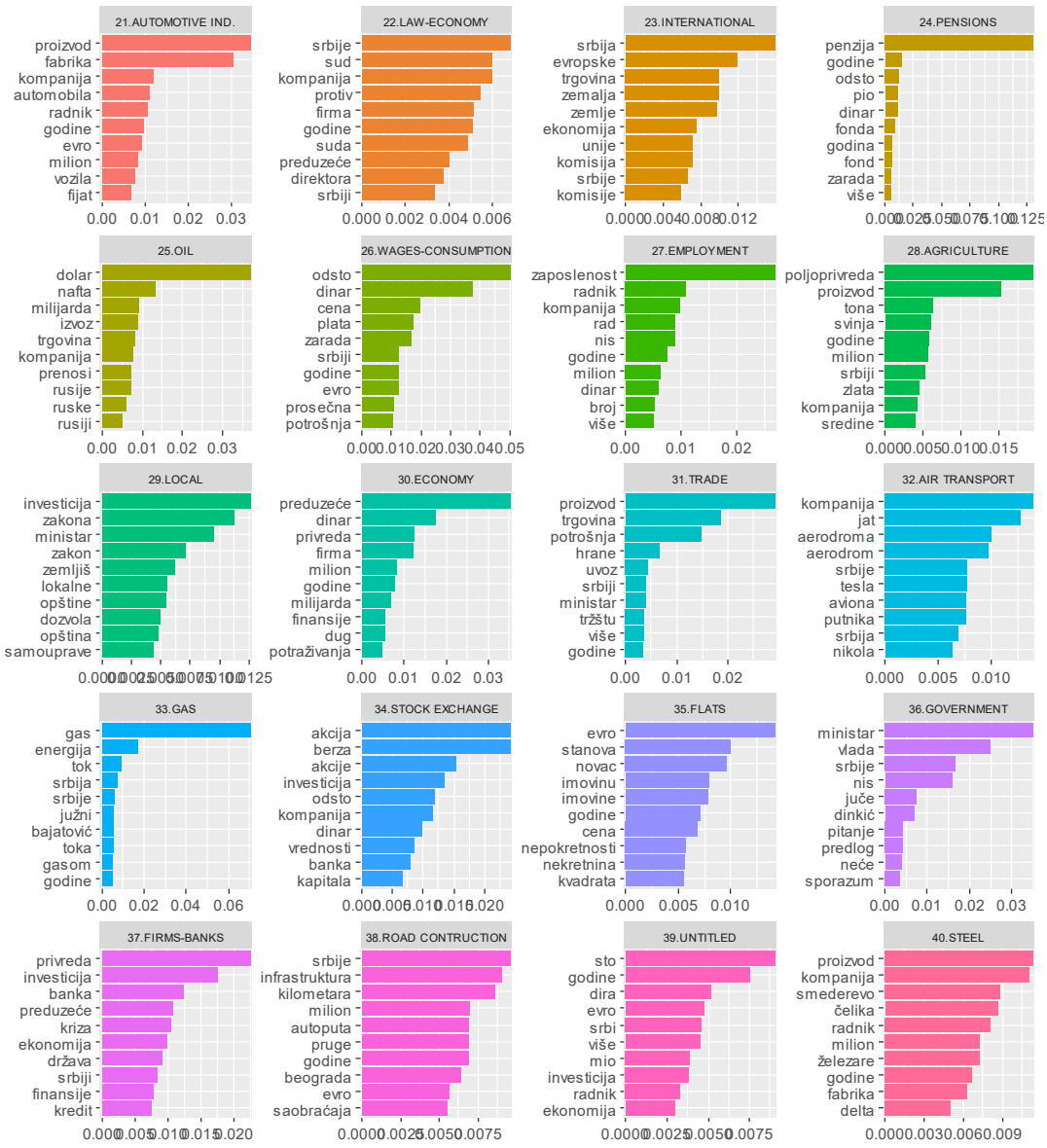


Table A3 **Results of LASSO regression of inflation expectations to topics from newspaper articles**

Lasso linear model — No. of obs = 180
No. of covariates = 40
Selection: Cross-validation — No. of CV folds = 10

| ID | Description | lambda | No. of nonzero coef. | Out-of-sample R-squared | CV mean prediction error |
|---|---|---|---|---|---|
| 1 | first lambda | 2.516326 | 0 | 0.0009 | 19.49777 |
| 29 | lambda before | .1859746 | 17 | 0.5599 | 8.587741 |
| * 30 | selected lambda | .1694532 | 17 | 0.5601 | 8.583828 |
| 31 | lambda after | .1543994 | 17 | 0.5596 | 8.594913 |
| 33 | last lambda | .1281851 | 19 | 0.5568 | 8.649741 |

* lambda selected by cross-validation.

| | active |
|---|---|
| t3 | -35.07961 |
| t6 | 36.57279 |
| t15 | 33.34805 |
| t16 | 40.01757 |
| t18 | -18.47686 |
| t21 | 9.089888 |
| t22 | -16.58271 |
| t25 | 70.2822 |
| t26 | 82.99576 |
| t27 | -30.5818 |
| t29 | 27.18242 |
| t32 | -11.97298 |
| t35 | 50.08295 |
| t36 | 59.94977 |
| t37 | 43.70352 |
| t38 | -10.29667 |
| t40 | -23.55415 |
| _cons | 1.135015 |

Penalized coefficients

| MSE | R-squared | Obs |
|---|---|---|
| 6.447942 | 0.6696 | 180 |

## Literature

Angelico C., Marcucci J., Miccoli M. & Quarta F., (2021). "Can we measure inflation expectations using Twitter?," Temi di discussione (Economic working papers) 1318, Bank of Italy.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). "On finding the natural number of topics with latent Dirichlet allocation: Some observations" In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 391–402). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). "Latent Dirichlet Allocation". Journal of Machine Learning research, 3(Jan), 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). "A density-based method for adaptive LDA model selection". Neurocomputing, 72(7–9), 1775–1781.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). "Accurate and effective latent concept modeling for ad hoc information retrieval". Information Retrieval Journal, 17(2), 175–198.

Đukić, M. (2022). "Assessment of inflationary pressures using newspaper text analysis". Working Papers Bulletin. National bank of Serbia, September 2022. 41–66.

Griffiths, T. L., & Steyvers, M. (2004). "Finding scientific topics". Proceedings of the National Academy of Sciences of the United States of America, 101 Suppl 1(suppl 1), 5228–5235.

Kešelj V. and Šipka D. (2008) "A Suffix Subsumption-based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources". INFOTHECA, Journal of Informatics and Librarianship, vol. IX, no. 1–2, pp. 23a–33a, 21–31 May 2008.

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). "News-driven inflation expectations and information rigidities". Journal of Monetary Economics, 117, 507–520.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990), "Indexing by latent semantic analysis", Journal of the American Society for Information Science.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

# PAYMENT SYSTEMS MIGRATION TO THE ISO 20022 ELECTRONIC MESSAGING STANDARD

Ivan Radanović

© National Bank of Serbia, March 2024

Available at www.nbs.rs

The views expressed in the papers constituting this series are those of the author(s), and do not necessarily represent the official view of the National Bank of Serbia.

Payment System Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

## Payment systems migration to the ISO 20022 electronic messaging standard

Ivan Radanović

**Abstract:** The paper aims to analyse the projects of payment systems migration from the current ISO 15022 to the new ISO 20022 standard globally and in Serbia. One of the main project objectives is to facilitate cross-border payments, still largely characterised by high costs, low speed and insufficient transparency. This objective has been acknowledged globally as testified by the G20 roadmap designed in October 2020. National central banks are implementing their own migration projects based on keeping up with good practices and operating in accordance with the most up-to-date standards. The National Bank of Serbia also aims to achieve the compatibility required for potential connection with other payment systems (e.g. TARGET services of the European Central Bank) and connection to the SEPA geographical scope. New electronic messages are up to three times larger and structured in a way to offer greater flexibility, accommodation to economic conjuncture and complex requirements of AML/CFT, KYC, fraud prevention regulations, and the possibility for an almost one hundred percent straight-through processing rate. The analysis combines descriptive, comparative and case study methods to present in detail the characteristics of payments systems as the fundamental public infrastructure, payment trends, as well as the phenomenon of the electronic messaging standard and the XML pattern as the syntactic basis of the ISO 20022 standard. The paper also looks into the experiences of international payment systems and their operators, migration methods in the SWIFT network, as well as the work of the SWIFT central service for translation of MT and MX messages. Potential characteristics of the future software platform of the National Bank of Serbia for the NBS RTGS and NBS Clearing payment systems are also discussed in the paper. Payment systems migration will be completed in November 2025. As for the SWIFT network, the coexistence period started in March 2023 when messaging was possible under both standards. The NBS, as the operator of the payment systems which will switch to the new messaging format, will enable the coexistence of two messaging formats until the end of 2024 as one of the measures for ensuring the continuity of their work.

**Key words:** migration, ISO 20022 standard, electronic messages
**[JEL Code]:** E42, E58, F30, F33, G20

## Non-technical summary

In the prior period, most payment systems in the world used the ISO 15022 standard as the basis for electronic messages exchanged by participants: central banks, banks, non-banking payment service providers and other payment system operators. This is the SWIFT MT message format most often modified by payment system operators primarily to adapt to local market needs. Despite various shortcomings of MT messages which became clear during the years of their use, e.g. in cross-border payments, this implies certain fragmentation and impossibility to optimise the exchange of information among financial institutions. Amid expanding, more globalised and connected financial markets, the need of the financial community for improvement of electronic messages is rising.

The focus of efforts made at improving cross-border payments is the switch to the new, ISO 20022 standard of electronic messages in the MX format. These messages can contain several times richer and more structured information, which significantly reduces the potential for misinterpretation, at the same time increasing the success of straight-through data processing. A wealth of information will be used in several financial industries, not only the payment area, implying greater interoperability among diverse market participants. This should additionally enhance the quality of information about financial transactions, financial products and services for which it will be used, as well as help meet ever stricter regulatory requirements as the complexity of the payment services market increases.

The transition to the new standard is a global project headed by the International Organization for Standardization (ISO), Society for Worldwide Interbank Financial Telecommunication (SWIFT*)* and operators of the largest international payment systems, primarily the European Central Bank. The official transition deadline in the SWIFT network is November 2025 while the two message formats coexistence period started in March 2023. This means that as of March 2023 all institutions are required to have the technical capacity to receive the MX format messages, while as of 2025 they will be required to send messages in this format as well. The new message format is currently used in more than 70 countries. In November 2025, SWIFT will decommission MT messages.

The paper aims to present the characteristics of this migration, first in introductory sections on payment systems and trends. The analysis then observes electronic messages, differences between various formats and finally the ISO 20022 standard, as well as its syntactic basis, a standardised XML-based syntax for textual messages. The last section provides an overview of international migration experiences, interoperability of payment systems in the coexistence period, as well as domestic payment systems. Out of eight payment systems in the Republic of Serbia, the National Bank of Serbia (NBS) is the operator of six – NBS RTGS System, NBS Clearing System, NBS Interbank Clearing of FX Payments, International Clearing of FX Payments, DinaCard Clearing System and IPS System. As the last-mentioned system has operated under the ISO 20022 standard since 2018, it is the NBS RTGS System, NBS Clearing System, NBS Interbank Clearing of FX Payments, International Clearing of FX Payments that will migrate to the new standard. The scope of this paper covers the potential characteristics of the software platform which, once the migration is concluded, will be the basis for the operation of the NBS RTGS and Clearing payment systems, as important systems entirely concerning domestic payments in dinars. A comparative overview will be presented of the currently used MT ISO 15022 messages and their potential MX ISO 20022 equivalents. Also, this paper exemplifies the NBS's support to all payment system participants during the migration project.

## Contents:

# 1 Introduction

A payment system is a set of systems for transferring monetary assets, which facilitates money circulation. It is an indispensable part of the public infrastructure as well as a precondition for monetary policy implementation and economic activity. Instruments and procedures enabling the above also make a part of payment systems. As this is a structure with a significant economy of scale and a unique natural monopoly, the most frequent model is the one where central banks are payment system founders, owners and operators. Models differ not only in management structure but also in other criteria depending on payment system characteristics. Their heterogeneity concerning various issues – such as ownership of settlement assets, purpose of establishment, the requirements to be satisfied, etc. – conditions the diversity of their operations.

The principal activity of payment systems is to enable payments, as the transfer of funds from a payer to a payee after which mutual obligations of both parties are discharged. It is precisely because of this dynamic evolution of payment, which has moved away from the use of cash in the last twenty or so years, that the significance of payment system operators is growing. Digital payments are continuously developing owing to the constant interaction between payment service providers, ICT development and the needs of end users. These needs are diverse, but they all have in common the constant striving towards acceleration, simplicity and security of the payment process. This refers to both internal and international payment traffic.

Cross-border payments, as payments between two financial entities from different countries and different payment systems, are methods for executing international financial flows. To ensure economic growth, globalisation demands speed from capital. This implies the perfection of payment methods between parties in the same but also in different payment systems. Payments are made by exchanging a great number of different electronic messages between financial institutions, their clients and payment system operators. This takes place through computer networks which can be international (e.g. SWIFT) and local, i.e. tailor-made to individual payment systems. An electronic message is a set of information exchanged in a transaction by financial institutions: information on institutions, client information, information about amounts and types of used payment instruments, etc. These data have a certain structure established by conventions. These conventions are named standards. Formulating, confirming and promoting standards in the area of financial services, and many other standards, falls under the mandate of the International Organization for Standardization (ISO).

The bulk of payment systems worldwide, including those operated by the NBS (except the NBS IPS System, compatible with ISO 20022), have so far used SWIFT MT messages based on the ISO 15022 standard as the principal and most frequently used standard for the exchange of financial messages between financial institutions. National payment system operators, including the NBS, mainly opted for a minor or major modification of the content of this message format, to meet the needs of their local markets. This increases their usefulness in internal use by domestic customers but compromises it in international exchange, particularly when national payment systems are increasingly striving towards integration stimulated by globalisation requirements. As message types are rather diverse, this calls for optimisation, simplification, and acceleration of payments.

The new ISO 20022 electronic messaging standard should be able to satisfy this need. This is an improved system of recording information in electronic messages whose structure, scope and technical rigour have been significantly improved in relation to the ISO 15022 version. Influential international organisations, private sector and operators of globally significant payment systems (the

European Central Bank, the Federal Reserves, etc. as well as the SWIFT network) are the hubs of global migration to the new electronic messaging standard, which should be completed by November 2025. This is a part of the G20 roadmap designed in October 2020 with the objective to facilitate cross-border payments. This should be implemented by solving persistent problems such as high costs, low speed and insufficient transparency of these payments. Central banks and financial institutions which are payment system operators as well as payment system participants must finish the migration by the set deadline, when the SWIFT network will decommission the MT message format. This particularly refers to entities using the SWIFT network as the primary communication channel. As it entails substantial software-hardware adjustments, this is a technically and organisationally demanding process.

The paper is aimed at analysing this migration. A combination of descriptive, comparative and case study methods provides an overview of all relevant aspects of the migration ahead of payment systems in Serbia and their participants. They concern the nature of the payment systems, their technical and organisational features, digital payment trends, electronic message phenomena and syntactic basis of the ISO 20022 standard.

The paper is structured in the following way. Below the readers will be informed about the basic concepts concerning payment systems such as payment, clearing, settlement, etc. Several classification methods for payments systems are presented indicating the basic features of three out of six payments systems operated by the NBS: NBS RTGS System, NBS Clearing System and NBS IPS System. In addition to the mentioned three, the NBS also operates the following systems: NBS Interbank Clearing of FX Payments, International Clearing of FX Payments and DinaCard Clearing System. Payment systems are the key financial infrastructure benefiting all stakeholders – from government through banks to end users of payment services. Hence, it is important to also introduce the future trends of payment service provision – accelerated digitalisation and integration with advanced technologies such as artificial intelligence. The third part of the paper presents the phenomenon of electronic messages and their standardisation, with an overview of the XML pattern as the syntactic basis of the ISO 20022 standard. The fourth part of the paper concerns the migration to the new standard both worldwide and in our country – with a focus on the NBS RTGS System and NBS Clearing System as important payment systems.

As the intention of this paper is to symbolise the onset of payment systems migration in the Republic of Serbia, as the project of first-class importance for both the operator and participants – the paper ends without a conclusion in the narrow sense of the word. This is a result of its somewhat specific role. It is primarily informational. This means providing the main set of information, which is not final, but is useful for all participants in the project and for the success of its implementation. An equally important idea behind this paper was to exemplify the NBS's constant support to all participants in the payment systems of which it is the operator.

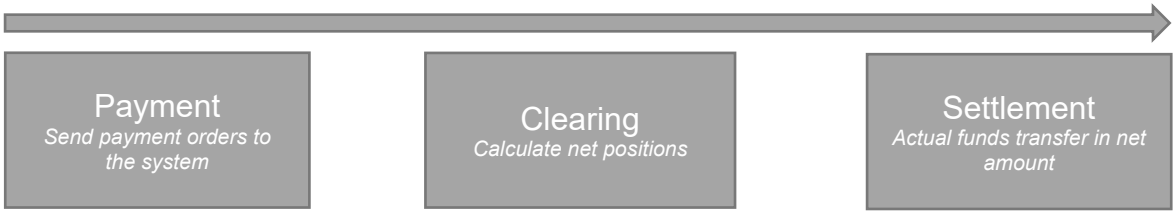
## 2 Payment systems: the bloodstream of modern-day economy

There are several ways to define payment systems. Most often they are defined as a complete set of instruments, intermediaries, rules, procedures and interbank funds transfer systems which facilitate the circulation of money in a country or currency area (ECB, 2010). It can also be said that the above set guarantees such circulation of money (Banco de México, 2024). Put more concisely, a payment system is an organised arrangement for transferring monetary value between two parties (Nakajima, 2011). According to our Law on Payment Services, a payment system is "a system for

the transfer of funds between its participants with written and standardised procedures and rules for the processing, netting and/or settlement of transfer orders, applied to all participants in the system" (RS Official Gazette, Nos 139/2014 and 44/2018).

Therefore, a transfer of monetary value (payment) is a transfer of funds which discharges an obligation on the part of the payer vis-à-vis a payee (ECB, 2010). The mutual obligations between the two parties are discharged once payment is done. This is referred to as settlement. Settlement is preceded by processing and clearing. Clearing is the process of determining the difference between the sum of all transfers received and the sum of all transfers sent on individual accounts of payment system participants – the calculation of the net position. If the difference is positive, the participant has a net positive (credit) position. If it is negative, it has a net negative (debit) position. During settlement, participants with a negative net position send an amount equal to their net position into the system. This is then summed up and distributed to participants with a positive net position. Though in theory clearing is one of payment system processes, not all payment systems function according to this principle.

Figure 1 **Functioning of payment systems with clearing**




According to: Nakajima, M., "Payment System Technologies and Functions", p. 6.

Payment systems are a vital part of each country's public infrastructure because they affect the speed of economic flows, the results of monetary policy implementation and transmission, and the costs and liquidity of participants. They must therefore stay reliable. In fact, no economic activities are possible without the transfer of money (Nakajima, 2012),1 and no transfers are possible without payment systems.[2]

## 2.1 Classification of payment systems

In order to get a broader picture of payment systems, we will classify them according to several criteria. The classification criteria differ depending on the information to be provided by a specific classification. Payment systems can therefore be classified by: operator, method and frequency of settlement, value of transactions settled in this system and settlement assets (Figure 2).

According to the operator, there are central bank payment systems and private payment systems. The former are owned and operated by central banks, which means that central banks have the role of payment system operators.[3] In the Republic of Serbia, the NBS is the operator of the following payment systems – NBS RTGS, NBS Clearing System, NBS IPS System, NBS Interbank Clearing of FX Payments, International Clearing of FX Payments and DinaCard Clearing System. Private

[1] A well-known example is the reaction of Alan Greenspan, chairman of the Federal Reserve Board, when he heard of the September 11 attack. As he later said, his immediate concern was not the inflation rate or the unemployment rate, but the "Fedwire" – the largest payment system in the US which transfers more than USD 4 trillion a day.

[2] Except in case of cooperation with correspondent banks.

[3] Such as the American Fedwire, the ECB's TARGET2, Japanese BOJ-NET etc.

payment systems are founded and operated by private market players, most often associations of banks and other financial institutions (Clearing Houses). In professional literature, these payment systems are referred to as "clearing systems". In Serbia, the Association of Serbian Banks is the operator of two systems – ASB Cheque Clearing System and ASB Direct Debit Clearing.

According to the method of settlement, there are net and gross settlement systems. In net settlement systems, system participants' positions are in the net amount. Each participant's net position is calculated either relative to any other participant (bilateral settlement) or relative to all other participants (multilateral settlement). The NBS Clearing System is based on the multilateral settlement principle. In gross settlement, payments are made in the full amount, by order of entry into the system. The classification of payment systems by frequency of settlement is closely related. There is real-time settlement and designated-time settlement. The first group includes Real-time Gross Settlement Systems – RTGS, and the other Designated-Time Net Settlement (DTNS) or Deferred Net Settlement Systems (DNS). The clearing cycles in the NBS Clearing System are initiated four times during the system's working hours, which means that it belongs to the latter group.

According to the criterion of value of transactions processed through the payment system, we differentiate between Large Value Payment Systems (LVPS) and Retail Payment Systems (RPS). These designations – "large-value" and "retail" – should not necessarily be taken literally, but may also be understood as the type of payments for the processing of which these systems were designed. As a type of payment, retail payments are payments between bank clients – payment system participants, and refer to the transfers and purchase of goods and services by natural and legal persons and entrepreneurs. In other words, these are payments between two natural persons (P2P), between one natural and one legal person (P2B) or commercial transactions between legal persons and entrepreneurs (B2B) (Bech & Hancock, 2020). We should also add payments to government accounts – P2G and B2G. Such payments have relatively low value, but are very numerous. In theory, the systems processing retail payments do not need to carry out settlement on the same day when the payment order was received. They are most often doing so, however, in order to minimise operational risks, and this is the practice adopted by the payment systems operated by the NBS. Wholesale payments, on the other hand, are payments between financial institutions – payment system participants: payments to settle securities and foreign exchange trades, and other interbank transactions. While there are significantly fewer compared with retail payments, their value is much larger. The average daily number of payments in the NBS RTGS System – the system for real time gross settlement of large-value payments[4] – is around 840,000, and the average value of an individual payment is around RSD 936,000.[5] Conversely, the NBS IPS system for retail payments processes around 185,000 payments a day, and their average value is around RSD 11,000.

The final classification is made according to ownership of settlement assets. There we differentiate between payment systems which settle in central bank money – money in payment system participants' accounts with the central bank – and in commercial banks money.[6] The finality

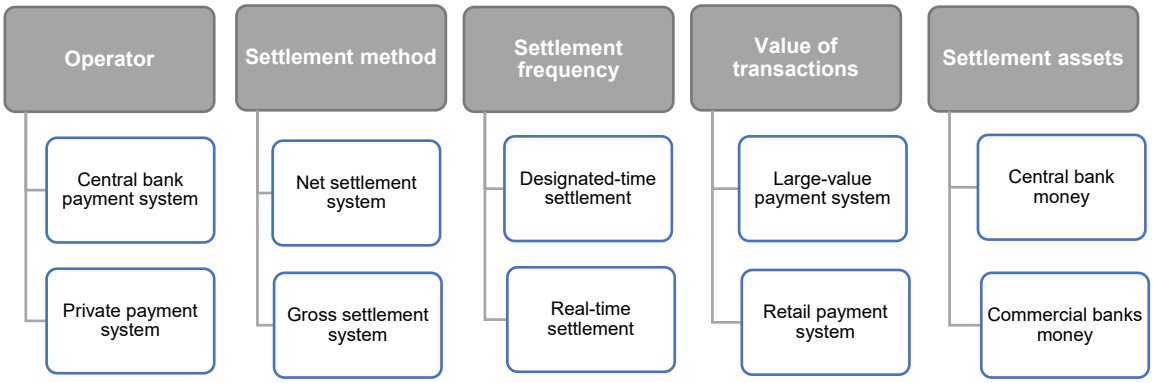
[4] Large-value payments are payments exceeding RSD 300,000. Not all payments in the NBS RTGS system are large-value payments in the wholesale sense, however. Retail payments are also effected in real time in this system. Their individual amounts are lower than RSD 300,000 but they are made in bulk. This is a service which aims to advance retail payments in the country. The NBS IPS system processes payments below RSD 300,000.

[5] "General indicators of RTGS and Clearing system in 2023", available at: https://www.nbs.rs/export/sites/NBS_site/documents-eng/platni-sistem/statistika/rtgs/stat_23.pdf

[6] It is often forgotten that natural persons' balances in banks do not represent their ownership but only, in a strictly theoretical sense, an "obligation" of the banks towards their clients.

of settlement liquidity risk, associated with balances with commercial banks, is minimal in the case of central bank balances. For this reason, the relevant global standards recommend that such systems be settled in central bank money (BIS, 2001).

Figure 2 **Classification of payment systems**




Source: author's analysis.

## 2.2 Significance of payment systems

The above is reflected in the fact that the ECB's RTGS system settled the equivalent of annual GDP of euro area countries in six days of operation, i.e. around EUR 13 trillion (ECB, 2023). The equivalent payment system in the UK settled an average of over GBP 720 billion each working day, equivalent to the UK's GDP every three days (BOE, 2022). The NBS RTGS system, as the most important payment system in the Republic of Serbia, processes around EUR 6.7 bn daily. The annual value processed through this system is approximately 29 times the country's GDP.

This is not the only reason why payment systems are difficult to overvalue. Each country relying on a stable macroeconomic environment, where the public has confidence in the financial and banking systems as hubs of economic activity, has to make sure that payments can be made at all times.

The functioning of payment systems costs money. The social cost of payment systems includes all costs sustained in order to make payments using all payment instruments[7] less charges on the transactions between system participants. Examples include the costs for the production of banknotes, the time it takes households to obtain cash and make payments or the costs of processing payment orders or cheques at bank tellers (Krüger & Seitz, 2014). Taken together, these are the net costs of payment systems. This concept is useful, but with limited potential. The estimates of net costs of payment systems, both aggregate and by individual transaction, depend on the characteristics of the national economy and, most of all, on the frequency of using specific payment instruments. This is a key factor for the economy of scale, necessary for payment system efficiency. It depends on the scale of the acceptance network for cashless payments, and even on the interest rates used when calculating the opportunity costs of going to an ATM or a bank teller. For all these reasons, the studies were not consistent even when they referred to the same countries, the same methodology

[7] Cash, payment cards, credit transfers, etc.

and similar time periods. Since costs depend on too many factors, different studies are not easily comparable. In spite of efforts, there is only limited knowledge and information available for making valid comparisons across countries (Schmiedel et al., 2012).

Increasing technological complexity and, by extension, efficiency of payment systems (e.g. through the introduction of instant payments systems) delivered the liquidity required for economic activities to take place. The technological development of payment system is hence critically important to the functioning of globalised markets. In other words, an efficient payment system can promote economic growth and deliver long-term productivity improvements which are prerequisites to elevating living standards (CEBR, 2022). The more efficient a payment system is, the fewer costs it involves.

The outbreak of the economic crisis in 2008 spiked interest in the development of macroeconomic forecasting tools, particularly those based on monetary and financial data. This has created scope for the analysis of data on payment instruments which, being associated with economic transactions, represent a unique source of information for the purposes of short-term economic activity forecasts. Though this connection was known more than a century ago, thanks to economist Irving Fischer (1912) who wrote that the "equations [quantitative theories of money] mean that the currency paid for goods is the equivalent of the value of the goods bought", the renewed interest was encouraged by the ideas of the so-called New Monetarists.[8] In this regard, based on earlier research on the relation of GDP, private consumption and the statistics of payment instruments – mostly cards, a close relationship was established between the time series on retail payments and main economic aggregates.

That the relationship between aggregate production and payment system statistics of a country is not only descriptive has been confirmed by a recent Indian study (Rooj, Sangupta, 2020). The researchers concluded that large-value payment systems positively impact economic growth, and that economic growth also leads to an increase in the value and volume of payments within these systems. They also found that an increase in RTGS payments leads to an increase in money supply and price level as indicated by the CPI. In other words, when the economy and incomes are rising (pushing up aggregate demand), people tend to indulge more in cashless[9] payments and thus enhance economic growth.

The main financial system participants are payment systems, central banks and commercial banks. Central banks play a vital role in modern-day payment systems. They operate these systems and carry out settlement for banks and other financial institutions. They also take active part in monetary policy implementation through open market operations. Hence a sound payment system is a prerequisite for successful monetary policy (Bech, 2008). Commercial banks use payment systems for operations in the markets of money, foreign exchange or securities on their own or on their clients' behalf. The impact of payment systems on finances and banking is reflected in the fact that banks are the key players in the payment services market, i.e. key payment system participants. So far, the banks' role was reflected in the performance of transactions at a significant and rising volume and in using the benefits of the economy of scale with a tendency of decline in costs. For banks, payment

[8] For more on this school of thought which, by contrast to traditional monetarism (and Keynesianism), relies more on macroeconomic analysis, see: *https://oxfordre.com/economics/display/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-397*.

[9] This can refer to cash payments, but because of the immanent risk of grey economy, cashless payments are a better criterion.

operations are an important source of stable income, and are not subject to capital requirements.[10] They are also an important source of information on clients' conduct and the base for applying the most advanced analytics (Živković, 2019).

Payment systems tend towards so-called natural monopoly. This is a situation where, in an economy of scale, maximum efficiency of production and distribution can only be achieved through a single supplier. The natural monopoly is characterised by enormous fixed costs and negligible marginal costs[11] (Varijan, 2014). Also, similarly to mobile telephony networks, payment systems are characterised by the so-called network externalities – the more participants in a payment system, the greater the value transmitted through the system and the more system participants and end-users it reaches.

In view of the above, central banks and the broader social community are clearly highly interested in ensuring the reliability and efficiency of payment systems.

## 2.3 Trends in payment services and payment systems

With the robust technological advances powered by the Internet and later also smartphones, the consumer habits, preferences and conduct have shifted in the past fifteen years from using cash to making digital payments. In parallel, the scope of Internet-based payment services widened, including e-banking or e-money payments.[12] The acquiring network is becoming modernised – ever more functional POS and ATM terminals are increasingly available, while the number of ATMs offering only cash withdrawal is dwindling. In addition to digitalised communication, digitalised purchase methods are another requirement of the technologically-mediated society, leading to higher demand for real-time retail payment instruments.

The general shift away from cash, cheques and credit transfers based on a paper order (payment slip), and towards cashless instruments, such as online credit transfers, cards or e-money, is particularly evident in the statistics on payment services provision. The average number of digital payments per person in the group of countries whose statistics are published by the Bank for International Settlements (*Red Book statistics*) increased from 179 to 332 in the 2012–2021 period.[13] Cash lost further popularity in the pandemic over fear of pathogens transmission via paper banknotes.[14] These trends are not universal, however, as the demand for cash, even in the richer countries of the capitalist centre, remains high. According to the last year's report of the Bank for International Settlement, currency in circulation as a share of GDP grew to an all-time high in the first year of the pandemic (2021), and still exceeds its pre-pandemic levels (Glowka et al., 2023). An important reason is that cash became significant as a store of value, which lead to its "hoarding" (Auer et al*.,* 2022).

[10] The scale of this income is best reflected in the fact that total net income from fees and charges in 2015 measured 65% of operating income of European banks.

[11] The establishing of a payment system involves very high initial costs independent from the number of payment system participants. Once a payment system has been set up, the additional cost of individual participants joining the system is negligible from the viewpoint of the owner of the payment system.

[12] The Serbian Law on Payment Services defines electronic money as electronically and magnetically stored monetary value as represented by a claim on the issuer of such money. It is not a synonym for balances in ordinary payment accounts. Like cash and account balances, electronic money is a means of payment and as such is regulated by the above Law.

[13] The list of countries is available at: *https://stats.bis.org/statx/toc/CPMI.html*.

[14] In the meantime, research revealed that banknotes and coins are no more dangerous than other frequently touched surfaces (Tamele, B., A. et al., 2021).

Regulators are adapting to the above trends. The central banks are upgrading their legal and regulatory framework: the key trend is the lifting of barriers to market entry for new, non-banking payment service providers (non-banking financial institutions – NBFI).[15] It is equally important that, by end-2023, over 60 countries modernised their payment systems by launching instant payment systems.[16] The world leader in the number of instant payments per inhabitant above 15 years of age is Thailand, with 276 payments a year. The most successful European country is Sweden, with 114 such payments (ACI, 2023). In the Republic of Serbia, the NBS IPS system was launched on 22 October 2018. In 2023, the system processed 67 million transactions or 11.25 instant payment transactions per inhabitant above 15 years of age. Though this type of statistics can differ greatly from country to country, depending on the dominant payment instrument and type of transaction (P2P, B2B, P2B or P2G), research shows that the acceptance of instant payments is followed by a general pattern of advancing technological innovation in payments (Bech et al., 2017). This means that instant payments are becoming standard.

An indication of future trends was given by the Global Payments portal (2024), which held in-depth discussions with leaders from financial institutions, businesses and international payment bodies, as well as from financial markets in periphery countries. The prevailing attitude is that, whatever direction global economy may take, market players are taking the initiative to accelerate and remove friction from the payment process, deploy the latest technologies to reduce fraud, and deepen their relationship with consumers of payment and other services. Artificial intelligence clearly takes centre stage: "It's not very often in one's lifetime that a technology like this [AI] comes along with such a wide variety of use cases," says Vanessa Colella, head of innovation and digital partnerships at Visa. Businesses do not have a single answer to how they intend to use it, but a half of SMEs surveyed are enthusiastic about the impact of AI on their business, as are 80% of large and multinational enterprises. Specific changes are emerging – according to the above research, the top three expected applications for AI are customer service (43%), fraud detection (43%) and marketing (40%). According to McKinsey consulting firm (2023), generative AI could add USD 2.6 trillion to USD 4.4 trillion in value to the global economy annually. Another important factor behind payment trends, even more than economic conjuncture, is open finance[17] and the resulting concepts such as banking as a service (BaaS). This is a set of related services provided by non-financial companies to their clients: from merchant m-applications with the internal account and payment functionality, to payment cards issued by the specific company or even microloans from online stores.[18]

Though non-traditional financial institutions have been allowed market access and cross-border payment dynamics are robust – reaching about USD 150 trillion in 2022, a 13 percent increase relative to 2021 – challenges remain (McKinsey & Company, 2023). They persist despite global communication networks which include thousands of financial institutions. Some of these challenges are: high costs due to chains of intermediaries, currency conversion, different working hours of payment systems, security risks, etc.

[15] According to the World Bank definition, these are financial institutions that do not have a full banking licence and cannot accept deposits from the public, but they do facilitate alternative financial services (World Bank, 2016). Such services, regulated in particular by the revised Directive EU 2015/2366 (PSD2), include payment initiation services and account information services.#

[16] Lipis Advisors. (2023). *Overview of instant payments landscape today*.

[17] Open finance means a market structure where banking and other financial services are provided not only by traditional institutions (e.g. banks), but also by regulated third-party institutions in order for consumers to have greater control of their finances and for the market to be more competitive.

[18] A well-known example is the Buy now, pay later *(*BNPL) concept, whereby the bill is divided into several equal instalments and no interest is charged.

Payment systems function through the exchange of electronic messages between their operators and participants. An electronic message is a set of structured information exchanged by parties to a financial transaction. In the discharge of their activities, banks and other financial institutions exchange enormous quantities of data between themselves and with their customers. Such exchange is reliant upon both the sender and the receiver of a message having a common understanding of how to interpret the information and data they receive. This has become more and more important as computerisation has advanced and human participation diminished. To overcome this the financial industry has created standards on how to organise the data they want to exchange in structured formats (syntax) and meanings (semantics).

These conventions are known as standards. The structure and content of electronic messages varies greatly from area to area, indicating the importance of standardisation. The hubs of these efforts are the International Organisation for Standardisation and the SWIFT global network, as important levers of global standards acceptance. Standardisation of electronic messages is one of the key methods to improve the functioning of payment systems and facilitate cross-border payments.

Consistent data quality is also important for financial regulators. This is key in a situation where it is necessary to aggregate the data of different financial institutions from different markets. Without some uniformity in the generating and processing of information, conclusions based on their analysis cannot be reliable either. The lack of high-quality and accessible data can be risky, particularly at times of financial crises. Risk management systems and short-term financing mechanism can then face difficulties in identifying key information, terms and elements of the messages (Office for Financial Research, 2012).

## 3 Electronic message standards

Information is becoming an increasingly important factor in modern-day finance. Its scope, complexity and diversity are expanding hand in hand with the economy, and this is a trend which is expected to last. In other words, the more important information is, the greater the need for it to be high-quality. That is why the standardisation of electronic messages is so important (Powell, 2014).

The history of message standardisation is longer than half a century and it is associated with the banking markets which used to be highly fragmented at that time. The first message formats were created in the USA, where local formats such as BAI (Bank Administration Institute) or its upgraded variant BAI2 appeared. CFONB and other formats were created in France and STD18 in the UK. The first major step towards unification occurred with the MT messages of different generations formulated by the SWIFT global network in the early 1970s. Until these messages appeared, only telex international transfers were possible, and they came with numerous shortcomings. Payment took up to 4 days and the messages were unstructured – to the extent that transfers were described in entire sentences which the recipient was supposed to read. Because of the possibility of human error, this resulted in a low rate of successful processing. Communication multiplied after the introduction of the international MT format. Throughout 1979, the SWIFT network processed around 10 million messages (IR, 2021). It currently processes around 45 million messages daily.

Though technically obsolete, many older formats are still in use. Many countries also use internal, proprietary standards based on the international standard but adapted to the needs of the

local market.[19] This apparently illogical coexistence is illustrated by the famous quip that "the great thing about standards is that there are so many of them". It is illogical because a multitude of standards is a contradiction in itself, but only at first glance, as the potential costs of rapid harmonisation greatly outstrip the benefits.

When it comes to SWIFT MT messages, most payment systems use the ISO 15022 standard, introduced at the beginning of this century to replace the ISO 7775 standard. Once the latter standard was excluded, electronic messages became more informative and easier to structure, with greater automation. This is reflected in the fact that the straight-through processing (STP) rate increased from around 65% to 90% and more, which is the standard today. STP means that all sub-processes relating to payments and the associated information flows are computerised, minimising manual human work.

Almost a half of all messages exchanged every day within the SWIFT network are structured according to the ISO 15022 standard. Message standard ISO 8583 is dominant in the card systems, and the volume of their daily exchange is measured in hundreds of millions. Different financial activities use different standards, and one of the objectives of introducing the ISO 20022 standard was precisely to include them all. Presented below are elements of two messages relating to an ordinary credit transfer from a payer to a payee, but according to different standards: SWIFT ISO 15022 and the internal US standard. In this hypothetical example, ABC bank from Belgrade (BG), Republic of Serbia (RS) wishes to transfer USD 15,000 at the payment order of its payment service consumer DEF, head office in Terazije 25, Belgrade, on 1 March 2024, from his account 123456789.

Figure 3 **Comparison of electronic messages for credit transfer**

| MT103 | FEDWire Proprietary |
|---|---|
| :32A: 2410301USD15000<br>:50F:/123456789<br>1/DEF<br>2/Terazije 25<br>3/RS/Beograd<br>:52A: ABCRSBG | {1520}20240301xxxxxxxyyyyyyyy {2000}0000015000<br><br>{5000}D123456789 DEF.*TERAZIJE 25*BEOGRAD*<br>SRBIJA* {5100} BABCRSBG |

Source: author's analysis.

From the above, we see that the same data (payer's name, payer's payment service provider, head office, etc.) are structured differently, with different field elements. According to one standard, the bank identification code is specified in field 52A, while according to the other standard, it is entered in field 5100, and the content of the two fields differs. As these are just examples of parts of messages, which in practice may contain as many as twenty fields, the problem of communication becomes even more complex.

This means that communication using different standards can create problems with end-to-end automation. This problem's potential increases with the length of communication chains inside constantly expanding and increasingly interconnected financial markets.

Communication is significantly impeded if different semantics and syntax are used.

1. In this sense, *semantics* means the "vocabulary" used by different areas. Incompatibilities may arise if different words refer to the same thing or, even worse, if the same

[19] This is also the case with the messages used in the payment systems in the Republic of Serbia.

word means two different things. For example, a payer in a transaction may be denoted in English as the Ordering customer, Payer, Payor, Payment Originator, Initiator, etc. Initiator can mean two things, depending on the payment instrument. In a credit transfer, the initiator is at the same time the payer, while in a direct debit the initiator is the payee based on the payer's authorisation.

2. In this sense, *syntax* means the structure of information, that is the "language" of communication. If the message recipient does not understand the syntax used, he will not understand the content of the message either. Standard can be seen as an agreement on the content of electronic messages and the meaning and structure of data within messages. Each business model (e.g. payment) must contain elements (e.g. payer, payee, payment service provider, address) described so as to be intelligible for humans and software, enhancing interoperability and the potential for automation (SWIFT, 2015). Therefore, the standard defines *what* is communicated by the message, and the syntax *how/what with* this is communicated.

The standards are not adopted in order to be a source of competitive advantage for a specific organisation or company, but in order to offer added value to all financial system participants. This is reflected in the more comprehensive and higher-quality content of messages which will meet future market requirements as well, such as more data on cross-border payments (for users), more data on payment service consumers in order to improve the offer (for banks), easier assessment of payment orders' compliance with legal regulations (for regulators), etc.

As a natural solution for state-of-the art instant payment systems, ISO 20022 is a genuine catalogue with more than 800 different messages covering different business areas (Figure 3). Within each area, there is a large number of individual messages denoting different business relations, such as those between end-users and payment service providers, between payment service providers, and between payment service providers and payment system operators.

Figure 4 **Overview of business areas where the new electronic message format is used**

| | |
|---|---|
| **acmt:** Account management | **colr:** Collateral management |
| **auth:** Authorities communications | **setr:** Securities trade |
| **caaa:** Acceptor to acquirer card transactions | **secl:** Securities clearing |
| **catm:** Card terminal management | **sese:** Securities settlement |
| **pacs:** Payments clearing and settlement | **semt:** Securities management |
| **pain:** Payments initiation | **seev:** Securities Events |
| **camt**: Cash management | **tsin:** Trade services initiation |
| **remt:** Payments remittance advice | **tsmt:** Trade services management |
| **fxtr:** Foreign exchange trade | **reda:** Reference data |

According to: ISO20022 Business Areas, p. 2.

The use of common rules, within harmonised categories, makes it possible to focus organisational, material and other resources on value drivers, i.e. activities where the greatest value is created, instead of dissipating these resources on formatting, conversion and reinterpretation of data because of different standards. This does not mean that one standard is not translated into another, because it is, but the intention is to make this practice obsolete by transitioning to the ISO 20022 standard. Before we analyse the above standard, it is useful to look at its syntactic basis.

## 3.1 XML: Syntax of the new standard

The most widely used syntax within the ISO 20022 standard is XML – eXtensible Mark-up Language.[20] Since its design in 1998, this has been the most popular standard for marking up documents and messages online and offline, as it is also used in computer applications,[21] vector graphics, mail exchange, voice mail systems, etc. It defines a generic syntax used to mark up data with simple, human-readable tags (Harold & Means, 2006). This is the most robust, reliable and flexible document syntax ever invented.[22]

An XML document contains text. It is built from text content marked up with text tags between symbols "<" and ">". There are start-tags and end-tags. The use of the document is very flexible, but there are strict rules about where tags are placed and how they are written. This means that all XML documents must be well-formed, i.e. conform to a set of rules, such as:

- Every start-tag must have a matching end-tag;
- Elements may not overlap;
- Attribute values must be (single or double) quoted;
- Comments or processing instructions may not appear in tags.

These rules are unbreakable, which makes parsing XML documents easier and writing them a little harder. Below is an example of the simplest yet complete XML document:

```
<person>
      Petar Petrovic
</person>
```

The element in this document is "person". The element is delimited by the start-tag <person> and the end-tag </person>. This syntax is case-sensitive, so if you open an element with a <person> tag, you cannot close it with a </PERSON> tag. Everything between the start-tag and the end-tag of the element (exclusive) is called the element's content. The content of this element is "Petar Petrović".

Below is a somewhat more complex document which contains one *person* element, but also data on the name, surname and profession:

```
<person>
      <name_and_surname>
            <name>Petar</name>
            <surname>Petrovic</surname>
      </name_and_surname>
      <profession>intern</profession>
</person>
```

Within this syntax, tags give meaning and structure to the information which is exchanged. Data are included as strings of text. In the above example, the string is "Petar", "Petrović", "intern".

[20] The other syntax supported by the ISO 20022 standard is JSON.

[21] For example, MSWord files are saved in the .docx format, where the final "x" denotes XML, or extensible.

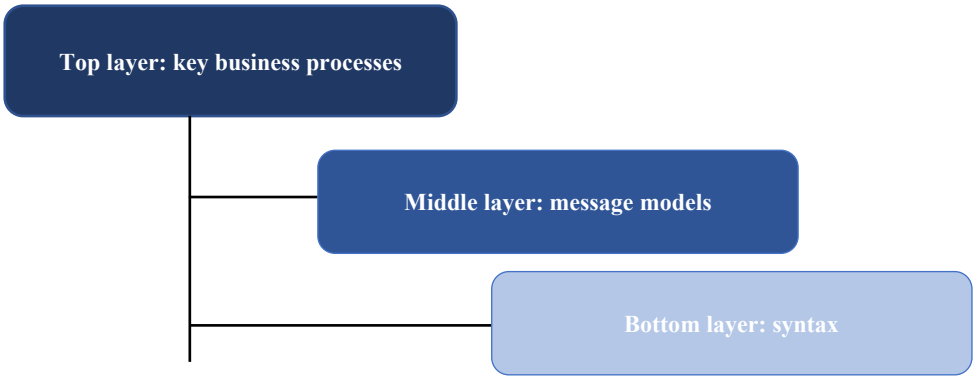
[22] XML is a descendant of SGML, the Standard Generalized Markup Language, invented at IBM in the 1970s and adopted as ISO 8879 standard in 1986. SGML was extremely powerful and was used in the U.S. military and government, strategic industries, and in other domains that needed ways of efficiently managing technical documents that were tens of thousands of pages long. However, SGML was very complicated — its official specification is over 150 very technical pages. It is so complex that almost no software has ever implemented it fully. Its simplification and trimming began in 1996, but so as to retain most of its useful functionalities. The result was XML 1.0. (Ibid, p. 9).

In the area of payments, these documents are much longer and more complex and the names of the elements are officially defined and unique. For example: <FinInstnId> for the identification code (BIC) of the payment service provider; <Nm>, <Adr> for data on parties to a transaction; <CCy> and <Amt> for transaction currency and amount, etc. These and many other elements are included in the three areas covered by the payment systems, shaded in Figure 3: payment clearing and settlement (pacs), payments initiation (pain) and cash management (camt). It is of key importance that element names are unique and unchangeable in all payment systems where the ISO 20022 standard is used, which is optimal for machine parsing of a large number of messages.

## 3.2 Characteristics and use of ISO 20022 in the area of payments

Business processes are the methodological basis of this standard. Information is needed in order for business processes to take place. It is organised in business components (message models) which, in turn, consist of elements described by syntax. In other words, this is a concept of three separate layers.

Figure 5 **Three layers of the ISO 20022 standard**




Source: author's analysis.

Business processes are relevant activities in the financial industry. They are also referred to as business domains and include operations with bonds, payment systems, trade services, foreign exchange market and card services (ISO, 2024). Message models are segments where specific activities take place (Figure 3), while the bottom layer contains the structure of the specific message model.

There are multiple differences between the MX messages based on the ISO 20022 standard and MT messages based on the ISO 15022 standard. The former have a better defined structure, with dedicated, structured elements for each detail important for the transaction. This facilitates parsing and enhances interoperability of payment systems across regions. Specifically, MX messages have element hierarchy with nested elements for logical grouping of data. For example, if we are interested in information on the payee (<Cdtr>), we will find it the lower branching order under name (<Nm>) and address (<PstlAdr>), then under address you can have further structured elements for street number (<StrNm>), postal code (<PstCd>) etc. This was not possible in the MT message format, as can be seen below.

Figure 6 **Simplified overview of a credit transfer according to the old and the new standard**

| MT103 | pacs.008 |
|---|---|
| {1:F01ABNANL2AAXXX1234012345}<br>{2:O1031511010606UBSWCHZHGXXX0000013085010549S}<br>{3:{108:UHBMT103001}{121:360f1e65-90e0-44d5-a49a-92b55eb3025f}}<br>{4:<br>:20:494931/DEV<br>:23B:CRED<br>:32A:011521USD10551,50<br>:33B:USD10551,50<br>:50K:/122267890<br>BIODATA GMBH<br>HOCHSTRASSE, 27<br>8022-ZURICH<br>SWITZERLAND<br>:59:/1234567890<br>CUBA SPORTS BAR GRILLE<br>1234 OCEAN DRIVE<br>90099 LA<br>:71A:SHA | <?xml version="1.0"?><br><Document<br>xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"<br>xmlns="urn:iso:std:iso:20022:tech:xsd:pacs.008.001.08"><br>FIToFICstmrCdtTrf><br><br><GrpHdr><br><MsgId>494931/DEV</MsgId><br><CreDtTm>2021-15-01T00:00:00</CreDtTm><br><NbOfTxs>1</NbOfTxs><br><SttlmInf><br><SttlmMtd>INDA</SttlmMtd><br></SttlmInf><br></GrpHdr><br><CdtTrfTxInf><br><PmtId><br><InstrId>4949431/DEV</InstrId><br><TxId>NOTPROVIDED</TxId><br><EndToEndId>NOTPROVIDED</EndToEndId><br><UETR>360f1e65-90e0-44d5-a49-a92b55eb3025f</UETR><br></PmtId><br>…<br><CdtTrfTxInf><br>…<br><Cdtr><br><Nm>Cuba Sports Bar & Grille</Nm><br><PstlAdr><br><StrtNm>Ocean Drive</StrtNm><br><BldgNb>1234</BldgNb><br><PstCd>90099</PstCd><br><TwnNm>Los Angeles</TwnNm><br><CtrySubDvsn>CA</CtrySubDvsn><br><Ctry>US</Ctry><br></PstlAdr><br>…<br></Cdtr> |

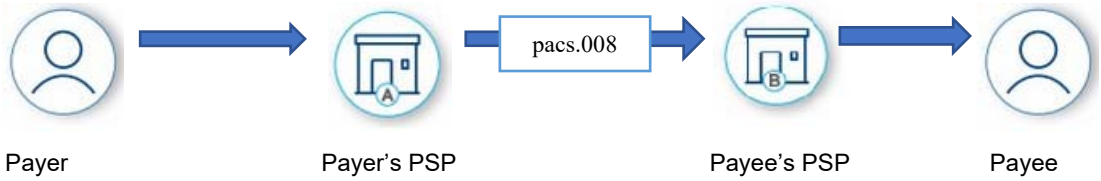
Source: McConnell, S., McAuliffe, R. (2020).

Data on the payee are shaded yellow and green in the two message formats. The difference in the structure of information is visible at first glance. There can be no confusion when interpreting the pacs.008 MX message, as each piece of information has special, dedicated elements. In the example on the left, incompletely structured data on the payee's address may result in "CUBA" being interpreted as the name of the country which is under financial sanctions, and not as a part of the street name. The reference number of the payer's payment service provider is shaded red in the old and the equivalent data is shaded light blue in the new message format. Whereas in the old format 16 characters could be included in field 20, in the new format the payer sends the unique message identifier <MsgId>, payment order identifier <InstrId>, end-to-end processing identifier <EndToEndId> and data on the payment <TxId>. Each of the elements can have 35 characters.

Three message models are used for payments: PACS, PAIN and CAMT. They cover the entire communication chain – customer to bank (pain), bank to bank (pacs) and different notifications for banks, clients and the payment system operator (camt). Different secondary functions are included such as bank account management, managing limits in bank accounts, managing direct debit authorisation, regulatory reporting, etc. The coverage is not final, and is constantly expanding to

respond to market requirements.[23] Given below are the descriptions of some of these models[24] with a simplified[25] illustration of the communication flow. The first message model is exchanged between financial institutions, i.e. payment system participants, to clear the transaction information and settle the related funds. Examples are:

1. **pacs.008** – FI to FI Customer Credit Transfer;
2. **pacs.003** – FI to FI Customer Direct Debit;
3. **pacs.009** – Financial Institution Credit Transfer;
4. **pacs.004** – Payment Return.




The PAIN message group supports the initiation of a payment from the ordering customer to an account-managing financial institution. This model is designed to support the flow of messages exchanged between customers – payers or payees – and their financial institutions – banks or non-banking payment service providers. Examples are:

1. **pain.001 –** Customer Credit Transfer Initiation;
2. **pain.008 –** Customer Direct Debit Initiation.


As the domestic instant payment system (IPS) which functions based on the ISO 20022 standard processes credit transfers and generates associated messages, pacs.003 and pain.008 messages are currently not used.[26] Those that are in use are regulated according to their purpose and type by means of the Guidelines on Messages used in the NBS IPS System.[27]

Finally, CAMT messages are designed for communication between payment system participants and customers of payment service providers in order to inform them on the status of an initiated transaction, on the arrival of incoming transactions, to communicate their accounts statements, etc.

[23] In that sense, perhaps the greatest advocate for ISO 20022 standard adoption, at least in Europe, is the Single Euro Payments Area.

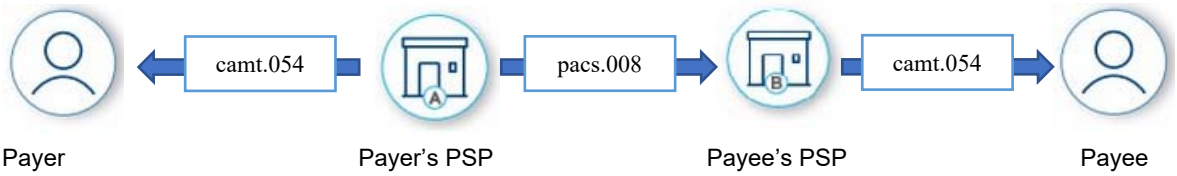
[24] A much greater number of messages are actually used, but their extensive description would be out of scope of this paper.

[25] Simplified primarily because it excludes payment system operators as the central hubs of such diagrams. The operators send instruction to participants and control the validity of the messages which, if inadequate, are returned to senders – participants. Payment in not executed in that case.

[26] These two message models refer to direct debits processed in the Direct Debit Clearing, a payment system operated by the Association of Serbian Banks.

[27] Available at: *https://www.nbs.rs/export/sites/NBS_site/documents/propisi/propisi-ps/Pravila_IPS_2018_prilog7.pdf*.

1. **camt.052** – Bank to Customer Account Report;
2. **camt.053** – Bank to Customer Statement;
3. **camt.054** – Bank to Customer Debit Credit Notification.




Though these messages do not exhaust the list of messages on payments, they are the most frequently used ones. An additional benefit of the ISO 20022 message standard is to reduce the diversity of the messages used by the previous standard. In other words, several MT messages are now functionally merged in fewer[28] MX messages, as can be seen in the table below.

Table 1 **Equivalence of some messages under the old and the new standard**

| MT ISO15022 | message name | MX ISO20022 |
|---|---|---|
| ***financial***[29] | | |
| MT101 | Request for Transfer | **pain.001** |
| MT102 | Multiple Customer Credit Transfer | **pacs.008** |
| MT103 | Single Customer Credit Transfer | |
| MT103/2 | Payment Return | **pacs.004** |
| MT202 | | |
| MT104 | Direct Debit and Request for Debit Transfer Message | **pacs.003** |
| MT200 | Financial Institution Transfer for its Own Account | **pacs.009** |
| MT201 | Multiple Financial Institution Transfer for its Own Account | |
| MT202/202COV | General Financial Institution Transfer | |
| MT203 | Multiple General Financial Institution Transfer | |
| MT205 | Financial Institution Transfer Execution | |
| ***non-financial***[30] | | |
| MT900 | Confirmation of Debit | **camt.054** |
| MT910 | Confirmation of Credit | |
| MT920 | Request Message | **camt.060** |
| MT940 | Customer Statement Message | **camt.053** |
| MT950 | Statement Message | |
| MT941 | Balance Report | **camt.052** |
| MT942 | Interim Transaction Report | |

Source: Citibank, N. A. (2021).

Though this is not an exhaustive list of messages, it is noticeable that, out of nine existing categories of MT messages, only the categories used in payment systems – 10xx, 2xx and 9xx – are preparing for migration.

[28] There are changes in the opposite direction as well. For example, different information sent by MT n99 messages is now included in camt.030, camt.031, camt.032, camt.038 etc.

[29] Financial messages are messages followed by a transfer of money.

[30] Their exchange has a communication purpose and does not result in a transfer of money.

# 4 Migration to the new standard: opportunities and challenges

An important detail about the entire ISO 20022 migration is that it is not a new standard – the International Organization for Standardization published it back in 2004, just as the global transition to ISO 15022 was completed. The fact that it took 21 years from its design to becoming a global solution is explained by the fact that even the best standards do not gain widespread acceptance until the needs of the market reach a sufficient level of complexity. That they have reached that level is evidenced not only by the omnipresent globalisation but also by the growing need for interoperability and, perhaps above all, the digitalisation of payment services. In the last quarter century, it has been marked by the development of internet traffic, smartphones, the emergence of electronic and mobile banking, real-time payments, contactless payment instruments, and similar – as well as the processing power of all necessary hardware components.[31] Therefore, it is not even remotely true to consider that the coronavirus was the key factor in the digitalisation of payment services, though it did highlight its importance, because this diminishes the complexity and duration of digitalisation.

Digitalisation of payment services requires improvement of payment systems and application payment solutions for end users. This, in turn, means the application of modern technologies which will ensure the execution of payments in a simple and efficient way through digital channels and the fulfilment of regulatory and other requirement by banks, especially in the field of preventing money laundering (AML), financing of terrorism (CTF) and fraudulent payments (Fraud Prevention). Also, the emergence of new and the rising complexity of existing business models and payment services will increase the resource and technical requirements for banks, since MX messages have up to three times the capacity for information transfer (Deutsche Bank, 2019).

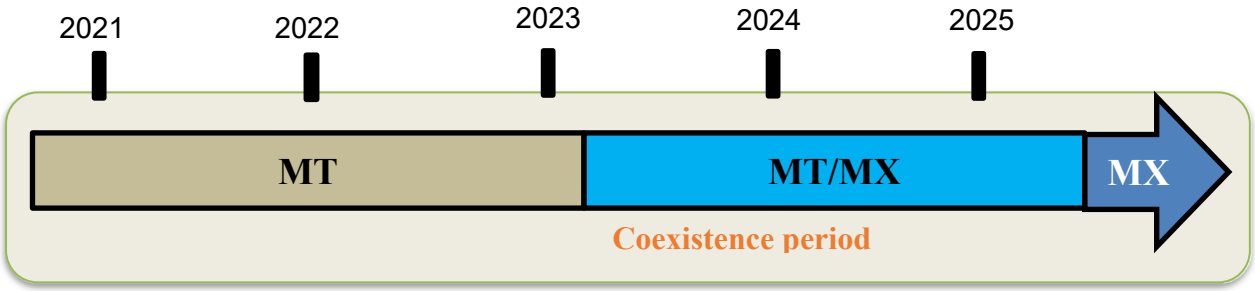
## 4.1 Beginnings of migration

The first explicit plans for global ISO 20022 transition were formulated in 2016, when SWIFT, together with representatives of the world leading banks and financial infrastructures, established the High Value Payments Systems Plus (HVPS+) workgroup. Its objective is the development of global guidelines for the application of this standard, which would contribute to further automation, greater transparency and content of information flows. In addition, SWIFT also encouraged the formation of a group for cross-border payments and reporting (Cross Border Payments and Reporting Plus, CBPR+) to develop guidelines for the harmonisation of cross-border payments and reporting. Those guidelines became the basis of the central SWIFT message translation service (see next section).

The global transition to the new standard takes place in two ways. On the one hand, the leading market infrastructures are in the midst of transition or have already completed it for internal payments. This particularly applies to retail payment systems and payment areas such as SEPA. On the other hand, SWIFT officially enabled the use of ISO 20022 for cross-border payments and within wholesale payment systems in March 2023. This means that participants who are ready can start exchanging data in this way, while those who are not can still use the old MT format. In other words, March 2023 saw the start of the so-called period of coexistence, i.e. parallel use of both formats. This phase will last until November 2025, when SWIFT will decommission MT messages. Below is a general migration plan.

[31] It is also a prerequisite for the so-called scalability, the ability to increase the volume of transactions without increasing costs.

Figure 7 **ISO 20022 migration plan (*CBPR+*)**




Source: author's analysis.

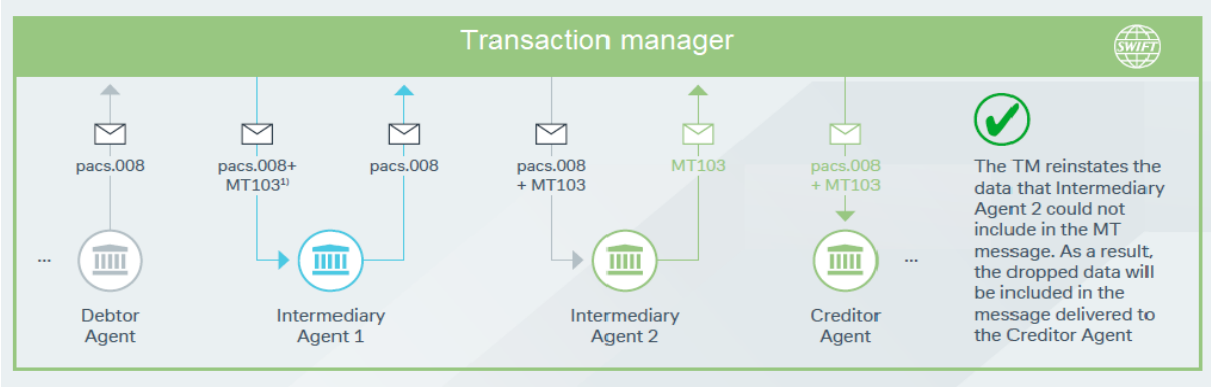
## 4.2 Ensuring interoperability in coexistence

Perhaps the biggest challenge of the entire migration is the coexistence of two standards. Throughout its duration (2023–2025), banks and other payment system participants have room to adjust their hardware–software solutions. As the new standard is adopted, many old and new format messages will remain in circulation. In order to prevent negative consequences, inconsistencies and data spillage along communication chains, SWIFT provides a mechanism which bridges such differences. In other words, it provides a kind of reinterpretation, i.e. translating messages from the old to the new standard and vice versa. This is crucial in the cases where in one transaction, i.e. in the same chain of communication, there are institutions which transitioned to the MX format and those that still use MT. This mechanism is called "Transaction Manager" (TM) and its objective is to ensure the interoperability of participants. It is a prerequisite that participants adopt the CBPR+ guidelines.

The TM mechanism works by "capturing" the initial MX message and preserving its integrity. In the next step, if the financial institution – intermediary in the chain uses the old MT standard, the mechanism ensures that such a message is delivered to it, but "enriched" as much as possible with the content from the original MX message, which the MT format would not initially recognise. Any subsequent participant using the MX message will receive the same message without reduced quality. This means that, from the beginning to the end of a transaction, the possibility of important data being lost, disintegrated or disappearing under a new record in the same fields is greatly reduced.

Generally, everyone benefits from this mechanism: institutions which have not yet migrated still receive a richer content of messages, and those which have can use the full potential of the new standard without any frictions, bottlenecks or other negative influences from other participants.

Figure 8 **A simplified TM mechanism operation model**




Adjusted according to: Deutsche Bank (2022), p. 16.

## 4.3 International experiences

Viewed in a wider context, the migration to the new standard is part of the G20 roadmap formulated in October 2020, one of the objectives of which is to facilitate cross-border payments (FSB, 2020). This programme aims to solve persistent problems which are characteristic for cross-border money transfers, such as high costs, low speed and insufficient transparency. The heterogeneity of electronic message standards is recognised as the main factor behind these problems (BIS, 2022). Although the global ISO 15022 standard was already dominant at the time, cross-border payments were made difficult by the fact that countries continued to use their own, internal message formats based on the aforementioned standard, but modified according to the local needs. Though this creates added value for the local market, it takes it away from the global market. The reason is that any cross-border payment in that case requires translation from the internal to the official SWIFT MT format, which can cause partial data loss and slow down the entire procedure.

The complete migration of the European infrastructure to ISO 20022 was perhaps best described by Christian Westerhaus, Head of Cash Products, Cash Management at Deutsche Bank, saying that "this is not just 'another IT project', but the most impactful payments industry undertaking since the introduction of SEPA" (Deutsche Bank, 2019). More specifically, SEPA has resulted in EUR 21.9 bn cost-savings per annum for euro area members (PwC, 2014). Also, liquidity rose by close to EUR 230 bn and over 970 thousand man-years[32] were unlocked by simplifying the netting and settlement processes between hundreds of payment infrastructure hubs within the European Union. Similar effects are expected from this project. The new electronic messaging standard is currently used in more than 70 countries (BOE, 2024).

Roughly speaking, the largest global payment systems and market infrastructures have migrated to the new standard. In the practice of the western financial institutions so far, two approaches to migration have crystallized:

1) in one step, the so-called big-bang approach and

2) gradual, so-called like-for-like approach.

The US Federal Reserve System began the project in 2018, when it officially proposed that the nation's largest payments system, Fedwire, accept the new standard in three phases starting in 2020. The acceleration of cross-border payments and rising competition between banks through the offer of new, more diverse payment services to clients were highlighted as the expected effects. Currently, this RTGS system uses an internal message format which supports different types of communication, so participants can send both financial and non-financial messages. During the migration, payment system participants are expected to develop their own application solutions. The initial plan to end the migration was at end-2023 (Sullivan & Cromwell, 2018). However, this plan was later modified, so the current plan places the final migration in March 2025. As shown below, the decisions of the payment systems were also influenced by the actions of other leading infrastructures. The Federal Reserve had its reasons too – in the summer of 2023, an instant payment service called FedNow was introduced in the States. The complexity of this project in the US is determined by the fact that the two main infrastructures – Fedwire and CHIPS[33] – adopted different

[32] A man-year is a unit of measure which indicates the amount of work done by an individual in the course of one year.

[33] Clearing House Interbank Payments System – This is the second largest system for processing large-value payments in the US. While the Fedwire mainly processes payments between US banks, CHIPS is also used for international payments. In 2021, a record 204.5 million transfers, with an average value of USD 4.9 mn, were processed through Fedwire. Around USD 1.8 bn is processed daily through CHIPS.

timelines, suggesting the so-called like-for-like approach to migration. The latter should complete its migration to the new standard in April this year. Until then, these two payment systems use internal, mutually different but compatible message formats.

As for the European Union, the key payment systems switched to the new standard in March 2023 – these are T2, an RTGS payment system operated by the European Central Bank, and the private payment system EURO1, operated by the European Banking Association (EBA)[34]. Although the original plan envisaged that the migration would take place at end-2022, technical and geopolitical circumstances postponed it. Basically, this was the so-called big-bang migration, where, once the participants tested the messages after a defined period, the migration was carried out in one day. On the other side of the Channel, the Bank of England completed the migration in June last year, also after several delays.

This does not imply that migration is a simple process – on the contrary. The European Central Bank is precisely one of the institutions which are the most responsible for multiple migration delays. The reason is that the European Central Bank should have been the first to make such a significant change, so that it would set an example for other banks and payment system operators. However, the European Central Bank had its own projects in the same period, primarily the consolidation of its TARGET2 large-value payments system and its TARGET2-Securities (T2S) system. This means that each time a European project requiring compliance with ISO 20022 was postponed because participants were not ready, SWIFT also postponed the migration. The biggest delay occurred in October 2022, when the European Central Bank decided to extend the deadline for launching the new consolidated RTGS system[35] to March 2023. The same delay was announced by SWIFT. For its part, the operator of Lynx, Canada's high-value payment system, decided to follow their lead and postponed the migration from November 2022 to March 2023. The initial deadline defined by SWIFT for the migration was November 2021.

While SWIFT and the rest of the western business community are trying to ensure safe migration of most payment systems and banks to the new standard, it is interesting that in PR China the largest RTGS system has been operating according to ISO 20022 since 2013. An important benefit experienced by Chinese banks is that the new standard can transmit information recorded in traditional, Mandarin characters, which was not possible with the MT format.[36] More importantly, China's CIPS cross-border payment system, with more than 1,400 participants from 113 countries, established in 2015, also uses this standard (CIPS, 2024). This is also the case with the Chinese system for instant payments Internet Banking Payments System IBPS[37], with more than 200 banks as participants. The Chinese leadership has expanded the scope of use of this standard beyond the field of payments, thus the Chinese and Japanese central securities registers are jointly based on ISO 20022 (Asianbondsonline). Despite the early adoption of advanced international standards, Chinese payment systems also use internal message formats, i.e. they do not use the SWIFT network for operations (Fintech Futures, 2020).

[34] Since its establishment in November 2018, the TIPS system for instant payments of the European Central Bank has been based on the ISO 20022 standard.

[35] Its name has now been abbreviated to T2. Until the consolidation, the T2S system was already using the new message standard. For more details, see: *https://www.ecb.europa.eu/paym/target/consolidation/html/index.en.html*.

[36] In the Serbian market, these messages will be able to convey information written in Cyrillic.

[37] For the sake of precision, the Chinese RTGS system and the aforementioned IBPS form the basis of what is called the China National Advanced Payments System, CNAPS (BIS, 2012).

## 4.4 Payment systems in the Republic of Serbia

In our country, the historical backbone of the payment system consists of the NBS RTGS System and the Clearing System, two important payment systems, which have been operating successfully since 2003. High reliability, both in terms of availability for participants and operational performance, was made possible by the technical upgrades that took place in the meantime, the last in March 2022. In terms of high operational performance of domestic payment systems, the contribution of the NBS IPS system for instant payments is indispensable, with 67 million processed payments last year and with almost 100% availability.

Like numerous financial institutions and operators of payment systems in the world, the NBS prescribed that domestic payment systems use an internal message format based on ISO 15022. As for the communication channel for data exchange between the system and the participants, an internal star-type computer network is used (the operator is in the centre, and the participants are the spokes) based on the IP protocol, whose communication resources are administered by the NBS at the network hub. There is also a connection to the SWIFT network, which may be used as an alternative communication channel. Payment system participants have the possibility to use both networks, so that interoperability within the system is preserved. Currently, all participants use the internal network.

The existence of an internal network made it possible for the messages used to differ to a certain extent from those used in the international network. This kind of independence is sustainable as long as there is technical support from the SWIFT community for the MT message format. In addition to the fact that the NBS follows the best business practices regarding the use of appropriate guidelines and standards in the field of payments[38], an important reason for starting the project of ISO 20022 migration, together with the participants of the payment systems of which it is the operator, is that the MT message format will be decommissioned at the end of the next year.

In addition to the above, it is important to migrate to a new updated set of MX messages which conform to the CBPR+ and HVPS+ guidelines for other reasons as well. First of all, it complies with the upcoming project of Serbia's joining the SEPA geographical scope, where banks communicate according to ISO 20022. Also, it is expected that a new participant will access the NBS RTGS System – Euroclear bank. As it uses the SWIFT network as the primary communication channel in its operation, there will be no other option but to use the MX message format after 2025. The importance of this project goes beyond payment systems, as the financial settlement of transactions related to the securities issued by the Republic of Serbia will be executed through the Euroclear bank account in the NBS RTGS System.

The project of migration to the new standard will be implemented by November 2025. By the end of 2024, the NBS, as an operator, will start the migration process of its own application platform which supports the operation of the NBS RTGS and Clearing System, and provide an appropriate test environment in which participants will test the new message format. Although the project is at an early stage, it is possible to sketch some of the likely features of the future application platform.

Since the so-called period of coexistence is ongoing, the implemented system will support parallel operations with both MT and MX messages, but in such a way that one participant (both in the NBS RTGS and in the Clearing System) may use only one format until it is able to fully transition

[38] As evidenced by the NBS IPS system for instant payments, based on the ISO 20022 standard from 2018.

to the newer one. Different message formats used by payment system participants will be harmonised thanks to the converter, which will perform conversion from one to another message format. This will be provided by the NBS as an operator in due time and it will be available to participants in the transition period. The platform itself will, as is the case now, support the operation of both payment systems through a single application. Technically, this means that it will enable real-time gross settlement, but also batch clearing. As such, it will imply the necessary scalability, that is, the capacity to support growing volumes of payments in the future, the possibility of operating in multiple currencies, as well as connectivity with direct and indirect participants of payment systems. As is the case now, the participants will be able to choose to operate either in the NBS internal network or in the SWIFT network at any time.

Given that the NBS RTGS System will be fully compliant with the HVPS+ guidelines, it will provide the basis for interoperability with other payment systems which operate under same principles, including the T2 payment system of the European Central Bank.

Below is the range of potential MX messages which will be used in domestic payment systems after a successful migration, as well as their MT equivalents. As mentioned, the messages which are currently in use at home are modified in relation to the MT messages prescribed by SWIFT, which means that the use of a certain number of fields is mandatory according to domestic regulations[39]. Those fields which are also in the official SWIFT specification, the use of which is not mandatory, may be used by payment system participants, but their content is not controlled and does not affect the execution of transactions. Since parallel operation with two message formats will be provided as an important measure to ensure the continuity of the payment system operation, of which the MT format contains a smaller volume of information – the use of the MX format will have certain limitations. It will be eliminated when all participants completely transition to the newer message format.

Table 2 **Potential MX messages in domestic payment systems and their MT equivalents**

| MX ISO20022 | Message description | MT ISO15022 |
|---|---|---|
| **pacs.004** | Payment return (revocation of a previously received payment order) | **MT103 \| MT202** |
| **pacs.008** | Individual credit transfer for the account of end users of payment services | **MT103** |
| **pacs.008** | Group credit transfer for the account of end users of payment services | **MT102** |
| **pacs.029** | Batch message for settlement of external payment systems within RTGS | **MT971** |
| **pacs.009** | Transfer of funds between participants in the payment system (wholesale). Tracks the transfer of funds from account to account in RTGS. | **MT202** |
| **pacs.010** | Direct debit between participants in the payment system | **MT204** |
| **camt.050** | Credit transfer between participants in the payment system | **MT202** |
| **camt.051** | Debt transfer between participants in the payment system | **MT202** |
| **pacs.028** | Request for information on payment status | **MTn95** |
| **camt.007 \| camt.087** | Request to change the priority of an incomplete transfer message | **MTn95** |
| **camt.008 \| camt.056** | Request for revocation of an incomplete transfer message previously sent to the system | **MTn92** |
| **camt.018** | Download business day notifications | **MT999** |
| **camt.019** | Return business day information | **MT999** |
| **camt.025** | Response to request or error notification | **MTn96** |
| **admi.002** | Message rejection notice | **MT996** |

[39] Guidelines on the Format and Purpose of Electronic Messages Exchanged in Payment Operations. Available at: *Microsoft Word - Uputstvo o formatu i nameni 2009 _2_.doc (nbs.rs).*

| MX ISO20022 | Message description | MT ISO15022 |
|---|---|---|
| **admi.004** | Password change notification | **MT996** |
| **pacs.002** | Payment status notification (MT196 or MT296) | **MTn96** |
| **camt.052** | Answer to the inquiry about the account status | **MT986** |
| **camt.052** | Account status | **MT941** |
| **camt.053** | The final statement with all the details related to the specific account during the current business day. It concerns a payment that has been settled. | **MT940 \| MT950** |
| **camt.054 CR** | Account credit confirmation. It is used after the related transactions have been completed within the RTGS. | **MT910** |
| **camt.054 DR** | Account debit confirmation. It is used after the related transactions have been completed within the RTGS. | **MT900** |
| **camt.060** | Request for account balance information or account statement. With one message, it is possible to request several reports related to several accounts. | **MT920** |
| **camt.060** | Account status inquiry | **MT985** |
| **camt.998** | Free format message | **MTx99** |

# 5 Conclusion

A fast and smooth flow of capital is needed for national and international financial markets to work successfully. One of the most important prerequisites for that is the optimisation of information exchange between financial institutions and the operators of the payment systems in which they participate. This primarily concerns electronic messages which are exchanged between different participants in large volumes and on a daily basis, following each transaction. An electronic message is a set of structured information providing necessary knowledge about elements of the transaction – parties, transaction amount, end users of payment services, etc. The exchange of electronic messages has always been a part of automation trend, which means that the share of human work is decreasing, and the importance of computer data processing is increasing.

To make this processing as successful as possible, especially in the light of globalised markets and rising volume of financial transactions, it is crucial that financial institutions produce and exchange harmonised information. The information must also be of higher quality. That is why the financial industry is constantly formulating conventions – standards about the structure and meaning of this information. Financial markets are currently transitioning to using the latest ISO 20022 standard of electronic messages in the MX format, which replaces the previously used MT format messages, based on ISO 15022.

The objective of this paper is to analyse these migrations. There are several reasons why payment systems decide to transition – more content and better structured messages, greater flexibility and adaptability of data to regulatory requirements, but also better automation of their processing. The analysis began with an introduction on payment systems and trends, followed by an overview of electronic messages as the basic method of interbank communication in modern payment systems, the difference between various message formats and, finally, the messages according to the mentioned new standard and its syntactic basis. The final part concerns the payment systems operated by the NBS – primarily the NBS RTGS System and the NBS Clearing System, as important payment systems – which are also expected to migrate to the use of MX messages according to ISO 20022 by end-2025.

Although uncertain at the moment due to the early stage of the migration project, the architecture of the future application platform which will support the operation of domestic payment systems has several probable characteristics. In addition to technical support for the parallel use of both MT and MX message formats – whereby an individual participant can use only one of them in production

work – the system will enable the conversion of messages through a centralised or indirect mechanism. Both the RTGS System for the real-time settlement of large-value payments and the Clearing System for the settlement of group payments at a certain time will operate, as before, on a single platform – while International and Interbank Clearing of FX Payments will be located on a separate platform. Considering the current projects of the NBS, and thus the payment systems it operates and their participants, that platform should be as compatible as possible with the technical requirements of potential participation in the TARGET systems of the European Central Bank, and the SEPA geographical scope – including working with messages based on the ISO 20022 standard, in accordance with the HVPS+ and CBPR+ guidelines.

## Literature

"2024 Commerce and Payments Trends Report", website "Globalpayments", available at: https://www.globalpayments.com/commerce-payment-trends

ACI Worldwide & Global Data (2023). It's Prime Time for Real-Time. The Global Real-Time Payments Report.

Auer, R., Cornelli, G., Frost, J. (2022). "The pandemic, cash and retail payment behaviour: insights from the future of payments database". BIS Working Papers, No 1055.

Banco de México, "Introduction to payment systems, transfers", website of Mexico's central bank, available at: https://www.banxico.org.mx/payment-systems/introduction-to-payment-syste.html

Bank of England (2022). "Real-Time Gross Settlement system and CHAPS Annual Report 2021/22". Website of the Bank of England, available at: https://www.bankofengland.co.uk/report/2022/rtgs-and-chaps-annual-report-2021-22

Bank for International Settlements. (2001). "Core Principles for Systemically Important Payment Systems".

Bank for international Settlements. (2012). Payment, clearing and settlement systems in China.

Bank for international Settlements. (2022). Harmonisation of ISO20022: partnering with industry for faster, cheaper, and more transparent cross-border payments.

Bech, M. (2008). The Diffusion of Real-time Gross Settlement. In Haldane, A., Millard, S., & Saporta, V (Eds.). The Future of Payment Systems. Routledge.

Bech, M., Shimizu, Y., Wong, P. (2017). "The quest for speed in payments". BIS Quarterly Review.

Bech, M., Hancock, J. (2020). "Innovations in payments", BIS Quarterly Review.

"Business Model", website ISO20022, available at: https://www.iso20022.org/iso20022-repository/business-model

Centre for Economics and Business Research. (2022). The Economic Impact of Real-Time Payments. Presentation, available at: https://cebr.com/wp-content/uploads/2022/04/Real-Time-Report_v8.pdf

Citibank, N.A. (2021). "ISO20022: Survival Guide", website of Citibank, available at: ISO 20022: Survival Guide (citibank.com)

"CIPS Participants Announcement No. 92". Website of CIPS, available at: https://www.cips.com.cn/en/participants/participants_announcement/60849/index.html

"China leads the way in adoption of ISO20022 for payments", Fintech Futures, website, available at: https://www.fintechfutures.com/2020/03/china-leads-the-way-in-adoption-of-iso-20022-for-payments/

Deutsche Bank. (2019). Ultimate guide to ISO20022 migration.

European Central Bank (2023). "TARGET Annual Report 2022". Website of the European Central Bank, available at: https://www.ecb.europa.eu/pub/targetar/html/ecb.targetar2022.en.html

Financial Stability Board. (2020). Enhancing Cross-border Payments. Website of the Financial Stability Board, available at: Enhancing Cross-border Payments: Stage 3 roadmap (fsb.org)

Glowka, M., Kosse, A., Szemere, R. (2023). „Digital payments make gains but cash remains". CPMI Bried No 1.

"Harmonization and Standardization of Bond Market Infrastructure in ASEAN+3 (Phase 3 Report)". Website of Asianbondsonline, available at: https://asianbondsonline.adb.org/documents/abmg/abmf-sf2-chp4.pdf

"ISO20022: the new messaging standard for CHAPS and RTGS". Website of the Bank of England, available at: https://www.bankofengland.co.uk/payment-and-settlement/rtgs-renewal-programme/iso-20022

Kokkola, T. (ed). 2010. The Payment System. Frankfurt: European Central Bank.

Krüger, M., Seitz, F. (2014). Costs and Benefits of Cash and Cashless Payment Instruments.

Lipis Advisors. (2023). Overview of instant payments landscape today.

McConnell, S., McAuliffe, R. (2020). "ISO 20022 Migration: Industry Strategy, Insights and Best Practices". Citi TTS Online Academy.

Nakajima, M. (2011). Payment System Technologies and Functions. Hershey, PA: Business Science Reference.

Nakajima, M. (2012). "The Evolution of Payment Systems", The European Financial Review, available at: http://www2.bbwebarena.com/nakaji8/European%20Financial%20Review%20Feb2012.pdf

"Nonbanking financial institution", website of the World Bank. Available at: https://www.worldbank.org/en/publication/gfdr/gfdr-2016/background/nonbank-financial-institution

"On the cusp of the next payments era: Future opportunities for banks" (2023). Website of McKinsey, available at: https://www.mckinsey.com/industries/financial-services/our-insights/the-2023-mckinsey-global-payments-report#/

Powell, L. (2014). Office of Financial Research, U.S. Department of the Treasury, GS1 Global Forum, Brussels.

PwC. (2014). Economic analysis of SEPA. Website of Price Waterhouse Cooper, available at: https://www.pwc.com/gx/en/audit-services/corporate-treasury-solutions/assets/pwc-sepa-benefits-and-opportunities-ready-to-be-unlocked-by-stakeholders.pdf

Rooj, D., Sangupta, R. (2020). The Real-time impact on real economy – A multivariate BVAR analysis of digital payment systems and economic growth in India. Asian Development Bank Institute. ADBI Working Paper 1128.

Sullivan & Cromwell LLP (2018). "Federal Reserve Proposes ISO20022 Message Format for Fedwire Funds Service".

Schmiedel, H., Kostova, G. & W. Ruttenberg (2012), The Social and private Costs of Retail Payment Instruments: A European Perspective, ECB Occasional Paper Series No. 137, September.

Tamele, B., A. et al. (2021). “Catch me (if you can): assessing the risk of SARS-CoV-2 transmission via euro cash”. ECB Occasional Paper Series, no 259.

“The economic potential of generative AI: The next productivity frontier”, website of McKinsey Digital, available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

“What Businesses Need To Know About The SWIFT Messaging Format”, website of IR, available at: SWIFT Messaging Format: What Businesses Need To Know | IR

Varijan, H. (2014). Mikroekonomija. Beograd: Centar za izdavačku delatnost Ekonomskog fakulteta.

Živković, S. (2019). „Uticaj PSD2 regulative na budući razvoj bankarskog poslovanja”. Zbornik radova EkonBiz, str. 203–213.

**CIP** - Каталогизација у публикацији
Народна библиотека Србије, Београд

336.71

**WORKING Papers Bulletin** / editor-in-chief Jorgovanka
Tabaković. - [Engl. ed.]. - 2021, No. 1- . - Belgrade : National Bank
of Serbia, 2021- (Belgrade : Institute for Manufacturing Banknotes
and Coins - Topčider). - 30 cm

Ima izdanje na drugom jeziku: Зборник радова (Народна банка Србије) = ISSN 2787-3226
ISSN 2787-3315 = Working Papers Bulletin
COBISS.SR-ID 47831561