Accounting and Management Information Systems
Vol. 24, No. 4, pp. 651-672, 2025
DOI: http://dx.doi.org/10.24818/jamis.2025.04005

Enhancing credit risk forecasting using time-series models and R programming: A comparative analysis

Alexey Litvinenko[1, a], Anna Litvinenko[b] and Samuli Saarinen[c]

[a]*University of Tartu, School of Economics and Business Administration, Estonia*
[b]*Tallinn University of Technology, School of Business and Governance, Department of Business Administration, Estonia,*
[c]*Estonian Business School, Estonia*

Abstract

Research Question: Which of the four models (MLR, IV, ARIMA, ES) performed through R programming are more precise in credit risk forecasting based on financial ratios and possess improved robustness and generalizability as well as being less prone to overfitting?
Motivation: Traditional econometric models used in credit risk forecasting often suffer from overfitting, particularly when applied to financial ratio data with low variance. This challenge is especially pronounced in small sample settings typical of emerging markets or firm-level analysis. Exploring alternative, more adaptive models is necessary to improve forecasting reliability under such constraints.
Idea: This study evaluates whether transforming financial statement data into time-series ratio formats and applying ARIMA and ES models can enhance forecasting robustness and reduce overfitting, compared to conventional linear models.
Data: The historical panel data for 7 years from the annual reports of two production companies listed on the Baltic Stock Exchange, processed into financial ratios for forecasting 3-year horizons.
Tools: All four models are developed using R programming. Forecast performance is evaluated using Akaike Information Criterion (AIC) and other diagnostic measures for predictive accuracy, robustness, and resistance to overfitting.
Findings: ARIMA and ES models demonstrate superior predictive accuracy and robustness, especially in small-sample conditions. They respond better to structural changes and recent

[1] *Corresponding author:* School of Economics and Business Administration, University of Tartu, Narva mnt 18, Tartu, Estonia, email addresses: alexey.litvinenko@ut.ee; alx199@gmail.com.
Funding: there is no funding for this research.

Article History: Received August 11, 2025; Accepted 2 December 2025.
Accepted by: Nadia Albu.

data trends than Multiple Linear Regression (MLR) and Instrumental Variable (IV) models. This suggests ratio-based forecasting benefits from dynamic, time-sensitive modelling. The findings challenge linear assumptions and emphasize the value of time-series approaches in improving credit risk estimation under constrained data conditions.
Contribution: The study offers a replicable, R-based framework for robust credit risk forecasting, advancing time-series methods in small-sample financial analysis.

Keywords: R programming; Econometric Analysis; Financial Analysis; Credit Risk Forecasting.

JEL codes: C22, G32.

1. Introduction

In credit risk forecasting based on financial ratios, a persistent challenge is the tendency of Multiple Linear Regression (MLR) and Instrumental Variable (IV) models to overfit the data, thereby reducing the robustness and generalizability of their predictions. This problem stems from the inherent properties of ratio-based financial data. Many of the ratios consist of the same components in the numerator and denominator (variables) obtained from the financial statements, making the ratios have characteristics close to each other. With this type of variable, where the variance between points of measurement is low, different types of methods can be applied. To solve this problem with overfitting in this research, we used two alternative forecasting methods: ARIMA and Exponential Smoothing (ES).

This paper answers the following research question: Which of the four models (MLR, IV, ARIMA, ES) performed through R programming are more precise in credit risk forecasting based on financial ratios and possess improved robustness and generalizability, as well as being less prone to overfitting?

In this paper, we show how to transform the financial data from the annual financial statements into ratios and time series. We provide a step-by-step guide on cleaning and preparation of the data for MLR, IV, ARIMA, and ES forecasting models implemented through R programming.

The paper contributes to the ongoing academic discussion as it aims to expand Kowal's (2016) research on a foundational aspect, extending the multi-linear regression model's forecasting efficiency. It became evident that when applying the method to financial data, using absolute values as dependent variables (rather than financial ratios), the cash-based and accrual-based methods of financial analysis become prone to overfitting due to the small variance in the data. To address this issue, we used ARIMA and ES methods, treating the individual characteristics of the financial ratios as independent variables, with the ratio itself serving as the dependent variable.

Our finding suggests that through the newly developed and enhanced R-script, ARIMA and ES methods showed their efficiency in short-term forecasting, due to their ability to adapt fast to the changes in the data across the recent timeline. They proved themselves as efficient approaches to forecast the results of financial ratios and their possible shifts, even based on a small dataset, thanks to the adaptability of the models.

The research contributes to practice, assisting credit risk practitioners with a comprehensive step-by-step explanation of the enhanced R-script construction and proposition of its use in financial forecasting with different types of data. We plan to extend the research by exploring the characteristics of financial ratio forecasting; however, the primary focus of this paper is on developing a method for obtaining this information using R programming.

The literature review in Section 2 introduces important concepts and methods of forecasting. The research design and research method in Section 3 describe the logic behind the technical implementation of empirical research. Section 4 describes the step-by-step development and implementation of the forecasting methods. Section 5 discusses findings and concludes the study.

2. Literature Review

To build up sufficient theoretical grounds for the present paper, it is important to bring out several key concepts and works contributing to the knowledge base. This includes the background behind the credit risk forecasting, IV, MLR, ARIMA and ES models, which are presented further.

2.1 Credit risk in forecasting

Credit risk in forecasting has been a widely discussed and burning topic for decades, as it significantly affects decision-making in banking and finance. Nowadays, fintech and machine learning methods in credit risk forecasting play a crucial role in the development of scientific and practical knowledge in the domain. There are some areas of consensus and contention between past and modern research. The early models in credit risk forecasting, for instance, CreditMetrics and KVM, relied mostly on statistical methods like credit scoring and regression analysis (Crouhy *et al.,* 2000). Traditional methods like Altman Z-scores prioritise simplicity and interpretability while not always capturing the non-linear relationships and time-varying covariates necessary for credit risk prediction (Medina-Olivares *et al.,* 2023; Chen *et al.,* 2015). The largest advancement came with machine learning (ML) methods and techniques such as neural networks, support vector machines, and ensemble methods, which offered superior predictive accuracy and began replacing traditional approaches (Zhu *et al.,* 2019). On the other hand, ML models are

criticised for their "black box" nature, complexity of algorithms and lack of interpretability, which is critical for highly regulated financial industries and remains a subject of tense discussions of a trade-off between accuracy and interpretability in literature (Chen *et al.,* 2024). Hybrid models were introduced to answer the problem, combining the temporal focus on lifecycle models with forward-looking adaptability to capture the age-related dynamics and macroeconomic variations (Luong & Scheule, 2022). The significant challenge for credit risk modelling comes from the high dimensionality of financial data. However, techniques like PCA and ISOMAP are increasingly employed to combat this issue and improve computational efficiency without sacrificing accuracy (Chen *et al.,* 2015). It is also important to mention the regulatory shift under IFRS 9 and CELS standards from one-year expected loss models to multi-period frameworks with the integration of economic capital and loan loss provisioning (Luong & Scheule, 2022).

Despite significant advancements, there are still critical gaps underlined by the researchers. Kedia and Mishra (2024) highlighted the need for standardised interpretability frameworks, improved handling of data scarcity and imbalance, and integration of non-financial data (behavioural and ESG factors).

2.2 Instrumental Variable Model

IV models offer a solution to the endogeneity problem by providing a means to obtain consistent estimators (Wooldridge, 2013). An instrumental variable is correlated with the endogenous explanatory variable but uncorrelated with the error term, satisfying two critical conditions for effective IV use, which is particularly useful in situations where controlled experiments are not feasible (Angrist & Krueger, 2001). However, a weak instrument, poorly correlated with the endogenous variable, can lead to unreliable estimates; therefore, selecting appropriate instruments is critical, making testing for the validity of instruments an essential step in IV analysis (Bound *et al.,* 1995).

Recent studies contributed to these concepts. Horowitz (2011) highlighted the limitations of linear and other finite-dimensional parametric models in capturing the complexity of economic phenomena, suggesting that nonparametric methods can lead to substantive conclusions that differ significantly from those obtained using standard parametric estimators. Imbens (2014) reviews the assumptions behind IV methods, emphasising their relevance in both classical applications and modern contexts like randomised experiments with noncompliance. The study highlights the need for a theoretical foundation when selecting instruments. Swamy *et al.* (2015) argue that no instrument can simultaneously satisfy both exogeneity and relevance, challenging the viability of standard IV techniques. The critique emphasises the need for a comprehensive approach to model misspecification and measurement error. Andrews *et al.* (2019) highlighted the problem of weak instruments in linear IV

regression, particularly under non-homoscedasticity, and called for robust confidence sets and diagnostic procedures, underscoring the ongoing relevance of addressing weak instruments in empirical research.

2.3 Multiple Linear Regression

MLR is a statistical technique that allows isolating the effect of each independent variable while controlling for the influence of others, offering a clearer insight into the factors underlying economic phenomena (Greene, 2018). It has proved ability to handle complex, real-world situations where multiple variables interact is crucial in econometrics (Greene, 2018; Gujarati & Porter, 2009).

Recent studies have further explored the practical use and challenges of MLR discussing the interpretability of regression coefficients (Schielzeth, 2010), the correlation between variables using MLR, underscoring the method's preference due to the influence of multiple factors on resultative variables (Cruceru *et al.,* 2016) aligning with the econometric principle of considering various proportions of factors in each economic outcome. Kowal (2016) examined the efficiency of Ordinary Least Squares (OLS) estimators in simple linear regressions, providing a basis for evaluating forecasting performance in multiple regression. However, Varian (2014) highlighted that while traditional methods like regression remain effective, they may need adaptation for big data and complex relationships. This perspective is crucial for modern econometric analysis, where the volume and variety of data have expanded dramatically.

2.4 ARIMA

The effectiveness in modelling credit risk by the ARIMA approach through time series analysis demonstrates its versatility beyond traditional econometric applications, with the ability to accommodate both upward and downward movements in forecast outcomes (Tsay, 2010). By assigning equal probabilities to movements in both directions, the ARIMA model works as a good tool on the side of more simplified forecasting models such as MLR and ES, although it also has its limitations in capturing the fat tails and volatility clustering characteristic of credit default time series underlining the need to produce the forecasts in multiple different methods so that the best fitting for the specific data could be picked (Duffie *et al.,* 2009).

The Hyndman & Khandakar (2008) algorithm, implemented in the auto.arima and forecast functions of the "forecast" package in R, automates ARIMA model selection. It uses unit root tests and optimises model parameters by minimising the Akaike Information Criterion (AIC) and applying Maximum Likelihood Estimation (MLE). The AIC evaluates quality by balancing goodness-of-fit and complexity,

where lower AIC values indicate models that better capture the data with minimal information loss.

Unit root tests assess whether a time series is stationary. If not, differencing is applied to achieve stationarity, which is essential for accurate forecasting. This step is particularly valuable in credit risk assessment, where ARIMA models have been used to estimate credit default swap spreads and assess financial stability, demonstrating their practical relevance in real-world risk management (Cifter *et al.,* 2009).

2.5 Exponential Smoothing

ES is a forecasting technique in finance and accounting demonstrating its superior accuracy compared to other naïve forecasting methods (Hyndman & Khandakar, 2008), recognised for its effectiveness in removing the white noise on time series data (Kourentzes *et al.,* 2014) and its application in seasonal adjustment and trend forecasting (Taylor, 2004). ES models use previous forecasts as a basis and use forecast errors to continuously refine the outcome based on historical data (Hyndman & Athanasopoulos, 2018). In financial planning, these models are used for detecting significant changes in stochastic processes (Chatfield, 2003) and, with their adaptability, are used in predicting the volatility of financial returns, accommodating shifts in time series main characteristics (Taylor, 2004). Brooks and Buckmaster (1976) used this application in identifying systematic patterns in income time series, which have impacts on firm survival and income manipulation. Moreover, ES is widely applied for economic forecasting (Lin & Koo, 2007; Snyder *et al.,* 2002; Makridakis *et al.*, 1998). In summary, ES is an effective tool in finance and accounting, used for forecasting, inventory control, financial planning, and analyzing financial data trends. Its adaptability and accuracy make it valuable for handling various financial and economic time series data.

3. Research Methodology

3.1 Research design

This study focuses on assessing the credit risk forecasting capabilities of models developed using R programming within the domain of financial data analysis. Specifically, it focuses on transforming financial ratios derived from absolute values in financial statements into time series for credit risk forecasting purposes. These ratios are created from the absolute values retrieved from financial statements. The R script created in this paper employs several forecasting techniques, which we evaluate, namely ARIMA, ES, IV regression, and MLR. The research provides a detailed, step-by-step explanation of the development of the R script and the functions used. The script could be divided into the following sections: data cleaning

and filtering, followed by the creation of ratios, which are then transformed into time series to facilitate forecasting and subsequent analysis. Of the selected forecasting methods, ARIMA and ES form a pair that is tested against more traditional forecasting methods, IV and MLR models, to overcome the overfitting problem in credit risk forecasting, which is a very common phenomenon with IV and MLR models used in financial short-term forecasting. This research design is constructed to support and fit an explanatory investigation in the application of advanced econometric models in financial forecasting using a technical, data-driven approach.

3.2 Research method

The research adopts a quantitative methodological approach. Primary financial data were collected from the annual reports of two production companies listed on the Baltic Stock Exchange, forming a seven-year panel dataset. Panel data structures of this type enable the examination of both cross-sectional and temporal variation, thereby improving the robustness of empirical inferences (Baltagi, 2021). Following the data collection, the dataset was cleaned, structured and prepared for analysis. The transformed data were subsequently used to generate three-year forecasting results through a custom R script. The development and application of the script constitute a central methodological contribution of the study, as it enables the systematic evaluation of cash-flow-based forecasting within the selected firms.

The target population for this study is collected from large and medium-sized businesses operating in the Baltic region. From this population, the study uses Linas Agro Group as a representative case because of its long operating history, availability of consistent public financial disclosures, and its relevance as one of the largest vertically integrated agricultural groups in the Baltics. The empirical sample includes annual financial observations from 2016 to 2022, producing a seven-year panel dataset.

All numerical inputs used in the analysis were collected from publicly accessible sources, including the company's consolidated annual reports, audited financial statements, and investor materials. The information was manually extracted and transferred into a structured Excel dataset to ensure accuracy, comparability across years, and transparency in the construction of financial indicators. No external databases for processed data were used. The dataset, therefore, reflects strictly verifiable, publicly reported financial information.

The empirical analysis employs a set of 24 financial ratios, grouped into categories commonly applied in corporate financial analysis. These variables capture multiple dimensions of firm performance:

1. Leverage and Capital Structure:
   Total liabilities, debt-to-equity ratio, equity-to-assets ratio, and cash-flow-

to-debt ratio. These indicators allow assessment of long-term solvency and capital composition.

2. Coverage Measures:
   Interest coverage and cash interest coverage ratios which evaluate the firm's capacity to service its financing obligations through earnings and cash flows.
3. Liquidity Metrics:
   The current ratio and selected cash-flow-based liquidity indicators reflect short-term solvency and the firm's ability to meet immediate obligations.
4. Profitability Indicators:
   Net profit margin, return on assets (ROA), return on equity (ROE), return on capital employed, and related measures that describe efficiency in generating returns from assets, equity, and invested capital.
5. Cash-Flow Indicators:
   Operating cash flow, free cash flow, capital expenditure, cash-flow per share, and quality-of-income and quality-of-sales ratios. These variables capture the stability and reliability of earnings as reflected in actual cash generation.
6. Efficiency Ratios:
   Asset turnover, capital turnover, and cash turnover reflect the effectiveness with which the firm employs its assets to generate revenue and cash inflows.

The R script was constructed using several function packages, such as dplyr, ggplot2, forecast, and readxl. These packages are used for data cleaning, transformation and visualisation of the results. The credit risk forecasts produced through the R script are done through the following models: ARIMA, ES, IV regression, and MLR.

ARIMA and ES time-series forecasting models are applied to financial ratios after transforming the data into a time series. The ARIMA model is used to assess trends and seasonality in the data, while ES handles more recent data points to predict future trends with less noise. IV and MLR traditional econometric models are applied to assess the relationships between dependent financial variables and independent financial ratios. The MLR model is used to handle multiple predictors, while IV regression addresses endogeneity issues within the financial ratio data. Robustness checks and validation methods are used to see the magnitude of overfitting and ensure the accuracy of forecasts through statistical metrics like the Akaike Information Criterion (AIC). By comparing the results of the time-series models with traditional regression methods, the study aims to evaluate the efficiency of each approach in reducing overfitting and improving forecast accuracy. This methodical approach enables a comprehensive comparison of various forecasting techniques and their implementation in a functional R script, which can later be used in financial forecasting with data exhibiting diverse characteristics.

4. Research Results and Discussion

This chapter describes the development and implementation of the forecasting models using R programming: IV model, MLR, ARIMA and ES. Each model is discussed step-by-step from data preparation, model creation and use of particular R packages to analyse financial data. The R code for the process is available via link in Appendix 1.

4.1 IV Model

The following formula (1) was used for the IV Model:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2\, x_{2,t} + u_t \quad (1)$$

where

- y_t is the dependent variable in year t,
- $x_{1,t}$ and $x_{2,t}$ are the explanatory financial ratios,
- $\beta_0, \beta_1, \beta_2$ are the parameters to be estimated, and
- u_t is the error term capturing unobserved factors and random disturbances.

In the IV model, the process began with uploading the dataset and loading the necessary packages into the script, which can be seen between lines 12 and 21 of the code. The readxl package was used to import data from an Excel sheet via the read_excel function. The imported dataset was then stored in the script under the name "data" for further analysis. The Stargazer package was used to generate regression tables and format them in LaTeX for a polished final presentation. The research data was imported from an Excel file using the script and stored under the name "data" for further analysis. Next, we examined the correlations between the variables using the cor function. Following this, we calculated the Variance Inflation Factor (VIF) for the model, assessing both cash and accrual variables. This process is documented between lines 23 and 31 of the code. Further, the IV models were created by the ivreg function. The first stage of the IV model is visible in the code between lines 34 and 40. The second stage for the IV model, where we incorporated the instrument variables derived from the first stage, can be seen in the code between lines 34 and 46. The check for robustness was performed between lines 49 and 70 in the code. The full code is available in Appendix 1.

4.2 Multiple linear regression

The following formula (2) was used for multiple linear regression:

$$\Delta \mathrm{TA}_t = \beta_0 0 + \beta_1 X_{1,\mathrm{t}} + \beta_2 X_{2,\mathrm{t}} + \beta_3 X_{3,\mathrm{t}} + \varepsilon_t \quad (2)$$

where

- ΔTA_t is the change in total assets in the year t,
- $X_{i,t}$ represents the financial ratios used as predictors,
- β_i are the estimated parameters,
- ε_t is the stochastic error term.

For the MLR analysis, the process began with uploading the dataset and loading the necessary packages into the script, as shown between lines 12 and 23 of the code. The dplyr package was selected for data cleaning due to its extensive functionality, including filtering, summarizing, grouping, and mutating variables. However, in this case, only filtering was required.

Package Ggplot2 was selected for the graphical presentation of the data due to its widespread use and its ability to customize graphs with parameters which are not available in Microsoft Excel. Package ggthemes was chosen because it offers a wide selection of templates for graphs made on ggplot2. Package psych was chosen due to the "describe" function producing well-structured summary tables from all variables in the dataset. The "forecast" package was selected for forecasting, with its "predict" function being used specifically for linear regression models in the case of MLR. The package readxl was used to extract the data from the Excel sheet. This was done through the read_excel function, and the Excel sheet was uploaded into the script named "data".

In the next step of the script, we generated the absolute numbers for ratio analysis, which required calculating the average capital and average working capital variables. This was done between lines 25 and 40 in the code. In the script, the data for calculation of working capital was initially retrieved from the data variable by dplyr's filter function. First, we filtered the years we wanted to analyse, and then we selected variables to include in our new variable that is saved with the name data.ac.

Next, we created the variable for average_capital by using the base R function "sum" for every column to calculate the sum of each row in the data.ac variable. The result was then divided by the number of years, which, in this case, was 2. The outcome was saved as a variable named "average_capital". The next step in our code was to create the ratio table and calculate the ratios to be included in it. This is done in the code between lines 42 and 70.

To construct the table / dataframe, we chose the Year column from the data table as the first column in our table. This was chosen because it is a variable where all the other variables have independent dependency, meaning that the year is a variable that does not change if other variables are changing. This was done by as.data.frame function where we inserted the Year column from the data variable. The result was saved as a new variable named "ratios". Since as.data.frame function uses the

column location as a name (`data$Year`), we applied the "mutate" function from dplyr package to rename the column as "Year" for simplicity. We then ensured that only this variable was included in the "ratios" dataset.

In the next step, we formed the ratios. These ratios were formed from their mathematical equations, and the variables for equations were taken from the data table. The outcome of these calculations is saved in the table with the ratios in Appendix 1. After the calculation of ratios, we proceeded with the calculation of the differences in variables between the years. Then, the outcome was inserted into the regression models to predict the movement of the ratios for the coming years. This is visible between lines 71 and 100 in the code. The first variable from the ratios table is the debt-to-equity ratio, which can be found in the ratios.diff table. The difference between each row in the debt-to-equity column was calculated using the diff function in R, and the result was stored as a data frame. To ensure consistency and simplify future use, the same renaming process applied to the Year column in the ratios table was also applied to the debt-to-equity column. When all the differences have been calculated, the outcome is saved into the Excel file by using the write.xlsx function. This was done to save the obtained outcome from our script.

The next step in our script is the forming of the regression models. This is done by the lm function. In the code, this is visible between lines 126 and 179. When running the regression model, we faced an overfitting issue. Initially, both dependent and independent variables in the model were expressed solely as ratios, which led to significant overfitting. This overfitting was caused by the low variance in the data and by minimal differences between the variables used to calculate many of the ratios. Consequently, the changes between these ratio-based variables were extremely limited. To address this issue and overcome this problem, we adjusted the model by converting the dependent variable to absolute values from the original dataset while keeping the independent variables as ratios. This approach allowed the ratios to predict the movement of these absolute values, effectively mitigating the overfitting issue. The code for this section is available in Appendix 1.

4.3 ARIMA

The ARIMA model investigates the progress of a variable through time; this type of variable is referred to as a time series. As a first step, each time series $y_t^{(k)}$ was decomposed using an additive decomposition scheme:

$$y^{(k)} = T_t^{(k)} + S_t^{(k)} + e_t^{(k)} \quad (3)$$

where

- $T_t^{(k)}$ is the trend component,
- $S_t^{(k)}$ is the seasonal (or systematic cyclical) component, and

- $e_t^{(k)}$is the irregular (error) component.

Following decomposition, each ratio $y_t^{(k)}$was modelled using an autoregressive integrated moving-average (ARIMA) process. The general ARIMA(p, d, q) specification for ratio $k$is:

$$\phi(B)(1-B)^d y_t^{(k)} = \theta(B)\varepsilon_t^{(k)} \quad (4)$$

where

- $B$is the backshift operator ($By_t^{(k)} = y_{t-1}^{(k)}$),
- $d$is the order of differencing,
- $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$is the autoregressive (AR) polynomial with parameters $\phi_1, \ldots, \phi_p$,
- $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$is the moving-average (MA) polynomial with parameters $\theta_1, \ldots, \theta_q$, and
- $\varepsilon_t^{(k)}$is a white-noise error term with mean zero and constant variance, $\varepsilon_t^{(k)} \sim$ i.i.d.$(0, \sigma_k^2)$.

For each ratio time series $y_t^{(k)}$, the selected ARIMA model was then used to generate out-of-sample forecasts for three future periods:

$$\hat{y}_{t+1}^{(k)}, \hat{y}_{t+2}^{(k)}, \hat{y}_{t+3}^{(k)}, \quad (5)$$

For ARIMA forecasting, the process begins with loading the required packages and importing the dataset into the script, as shown between lines 14 and 30 of the code. The dplyr package is used for data cleaning due to its comprehensive set of functions for filtering, summarizing, grouping, and mutating variables. However, in this case, only the filtering function is utilized. Additionally, the lubridate package is employed to manage date variables, converting them into a format suitable for constructing a time series. The psych package is used specifically for its "describe" function, which generates well-structured summary tables for all variables in the dataset. The forecast package is selected for forecasting purposes, with its predict function applied in the case of MLR to generate predictions from linear regression models. Additionally, the readxl package is utilized to import data from an Excel sheet using the read_excel function. The imported dataset is then assigned the name "data" within the script for further analysis.

The tidyverse package ensures seamless integration of the lubridate and dplyr packages, as both belong to the same package ecosystem. The ggthemes package is chosen for its extensive collection of pre-designed graphical templates, enhancing the visual presentation of ggplot2 graphs. The ggplot2 package itself is selected for data visualisation due to its widespread use and flexibility, allowing for extensive customisation beyond what is available in Microsoft Excel. Finally, the Stargazer

package is used to generate regression tables and format them in LaTeX for a polished final presentation.

The xts package is loaded into the script, though it is not utilised in the ARIMA section. Instead, it is employed in the ES section, where its ses function is used to generate ES forecasts, as explained in the corresponding subchapter. To optimise readability, the scipen value is set to 999 in the options settings. This adjustment prevents scientific notation from being applied unless a number contains 999 or more digits after the decimal point, ensuring clearer result interpretation. Next, the research data is imported using the read_excel function and is assigned the name linas_agro within the script. Additionally, a separate table is created with a Year column formatted in day/month/year format. To simplify both the code and the construction of the time series variable, the starting date of each year is used. Finally, the values in the Year column are converted into date variables using the dmy function from the lubridate package. Next, the time series variables are created using the ts function from the forecast package. This process is implemented in the code between lines 61 and 85. The time series is constructed for the period 2014 to 2022, as the calculation of year-over-year ratios results in the loss of data for the t-1 year. To ensure that the time series accurately represents data collected over a full calendar year (365 days), the frequency is set to 365. With the time series successfully generated, the ARIMA modelling process can begin. The ARIMA implementation for each variable is detailed in the code between lines 88 and 230.

The first part in the ARIMA process involves identifying the trend, variance and frequency within the dataset. In this project, the decompose function is used from the forecast package. An additive decomposition format is selected to treat all the components (trend, seasonality, and residuals) as independent elements that combine to form observed values without multiplicative interaction between them.

When we run the decomposing process, we are optimising the outcome through the minimisation of the Akaike Information Criterion (AIC) and maximising the likelihood estimation (MLE) using the auto.arima function from the forecasts package. Running the code for this section might take time from a couple of minutes to 10 minutes, depending on the number of CPU cores available for use. The code in this project is designed and tested on hardware with 8 cores fully available to use.

With all the components available and the settings optimised, the Arima forecast for the variables can now be performed. This is achieved using the forecast function from the forecast package. In our case, the forecast is generated for the next 3 years.

The forecast function generates its output as a list, which must be converted into a data frame using the as.data.frame function. However, this data frame does not initially include a corresponding time variable for the forecasted periods. To resolve this, a new year variable is created to represent the forecasted years 2023, 2024, and

2025. In the final section of the code, the focus shifts to visualising the forecast results. This is accomplished using the ggplot2 and ggthemes packages, which allow for the creation of well-structured and visually appealing graphs. The implementation of this visualisation process is documented in the code between lines 232 and 423.

The graph creation process begins with the ggplot function, where the first argument specifies the dataset to be visualised. The aes function is used to define the aesthetics, mapping the x-axis to the year variable and constraining the y-axis to values between 0 and 2.5. This constraint ensures that the graph remains focused and precise, allowing even minor variations to be visually discernible. To visualise the forecast, the geom_line function is applied, mapping the y-axis to the Point Forecast column from the forecasted data while assigning a specific colour to differentiate the forecasted trend. The scale_x_continuous function is used to establish evenly spaced breaks along the x-axis, ensuring that the timeline aligns with yearly intervals. The graph's theme is set to economist, a style specifically designed for financial and economic data visualisation. Finally, the labs function is used to incorporate a title, axis labels, and a caption, enhancing clarity and contextual understanding. The complete code for this section is provided in Appendix 1.

4.4 Exponential Smoothing

For exponential smoothing calculation, the following formula was used:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \quad (6)$$

where

- $\hat{y}_{t+1}$is the one-step-ahead forecast,
- $\alpha \in (0,1)$is the smoothing parameter,
- $y_t$is the observed value in period t, and
- $\hat{y}_t$is the smoothed level estimate for the period t.

The ES and ARIMA models are implemented within the same script, ensuring both methods utilise the same dataset and time series. This approach enhances efficiency by avoiding redundant processes, making the script more resource-friendly, particularly for older hardware or machines with limited computational power. Since the data uploading and cleaning procedures are identical to those used for ARIMA, they are not repeated in this section. Once the dataset is cleaned and the same time series is established, the ES forecast is executed. This process is implemented in the script between lines 425 and 592. The ES forecast is performed using the ses function from the xts package. The function's first argument specifies the dataset for forecasting, while the alpha parameter controls the smoothing factor, reflecting the expected level of fluctuation in the data. The h parameter is set to 10, generating forecasts for 2015 to 2022 and extending into the next three years (2023–2025),

resulting in a total of 10 forecast points. The forecast output is initially generated as a list, which is then converted into a data frame using the data.frame function. From this data frame, the first three rows—corresponding to the forecasted values for 2023, 2024, and 2025—are extracted. These values, initially stored as character strings, are then converted to a numeric format using the as.numeric function. The final step involves visualising the results in a graphical format, following the same structure as described in the ARIMA section. This visualisation process is implemented in the code between lines 594 and 605.

4.5 Outcome from the script

The results of IV regression for accrual-based ratios are presented below in Table 1.

Table 1. The result of IV regression for accrual ratios

Dependent Variable	Intercept	Debt_to_equity	Equity_to_assets	Summary Statistics:
				Residual Std. Error: 32060 on 4 degrees of freedom
Estimate	2619644	-471784	-4150176	Multiple R-Squared: 0.9805
Std. Error	229166	122572	351298	Adjusted R-Squared: 0.9707
t value	11.431	-3.849	-11.814	Wald Test: 100.5 on 2 and 4 DF,
p-value	0.000334 ***	0.018318 *	0.000294 ***	p-value: 0.0003807

Table 1 presents the results of an IV regression examining the relationship between financial structure and total assets. The intercept is 2,619,644 ($p < 0.001$), representing the baseline level of total assets when both predictors—debt-to-equity and equity-to-assets—are zero. The coefficient for debt-to-equity is -471,784 ($p < 0.05$), indicating that higher leverage relative to equity is associated with a reduction in total assets, potentially reflecting financial constraints. Similarly, equity-to-assets has a highly significant negative coefficient of -4,150,176 ($p < 0.001$), suggesting that as the proportion of equity in total assets increases, total assets decrease substantially. The model demonstrates a high explanatory power, with an R^2 of 0.9805 and an adjusted R^2 of 0.9707, indicating that approximately 98% of the variance in total assets is accounted for by the predictors. The Wald test statistic (100.5, $p < 0.001$) further confirms the joint significance of the explanatory variables. However, the high R^2 in combination with a very low number of observations (4 degrees of freedom) raises concerns about overfitting, suggesting that the model may not generalise well to a larger dataset. These findings highlight the strong influence of financial structure on total assets, though further research

with a larger sample size is warranted to validate these relationships and mitigate potential overfitting concerns.

The regression results presented in Table 2 show the relationship between cash flow and three predictors: cash interest coverage, cash flow to debt, and capital expenditure.

Table 2. The results of IV regression for cash-based ratios

Dependent Variable	Intercept	cash_ interest _coverage	Cashflow _debt	capital_ expenditure	Summary Statistics:
					Residual Std. Error: 5466 on 3 degrees of freedom
Estimate	-6966.9	1683.0	12293.9	4861.9	Multiple R-Squared: 0.9749
Std. Error	2970.5	943.6	50867.4	3584.2	Adjusted R-Squared: 0.9497
t value	-2.345	1.784	0.242	1.356	Wald Test: 38.79 on 3 and 3 DF, p-value: 0.006712
p-value	0.101	0.172	0.825	0.268	

Table 2 presents the results of an IV regression analysing the relationship between cash-based financial ratios and the dependent variable. The intercept is -6,966.9, which is not statistically significant ($p = 0.101$), indicating no meaningful baseline cash flow when the predictors are zero. Among the predictors, cash interest coverage has a positive coefficient of 1,683.0, but it is not statistically significant ($p = 0.172$).

Similarly, cash flow to debt (coefficient: 12,293.9, $p = 0.825$) and capital expenditure (coefficient: 4,861.9, $p = 0.268$) do not show significant relationships with cash flow, suggesting weak explanatory power for these variables in the model. Despite the lack of significance for individual predictors, the model itself demonstrates a high explanatory power, with an R^2 of 0.9749 and an adjusted R^2 of 0.9497, meaning that 97.49% of the variance in cash flow is accounted for by the predictors. However, given the low significance of the individual coefficients, this high R^2 may be primarily driven by the model's overall structure rather than the contribution of specific variables. The Wald test statistic (38.79, $p = 0.0067$) confirms that the model is jointly significant, implying that the predictors, when considered together, influence cash flow. However, the lack of statistically significant individual coefficients suggests potential model limitations, such as omitted variables or multicollinearity, that may be influencing the results.

Table 3 presents the results of an MLR examining the relationship between Total Assets and three predictors: debt-to-equity, equity-to-assets, and interest coverage. The constant term is statistically significant at the 5% level (20,756.260, $p < 0.05$),

suggesting a baseline level of total assets when all predictors are zero. However, none of the predictor variables exhibits statistical significance. The coefficient for debt-to-equity is 63,560.210 ($p > 0.05$), indicating a positive but statistically weak relationship with total assets. Similarly, equity-to-assets (-57,677.680) and interest coverage (3,071.088) also lack statistical significance, suggesting these variables do not independently explain variations in total assets within this sample.

Table 3. The results of MLR from accrual-based ratios

Variable	Value	Standard Error / Notes
debt_to_equity	63560.21	(73,642.230)
equity_to_assets	-57677.68	(284,469.900)
interest_coverage	3071.088	(3,132.650)
Constant	20756.26	(6,168.142)**
Observations	11	
R^2	0.541	
Adjusted R^2	0.345	
Residual Std. Error	19,995.350 (df=7)	
F Statistic	2.755 (df=3;7) (p=0.122)	

Note: *p<0.1; **p<0.05; ***p<0.01

The model accounts for 54.1% of the variance ($R^2 = 0.541$), with an adjusted R^2 of 0.345, indicating moderate explanatory power but a substantial degree of unexplained variance. The residual standard error (19,995.350) represents the average prediction error, highlighting model imprecision. Additionally, the F-statistic (2.755, $p = 0.122$) suggests that the model, as a whole, is not statistically significant at conventional significance levels. These findings indicate that while the model provides some insight into total assets, the lack of statistically significant predictors suggests that additional explanatory variables may be needed to improve predictive accuracy and robustness. Further model refinement, such as incorporating additional financial indicators or interaction terms, could enhance the explanatory power.

Table 4 presents the results of an MLR examining the relationship between net profit (dependent variable) and three predictors: capital expenditure, cash flow to debt, and cash interest coverage. The constant term is statistically significant at the 1% level (21,655.650, $p < 0.01$), indicating the baseline level of net profit when all predictors are zero. However, none of the independent variables show statistical significance at conventional levels.

Table 4. The results of MLR from cash-based ratios

Variable	Value	Standard Error / Notes
capital_expenditure	-13268.42	(7,771.108)
cashflow_to_debt	-36520.59	(78,259.380)

Variable	Value	Standard Error / Notes
cash_interest_coverage	2399.6	(1,412.555)
Constant	21655.65	(3,921.726)***
Observations	11	
R^2	0.744	
Adjusted R^2	0.634	
Residual Std. Error	12,570.180 (df=7)	
F Statistic	6.778 (df=3;7) (p=0.018)	

Among the predictors, capital expenditure has a negative coefficient of -13,268.420 but is not statistically significant ($p > 0.05$), suggesting a weak inverse relationship with net profit. Similarly, cash flow to debt has a negative coefficient of -36,520.590, indicating a potential negative association, though it is also not significant ($p > 0.05$). Conversely, cash interest coverage has a positive coefficient of 2,399.600, but this effect is not statistically significant ($p > 0.05$), implying that its influence on net profit is not strong in this model. The model demonstrates moderate explanatory power, with an R^2 of 0.744, meaning that 74.4% of the variance in net profit is explained by the predictors. The adjusted R^2 of 0.634 accounts for the number of predictors, suggesting a reasonably strong model fit despite the lack of significant individual coefficients. The residual standard error (12,570.180) reflects the average deviation between observed and predicted net profit values. The F-statistic (6.778, $p = 0.018$) indicates that the model is jointly statistically significant, even though the individual predictors do not reach significance. This suggests that while the variables collectively contribute to explaining net profit variations, their individual effects may be weak or confounded by other unmeasured factors.

4.6 Methodological limitations and model comparability

A key methodological limitation of the analysis concerns the restricted comparability of the four modelling approaches tested: instrumental-variables regression, multiple linear regression, ARIMA models, and exponential smoothing. Although each model independently generates interpretable results related to credit-risk behaviour, their underlying assumptions and data structures differ substantially. The IV regressions use ratio-based explanatory variables within a very small sample, while the MLR models combine ratio-based predictors with absolute-value dependent variables to reduce overfitting. On the other hand, ARIMA and ES are univariate time-series frameworks applied to individual ratios over longer and more consistent temporal spans without incorporating the multivariate structure central to the regression models. These structural differences in sample size, variable formation and temporal granularity limit the extent to which these results can be interpreted in a directly comparative manner.

Within this context, the findings show model-specific outcomes rather than systematic similarities or divergences across modelling techniques. For instance, the IV models reveal negative associations between leverage indicators and asset levels, yet these relationships do not appear in the MLR results, where differences in variance and variable transformations weaken coefficient significance. Similarly, the forecasts generated by ARIMA and ES describe the temporal evolution of individual ratios but cannot be replicated in regression models that rely on cross-sectional or differenced information. Consequently, while the results are informative within each methodological domain, they do not collectively yield a coherent comparative assessment of credit-risk determinants.

These limitations, alongside the findings of the present paper, point to potential omitted-variable bias, multicollinearity and model-specification issues, particularly for the regression models. Future research could expand the explanatory set, incorporate interaction effects or apply multivariate time-series techniques to improve predictive accuracy and capture interdependencies among financial ratios. Such extensions would help to overcome the constraints observed in the present analysis and support more robust comparative modelling.

5. Conclusions

In this research, we explored the effectiveness of R programming in forecasting financial information using financial ratios and absolute numbers. The empirical research demonstrates that the ARIMA and ES models outperform traditional IV and MLR models in forecasting financial ratios. The ability of ARIMA and ES models to adapt to the unique characteristics of financial data provides a more stable and accurate forecasting tool, significantly reducing the issue of overfitting encountered with absolute numbers. This is attributed to their bottom-up approach, which allows forecasts to be formed based on the individual characteristics of a variable rather than on the absolute value or simple change between two points.

In response to the research question, we conclude that it is possible to forecast financial ratios more precisely using R programming for ARIMA and ES, as indicated by our findings. When comparing the result of the script presented in this paper with the information from the outcome, which was presented in a publication by Litvinenko, Litvinenko, and Saarinen (2025), it is evident that ARIMA and ES forecasts performed through R programming are producing results that are able to tackle the existing overfitting problem with financial ratios forecasting.

Building on Kowal's (2016) research, which extends the forecasting efficiency of MLR models, we addressed the enhancement of such models for financial ratios by focusing on individual characteristics of variables such as variance, frequency, and trend. Our findings with financial data indicated that when absolute values are used

as dependent variables instead of financial ratios, models for both accrual and cash flows are prone to overfitting due to the low variance in the data. To address this issue, we shifted our approach towards methods like ARIMA and ES, where the individual characteristics of the ratios serve as independent variables and the ratio itself as the dependent variable. This adjustment was made to simplify the forecasting process.

Our findings demonstrate that the ARIMA method can be applied to financial data, not only to macroeconomic data, as shown in Zhu's (2018) research. In the context of financial data, ARIMA forecasting outperformed other methods due to its capacity to adapt to data changes even with a limited amount of training data.

In this research, we utilised the extracted information to explain financial ratio data using various forecasting methods. In future, this approach could be modified to develop more efficient credit and bankruptcy risk models.

Although this research focused on R programming, future research could compare the effectiveness of similar forecasting methods implemented in other programming languages, such as Python and MATLAB, potentially providing insights into the most efficient tools for financial forecasting. Another avenue for future research could involve the development of real-time forecasting models using R programming, which would entail leveraging streaming financial data to make instantaneous forecasts – a critical capability in today's fast-paced financial markets.

References

Andrews, I., Stock, J. H., & Sun, L. (2019) "Weak instruments in instrumental variables regression: Theory and practice", *Annual Review of Economics,* vol. 11, no. 1: 727-753

Angrist, J. D., & Krueger, A. B. (2001) "Instrumental variables and the search for identification: From supply and demand to natural experiments", *Journal of Economic Perspectives,* vol. 15*,* no. 4: 69-85

Baltagi, B. H. (2021) *Econometric Analysis of Panel Data (6th ed.).* Springer.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995) "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak", *Journal of the American Statistical Association,* vol. 90, no. 430: 443-450

Brooks, L. D., & Buckmaster, D. A. (1976) "Further evidence of the time series properties of accounting income", *Journal of Finance,* vol. 31, no. 5: 1359-1373

Chatfield, C. (2003) *The Analysis of Time Series: An Introduction*, Boca Raton: Chapman and Hall, CRC Press

Chen, N., Ribeiro, B., & Chen, A. (2015) "Financial credit risk assessment: A recent review", *Artificial Intelligence Review,* vol. 45*,* no. 1: 1-23

Chen, Y., Calabrese, R., & Martin-Barragán, B. (2024) "Interpretable machine learning for imbalanced credit scoring datasets", *European Journal of Operational Research,* vol. 312*,* no. 1: 357-372

Cifter, A., Yilmazer, S., & Cifter, E. (2009) "Analysis of Sectoral Credit Default Cycle Dependency with Wavelet Networks: Evidence from Turkey", *Economic Modelling,* vol. 26: 1382-1388

Crouhy, M., Galai, D., & Mark, R. (2000) "A comparative analysis of current credit risk models", *Journal of Banking & Finance,* vol. 24, no. 1-2: 59-117

Cruceru, D., Anghel, M., & Diaconu, A. (2016) "The multiple linear regression used to analyze the correlation between variables", *Romanian Statistical Review Supplement,* vol. 64*,* no. 1: 114-117

Duffie, D., Eckner, A., Horel, G., & Saita, L. (2009) "Frailty Correlated Default", *The Journal of Finance,* vol. 64, no. 5: 2089-2123

Greene, W. H. (2018) *Econometric analysis* (8th ed.). Pearson Education Limited

Gujarati, D. N., & Porter, D. C. (2009) *Basic econometrics* (5th ed.). McGraw-Hill

Horowitz, J. L. (2011) "Applied nonparametric instrumental variables estimation", *Econometrica,* vol. 79*,* no. 2: 347-394

Hyndman, J., & Athanasopoulos, G. (2018) *Forecasting: Principles and Practice* (2nd ed.). Melbourne, Australia

Hyndman, R. J., & Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", *Journal of Statistical Software,* vol. 27*,* no. 3: 1-22

Imbens, G. W. (2014) "Instrumental variables: An econometrician's perspective", *Statistical Science,* vol. 29, no. 3: 323–358

Kedia, P., & Mishra, L. (2024) "Credit risk management: A systematic literature review and bibliometric analysis", *Journal of Credit Risk,* vol. 20*,* no. 1: 51-76

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014) "Improving forecasting by estimating time series structural components across multiple frequencies", *International Journal of Forecasting,* vol. 30*,* no. 2: 291-302

Kowal, R. (2016) "The efficiency of OLS estimators of structural parameters in a simple linear regression model in the calibration of the averages scheme", *Folia Oeconomica Stetinensia,* vol. 16, no. 1: 236-249

Lin, L. T., & Koo, T. Y. (2007) "Applying exponential smoothing in quality function deployment to analyze dynamic customer needs", *Chinese Journal of Management,* vol. 8, no. 3: 59-69

Litvinenko, A., Litvinenko, A., & Saarinen, S. (2025) "Applying forecasting methods to accrual-based and cash-based ratio analysis", *Journal of Accounting and Management Information Systems,* vol. 24, no. *2:* 328-360

Luong, T. M., & Scheule, H. (2022) "Benchmarking forecast approaches for mortgage credit risk for forward periods", *European Journal of Operational Research,* vol. 299, no. 3: 750-767

Makridakis, S., Wheelwright, S.C., & Hyndman, R.J. (1998) *Forecasting: Methods and Applications*, New York: John Wiley & Sons

Medina-Olivares, V., Calabrese, R., Crook, J., & Lindgren, F. (2023) "Joint models for longitudinal and discrete survival data in credit scoring", *European Journal of Operational Research,* vol. 307, no. 3: 1457-1473

Schielzeth, H. (2010) "Simple means to improve the interpretability of regression coefficients", *Methods in Ecology and Evolution,* vol. 1, no. 2: 103-113

Snyder, R., Koehler, A., & Ord, J. (2002) "Forecasting for inventory control with exponential smoothing", *International Journal of Forecasting,* vol. 18, no. 1: 5-18

Swamy, P., Tavlas, G., & Hall, S. (2015) "On the interpretation of instrumental variables in the presence of specification errors", *Econometrics,* vol. 3, no. 1: 55-64

Taylor, J. W. (2004) "Volatility forecasting with smooth transition exponential smoothing", *International Journal of Forecasting,* vol. 20, no. 2: 273-286

Tsay, R. S. (2010) *Analysis of Financial Time Series*. 3rd Edition, John Wiley & Sons, Hoboken

Varian, H. (2014) "Big data: New tricks for econometrics", *Journal of Economic Perspectives, vol. 28*, no. 2: 3-28

Wooldridge, J. M. (2013) *Introductory Econometrics: A Modern Approach* (5th ed.). South-Western Cengage Learning

Zhu, Y., Wang, Y., Liu, T., & Sui, Q. (2018) "Assessing macroeconomic recovery after a natural hazard based on ARIMA: A case study of the 2008 Wenchuan earthquake in China", *Natural Hazards, vol. 91*, no. 3: 1025-1038

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J., & Nguyen, T. V. (2019) "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach", *International Journal of Production Economics,* vol. 211: 22-33

Appendix 1

The data and code are available via the link below.
https://github.com/publicationcodes/Different-Methods-and-Techniques-of-Forecasting-Written-in-R-Studio.

Inside the depository, the codes are divided into 4 different sections: ARIMA, Exponential Smoothing, Multiple Linear Regression and IV regression.

The Excel files contain ratio calculations.

Since the files represent intellectual property, the depository is private. Therefore, to access the content, anyone has to request access. To request access, the email or GitHub credentials of the requester have to be sent to the corresponding author.