VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Magistro darbas

Kredito unijų pelningumo vertinimas ir prognozavimas: Lietuvos atvejis

Credit union profitability estimation and prediction: the case of Lithuania

Agnė Griškevičiūtė

VILNIUS 2020

MATEMATIKOS IR INFORMATIKOS FAKULTETAS
STATISTINĖS ANALIZĖS KATEDRA

Darbo vadovas doc. dr. Jurgita Markevičiūtė ______________

Darbas apgintas 2020 m. sausio 13 d.

Registravimo NR. ______________

Kredito unijų pelningumo vertinimas ir prognozavimas: Lietuvos atvejis

Santrauka

Magistro darbo tikslas ištirti veiksnius, lemiančius Lietuvos kredito unijų pelningumą, bei prognozuoti būsimas kredito unijų pelningumo rodiklių reikšmes. Siekiant identifikuoti veiksnius, sąlygojančius kredito unijų pelningumo pasikeitimus, naudoti paneliniai autoregresiniai paskirstytų vėlavimų modeliai su paklaidų korekcija (ARPV). Prognozavimui naudoti įvertinti paneliniai ARPV modeliai, regresijos medžiai bei neseniai pasiūlyti medžiais pagrįsti algoritmai paneliniams duomenims. Rezultatai rodo, jog šalies ekonominis aktyvumas, bankų kainodara bei operacinių išlaidų dalies pajamose rodiklis yra reikšmingi ilgo laikotarpio kredito unijų pelningumą lemiantys veiksniai. Taip pat darbe parodyta, kad trumpuoju laikotarpiu kapitalo pakankamumo bei grynųjų palūkanų pajamų pokyčiai turi teigiamą poveikį pelningumui, o išaugusi prastesnės kokybės paskolų dalis turi neigiamą poveikį kredito unijų pelningumui. Įvertintų panelinių ARPV modelių bei trijų medžiais pagrįstų metodų prognozių tikslumų palyginimas skirtingiems prognozavimo horizontams leidžia teigti, kad daugeliu atvejų mažiausias prognozavimo paklaidas turėjo padidintų medžių metodas. Taip pat pastebėta, kad kai kuriais atvejais, statistinių panelinių ARPV modelių prognozių paklaidos buvo artimos mažiausias prognozavimo paklaidas turinčiam metodui. Tyrimo rezultatai suteikia geresnį suvokimą apie kredito unijų pelningumą, o bendros darbo įžvalgos gali būti naudingos priežiūros institucijoms.

Raktiniai žodžiai : Lietuvos kredito unijos, pelningumas, paneliniai duomenys, regresijos medžiai.

Credit union profitability estimation and prediction: the case of Lithuania

Abstract

The purpose of this master thesis is to investigate main determinants of Lithuanian credit unions profitability and to predict future credit union profitability values. The examination of main profitability factors has been done by panel autoregressive distributed lags model with the error correction term (ARDL). Estimated panel ARDL model, regression trees and more novel tree-based algorithms have been applied to predict future profitability ratios. The results indicate that country economic activity, average loan interest rates of banks and ratio of operating expenses to total income are significant long-term profitability determinants. Additionally, it was found that in short-term, changes in capital adequacy ratio and net interest income have positive impact to credit unions' profitability, while growth of worsened loan portfolio quality, have negative consequences on credit unions profitability. Comparison of predictions for different forecasting horizons of the fitted statistical panel ARDL model and three tree-based methods, indicate that boosted trees predicts credit union profitability with the lowest prediction error. Moreover, it was shown that in some cases, prediction errors of panel ARDL models were sufficiently close to the boosted trees. The findings of the study provide better understanding of internal and external credit union profitability determinants and the overall insights can be useful for supervisory institutions.

Key words : Lithuanian credit unions, profitability, panel data, regression trees.

Contents

1 Introduction

Credit unions are cooperative institutions what provide main banking services solely to their members. Each member has an amount of credit union shares and is an owner of credit union. The main purpose of such institution is to satisfy the needs of their members. In theoretical literature credit unions are presented as non-profitable organisations, since to satisfy the needs of their participants, the access of credit union services are usually at competitive rates comparing to banks. However, the consequences of operating non-profitably has been seen in various countries, such as Lithuania, in the past few decades. The downturns of many credit unions have changed thinking about credit union activities and purposes. It was understood, that in order to satisfy the needs of their members, credit unions have to be financially healthy, ensure financial stability and soundness to their members. One of financial health indicators are profitability ratios. Literature review showed that only several researches investigated profitability of credit unions. Mainly, these studies have focused on only few possible profitability determinants. However, there were no studies on credit union profitability prediction and investigation on how macroeconomic and credit union financial activities affect profitability of credit unions. Moreover, it was noted that recent studies suggest superior results on panel data prediction using tree-based methods. Therefore, the downturns of credit union sector and the lack of researches investigating credit union profitability, highlights the importance of the topic.
The results showed significant long term impact of country's economic activity, banks' loan pricing and ratio of operating expenses to total income to credit union profitability. It was indicated that economic activity and bank loan interest rates have positive long-run impact, ratio of operating expenses to total income have negative effect. Significant short term relationships were found between capital adequacy ratio, loan portfolio quality, net interest income and profitability ratios. Prediction accuracies of the fitted model and three tree-based methods were superior to the novel boosted trees algorithm for panel data. However, in some cases the fitted model prediction errors similar as boosted tress, therefore we can not conclude absolute superiority of tree-based methods to traditional statistical model.

The **aim** of the master thesis is to investigate main determinants of Lithuanian credit unions profitability and to predict future credit union profitability values.

The main **tasks**:

1. Perform literature review on credit institutions' profitability and indicate

the methods used for further profitability estimation and prediction analysis.

2. Estimate panel auto-regressive distributed lags model with error correction on two profitability ratios: return on assets and real interest margin and check the model assumptions.

3. Perform predictions on return on assets and real interest margin using estimated model and tree-based methods and compare prediction accuracies.

The further structure of this master thesis is the following: section 2 covers literature review on past studies including credit institutions' profitability modeling and forecasting, section 3 represents methods and data that will be used in a study, section 4 focuses on model estimation, prediction and results. The last section conveys conclusions.

2 Literature Review

The purpose of this section is to investigate the methods used by related literature for both profitability evaluation and prediction. Because of the lack of researches on credit union profitability, literature review focuses on studies on profitability of all credit institutions, mainly, banks. The methods used in studies that have analyzed financial institutions' profitability can be divided into two specific groups: traditional statistical models and machine learning methods.

Traditional models, such as multiple linear regression, static and dynamic panel data regressions have been broadly examined in many studies searching profitability determinants. Early investigations on profitability have focused on multiple linear regression models using ordinary least squares and mainly were used in the past few decades. Sanusi, Mohamed (2007), Nataraja *et al.* (2018) used multiple linear regression in order to find the main profitability determinants of domestic banks', Haron (1997), Capraru, Ihnatov (2014), Jilkova *et al.* (2017) analyzed cross-country data multiple regressions. The results of these multiple regression studies show that bank profitability determinants were bank specific-financial and country macroeconomic variables. Dandapani *et. al* (2008) investigated the impact of internet banking services on profitability of credit unions in USA. The researchers used multiple cross-sectional regressions. They have found similar average profitabilities between credit unions that provide ability to the internet access and those, that do not have web accounts. Multiple linear regression has an advantage of being easily understandable and interpretable; however, multiple regression faces significant disadvantages: estimates can be inconsistent and biased because of collinearity and autocorrelation problems, outliers and high-leverage points can meaningfully affect regression results (Baltagi, 2001; Mazlina, Bakar, 2009). Additionally, inclusion of more explanatory variables requires sufficiently large size of the data; otherwise, estimators might be unstable. On contrary, another traditional approach, mainly used in financial institutions' profitability studies – panel data models, usually have more degrees of freedom, can control the impact for both individual heterogeneity and intemporal dynamics (Baltagi, 2007). Therefore, panel data can improve accuracy of the estimates and predictions (Baltagi,2008)). Hsiao (1989, 1993) noted that in cases where individual behaviors are similar on certain variables, panel data might possibly learn individual's behavior by observing behavior of other individuals. Therefore, panel data can obtain more accurate portrait about individuals' manner. In such cases, if data are fitted as pooled time series data, estimates would suffer from aggregation bias.

Static panel data models were rarely used for profitability evaluation comparing to dynamic ones. A few to mention were Du (2018), Hadi *et al.* (2018), Cetin (2019) papers. The study of Du (2018) was one of the first studies that used panel data model for credit union data. The author investigated the impact of channel and product diversity to 7577 credit unions profitability and profit volatility during 2009-2016 time period. It was found that variety of both offline and online services lead to higher profitability of credit unions and lower profit volatility. However, more determinants have not been taken into account.

Another type of panel data regressions - dynamic panel data models are commonly used in profitability literature. According to (Hsiao, 2007), dynamic panel regression estimates can give more accurate model parameters since they usually reduce collinearity between current and lagged variables. Also, it was noted by (Nerlove, 2002) that dynamic relationships are inherent in economic behaviour. Consequently, dynamic panel data models were more popular while assessing profitability. Since standard static panel data procedures, such as random effects and fixed effects, in dynamic case are biased (Anderson and Hsiao, 1981; Anderson and Hsiao, 1982, Nickel, 1981), generalized methods of moments (GMM) estimator, first proposed by Hansen (1982), is common in dynamic panel data studies. GMM estimator allows for efficient estimation even in the presence of heteroscedasticity. Dynamic GMM estimator uses lagged values of dependent variables as well as lagged values of independent variables and a certain number of moment conditions. This approach was broadly used in recent studies by the researchers that investigated performance of domestic banks (Athanasoglou, Brissimis, Delis (2006), Tan (2015)) and cross-country banks (Albulescu (2014), Dietrich, Wanzenried (2014), Petria *et. al* (2015), Bouzgarrou *et. al* (2017), Campmas (2018)). As in multivariate model cases, it was found that main banks' profitability determinants are bank-specific and macroeconomic variables.

Another dynamic panel data approach was adopted by Naruševičius (2018) in order to find long-term and short-term determinants of Lithuanian banks' revenue and expense statements. The author used panel ARDL error correction model with pooled mean group (PMG) estimation technique that allows for heterogeneity in short-term coefficients, intercepts, error correction terms while using homogeneity in long-term coefficients. The author was one of the first that introduced this method on evaluation of financial data. Recently, Ali *et. al* (2019) used dynamic panel ARDL model for evaluation of non-traditional income effect to profitability of three major banks in South Asia. The main advantage of this estimator, proposed by Pesaran *et. al* (1999), is that in case where both number of time of observations (T) and cross-sections (N) are sufficiently large, PMG estimator, under certain assumptions, can be consistent and efficient. Additionally, Pesaran and Smith (1995) showed that in dynamic case, when T is larger and panels are heterogeneous, pooled models with estimations of fixed effects, IV, GMM estimators might give inconsistent and potentially very misleading estimates, when coefficients differ across groups.

Another group of methods – machine learning (ML), have gained popularity in predictive researches. There were several studies that used ML for predicting financial institutions' profitability. Lin et. al. (2009) used particle swarm optimization on support vector machines (SVM) and decision tree methods to select subset of parameters for accurate prediction of Taiwanese commercial banks' performance. The results indicated that proposed approaches are helpful to reduce unnecessary features and improve classification accuracy. Another study of Erdal, Karahanoglu (2016) examined main determinants of profit on Development and Investment Banks in Turkey. The bagging method on tree-based methods (Decision Stump, Random Tree, Reduced Error Pruning Tree) were used. The results have shown that bagging approach improves the prediction accuracy of tree-based methods. Ten-fold cross validation was used in order to predict generalization error. The bagging that was used on a random tree performed best.

Comparisons of prediction accuracy between traditional models and non-traditional machine learning methods were frequent and mostly superior to machine learning methods. Dincer, Hacioglu, Emir (2014) analyzed bank profits of the Turkish Banking Sector by comparing support vector regression (SVR) and linear regression predictions. The findings have shown that bank profits are better predicted by SVR method than linear regression. Haskamp (2017) forecasted profitability of 2000 regional banks using traditional (random walk, autoregression, OLS) and machine learning (random tree, random forest, gradient boosting) methods. All of machine learning methods were superior to traditional ones. Best results were gained by gradient boosting algorithm. Mazlina, Bakar (2009) in their research used multiple linear regression and neural network methods while using data of thirteen banks for the period 2001-2006. The study aims to predict performance of banks. The results have shown that artificial neural network method has lower prediction error comparing to traditional multiple regression method. Paula *et. al* (2019) used classification trees and logistic regression in order to compare prediction accuracies of Brazilian credit unions clients by measuring their profitability and default probability. The result have showed that regression trees provide much better prediction accuracy compared to traditional logistic regression. However the researchers concluded that the results of the study might not be very robust. It is seen that the results of majority of papers are in a favor of machine learning techniques while predicting banks' profitability, since these methods do not require distribution of the variables to be specified.

Additionally, it was found in recent literature in other fields of study that machine learning methods applied on panel (longitudinal) data also might improve prediction accuracies, comparing to traditional methods. Recent papers by Adler *et. al* (2011), Sela, Simonof (2009, 2012), Loh, Zheng (2013), Pande *et. al* (2017), Ngufor *et. al* (2018) compared prediction accuracies of various tree-based methods constructed on longitudinal data with traditional statistical panel data regressions. Adler *et. al* (2011) proposed bagged classification trees and random forests for classification of medical data and compared predictions with statistical logistic regression. Pande et. al (2017) used linear regression with generalized least squares on longitudinal data and multivariate data as a benchmark, while comparing predictions of proposed multivariate regression trees methods. Ngufor et. al (2018) proposed mixed-effect machine learning framework to predict longitudinal change in glycemic control. The predictions were compared to longitudinal generalized linear mixed-effect models. Sela, Simonof (2012) compared predictions of traditional panel data models: linear pooled model, linear mixed effects regression, linear mixed effects regression with auto-correlation and tree-based methods, applied on panel (longitudinal) on transactions data: decision tree, semi-parametric Random Effects/EM tree. Most of the studies, especially then data set is large, conclude that tree-based methods might better predict panel (longitudinal) data comparing to traditional statistical methods.

In contribution to the literature and data properties, panel dynamic autoregressive distributed lags model with the error correction will be used to find main profitability determinants. Additionally to the mentioned model, tree-based methods will be used for predictions: regression trees and novel random forests and boosted trees algorithm for longitudinal data, proposed in Sexton (2018), that were not yet considered in the literature.

3 Data and Methodology

In this section data and methodology of panel dynamic auto-regressive distributed lags model with the error correction (ARDL) and other methods that will be used to predict credit union profitability are reviewed . In contribution to the literature, ARDL model with pooled mean group (PMG) estimator will be used to fit credit union-specific and macroeconomic data. As a robustness check mean group (MG) estimator of the panel ARDL model was additionally performed. The estimates of PMG have important properties that enable to highlight long-term and short-term relationships between credit union profitability and other financial and macroeconomic variables. As was described in section 2, in cases where N and T are relatively large and panels are heterogeneous in the short run (and long run for MG estimator) coefficients, these estimators can provide more accurate estimates and, consequently, predictions, comparing to other traditional dynamic panel data models. The subsections bellow describe procedures that will be used in order to estimate panel data ARDL model with PMG and MG estimators. In this thesis ARDL model with PMG and MG estimators will be called traditional panel data models. On contrary, to compare prediction accuracy of these models, additionally, regression tree method, and more novel, tree-based algorithms, constructed on longitudinal data, will be estimated on credit union panel data in this thesis. Prediction accuracies of the traditional model and tree-based methods is compared by the mean squared error (MSE). Therefore, the later subsections describe the data of the research and methodology of the previously mentioned panel ARDL model and tree-based methods.

3.1 Data

The data of this master thesis contains main credit union-specific financial variables and country macroeconomic variables. Lithuanian credit union-specific financial data were collected form the Bank of Lithuania database. Most of the latter variables are not available in public. Country macroeconomic variables were extracted from Lithuanian Department of Statistics and data of Lithuanian banks were imported for ECB Statistical Data Warehouse. The data of the study contains 2009 Q2 - 2019 Q2 time period, meaning that each credit union have 41 time observations. Under the period of the study 58 credit unions were operating. In order to have greater number of observations, the missing data of the last two quarters of one credit union was full-filled. Since one credit union had reorganized and united to the other credit union, the data of the latter was demean under assumption that the data of the credit union, that was attached to another one, would be growing by the same path as other, very similarly operating credit union, that had highest correlations with the one attached. Therefore, panel data is balanced and the total time observations T=41 and number of panels N=58, are relatively large. Description of all of the variables used in the study is provided in Appendix A. The data covers main credit union specific financial variables, that indicates efficiency of credit union (OIP, PIP, PIKP, GPMVT, IPVPKTA, GPP, RPM, BAIP, PIKI), loan portfolio quality (SAPL, NPL, VSI), liquidity (LTOD, LR, VVPTA), amount of capital (KPR, PERSK), asset parts (LOANS), commitments (I), concentration (CONC5). Macroeconomic variables were GDP, Infl. Additionaly bank rates wrere considered (BINT, EURIBOR6). Dummy variables: crisis, region.

Variables for panel ARDL model were chosen taking into consideration results of panel co-integration tests, findings from inclusion of variables in the model one-by-one, leaving only significant ones. On contrary, since tree-based methods are not restricted by the number of variables, all of the available variables from the data set were used.

3.2 Determinants of credit unions profitability

In this subsection, more closely will be described differences between chosen profitability indicators and possible economic relationships with depended variables, chosen for the study of panel ARDL model.

3.2.1 Profitability indicators

Two credit union profitability ratios are included in this study: return on assets and real interest margin.

Return on assets ratio is one of the most commonly used profitability indicators for credit institutions. Return on assets ratio shows how efficient credit union asset are used to generate profit. In other words, this ratio shows how much net profit has one amount of average credit union assets. The ratio is calculated by dividing net profit by average assets and multiplying by 4, 2, 3/4 and 1 in the first, second, third and fourth quarters respectively. The larger return on assets ratio is, the more efficiently credit union assets are used.

One of the most popular profitability measures for banks is net interest margin. For banks, this ratio includes only net interest income in the numerator. However, for credit unions this ratio was expanded to the sum of net interest income, net income from commissions and services and interest income from government securities. There were two aspects to do so. One is that, there is a part of credit unions that does not require to buy additional shares in order to get loan, but ask to pay larger administrative taxes, that are accounted as commissions and services income. Additionally, historically, sufficient part of credit union assets had government securities. Therefore, additional statements was included, as they show real interest margin for credit unions and the ratio is called real interest margin. This ratio is calculated by dividing real net income: net interest income from loans, commissions and services, government securities, by the total credit union income. Real interest margin ratio shows how much real net income are generated by one amount of total income. The greater the amount of expenses comparing to amount of income, the lower the real interest margin, consequently, the less profitable credit union is. Differently from return on assets ratio, real interest margin shows efficiency of credit union pricing model.

3.2.2 Profitability determinants

Credit union profitability determinants were chosen by, panel co-integration tests results and estimation of the panel ARDL model to search whether statistical significant relationship and economically logical sign exist. Various types of variables were included as potential credit union profitability determinants.

As credit union efficiency ratio, ratio of operational expenses to total income has been chosen. The ratio shows how much total income are used to cover credit union expenses form operation (expenses for employees salaries, IT, marketing, rent expenses). The lower the ratio, the more efficiently credit union operates. It is expected that the ratio of operational expenses to total income has negative sign on credit union profitability, since greater the ratio, the less income are left to cover other credit union expenses, such as interest expenses, and less probable to operate at least with profit.

For leverage, the ratio of loan loss provisions to total loan portfolio was used. This ratio shows what part of loan portfolio is at the moment probably lost (not payed-back or not fully payed back). This is a ratio of loan portfolio quality: the lower the ratio, the better loan portfolio quality is. It is expected that the growth of this ratio would have negative effect on credit union profitability, since lower loan quality show the risk that the main credit union income from interest of loans are not gained.

Another leverage measure is ratio of non-performing loans to total loans. The main difference from loan loss provisions ratio is that non-performing loans are all loans that are late to pay-back at least for 60 days and there are or there are no loss provisions made by credit union. Therefore, the ratio non-performing loans to total loans is more strict loan quality measure. The greater the ratio, the lower loan portfolio quality. As in loan loss provision case, it is expected that non-performing loans have negative effect on credit union profitability.

Other ratio, used as profitability determinant is capital adequacy ratio, that is one of supervisory requirements. This ratio had to be not smaller than 13 percent until 2018 and after 2018, when methodology has changed, has to grow sequentially to 10,5 percent till 2028. The ratio is calculated by dividing recalculated credit union capital by risk weighted assets. The adequate amount of capital helps to save deposits from possible losses, additionally, the control of this ratio helps to ensure enough coverage of the main credit union risks: credit, market, operational risks. Capital adequacy ratio shows credit union's ability to absorb losses and handle risk exposures, therefore indicate safety and soundness. It is expected, that greater capital adequacy has positive effect on credit union profitability.

Net interest income is calculated as the difference between loan interest income and loan interest expenses. As was mentioned, credit unions, as cooperative institutions, often provide services at more competitive rates comparing to banks. Therefore, it is important to see how strong is the effect of the main intermediary credit union activity: provision of loans and acceptance of deposits. The greater the difference between the loan interest income and loan interest expenses, the more rational credit union pricing from the main intermediary activity is. It is expected that changes in net interest income have positive effect to credit union profitability.

Gross domestic product is used as a measure of country's economic activity. This indicator was often used in studies of bank profitability and was not considered in credit union profitability researches. According to Athanasoglou et. al. (2008), Narusevicius (2018), the reasons why banks' profitability are positively affected by greater economic activity is because under growing economic cycle demand for credit, transactions and other operations usually increase. Moreover, the greater demand of credit allows raising interest margin. It is believed, that economic activity might similarly affect profitability of credit unions, since services such as credit, payments are determinants of credit unions' income.

Another variable used as credit union profitability determinant is average banks' loans interest rates. It is known that banks are main credit providers and deposit holders in most of the countries in the world. Additionally, it was shown in Narusevicius (2018) that Lithuanian banks have economies of scale and scope. As credit unions provide similar services as banks, consequently banks are main credit union competitors. Therefore, it is expected that changes in bank pricing for loans might have a positive effect to credit union profitability: the greater the loan interest are provided by banks, the greater interest rate can be set by credit unions.

3.3 Methodology

3.3.1 Panel data with pooled mean group and mean group estimators

Pesaran *et. al* (1997, 1999) proposed estimator for dynamic panel data, called pooled mean-group (PMG) estimator for panel ARDL model with error correction. The proposed estimator constrains the long term coefficients to be the same across credit unions and allows only the short-term coefficients, error variances and intercepts to vary. The meaning of pooled mean-group is that the estimators are fitted on separate equations for each credit union and coefficients are pooled imposing homogeneity in the long run equation and averaged from individual short run equations. The PMG estimator is similar, to the MG estimator, proposed by Pesaran and Smith (1995), that averages the estimates of separate models for each group and allows to vary short run, long run coefficients, error variances and intercepts.

From Pesaran *et. al* (1999), consider a general panel ARDL$(p, q_1, \dots, q_n)$ model for our data set:

$$y_{it} = \sum_{j=1}^{p} \lambda_{ij} y_{i,t-j} + \sum_{j=0}^{q_l} \delta_{ij}^{'} C_{i,t-j} + \sum_{j=0}^{q_l} \gamma_{ij}^{'} M_{t-j} + \sum_{j=0}^{q_l} \alpha_{ij}^{'} B_{t-j} + \mu_i + \epsilon_{it}, \qquad (1)$$

where $i = 1, 2, \dots, N$ is number of groups (credit unions); $t = 1, 2, \dots, T$ is number of time periods, y_{it} is dependent variable (return on assets, real interest margin) of i-th group, explanatory variables: C_{it} credit union specific variables, M_t macroeconomic variables, B_t bank variables, $y_{i,t-j}$ are lagged dependent variables, λ_{ij} are coefficients of lagged dependent variables (scalars), δ_{ij},γ_{ij}, α_{ij} are coefficient vectors of explanatory variables, μ_i are unobserved fixed effects, ϵ_{it} is the error term.

The following reparametrization of the (1) in order to imply the error correction equation:

$$\Delta y_{it} = \phi_i (y_{i,t-1} + \theta_i^{'} C_{it} + \beta_i^{'} M_t + \kappa i^{'} B_t) + \sum_{j=1}^{p-1} \lambda_{ij}^{*} \Delta y_{i,t-j} + \sum_{j=0}^{q_l-1} \delta_{ij}^{*'} \Delta C_{i,t-j} + $$
$$+ \sum_{j=0}^{q_l-1} \gamma_{ij}^{*'} M_{t-j} + \sum_{j=0}^{q_l-1} \alpha_{ij}^{*'} \Delta B_{t-j} + \mu_i + \epsilon_{it}, \qquad (2)$$

where $\phi_i = -(1 - \sum_{j=1}^{p} \lambda_{ij})$ is the error-correcting speed of adjustment, $\theta_i = \sum_{j=0}^{q} \delta_{ij} / (1 - \sum_k \lambda_{ik})$, $\beta_i = \sum_{j=0}^{q} \gamma_{ij} / (1 - \sum_k \lambda_{ik})$, $\kappa_i = \sum_{j=0}^{q} \alpha_{ij} / (1 - \sum_k \lambda_{ik})$ are vectors containing long-run relationships, $\lambda_{ij}^{*} = - \sum_{m=j+1}^{p} \lambda_{im}, j = 1, \dots, p -$

1, $\delta^*_{ij} = -\sum_{m=j+1}^{ql} \delta_{im}, j = 1, 2, \ldots, q_l - 1$, $\gamma^*_{ij} = -\sum_{m=j+1}^{q_l} \gamma_{im}, j = 1, 2, \ldots, q_l - 1$, $\alpha^*_{ij} = -\sum_{m=j+1}^{q_l} \alpha_{im}, j = 1, 2, \ldots, q_l - 1$.

Pesaran *et. al* (1999) formed following assumptions for this model:

1. the disturbances ϵ_{it} are independently distributed across i, t and regressors, with means equal to 0, variances equal to $\sigma_i^2 > 0$ and finite fourth-order moments.
   However, it was shown in Pesaran and Smith (1995) that it is relatively straightforward to allow for possible dependence of regressors to ϵ_{it} when estimating long-run coefficients, as long as independent variables have finite-order autoregressive representations. Therefore, the following assumption of independence of regressors to $\epsilon_{it}$is not considered.

2. $ARDL(p, q, ..., q)$ is stable: the roots $\sum_{j=1}^{p} \lambda_{ij} z^j = 1, i = 1, 2, \ldots, N$ lie outside unit circle. Therefore, $\phi_i < 0$ and the long-term relationship exists.

3. Long-term coefficients θ_{it} are the same across groups (only for PMG estimator), namely
   $\theta_i = \theta, \beta_i = \beta, \kappa_i = \kappa$.

If variables in equation (1) are I(1) and are co-integrated, ϵ_{it} is I(0) $\forall i$. Additionally, Pesaran *et. al* (1999) noted, that variables might also be I(0) processes, since consistent long-term estimates are also ensured. Co-integration between variables ensure the response to deviations from long-run equilibrium. The parameter ϕ_i is expected to be statistically significant and negative under assumption that variables show a return to long-run equilibrium.

Pesaran *et. al* (1999) proposed maximum likelihood approach to estimate parameters. For likelihood approach Pesaran *et. al* (1999) assume that the disturbances ϵ_{it} are normally distributed, but note that this assumption is not required for the asymptotic results. The authors write likelihood as the product of the likelihoods for each group:

$$l_T(\psi', \varphi', \sigma') = -\frac{T}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{N} \frac{1}{\sigma_i^2}(\Delta y_i - \phi_i \xi_i(\psi))' H_i (\Delta y_i - \phi_i \xi_i), \quad (3)$$

where $\xi_i(\psi) = y_{i,t-1} - X_i\psi_i$, where $X_i = (C_i, M, B)$ is matrix of observations, $\psi_i = (\theta_i, \beta, \kappa)$ coefficients vector, $H_i = I_T - W_i(W_i' W_i)W_i$, where I_T is identity matrix, $W_i = (\Delta y_{i,t-1}, \ldots, \Delta y_{i,t-p+1}, \Delta X_i, \Delta X_{i,t-1}, \ldots, \Delta X_{i,t-q+1})$.

The likelihood is maximized by back-substitution method: starting with initial long-run coefficient vector estimates $\hat{\psi}$, short-run coefficients, speed of adjustment terms, intercepts are estimated under regressions of Δy_i on $(\hat{\xi}_i, W_i)$. Then, the estimates are used for updating $\hat{\psi}$ and the process iterate until convergence (Blackburne and Frank, 2007).

Another, MG estimator, proposed in Pesaran and Smith (1995) is used is this thesis for parameters robustness check. Differently from PMG estimator, in MG estimator intercepts, slope coefficients, and error variances are allowed to differ across groups. The estimator parameters are simple arithmetic means of individual coefficients. As an example, error correction coefficient is estimated:

$$\hat{\phi} = N^{-1} \sum_{i=1}^{N} \hat{\phi}_i. \tag{4}$$

While variance:

$$\hat{\Delta}_{\hat{\phi}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\hat{\phi}_i - \hat{\phi})^2. \tag{5}$$

Other MG coefficients and variances are calculated similarly.

To test which estimator is more suitable for data, therefore, if the long-run homogeneity assumption holds, Hausman (1978) test is applied. This test is used to compare two different estimators, for example, β_0 and β_1, where β_0 is mean group long-term estimator and β_1 is pooled mean group estimator. Hausman statistic is calculated the following:

$$H = (\beta_1 - \beta_0)' (Var(\beta_0) - Var(\beta_1))^{-1} (\beta_1 - \beta_0).$$

The statistics is then compared with critical χ^2 value. Under the null hyphothesis, both estimators are consistent, but second (β_1) is efficient (with smallest variance) and under alternative, the β_0 estimator is solely consistent. Therefore, Hausman test will allow to check long-run homogeneity assumption.

3.3.2 Panel Unit Root and Co-integration Tests

In order to test stationarity and order of integration of panel data, panel unit root and traditional unit root tests are used. Panel unit root tests are specified for panel data while ordinal unit root tests - for time series data.

The main difference between various panel unit root tests is whether the common unit root or individual unit root processes are assumed and tested. Most commonly used panel unit root tests are Levin, Lin and Chu, Im, Pesaran and Shin, Fisher ADF, Fisher PP tests.

Levin, Lin and Chu (2002) proposed a unit root test that employs the assumption about a common unit root process. Under the null hypothesis of this test, each time series contains a unit root and under alternative hypothesis - each time series is stationary. For evaluation of this test, consider a general augment Dickey-Fuller (ADF) process for each credit union:

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{p_i} \theta_{iL} \Delta y_{it-L} + \alpha_i d_t + \epsilon_{it}, \tag{6}$$

where d_t is a vector of one's, α_i is vector of coefficients form the assumed model in Levin, Lin and Chu (2002): $\Delta y_{it} = \alpha_{0i} + \delta y_{it-1} + \zeta_{it}$, where ζ_{it} are independently distributed across groups and stationary invertible ARMA error process. On the next step, two auxiliary regressions are estimated to obtain residuals:

1. $\hat{e}_{it} = \Delta y_{it} - \sum_{L=1}^{p_i} \hat{\pi}_{iL} \Delta y_{it-L} - \hat{\alpha}_i d_i$,
2. $\hat{v}_{i,t-1} = y_{it-1} - \sum_{L=1}^{p_i} \tilde{\pi}_{iL} \Delta y_{it-L} \tilde{\alpha}_i d_t$.

The obtained residuals are standardized by standard error $\hat{\sigma}_{\epsilon_i}$ from (6) equation for every i:

$\hat{e}_{it} = \frac{\hat{e}_{it}}{\hat{\sigma}_{\epsilon_i}}$, $\hat{v}_{i,t-1} = \frac{\hat{v}_{it}}{\hat{\sigma}_{\epsilon_i}}$.

On the next step, ratio of long-run to short-run standard deviations are calculated. Under null hypothesis of unit root, long run variance of (6) is calculated:

$$\hat{\sigma}_{yi} = \frac{1}{T-1} \sum_{t=2}^{T} \Delta y_{it}^2 + 2 \sum_{L=1}^{\overline{K}} w_{\overline{K}L} \left[\frac{1}{T-1} \sum_{t=2+L}^{T} \Delta y_{it} \Delta y_{it-L} \right],$$

where $\overline{K}$ is truncation lag parameter, shown in Levin, Lin and Chu (2002), $w_{\overline{K}L}$ is sample covariance weights, $w_{\overline{K}L} = 1 - \frac{L}{\overline{K}+1}$.

Then, long-run standard deviation to innovation standard deviation is calculated:

$\hat{s}_i = \frac{\hat{\sigma}_{yi}}{\hat{\sigma}_{\epsilon_i}}$.

On the final step, pooled OLS regression is fit:

$\tilde{e}_{it} = \rho \tilde{v}_{i,t-1} + \tilde{\epsilon}_{it}$

t statistic for testing $\rho = 0$:

$$t_\rho = \frac{\hat{\rho}}{STD(\hat{\rho})},$$

where $\hat{\rho} = \frac{\sum_{i=N}^{N} \sum_{t=2+p_i^T} \tilde{v}_{it-1} \tilde{e}_{it}}{\sum_{i=N}^{N} \sum_{t=2+p_i}^{T} \tilde{v}_{it-1}^2}$, $STD(\hat{\rho}) = \hat{\rho}_\epsilon \left[\sum_{i=N}^{N} \sum_{t=2+p_i}^{T} \tilde{v}_{it-1}^2 \right]^{-1/2}$,

$\hat{\rho}_\epsilon = \left[\frac{1}{N\tilde{T}}\sum_{i=N}^{N}\sum_{t=2+p_i}^{T}(\tilde{e}_{it} - \hat{\rho}\tilde{v}_{it-1})^2\right].$

Under the null hypothesis, $\rho = 0$, meaning that all groups have unit root, and under alternative $\rho < 0$, all units are stationary. As was noted by Levin, Lin and Chu (2002), statistic works well when N is between 10 and 250, T between 5 and 250.

Another test proposed by Im-Pesaran-Shin (2003), allows ρ coefficients to vary across cross-sections and is called an individual unit root process. General form of the following ADF equation:

$$\Delta y_{it} = \alpha_i + \rho_i y_{i,t-1} + \sum_{j=1}^{p_i} \beta_{ij}\delta y_{i,t-j} + \epsilon_{it},$$

$i = 1, ..., N, t = 1, ..., T.$
For this equation ADF unit root test statistics are calculated for each credit union i and averaged across groups. When t statistic is the following:

$$t_{NT} = \frac{1}{N}\sum_{i=1}^{N} t_{iT}(\rho_i\beta_i),$$

$$t_{iT}(\rho_i\beta_i) = \frac{\sqrt{T-\rho_i-2}(\mathbf{y}_{i,-1}^{'}\mathbf{M}_{Q_i}\Delta\mathbf{y}_i)}{(\mathbf{y}_{i,-1}^{'}\mathbf{M}_{Q_i}\mathbf{y}_{i,-1})^{-1/2}(\Delta\mathbf{y}_i^{'}\mathbf{M}_{\mathbf{x}_i}\Delta\mathbf{y}_i)^{1/2}},$$

where:

$\beta_i = (\beta_{i1}, \ldots, \beta_{ip_i})^{'}$, $\mathbf{y}_{i,-1} = [y_{i0}, \ldots, y_{iT-1}]^{'}$, $\Delta\mathbf{y}_{i,-s} = [\Delta y_{i1-s}, \ldots, \Delta y_{iT-s}]^{'}$, $s = 0, \ldots, \rho_i$, $\Delta\mathbf{y}_i = \Delta\mathbf{y}_{i,-0}$, $\mathbf{Q}_i = [(1, \ldots, 1)^{'}, \Delta\mathbf{y}_{i,-1}, \ldots, \Delta\mathbf{y}_{i,-p_i}]$, $\mathbf{M}_Q = \mathbf{I}_T - \mathbf{Q}_i(\mathbf{Q}_i^{'}\mathbf{Q}_i)^{-1}\mathbf{Q}_i$, $\mathbf{X}_i = [\mathbf{y}_{i,-1}, \mathbf{Q}_i]$, $\mathbf{M}_{\mathbf{X}_i} = \mathbf{I}_T - \mathbf{X}_i(\mathbf{X}_i^{'}\mathbf{X}_i)^{-1}\mathbf{X}_i$.

Under the null hypothesis of this test, all individuals have unit root: $\rho_i = 0, \forall i$ and under alternative - some individuals have unit root:

$H_1 : \begin{cases} \rho_i < 0, i = 1, 2, \ldots, N_1 \\ \rho_i = 0, i = N_1 + 1, \ldots, N \end{cases}$

Other alternative of tests - Fisher ADF and Fisher PP tests, proposed by Maddala and Wu (1999), combine p-values from individual unit root tests.

The formula of the test:

$$P = -2\sum_{i=1}^{N} log(\pi_i) \rightarrow \chi^2_{2N}$$

where, π_i are significance levels (p-values) from individual unit root test (ADF or PP) for cross-section i.

For Fisher ADF test, individual unit root tests are calculated considering equation:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{s=1}^{m} a_s \Delta y_{t-s} + \varepsilon_t. \quad (7)$$

Under null hypothesis, $\gamma = 0$, under alternative $\gamma < 0$. When t-ratio is calculated for γ:

$$t_\gamma = \frac{\hat{\gamma}}{se(\hat{\gamma})}, \quad (8)$$

where $se(\hat{\gamma})$ indicates coefficient standard error.

For Fisher PP test, individual unit root tests are calculated considering equation:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \varepsilon_t. \quad (9)$$

t-ratio is then calculated using non-parametric method:

$$\tilde{t}_\gamma = t_\gamma \left(\frac{\alpha_0}{f_0}\right)^{\frac{1}{2}} - \frac{T(f_0 - \alpha_0)(se(\hat{\gamma})}{2 f_0^{\frac{1}{2}} s}, \quad (10)$$

where $se(\hat{\gamma})$ is coefficient standard error, s is the standard error of the test regression, $\alpha_0 = \frac{(T-k)s^2}{T}$ is estimate of the error variance in (9), where k is number of regressors and f_0 is estimator of the residual spectrum at frequency zero, calculated using kernel-based sum of covariances. The latter estimator is calculated using the weighted sum of autocovariances:

$$f_0 = \sum_{j=-(T-1)}^{T-1} \hat{\gamma}(j) K(j/l), \quad (11)$$

where K is a Bartlett Kernel function, l is a bandwidth, $\hat{\gamma}(j) = \sum_{t=j+1}^{T} (\varepsilon_t \varepsilon_{t-j})/T$ is autocovariance of j-th sample of residuals ε_t.

For macroeconomic variables, ADF, PP and KPSS tests are applied. ADF and PP tests' procedures are shown in (8) - (10). Under the null hypothesis of these tests, unit root is present in time series, while under alternative, time series is stationary. Another type of unit root tests, KPSS test, use the following equation:

$$y_t = r_t + \beta t + \vartheta_1,$$

where βt is is deterministic part, r_t is a random walk and ϑ_1 is stationary error.

Then, LM statistic is calculated:

$$LM = \sum_t S(t)^2/(T^2 f_0)$$

$S(t) = \sum_{r=1}^{t} = \vartheta_r$ is cumulative residual function, f_0 is estimator of the residual spectrum at frequency zero from (11), only with residuals ϑ_t. For KPSS test, the null hypothesis is that time series is (trend-) stationary and under alternative - unit root exists.

For co-integration testing, seven residual-based statistics proposed by Pedroni (1999) are computed. From hypothesized cointegration regressions residuals are computed. In general case:

$$y_{i,t} =_i + \beta_{1i} x_{1i,t} + \cdots + \beta_{Mi} x_{Mi,t} + e_{i,t}$$

$$\Delta y_{i,t} = \sum_{m=1}^{M} \beta_{mi} \Delta x_{1m,t} + \eta_{i,t}$$

$$\hat{e}_{i,t} = \hat{\gamma}_i \hat{e}_{i,t-1} + \hat{\mu}_{i,t}$$

$$\hat{e}_{i,t} = \hat{\gamma}_i \hat{e}_{i,t-1} + \sum_{k=1}^{K} \hat{\gamma}_{i,k} \Delta e_{i,t-k} \hat{+} \hat{\mu}^*_{i,t},$$

where $m = 1, 2, \ldots, M$ is number of regressors, $k = 1, 2, \ldots, K$ is number of lags.

$y_{it} = \alpha + \delta_i t + \beta_{1i} x_{1it} + \cdots + \beta_{Mi} x_{Mit} + e_{it}$, $t = 1, \ldots, T$; $i = 1, \ldots, N$, M is the number of regression variables, T number of observations, N number of individual members. From these equations estimated residuals are tested over seven Pedroni's statistics:

1. Panel v statistics:

$T^2 N^{\frac{3}{2}} Z_{\hat{v}_{Nt}} \equiv T^2 N^{\frac{3}{2}} (\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{it-1}^2)^{-1}$,

2. Panel ρ statistics:

$T\sqrt{N} Z_{\hat{\rho}_{NT-1}} \equiv T\sqrt{N} (\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{it-1}^2)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} (\hat{e}_{it-1} \Delta \hat{e}_{it} - \hat{\lambda}_i)$,

3. Panel t-statistics:

$Z_{t_{NT}} \equiv (\hat{\sigma}_{NT}^2 \sum_{i=1}^{N} \sum_{i=1}^{T} \hat{L}_{11i}^{-2} e_{it-1}^2)^{\frac{-1}{2}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{2} (\hat{e}_{it-1} \Delta \hat{e}_{it} - \hat{\lambda}_i)$,

4. Panel ADF statistics:

$Z^*_{t_{NT}} \equiv (\hat{s}^{*2}_{NT} \sum_{i=1}^{N} \sum_{i=1}^{T} \hat{L}^{-2}_{11i} e^{*2}_{it-1})^{\frac{-1}{2}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}^{2}_{11i} (\hat{e}^*_{it-1} \Delta \hat{e}^*_{it})$,

5. Group ρ statistics:

$TN^{\frac{-1}{2}} \hat{Z}_{\hat{\rho}_{NT-1}} \equiv TN^{\frac{-1}{2}} \sum_{i=1}^{N} (\sum_{t-1}^{T} \hat{L}^{-2}_{11i} \hat{e}^2_{it-1})^{-1} \sum_{t=1}^{T} (\hat{e}_{it-1} \Delta \hat{e}_{it} - \hat{\lambda}_i)$,

6. Group t statistics:

$N^{\frac{-1}{2}} \hat{Z}_{t_{NT}} \equiv N^{\frac{-1}{2}} \sum_{i=1}^{N} (\hat{\sigma}^2_i \sum_{t=1}^{T} \hat{e}^{*2}_{it-1})^{\frac{-1}{2}} \sum_{t=1}^{T} (\hat{e}_{it-1} \Delta \hat{e}_{it} - \hat{\lambda}_i)$,

7. Group ADF statistics:

$N^{\frac{-1}{2}} \hat{Z}_{t^*_{NT}} \equiv N^{\frac{-1}{2}} \sum_{i=1}^{N} (\sum_{t=1}^{T} \hat{s}^{*2}_i \hat{e}^{*2}_{it-1})^{-1} \sum_{t=1}^{T} (\hat{e}_{it-1} \Delta \hat{e}_{it})$,

where:

$\hat{\lambda}_i = \frac{1}{T} \sum_{s=1}^{k_i} (1 - \frac{s}{k_i+1}) \sum_{t=s+1}^{T} \hat{\mu}_{it} \mu_{it-s}$,

$\hat{s}^2_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}^{*2}_{it}$,

$\hat{\sigma}^2_i = \hat{s}^2_i + 2\hat{\lambda}_i$,

$\hat{\sigma}^2_{NT} = \frac{1}{N} \sum_{i=1}^{N} \hat{L}^{-2}_{11i} \hat{\sigma}^2_i$,

$\hat{s}^{*2}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}^{*2}_{it}$,

$\hat{s}^{*2}_{NT} = \frac{1}{N} \sum_{i=1}^{N} \hat{s}^{*2}_i$,

$\hat{L}^{-2}_{11i} = \frac{1}{T} \sum_{t}^{1} \hat{\eta}^2_{it} + \frac{2}{T} \sum_{s=1}^{k_i} (1 - \frac{s}{k_i+1}) \sum_{s+1}^{T} \hat{\eta}_{it} \hat{\eta}_{it-s}$.

The proposed tests are constructed for non-stationary heterogeneous panels with large T and large N.

For the lag selection of panel ARDL model, BIC (Bayesian information criterion) criterion is used. BIC criterion has a penalty term for the growth of number of parameters (model complexity), that allows to filter unnecessary number of parameters. The model with the lowest BIC value is preferred. BIC is used for (1) for each individual credit union equation. The lag of general pooled mean group and mean group models is selected from the most common lags among credit unions.

BIC formula:

$$BIC = \ln(n)k - 2\ln(\hat{L}),$$

where $\hat{L} = p(x|\hat{\theta}, M)$ is maximized value of likelihood function of the model M, where $\hat{\theta}$ are values that maximize the likelihood function, x are data, n is sample size, k is number of parameters in the model.

For residual analysis, residual independence across cross-sections and time is tested using Pesaran CD test and autocorrelation plots respectively. Pesaran

(2004) proposed a CD test for residual independence across cross-sections:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij} \right),$$

where $\hat{\rho}_{ij}$ is pairwise correlation of model residuals:

$$\hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^{T} \hat{\epsilon}_{it}\hat{\epsilon}_{jt}}{(\sum_{t=1}^{T} \hat{\epsilon_{it}^2})^{1/2}(\sum_{t=1}^{T} \hat{\epsilon}_{jt}^2)^{1/2}}$$

This test is directly applied on panel data and under null hypothesis state the presence of residual cross-sectional independence. An alternative is cross-sectional dependence in residuals. The test is proper for larger T and panel data models for both homogeneous or heterogeneous panels, nonstationary models (Hoyos, Sarafidis, 2006).

Additionally, to check, whether residuals are independently distributed across t, autocorrelation plots are used for each credit union residuals. Autocorrelation at the j-th time lag is calculated: $r_j = c_j / c_0$,
where $c_j = \frac{1}{N} \sum_{t=1}^{N-j} (Y_t - \bar{Y})(Y_{t+j} - \bar{Y})$, $c_0 = \frac{1}{N} \sum_{t=1}^{N} (Y_t - \bar{Y})$.
Bounds of α significance level is calculated: $B = z_{1-\alpha/2} se(r_j)$, where $se(r_j)$ is standard error of autocorrelation at the j-th lag.

3.3.3 Tree-based methods

Other methods, used for credit unions profitability prediction are tree-based methods. The most common tree-based method is CART (classification and regression trees), proposed in Breiman *et. al* (1984). The idea of fitting CART starts from division of all predictors x_i into J distinct non-overlapping regions (groups) R_j, where $j = 1, \ldots, J$. Divisions into regions are such, that sum of squares error are minimized by parameters j, c:

$$\sum_{i:x_i \in R_1(j,c)} (y_i - \hat{\mu}_{R_1})^2 + \sum_{i:x_i \in R_2(j,c)} (y_i - \hat{\mu}_{R_2})^2, \quad (12)$$

where $\hat{\mu_{R_1}}$ and $\hat{\mu_{R_2}}$ are the mean responses of observations within regions R_1, R_2 , c is a cut-point for splitting predictor into regions.

At first step, data is split into two regions by some c. Then the process is repeated to split the data from two previous regions further to minimize sum of squares error, taking into account rest of predictors. The splitting process continues until a stopping criterion is reached. Since it is known that regression trees are likely to overfit the data, tree-pruning is used. The pruning of a tree is done by introducing complexity parameter α to the tree size. This parameter controls between data fit and tree complexity and is written:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \to R_m} (y_i - \hat{\mu}_{R_m})^2 + \alpha |T|,$$

where $|T|$ is a number of terminal nodes of tree T, $\mathrm{T} \subset \mathrm{T}_0$ and T_0 is very large tree.

Another algorithms including random forests and boosted trees methods had been recently developed to longitudinal data in Sexton (2018). These methods are available in **htree** package in **R**. According to Sexton (2018) **htree** (in other words, historical regression trees) algorithm is an extension of standard regression trees, since the split is done under known tree method and ensembles of such trees are formed either via boosting or as in random forest. Alternatively, this algorithm divides predictors in to two types of categories: concurrent (at time t_{ij}) variables and historical (prior to time t_{ij}). In other words, the response for this method depends on all of its historical realizations as well as time-varying predictor variables.

Data are assumed to be in form:

$$z_{ij} = (y_{ij}, tij, x_{ij}),$$

where $i = 1, ..., n$ are subjects (credit unions), $j = 1, ..., n_i$ observations, y_{ij} is dependent variable (return on assets, real interest margin), t_{ij} is the time of the j-th observation on the i-th credit union and x_{ij} is a vector of independent variables, varying in time t_{ij}. The concurrent predictors for y_{ij} are the elements of the vector (t_{ij}, x_{ij}) while a historic predictor is the set of all values prior to time t_{ij} of a given element of (y_{ij}, t_{ij}, x_{ij}) for credit union i.

The set of all observations for credit union i prior to its j-th observation can be written:

$$\bar{z}_{ij} = \left[z_{ik} : t_{ik} < t_{ij} \right],$$

therefore, $\bar{z}_{ij}$ is the history of i up to t_{ij}.

The history of each variable up to some point for a given credit union i, is represented using a summary function. Letting $\bar{z}_{ijk}$ denote the set of historical values of the k-th element of of z_{ij}, the summary function is denoted $s(\eta; \bar{z}_{ijk})$ where η is the argument vector of the summary function. The summary function is calculated the following:

$$s(\eta;\bar{z}_{ijk}) = \sum_{h:t_{ij}-\eta_1 \le t_{ih} \le t_{ij}} I(z_{ijk} < \eta_2)$$

This summary function counts the number of past values of $\bar{z}_{ijk}$ that are less than η_2, but within η_1 units of time of t_{ij}. For a fixed value of η, each observation history is reduced to a single number and finding the best split for that (fixed) η value is done in the same manner as for a standard regression tree. Which value of η that will produce the most beneficial split is not known. Therefore, the argument η is determined at the node partitioning. η_1 and η_2 are calculated by quantile method. By the default, this method samples 20 quantiles from the time vector $t = 1, 2, \ldots, 41$. These sampled quantiles are possible upper limits of the historical time observations. The lower limits are vector of replicated value lowest time period from quantile method. For credit union data, the algortihm generates about 5 billion various numbers of past observations credit union time vector, with different upper limits of length 20.

Node splitting based on a historical predictor is done by solving:

$$(\hat{k}, \hat{\eta}, \hat{c}, \hat{\mu}) = \text{argmin} \sum_{(ij) \in Node} (y_{ij} - \mu_L I(s(\eta;\bar{z}_{ijk} < c) - \mu_R I(s(\eta;\bar{z}_{ijk} \ge c))^2$$

where indexes $_L$ and $_R$ shows whether node is split to the left or to the right.

The solution to equation above not only selects which predictor history to split the node, but it also determines which summarization of the history produces the most beneficial split. Node splitting on a concurrent predictor follows the approach in standard regression trees, mentioned above.

Ensemble of trees for the above algorithm is formed using boosting and random forests methods. For random forest splitting is done for bootstraped random samples of training set B times and a size of randomly sampled subset of potential predictors is equal to $\sqrt{p}$, where p is number of all predictors. When ensemble of trees and therefore, separate prediction models, are constructed, the prediction values from trees are averaged. The optimal accuracies of random forest are selected using cycle, with different node size and number of trees parameters. For this method, time periods are not sampled, only credit unions.

Boosted trees method is initialized with constant value, which is predicted by:

$F_0(x) = \underset{\gamma}{\text{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$,

where γ is predicted value that minimizes loss function $L(y_i, \gamma)$.

Next, four further steps are looped for $m = 1, \ldots, M$, where M is the last tree.

1. Pseudo residuals are calculated by computing derivative of a loss function:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i}\right]_{F(x)=F_{m-1}(x)},$$

   where $i = 1, \ldots, n$, $L(y_i, F(x_i))$ is a differentiable loss function.

2. On the next step, a regression tree to the pseudo residual values are fit in order to create terminal regions $R_{jm}, j = 1, \ldots, J_m$. From this step, method predicts not response, but pseudo residuals.

3. Then, new predictions $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij} L(y_i, F_{m-1}(x_i + \gamma)}$ are computed from each leaf in a new tree. In this step, previous prediction and new γ is taken into account.

4. $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m \gamma_{jm} I(x \in R_{jm})}$, where ν is learning rate value between 0 and 1, that reduces the effect each tree has on final prediction.

Then the process is repeated till the M-th tree, the output is $F_m(x)$.

After calculating predictions from the above methods and models, predictive accuracies are calculated using mean squared error:

$$MSE = \frac{1}{q} \sum_{i=1}^{n+q} (y_i - \hat{y}_i)^2,$$

where q are data points that were not used in estimating the model/method (test data), n are data points used for model/method estimation.

4 Empirical results

4.1 Model estimation and residual analysis

4.1.1 Model estimation

Before estimation of panel ARDL model with PMG and MG estimators, the order of integration and existing co-integrating relationships are investigated.

To examine stationarity properties of data, panel unit root tests are applied to credit union-specific data and ordinal unit root tests are used for macroeconomic data. As was noted in Pesaran *et. al* (1999), ARDL model provides consistent estimates of the long-term coefficients, for both I(1) and I(0) regressors. Therefore, I(2) regressors should not be included in the model. The main tests for panel unit root testing are Levin, Lin and Chu, Im, Pesaran and Shin, Fisher ADF and Fisher PP tests. Levin, Im and Chu test assume for a common unit root process and is more restrictive. Under the null hypothesis of this test all cross sections have a unit root and under alternative hypothesis, test suggests that each of the time series is stationary. Another test by Im, Pesaran and Shin test is more flexible since it allows for parameters ρ to vary across cross sections. Alternatively, Fishers' ADF and PP tests use p-values from unit root tests for each cross-section i. Under the null hypothesis of Im, Pesaran and Shin, Fisher ADF, Fisher PP tests, all individuals follow a unit root process, under alternative - a some part of the individual processes are stationary.

The above mentioned tests have been applied to profitability ratios and other credit union-specific variables that will be included the in panel ARDL model. The results are shown in Table 1. The findings from results of panel unit root tests show that in almost all cases, tests for return on assets and real interest margin do not reject the null hypothesis, that suggests the existence of unit root in data at levels. On contrary, when data are integrated of order one, p-values of all of the tests' values are smaller than critical 0.05 level, and, therefore, allows to conclude that dependent variables are stationary under integration of order one. Similar results are seen for other credit union-specific variable panel unit root test in Table 1. Majority of the tests show the existence of unit root in data at levels and stationarity under first differences. Therefore, all of the variables, shown in Table 1, are stationary under integration of order one (I(1)).

Table 1: Panel Unit root tests for credit union-specific data

Test	Levin, Lin, Chu		Im, Pesaran, Shin		Fisher ADF		Fisher PP	
Variable	Level	FD	Level	FD	Level	FD	Level	FD
ROA	-0.57	-29.82*	1.38	-34.96*	146.00*	1134.72*	142.34	1345.43*
RPM	-1.1	-31.1*	1.27	-36.7*	146.1*	1177.2*	143.0*	1362.6*
SAPL	-1.5	-18.0*	0.7	-22.9*	102.1	729.7*	98.1	1303.5*
KPR	3.8	-19.4*	6.3	-24.8*	62.4	814.8*	76.8	1717.3*
OIP	-2.2*	-19.3*	0.4	-26.1*	95.5	855.6*	105.1	1748.6*
GPP	11.1	-28.2*	-4.8	-32.2*	243.5	-1113.5	1292.7*	1292.7*
NPL	-10.2*	0.9*	0.7	-33.0*	67.9	1123.3*	7.8	1465.6*

* indicates what p-value is smaller than 0.05 significance level.

For macroeconomic and bank-specific variables, such as logarithm of gross domestic product and bank loan interest rates, augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests have been used. ADF and PP tests under the null hyphothesis state that time series has unit root, under alternative - time series is stationary. Alternatively, for KPSS test, the null hyphothesis is that time series is stationary around a deterministic trend, alternative - the presence of unit root in times series. From the Table 2, it is seen that all three tests support the existence of unit root process at levels, for both logarithm of gross domestic product and bank loan interest rates. However, when macroeconomic variables are first differenced, all three tests show that data are stationary (for ADF and PP tests) and trend stationary (KPSS test). It is seen that there are no contradictions between tests. Therefore, it can be concluded, that credit union-specific financial data and macroeconomic data are I(1) processes.

Table 2: Unit root tests for macroeconomic variables

Test	ADF		PP		KPSS	
Macroeconomic variables	Level	FD	Level	FD	Level	FD
log(GDP)	-2.299	-3.4338*	-7.8766	-33.407*	1.0875*	0.073245
BINT	-2.874	-3.3057**	-7.9128	-44.338*	0.98613*	0.11041

* p-value is smaller than 0.05 significance level; ** p-value is smaller than 0.1 significance level.

On the next step, to check whether the long term co-integrating relationships between variables exist, panel data co-integration tests, proposed by Pedroni (1999), have been used. Pedroni (1999) derived seven residual-based co-integration statistics. Panel v-statistic, Panel rho-statistic, Panel PP-statistic and

Panel ADF-statistics are based on within-dimension while Group rho-statistic, Group PP-statistic, Group ADF-statistics are based on between-dimension. Therefore, the first set of statistics are for testing co-integration in homogeneous panels and the second set of statistics are for testing cointegration in heterogeneous panels. The null hypothesis of no co-integration for the panel co-integration test is the same for each statistics. Under alternative hypothesis for within statistics, there is a common co-integration between variables of interest, and for between statistics the individual co-integration between variables exist. Following Naruševičius (2018), panel co-integration tests have been applied to bivariate cases: for credit union profitability ratios and each of explanatory variable and, additionally, to credit union profitability ratios and group of potential long-run equation explanatory variables.

Table 3 provides results of bivariate panel co-integration statistics for profitability and each of chosen explanatory variables. It is seen that at a 5 percent significance level, for logarithm of gross domestic product and loan loss provisions all of seven statistics reject null hypothesis of no co-integration and accepts alternative of existing co-integration in the data. For the ratio of operating expenses to total income and capital adequacy ratio six out of seven statistics reject the null hypothesis of no cointegration and accept alternative. Therefore, between credit unions return on assets and logarithm of gross domestic product, ratio of operating expenses to total income, loan loss provisions and capital adequacy ratio, bivariate panel co-integration exist.

Table 3: Bivariate panel cointegration tests (return on assets)

Variable name	Panel v-statistic	Panel rho-statistic	Panel PP-statistic	Panel ADF-statistic	Group rho-statistic	Group PP-statistic	Group ADF-statistic
log(GDP)	4.204*	-23.829*	-19.221*	-14.464*	-17.927*	-21.367*	-16.066*
OIP	1.146	-23.541*	-19.103*	-13.916*	-18.176*	-21.611*	-16.186*
SAPL	12.021*	-22.358*	-18.246*	-12.347*	-14.934*	-18.411*	-14.118*
KPR	-0.231	-20.087*	-16.649*	-11.109*	-15.661*	-18.917*	-13.967*

* indicates what p-value is smaller than 0.05 significance level.

Table 4 demonstrates the results of bivariate panel co-integration statistics for another profitability measure - real interest margin, and each of the chosen explanatory variables. At a 5 percent significance level, the null hypothesis of no co-integration is rejected according to all seven statistics for logarithm of gross domestic product, bank loan interest rates and net interest income. For non-performing loans, six out of seven statistics reject the null hypothesis and accept the alternative of existing co-integration in the data. It is seen that for both profitability ratios co-integration between chosen variables exist, thus, panel ARDL model with the error correction can be further estimated.

Table 4: Bivariate panel cointegration tests (return on assets)

Variable name	Panel v-statistic	Panel rho-statistic	Panel PP-statistic	Panel ADF-statistic	Group rho-statistic	Group PP-statistic	Group ADF-statistic
log(GDP)	7.284*	-6.115*	-6.615*	-7.935*	-2.170*	-4.248*	-6.783*
BINT	2.903*	-5.866*	-6.531*	-7.972*	-2.555*	-4.834*	-7.071*
NPL	2.299*	-3.628*	-5.807*	-6.224*	-0.693	-3.340*	-4.993*
GPP	3.418*	-7.463*	-8.527*	-7.944*	-3.867*	-6.346*	-6.558*

* indicates what p-value is smaller than 0.05 significance level.

Additionally, in order to choose long term variables, panel co-integration tests have been applied to credit union profitability ratios and group of potential long term variables. Also, panel ARDL models with PMG were estimated in order to see primary information of statistical significance of variables (including error correction term). It was seen that most appropriate long term variables for return on assets are logarithm of gross domestic product and ratio of operating expenses to total income. In the Table 11 in Appendix B co-integration tests results for credit union return on assets and both logarithm of gross domestic product and ratio of operating expenses to total income are shown. The results of majority of tests indicate the presence of long run relationship between variables. Same steps are done for credit union real interest margin and other potential long term variables. It was indicated that logarithm of gross domestic product and loan interest rates of banks are important long term determinants. Moreover, co-integration test between these variables in Table 12 in Appendix B show that six out of seven tests rejects null hypothesis and accepts alternative of co-integration.

After finding co-integrating relationships between variables, the next step before model estimation is setting the proper lag order of the ARDL equations (long term variables). As in Pesaran *et. al* (1999), Bayesian informational criterion (BIC) criterion was exploited in order to find the best model. This criterion estimate the quality of each model, relative to other models and, therefore, allow to choose a better model. The lag order is estimated for each credit union separately and in contribution to Pesaran *et. al* (1999), the most common lag order for majority of credit unions has been chosen. Lag orders provided by BIC criterion for each credit union and different profitability ratios are shown in the Table 13 in Appendix C. It is seen that most common lag order for the ARDL equation was ARDL(1,0,0) for 33 credit unions for return on assets panel ARDL model with error correction, while for real interest margin model 34 credit unions have lag order equal to ARDL(1,0,0). Therefore, only the lag of dependent variable will be included in the models.

The results of fitted ARDL models with pooled mean group (PMG) and mean group (MG) estimators for return on assets are shown in Table 4, while the results for real interest margin model are in Table 5 . As expected, both estimators in Table 4 show that the error correction term is negative and statistically significant, meaning that the speed of adjustment to the long run equilibrium exists and is equal to -0.50 and -0.56 respectively. Negative sign indicates the return of short run disequilibrium to the long run equilibrium. Results of PMG estimator show that credit union return on assets in short term is positively affected by changes in capital adequacy ratio. Alternatively, greater changes in the loan loss provisions, the lower is credit union profitability. Interpretation can be that if credit union has larger amount of capital to cover possible risks, meaning greater ability to absorb losses, the more profitable credit union is. On contrary, if credit union faces growing part of loans that are not payed-back, this have negative consequences in the short run credit union profitability. In the long run equation of PMG estimator, the results show that growth of country's economic activity, estimated by logarithm of GDP, has positive and statistically significant effect to credit union profitability. Meanwhile the ratio of operating expenses to total income has negative effect on the long run profitability, meaning that growth in part of operating expenses that are not fully, or significantly, covered by income gains lowers credit union profitability in the long run. MG estimator coefficients in short run and long run equations share similar coefficient signs and most of statistics significance, as in model with PMG estimator. In model with MG estimator ratio of operating expenses to total income does not have statistically significant relationship to profitability. However, it is seen that in most of the cases, standard errors are larger in the model with MG estimator.

Table 5: ARDL(1,1,1) PMG and MG estimator results (ROA)

	PMG estimator		MG estimator	
Variable	Coefficient	S.E.	Coefficient	S.E.
Long run equation				
Log(GDP)	1.451046*	0.354782	1.55789**	0.897283
OIP	-0.012656*	0.003808	-0.0049306	0.009869
Short run equation				
EC	-0.50248*	0.024533	-0.5598179*	0.027897
ΔSAPL(-1)	-0.671166*	0.072961	-0.640218**	0.071252
ΔKPR	0.100498*	0.020698	0.1069035*	0.021902
Intercept	-6.338568*	0.322095	-10.15247*	3.702293

* significant at 5 percent significance level; ** significant at 10 percent significance level

From the Table 5 it is seen that for PMG estimator, economic activity and average banks' loan interest rates have positive and statistically significant effect on real interest margin. As in previous models, the error correction term is negative and statistically significant for both PMG and MG estimators. Therefore, the existence of the estimated long run relationship is confirmed. The positive sign of economic activity coefficient allows to assume that in upward economic cycle, the growing demand of credit enables to increase loan interest rates, income of commissions and services at a faster pace comparing to expenses. The sign of average banks' loan interest rates indicates that credit unions' pricing model is positively related to banks, that have large part of the loan market. Short term equation shows that a change in credit unions' net interest income has positive and statistically significant effect in the short term, meaning that the better credit unions' pricing model from the main credit unions' activity - intermediation is, the greater real interest margin is. Therefore, good pricing model of the main activity is important real interest margin determinant in the short run. Alternatively, a change of ratio of non-performing loans in the short run, negatively affects real interest margin, meaning that lower the quality of loans, lessens credit union ability to have better pricing model. Similarly, in the case of MG estimator, coefficients, signs, statistical significance do not differ much from PMG estimator. However, it is also seen that majority of standard errors are larger in model with MG estimator.

Table 6: ARDL(1,0,1) PMG and MG estimator results (RIM)

	PMG estimator		MG estimator	
Variable	Coefficient	S.E.	Coefficient	S.E.
Long run equation				
Log(GDP)	3.169042*	0.391482	4.418961*	1.4325
BINT	0.171689*	0.014064	0.259001*	0.066543
Short run equation				
EC	-0.333212*	0.025750	-0.391316*	0.028388
ΔGPP	0.58675*	0.151507	0.6102768*	0.169929
ΔNPL(-1)	-0.037705*	0.005319	-0.036374*	0.005146
Intercept	-3.514292*	0.277822	-9.770979*	4.594521

* significant at 5 percent significance level

In order to check which model estimates, PMG or MG, are more proper, the Hausman test is applied. Hausman (1978) proposed a test that enables comparisons between different estimators. This test allows to estimate the difference between the MG and PMG estimators and, therefore, to test which of two estimates should be used. Under the null hypothesis, the difference between coefficients is not systematic and PMG estimator is efficient, under alternative - the difference between coefficients is systematic, meaning that PMG estimator is inconsistent and MG estimator should be used instead. From the Table 7 it is seen that for both return on asset ratio and real interest margin, the null hypothesis is rejected with a 5 percent significance level. The results allow to conclude that PMG estimator is efficient and the long term homogeneity assumption is proper on credit unions. Therefore, further analysis (prediction) is done only on focus on ARDL models with PMG estimator.

Table 7: Hausman test results

Panel ARDL for ROA				
	Coefficients (MG)	Coefficients (PMG)	Difference	SE
log(GDP)	4.797335	4.284937	0.5123979	0.7906165
OIP	-0.0480416	-0.0527557	0.0047141	0.010959
Chi.2 3.22		P-value 0.1999		
Panel ARDL for RIM				
	Coefficients (MG)	Coefficients (PMG)	Difference	SE
log(GDP)	4.418961	3.169042	1.24992	1.539218
BINT	-0.2590009	0.1716887	0.0873122	0.0724239
Chi.2 2.92		P-value 0.2317		

The final model estimation step is to check if residual assumptions are met in the model. One of the PMG estimator assumptions is that residuals are independently distributed across groups and time, with mean equal to zero and variances across cross-sections greater than zero. To check for residual independence across cross-sections, Pesaran CD test has been chosen, since CD statistic is valid for larger number of N. As is seen in Table 8, for estimated ARDL model on return on assets, the null hypothesis of cross-sectional independence is not rejected at 1 percent significance level. For estimated ARDL model on real interest margin, the null hypothesis of no cross-sectional dependence is not rejected at five percent significance level. Therefore, the assumption of residual independence between credit unions is assumed to be satisfied at a 1 percent significance level.

Table 8: Pesaran CD test results

	z	p-value
ROA model	2.1706	0.02996
RIM model	-0.84994	0.3954

Residual independence across time is checked by individual auto-correlation plots for identification of presence o higher-order serial correlation. Auto-correlation plots are shown in figures in the Appendix E for return on assets model and Appendix F for real interest margin. It is seen, that major part of credit unions residuals are not serially correlated for both return on assets and real interest margin models.

Additionally, in Figures in Appendix G individual variance plots are shown. The Figures indicate that individual variance in greater that zero. Meanwhile residual means are very close to zero: $1,001192 * 10^{-16}$ and $9.69681 * 10^{-18}$ for return on assets and real interest margin models respectively.

4.2 Comparison of prediction accuracy

As was mentioned in section 2, the recent literature shows that some of tree-based methods applied to panel (longitudinal) data can provide better predictions, comparing to some of traditional models. This subsection concentrates on comparison of prediction abilities of estimated panel ARDL model and few of tree-based methods. In this subsection, prediction accuracy of the previously fitted ARDL model with PMG estimator and tree-based methods is compared by calculating mean squared error of predictions. Credit union data were split in two samples: training and test. Training sample covers all credit union data from 2009 Q3 to 2018 Q2 time period for tree-based methods and for ARDL model, since for panel ARDL model, data starts from 2009Q3 because of the time lag. The test sample covers data from the last four quarters: 2018Q3 - 2019 Q2 period. In order to check robustness of predictions, a comparison of prediction accuracies are also made for smaller test data samples: 2018Q4-2019Q2, 2019Q1-2019Q2, 2019Q1 periods. Therefore, training data sample is respectively: 2009Q2-2018Q3, 2009Q2-2018Q4, 2009Q2-2019Q2. The fitted panel ARDL models with PMG estimator and tree-based methods are fit on the training data sample and are predicted on test data sample for both return on asset and real interest margin.

First, the tree-based methods have been applied to credit union panel data. Since these methods are not restricted by the number of parameters, all of the available data set, mentioned in subsection 3.1., had been used. In contribution to some of the mentioned papers (Sela, Simonof (2009,2012)), regression trees (RT) have been applied directly to credit union data, ignoring panel data structure. Another tree-based methods: random forests (RF) and boosted trees (BT) have been constructed by (Sexton, 2018) on panel (longitudinal) data. Specifically, these RF and BT algorithms, to our knowledge, have not been used in a literature. But as was found from some similar researches using tree methods on panel (longitudinal) data, might possibly provide better predictions. Variable importance of the fitted trees for four period ahead predictions are shown in Tables 13-15 in Appendix D. As it is seen, tree-based methods show that loan quality and efficiency indicators have most predictive power for return on assets.

In cases, where real interest margin is dependent variable, net interest income were the most important variable for predictions. Moreover, it is seen that historical values of real interest margin were also important and considered in a tree. It is seen that some of variables shown in Appendix D were included in panel ARDL models, however, some of them were not statistically significant and were not included in traditional statistical panel data model. All of tree-based methods were calculated by searching minimal MSE value, using cycle. For RT method, various combinations of complexity paramaters, node sizes and minimum number of observations that must exist in a node were examined. For RF method on panel data, the size of the node and number of trees in the forest have been tested. Finally, for BT method, the size of the node and number of trees in the forest and different learning rates have been applied.

Table 9 shows MSE results from the predictions of credit union return on assets of panel ARDL PMG estimator (ARDL), regression trees (RT), random forests (RF) and boosted trees (BT). It is seen that at four and three period ahead predictions, random trees and boosted trees have smaller MSE value, comparing to panel ARDL model. At two and one periods ahead predictions only boosted trees were superior to PMG estimation. The results of another tree-based method, random forests, at all four periods ahead forecasts show possible over-prediction of true profitability values. The interpretation of this over-prediction comparing to other tree-based methods might be because RF method algorithm does not have penalty parameter for tree size. Therefore, the RF method might still have high variance, no matter that the RF method reduces variance by aggregation. It was shown in Segal (2003) that random forest methods might over-predict because of lack of penalty for tree size. The penalty parameter in RF algorithm is not indicated. Comparing to all tree-based methods, it is seen that the difference at two periods ahead predictions between panel ARDL and boosted trees, and four periods ahead predictions between panel ARDL and regression trees, boosted trees, are sufficiently small.

Table 9: MSE of credit unions return on assets

Model	One period ahead	Two periods ahead	Three periods ahead	Four periods ahead
ARDL	1.911975	1.649453	2.694303	2.112619
RT	5.197251	5.936287	2.110583	1.932974
RF	3.586677	8.230472	6.813847	6.272054
BT	0.354996	1.042596	0.926949	2.103311

Table 10 show similar results for real interest margin prediction. All periods ahead predictions indicate that boosted trees and regression trees predicted real interest margin with lowest MSE. Panel ARDL model predicted with largest MSE for two periods ahead test data, however, for one ahead period, the fitted model predicted real interest margin with sufficiently small error, comparing to the best predicted methods. As in return on assets case, it is seen that for most of prediction periods random forest method predicted with the largest error.

Table 10: MSE of real interest margin

Model	One period ahead	Two periods ahead	Three periods ahead	Four periods ahead
ARDL	0.0322441	0.3714221	0.1053182	0.1208771
RT	0.0451754	0.0226083	0.0506260	0.0551690
RF	0.0511975	0.1414972	0.2006050	0.2279593
BT	0.02124686	0.0236406	0.0605382	0.0576616

Therefore, forecasts of credit unions return on assets and real interest margin indicate that boosted trees method were superior to other methods and models, however, in some cases, MSE of statistical panel ARDL model were not sufficiently worse.

5 Conclusions

1. Panel ARDL model results show that logarithm of GDP, ratio of operating expenses to total income and banks' interest rates are important long-term determinants of credit union profitability.

2. Model also indicates that change in capital adequacy ratio and net interest income have positive significant impact on credit union profitability in the short-term. On contrary, change of non-performing loans and ratio of loan loss provisions to total loan portfolio have negative significant impact.

3. Comparisons of prediction accuracies of panel ARDL and tree-based methods indicated that some of the later methods provide more accurate predictions of credit union profitability to traditional statistical panel ARDL model. However, in some cases, the differences between prediction accuracies are not major.

   The highlights of this thesis might be used by credit union supervisory authorities, such as central bank and central credit unions, for the use of future supervisory decisions, such as choice for inspections, verification of business model forecasts, stress-testing calculations.

   Suggestion for further predictive researches is to consider and compare predictions of random forests method implemented for longitudinal data with different shrinkage parameters. Moreover, further applications using other machine learning methods implemented on financial panel data should be done in order to provide useful highlights to predictive analysis of financial institutions' performance.

6 References

1. Adler, W., Potapov, S. Lausen, B. (2011). Classification of repeated measurements data using tree-based ensemble methods. *Computational Statisticss*, Vol. 26.

2. Ali R., Butt Z. Z., Butt S. U. (2019). Do Non-Traditional Income, Size, and Growth Affect the Performance of the Banks? Evidence from the Big Three Countries of South Asia. *Journal of Management*, Vol. 2, No. 3, p. 58-66.

3. Anderson T.W., Hsiao C. Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association*, Vol. 76, No. 375, p. 598-606.

4. Anderson, T. W., and Hsiao C. (1982). Formulation and Estimation of Dynamic Models Using Panel Data, *Journal of Econometrics*, 18, 47-82.

5. Athanasoglou P. P., Brissimis S. N., Delis M. D. (2006). Bank-specific, industry-specific and macroeconomic determinants of bank profitability. *International Financial Markets, Institutions and Mone*y, Vol. 18, p. 121-136.

6. Albulescu C. T. (2015). Banks' Profitability and Financial Soundness Indicators: A Macro-Level Investigation in Emerging Countries. *2nd Global Conference on Business, Economics, Management and Tourism*, 30-31 October, 2014, Prague, Czech Republic.

7. Arellano M., Bond S. (1991). Some tests of specification for panel data:onte Carlo evidence and an application to employment equations. *Review of Economic Studies*, Vol 58, p. 277-297.

8. Baltagi B. H. (2001). Multiple Regression Analysis. *Econometrics*, p. 77-98.

9. Baltagi B. H. (2008). Forecasting with panel data. *Journal of forecasting*, Vol. 27, p. 153-173.

10. Blackburne F. E., Frank M. W. (2007). Estimation of nonstationary heterogeneous panels. *Stata Journal*, No. 2, p. 197–208.

11. Bouzgarrou H., Jouida S., Louhichi W. (2018). Bank profitability during and before the financial crisis: domestics vs. foreign banks. *Research in International Business and Finance*, Vol. 44, p. 26-39.

12. Capraru B., Ihnatov I. (2014). Banks' Profitability in Selected Central and Eastern European. *Countries. 21st International Economic Conference*, 16-17 May 2014, Sibiu, Romania.

13. Campmas A. (2018). How do European banks portray the effect of policy interest rates and prudential behaviour on profitability? *Research in International Business and Finance*, p. 1-19 (?).

14. Cetin H. (2019). Inflation and Bank Profitability: G20 Countries Banks Panel Data Analysis. *Proceedings of the 2019 International Conference on Management Science and Industrial Engineering*.

15. Dandapani K., Karels G. V., Lawrence E. R. (2008). Internet banking services and credit union performance. *Managerial Finance*, Vol. 34 No. 6, p. 437-446.

16. Dietrich A., Wanzenried G. (2014). The determinants of commercial banking profitability in low-, middle-, and high-income countries. *The Quarterly Review of Economics and Finance*, Vol. 54, p. 337-354.

17. Dincer H., Hacioglu U., Emir S. (2014). Financial Determinants of Bank Profits: A Comparative Analysis of Turkish Banking Sector. *Globalization of Financial Institutions*, p. 109-123.

18. Djalilov K. Piesse J. Determinants of bank profitability in transition countries: What matters most? (2016). *Research in Internatioal Business and Finance*, Vol. 38, p. 69-82.

19. Driscoll J. C., Kraay A. C. (1997). Consistent covariance matrix estimation with spatially-dependent panel data. *Review of Economics and Statistics*, Vol. 80, p. 549-560.

20. Du K. (2018). The impact of multi-channel and multi-product strategies on firms' risk-return performance. *Decision Support Systems*, Vol. 109, . 27-38.

21. Erdal H., Karahanoglu I. (2016). Bagging ensemble models for bank profitability: An empirical research on Turkish development and investment banks. *Applied Soft Computing*, Vol. 49, p. 861-867.

22. Forgione A. F., Migliardo C. (2018). Forecasting distress in cooperative banks: The role of asset quality. *International Journal of Forecasting*, Vol. 34, p. 678-695.

23. Hadi A., Suryanto T., Hussain H. Yap E. (2018). Bank's performance and its determinants – Evidence from middle east, Indian sub-continent and African banks. *Polish Journal of Management Studies*, Vol. 17(1), p. 17-26.

24. Hansen L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, Vol. 50, No. 4, p. 1029-1054.

25. Haron S. (1997). Determinants of Islamic bank profitability: some evidence. *Jurnal Pengutusan*, Vol. 16, p. 33-46.

26. Haskamp U. (2017). Improving the Forecasts of European Regional Banks' Profitability with Machine Learning Algorithms. *Ruhraral Economics Papers*.

27. Hsiao C, Appelbe T., Dineen C. (1993). A general framework for panel data analysis—with an application to Canadian customer dialed long distance service. *Journal of Econometrics*, Vol. 59, p. 63–86.

28. Hsiao C., Luke C. M., Mountain D., Tsui, K. (1989) Modeling Ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics*, Vol. 19, p. 567–587.

29. Hsiao C. (2007). Panel data analysis - advantages and challenges. *Sociedad de Estadistica e Investigacion Operativa*, p. 1-22.

30. Hoyos R., Sarafidis V. (2006). Testing for cross-sectional dependence in panel-data models. *Stata Journal*, Vol. 6, p. 482-496.

31. Im K. S., Pesaran M. H., Shin Y. (2003). Testing for unit roots in heterogeneous panels, *Journal of Econometrics*, Vol. 115(1), p. 53–74.

32. Iturriaga F. J. L., Sanz I. P. (2014). Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Sustems with Applications*, Vol. 42, p/ 2857-2869.

33. Jilkova P., Stranska P. K. 2017. Multiple linear regression analyses of the performance and profitability of the Czech banking sector. *Working Papers*, Vol. 41, Institute of Economic Research.

34. Levin A., Lin C., Chu C.J. (2002). Unit root tests in panel data:asymptotic and finite-sample properties. *Journal of Econometrics*, 108, p. 1–24.

35. Lin S.W., Shiue Y. R., Chen S. C., Cheng H. M. (2009). Applying enhanced data mining approaches in predicting bank performance: A case on Taiwanese commercial banks. *Expert Systems with Applications*, Vol. 36, p. 11543-11551.

36. Loh W. Y., Zheng W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics* , Vol. 7, No. 1 (2013), p. 495-522.

37. Maddala G.S., Wu S. (1999) . A Comparative Study of Unit Root Tests with Panel Data and A New Simple Test. *Oxford Bulletin of Economics and*

Statistics, Vol. 61, p. 631-652.

38. Mazlina N., Bakar A. (2009). Applying multiple linear regression and neural network to predict bank performance. *International Business Research*, Vol. 2, No. 4, p.176-183.

39. Naruševičius L. (2018). Bank profitability and macroeconomy: evidence from Lithuania.*TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY*, Vol. 24(2), p. 383–405.

40. Nataraja N., Nagaraja R. C., Ganesh L. (2018).Financial Performance of Private Commercial Banks in India: Multiple Regression Analysis. *Academy of Accounting and Financial Studies Journal*, Vol. 22.

41. Nerlove M. (2002). *Essays in Panel Data Econometrics*, Cambridge university press, Cambridge.

42. Ngufor C., Houten H., Caffo B. S., Shah N. D., McCoy R. G. (2018). Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c. *Journal of Biomedical Informatics*, Vol. 89, p. 56-67.

43. Nickel S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, Vol. 49., p. 393-410.

44. Pande A., Li L., Rajewsvaran J., Ehrlinger J. Kogalur U., Blackstone E., Ishwaran H. (2017). Boosted Multivariate Trees for Longitudinal Data. *Machine Learning*, Vol. 106(2), p. 277-305.

45. Paula D. A. V., Artes R., Ayres F., Minardi A. M. A. F. (2018). Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. *Management Journal*, Vol. 54, No. 3, p. 321-336.

46. Pedroni, P. (1999). Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics and Statistics*,

Vol. 61, p. 653–670.

47. Pesaran, H., Smith, R. P. (1995). Estimating long-run relationships from dynamic heterogeneous panels, *Journal of Econometrics*, Vol. 68(1), p. 79–113.

48. Pesaran, M. H., Shin, Y., Smith R. P. (1999).Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association*, Vol. 94, No. 446, p. 621-634.

49. Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. *Cambridge Working Papers in Economics*, No. 0435.

50. Petria N., Capraru B., Ihnatov I. (2015). Determinants of banks' profitability: evidence from EU 27 banking systems. *Procedia Economics and Finance*, Vol. 20, p. 518-524.

51. Sanusi N. A. and Mohammed N. (2007). Profitability of an Islamic bank: panel evidence form Malaysia. Readings in Islamic Economics and Finance, p. 97-116.

52. Sela R. J., Siminoff J. S. (2009). Re-Em Trees: A New Data Mining Approach for Longitudinal Data. *NYU Working Paper No. 2451/28094*.

53. Sela R. J., Siminoff J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, Vol. 86, p. 169–207.

54. Segal M. R. (2003). Machine Learning Benchmarks and Random Forest Regression, Kluwer Academic Publisher.

55. Sexton J. (2018). htree: Historical Tree Ensembles for Longitudinal Data, R package. R package version 2.0.0.

56. Tan Y. (2016). The impacts of risk and competition on bank profitability in China. *Journal of International Financial Markets, Institutions and Money*, Vol. 40, p. 85-110.

57. Tanaka K., Kinkyo T., Hamori S. (2016). Random forests-based early warning system for bank failures. *Economics letters*, Vol. 148, p. 118-121.

58. Wintgens M. (2017). Predicting the probability level of companies regarding the five comparability factors. MSc. Business Analytics.

59. Zhang G. Hu M. Y., Patuwo E., Indro D. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis.*European Journal of Operational Research*, Vol. 116, p. 16-32.

Appendices

A Appendix: data set

Variable name	Description
Dependent variable:	
ROA	Return on assets.
RIM	Real interest margin.
Country macroeconomic variables:	
BVP	Gross domestic product
BINT	Average loan interest rates of Lithuanian banks
Infl	Inflation
EURIBOR 6	Average 6 months inter-bank interest rates of European banks
Credit union-specific financial variables:	
KPR	Capital adequacy ratio
PERSK	Sustainable credit union capital
LOANS	Logarithm of total loans
SAPL	Ratio of loan-loss provisions to total loans
NPL	Ratio of non-performing loans to total loans
VSI	Ratio of assets loss provisions to total expenditures
OIP	Ratio of operating expenses to total income
CONC5	Ratio of five largest credit unions' assets to total credit unions assets
BAIP	Ratio of gross administrative expenses to total assets
I	Total commitments
GPP	Net interest income
PIP	Ratio of interest expenses to total income
PIKI	Ratio of interest expenses to total expenditures
IPVPKTA	Ratio of interest expenses to average interest commitments
GPMVT	Ratio of net interest margin to average assets
PIKP	Ratio of income on services and commissions to total income
VVPTA	Ratio of government securities to total assets
LTOD	Ratio of loans to total deposits
LR	Liquidity ratio
Dummy variables:	
Crisis	Variable indicating financial crisis (1 in 2009 Q2, Q3, Q4, else - 0)
Region	1 if credit union is located is one of biggest cities, 0 if credit union is located in smaller cities and rural area

B Appendix: panel co-integration results

Table 11: Panel cointegration tests (return on assets)

Panel v-statistic	Panel rho-statistic	Panel PP-statistic	Panel ADF-statistic	Group rho-statistic	Group PP-statistic	Group ADF-statistic
5.438*	-16.028*	-15.126*	-11.125*	-15.849	-22.104*	-16.657*

* indicates what p-value is smaller than 0.05 significance level.

Table 12: Panel cointegration tests (real interest margin)

Panel v-statistic	Panel rho-statistic	Panel PP-statistic	Panel ADF-statistic	Group rho-statistic	Group PP-statistic	Group ADF-statistic
3.626*	-3.625*	-5.340*	-6.408*	-1.14	-4.912*	-7.128*

* indicates what p-value is smaller than 0.05 significance level.

C Appendix: BIC lag order results

Lag order selected for ROA model	Lag order selected for RPM model
ARDL(1,0,0) regression	ARDL(1,0,1) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,1,1) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,1) regression
ARDL(1,0,0) regression	ARDL(1,1,1) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,0,1) regression
ARDL(1,1,0) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,1) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,0,0) regression
ARDL(1,0,1) regression	ARDL(1,1,0) regression
ARDL(1,1,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,0,0) regression
ARDL(1,1,0) regression	ARDL(1,1,0) regression
ARDL(1,0,0) regression	ARDL(1,1,0) regression
ARDL(1,1,0) regression	ARDL(1,1,0) regression

D Appendix: Variable importance

Table 13: Largest variable importance: Regression trees*

Variable name	ROA	Variable name	RIM
SAPL	10741.037	GPP	2071.162
OIP	4739.376	DTA	4276.817
NPL	4057.914	LTOD	248.858
KPR	3580.517	BINT	150.157
LR	2370.839	LTTA	45.638

* Calculated as the sum of the decrease in error

Table 14: Largest variable importance: Random forest*

Variable name	ROA	Variable name	RIM
SAPL	9.0606	GPP	0.6951
PIP	7.7819	RIM	0.3814
OIP	7.7186	PIP	0.3141
PIKI	7.3670	LTOD	0.2886
SAPL	7.3184	BINT	0.2874

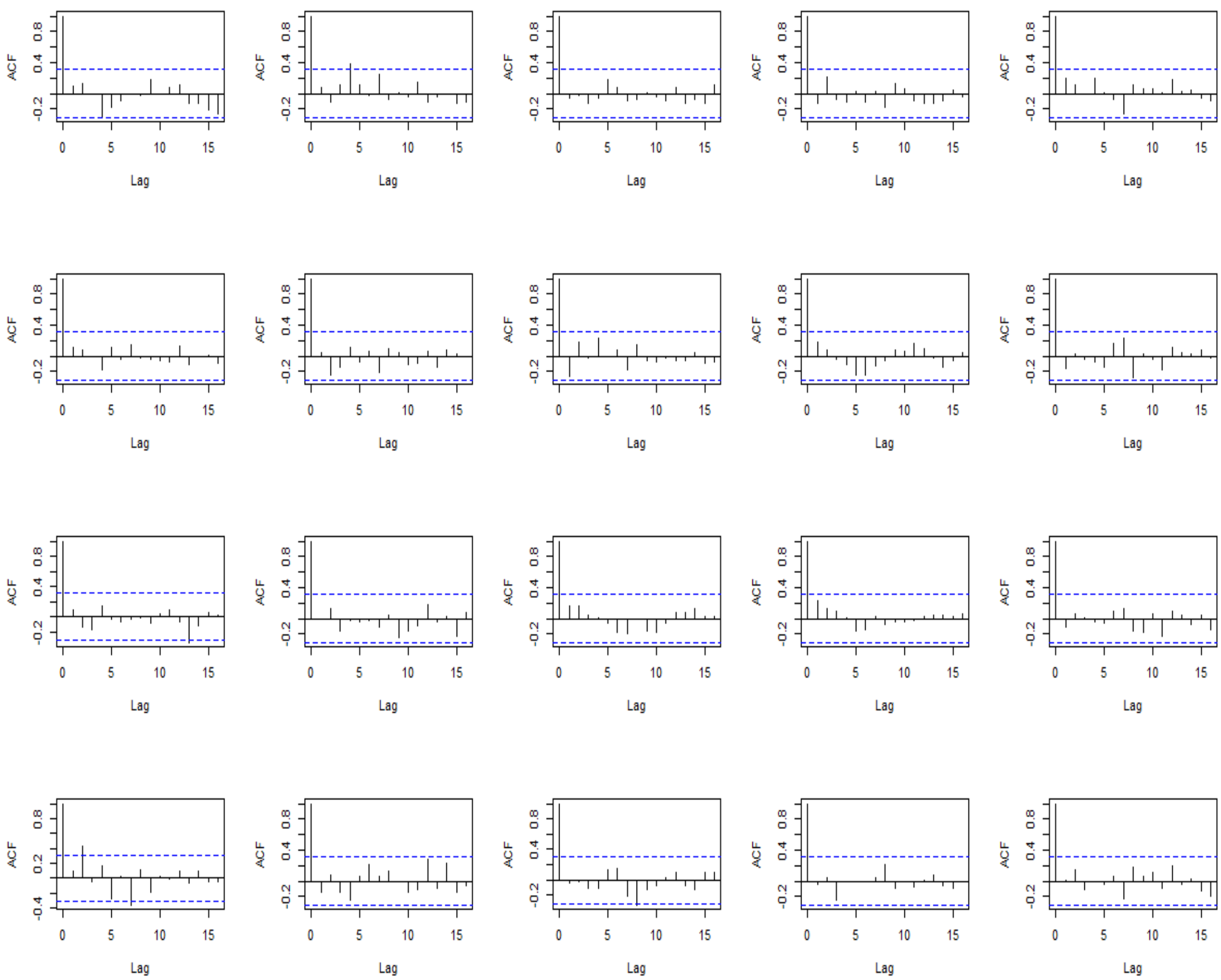
* Calculated as the marginalized error

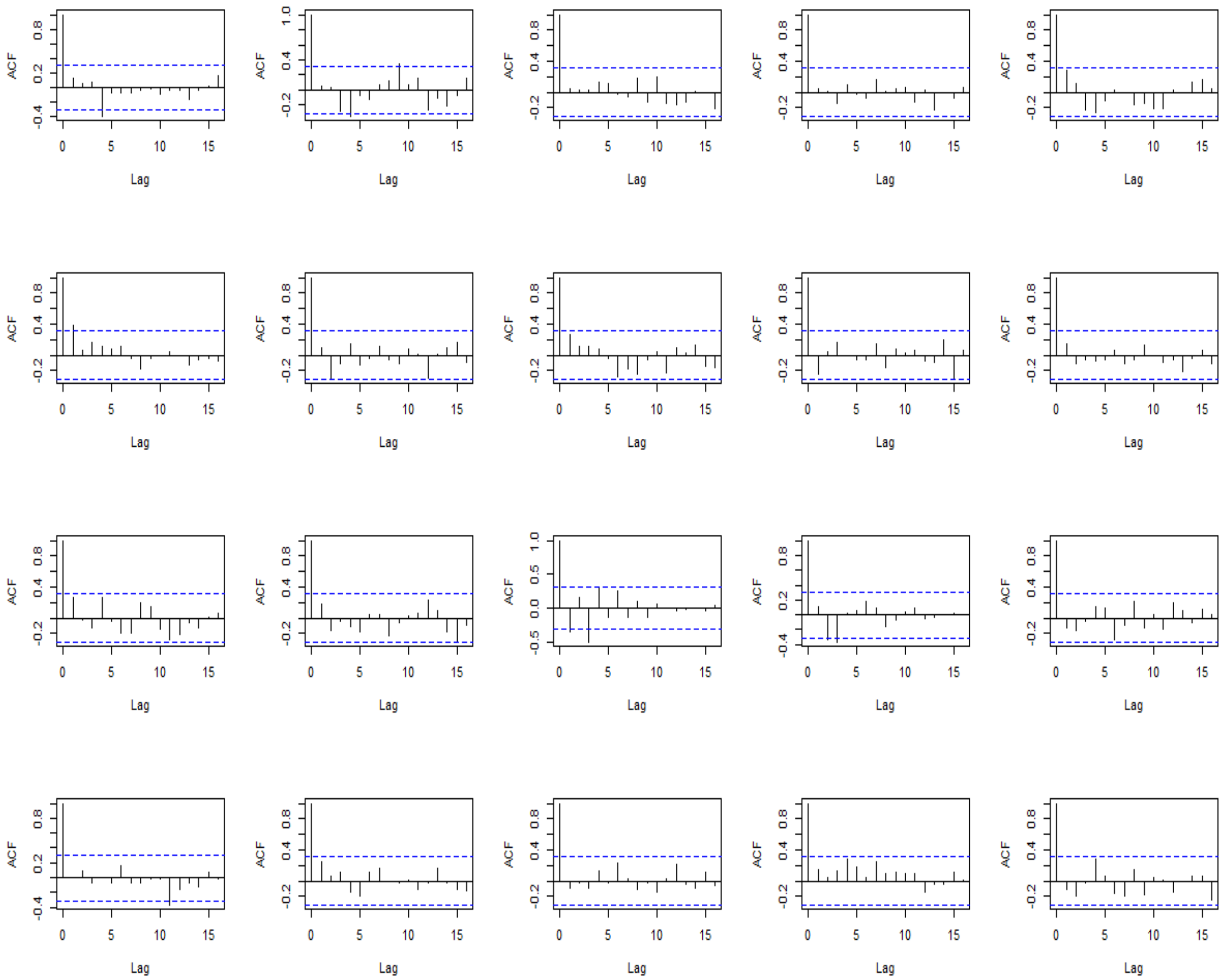
Table 15: Largest variable importance: Boosted trees*

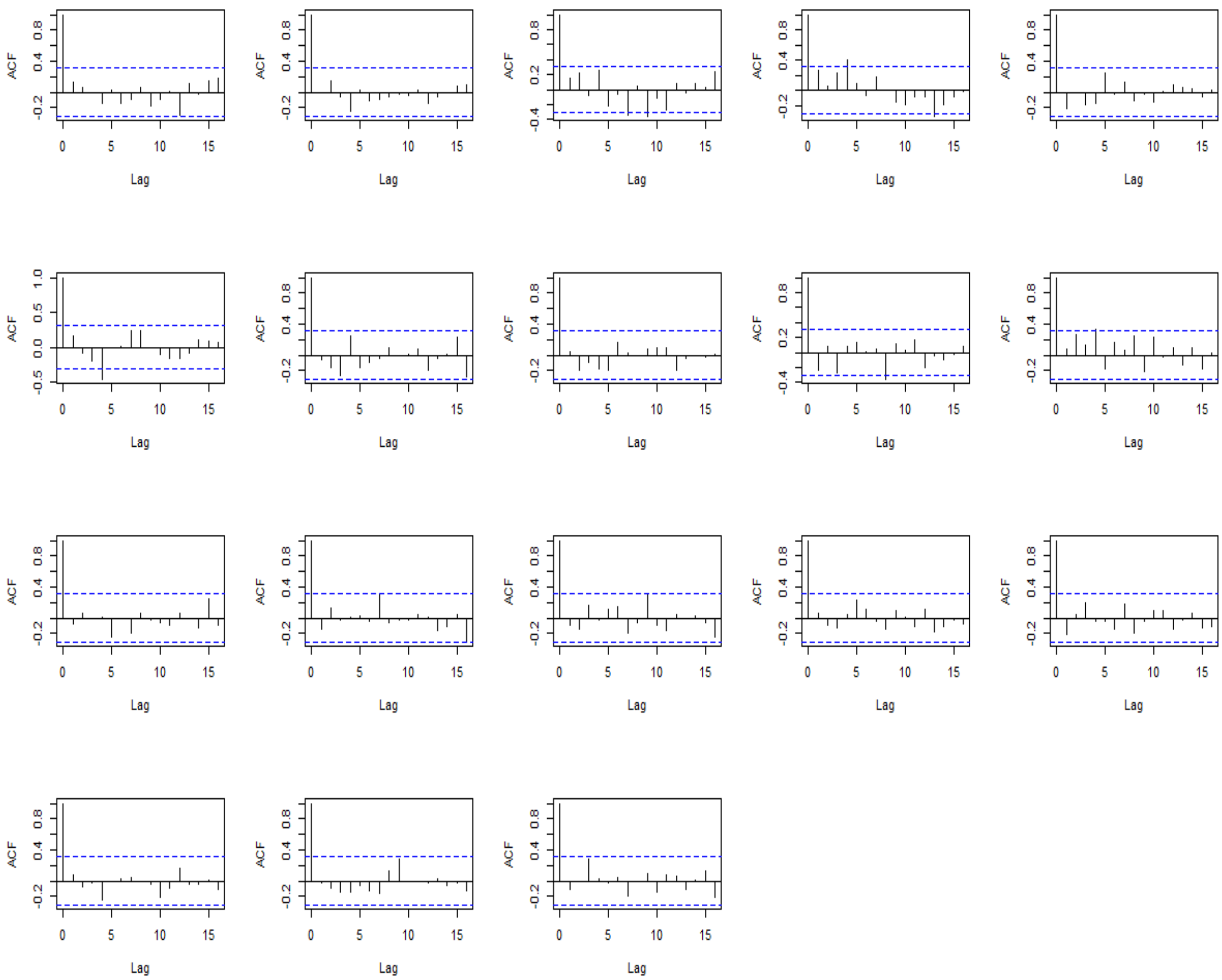
Variable name	ROA	Variable name	RIM
SAPL	12.6113	GPP	0.9100
PIP	8.1809	RIM	0.1861
OIP	7.9313	NPL	0.1741
VSI	7.3973	LTOTA	0.1706
PERSK	7.1211	PERSK	0.1673

* Calculated as the marginalized error

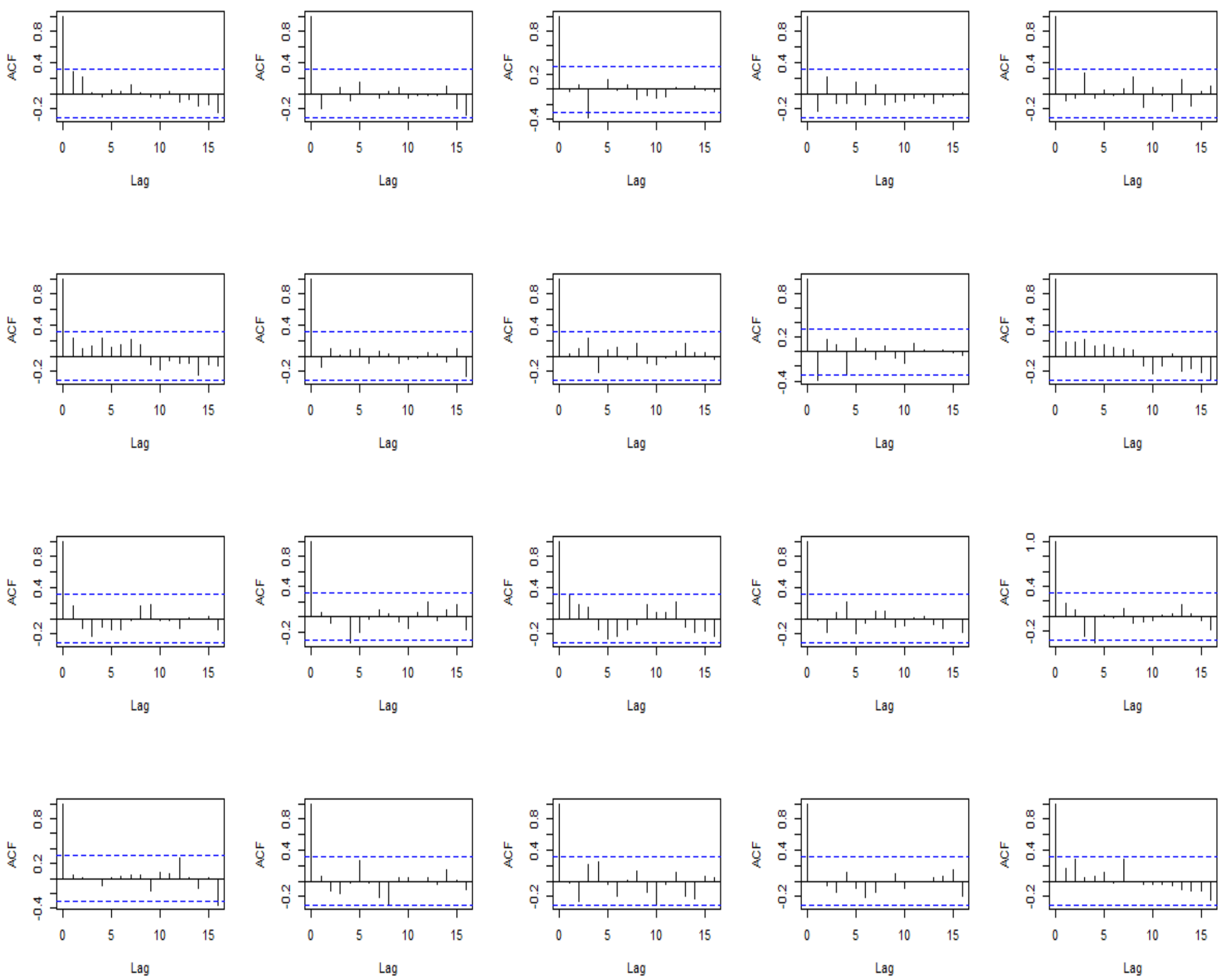
E Appendix: Residual ACF of ROA model

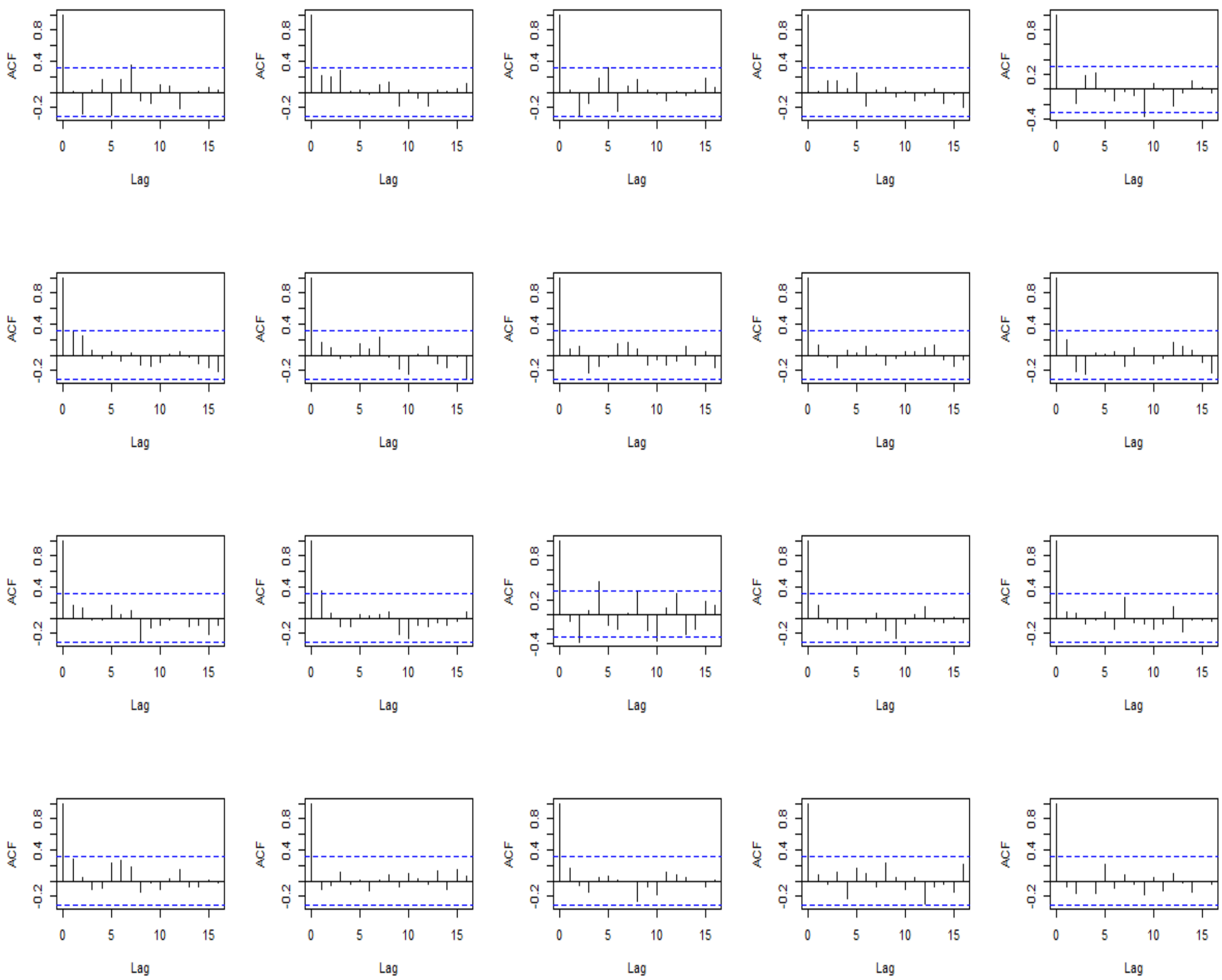


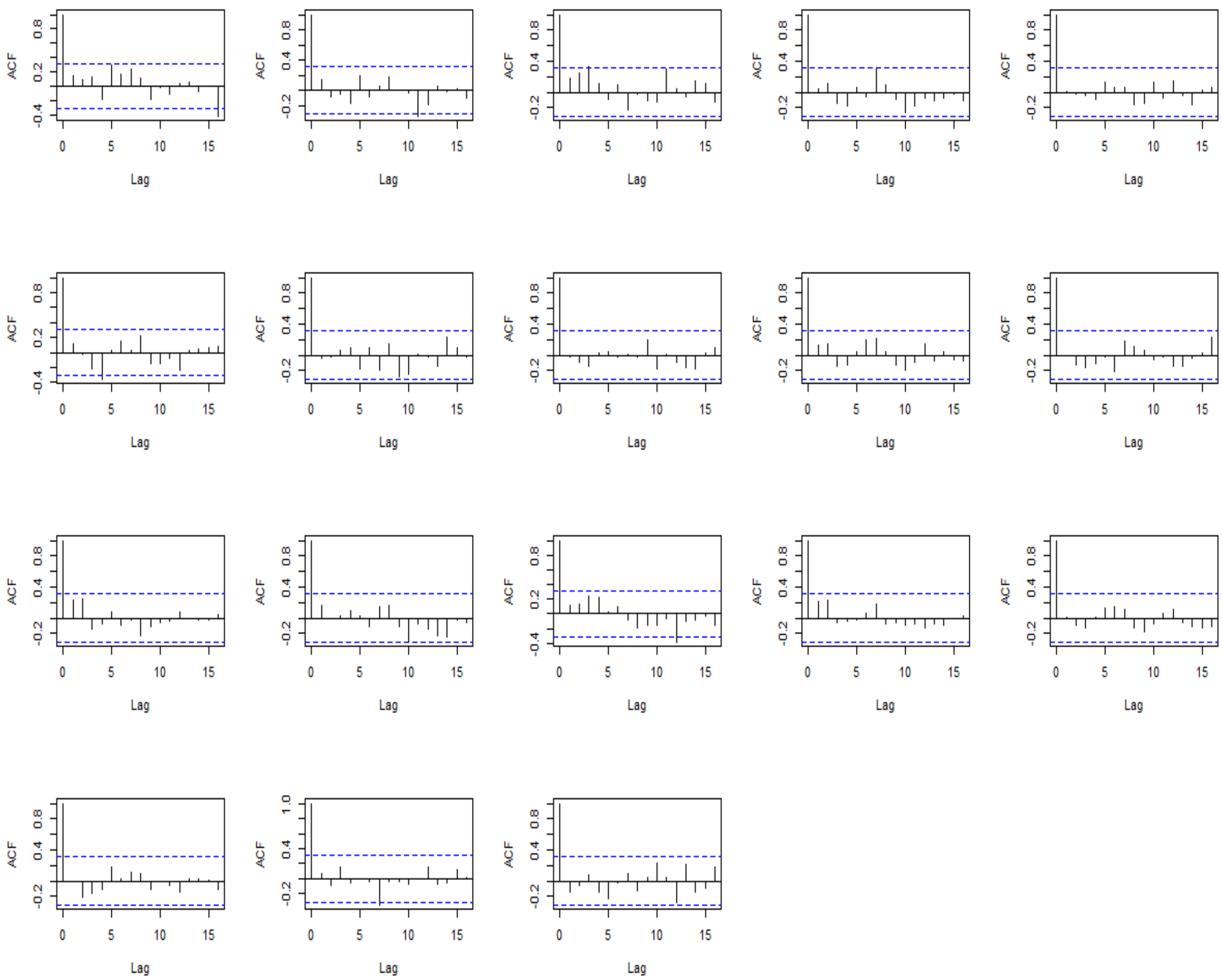

ACF
Lag
0
5
10
15


ACF
Lag
0 5 10 15

F Appendix: Residual ACF of RIM model




ACF
Lag


ACF
Lag

G Appendix: Residual variance