Charles University in Prague

Faculty of Social Sciences

Institute of Economic Studies

MASTER THESIS

What Drives the Aggregate Credit Risk: The Case of the Czech Republic

Author: **Bc. Jan Málek**

Supervisor: **PhDr. Jakub Seidler, Ph.D.**

Academic Year: **2012/2013**

Declaration of Authorship

1. Hereby I declare that I have compiled this master thesis independently, using only the listed literature and sources.
2. I declare that the thesis has not been used for obtaining another title.
3. I agree on making this thesis accessible for study and research purposes.

Prague, July 31, 2013

Signature

Acknowledgments

I would like to express my gratitude to my thesis advisor Jakub Seidler (Czech National Bank, Charles University in Prague) for supervising my work and for continuous encouragement. His useful comments made the origination of thesis much easier. I also thank to other people who provided me with any advice regarding different issues about this thesis: in particular to Chang-Jin Kim (University of Washington) for technical advice in the final chapter of this thesis and to Marek Soukal (Santa Monica College) for the help with final language correction. However, all remaining errors and omissions are my own.

A special word of thanks belongs of course to my family and friends for their patience throughout the process of writing.

Abstract

There has been a long discussion about macroeconomic variables influencing the level of aggregate credit risk in the economy. While literature provides both empirical evidence and theoretical explanation of the influence of the business cycle on credit risk, the effect of other macroeconomic variables has not been explored sufficiently. In addition, recent literature suggests the existence of a latent risk factor behind aggregate credit risk, which is regularly interpreted as the latent default cycle.

This thesis provides in its first part a discussion of potential aggregate credit risk drivers, which have been previously suggested in literature. We verify using a linear regression model whether the effect of these macroeconomic variables is also apparent in the Czech Republic. Results seem to be stable for both different model specifications and different clients segments and are in line with previous studies.

The second part of this thesis explicitly models the latent factor that is assumed behind aggregate credit risk by adding an unobserved component to the already existing model constructed earlier in this thesis. The unobserved component can be estimated by applying Kalman filter. We subsequently discuss the sources of the latent component and whether it can be interpreted as the default cycle.

The contribution of this diploma thesis is due to our belief twofold. First, we add a latent component to the linear regression model. Secondly, we analyze if and under which circumstances the latent component extension improves the fit of the regression model and discuss whether the explicit estimate of the unobserved component has a feasible interpretation as the default cycle.

Keywords credit risk, credit cycle, business cycle, segmented regression, GLS, Kalman filter

Author's e-mail janmalek@centrum.cz
Supervisor's e-mail seidler@email.cz

Abstrakt

Literatura poskytuje obsáhlou diskuzi na téma, jaké makroekonomické proměnné ovlivňují míru agregátního úvěrového rizika. Zatímco efekt hospodářského cyklu byl již teoreticky zdůvodněn a prokázán na dostupných datech, efekt ostatních makroekonomických proměnných prozatím nebyl analyzován dostatečně. Navíc, nedávné studie naznačují existenci skrytého faktoru, který míru agregátního úvěrového rizika také ovlivňuje a je často interpretován jako skrytý cyklus úvěrového selhání.

První část této práce diskutuje potenciální makroekonomické faktory ovlivňující agregátní úvěrové riziko, které byly již diskutovány v předešlých studiích na toto téma. Efekt těchto proměnných je ověřován na datech pro Českou republiku a výsledky se zdají být stabilní jak pro rozdílné úvěrové segmenty, tak pro rozdílné specifikace regresního modelu a jsou v souladu se závěry předchozích studií.

Druhá část práce modeluje explicitně skrytý faktor přidáním dodatečného komponentu do již existujícího regresního modelu odhadnutého v předešlé části této práce. Ukazujeme zde, že vývoj tohoto dodatečného komponentu během sledovaného období může být odhadnut pomocí Kalmanova filtru. Následně diskutujeme možné důvody existence skrytého faktoru a to, zda může být interpretován jako cyklus úvěrového selhání.

Přínos této diplomové práce je dle našeho přesvědčení dvojí. Zaprvé je jím přidání skrytého komponentu do lineárního regresního modelu. Zadruhé může být za přínos považována diskuze, jak a za jakých podmínek toto rozšíření prostého lineárního regresního modelu zlepší jeho fungování a zda může být explicitní odhad latentního komponentu interpretován jako cyklus úvěrového selhání.

Klíčová slova uverove riziko, uverovy cyklus, hospodarsky cycklus, segmentova regrese, zobecnena metoda nejmensich ctvercu, Kalman filtr

E-mail autora janmalek@centrum.cz
E-mail vedoucího práce seidler@email.cz

Contents

List of Tables

List of Figures

Acronyms

ACR Aggregate Credit Risk
ADF Augmented Dickey-Fuller test
AIC Akaike Information Criterion
AR Autoregressive Process
ARMA Autoregressive-Moving-Average Process
BIC Bayesian Information Criterion
BP Breusch-Pagan Test
CDO Collateralized Debt Obligations
CDS Credit Default Swaps
CNB Czech National Bank
CPI Consumer Price Index
EAD Exposure at Default
F,F-distribution Fisher-Snedecot Distribution
GDP Gross Domestic Product
GLS Generalized Least Squares Estimation
IID Independently and Identically Distributed
IP Industrial Production
JB Jarque-Bera Test
LB Ljung-Box Test
LGD Loss Given Default
MA Moving-Average Process
MLE Maximum Likelihood Estimation
NPL Non-performing Loans
NSA Not Seasonally Adjusted
OLS Ordinary Least Squares Estimation
PD Probability of Default
PPI Producer Price Index
PRIBOR Prague InterBank Offered Rate
SA Seasonally Adjusted
SSE Error Sum of Squares
SSR Residual Sum of Squares
SST Total Sum of Squares
t,t-distribution Student Distribution
VAR Vector Autoregressive Model
VECM Vector Error Correction Model
VIF Variance Inflation Factors

Chapter 1
Introduction

The immense increase in number of publications regarding credit risk on the aggregate level in the last decade declares an increasing interest in this topic. From a systemic point of view of central bankers or other regulatory authorities, the aggregate level of credit risk in the economy is the major concern. This dimension of credit risk (in contrast to ideosyncratic, client-specific credit risk) can't be diversified away and can therefore contribute to instability of banking sector. Numerous stress tests are being carried out by regulatory authorities around the world in order to assess resistence of commercial banks (and the banking sector as a whole) to unexpected events including large-scale realization of credit risk in their portfolios.

Although the nature of this type of risk suggests that it is mainly of concern of central banks or other institution being responsible for keeping financial stability, many commercial banks and other financial institutions become interested in a proper understanding of aggregate credit risk drivers because it appears to be essential for managing banking risks and conducting prudential stress testing and is therefore vital for sound financial planing.

One important reason for commercial banks to pay attention to the aggregate credit risk dynamics lies in the adoption of a portfolio perspective to credit risk. The standard approach makes distinction between idiosyncratic and systemic risk. Ideosyncratic credit risk is dominant on at client level when commercial banks assess credibility of a client and consequently make the decision on approving the loan. On contrary, the systemic credit risk is most important at a portfolio level because idiosyncratic risk can be largely diversified. Systematic credit risk factors are documented to be correlated with macroeconomic conditions (see Zeman and Jurca [2008], Kalirai and Scheicher [2002], Couderc et al. [2004] and many other papers) and also backed up by theoretical models on real business cycle theory (see Kiyotaki and Moore [1997], Bernanke, Gertler, and Gilchrist [1999]). Therefore, if we can establish a link between macroeconomic environment and aggregate default rates, the knowledge on the current state of macroeconomic variables can help to improve the ability to assess portfolio credit riskiness.

Another reason for the interest in aggregate credit risk dynamics is the emergence of new financial market products with ability to speculate in or hadge against credit risk. Well-illustrating examples of these products are asset-backed securities such as Collateralized Debt Obligations (CDO) with their value largely determined by the inherent systemic component. Appropriate pricing of these products requires an adequate understanding of their value drivers. For the purpose of identifying the key

value drivers, one can either use directly observed historical data on the variables themselves or use implied models based on prices of derived credit sensitive instruments with high liquidity such as Credit Default Swaps (CDS). The resulting flexibility in managing a portfolio through these derivatives complements the ability of commercial banks to differentiate between desired and undesired borrowers. In addition, it shifts the attention from cross-sectional assessment of borrowers' quality ti a dynamic, inter-temporal perspective.

Finally, commercial banks are obliged to fulfil regulatory requirements set both by the Basel Capital Accord and central banks. Majority of capital requirements stem directly from the creditworthiness of counterparties. The creditworthiness is assessed either by banks' internal rating models or by external ratings issued by rating agencies. There have been concerns that this scheme can lead to pro-cyclical capital requirements (see Repullo and Suarez [2012], Kashyap and Stein [2004]), and thereby amplify business cycle fluctuations. The argument is that during an upswing of the economy, commercial banks may lower their capital reserves. Such a decrease in reserves may be enabled by improving ratings of their counterparties based on estimates of default probabilities and resulting lower capital requirements. As a result, capital reserves may be too low at the peak of the cycle to cope with subsequent downswing. Moreover, the increase in capital reserves during a downswing can aggrevate already poor economic conditions. Regulatory authorities become increasingly aware of this issue. Recent version of Basel Capital Accord (i.e. Basel III) includes instruments to deal the pro-cyclicality of capital requirements and central banks have to competence to increase and decrease mandatory capital reserves throughout different stages of business cycle. However, the isuue of pro-cyclicality results in a need to assess how default rates and other credit risk drivers are affected by business cycle.

Hence, a deep understanding and identification of aggregate credit risk drivers is of high relevance for both central and commercial banks. Many researchers tried to estimate the drivers of aggregate credit risk by simultaneous equations models with multiple endogenous variables where shocks in one equation are propagated to other equations. The most commonly used techniques are vector autoregressive models (VAR, see Lucas and Koopman [2005]) or alternatively their modifications, such as vector error correction models (VECM). Nonetheless, such models require a clear-cut devision of explanatory variables between endogenous (i.e. determined by the system with shock in one variable being propagated to other variables) and exogenous (i.e. being used only as an input). Given the lack of a clear-cut theoretical framework explaining the driving factors of aggregate credit risk, a long list of macroeconomic variables is a priori available to serve as explanatory variables for aggregate credit risk.

This diploma thesis uses a linear regression model with 3-month aggregate default rates and 3-month changes in share of non-performing-loans (NPL) for different sectors of Czech economy as the explained variable whereas all explanatory variables are exogenous and we can choose a subset of variables that guarantees the best fit. We argue that data for 3-month aggregate default rates and 3-month changes in share of NPL are fully comparable as well as results that can be deduced from them. To choose the best model, we use the method of forward stepwise selection. We start building our model including one explanatory variable by adding industrial production[1] (IP) as a reasonable proxy for business cycle. There has been remarkable evidence that aggregate credit risk reacts asymmetrically to fluctuations in business cycle (see Marcucci and Quagliariello [2009], Morales and Gasha

[1] In fact, we use HP-filtered deviation from the potential value of industrial production.

[2004]) with positive fluctuations in business cycle having significantly lower impact than negative fluctuations. We test this hypothesis using a regression approach allowing for possible asymmetries

Moreover, there is mounting evidence of latent factors driving aggregate credit risk (see Boss, Fenz, Pann, Puhr, Schneider, and Ubl [2009], Kerbl and Sigmund [2011]). We incorporate this idea into the regression model by adding an unobservable component, which can be estimated using Kalman filter. It is argued that adding an unobservable component can improve the fit of the model with various degree of improvement for different sectors of the economy.

The rest of this thesis is structured as follows. Chapter 2 presents all data considered in this thesis: aggregate default rates and changes in total amount of non-performing loans as proxies to aggregate credit risk and various macroeconomic variables as potential explanatory variables. Chapter 3 introduces first the methodology of the linear regression model used in this thesis and suggest ways how to solve issues that arise when such a model is applied to time series data. Secondly, it describes rusults of the model with one explanatory variable, which tests for assymetry in the effect of business cycle to aggregate default rates using joinpoint regression approach. Thirdly, it presents results of the model with multiple explanatory variable and tries to determine default drivers for different segments of the economy. Chapter 4 argues that the results of the regression model can be improved by adding an unobservable component and tries to estimate the evolution of this componenent using Kalman filter.

Chapter 2
Discussion of Variables

This diploma thesis focuses on the aggregate credit risk in the Czech economy. Its aim is both to find aggregate credit risk drivers and to produce a model allowing us to estimate the expected development of this risk in response to evolution of key macroeconomic indicators.

First, we define a proxy variable for measurement of aggregate credit risk. Many studies have previously used the share of non-performing-loans (NPL) on total amount of commercial banks' loans (see Jakubik [2007], Jakubik and Schmieder [2008]). A preferable alternative is to use aggregate default rates that are an empirical counterpart of the probability of default (PD). PD is among loss given default (LGD) and exposure at default (EAD) the most commonly used parameter to quatify the credit risk on both individual and aggregate level. Under the condition of equal distribution of credit risk among clients of a bank, any forecasting model that predicts aggregate rates of default should also predict average clients PD with weights corresponding to their outstanding amount of credit . It is argued (see Kerbl and Sigmund [2011]) that using aggregate default rates is due to its nature of a flow measure preferable to a stock measure of NPL. Using a stock measure of credit risk such as NPL can lead to dubious conclusions due to unstable rates of write-offs and recoveries that might be dependent on cyclical position of the economy. In addition, some share of NPLs is being repaid on schedule despite the fact that they are denoted as non-performing in banking portfolios. However, Section 2.1 argues that data for aggregate default rates and for first differences of NPL match each other well. Thus the results that can be deduced from them are also fully comparable.

Second, there has been a long discussion about what macroeconomic variables can explain the aggregate credit risk. Due to the lack of a clear-cut theoretical framework explaining[1] the causes and evolution of aggregate credit risk, a long list of macroeconomic variables is available to be added into the model as explanatory variables. The selection among them would become even more challenging if we took into account their possible dynamic lag structure and/or their collinearity. Almost all previously published studies propose to add an arbitrary measure of business cycle such as GDP or industrial production (IP) into the model. Because it is strongly advisable to have both explained and explanatory variables in the model stationary, most studies use proportional deviations from HP-filtered potential values of GDP or industrial productions. On top of the business cycle measures, literature (see Couderc et al. [2004], Kalirai and Scheicher [2002]) includes examples of variables

[1] The theoretical effect of business cycle on aggregate credit risk is explained by many studies (see Kiyotaki and Moore [1997], Bernanke et al. [1999], Miao and Wang [2010]). In contrast, there is an apparent lack of theory explaining the influence of other variables.

that can be added into the model as an indicator of financial market situation, external situation, price stability and household situation. A discussion of variables that can be added into the model is provided in Section 2.2.

2.1. Variables to Explain

Our dataset comprises corporate default rates for the Czech Republic on monthly basis for a period 11/2002 - 3/2013. These time series is collected by the the Czech National Bank (CNB) and is presented regularly in the CNB Financial Stability Report. Due to strict reporting rules stemming from the Basel Accord, data for the flow of aggregate defaults are reliable with sufficient timeliness.

As the variable to explain, we employ 3-month ex-ante aggregate defaul rates as a share of the amount of non-defaulted loans at the beginning of the period. Formally,

$$DR_{1,t} = \frac{\sum_{i=t+1}^{t+3} Defaults_i}{loans_t} \tag{2.1}$$

where $DR_{1,t}$ is aggregate 3-month ex-ante default rate as a share of non-defaulted loans, $Defaults_i$ are amounts of of defaulted loans at respective months and $loans_t$ are amounts of non-defaulted loans at the beginning of 3-month period. Due to possible seasonality of aggregate default rates, we employ the TRAMO/SEATS procedure (see Gomez and Maravall [1998], Maravall [1996] and Section 2.3) in order to filter out the seasonality bias.

It is important to stress out the forward-looking property of 3-month ex-ante default rates. Any model that estimates $DR_{1,t}$ based on informations available at time t in fact forecast the development of default rates over next 3-months. Another desirable features of using 3-months aggregation are lower fluctuations resulting in possible higher significance of employed mathematical models and easier interpretability of 3-months aggregate default rates. A partial drawback is related to the forward-looking property. Because this statistics is computed by summing defaulted loans over the next 3 month, we have to wait until time $t+3$ until it can be computed, which results in a loss of the 3 most recent observation and the statistics is available only for 11/2002-12/2012.

Although we consider 3-month default rates to be the most reasonable proxy for the aggregate credit risk, there is a problem with their unsufficient specificity. For the purpose of elaborating on this thesis, aggregate corporate default rates were available only as one time series for the Czech economy as a whole. Because there is a need for more specific data allowing us to search for sector specific aggregate credit risk drivers, we enlarge our analysis by adding the statistics for non-performing loans (NPL). Czech National Bank (CNB) collects statistics of NPL for 17 different sectors of corporate clients. Resulting amount of NPL aggregates all loans classified as non-performing[2] for the particular sector of the Czech economy. As noted earlier in this thesis, the statistics of NPL has a nature of a stock measure with newly defaulted loans as an inflow into the common pool of NPL and recoveries[3]

[2] For the purpose of this thesis, we consider non-performing loans to be roughly identical to defaulted loans.

[3] Recoveries include both loans that are denoted as non-performing at the beginning of the period but become performing again and also loans that are being repaid on schedule despite being classified as non-performing..

and write-offs[4] as an outflow from the common pool of NPL. Our goal is to find a modification of the statistics for NPL that is consistent with 3-month default rates from Equation 2.1. Thus, we define 3-month ex-ante changes in NPL as a share of performing loans at the beginning of this period and denote them as $DR_{2,t}$[5][6]. Formally,

$$DR_{2,t} = \frac{NPL_{t+3} - NPL_t}{loans_t} \tag{2.2}$$

where NPL_{t+3}, NPL_t are amounts of non-performing loans (NPL) at the end of the 3-month period, and at the beginning of the period respectively. $loans_t$ is the amount of performing loans at the beginning of the period. It is straightforward to show the relation to $DR_{1,t}$ from Equation 2.1. As the changes in NPL in each period are defined as the difference between an inflow of newly defaulted loans and an outflow of recoveries and write-offs, we can decompose the amount of the change in NPL over 3-month interval into sum of inflows and sum of outflows respectively. Formally,

$$\begin{aligned} DR_{2,t} &= \frac{\sum_{i=t+1}^{t+3}(NPL_i - NPL_{i-1})}{loans_t} = \frac{\sum_{i=t+1}^{t+3}(Defaults_i - Recoveries_i - Writeoffs_i)}{loans_t} = \\ &= \frac{\sum_{i=t+1}^{t+3} Defaults_i}{loans_t} - \frac{\sum_{i=t+1}^{t+3}(Recoveries_i + Writeoffs_i)}{loans_t} = DR_{1,t} + \varepsilon_t \end{aligned}$$

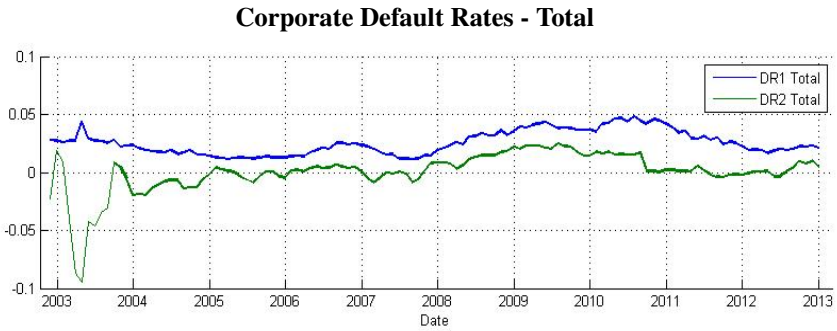
where $Defaults_i$, $Recoveries_i$, $Writeoffs_i$ are sources of inflow and outflow into NPL and ε_t is the error component that is not feasible to filter-out from the data. It has been documented (see Khieu et al. [2012]) that the rate of recoveries can significantly fluctuate over time. Moreover, the rate of write-offs is given by an arbitrary choice of commercial banks to remove bad loans from their portfolios. The argument is that, if there is a period with changes of NPL ($DR_{2,t}$) closely following the aggregate default rates ($DR_{1,t}$), it is possible to consistently compare the result from performing an analysis on aggregate default rates and changes in NPL or alternatively to state that an inference drawn from an anylysis of changes in NPL is also valid for aggregate default rates.

Figure 2.1 compares 3-month ex-ante aggregate default rates (DR_1) and 3-month ex-ante changes in non-performing Loans (DR_2). Data are annualized (i.e. multiplied by 4) in order to maintain comparability with 12-month rates. For the feasibility of comparison, both statistics are aggregated for corporate, non-financial clients in the Czech Republic. The enclosed table contains descriptive statistics of 3-month ex-ante aggregate default rates (DR_1) and 3-month ex-ante changes in non-performing loans (DR_2) for corporate, non-financial clients in the Czech Republic. The values for 3-month changes in non-performing Loans (DR_2) are plotted in bold for a period of 9/2003-12/2012 because their values very closely follow the aggregate default rates and an analysis

[4] Write-offs include on one hand accumulated non-performing loans that are so unlikely to be repaid that banks decide to remove them out of their portfolios.

[5] Although $DR_{2,t}$ doesn't represent the aggregate default rates directly, it is is very closely related to it. As shown later in this thesis in Figure 2.1, $DR_{1,t}$ and $DR_{2,t}$ are linearly shifted (i.e differ only by a constant) for most of the analyzed period and therefore include the same information about how individual explanatory variables influence aggregate credit risk. We use this notation in order to emphasize this relation.

[6] The resulting statistics of 3-month ex-ante changes in NPL are available for the time period of 11/2002-12/2012, which constitutes a loss of the 3 most recent observations as a result of the forward-looking property.



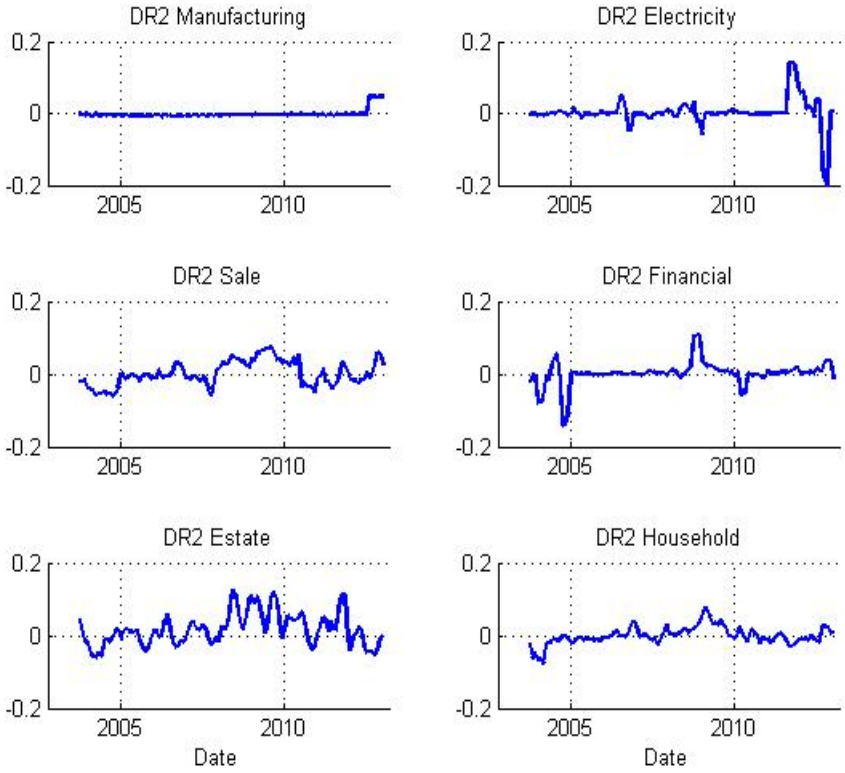
	Start	End	Min	Max	Mean	Median	St.dev
$DR_{1,Total}$	9/2003	12/2012	0.0114	0.0481	0.0253	0.0226	0.0106
$DR_{2,Total}$	9/2003	12/2012	-0.0199	0.0248	0.0038	0.0017	0.0102

Figure 2.1.: **3-Month Ex-ante Aggregate Defaul Rates (DR_1) and 3-Month Ex-ante Changes in Non-performing Loans (DR_2) for Corporate Clients in the Czech Republic**
Source: Czech National Bank + Author's Calculation

of aggregate credit risk via changes in NPL is feasible. Both time series are seasonally adjusted by TRAMO/SEATS procedure, the comparison of seasonally adjusted (SA) and not seasonally adjusted (NSA) can be found in Section A.1 of the Appendix. The degree to which seasonal adjustment alters values of our data differs from segment to segment. The most notable alteration is carried out in case of manufacturing ($DR_{2,Man}$). Although the original values of $DR_{2,Man}$ exhibit substantial deviations, most of these deviations dissapear with the seasonal adjustment.

Because a comovement of DR_1 and DR_2 was found only for the period 9/2003 - 12/2012, we will limit our analysis of sector specific 3-month ex-ante changes in non-performing loans ($DR_{2,Sector}$) to this period. Figure 2.2 shows the evolution of this variable for 6 selected segments of corporate clients. In addition, descriptive statistics of DR_2 for 6 sectors of the Czech Economy with biggest exposure on commercial banking portfolios can be found in Table A.2 in the Appendix. All time series are seasonally adjusted by TRAMO/SEATS procedure, the comparison of seasonally adjusted (SA) and not seasonally adjusted (NSA) can be found in Section A.1 of the Appendix.

3-Month Changes in NPL - Sector Specific



Abbreviation	Name of Sector
$DR_{2,Man}$	MANUFACTURING
$DR_{2,El}$	ELECTRICITY, GAS, STEAM, WATER SUPPLY
$DR_{2,Sal}$	WHOLESALE AND RETAIL TRADE
$DR_{2,Fin}$	FINANCIAL AND INSURANCE ACTIVITIES
$DR_{2,Est}$	REAL ESTATE ACTIVITIES
$DR_{2,HH}$	ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS

Figure 2.2.: **3-Month Ex-ante Changes in Non-performing Loans (DR_2) for Specific Sectors of Corporate Clients in the Czech Republic**
Source: Czech National Bank + Author's Calculation

2.2. Explanatory Variables

In this section we present potential determinants of aggregate credit risk that will be used to construct our regression models.

Table 2.1 at the end of this section summarizes all proposed explanatory variables clustered into thematic groups. Its last column includes all modification of actual values of a particular variable: this covers the type of stationarization (HP-filter or log-differencing) and whether the variable is seasonally adjusted or not. Finally, all explanatory variables are normalised so that they have the sample mean of 0 and sample standard deviation of 1. All modifications of explanatory variables are described in Section 2.3.

Table 2.1 also summarized expected signs of the regression coefficients. However, these expectations don't take into account possible instability of regression coefficients due to existing multicollinearity among explanatory variables. Also the logic behind is very simplified and does not have to reflect the reality.

In many cases, it is not possible to infer about the direction of causality between variables. The problem of linear regression models is that the sign of regression coefficients only provides evidence about correlation between explanatory variables and the variable that was arbitrarily chosen as the variable to explain whereas there can also be an opposite source of causality. This is addressed in literature as the endogeneity problem (see Greene [2012]).

2.2.1. Business Cycle Indicators:

Almost all previously published studies propose to add an arbitrary measure of business cycle such as GDP or industrial production (IP) into the model. As argued in Subsection 2.3.2, it is strongly advisable to have both explained and explanatory variables in the model stationary, most studies use proportional deviations from HP-filtered potential values of GDP or industrial productions. Some of the studies (see Marcucci and Quagliariello [2009]) also suggest to add a modification of the business cycle measure such as in the form of a censored variable allowing for an assymetric effect of business cycle and/or a lagged value of proportional deviation from HP-filtered potential values of GDP (or alternatively industrial production). The argument is that the effect of business cycle to aggregate credit risk is so prominent and possibly non-linear that it doesn't have to be fully captured in only one variable.

The assumption is that when the economy is currently in a phase of economic recovery, the business activities are expanding and general well-being is increasing as well. Positive business cycle conditions thus reduces the volume of aggregate credit risk. Therefore a negative sign between the measure of aggregate credit risk DR_t and business cycle conditions can be expected (see Couderc et al. [2004]).

An important issue for building up our analysis is the adequate choice of the business cycle indicator. Regardless of the indicator variable (either GDP or IP), we focus on the percentage gap grom the Hodrick-Prescott filter trend value of the variable (see Hodrick and Prescott [1997]). An obvious wirst choice would be to use real, seasonally adjusted GDP. The problem is that the data for GDP in the Czech Republic are only available with quarterly frequency. Although it is possible to interpolate the values of GDP to monthly frequency, such a procedure will not be capable of reflecting rush changes

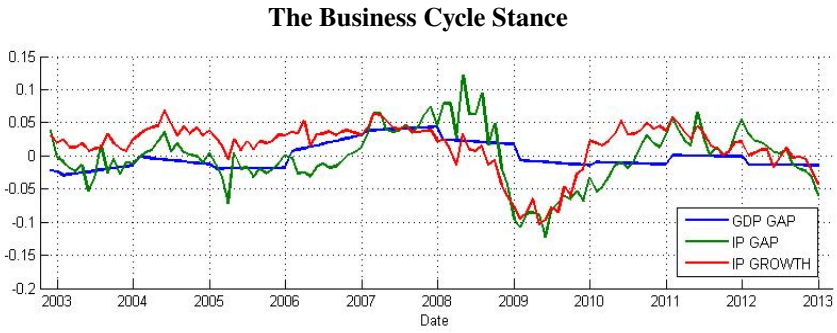


Figure 2.3.: **Different Measures for the Business Cycle Stance**
Source: Czech Statistical Office (CSU) and Organization for Economic Cooperation and Development (OECD) + Author's Calculation

in the business cycle. For this reason, we suggest to utilize data for seasonally adjusted industrial production (IP). Moreover, as noted in Kalirai and Scheicher [2002], the development of IP usually follows the development in GDP.

Figure 2.3 depicts the evolution of the business cycle measured as proportional deviation from its trend values of GDP and IP. Both measures are consistent, with a correlation of 0,91. Using IP apparently overestimates the amplitude of business cycle compared to GDP. Nevertheless, we suggest to use IP due to its better timeliness at the expense of loosing some share of accuracy. The main reason is that the time series for IP better captures fast economics deterioration at the end of year 2008.

Marcucci and Quagliariello [2009] suggest to use annual GDP/IP growth rates. They argue this measure of business cycle is timely shifted ahead by sevelal months compared to proportional deviation from HP-filtered trend value. The point is that for HP-filtered values to jump above zero (the average by definition), the annual growth rate must have been above average for several preceding months. Other advantage of using growth rates compared to HP-filtered deviations from the potential value of GDP/IP lies in the impossibility to have accurate results for HP-filter at time t available immediately at time t . It usually takes several subsequent periods to allow the Hodrick-Prescott filter to accurately calculate the potential value at time t.

For the reasons explained above, we decided to choose the following representative variables from the group of business cycle indicators:

1. Proportional Deviation of Industrial Production from Its HP-filtered Value (IP_GAP)
2. Annual Industrial Production Growth Rate (IP_GROWTH)

2.2.2. Price Stability Indicators:

As representative variables from the group of price stability indicators, we have chosen the following variables:

1. Consumer Price Index (CPI)
2. Producer Price Index (PPI)

3. Monetary Aggreagate M1 (M1)

As mentioned in Zeman and Jurca [2008], a discrepancy between inflation expectations and realized inflation can cause serious problems to both firms and individuals. An increase in inflation is associated with the reduction of the real value of debt. In short term, it also decreases real interest rate and supports[7] a higher economic activity. This would arque that and increase in inflation should decrease the aggregate credit risk in terms of default rates and negative relationship between DR_t and inflation should be expected.

The logic Zeman and Jurca [2008] follows is bit controversial. It might hold for a significantly indebted household/firm that has a fixed stream of capital outflow (in nominal terms) due to obligations to repay its debt and on the other hand variable money inflow, which is positively influenced by inflation. In contrast, the credibility of a household/firm with variable money outflow that increases due to inflation and fixed money inflow (in nominal terms) could be influenced adversely. The impact of increased inflation on firms can be similarly ambiguous. The change of individual credit risk depends on individual exposition on risk resulting from the change of inflation.

We should stress out that this thesis points at credit risk among corporate clients who often provide services to consumers and thus CPI rougly determines their income whereas PPI roughly determines their expenses. For this reason, we expect negative correlation for CPI and positive for PPI. The monetary aggregate M1 will be probably negatively correlated with aggregate credit risk due to the fact that the more money are spread in the economy the better the economy should prosper in short term.

A slightly different type of intuition behing the effect of price stability on aggregate credit risk is presented by Couderc et al. [2004]. Their article arques that the correlation between inflation and aggregate credit risk should be positive because high inflation is usually associated with growth and growth (i.e busines cycle conditions) is documented (see Couderc et al. [2004] or many other studies) to be negatively correlated with aggregate credit risk. However, this example shows that correlation does not have to be related to direct causality.

2.2.3. External indicator

As representative variables from the group of external indicators, we have chosen the following variables:

1. Nominal Exchange Rate vis-a-vis Euro in Direct Quotation (ER)
2. Current Account Balance (CA)

Kalirai and Scheicher [2002] argue that nominal exchange rate could have an impact on the financial sector through foreign trade linkages depending on whether a firm is an net importer or a net exporter. Depreciation of the currency (i.e. an increase of nominal exchange rate (ER)) should improve credibility of net exporters and impair credibility of importers. Appreciation of the currency should have the opposite effect. For the purpose of this thesis, we assume that the effect of currency change on net

[7] The relationship between inflation and economic activity is a controversial topic. This line of statement follows the logic that an unexpected increase in inflation provides an incentive for higher economic activity until inflation expectation is increased and the incentive ceases to exist.

exporters prevails, which would result in negative correlation of nominal exchange rate and aggregate credit risk.

On the other hand, the debt burden of some firms denominated in foreign currency increases in case of home currency depreciation. This fact contributes to positive correlation between nominal exhange[8] rate and aggregate credit risk. For this reason, we don't provide a clear idea about expected sign of correlation between nominal exchange rate and the measure of aggregate credit risk DR_t.

In contrast, the correlation between current account balance (CA) and the measure of aggregate credit risk DR_t should be negative. An improvement of current account results in additional inflow of liquidity into the economy, which should decrease the aggregate credit risk.

2.2.4. Credit Market Indicators:

As representative variables from the group of household indicators, we have chosen the following variables:

1. 3-Month Prague InterBank Offered Rate (PRIBOR3M)
2. Total Amount of Loans on Commercial Banks' Portfolios (CREDIT)

Regarding the correlation of 3-Month Prague Interbank Offered Rate (PRIBOR3M) and DR_t, we first assume that the level of interest rate relates to direct costs of credit. An increase in interest rate causes that the obligors' debt burden rises as well and obligated parties are more likely to default. This would provide a justification of positive correlation between PRIBOR3M and the measure of aggregate credit risk. However, interest rates tend to be lower in contraction periods and higher in expansions (see Friedman [1986]). Therefore, we don't provide an idea about expected sign of correlation between Pribor 3M and DR_t.

Total amount of loans on commersial banks' portfolios (CREDIT) is expected to be negatively correlated with DR_t. Increased amount of loan provision is a sign of good short-term economic prospects. On the other hand, increased lending to clients of dubious quality can result in an impairment of overal economic outlook. However, we forecast the realization of aggregate credit risk only over the period of next 3 month and the latter effect will probably have a negligible influence.

2.2.5. Financial Market Indicators:

As representative variables from the group of financial market indicators, we have chosen the following variables:

1. Return on PX50 over Last 200 Trading Days (RET_PX50)
2. Volatility of PX50 over Last 30 trading Days (VOL_PX50)
3. 3 Year Treasury Yield (GOV3)
4. Slope of Term Structure (GOV10-GOV3) defined as 10 year treasury yield minus 3 year yield

[8] Because we use direct quotation for nominal exchange rate (ER), an increase of ER is understood as depreciation of Czech currency against Euro. A decrease of ER is understood as an appreciation of Czech currency against Euro.

Return on PX50 (RET_PX50) is expected to be negatively correlated with DR_t. Short and mid-term economic performance should be positively correlated with past returns on stock market indices and therefore we expect a negative impact on the measure of aggregate credit risk. Furthermore, an increase in equity prices tends to decrease firm leverage and therefore also push down the aggregate credit risk. We use the realized return over the last 200 trading days.

Volatility of PX50 (VOL_PX50) returns is expected to be positively correlated with DR_t. In a traditional Merton model (see Merton [1974]), there are two drivers of default probability: leverage and the volatility of firms' assets. The volatility of equity returns is often used as a proxy for the latter and we expect it to have a positive impact on the realization of aggregate credit risk. We use the realized volatility over the last 30 trading days.

3 year treasury yield (GOV3) is determined by the cost of capital plus a risk premium due to the possibility that the governmental debt will not be repaid in full. Higher treasury yield may be caused by increased cost of capital, which negatively influence the macroeconomic environment and impair the aggregate credit risk. This would provide justification for an effect of increased treasury yield on increased aggregate credit risk. However interest rates tend to be lower in contraction periods and higher in expansions. On top of that, risk premium as a component of 3 year treasury yield may be in case of developed, highly credible countries (as the Czech Republic might be) can fall in case of an economic downturn due to decreased risk appetite of investors (see Gonzalez-Hermosillo [2008]). Thus, we don't provide an idea about expected sign of correlation between GOV3 and DR_t.

Slope of term structure (GOV10-GOV3) is expected to be negatively correlated with DR_t. As mentioned in Couderc et al. [2004], steep term structures of interest rates are usually associated with strong growth prospects and we expect this variable to indicate sound mid-term economic conditions consistent with low aggregate credit risk.

2.2.6. Household Indicators:

As representative variables from the group of household indicators, we have chosen the following variables:

1. Number of Unemployed Workers (UNEMP)
2. Number of Vacant Working Positions (VACANT)

Number of unemployed workers (UNEMP) is expected to be positively correlated with DR_t. It is rational to expect that people get into great difficulties to repay loans if they lose the job. Therefore a higher unemployment is consistent with higher aggregate credit risk. On firm level, lower number of unemployed workers signalises good economic prospects and firms' intention to expand their economic activity.

Number of vacant working positions (VACANT) is expected to be negatively correlated with DR_t. It is a sign of short/mid-term economic recovery. Increased number of vacant working positions is also a guarantee that employees being released from work find a new job more easily and don't fail to fulfil their financial obligations. As in case of UNEMP, higher number of vacant working positions indicates firm's intention to expand their economic activity.

Table 2.1.: **List of All Candidate Explanatory Variables with Expected Signs of Their Effect on Aggregate Credit Risk**

Group	Variable	Expected Sign of Correlation	Remarks
Business	IP_GAP	-	Norm, HP
	IP_GROWTH	-	Norm, 12M-log-diff
Household	UNEMP	+	Norm, SA
	VACANT	-	Norm
Financial	RET_PX50	-	Norm
	VOL_PX50	+	Norm
	GOV3	+/-	Norm
	GOV10-GOV3	-	Norm
Credit	PRIBOR3M	+/-	Norm
	CREDIT	-	Norm, 12M-log-diff
External	ER	+/-	Norm, 12M-log-diff, SA
	CA	-	Norm, SA
Price	CPI	-	Norm, SA
	PPI	+	Norm, SA
	M1	-	Norm, 12M-log-diff, SA

SA : seasonally adjusted (if not stated explicitly, variables are seasonally unadjusted)

12M-log-diff : variable log-differenciated subject to its values lagged by 12 months (see Subsection 2.3.2)

HP : variable detrended by computing proportional deviations from HP-filtered trend value (see Subsection 2.3.2)

Norm : variable normalized so that it has a sample mean of 0 and sample variance of 1

2.3. Modification of Variables

2.3.1. Seasonal Adjustment of Variables Being Explained and Selected Explanatory Variables

Some time series might exhibit a seasonal pattern. It is vital to filter out such pattern otherwise there would be a systematic bias in the estimation. We deal with seasonality in case of all variables to explain and some explanatory variables, for which the seasonal effect seems to be significant. The correction of seasonal effect in the variables to explain can be seen in Section A.1 of the Appendix. Although seasonal adjustment of the explanatory variables is not graphically represented, remarks about what explanatory variables were seasonally adjusted can be found in Table 2.1.

We apply the TRAMO/SEATS procedure elaborated by the Bank of Spain, which works almost automatically in many econometric programs. This methodology goes further beyond the scope of this thesis but can be found in Gomez and Maravall [1998] or Maravall [1996].

2.3.2. Stationarity Testing and Stationarity Adjustment of Both Variables Being Explained and Explanatory Variables

Phillips [1985] and Granger and Newbold [1974] illustrate that even absolutely unrelated variables can demonstrate very high fit in terms of the coefficient of determination (R^2) under the assumption that they all have a common (or reversed) trend. This problem is in econometrics referred as the spurious relationship and can potentially constitute a big problem in the least squares estimation leading to false inference about relationships among variables.

The spurious relationship problem occurs for a wide range of data generating processes, such as driftless unit roots, unit roots with drift, long memory, trend and broken-trend stationarity. As noted in Cipra [2008], the majority of time series is in fact non-stationary. All these processes should be recognised by a unit-root-type test. We apply the Augmented Dickey-Fuller test (ADF[9]) with the null hypothesis H_0 of stationarity as explained in Cipra [2008]. This test can be formally written in the following way:

$$\triangle X_t = \psi t + \Pi X_{t-1} + \sum_{j=1}^{p} \varphi_j \triangle X_{t-j} + \varepsilon_t$$

$$\begin{aligned} H_0: &\quad \psi = 0, \Pi = 0 \quad stationary \\ H_1: &\quad \psi \neq 0, or \Pi \neq 0 \quad non-stationary \end{aligned}$$

where ψ capures the deterministic trend of the variable, $\triangle X_{t-i}$ is the difference of i-th order of variable X and φ_j captures the remaining autocorrelation in residuals.

In case ADF test indicates non-stationarity of a particular variable, we must perform a modification that solves this problem. According to Cipra [2008], we distinguish between two types of non-stationarity. The trend non-stationarity can be eliminated by detrending the series (preferably by the Hodrick-Prescott filter). The difference non-stationarity is made stationary by applying differences. According to a nature of the particular variable, we decide for a particular type of nonstationarity and the corresponding type of adjustment.

It turns out that non-stationarity in case of all variables to explain is not a problem (i.e. stationarity is not rejected). In contrast, some explanatory variables exhibit a significant trend and stationarity is rejected on very low levels of significance. Thus there is a need to adjust them. In case non-stationarity of the original form of an explanatory variable is diagnosed, we don't provide any advanced testing whether it is of the nature of difference or trend non-stationarity. Such testing would go far behind the reach of this thesis. We choose for the type of adjustment according to a common sense about the the nature of the particular explanatory variable. The exact type of adjustment of a particular variable is noted in Table 2.1.

We eliminate the difference non-stationarity by applying log-differenciation of the variable subject to its value lagged by n months. Formally

$$y_t^{log-diff} = ln\left(\frac{y_t}{y_{t-n}}\right) \approx \frac{y_t - y_{t-n}}{y_{t-n}}$$

where y_t is the actual value of the variable at time t , y_{t-n} is its value lagged by n months and $y_t^{log-diff}$ the proportional change of the variable over last n months that is assumed to be stationary. The value of n is for convinience set to 12 because in case of other values not being a multiple of 12 we would have to deal with the seasonality problem.

[9] The Augmented Dickey-Fuller (ADF) test is usually used for testing the stationarity condition, which is implied by the time series having a constant mean (i.e. not having a trend) and a constant variance and covariance structure. However, we only require the variables to have a constant mean.

The trend non-stationarity will be eliminated by applying the Hodrick-Prescott filter (HP-filter) as described in Hodrick and Prescott [1997] for finding the trend value of the particular variable and subsequently by computing proportional deviations from this trend value. Formally

$$\underset{\tau_1 \dots \tau_T}{argmin} \left(\sum_{t-1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} [(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2 \right)$$

$$y_t^{HP} = \frac{y_t - \tau_t}{\tau_t}$$

where T is the number of observations, y_t is the actual value of the variable at time t, τ_t is the HP-filtered value of the variable, λ is the smoothing parameter and y_t^{HP} is the proportional deviation from the trend value that is assumed to be stationary. We choose the smoothing parameter λ in accordance with Ravn and Uhlig [2001] as 1600 for data with quarterly frequency and 129600 for data with monthly frequency.

2.3.3. Normalization of Explanatory Variables

As further discussed in Sections 3.2 and 3.3, multicolinearity among explanatory variables can constitute a big problem for the stability of regression coefficients. Greene [2012] suggests a way to moderate the problem multicollinearity causes to regression estimation by normalizing explanatory variables from Section 2.2 by applying the following formula

$$X_{i,t} = \frac{X_{i,t} - \hat{\mu}_{X_i}}{\hat{\sigma}_{X_i}} \tag{2.3}$$

where $X_{i,t}$ is the actual value of the explanatory variable i at time t, μ_{X_i} is the sample mean of the explanatory variable X_i over the examined period and $\hat{\sigma}_{X_i}$ is the sample standard deviation over the examined period. This transformation results in all explanatory variables having the sample mean of 0 and sample standard error of 1.

It should be noticed that transforming explanatory variables as described in Equation 2.3 does not reduce multicollinearity in terms of variance inflation factors (VIF[10]). However, it makes the calculation of the inverse $(X_t^T (\hat{\Omega}^{OLS})^{-1} X_t)^{-1}$ in generalized least squares estimation (GLS) in Equation 3.5 and in the formula for the variance of this estimate 3.6 more stable and less prone to rounding errors.

In addition, the estimated values for regression coefficients $\beta_{2,\dots,}\beta_k$ in Equation 3.1 can be more easily compared among themselves due to the fact that normalizing causes all explanatory variables to have the same variance. As a result, the proportions of absolute values among $\beta_{2,\dots,}\beta_k$ reflect the relative affects of explanatory variables on the endogenous variable. Also the intercept coefficient β_1 in Equation 3.1 will provide a rough information about the long-term average of DR_t.

We should also note that normalizing is always carried out as the last modification so that all explanatory variables that enter the regression models constructed later in this thesis really have sample variance of 1 and sample mean of 0.

[10] Variance inflation factors (VIF) for an individual explanatory variable and a model with k explanatory variables explaining DR_t can be computed as the coefficient of determination R^2 in a model with this variable as the variable to explain and all other k-1 varibles as the explanatory variables. See Trefethen [1997] and Greene [2012] for further details.

Chapter 3
Regression Model: Determinants of Aggregate Credit Risk

The influence of business cycle on aggregate credit risk has been both theoretically explained (see Kiyotaki and Moore [1997], Bernanke et al. [1999], Miao and Wang [2010]) and empirically evidenced (see for example Rosch [2003], Lucas and Koopman [2005]). On the other hand, the influence of other macroeconomic variables hasn't been due to our belief explored to a satisfactory extent[1] and the theory still doesn't provide sufficient background. Therefore, we start building our model by including a business cycle measure. Then, we examine if constructing a more complex model by adding additional explanatory variables can improve the performance of the model.

In this chapter, we first construct a model with only business cycle measure and its non-linear component as the explanatory variable and present its results in Section 3.4. Subsequently, we construct models that includes beside the business cycle measure also additional explanatory variables and present the results in Section 3.5. As previously noted, the list of candidate predictors can be extensive even if we for computational feasibility don't take into account their possible dynamic lag structure. On previous work on the this topic, most studies have picked macroeconomic predictors by qualitative reasoning (see Kalirai and Scheicher [2002], Zeman and Jurca [2008], Boss et al. [2009]). Boss et al. [2009] group macroeconomic predictors into thematic sets and allow only one variable from each set to be selected. We depart from these qualitative approach and opt for a data-driven selection mechanism where individual variables are added into the model successively based on their ability to improve its fit. This is done by the forward stepwise selection[2].

In Section 3.1, we describe models that are assumed behind the dynamics of aggregate credit risk. We outline the econometric theory and estimation procedures in Section 3.2. It is argued that using ordinary least squares estimation (OLS) can lead to dubious results and that using generalized linear squares estimation (GLS) allowing for autocorrelation in residuals is preferable. Section 3.3 describes criteria for choosing the optimal model.

[1] The insufficiency of theory explaining aggregate credit risk drivers is discussed for example in Kerbl and Sigmund [2011].

[2] Forward stepwise selection method adds first into the model the variable that single-handedly guarantees the highest fit in terms of the coefficient of determination (R^2). Subsequently, it adds to the model the variable that improves the fit to the highest extend subject to the variables that are already included in the model.

3.1. Model Description

We use a model where the measure of aggregate credit risk (DR_1 or DR_2) as the endogenous variable is modeled as a linear combination of macroeconomic variables that are regarded as exogenous. Formally, the model has the following general form:

$$DR_t = X_t\beta + \varepsilon_t = \beta_1 + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \varepsilon_t \tag{3.1}$$

where DR_t is the aggregate credit risk measure (either $DR_{1,t}$ or $DR_{2,t}$), $X_{2,t} .... + X_{k,t}$ are macroeconomic predictors and ε_t is an error term that is assumed to follow a martingale difference sequence with respect to sigma algebra generated by the past history of explanatory variables X[3]. Some previous papers (see Kerbl and Sigmund [2011], Klein [2013] and many others) used logit-transformed[4] value of the endogenous variable DR_t, which solves a major problem of Model 3.1 that the fitted values for DR_t have (unlike the actual values) neither lower nor upper bound. However, it is not feasible to perform logit transformation on values of $DR_{2,t}$ because given its definition in Equation 2.2, it can reach negative values. Therefore, in order to preserve consistency in modelling 3-month ex-ante aggregate default rates ($DR_{1,t}$) and 3-month ex-ante changes in non-performing loans ($DR_{2,t}$), we model them in their actual values.

Almost all previously published studies propose to add an arbitrary measure of business cycle into the model. For the reasons explained in Subsection 2.2.1, we choose the proportional deviation from HP-filtered industial production (IP) as the measure of business cycle. Some of the studies (see Marcucci and Quagliariello [2009]) also suggest to add a modification of the business cycle measure such as in the form of a censored variable allowing for an assymetric effect of business cycle. The argument is that the effect of business cycle to aggregate credit risk is so prominent and also documented to be non-linear that it can't be fully captured in only one variable. The assymetric affect of business cycle with more significantly negative effect in times of recessions on aggregate credit risk has been documented for example in Morales and Gasha [2004], Marcucci and Quagliariello [2009] and Nickell et al. [2001].

As a starting point for the analysis, we construct a simple model with 2 explanatory variables, both being a measure of business cycle. This model has the form of:

$$DR_t = \beta_1 + \beta_2 IP_{GAP,t} + \beta_3 (IP_{GAP,t} - \psi)^- + \varepsilon_t \tag{3.2}$$

where DR_t is either 3-month ex-ante aggregate default rate ($DR_{1,t}$) or 3-month ex-ante change in non-performing loans ($DR_{2,t}$), $IP_{GAP,t}$ is proportional deviation from HP-filtered industial production,

[3] The error term ε_t is assumed to be a martingale difference sequence with respect to the sigma algebra generated by the past history of the explanatory variables. This is a more realistic assumption than taking error terms as independently and identically distributed (IID) with N(0,σ^2). It allows for autocorrelation in disturbance, but their conditional expectation with past values of explanatory variables is zero. In case of Model 3.1, this condition is E($e_t|X_t, X_{t-1}, ...$)=0.

[4] Logit-transformation of DR_t would be executed as $log\frac{DR_t}{1-DR_t}$. Consequently, the aggregate credit risk would be modelled as

$$log\frac{DR_t}{1-DR_t} = X_t\beta = \beta_1 + \beta_2 X_{2,t} + ... + \beta_k X_{k,t}$$

It is straightforward to show that after the estimation of regression coefficients, the dependent value can be transfered back to non-transformed values as

$$DR_t = \frac{1}{1+e^{-X_t\beta}}$$

$(IP_{GAP,t} - \psi)^- = min(IP_{GAP,t} - \psi, 0)$ is truncated form of $IP_{GAP,t}$ allowing for the asymmetric effect of business cycle, ψ is the threshold value and ε_t is the error term. The slope of the regression line is $\beta_2 + \beta_3$ for values of explanatory variable bellow threshold and β_2 for values of explanatory variable $IP_{GAP,t}$ above threshold. This model belongs among a class of joinpoint regression models (see Kim et al. [2004], Muggeo [2003]) that were originally used in medical research. A preferable alternative to the model described in Equation 3.2 would be to use a threshold regression model[5] (see Hamilton [1989], Franses and Dijk [2000]). However, using a threshold models would not be consistent with the estimation of the model described in Equation 3.1 and testing for significance of the non-linear component would become more complicated (see Davies [1987]).

Literature suggests (see Muggeo [2003]) to estimate model described in Equation 3.1 by maximum likelihood estimation method (MLE). However, such estimation goes beyond the scope of this thesis and we will use the ordinary least squares method (OLS) or alternatively the generalized least squares method (GLS). Testing for significance of the non-linear component in Model 3.2 is then done simply based on the estimate of coefficient β_3 and the corresponding element of variance-covariance matrix as described in Equation 3.8. There is an apparent problem that we have to choose value for threshold ψ arbitrarily. Nonetheless, a grid search method as proposed in Lerman [1980] can be utilized to find its optimal value. In this procedure, we estimate the model for multiple values of threshold and subsequently select the value, which leads to the highest explanatory power of the model[6]. Our situation is simplified by the assumption that there exists only one threshold despite the fact that the influence of business cycle to aggregate credit risk could be theoretically more non-linear and a multiple-threshold model would be more appropriate. For the stability of results, we limit the search for threshold value ψ between 15% and 85% percentile values of $IP_{GAP,t}$. The optimal model selected by grid search is then compared to a simple model without the non-linear component in the form of:

$$DR_t = \beta_1 + \beta_2 IP_{GAP,t} + \varepsilon_t \tag{3.3}$$

.

3.2. Model Estimation

The standard way to estimate the vector β in model $DR_t = X_t\beta + \varepsilon_t$ is to use the ordinary least squares method (OLS), which computes an estimate of β by the following formula:

$$\hat{\beta}^{OLS} = (X_t^T X_t)^{-1} X_t^T DR_t \tag{3.4}$$

where X_t is the matrix of explanatory variables and DR_t is either 3-month ex-ante aggregate default rate ($DR_{1,t}$) or 3-month ex-ante change in non-performing loans ($DR_{2,t}$).

Section A.3 in the Appendix summarises assumptions of the Gauss-Markov theorem, i.e conditions

[5] An alternative threshold model based on Hamilton [1989], Franses and Dijk [2000] would have a form of

$$DR_t = (\beta_{01} + \beta_{11} IP_{GAP,t}) I(IP_{GAP,t} < \psi) + (\beta_{02} + \beta_{12} IP_{GAP,t}) I(IP_{GAP,t_t} \geq \psi) + \varepsilon_t$$

where $I()$ is the indicator function with the value of 1 when the condition in the brackets is fulfilled and 0 otherwise. The regression line estimated by this regression does not have to be continuous.

[6] The highest explanatory power of the model is reached through minimization of the error sum of squares (SSE) of equivalently through maximization of log-likelihood in case of MLE estimation.

for $\hat{\beta}^{OLS}$ to be a minimal-variance linear unbiased estimator. (Ass. 1),(Ass. 2),(Ass. 3) are sufficient conditions for $\hat{\beta}^{OLS}$ to be unbiased. (Ass. 4),(Ass. 5) are additional sufficient conditions for $\hat{\beta}^{OLS}$ to attain the lower bound for the variance of $\hat{\beta}^{OLS}$ given by the Cramér-Rao theorem.We will regard correlation between explanatory variables and error terms (Ass. 1) as a sign of ill-specification of the model (e.g. of omitted explanatory variables). Random sampling of X and DR_t (Ass. 2) can cause a systematic bias in the estimation, this is particularly dangerous in systemic bias in the choice of endogenous variable DR_t. Our data cover the period of 9/2003-12/2012, i.e. they have a length of cca. 9 years. So, if we accept the theory that aggregate credit risk follows a cycle and the lenght of the cycle is different from $9/k$ (with $k\varepsilon N$), Assumption 2 is not fulfilled and there is a systemic bias in $\hat{\beta}^{OLS}$. However, the verification of this assumption is beyond the scope of this thesis. Perfect collinearity (Ass. 3) is highly improbable and should not be a problem for out estimation procedure. However, it turns out that imperfect collinearity constitutes a significant problem through causing t-statistics for significance of individual variables in the model to be invalid. Homoscedasticity (Ass. 4) doesn't seem to be violated, which is proven by testing in empirical parts of this thesis. Finally, autocorrelation of residuals (Ass. 5) is probably the most immediate problem causing the estimation to be inefficient.

Greene [2012] shows the properties of $\hat{\beta}^{OLS}$ under the assumption that error terms ε_t follow an autoregressive process of order 1 (AR(1)), i.e. $\varepsilon_t = \rho\varepsilon_{t-1} + e_t$, where $|\rho| < 1$ and $e_t \overset{iid}{\approx} N(0, \sigma_e^2)$. Literature provides numerous evidence about error terms in models explaining aggregate credit risk being serially correlated (see Kerbl and Sigmund [2011], Boss et al. [2009]). As shown later in this thesis in Chapter 4, the assumption of error terms ε_t following AR(1) or a higher autoregressive process is not only suggested by literature but also backed up by our data. Under the assumption of error terms ε_t being uncorrelated with explanatory variables X (Ass. 1) unbiased regardless of serial autocorrelation of ε_t, $\hat{\beta}^{OLS}$ is still unbiased. The proof is provided in Section A.4 of the Appendix:

Variance of $\hat{\beta}^{OLS}$ under the assumption of error terms ε_t being homoscedastic and following AR(1) is also derived in Section A.4 of the Appendix. As can be seen there, $\hat{\beta}^{OLS}$ is inefficient and t-statistics for significance of individual variables in the model are invalid.

Therefore, we perform a modification of estimation procedure described in Equation 3.4 and estimate β by the generalized least squares method (GLS), which computes the estimate of β by the following formula:

$$\begin{aligned}\hat{\beta}^{GLS} &= (X_t^T\Omega^{-1}X_t)^{-1}X_t^T\Omega^{-1}DR_t = \\ &= \left[X_t^T\begin{bmatrix}1 & \rho & \dots & \rho^{n-1}\\ \rho & 1 & \dots & \rho^{n-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{n-1} & \rho^{n-2} & \dots & 1\end{bmatrix}^{-1}X_t\right]^{-1}\left[X_t^T\begin{bmatrix}1 & \rho & \dots & \rho^{n-1}\\ \rho & 1 & \dots & \rho^{n-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{n-1} & \rho^{n-2} & \dots & 1\end{bmatrix}^{-1}DR_t\right]\end{aligned} \tag{3.5}$$

where Ω is the correlation matrix of ε_t following AR(1) process in the form of $\varepsilon_t = \rho\varepsilon_{t-1} + e_t$, $|\rho| < 1$. For the estimation of β via GLS, it is essential to estimate parameter ρ in the correlation matrix of ε_t. Such estimate provided by using residuals $\hat{\varepsilon}_t^{OLS}$ from estimation of the model by OLS has the form of $\hat{\rho} = \hat{\rho}^{OLS} = \left(\sum_{i=2}^{n}\varepsilon_i^{OLS}\varepsilon_{i-1}^{OLS}\right)/\left(\sum_{i=2}^{n}(\varepsilon_{i-1}^{OLS})^2\right)$. For the purpose of GLS estimation, an estimate $\hat{\Omega}^{OLS}$

with off-diagonal elements given by respective powers of $\hat{\rho}^{OLS}$ is computed and the estimation of β is carried out by

$$\hat{\beta}^{GLS} = (X_t^T(\hat{\Omega}^{OLS})^{-1}X_t)^{-1}X_t^T(\hat{\Omega}^{OLS})^{-1}DR_t$$

.

Greene [2012] shows that $\hat{\beta}^{GLS}$ from Equation 3.5 is an unbiased estimate of β. Moreover, $\hat{\beta}^{GLS}$ is an efficient[7] estimate of β with variance computed as:

$$Var(\hat{\beta}^{GLS}) = \hat{\sigma}_\varepsilon^2(X^T(\hat{\Omega}^{OLS})^{-1}X)^{-1} = \tag{3.6}$$

where σ_ε^2 is computed as $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k}(\varepsilon_t^{GLS})^T\varepsilon_t^{GLS}$

3.3. Optimal Model Selection Procedure and Model Testing

In this section, we describe the approach for finding the optimal model. In case of K candidate explanatory variables, the number of possible model specifications is 2^K [8]. It is therefore highly desirable to have a criterion for comparing performance ability of diferent models at hand. This allows us to rank different specifications from the best to the worst and to present only results of the optimal model.

Because estimating the model for all possible 2^K specifications would not be feasible, we choose the forward stepwise selection method. This method first adds into the model the variable that single-handedly guarantees the highest fit. Subsequently, it adds to the model the variable that improves the fit to the highest extend subject to the variables that are already included in the model. This approach reduces the number of models to estimate from 2^K to $\frac{K(K+1)}{2}$ [9]. At the end, we have K models, one that was indicated as optimal by forward stepwise selection for given number of included exogenous variables ranging from 1 to K. Out of K models, we pick the one that maximizes the optimality criterion in Equation 3.7. For this model we present the results and test if it satisfies properties of a good model. If it doesn't satisfy these properties, we may consider switching to a more/less complex model.

Alternatives to the forward stepwise selection method would be the backward stepwise selection, which operates similarly as the method we use (accept that in the reversed order), or the best subset selection method (see Hastie et al. [2009]), which would however be excessively computationally demanding.

A natural way to assess the performance of different model specifications is based on proportion of total sum of squares (SST) they are able to explain, which is captured in a statistics called coefficient

[7] Efficiency is of $\hat{\beta}^{GLS}$ is reached through elimination of autocorrelation in ε_t. Generalized least squares method (GLS) relies on Cholesky decomposition of Ω. As described in Trefethen [1997], each non-singular symmetric matrix can be via LU factorization decomposed into a lower triangular matrix and its inverse, i.e $\Omega^{-1} = PP^T$, where P is a lower tringular matrix. Pre-multiplying the equation $DR_t = X_t\beta + \varepsilon_t$ by P eliminates autocorrelation because modified residuals $P\varepsilon_t$ generate the correlation matrix $E[(P\varepsilon_t)(P\varepsilon_t)^T] = E[P\varepsilon_t\varepsilon_t^T P^T] = \sigma_\varepsilon^2\mathrm{I}$, which only has elements on its diagonal.

[8] The number of possible model specifications is 2^K because each candidate explanatory variable can either be included or not. Therefore, every additional variable doubles the number of possible specifications.

[9] In case of forward stepwise selection and K candidate explanatory variables, the number of models to estimate is under $K + (K-1) + ... + 2 + 1$, which equals $\frac{K(K+1)}{2}$.

of determination (R^2)[10]. However, any model selection procedure based on R^2 tends to include additional variables into the model even if their contribution to the model performance negligible. Due to the need for a criterion weighing between improvement of the model fit and the model parsimony, we employ a modification of R^2 that imposes a penalty for any additional variable included. So as suggested in Ohtani and Hasegawa [1993], we set the criterion of optimality as the adjusted coefficient of determination ($\bar{R}_k^2$)[11]:

$$\overline{R}_k^2 = 1 - \frac{SSE_k/(n-k)}{SST/(n-1)} \tag{3.7}$$

where $SSE_k = \sum_{i=1}^{n}(y_i - (\beta_0^{GLS} + \sum_{j=1}^{k}\beta_j^{GLS}x_j))^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i^{GLS})^2$ is error sum of squares defined as the sum of squared residuals from the model, $SST = \sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is total sum of squares defined as the sum of squared deviations from the average value of endogenous variable, n is number of of observations in our dataset and k is number of explanatory variables in the particular model. It can be shown that a model that maximizes $\overline{R}_k^2$ also minimizes unbiased estimates of error terms in form of $\sigma_\varepsilon^2 = \frac{1}{n-k}(\varepsilon_t^{GLS})^T\varepsilon_t^{GLS}$. The choice of adjusted coefficient of determination ($\overline{R}_k^2$) will probably result in a more complex model than using other potential criteria, such as Akaike information criterion (AIC) or Bayesian information criterion (BIC), because the number of regression parameters (k) is penalized less strongly in case of $\overline{R}_k^2$.

Throughout the estimation procedure, we can infer about the influence of individual variables based on the sign of coefficients from regression estimation. Nonetheless, an analysis of significance of such influence should be carried out. Even if the true value of β_i is zero, stochastic components of the model can cause the explanatory variable to have a nonzero estimate of the coefficient (i.e. $\hat{\beta}_i \neq 0$). For this reason, there is a need for testing whether the estimated coefficient corresponding to an exlanatory variable differs significantly from zero. Such testing is carried out by computing t-statistics for each regression coefficient by comparing it's estimated value $\hat{\beta}_i$ with a hypothetical zero value prevailing under the hypothesis that the explanatory variable is not relevant for explaining aggregate credit risk (i.e. $H_0 : \beta_i = 0$). The t-statistics is computed in the following way:

$$t_i = \frac{\widehat{\beta}_i^{GLS} - \beta_i^0}{\hat{\sigma}[(X^T(\hat{\Omega}^{OLS})^{-1}X)^{-1}]_{ii}^{\frac{1}{2}}} = \frac{\widehat{\beta}_i^{GLS}}{\hat{\sigma}[(X^T(\hat{\Omega}^{OLS})^{-1}X)^{-1}]_{ii}^{\frac{1}{2}}} \approx t_{n-k} \tag{3.8}$$

where $\hat{\beta}_i$ is the estimate of β_i from linear regression, β_i^0 is the value of β_i consistent with H_0, $\hat{\sigma}$ is the estimate of standard deviation and $\hat{\sigma}[(X^T(\hat{\Omega}^{OLS})^{-1}X)^{-1}]_{ii}$ is the element of variance-covariance matrix from Equation 3.6 corresponding to coefficient β_i . It can be shown that the t-statistics should theoretically follow student distribution with $n-k$ degrees of freedom (t_{n-k}) and therefore each variable with it's t-statistics exceeding te corresponding critical value ($|t| > t_{n-k}(1-\frac{\alpha}{2})$) is deemed significant on the α level of significance.

[10] Coefficient of determination (R^2) for a model of k variables is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - (\beta_0^{GLS} + \sum_{j=1}^{k}\beta_j^{GLS}x_j))^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i)^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i^{GLS})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE_k}{SST}$$

where SSE_k is the residual sum of squares defined as the sum of squared residuals from the model and SST is the total sum of squares defined as the sum of squared deviations from the average value of the endogenous variable.

[11] The form of $\bar{R}_k^2$ has a strong connection to the fact that $\hat{\sigma}_{\varepsilon,k}^2 = \left(\sum_{i=1}^{n}(y_i - (\beta_0^{GLS} + \sum_{j=1}^{k}\beta_j^{GLS}x_j))^2\right)/(n-k)$ is an unbiased estimate of error term variance if the model with k regressors holds and $\hat{\sigma}_{\varepsilon,1}^2 = \left(\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i)^2\right)/(n-1)$ is an unbiased estimate of error term variance if the model with intercept only holds.

Explanatory variables may seem to be insignificant due to correlation between individual variables. Any explanatory variables added to the model contributes to the multicollinearity problem. As a result, some elements of matrix $(X^T(\hat{\Omega}^{OLS})^{-1}X)^{-1}$ may be of large values. Some degree of multicullinearity might be tollerable because it doesn't violate assumptions for $\hat{\beta}^{GLS}$ in Equation 3.5 to be unbiased. On the other hand, it can make the estimates inefficient and the results unstable. For testing the degree of collinearity being present in explanatory variables, we apply a rule of thumb as presented in Trefethen [1997], which relies on comparing the largest and the smallest (both in absolute terms) values of matrix $(X^T(\hat{\Omega}^{OLS})^{-1}X)$

$$\kappa = \frac{\lambda_{max}\left(X^T(\hat{\Omega}^{OLS})^{-1}X\right)}{\lambda_{min}\left(X^T(\hat{\Omega}^{OLS})^{-1}X\right)} \in \langle 1, +\infty)$$

where κ is condition number of matrix $X^T(\hat{\Omega}^{OLS})^{-1}X$, λ_{max} (λ_{min}) is the largest (smallest) eigenvalue of matrix $(X^T(\hat{\Omega}^{OLS})^{-1}X)$. Trefethen [1997] states that for values of condition number between 1 and 100, multicollinearity problem can be ignored. Values between 100 and 1000 imply a moderate problem of multicollinearity. For values greater than 1000, the multicollinearity problem gets severe. We compute condition number for each model that is indicated as optimal by adjusted coefficient of determination ($\overline{R}^2_k$). However, it is beyond the scope of this thesis to solve the problem of multicollinearity[12] by other means than normalizing explanatory variables as descried earlier in this thesis. If the multicollinearity problem despite normalized explanatory variables still remains severe, we should consider applying a model with fewer explanatory variables and therefore with lower predisposition to correlation .The extreme case of perfect multicollinearity among exogenous variables, which violates the assumptions of Gauss-Markov theorem (Section A.3 of the Appendix), is however extremely unlikely to occur in the data and its prospective incidence would be immediately recognized due to non-existence of $(X^TX)^{-1}$ and resulting failure of regression estimation

Although individual variables may not seem significant, the model might still be highly significant as a whole and thus highly relevant for explaining aggregate credit risk. We can demonstrate this by constructing an F-test relying on the assumption that $\hat{\beta}^{GLS}$ defined in Equation 3.5 is an unbiased estimate of β with variance $Var(\hat{\beta}^{GLS})$ defined in Equation 3.6. Greene [2012] presents slightly modified form of F-test for joint significance of regression coefficients by comparing the sum of squared errors from the restricted model (H_0) with $k-1$ restrictions on $\beta_2 = 0, \beta_3 = 0, ..., \beta_k = 0,$. and the unrestricted model allowing the slope coefficients to be non-zero ($\beta_2 \neq 0, \beta_3 \neq 0 ...., \beta_k \neq 0$). We construct this F-statistics in the following way:

$$F = \frac{\frac{SST-SSE}{k-1}}{\frac{SSE}{n-k}} \approx F_{k-1,n-k} \tag{3.9}$$

where $SST = \sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is total sum of squares from model with $k-1$ restriction allowing for the intercept only, $SSE = \sum_{i=1}^{n}(y_i - (\beta_0^{GLS} + \sum_{j=1}^{k}\beta_j^{GLS}x_j))^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i^{GLS})^2$ is error sum of squares from unrestricted model with k regression coefficients, n is the number of observations and $k-1$ is total number of restrictions. In order to test for the joint significance of the model, it can be shown that the F-statistics should theoretically follow Fisher-Snedecot distribution with $k-1$ and

[12] Literature suggests several ways to tackle the multicollinearity problem. Some of the ways are Ridge regression (see Hoerl and Kennard [1970]), LASSO regression (see Tibshirani [1994]) or famous method of principal component analysis.

$n-k$ degrees of freedom ($F_{k-1,n-k}$) and each model with it's F-statistics exceeding the corresponding critical value (i.e. $F > F_{k-1,n-k}(1-\alpha)$) is deemed jointly significant on the α level of significance.

At the end of forward stepwise selection, we present results for the model that generates the highest adjusted coefficient of determination ($\bar{R}^2_k$) and provide testing whether residuals satisfy properties of a good model.. We concentrate on homoscedasticity of residuals (Breusch-Pagan test, see Breusch and Pagan [1979]), normality of residuals (Jarque-Bera test, see Jarque and Bera [1987]) and zero autocorrelation in residuals (Ljung-Box test, see McLeod and Li [1983]). Methodics of all tests is presented in Section A.5 of the Appendix.

3.4. Model with Business Cycle Measures as the Explanatory Variable

In this section, we present results for the model with business cycles measures (IP_GAP) and its censored component as the explanatory variables. This leads to two alternative model specification (one with the censored component and one without), which have the following forms

$$DR_t = \beta_1 + \beta_2 IP_{GAP,t} + \varepsilon_t \quad (1)$$

$$DR_t = \beta_1 + \beta_2 IP_{GAP,t} + \beta_3 (IP_{GAP,t} - \psi)^- + \varepsilon_t \quad (2)$$

First and foremost, it turns out that despite the arguments stated earlier in this thesis, it is inappropriate to estimate Model (1) and Model (2)[13] by generalized least squares estimation (GLS) because such a method of estimation would lead to low explanatory power of the model. In the subsequent section of this thesis (Section 4.4), we argue that omission of relevant variables through applying an insufficiently extensive model can lead to a cycle in residuals with an extremely long period. To express this argument more specifically, business cycle conditions cannot be expected to capture the full macroeconomic environment driving the aggregate credit risk. In the methodics presented in Section 3.2, such a long period in residuals results in very dense correlation matrix of residual $\hat{\Omega}$, i.e. having non-zero off-diagonal elements with insufficient rate of convergence towards zero. This makes the generalized least squares estimation not work well in terms of achieving high coefficient of determination (R^2).

Earlier in this thesis, we have also pointed at the unbiasedness of ordinary least squares estimation (OLS) of vector β under serially autocorrelated residuals. The only problem will constitute the subsequent inference about significance of individual variables in the model. Because we have also derived the inefficency of OLS estimation (see Section A.4 in the Appendix) resulting in non-validity of t-test[14] for significance of individual variables, the results for t-tests should be taken with caution.

[13] The variable denoted as $(IP_{GAP,t} - \psi)^-$ is defined as $min(IP_{GAP,t} - \psi, 0)^-$.

[14] t-test for individual significance of the explanatory variable i is derived as

$$t_i = \frac{\widehat{\beta}_i^{OLS}}{\hat{\sigma}[(X^T X)^{-1}]_{ii}^{\frac{1}{2}}}$$

Because we have concluded that $\sigma_\varepsilon^2 (X^T X)^{-1}$ is not the unbiased estimator of the variance of β^{OLS}, the denominator in the expression for t_i is biased and the statistics does not hold.

Secondly, we should remind of the fact that all explanatory variables used in this analysis are normalised. An unpleasant consequence of this modification is the imposibility of interpreting slope coefficients β_2 and β_3 as the percentage increase of endogenous variables with one percent increase of the business cycle conditions.

Table 3.1 shows the results of ordinary least squares estimation (OLS) of Model (1) and Model (2) for $DR_{1,Total}$ and $DR_{2,Total}$. The results are in line with our expectations. The intercept coefficient β_1 is more significantly positive for $DR_{1,Total}$, which reflects the fact that $DR_{1,Total}$ is shifted upwards by a constant in comparison with $DR_{2,Total}$. For Model (1), the slope coefficient β_2 is estimated as negative and therefore indicates a negative correlation between business cycle and aggregate credit risk.

For Model (2), we can observe a difference in the sign of β_2 and β_3. In line with the definition of the censored component, a unit increase in IP_{GAP} (after normalization) bellow threshold is transfered to $DR_{1,Total}$/$DR_{2,Total}$ as a $\beta_2+\beta_3$ unit increase. In contrast, a unit increase in IP_{GAP} (after normalization) above threshold is transfered as a β_2 unit increase. Our expectation is that the estimate for coefficient β_3 should be significantly negative and so causing more negative effect of business cycle under the threshold. The t-statistics for β_3 would normally represent the significance of the assymetric effect of business cycle to aggregate credit risk but we know it is biased. The question is by how much biased. Because in Table 3.1 can be seen that the value of t-statistics for β_3 is approximately 6.5 for both $DR_{1,Total}$ and $DR_{2,Total}$, there is in fact a significant buffer against the bias in the OLS estimation for variance. Even if the actual variance was 3 times higher than indicated by OLS, the t-statistics computed by using the real variance would still exceed the critical value that indicates β_3 as significant. Thus, we can include that there is probably a nonlinear effect of business cycle on aggregate credit risk. To make a conclusion about the estimation of Model (2), the effect of business cycle is indicated as negative bellow the threshold ($\beta_2+\beta_3<0$) and surprisingly, though not significantly, positive ($\beta_2>0$) above the threshold.

Results for Model (1) and Model (2) for different segments of corporate clients are included in Section A.6 and A.7 of the Appendix. As can be seen there, the significance of the assymetric effect of the business cycle is not as high as in the preceeding case. Nonetheless, most sectors follow the similar pattern of business cycle having significantly negative effect on aggregate credit risk bellow the threshold and less significantly negative effect above the threshold.

The most important difference among analyzed segments is probably in the optimal value of the threshold. We have selected the value of threshold that maximizes the coefficient of determination (R^2) of the regression model by applying a grid search. Results for the grid search can be found in Section A.8 of the Appendix. The coefficient β_3 has an effect only for values of explanatory variable bellow the threshold and thus different values of the threshold select between low values of the business cycle measure having an effect on fewer observations with higher weight or on more observations with lower weight. However, we don't consider results obtained by applying grid search for finding the optimal value of threshold to be robust.

The inference between the value of threshold ψ and assymetric effect of business cycle through the coefficients $\beta_2$and β_3 can be seen in Figure 3.1. For $DR_{1,Total}$ the value of threshold ψ is significantly different from the case of $DR_{2,Total}$. The resulting number of observations bellow and above the threshold are denoted as N_1 and N_2 in Table 3.1. Model (2) fully captures the deterioration in $DR_{1,Total}$ in 2009 and 2010 while for the rest of the analysed period it has very low explanatory power, which

Table 3.1.: **Results for the Model with Business Cycle Measures as the Explanatory Variable**

	$DR_{1,Total}$		$DR_{2,Total}$	
Model	(1)	(2)	(1)	(2)
β_1	0.0253	0.0232	0.0038	-0.0035
	(26.17,***)	(4.89,***)	(4.31,***)	(-1.44,)
β_2	-0.0030	0.0008	-0.0044	0.0031
	(-3.08,***)	(1.02,)	(-4.97,***)	(1.65,*)
β_3	-	-0.0115	-	-0.0165
	-	(-6.53,***)	-	(-6.79,***)
ψ	-	-0.5465	-	0.1150
N_1	112	85	112	51
N_2	-	27	-	61
R^2	0.0802	0.1832	0.1850	0.4529
F	0.0025	0.0000	0.0000	0.0000
LB	0.0000	0.0000	0.0000	0.0000
JB	0.0445	0.0152	0.5000	0.5000
BP	0.1321	0.1628	0.3711	0.1211

* indicates significance on 10% level

** indicates significance on 5% level

*** indicates significance on 1% level

N_1 : number of observations for which the explanatory variable is above threshold ψ

N_2 : number of observations for which the explanatory variable is bellow threshold ψ

F: p-value for the F-test

LB: p-value for Ljung-Box test

JB: p-value for Jarque-Bera test

BP: p-value for Breusch-Pagan test

ψ: threshold in the effect of business cycle

Source: Author's Calculation

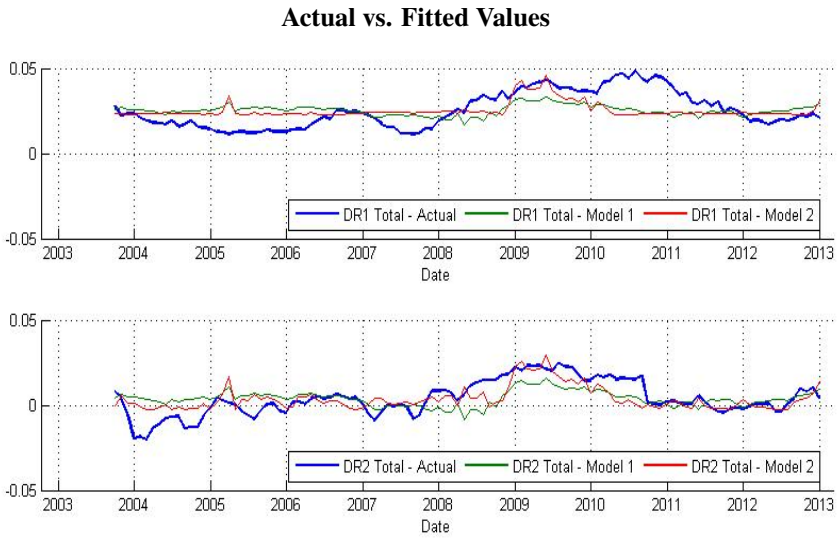


Figure 3.1.: **Actual vs. Fitted Values of $DR_{1,Total}$ and $DR_{2,Total}$ for the Model with Multiple Variables**

Source: Author's Calculation

is a direct consequence of selecting a very low threshold value. In contrast, the threshold in case of $DR_{2,Total}$ is selected much higher and its explanatory power is more well-proportioned, i.e. it doesn't fully capture the deterioration in 2008-2010 but perfoms better on the rest of the analyzed interval than in case of $DR_{1,Total}$.

3.5. Model with Multiple Explanatory Variables

This section presents results for the model with multiple explanatory variables as well as an analysis of the most frequently selected macroeconomic variables across analyzed sectors. The model to estimate is in the form of:

$$DR_t = X_t\beta + \varepsilon_t = \beta_1 + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \varepsilon_t$$

where X_2, ..., X_k are various explanatory variables added into the model. The entire list of candidate variables can be found in Table 2.1

As discussed earlier in Section 3.3, we use the forward stepwise selection method to select the best subset of explanatory variables. One important distinction between this approach and the approach followed for example in Kalirai and Scheicher [2002] and Koopman et al. [2009] is that they added variables to the model by mere qualitative reasoning and thus came to more parsimonious models with lower explanatory variables because their sets of candidate explanatory variables were more limited. In contrast, we build our model in an atheoretical way by adding variables solely according to their ability to improve the coefficient of determination (R^2).

Results for $DR_{1,Total}$ and $DR_{2,Total}$ are included in Table 3.2. For better understandability, we present only signs of regression coefficients and its significance. Complete results with the exact

Table 3.2.: **Results for the Model with Multiple Explanatory Variable**

	$DR_{1,Total}$		$DR_{2,Total}$	
	Sign	**Significance**	**Sign**	**Significance**
IP_GROWTH	not included		-	***
VACANT	not included		+	***
RET_PX50	-	***	-	***
GOV3	+	**	not included	
GOV10-GOV3	+	**	not included	
CREDIT	-	***	-	***
ER	-	***	-	***
CA	-	***	not included	
PPI	-	***	-	**

* indicates significance on 10% level

** indicates significance on 5% level

*** indicates significance on 1% level

Source: Author's Calculation

size of the regression coefficients and some statistics for model diagnostics are presented in Section A.9 of the Appendix. As can be seen in Table 3.2, the optimal model specifications for $DR_{1,Total}$ and $DR_{2,Total}$ include 4 common variables. This further supports our assumption that these 2 measures of aggregate credit risk are fully comparable.

For $DR_{1,Total}$ and $DR_{2,Total}$, the optimal size of the subset of explanatory variables is determined as 7 and 6 respectively. For the rest of analyzed sectors, the optimal subset size varies between 6 and 10. This is far from a parsimonious model specification because as we have already noted, using the adjusted coefficient of determination ($\overline{R}^2_k$) as a criterion for optimal model selection may lead to more extensive specification than using other criteria.

The plot of fitted values of $DR_{1,Total}$ and $DR_{2,Total}$ against their actual values can be seen in Figure 3.2. As can be seen there, the optimal regression models fits actual data very well with no substantial shortcomings in any part of the analysed period. The same plots for specific segments of corporate clients can be seen in Section A.10 in the Appendix. We can notice substantial cyclical fluctuations around the fitted values for some segments, such as $DR_{2,Est}$ (estate) and $DR_{2,Sal}$ (sale).

An important question is what macroeconomic variables are selected by the forward stepwise selection. Table 3.3 provides an analysis of the most frequently selected macroeconomic variables and thus indicates how important a particular variable is in explaining the aggregate credit risk across different sectors. Its first numerical column indicates the number of optimal model specifications the variable appears in. There are in total 8 analyzed segments (2 aggregate categories $DR_{1,Total}$, $DR_{2,Total}$ and 6 specific segments of corporate clients) and thus a frequency of 4 indicates that the variable is included in a half of optimal model specifications. The respective 2 columns on the right represent the frequency, in which the variable was indicated as significantly positive and significantly negative on 5% level.

The results for the regression coefficients are surprisingly significant. Out of 59 estimates of regression coefficents for 8 different models, 46 are significant on 5% level. Estimates are made by generalized least squares estimation (GLS) and therefore inference about individual significance of variables from respective t-statistics is asymptotically valid. Thus, our concerns made earlier in this thesis that substantial multicollinearity among explanatory variables will result in insignificance of

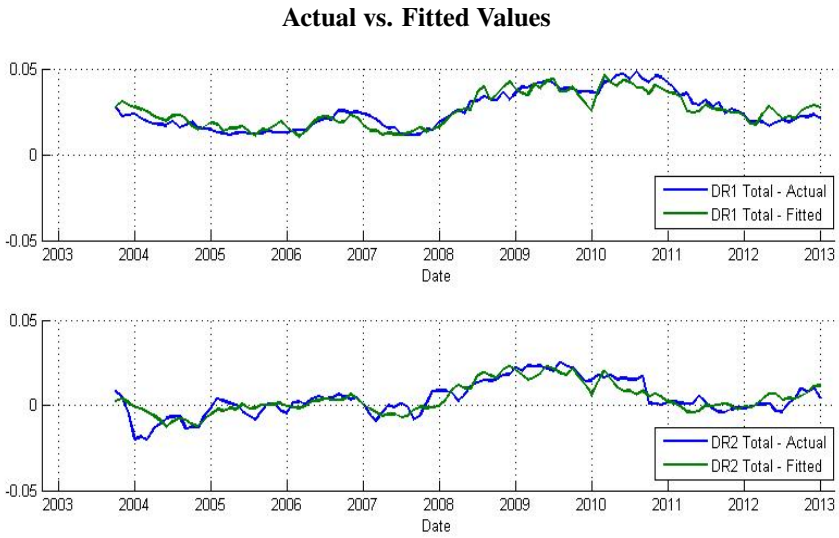


Figure 3.2.: **Actual vs. Fitted Values of $DR_{1,Total}$ and $DR_{2,Total}$ for the Model with Business Cycle Measures as the Explanatory Variable**
Source: Author's Calculation

estimation did not realize.

A particularly interesting finding is that the growth of industrial production (IP_GROWTH) is more frequently added to optimal models specification than the proportional deviation of industrial production from its equilibrium value (IP_GAP). Among other frequently included variables, we should name return on stock market index PX50 (RET_PX50), 3 year treasury yield (GOV3) and exchange rate (ER).

Results are generally in line with our expectations. There are only a few cases, in which the resulting sign of correlation wasn't predicted correctly. One of possible explanations for these exceptions is taking wrong assumptions about the sample of corporate clients we analyze. For example, exchange rate (ER) was indicated as having significantly negative effect on aggregate credit risk in 4 out of 4 cases. Earlier in this thesis, we didn't provide a clear prediction about the sign of correlation because we were not sure whether in case of currency depreciation the positive effect on exporters or the negative effect on importers prevails. To sum up, significantly negative sign of correlation between ER and DR_t in all cases indicates that the positive effect on exporters and on better compatitiveness of the Czech economy provails. Also the effect of producer prices index (PPI) wasn't anticipated correctly. This might be a consequence of changes in producer prices index (PPI) not having as clear effect as we have expected earlier in this thesis. Further research would be necessary to explain interactions between producer prices index (PPI) and aggregate credit risk.

On the other hand, the correlation sign in case of number of unemployed workers (UNEMP) and number of vacant positions (VACANT) is puzzling. We have expected a straightforward relationship between UNEMP and VACANT on one side and aggregate credit risk on the other side, but results are rather opposite to our assumptions.

To sum up, we have found several variables that drive aggregate credit risk across different sectors of the economy and thus are particularly useful for its understanding (and forecasting). Among these variables, we stress out growth of industrial production (IP_GROWTH), return on the stock market

Table 3.3.: **Frequency of Selected Explanatory Variables for the Model with Multiple Explanatory Variables**

Group	Variable	Frequency	Significantly Positive	Significantly Negative	Expectation
Business	IP_GAP	3	1	1	-
	IP_GROWTH	6	0	4	-
Household	UNEMP	2	0	2	+
	VACANT	4	2	0	-
Financial	RET_PX50	6	1	5	-
	VOL_PX50	3	2	1	+
	GOV3	5	3	2	+/-
	GOV10-GOV3	3	1	1	-
Credit	PRIBOR3M	3	0	2	+/-
	CREDIT	4	0	3	-
External	ER	4	0	4	+/-
	CA	4	1	3	-
Price	CPI	4	1	3	-
	PPI	4	0	3	+
	M1	4	0	1	-

Variable is deemed significant if it is significant on 5% level by the t-test.

Source: Author's Calculation

index (RET_PX50), treasury yield (GOV3) and exchange rate vis-a-vis Euro (ER). We limited our analysis to explanatory variables that were expected to have a significant influence across all sectors. In case of a more detailed analysis specialized in a particular sector, we could certainly use other, sector-specific explanatory variables (i.e. different, specialized variables for each sector) to guarantee a better fit in each sector. Taking sectoral differences into account would be certainly a way how to guarantee better explananatory power of the models. However, such analysis would go far beyond the scope of this thesis.

Chapter 4
Latent Factor Extension: Search for the Default Cycle

In this section, we develop an idea that models from the previous section can be further improved by adding a latent factor, which is not observable in explanatory variables.

Litetature provides mounting evidence of latent factors driving aggregate credit risk . Some studies (see Bruche and Gonzalez-Aguado [2008], Koopman et al. [2008]) assume that the state variable driving credit risk is descrete and the number of states is at least two, i.e. a "good and " a "bad" state. These models are commonly referred to as hidden Markov models. In contrast, some other papers (see Boss, Fenz, Pann, Puhr, Schneider, and Ubl [2009], Kerbl and Sigmund [2011], Jimenez and Mencia [2007]) choose a more general approach allowing the latent factor influencing credit risk to be a continuous variable.

For the purpose of this thesis, we select the latter approach of modelling the latent factor as a continuous variable following an autoregressive process. Such idea is being referenced in literature as a default cycle. We incorporate this idea into the already-existing regression model from Chapter 3 by adding an unobservable component. The description of the model is introduced in Section 4.2 with the estimation methodology being introduced in Section 4.3.

One theory relates the existence of the default cycle to the leverage cycle (see Geanakoplos [2009], Fostel and Geanakoplos [2008]). Fostel and Geanakoplos [2008] argue that even a small drop in the value of assets used as collateral can cause a significant decline in wealth of leveraged investors, which is further amplified by forced sales carried out by investors with insufficient funding liquidity.

The second default cycle theory assumes that too lenient standards in providing new loans in times of economic prosperity result in an accumulation of risk, which materializes in times of subsequent economic downturn. This phenomenon can be reinforced by profit-maximizing strategies of commercial banks in times of plentiful liquidity (see Ruckes [2004]).

Third, some papers (see Giesecke [2004], Eisenberg and Noe [2001]) argue that the default cycle is driven simply by itself with periods of high default rates having a tendency to prevail. Because defaulted companies usually have some obligations to other companies, which are very improbable to be met, current high default rates may results in secundar insolvency of other companies and thus in an existence of the default cycle.

Beside the three above-mentioned theories regarding the existence of the default cycle, there is also an alternative explanation. The presence of a default cycles in data can be caused simply by omission of relevant explanatory variables due to our imperfect understending of aggregate credit risk drivers and resulting failure to propose appropriate candidate explanatory variables. It can not only trigger a bias in the ordinary least squares (OLS) estimation, but it also causes an illusion of a cycle beyond the ordinary linear regression model. This thesis has already presented a well-illustrating example in Section 3.4.

4.1. Latent Factor Extension of Linear Regression Model

The generalized least squares estimate β^{GLS} from Equation 3.5 in Section 3.2 relies on the assumption that residuals ε_t follow an autoregressive process of order 1 (AR(1)). Formally

$$DR_t = X_t\beta + \varepsilon_t \tag{4.1}$$

$$\varepsilon_t = z_t + e_t \tag{4.2}$$
$$z_t = \phi z_{t-1} + \xi_t \tag{4.3}$$

where DR_t is the aggregate credit risk measure (either $DR_{1,t}$ or $DR_{2,t}$), X_t are macroeconomic predictors with corresponding effects on DR_t captured in vector β (which is already consistently estimated by β^{GLS} [1]). The error term ε_t is now assumed to follow an AR(1) process with parameter ϕ and error terms e_t, ξ_t being white noise[2].

We can allow z_t to follow a more complicated autoregressive process of order p (AR(p)). Formally

$$z_t = \phi_1 z_{t-1} + ... + \phi_p z_{t-p} + \xi_t \tag{4.4}$$

where ϕ_1, ϕ_2, ..., ϕ_p are parameters of the AR (p) process.

Due to the definition of $DR_{1,t}$/$DR_{2,t}$ as 3-month aggregate default rates / 3-month aggregate changes in NPL, each default event (i.e. realization of aggregate credit risk) should appear in data for 3 consecutive months. This comes directly from the fact that any loan that default at time t was non-defaulted at $t-1, t-2$ ... . Such a pattern corresponds to error terms e_t in Equation 4.2 having an effect at time t, $t+1$ and $t+2$. This together with the assumption that the latent factor z_t follows AR (p) process results in the model form described in Equations 4.5-4.6 at the beginning of next section.

[1] A partial drawback of computing $\hat{\varepsilon}_t$ as the residuals from a linear model is the fact that β^{GLS} is estimated using informations, which are not available at time t. β^{GLS} could be possibly estimated at each period. However, it would make this analysis not feasible.

[2] White noise is an independently and identically distributed realization of normal distribution with $\mu = 0, \sigma = 1$ (i.e. $e_t \overset{iid}{\approx} N(0,1)$) and $\xi_t \overset{iid}{\approx} N(0,1)$)

4.2. Model Description

In this section, we come out from the idea that residuals $\hat{\varepsilon}_t^{GLS}$ from the Equation 3.5 in Section 3.2 are driven by a latent factor z_t, which follows AR (p) process, and describe a procedure for its estimation. The system to estimate can be formally described as

$$\hat{\varepsilon}_t^{GLS} = z_t + e_t + e_{t-1} + e_{t-2} \tag{4.5}$$

$$z_t = \phi_1 z_{t-1} + ... + \phi_p z_{t-p} + \xi_t \tag{4.6}$$

where $\phi_1, \phi_2, ..., \phi_p$ are parameters of the AR (p) process and e_t, ξ_t are error terms, which are both assumed to be independently and identically distributed (IID) .

Parameters $\phi_1, \phi_2, ..., \phi_p$ could be alternatively estimated by maximum likelihood estimation (MLE). However, such a procedure would not be able to distinguish between the nature of e_t having an effect at time $t, t+1, t+2$ only and ξ_t having an effect at time t directly and at more distant periods indirectly through its propagation by the autoregressive effect.

Kalman [1960] describes a recursive solution to the discrete data filtering problem, which is able to solve this problem. This method, being commonly refered to as Kalman filter, will directly result in explicit estimation of the latent factor z_t representing the default cycle. Kalman filter requires a modification of the system described in Equations 4.5-4.6 into a state space representation, which will allow its estimation.

Any state space model consists of two equations: a measurement equation and a transition equation. Measurement equation describes the relation between observed variables and unobserved state variables (i.e. latent factor z_t). Transition equation describes the dynamics of the state variables. The transition equation has the form of a first-order difference equation. The measurement and transition equation of the system to be estimated in this thesis by the Kalman filter are represented by Equations 4.7 and 4.8 respectively.

We use the state space representation[3] for AR (p) model as described in Durbin and Koopman [2001] and Hamilton [1994], which has the following form

$$\hat{\varepsilon}_t^{GLS} = Z\alpha_t + Ge \tag{4.7}$$

$$\alpha_{t+1} = T\alpha_t + H\xi_t \tag{4.8}$$

$$\xi_t \overset{iid}{\approx} N(0, \sigma_\xi^2)$$

$$e_t \overset{iid}{\approx} N(0, \sigma_e^2)$$

$$\alpha_1 \approx N(0, P_1)$$

Denote $m = min(p, 3)$. Then T, Z, G, H[4] are the following time invariant system matrices

[3] We use this particular representation described in Durbin and Koopman [2001] and Hamilton [1994] because there is a sofftware package for MATLAB presented in Peng and Aston [2011], which uses this representation. However, there would be other possible options of representing the same system. One of the other options would be to use the representation described in Harvey [1993].

[4] 0_{mxn} refers to a zero matrix of dimension mxn (m rows and $n$columns).

$$T = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ \phi_{m-1} & 0 & 0 & & 1 \\ \phi_m & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 0_{1x(m-1)} \end{bmatrix}$$

$$G = \begin{bmatrix} 1_{1x3} & 0_{1x(m-3)} \end{bmatrix}$$

$$H = \begin{bmatrix} 1 \\ 0_{(m-1)x1} \end{bmatrix}$$

The state vector α_t has the form of

$$\alpha_t = \begin{bmatrix} z_t \\ \phi_2 z_{t-1} + \dots + \phi_p z_{t-m+1} \\ \phi_3 z_{t-1} + \dots + \phi_p z_{t-m+2} \\ \vdots \\ \phi_m z_{t-1} \end{bmatrix}$$

and error term vector e has the form of

$$e = \begin{bmatrix} e_t \\ e_{t-1} \\ \vdots \\ e_{t-m} \end{bmatrix}$$

For an estimation via Kalman filter, it is important to come up with an initial probabilistic distribution (i.e. at time $t = 1$) of the state variable. Thus, α_1 is the initial value for the state variable being assumed to follow a normal distribution with mean zero and variance P. Matrix G controls for the effect of DR_t being computed as 3-month aggregate statistics where each shock appears in 3 consecutive periods.

This state space representation described in Equations 4.7-4.8 of the system described in Equations 4.5-4.6 may not seem parsimonious and transparent. As noted in Gilbert [1993], it should not be a problem because the relationship of the original representation and its state space representation created for the purpose of estimation via Kalman filter has a character of mathematical equivalence.

1_{mxn} refers to a matrix of ones mxn (m rows and $n$columns).

4.3. Model Estimation

The tool that allows us to deal with the state-space representation described in Section 4.2 is the Kalman filter (see Kalman [1960], Kim and Nelson [2003], Hamilton [1994]), a recursive procedure for computing the estimator of an unobserved variable at time t, based on available information at time t. This approach will be used to estimate the latent factor z_i in Equation 4.5. More precisely, it will estimate the mean and the variance of vector α_t at time t conditional on the up to time t.

First, we estimate the model described in Equation 4.5 with maximum-likelihood estimation for various p (i.e. degrees of autoregressive process that the latent factor follows). We use the estimates for $\phi_1, \phi_2, \ldots, \phi_p$ to construct the matrix T, which is needed before we start applying the Kalman filter. We use the order p of the autoregressive process, which guarantees the lowest Bayesian information criterion (BIC).

Subsequently, we start applying Kalman filter step by step. Each step consists of the updating part (represented by Equations 4.9-4.10) and the prediction part (represented by Equations 4.11-4.12).

Updating: Once $\hat{\varepsilon}_t^{GLS}$ is realized at the end of time t, the error in the prediction can be calculated as $v_t = \hat{\varepsilon}_t^{GLS} - E\left[z_t|\hat{\varepsilon}_t^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right]$. This error in the prediction contains new information about the latent component z_t beyond informations being available previously. Thus, after observing $\hat{\varepsilon}_t^{GLS}$, a more accurate inference can be made about z_t.

Prediction: At the beginning of time t, we may want to derive an optimal predictor of latent component z_t and thereby also an optimal predictor of $\hat{\varepsilon}_t^{GLS}$ based on all the available information up to time $t-1$.

Let $a_{t|t-1} = E\left[\alpha_t|\hat{\varepsilon}_{t-1}^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right]$ denote the conditional mean of α_t based on information available at time $t-1$ and let $P_{t|t-1} = var\left(\alpha_t|\hat{\varepsilon}_{t-1}^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right)$ denote the conditional variance of α_t.

The updating equations of the Kalman filter compute $a_{t/t} = E\left[\alpha_t|\hat{\varepsilon}_t^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right]$
and $P_{t|t} = var\left(\alpha_t|\hat{\varepsilon}_t^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right)$ using

$$a_{t/t} = a_{t|t-1} + K_t v_t \tag{4.9}$$
$$P_{t|t} = P_{t|t-1} - P_{t|t-1} Z^T K_t^T \tag{4.10}$$

where

$$\begin{aligned} v_t &= \hat{\varepsilon}_t^{GLS} - Z a_{t|t-1} = \hat{\varepsilon}_t^{GLS} - E\left[z_t|\hat{\varepsilon}_t^{GLS}, ..., \hat{\varepsilon}_1^{GLS}\right] \\ F_t &= Z P_{t|t-1} Z^T + \sigma_e^2 G G^T \\ K_t &= P_{t|t-1} Z^T F_t^{-1} \end{aligned}$$

The variable v_t is the error in prediction, $F_t = var(v_t)$ and K_t is the Kalman gain matrix, which determines the weight assigned to new information about the state vector α_t contained in the prediction

error v_t. The updating equation $a_{t/t} = a_{t|t-1} + K_t v_t$ suggests that $a_{t/t}$ is computed as a weighted average of $a_{t|t-1}$ and new information contained in the error of prediction v_t. [5].

The prediction equations equations of the Kalman filter compute a_{t+1} and P_{t+1} using

$$a_{t+1|t} = Ta_{t|t} \tag{4.11}$$
$$P_{t+1|t} = TP_{t|t}T^T + \sigma_\xi^2 HH^T \tag{4.12}$$

Out-of-sample forecasting of α_t can be computed from the updating and prediction equations by setting $v_t = 0$, $F_t^{-1} = 0$ and $K_t = 0$.

An important issue related to functioning of the Kalman filter is its innitialization. The above mentioned recursive procedure implemented by repeated updating and prediction relies on having estimates for the state vector subject to informations from the previous period (i.e $a_{t|t-1}$) and its variance (i.e. $P_{t|t-1}$). At the initial period, we assume $a_{1|0} = 0$ and $P_{1|0}$ being equal to the unconditional mean of $P_{t|t-1}$[6].

4.4. Results of the Latent Factor Extension

In this section we first present evidence of the relevance of the latent factor in our datasets. For this purpose we first estimate the linear model from Section 3.5 for each sector of corporate clients with a varying number of explanatory variables. The explanatory variables are chosen by applying the forward stepwise selection described in Section 3.3. By applying the forward stepwise procedure, we get one model for each number of explanatory variables ranging from 1 to 15. For each model, we calculate the p-value of the Ljung-Box[7] test, which refers to significance of the latent component.

The plot of p-values of the Ljung-Box test against number of explanatory variables in the regression model is included in Figure 4.1 in this section and in Section A.12 in the Appendix. In order to infer, whether an inclusion of a latent factor following an autoregressive process can improve the model, we also plot a horizontal line representing the critical value for the hypothesis of autocorrelation to be rejected on 1% level of significance. As can be seen in the plot, p-values increase with the increasing number of explanatory variables included in the model and thus the relevance of the latent factor decreases with increasing number of explanatory variables. This is in line with the hypothesis that the default cycle is partly caused by an omission of relevant explanatory variables. Nonetheless, even if all 15 available explanatory variables are included, the p-value of the Ljung-Box test is still far from not-rejecting the zero hypothesis of no autocorrelation in residuals.

In addition, Figure 4.1 in this section and Section A.12 in the Appendix show a various relation between number of explanatory variables and significance of the latent component for different sectors

[5] Kalman gain K_t is an optimal matrix that minimizes the mean-square error of $E\left[||\alpha_t - a_{t|t}||^2\right]$. We can notice it is an inverse function of F_t^{-1} (i.e. the variance of the errors in prediction) and positive function of $P_{t|t-1}$ (i.e. the uncertainty underlying α_t). However, it is beyond the scope of this thesis to derive it.

[6] The unconditional mean of $P_{t|t-1} = P$ can be obteined by solving for P from the following equation:

$$P = TPT^T + \sigma_\xi^2 HH^T$$

The solution is derived for example in Kim and Nelson [2003].

[7] Ljung-Box (LB) test verifies the zero hypothesis H_0 of no autocorrelation against the alternative H_1 of autocorrelation being present. See Section A.5 in the Appendix for further details about the LB test.

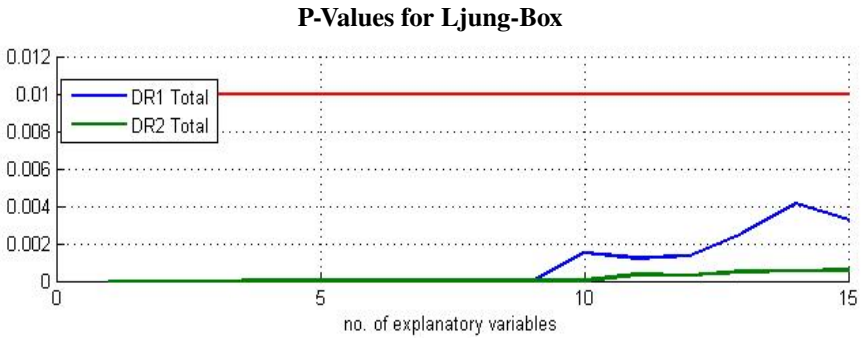


Figure 4.1.: **P-Values of Ljung-Box (LB) Test for Different Number of Explanatory Variables in the Regression Model**
Source: Author's Calculation

of corporate clients. This might be due to special character of particular sectors where very specific explanatory variables should be applied. Omission of such variables can result in autocorrelation of residuals and seeming existence of a default cycle. However, the significance of the latent factor for all analyzed sectors is so apparent that it can't be explained solely by the omission of explanatory variables. This is due to the fact that there is some degree of correlation between almost all possible explanatory variables and at least some share of information included in omitted explanatory variables must be captured in the 15 variables that are available for our analysis.

Thus, there must be an additional reason for existence of the default cycle in our dataset. Unfortunately, we cannot infer about the two theories introduced earlier in this chapter that the default cycle is caused by the leverage cycle (see Fostel and Geanakoplos [2008]) or alternatively by too lenient credit standards in times of economic prosperity (see Ruckes [2004]) because we miss sufficient data to provide such analysis. Yet, we think the theory presented in Ruckes [2004] that too lenient credit standards during an economic upturn results in an accumulation of credit risk, which then materializes in times of economic downturn, does not apply to prudential banking environment in the Czech Republic. We will therefore assume that the rest of the latent variable development is caused by serial autocorrelation in defaults (see Giesecke [2004], Eisenberg and Noe [2001]).

The following table summarizes the performance of the enriched models with latent component estimated by Kalman filter against the optimal models without latent component constructed earlier in Section 3.5. For $DR_{1,Total}$ and $DR_{2,Total}$, the optimal order of autoregressive process determining the latent component were set as 1 and 2 respectively. We can notice that inclusion of the latent component increased the coefficient of determination (R^2) significantly in case of $DR_{1,Total}$ but only slightly in case of $DR_{2,Total}$.

Comparison of the Models with Latent Component and Without Latent Component

	$DR_{1,Total}$		$DR_{2,Total}$	
no. of variables	7		6	
model	**without latent component**	**latent component**	**without latent component**	**latent component**
order of AR	-	1	-	2
R^2	0.8122	0.9404	0.7252	0.7853
LB(p-value)	0.0003	0.0082	0.0002	0.0038

Source: Author's Calculation

The same results for different sectors of corporate clients are presented in Section A.13 in the Appendix. We can notice that the different sectors demonstrate different improvement after adding the latent component. Some segments, such as $DR_{2,Sal}$ (sale), $DR_{2,Fin}$ (financial) and $DR_{2,HH}$ (household) demonstrate a significant increase in coefficient of determination (R^2). In case of the rest of the segments, inclusion of latent component doesn't increase the fit significantly. In addition, Ljung-Box test still indicates autocorrelation in case of most segments, which is however a direct result of the procedure how Kalman filter estimates the latent component via updating the old estimate with new information and the fact that Kalman filter estimates the latent component at time t using informations only up to time t. The degree of autocorrelation preserved in the data after applying the filter would be probably lower if we used Kalman smoothing[8] instead of ordinary Kalman filter.

Although we don't provide results of the model with latent component model for different numbers of explanatory variablesl, there should probably be some dependance. In case of a more parsimonious model, there is clearly more space for an improvement through adding the unobserved factor. This also explaines why some of the related studies with limited numbers of candidate explanatory variables (see. Jimenez and Mencia [2007], Gonzalez-Aguado and Bruche [2006]) find strong support for an inclusion of latent factor in an aggregate credit risk model.

Better understanding of the degree of improvement can be gained by analyzing Figure 4.2. The estimate for latent component follows a longer period in case of $DR_{1,Total}$. In contrast, the estimate for latent component is influenced by frequent shocks in case of $DR_{2,Total}$.

The same results for different segments of corporate clients are plotted in Section A.11 of the Appendix. Some segments, such as $DR_{2,Sal}$ (sale), $DR_{2,Fin}$ (financial) and $DR_{2,HH}$ (household), express desirable evolution of the latent factor with a cyclical period lasting several years . In case of other segments, there is also an apparent short cycle with the period of one year. Although the endogenous variable was seasonally adjusted by the TRAMO/SEATS procedure before we applied the linear regression model, there is still probably some degree of seasonality, which impairs the ability of Kalman filter to find the latent component and causes the Ljung-Box test to diagnose autocorrelation. Although the discovery of a seasonal default cycle is itself interesting, it is not the result we intended to find. The seasonality problem could be certainly remedied by numerous methods, which would filter out

[8] Kalman smoothing uses the information for the entire period for estimation of the latent component at time t. It would therefore guarantee a higher fit and lower autocorrelation in residuals due to its improved ability to distinguish between the different nature of shocks (i.e. between ξ_t and e_t in Equations 4.5 and 4.6).

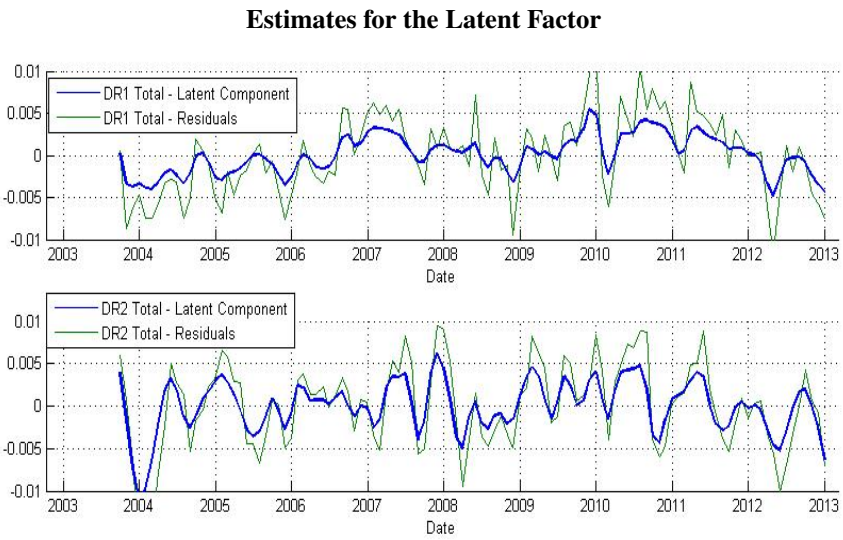


Figure 4.2.: **Estimates for the Latent Factor in $DR_{1,Total}$ and $DR_{2,Total}$**
Source: Author's Calculation

from the residuals the cycle with a period of one year corresponding to seasonality. Then, we would probably find indication of cycles with longer periods, which could be interpreted as the default cycle. However, it would require further research seeking for the source and nature of the seasonality.

Chapter 5
Conclusion

This thesis focuses on the determinants of aggregate credit risk in the Czech Republic. On one hand, we find the influence of various macroeconomic variables for different corporate sectors by constructing a linear regression model. One the other hand, we explicitly model a latent factor that influences the level of aggregate credit risk beyond macroeconomic environment and analyze how addition of this component to the linear regression model improves its fit.

First, we find support for the hypothesis that the effect of business cycle on aggregate level of credit risk is nonlinear. Whereas the effect of business cycle stance on the level of aggregate credit risk is generally negative (i.e. deteriorations of business cycle causing an increase in aggregate credit risk), negative fluctuations seem to have significantly higher impact than positive fluctuations. Results are stable for all analyzed segments with various significance of non-linearity for different segments.

Secondly, we construct a linear regression model consisting of multiple explanatory variables. We find several variables that drive aggregate credit risk simultaneously in multiple sectors of corporate clients. These variables include growth of industrial production, return on stock market index, treasury yield and exchange rate vis-a-vis EURO. Most of the selected variables show the anticipated sign of the effect on aggregate credit risk, but there are a few exceptions not following general intuition about the explected sign of their effect. For different analyzed segments, the forward stepwise selection procedure selects different sizes of the model with different model performance. The performance of the model would probably increase if additional sector-specific candidate explanatory variables were available.

Thirdly, we add a latent component to the linear models constructed earlier in this thesis. We conclude that significance of the latent factor depends on the number and suitability of explanatory variables that are selected into the model specification. For a parsimonious model with only a few explanatory variables, there is clearly more space for an improvement through adding the latent factor. Therefore in case of insufficient number of appropriate candidate explanatory variables being available, it seems to be reasonable to add a latent factor into the model. In additon, we explicitly model the evolution of the latent factor for multiple corporate segments and highlight differences among different sectors. For some sectors, we conclude that despite seasonal adjustment there is still some seasonal patern in the data, which is apparent in latent component estimation.

Bibliography

Ben S. Bernanke, Mark Gertler, and Simon Gilchrist. The financial accelerator in a quantitative business cycle framework. Handbook of macroeconomics, Elsevier, 1999.

Michael Boss, Martin Fenz, Johannes Pann, Claus Puhr, Martin Schneider, and Eva Ubl. Modeling credit risk through the austrian business cycle: An update of the OeNB model. *Financial Stability Report*, (17), 2009.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1979.

Max Bruche and Carlos Gonzalez-Aguado. Recovery rates, default probabilities and the credit cycle. SSRN Scholarly Paper ID 934348, Social Science Research Network, Rochester, NY, 2008.

Tomas Cipra. *Financni ekonometrie*. Ekopress, Praha, 2008.

Fabien Couderc, Olivier Renault, F. Couderc, and O. Renault. Times-to-default: Life cycle, global and industry cycle impacts. In *FAME Research Paper No.142*, 2004.

Robert B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74(1), March 1987.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, June 2001. ISBN 9780198523543.

Larry Eisenberg and Thomas H. Noe. Systemic risk in financial systems. *Management Science*, 47 (2):236–249, 2001.

Ana Fostel and John Geanakoplos. Leverage cycles and the anxious economy. *American Economic Review*, 98(4):1211–1244, 2008.

Philip Hans Franses and Dick van Dijk. *Non-Linear Time Series Models in Empirical Finance*. Cambridge University Press, 2000.

Benjamin M. Friedman. Money, credit, and interest rates in the business cycle. NBER chapters, National Bureau of Economic Research, Inc, 1986.

John Geanakoplos. The leverage cycle. SSRN Scholarly Paper ID 1441943, Social Science Research Network, Rochester, NY, 2009.

Kay Giesecke. Correlated default with incomplete information. *Journal of Banking & Finance*, 28 (7):1521–1545, 2004.

Paul Douglas Gilbert. *State Space and ARMA Models: An Overview of the Equivalence*. Bank of Canada, 1993.

Victor Gomez and Agustin Maravall. Seasonal adjustment and signal extraction in economic time series. Banco de Espana Working Paper 9809, Banco de Espanaa, 1998.

Carlos Gonzalez-Aguado and Max Bruche. Recovery rates, default probabilities and the credit cycle. FMG Discussion Paper dp572, Financial Markets Group, 2006.

Brenda Gonzalez-Hermosillo. *Investors Risk Appetite and Global Financial Market Conditions*. International Monetary Fund, 2008.

C. W. J. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of Econometrics*, 2 (2):111–120, 1974.

William H. Greene. *Econometric analysis*. Prentice Hall, Boston, 7th ed edition, 2012.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 1989.

James D. Hamilton. *Time series analysis*. Princeton University Press, Princeton, N.J, 1994.

A. C Harvey. *Time series models*. MIT Press, Cambridge, Mass., 1993.

Trevor Hastie, Robert Tibshirani, and J. H Friedman. *The elements of statistical learning data mining, inference, and prediction*. Springer, New York, 2009.

Robert J. Hodrick and Edward C. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 1970.

Petr Jakubik. Credit risk in the czech economy. Working Papers IES 2007/11, Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies, 2007.

Petr Jakubik and Christian Schmieder. Stress testing credit risk: Is the czech republic different from germany? Working Paper 2008/9, Czech National Bank, Research Department, 2008.

Carlos M. Jarque and Anil K. Bera. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2), 1987.

Gabriel Jimenez and Javier Mencia. Modeling the distribution of credit losses with observable and latent factors. Banco de Espana Working Paper 0709, Banco de Espana, 2007.

Harvir Kalirai and Martin Scheicher. Macroeconomic stress testing: Preliminary evidence for austria,. *Oesterreichische Nationalbank, Financial Stability Report 3*, 2002.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, 82(Series D):35–45, 1960.

Anil Kashyap and Jeremy C. Stein. Cyclical implications of the basel II capital standards. *Economic Perspectives*, 2004.

Stefan Kerbl and Michael Sigmund. What drives aggregate credit risk? *Financial Stability Report*, (22), 2011.

Hinh D. Khieu, Donald J. Mullineaux, and Ha-Chin Yi. The determinants of bank loan recovery rates. *Journal of Banking and Finance*, 36(4):923–933, 2012.

Chang-Jin Kim and Charles R Nelson. *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. the MIT Press, 2003.

Hyune-Ju Kim, Michael P Fay, Binbing Yu, Michael J Barrett, and Eric J Feuer. Comparability of segmented line regression models. *Biometrics*, 60(4), 2004.

Nobuhiro Kiyotaki and John Moore. Credit cycles. *Journal of Political Economy*, 105(2), 1997.

Nir Klein. *Non-Performing Loans in CESEE: Determinants and Impact on Macroeconomic Performance*. International Monetary Fund, 2013.

Siem Jan Koopman, Andre Lucas, and Bernd Schwaab. *Forecasting Cross-sections of Frailty-correlated Default*. Tinbergen Institute, 2008.

Siem Jan Koopman, Roman Kraussl, Andre Lucas, and Andre Monteiro. Credit cycles and macro fundamentals. *Journal of Empirical Finance*, 16(1):42–54, 2009.

P. M. Lerman. Fitting segmented regression models by grid search. *Applied Statistics*, 29(1), 1980.

Andre Lucas and Siem Jan Koopman. Business and default cycles for credit risk. *Journal of Applied Econometrics*, 20(2), 2005.

Agustin Maravall. Unobserved components in economic time series. Banco de Espana Working Paper 9609, Banco de Espana, 1996.

Juri Marcucci and Mario Quagliariello. Asymmetric effects of the business cycle on bank credit risk. *Journal of Banking and Finance*, 2009.

A. I. McLeod and W. K. Li. Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, 4(4), 1983.

Robert C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449, 1974.

Jianjun Miao and Pengfei Wang. *Credit Risk and Business Cycles*. 2010.

Armando Morales and Jose Giancarlo Gasha. Identifying threshold effects in credit risk stress testing. IMF Working Paper 04/150, International Monetary Fund, 2004.

Vito M. R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19), 2003.

Pamela Nickell, William Perraudin, and Simone Varotto. Stability of ratings transitions. Bank of England working paper 133, Bank of England, 2001.

Kazuhiro Ohtani and Hikaru Hasegawa. On small sample properties of r2 in a linear regression model with multivariate t errors and proxy variables. *Econometric Theory*, 9(03), 1993.

Jyh-Ying Peng and John A. D. Aston. The state space models toolbox for MATLAB. *Journal of Statistical Software*, Vol.41(No.6):1–26, 2011.

Peter C. B. Phillips. Understanding spurious regressions in econometrics. Cowles Foundation Discussion Paper 757, Cowles Foundation for Research in Economics, Yale University, 1985.

Morten O. Ravn and Harald Uhlig. On adjusting the HP-Filter for the frequency of observations. CESifo Working Paper Series 479, CESifo Group Munich, 2001.

Rafael Repullo and Javier Suarez. The procyclical effects of bank capital regulation. Working Paper wp2012_1202, CEMFI, 2012.

Daniel Rosch. Correlations and business cycles of credit risk: Evidence from bankruptcies in germany. University of Regensburg Working Papers in Business, Economics and Management Information Systems 380, University of Regensburg, Department of Economics, 2003.

Martin E. Ruckes. Bank competition and credit standards. SSRN Scholarly Paper ID 903955, Social Science Research Network, Rochester, NY, 2004.

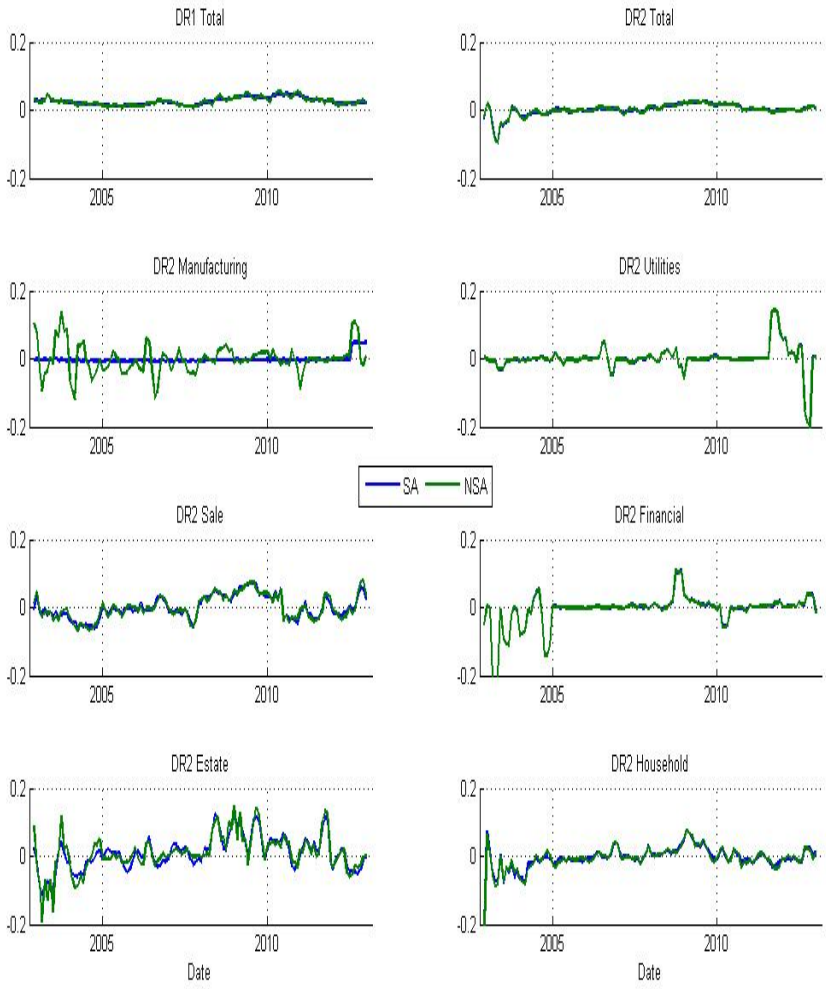
Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 1994.

Lloyd N. Trefethen. *Numerical linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997.

Juraj Zeman and Pavol Jurca. Macro stress testing of the slovak banking sector. Working and Discussion Paper WP 1/2008, Research Department, National Bank of Slovakia, 2008.

Appendix A
Accompanying Remarks

A.1. Seasonal Adjustment of Variables to Explain



SA = seasonally adjusted

NSA = not seasonally adjusted

Source: Czech National Bank + Author's Calculation

A.2. Descriptive Statistics of Variables to Explain for Different Segments of Corporate Clients in the Czech Republic

	Start	End	Min	Max	Mean	Median	St.dev
$DR_{2,Man}$	9/2003	12/2012	-0.0120	0.0479	-0.0018	-0.0042	0.0118
$DR_{2,El}$	9/2003	12/2012	-0.2000	0.1405	0.0013	0.000	0.0412
$DR_{2,Sal}$	9/2003	12/2012	-0.0616	0.0734	0.000	-0.0058	0.0342
$DR_{2,Fin}$	9/2003	12/2012	-0.1427	0.1081	0.000	0.0019	0.0343
$DR_{2,Est}$	9/2003	12/2012	-0.0597	0.1240	0.0154	0.0128	0.0444
$DR_{2,HH}$	9/2003	12/2012	-0.0785	0.0771	-0.0018	-0.0042	0.0238

Source: Czech National Bank + Author's Calculation

A.3. Gauss-Markov Theorem

For a linear model $DR_t = X_t\beta + \varepsilon_t = \beta_0 + \beta_1 X_{1,t} + ... + \beta_k X_{k,t} + \varepsilon_t$,

$\hat{\beta}^{OLS} = (X_t^T X_t)^{-1} X_t^T DR_t$ is a minimal-variance linear unbiased estimator of β if it satisfies following conditions:

1. E[$\varepsilon_t|X$]=0. The expected value of ε_t conditional on values explanatory variable X is zero, i.e error terms in any given period are not correlated with explanatory variables in all periods. It implies that explanatory variables are fully exogenous.
2. Explanatory variables X and endogenous variable DR_t are randomly sampled from their distributions.
3. No perfect collinearity of explanatory variables X, i.e. any explanatory variable can't be a perfect linear combination of the others.
4. Homoscedasticity. Conditional on X, the variance of ε_t is the same for all t: Var($\varepsilon_t|x$)=Var(ε_t)=$\sigma^2 < \infty$
5. No autocorrelation of residuals. Conditional on X, the errors in two different time periods are uncorrelated: Corr($\varepsilon_t, \varepsilon_{t-s}$)=0 for t$\neq$s

(1),(2),(3) are sufficient conditions for $\hat{\beta}^{OLS}$ to be unbiased. (4),(5) are additional sufficient conditions for $\hat{\beta}^{OLS}$ to attain the lower bound for the variance of $\hat{\beta}^{OLS}$. If in addition, error terms ε_t are normaly distributed, $\hat{\beta}^{OLS}$ is also normally distributed, i.e. $\varepsilon_t \approx N(0,\sigma^2) \Rightarrow \hat{\beta}^{OLS} \approx N(\beta, \sigma^2(X^T X)^{-1})$

A.4. Unbiasedness and Inefficiency of $\hat{\beta}^{OLS}$ under Auto-Correlation in Error Terms

The unbiasedness of $\hat{\beta}^{OLS}$ is proven by the following procedure

$$\begin{aligned} E[\hat{\beta}^{OLS}] &= E[(X_t^T X_t)^{-1} X_t^T DR_t] = E[(X_t^T X_t)^{-1} X_t^T (X_t\beta + \varepsilon_t)] = \\ &= \beta + (X_t^T X_t)^{-1} X_t^T E[\varepsilon_t] = \beta \end{aligned}$$

i.e $\hat{\beta}^{OLS}$ is under the assumption of error terms ε_t being uncorrelated with explanatory variables X (ass. 1) unbiased regardless of serial autocorrelation of ε_t. To show variance of $\hat{\beta}^{OLS}$, we have to define the variance covariance matrix of error terms defined as $E[\varepsilon_t \varepsilon_t^T]$. It is easy to show that under the assumption of ε_t being homoscedastic and following an AR(1) process, the theoretical variance covariance matrix of error terms has a form of

$$E[\varepsilon_t \varepsilon_t^T] = \sigma_\varepsilon^2 \Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{bmatrix} \neq . \sigma_\varepsilon^2 \mathrm{I}$$

The theoretical variance of $\hat{\beta}^{OLS}$ is then derived as follows:

$$\begin{aligned} Var(\hat{\beta}^{OLS}) &= E[(\hat{\beta}^{OLS} - \beta)\hat{\beta}^{OLS} - \beta)^T] = (X^T X)^{-1} X^T E[\varepsilon_t \varepsilon_t^T] X (X^T X)^{-1} = \\ &= \sigma_\varepsilon^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1} = \sigma_\varepsilon^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1} \neq \\ &\neq \sigma_\varepsilon^2 (X^T X)^{-1} \end{aligned}$$

i.e $\hat{\beta}^{OLS}$ is inefficient and t-statistics for significance of individual variables in the model are invalid.

A.5. Methodics of Residual Testing

A.5.1. Breusch-Pagan Test for Homoscedasticity

Suppose we estimate the following regression model

$$DR_t = X_t\beta + \varepsilon_t = \beta_1 + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \varepsilon_t$$

and estimate a vector of residuals $\hat{\varepsilon}_t$

The Breusch-Pagan test checks out a null hypothesis H_0 of homoscedasticity (i.e. that the variance of residuals from a linear model does not depend on the values of explanatory variables). Formally,

$$\varepsilon_t^2 = X_t\gamma + \varepsilon_t = \gamma_1 + \gamma_2 X_{2,t} + ... + \gamma_k X_{k,t} + u_t$$

$$\begin{aligned} H_0: &\quad \gamma_1 = 0, \gamma_2 = 0, ..., \gamma_k = 0 &\quad Homoscedasticity \\ H_1: &\quad \gamma_1 \neq 0, or\, \gamma_2 \neq 0, or ..., or\, \gamma_k \neq 0 &\quad Heteroscedasticity \end{aligned}$$

Breusch and Pagan [1979] define a test statistics BP, which is asymptotically χ_k^2 distributed (chi-square distribution with k degrees of freedom). For $BP > \chi_{k,95\%}^2$, the null hypothesis H_0 of homoscedasticity is is rejected on 5% level of significance.

A.5.2. Ljung-Box Test for Autocorrelation in Residuals

The Ljung-Box test of a null hypothesis H_0 that a vector of residuals ε_t exhibits no autocorrelation for first L lags (i.e. $H_0 : \rho(1) = 0, \rho(2) = 0, ..., \rho(L) = 0$), against an alternative that some of the autocorrelation terms[1] $\rho(i)$ for $i = 1, 2, ..., L$ is not zero.

$$\begin{aligned} H_0: &\quad \rho(1) = 0, \rho(2) = 0, ..., \rho(L) = 0 &\quad no\, autocorrelation \\ H_1: &\quad \rho(1) \neq 0, or\, \rho(2) \neq 0, or ..., or\, \rho(L) \neq 0 &\quad autocorrelation \end{aligned}$$

$$Q = n(n+2)\sum_{i=1}^{L}\left(\frac{\rho^2(k)}{n-i}\right) \approx \chi_L^2$$

where n is the sample size, L is the number of autocorreation lags that equal to zero according to the null hypothesis H_0 and $\rho(i)$ is the sample autocorrelation at lag i. If the null hypothesis H_0 holds, the asymptotic distribution of Q is χ_L^2 (chi-square distribution with L degrees of freedom). See McLeod and Li [1983] for further details about the Ljung-Box test.

For $Q > \chi_{L,95\%}^2$, the null hypothesis H_0 of no autocorrelation for L first lags is rejected on 5% level of significance.

A.5.3. Jarque-Bera Test for Normality of Residuals

The Jarque-Bera test of a null hypothesis H_0 that a vector of residuals comes from a normal distribution with zero mean[2] and unknown variance, against an alternative that it does not come from a normal distribution..It is a two-sided goodness of fit test suitable when a fully specified null distribution (i.e.

[1] The k-th autocorrelation term $\rho(i)$ is computed as

$$\rho(i) = \frac{\frac{1}{n-i}\sum_{i=1}^{n-i}\varepsilon_t\varepsilon_{t+i}}{\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2}$$

. Note that the mean mean of residuals is automatically quaranteed to be zero by the least squares estimation.

[2] The zero mean of residuals is automatically quaranteed by the least squares estimation.

the distribution under the assumption that H_0 holds) is not known and its parameters must be estimated. The test statistic is

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right) \approx \chi^2_2$$

$$\begin{aligned} H_0: &\quad normal\,distribution &\quad (s = 0, k = 3) \\ H_1: &\quad other\,distribution &\quad (s \neq 0, or\, k \neq 3) \end{aligned}$$

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. For large samples, the test statistic has asymptotically a χ^2_2 distribution (chi-square distribution with two degrees of freedom). See Jarque and Bera [1987] for further details about the Jarque-Bera test. The construction of the JB statistics relies on the fact that the theoretical skewness is 0 and the theoretical curtosis is 3.

For $JB > \chi^2_{2,95\%}$, the null hypothesis H_0 of residuals being sampled from a normal distribution with zero mean and unknown variance is rejected on 5% level of significance.

A.6. Model with Business Cycle Measure as the Explanatory Variable Results for the Baseline Model

The following table presents results for the model

$$DR_t \quad = \beta_1 + \beta_2 IP_{GAP,t} + \varepsilon_t \quad (1)$$

where DR_t is 3-month ex-ante change in non-performing loans ($DR_{2,t}$) for 6 sectors of corporate clients with the biggest share on banks' portfolios, $IP_{GAP,t}$ is proportional deviation from HP-filtered industial production, $(IP_{GAP,t} - \psi)^- = min((IP_{GAP,t} - \psi, 0)$ is truncated form of $IP_{GAP,t}$ allowing for the asymmetric effect of business cycle, ψ is the threshold value and ε_t is an error term.

	$DR_{2,Man}$	$DR_{2,El}$	$DR_{2,Sal}$	$DR_{2,Fin}$	$DR_{2,Est}$	$DR_{2,HH}$
β_1	-0.0018	0.0013	0.0008	0.0002	0.0154	-0.0018
	(-1.62,*)	(0.34,)	(0.26,*)	(0.06,)	(3.68,***)	(-0.81,)
β_2	-0.0016	0.0067	-0.0130	-0.0013	-0.0076	-0.0080
	(-1.45,)	(1.72,*)	(-4.29)	(-0.40,)	(-1.80,*)	(-3.73,***)
R^2	0.0189	0.0265	0.1448	0.0015	0.0289	0.1134
F	0.1480	0.0864	0.0000	0.6867	0.0729	0.0003
LB	0.0000	0.0006	0.0000	0.3013	0.0000	0.0000
JB	0.0010	0.0010	0.1925	0.0010	0.0465	0.0033
BP	0.3123	0.2984	0.0928	0.2783	0.1432	0.1693

Figures in parenthesis are t-statistics for significance of individual variables

* indicates significance on 10% level

** indicates significance on 5% level

*** indicates significance on 1% level

F: p-value for the F-test

LB: p-value for Ljung-Box test

JB: p-value for Jarque-Bera test

BP: p-value for Breusch-Pagan test

Source: Author's Calculation

A.7. Model with Business Cycle Measure as the Explanatory Variable Results for the Model with Censored Variable

The following table presents results for the model

$$DR_t \quad = \beta_1 + \beta_2 IP_{GAP,t} + \beta_3 (IP_{GAP,t} - \psi)^- + \varepsilon_t \quad (2)$$

where DR_t is 3-month ex-ante change in non-performing loans ($DR_{2,t}$) for 6 sectors of corporate clients with the biggest share on banks' portfolios, $IP_{GAP,t}$ is proportional deviation from HP-filtered industial production and ε_t is an error term.

	$DR_{2,Man}$	$DR_{2,El}$	$DR_{2,Sal}$	$DR_{2,Fin}$	$DR_{2,Est}$	$DR_{2,HH}$
β_1	-0.0012	-0.0012	-0.0407	-0.0093	-0.156	-0.0145
	(-0.19,)	(-0.13,)	(-5.08,***)	(-1.12,)	(-1.41,)	(-3.04,***)
β_2	-0.0031	0.0112	0.0315	0.0123	0.0302	0.0107
	(-3.51,***)	(1.70,*)	(4.82,)	(2.06,**)	(3.85,***)	(2.99,***)
β_3	0.0107	-0.0128	-0.0641	-0.0267	-0.0613	-0.0383
	(3.48,***)	(-0.90,)	(-6.83,***)	(-2.58,***)	(-4.96,***)	(-6.03,***)
ψ	-1.3183	-0.4914	0.4458	-0.0504	0.2252	-0.1055
N_1	99	84	35	60	43	63
N_2	13	28	77	52	69	49
R^2	0.0441	0.0353	0.4290	0.0639	0.2102	0.3785
F	0.0856	0.1411	0.0000	0.0274	0.0000	0.0000
LB	0.0000	0.0000	0.0004	0.0374	0.9409	0.0018
JB	0.0010	0.0010	0.3622	0.0010	0.0317	0.0012
BP	0.1323	0.4838	0.1389	0.2354	0.2037	0.1846

Figures in parenthesis are t-statistics for significance of individual variables

* indicates significance on 10% level

** indicates significance on 5% level

*** indicates significance on 1% level

F: p-value for the F-test

LB: p-value for Ljung-Box test

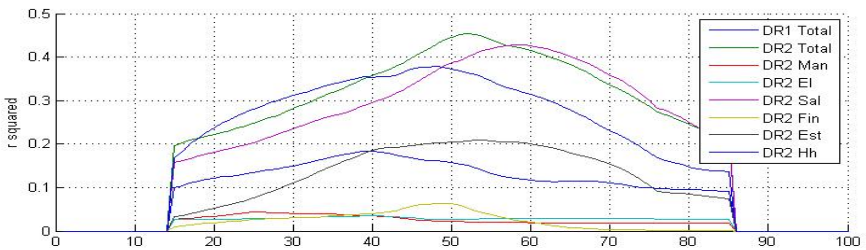
JB: p-value for Jarque-Bera test

BP: p-value for Breusch-Pagan test

Source: Author's Calculation

A.8. Grid-Search for the Model with Censored Variable

The following table depicts the coefficient of determination (R^2) for different sectors subject to the particular value of the threshold.



Source: Author's Calculation

A.9. Results for the Model with Multiple Explanatory Variables

The following table presents results for the model

$$\begin{aligned} DR_t &= \beta_1 + \beta_2 CA_t + \beta_3 CPI_t + \beta_4 CREDIT_t + \beta_5 ER_t + \beta_6 (GOV10 - GOV3)_t + \\ &+ \beta_7 GOV3_t + \beta_8 IP_{GAP_t} + \beta_9 IP_{GROWTH_t} + \beta_{10} M1_t + \beta_{11} PPI_t + \\ &+ \beta_{12} PRIBOR3M_t + \beta_{13} RET_{PX50_t} + \beta_{14} UNEMP_t + \beta_{15} VACANT_t + \\ &+ \beta_{16} VOL_{PX50_t} + \varepsilon_t \end{aligned}$$

where DR_t is the measure of aggregate credit risk ($DR_{1,t}$ or $DR_{2,t}$) for 2 aggregate categories and 6 sectors of corporate clients with the biggest share on banks' portfolios and $CA_t ... VOL_{PX50_t}$ are explanatory variables.

	$DR_{1,Total}$	$DR_{2,Total}$	$DR_{2,Man}$	$DR_{2,El}$	$DR_{2,Sal}$	$DR_{2,Fin}$	$DR_{2,Est}$	$DR_{2,HH}$
β_1	0.0253	0.0038	-0.0018	0.0013	0.0008	0,0002	0.0154	-0.0018
	(56.56,***)	(7.29,***)	(-2.71,***)	(0.39,)	(0.40,)	(0.08,)	(4.77,***)	(-1,13,)
β_2	-0.0034		-0.0054	0.0113			-0.0103	
	(-6.72,***)		(-6.88,***)	(2.13,**)			(-2.38,**)	
β_3				0.0164		-0.0073	-0.0191	-0.0073
				(2.38,**)		(-1.97,**)	(-3.42,***)	(-2.86,***)
β_4	-0.0072	-0.0080			-0.0150	0.0079		
	(-7.02,***)	(-5.81,***)			(-2.66,***)	(1.52,)		
β_5	-0.0060	-0.0037			-0.0084		-0.0134	
	(-10.25,***)	(-5.27,***)			(-2.83,***)		(-2.74,***)	
β_6	0.0017		-0.0061	0.0110				
	(1.99,**)		(-5.07,***)	(1.62,)				
β_7	0.0037		-0.0100	0.0337		-0.0189	0.0336	
	(5.28,***)		(-9.83,***)	(3.38,***)		(-4.95,***)	(3.68,***)	
β_8					-0.0033		0.0110	-0.0119
					(-0.73,)		(1.98,**)	(-3.74,***)
β_9		-0.0026	-0.0062	0.0117	-0.0120		-0.0178	0.0034
		(-3.08,***)	(-6.75,***)	(1.59,)	(-2.09,**)		(-2.46,**)	(0.98,)
β_{10}				-0.0126	-0.0039	0.0047	-0.0100	
				(-2.16,**)	(-1.36,)	(1.60,)	(-1.94,*)	
β_{11}	-0.0056	-0.0059		0.0026	-0.0145			
	(-10.00,***)	(-8.13,***)		(0.50,)	(-4.68,***)			
β_{12}				-0.0356	-0.0133		-0.0208	
				(-2.46,**)	(-2.74,***)		(-1.81,*)	
β_{13}	-0.0063	-0.0035	0.0030	0.0200	0.0062			-0.0072
	(-13.03,***)	(-4.78,***)	(3.30,***)	(-3.79,***)	(-2.03,**)			(-2.75,***)
β_{14}						-0.0181		-0.0167
						(-3.80,***)		(-5.99(***)
β_{15}		0.0076	0.0014	-0.0028	0.03			
		(5.10,***)	(0.91,)	(-0.26,)	(4.10,***)			
β_{16}						0.0111	0.0087	-0.0045
						(3.60,***)	(2.03,**)	(-2.06,**)
$no.of\,var$	7	6	6	10	9	6	9	6
R^2	0.81	0.72	0.66	0.30	0.65	0.46	46	0.55
$\bar{R}_k^2$	0.79	0.70	0.64	0.24	0.62	0.43	0.41	0.53
F	0	0	0	0	0	0	0	0
LB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
JB	0.500	0.006	0.001	0.001	0.500	0.001	0.070	0.09%
BP	0.137	0.039	0.372	0.143	0.103	0.365	0.384	0.463
κ	24.4	37.8	19.3	114.7	94.0	21.5	69.1	27.0

F: **p-value for the F-test**

LB: **p-value for Ljung-Box test**

JB: **p-value for Jarque-Bera test**

BP: **p-value for Breusch-Pagan test**

κ: **condition number as the indicator of multicollinearity among explanatory variables**

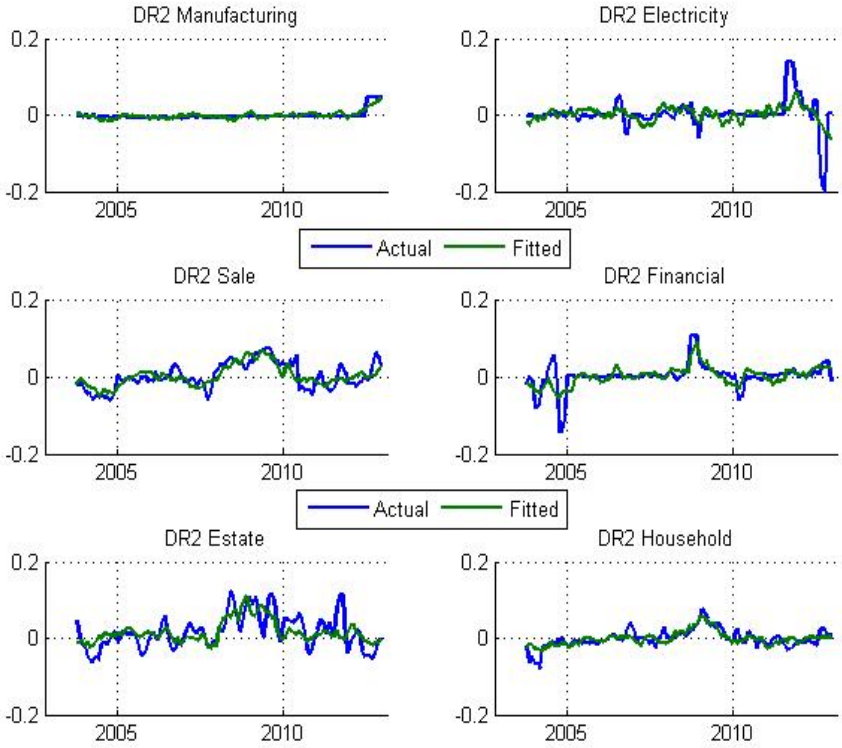
$no.of\,var$: **number of explanatory variables in the model**

number in parenthesis represent the t-statistics for the significance of individual variables

*/**/*** **indicate significance on 10%/5%/1% level**

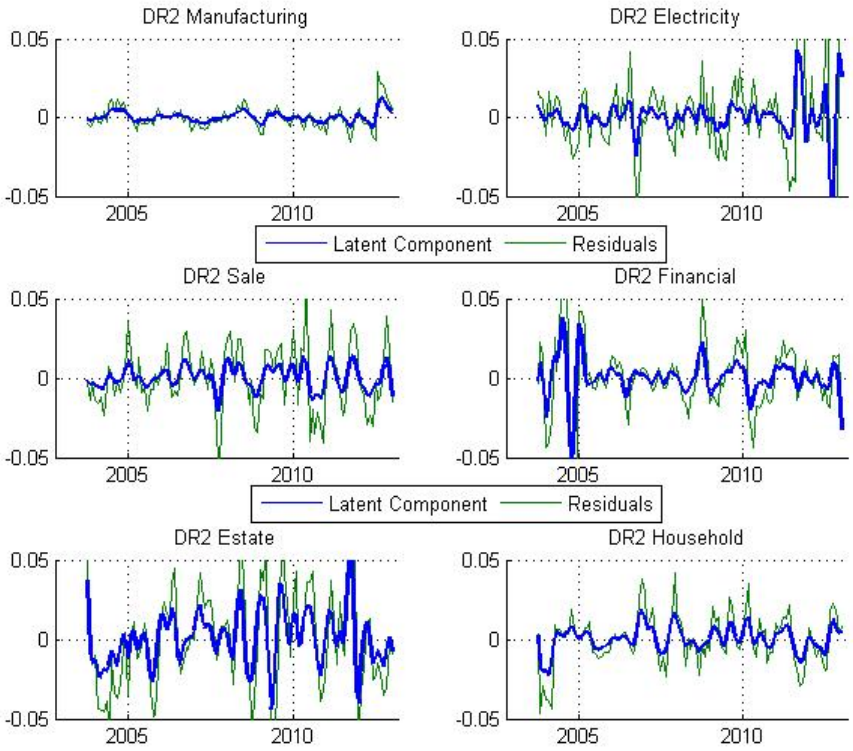
Source: Author's Calculation

A.10. Actual vs. Fitted Values of DR_2 for the Model with Multiple Explanatory Variables



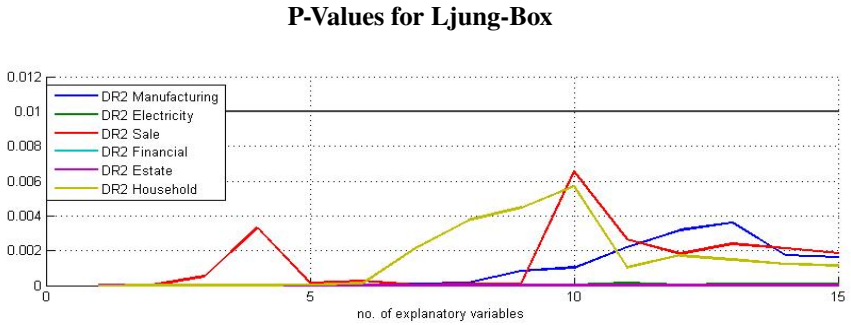
Source: Author's Calculation

A.11. Estimates of the Latent Component for 6 Sectors of Corporate Clients



Source: Author's Calculation

A.12. P-Values for Ljung-Box (LB) Test vs. Number of Explanatory Variables in the Regression Model



(black horizontal line depicts the 1% threshold for rejecting the significance of autoregressive effect)

Source: Author's Calculation

A.13. Performance of the Models with and without the Latent Component

	$DR_{2,Man}$		$DR_{2,El}$	
no. of variables	6		10	
model	**without latent component**	**latent component**	**without latent component**	**latent component**
order of AR	-	1	-	1
R^2	0.6610	0.7605	0.3092	0.5896
LB(p-value)	0.0001	0.0040	0.0000	0.0011

	$DR_{2,Sal}$		$DR_{2,Fin}$	
no. of variables	9		6	
model	**without latent component**	**latent component**	**without latent component**	**latent component**
order of AR	-	1	-	1
R^2	0.6595	0.8793	0.4625	0.7932
LB(p-value)	0.0008	0.0091	0.0000	0.0019

	$DR_{2,Est}$		$DR_{2,HH}$	
no. of variables	9		6	
model	**without latent component**	**latent component**	**without latent component**	**latent component**
order of AR	-	2	-	1
R^2	0.4600	0.7388	0.5578	0.8572
LB(p-value)	0.0001	0.0046	0.0002	0.0056

no. of variables: number of explanatory variables in the optimal regression model

order of AR: order of autoregressive process the latent component is expected to follow

LB (p-value): p-values for the Ljung-Box teest

Source: Author's Calculation

Appendix B
Content of Enclosed RAR folder

There is a RAR folder enclosed to this thesis containing empirical data and MATLAB source codes. The code uses the State Space Models Toolbox for MATLAB presented in Peng and Aston [2011], which is available at http://www.jstatsoft.org/v41/i06.

Open the README.txt file in enclosed RAR folder for hints about replicating the results presented in this thesis.

Master Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Author:	**Bc. Jan Málek**	Supervisor:	PhDr. Jakub Seidler
E-mail:	janmalek@centrum.cz	E-mail:	seidler@email.cz
Phone:	+420 732523687	Phone:	
Specialization:	*Finance, Financial Markets and Banking*	Defense Planned:	June 2013

Proposed Topic:

How Is the Credit Risk Influenced by the Business Cycle? - The Case of the Czech Republic

Topic Characteristics:

This thesis concentrates on an important factor of credit risk so called probability of default (PD) and it's empirical counterpart the share of non-performing loans (NPL) to total loans. The focus is namely given on it's relationship to the business cycle (denoted in percentage deviation from the potential value of GDP).

Firstly, we will concentrate on the pure relationship between the business cycle and credit risk. There is evidence that credit risk reacts to the business cycle non-linearly. We will make use of regime switching models (threshold modelling and markov-switching modelling) in order to capture this non-linearity.

Secondly, we will use the state space modelling, which assumes a latent factor behind the credit risk realization. The estimates of the latent factor can subsequently be put in relation with the cyclical position of the economy.

Thirdly, we will try to construct complex VAR models with several macroeconomic variables that will possibly be able to forecast the share of NPL. There are some variables that are expected to have substantial explanatory power (see the hypotheses)

By using abovementioned methods, the thesis will identify sensitivity of credit risk with respect to macroeconomic development, which is highly relevant for assessing the riskiness of the debtors.

Hypotheses:

1. Macroeconomic variables determine the aggregate default rate in the economy
   (we expect following variables to have explanatory power: GDP, nominal (real) interest rate, corporate indebtedness, GDP growth rate, real property prices, money growth)
2. For each sector of the economy, different variables are significant with different lags.
3. For each sector of the economy, the default rate is sensitive to the business cycle to different extend.
4. There is a substantial asymmetry in the effect of the business cycle on bank credit risk
5. The asymmetry (from 4) is different for each sector of the economy.

Methodology:

-Literature survey

-Empirical analysis on macroeconomic data from CNB ARAD Data Series System for non-performing loans (by Czech National Bank) and various macroeconomic variables using:

-Vector autoregressive models, estimating the credit risk parameters as a linear combination of past values of explanatory variables
-State space modelling, assuming a latent factor behind the credit risk realization
-Regime switching models:
a) threshold modelling, allowing for asymmetries in the realization of credit risk in different phases of the business cycle
b) markov-switching modelling assuming a latent factor influencing the credit risk realization

-as software, we will probably use R-Studio or TSM

Outline:

1. Introduction

2. Related literature
3. Modelling the assymetric effects of the business cycle on aggregate credit risk by regime-switching models
4. Modelling the aggregate credit risk by state-space models
5. Modelling the aggregate credit risk by vector-autoregression (VAR) models

6. Conclusion

Core Bibliography:

1. Altman E., Brady B., Resti A., Sironi A. (2003): "The Link between Default and Recovery Rates: Theory, Empirical Evidence and Implications", DefaultRisk.com, March 2003

2. Boss M., Fenz M., Pann J., Puhr C., Schneider M., Ubl E. (2009): "Modeling Credit Risk through the Austrian Business Cycle: An Update of the OeNB", Financial Stability Report, pp 85-101, Oesterreichische Nationalbank

3. Hansen M., Bruce E. (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis," Econometrica, Econometric Society, vol. 64(2), pages 413-30, March.

4. Jakubík P. (2006): "Credit Risk Models and Their Relationship to the Business Cycle ", Diploma thesis at the Institute of Economic Studies, Faculty of Social Sciences, Charles Univerity

5. Jakubik P., Schmieder Ch., (2008): "Stress Testing Credit Risk: Is the Czech Republic Different from Germany?", Working Papers 2008/9, Czech National Bank, Research Department.

6. Kerbl S., Sigmund M. (2011): "What Drives Aggregate Credit Risk?", Oesterreichische Nationalbank Financial Stability Report, Vol. 22, December 2011

7. Marcucci J., Quagliariello M. (2009): "Asymmetric effects of the business cycle on bank credit risk," Journal of Banking & Finance, Elsevier, vol. 33(9), pages 1624-1635, September.

8. Piger J. (2007): "Econometrics: Models of Regime Changes", prepared for: Springer Encyclopedia of Complexity and System Science.

9. Život E., Wang J. (2006): "Modeling financial time series with S-PLUS", Springer Publishing House, ISBN: 9780387279657

Author

Supervisor