Credit Risk and Regional Economic Disparities

Tomáš Vaněk [1] | *Mendel University in Brno, Czech Republic*
David Hampel | *Mendel University in Brno, Czech Republic*

Abstract

This paper aims to bridge the areas of credit risk and regional economic disparities, and investigates the relationship between credit risk and economic indicators in the Czech Republic at the regional (NUTS 3) level. This relationship is consecutively examined using graphical and correlation analysis, regression techniques, and different types of clustering methods. Regions are then clustered into three groups according to their economic similarities and disparities. Subsequently, it is shown on the real data that region-specific information has the potential to be utilizable in credit scoring and possibly other applications.

Keywords

Credit risk, regional disparities, cluster analysis, Bayesian information criterion, model selection, logistic regression

JEL code

R11, G32, C10, C25, C38

INTRODUCTION

Credit risk is one of the most fundamental and significant risks banks are exposed to, and is generally understood as the potential that a borrower or counterparty will fail to meet its contractual obligations (see BCBS, 2000). With the introduction of the Basel II capital requirements framework in 2004, attention paid to credit risk analysis and management has become even greater. Requirements in this area will be further augmented after January 1st, 2018, when the standard IFRS 9 *Financial Instruments*, introducing a new framework for credit impairment calculation, becomes effective.

Banks evaluate credit risk associated with potential and actual clients within credit scoring, which is a process for prediction of the probability that a client will default (Hand and Henley, 1997). For discussions on the historical context and development of credit scoring see Thomas (2000), or Abdou and Pointon (2011). As suggested above, credit risk evaluation (probability of default estimation) is important not only for internal credit decisions, but also for regulatory purposes – especially quantification of capital requirements within the internal ratings-based approach (see BCBS, 2004; and CRR, 2013) and calculation of credit impairment (loss allowances) within IFRS 9 (see IFRS Foundation, 2015).

Since the 1970s, a quantitative approach to credit scoring has been dominant, with statistical models playing a key role. Over time, logistic regression has become the standard and is usually used as a benchmark when estimating more sophisticated models (e.g. Crook et al., 2007). Li and Zhong (2012), or Lessmann et al. (2015) provide a good overview of the methods and models that have been used in credit scoring.

[1] Mendel University in Brno, Faculty of Business and Economics, Zemědělská 1, 613 00 Brno, Czech Republic. Corresponding author: e-mail: xvanek7@mendelu.cz, phone: (+420)721482751.

This paper aims to bridge credit risk and regional economic disparities representing a persisting development tendency in a majority of countries (e.g. Shankar and Shah, 2003), including the Czech Republic.

In the literature, there are a plenty of studies devoted to the analysis of credit risk in the macroeconomic context from various points of view. Pesaran et al. (2006) model conditional credit loss distributions, Pesaran et al. (2007) explore credit risk diversification, with both studies using 'global' macroeconometric models. In a recent study, Schwaab et al. (2016) also investigated credit risk from a global perspective, using a non-linear state-space model. Studies concerning credit risk and macroeconomy in the Czech Republic have been conducted by Jakubík and Heřmánek (2008), Jakubík (2008), Grešl et al. (2013) or Melecký et al. (2015), who deal with credit risk especially in the context of macroeconomic stress testing.

There are also many studies dealing with regional disparities from various perspectives. OECD (2016) can be considered as an up-to-date and comprehensive study with a global scope. In terms of more recent studies focusing on regional disparities within the Czech Republic, we can mention Kutscherauer et al. (2010), Kahoun (2010), Svatošová and Novotná (2012), Procházková and Radiměřský (2013), Kvíčalová et al. (2014), or Tuleja and Gajdová (2015).

As was outlined above, the primary goal of this paper is to provide in a sense an intersection between credit risk and regional economic disparities and investigate the relationship between credit risk and economic indicators in the Czech Republic at the regional (NUTS 3) level. Banks generally monitor credit risk (e.g. observed default rates) in relation to geographical locations (regions) as a part of their credit concentration risk[2] management, which is also required by Directive 2013/36/EU (CRD, 2013) or CEBS (2010). However, this monitoring is predominantly performed on an individual basis (meaning from an individual bank's or even portfolio's point of view), and to our knowledge, the relationship between credit risk and economic indicators at the regional level has not been paid a great deal of attention on a more comprehensive level. Therefore, this paper aims to address this issue and at the end also demonstrates that region-specific information may be utilizable in credit scoring models or possibly other applications.

1 DATA AND METHODOLOGY

Firstly, credit risk and economic indicators at the regional level were investigated using simple graphical and correlation analysis. After that, the relationship between credit risk and economic variables was analyzed with linear regression models. Subsequently, hierarchical cluster analysis and model-based cluster analysis were performed, the latter within a Gaussian finite mixture modelling framework. Finally, it is demonstrated that region-specific information can be utilized in credit scoring models (using logistic regression). The following data are used in the above-mentioned analyses (at the regional level):

- *Past_due:* a share of population (adult natural persons) with past due obligations in % (source: SOLUS Register – see SOLUS, 2016);
- *Une:* general unemployment rate in % (source: Czech Statistical Office);
- *GDPpc:* gross domestic product per capita in CZK (source: Czech Statistical Office);[3]
- *Wage:* average wage in CZK (source: Czech Statistical Office);
- *Educ:* a share of the population with a university-level education in % (source: Czech Statistical Office – Population and Housing Census 2011).

[2] Concentration risk can be understood to be a sub-risk of credit risk class – see e.g. Holub et al. (2015). It is one of the specific risks subject to supervisory review under Pillar 2 within Basel II, since it is not fully covered by Pillar 1 capital requirements.

[3] For an interesting methodological discussion on GDP per capita see Chlad and Kahoun (2011).

The share of the population with past due obligations of a given region serves as a proxy of credit risk in this paper. It should also be noted that by "regional level", the NUTS 3 level is implied. Therefore, 14 regions are considered: Zlínský (Zlín Region, ZL), Vysočina (Vysočina Region, VY), Jihomoravský (South Moravian Region, SM), Praha (Prague, PR), Pardubický (Pardubice Region, PA), Jihočeský (South Bohemian Region, SB), Olomoucký (Olomouc Region, OL), Královéhradecký (Hradec Králové Region, HK), Středočeský (Central Bohemian Region, CB), Plzeňský (Plzeň Region, PL), Moravskoslezský (Moravian-Silesian Region, MS), Liberecký (Liberec Region, LI), Karlovarský (Karlovy Vary Region, KV), Ústecký (Ústí nad Labem Region, UL).

For the credit scoring model (within the Discussion section), the real data (as of 2014) from a small bank operating in the Czech Republic is used – specifically, a sample of nearly 90 thousand clients (private individuals).

1.1 Model-based cluster analysis

Model-based cluster analysis can be considered as an alternative to traditional clustering methods, such as hierarchical clustering or partitioning clustering (k-means, partitioning around medoids etc.). Since model-based cluster analysis is not used as widely as more traditional methods, it will be briefly described in this section. As Fraley and Raftery (2007) note, together with the development of methods and software tools for model-based clustering, these techniques are becoming increasingly popular and preferred over the heuristic methods mentioned in the beginning of this paragraph. A prevailing statistical approach to clustering is the use of finite mixture models – see e.g. McLachlan and Peel (2000).

In model-based clustering, the data y are treated as coming from a mixture density $f(y) = \Sigma_{c=1}^{G} \varrho_c f_c(y)$, where f_c represents the probability density function of the observations in group c, and ϱ_c denotes the probability that an observation comes from the c-th mixture component. Therefore, $\varrho_c \in (0,1)$ and $\Sigma_{c=1}^{G} \varrho_c = 1$. Generally, the individual components (clusters) are modelled using the Gaussian (normal) distribution that is characterized by the mean vector μ_c and the covariance matrix Σ_c. Parametrization of Σ_c allows us to determine various geometric features of the clusters (shape, volume, orientation). The probability density function takes the following form (Fraley and Raftery, 2007):

$$p(y_i|\, \mu_c, \Sigma_c) = \frac{\exp\left\{-\frac{1}{2}(y_i-\mu_c)^{\mathrm{T}}\Sigma_c^{-1}(y_i-\mu_c)\right\}}{\sqrt{\det(2\pi\Sigma_c)}}\ ,\ i = 1, \dots, n. \tag{1}$$

A Gaussian mixture model with multivariate mixture components has the likelihood function of the form:

$$\mathcal{L}(\varrho, \mu, \Sigma|y) = \Pi_{i=1}^{n} \Sigma_{c=1}^{G}\, \varrho_c p(y_i|\, \mu_c, \Sigma_c). \tag{2}$$

The model parameters $(\varrho_c, \mu_c, \Sigma_c)$ are commonly estimated by the expectation-maximization (EM) algorithm initiated by hierarchical model-based clustering. However, alternative approaches for parameter estimation in these kinds of applications exist – for an overview see McNicholas (2011). For further technical details on model-based clustering (including the 'mclust' package in R that is used in this paper), see Fraley and Raftery (2002), Fraley et al. (2012, 2016), or Scrucca et al. (2016).

Model selection strategies can be based on several measures – for an overview see McLachlan and Peel (2000). However, as for example McNicholas (2011) notes, Bayesian Information Criterion (BIC) (Schwarz, 1978) is the most prevalent mixture model selection measure in the literature (and is considered in this paper as well). BIC adds a penalty term to the loglikelihood that takes the complexity of the model (number of parameters) into account. Therefore, the BIC has the form:

$$\mathrm{BIC} = 2\hat{\ell}_{\mathcal{M}}\big(y|\hat{\theta}\big) - m_{\mathcal{M}} \log(n), \tag{3}$$

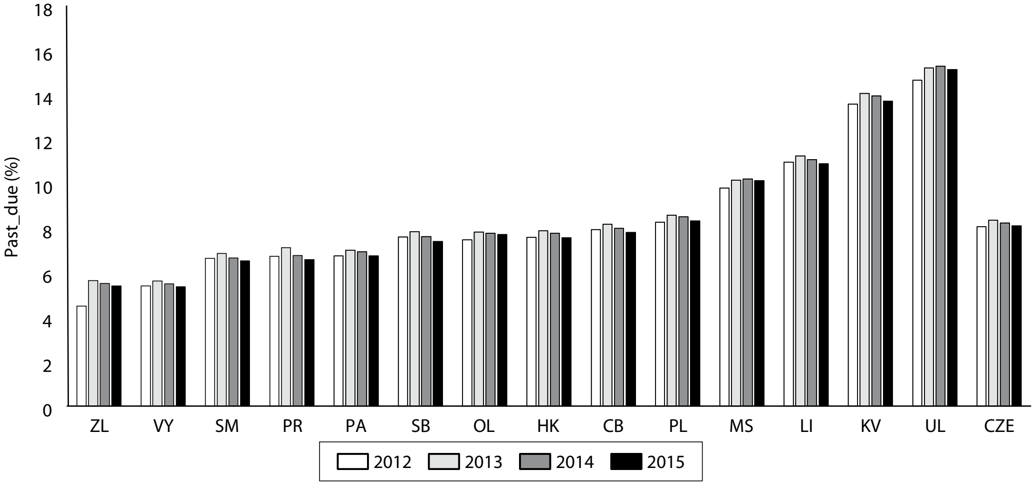
where $\hat{\ell}_{\mathcal{M}}$ is the maximized loglikelihood for model $\mathcal{M}$, $\hat{\theta}$,denotes the corresponding set of estimated parameters, and $m_{\mathcal{M}}$ represents the number of parameters of the model $\mathcal{M}$. Model selection in model-based clustering is discussed in more detail in e.g. Fraley and Raftery (1998), or Raftery and Dean (2006).

2 CREDIT RISK AND ECONOMIC INDICATORS AT THE REGIONAL LEVEL

2.1 Graphical and correlation analysis

Figure 1 depicts a share of the population with past due obligations in all of the considered regions and its development from 2012 to 2015. The last column represents the overall average. As it can be seen, the development of *Past_due* is quite stable over the observed time period.

Figure 1 Development of a share of the past due population in the individual regions from 2012 to 2015



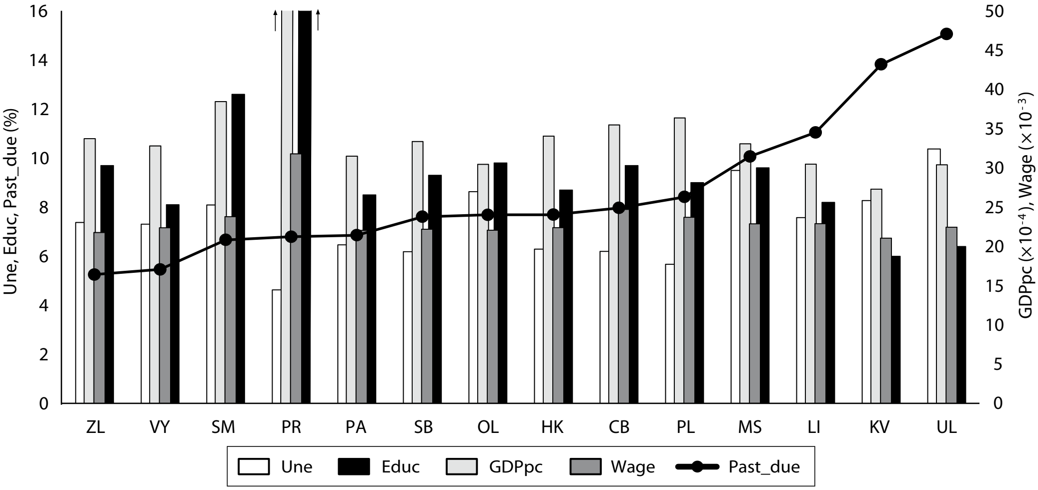
Source: SOLUS Register

The *Past_due* variable in the context of the considered economic indicators is presented in Figure 2. The values are obtained as averages of the variables over the observed period 2012–2015.[4] There is an exception in the case of the variable *Educ* – at the regional level the most current data is from the Population and Housing Census 2011. Even though there are some changes in the development of the variables from an absolute point of view, the proportions within the individual regions remain quite stable. Therefore, by averaging the values over years 2012–2015, no significant loss of information should occur – on the contrary, this procedure should assure that a more "long-term" view is provided by the performed analyses. Given the fact that relatively extreme values can be observed in the case of Prague (PR) – especially *Educ* 20.7% and *GDPpc* 813 thousand CZK – which would make differences among the other regions less visible in the graphical analysis, the scale of the vertical axes is adjusted in a corresponding manner to maintain lucidity.

Certain patterns can be observed from Figure 2. Generally, regions with relatively high shares of the population with past due obligations have higher unemployment rates, lower shares of the population with a university-level education and lower GDP per capita. The relationship between wages and past due rates is inconclusive. These observations are further elaborated below.

[4] *GDPpc* is averaged from 2012 to 2014 since the values of 2015 were not available when writing the paper.

Figure 2 The considered variables in the context of individual regions (averages from 2012 to 2015)



Source: SOLUS Register, Czech Statistical Office

Regarding unemployment, it turned out that it makes no considerable difference whether the general unemployment rate or registered unemployment rate is used in the analyses and estimations. Although there are some absolute changes between these two indicators, the proportions are similar. Moreover, based on the development of the standard deviations of all of the considered variables in the individual years, it can be said that no convergence is observed among the regions. In other words, the regional economic disparities in the discussed context do not become substantially less.

After the graphical analysis, a correlation analysis was performed. The correlations of the considered variables are summarized in Table 1. As above, the averages of the variables over the observed period 2012–2015 are used. Moreover, to avoid distorting the results with extreme values, Prague is excluded from the correlation analysis.

Table 1 Correlation analysis of the considered variables

Past_due	Une	GDPpc	Wage	Educ	
1	0.59*	–0.60*	–0.18	–0.68*	**Past_due**
	1	–0.43	–0.26	–0.21	**Une**
		–1	–0.74*	–0.80*	**GDPpc**
			–1	–0.53*	**Wage**
				–1	**Educ**

Note: Statistically significant correlation coefficients (*p*-value < 0.05) are marked with *.
Source: Own calculations

Focusing on the relationship between credit risk and economic variables, a high negative correlation is observed between *Past_due* and *Educ* (–0.68). This is logical, since it is expected that more educated people would naturally tend to have less problems with repaying their loans (they are expected to be more financial literate etc.). A high negative correlation can also be observed between *Past_due* and *GDPpc* (–0.60), which is also natural given the fact that in regions with higher *GDPpc* (i.e. with higher economic performance) there are higher wages and education rates, which is supported by high positive correlations between *GDPpc* and *Wage* (0.74), and *GDPpc* and *Educ* (0.80).

Nevertheless, the direct relationship between *Past_due* and *Wage* is insignificant. One might expect that regions with higher wages would tend to have smaller shares of the population with past due obligations; however, wages themselves do not provide much useful information in this context, since the value of loans is not taken into account. Therefore, ratio indicators such as debt-to-income (DTI) would be more evidential. The following sections provide a deeper insight into the relationship between *Wage* and *Past_due*.

Furthermore, a relatively high positive correlation between *Past_due* and *Une* can be seen from the analysis (0.59), which is also logical since it is naturally expected that regions with smaller unemployment rates will have smaller shares of population with past due obligations. Therefore, it can be said that the results from the correlation analysis correspond to the prior expectations and in a sense summarize what is depicted in Figure 2.

2.2 Regression analysis

After the graphical and correlation analysis, the relationship between credit risk and economic variables was further investigated using regression analysis. Specifically, cross-sectional and panel data regression analyses were performed. For the purpose of regression and cluster analyses (in Section 2.3), the considered variables are scaled to the same order due to statistical reasons (therefore *Wage* is in tens of thousands of CZK and *GDPpc* is in hundreds of thousands of CZK). Also, Prague is excluded since it can be considered as an extreme case (or outlier) that may distort the results.

As it was noted above, the development of the considered variables is quite stable over the observed time period in the individual regions. This is also shown in Table 2, which summarizes the main results of the five cross-sectional regressions performed (data as of 2012, 2013, 2014, 2015 and averages). Linear cross-sectional regression models were estimated by the ordinary least squares (OLS) method with heteroscedasticity robust standard errors.

Table 2 Cross-sectional regression results

	2012		2013		2014		2015		Averages	
	coef.	*p (t)*	coef.	*p (t)*	coef.	*p (t)*	coef.	*p (t)*	coef.	*p (t)*
const	−22.48	0.1271	−15.46	0.1263	−6.73	0.4970	−4.84	0.5703	−12.64	0.2354
Une	1.11	0.0010	1.05	0.0003	1.04	0.0001	1.19	0.0001	1.11	0.0002
Wage	16.28	0.0155	12.66	0.0042	8.63	0.0493	7.42	0.0554	11.20	0.0172
Educ	−1.51	0.0012	−1.41	0.0004	−1.35	0.0017	−1.29	0.0013	−1.39	0.0009
R^2	0.73		0.79		0.75		0.75		0.76	
p (F)	0.0004		0.0000		0.0000		0.0000		0.0000	

Source: Own calculations

The variable *GDPpc* was excluded from the regressions after the backward elimination procedure. This can be explained by the high correlations between *GDPpc* and *Educ,* and *GDPpc* and *Wage* (see Table 1).[5] However, as it can be seen in Table 2, the signs of *Une* and *Educ* are as expected and described above (in Section 2.1). The sign of the variable *Wage* is positive, nevertheless, as it was noted, in order to create a direct link with credit risk (*Past_due*), an indicator considering both wage and the volume of the loan would have to be used (e.g. DTI). With wage itself, the logic is incomplete. Further comments on this topic are given below.

[5] Based on the values of the calculated variance inflation factors (in all cases under 1.7 after exlusion of *GDPpc*), it can be concluded that no additional multicollinearity problems arised. For details on this topic and other relevant diagnostic tests of linear regression models see e.g. Heij et al. (2004), or Greene (2012).

Furthermore, several panel data regression models were estimated. Table 3 summarizes the final selected one. The null hypothesis that all regions have the same intercept was rejected, hence the individual-specific effects model was preferred. Also, the Hausman test (see Hausman, 1978) proved that the random effects generalized least squares (GLS) estimator is consistent and more efficient than the fixed effects one. Therefore, the random effects model was preferred to the fixed effects model. The random effects panel data regression model was estimated by the Arellano heteroscedasticity and autocorrelation robust GLS estimator (see Arellano, 2003) using the Nerlove method for estimating "within" and "between" variance (see Nerlove, 1971). However, the results were robust to the use of other methods, e.g. Swamy and Arora (1972). The Nerlove method was preferred due to slightly lower standard deviations of the obtained estimates. Although the GDP data at the regional level were not available for 2015, *GDPpc* was not considered in this estimation (similarly as above) – therefore, the data used covers the 'full' time span 2012–2015.

Table 3 Random effects panel regression results

	coef.	***p (t)***
Const	12.23	0.0275
Une	0.26	0.0000
Wage	2.61	0.0021
Educ	–1.27	0.0248
corr*(y, ŷ)*^{2}	0.60	
Hausman test *(p)*	0.33	

Source: Own calculations

To a large extent both types of regressions yield corresponding results. After analyzing the relationship between *Wage* and *Past_due*, the results obtained from cross-sectional and panel regressions are also in line – a positive sign can be observed. Since the regression analyses take indirect effects into account, the results obtained in this section are considered to be more plausible compared to the results from the simple correlation analysis (in Section 2.1).[6]

Therefore, the results imply that regions with higher wages tend to have a higher share of the population with past due obligations. The reasoning behind this statement could be that people with higher wages generally tend to take higher loans, and at the same time these people are more sensitive to adverse events, typically losing their job. For people with higher wages it may be more difficult to find a job with corresponding salary in a reasonably short time to cover the relatively high repayments of their loans. Another agrument could be that there are different price levels across regions implying that higher nominal wages do not necessarily mean higher real wages.

2.3 Cluster analysis

So far, we have dealt with the a more "overall" picture. The relationship between credit risk and selected economic indicators was investigated using region-level data. In this section, the individual regions will be directly worked with, in order to cluster them based on various shared economic characteristics. Two types of clustering are performed – hierarchical and model-based.

[6] This reasoning is also supported by the fact that the Pearson partial correlation coefficient (taking other variables into account) of P*ast_due* and *Wage* is positive.

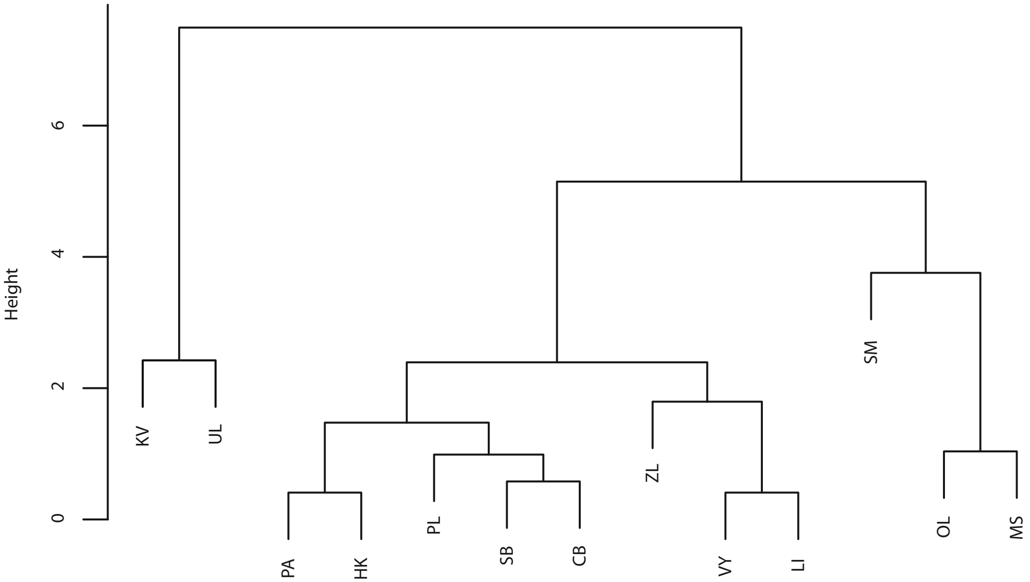
Figure 3 shows a dendrogram obtained from the hierarchical cluster analysis with a Euclidean distance matrix using the complete linkage method. Note, however, that the results yielded by using other methods (e.g. Ward) were very alike. Similarly as above, Prague was excluded from this analysis. Also, scaled averages of variables are used. All of the calculations in this section are performed in the R computational system (R Core Team, 2017).

The results of the model-based cluster analysis are summarized in Table 4. In this case, the analysis was performed in four versions – with 2, 3, 4, and 5 clusters. The mentioned options were set to be reasonable, given the application and subsequent interpretation. In every version, the final model was selected by running an optimization exercise within Gaussian finite mixture models using the 'mclust' package in R (Fraley et al., 2012, 2016). The optimization measure was the Bayesian information criterion (BIC) and the models were estimated using an expectation-maximization algorithm initialized by hierarchical model-based clustering. The selected models for every version have the following features:

- v1 (2 clusters): ellipsoidal distribution, equal volume and orientation;
- v2 (3 clusters): diagonal distribution, equal shape of clusters;
- v3 (4 clusters): diagonal distribution, equal volume and shape of clusters;
- v4 (5 clusters): diagonal distribution, equal volume and shape of clusters.

In this case, the data as of 2014 (most recent and complete) were used, excluding Prague.

Figure 3 Dendrogram of the considered regions



Source: Own calculations

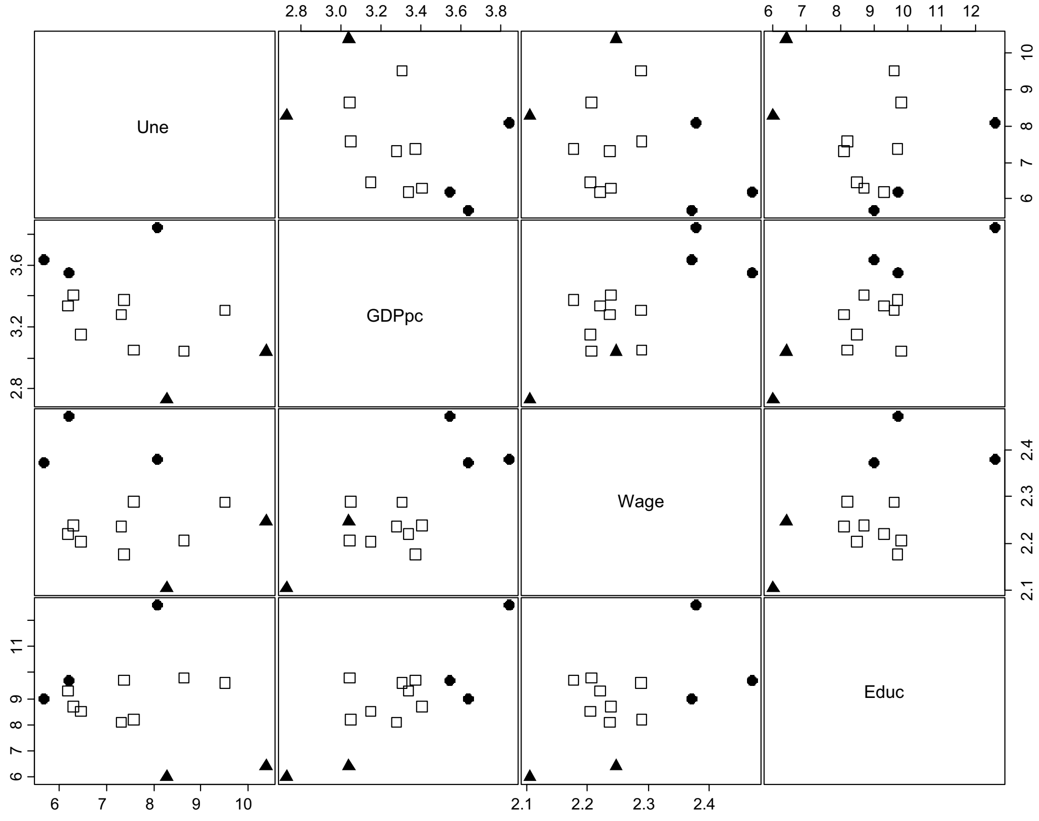
The version with three clusters (illustrated in Figure 4) seems to be the most appropriate. Table 4 suggests that the SM region tends to be left alone as a separate cluster when the number of clusters increases. On the other hand, only two clusters are not sufficient – looking also at Figure 3, it is clear that regions KV and UL (cluster Δ in Table 4 and Figure 4) stand on the side from both an economic and credit risk point of view. Therefore, from the performed analyses it can be seen that the regions are clustered mostly into three groups, and this version is used in the following text.

Table 4 Model-based cluster analysis summary

No.	Region	Past due	2 clusters (v1)	3 clusters (v2)	4 clusters (v3)	5 clusters (v4)
1	ZL	5.50	□	□	□	▪
2	VY	5.48	□	□	□	□
3	SM	6.65	○	○	+	+
4	PA	6.93	□	□	□	□
5	SB	7.61	□	□	□	○
6	OL	7.77	○	□	□	▪
7	HK	7.77	□	□	□	□
8	CB	7.98	□	○	○	○
9	PL	8.50	□	○	○	○
10	MS	10.20	○	□	□	▪
11	LI	11.08	□	□	□	□
12	KV	13.95	○	Δ	Δ	Δ
13	UL	15.28	○	Δ	Δ	Δ
14	PR	6.80	*Excluded*			

Source: Own calculations

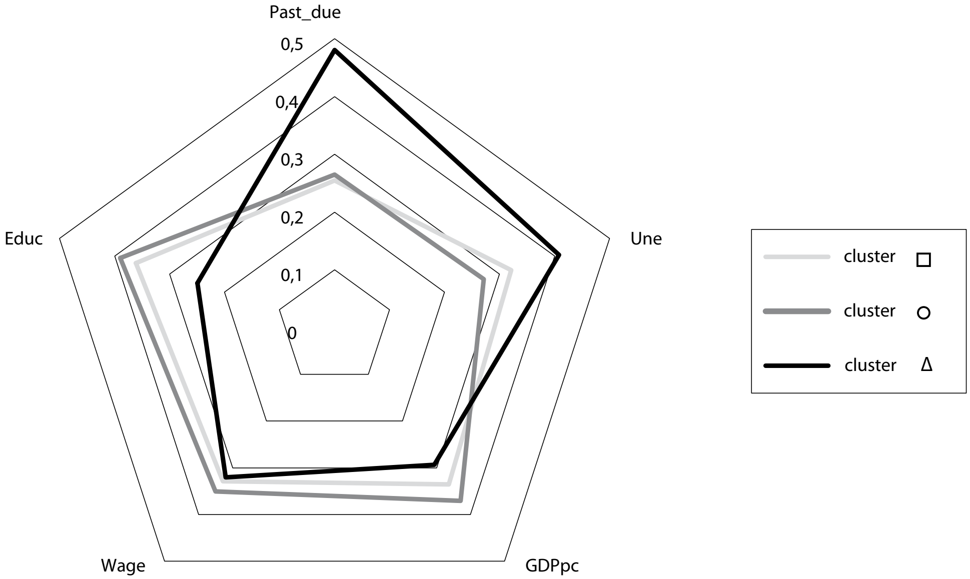
Figure 4 Result of the model-based cluster analysis (three clusters)



Source: Own calculations

The relationship between the considered economic variables and credit risk *(Past_due)* within the individual clusters is further investigated using a spider graph – see Figure 5. Note that the medians of the corresponding variables are used for each cluster of regions. Moreover, for illustrative purposes, the values are normalized across the clusters. The darkest line representing cluster Δ confirms the findings from the previous analyses. Regions in this cluster have the highest unemployment rate, lowest GDP per capita, lowest wages and lowest education, and at the same time the highest share of the population with past due obligations. Regarding the other two clusters, it can be seen that even though the regions in cluster □ (the lightest line) are "worse" than □ in terms of economic performance indicators, their share of the population with past due obligations is comparable (even slightly higher in the case of ○). One may suggest that two clusters could be sufficient due to similarities between clusters □ and ○; however, it should be noted that the scale of the spider graph is in a sense "extended" because of cluster Δ, which optically diminishes the differences. Moreover, the choice of three clusters will also prove to be appropriate in the next section.

Figure 5 Spider graph of individual clusters in the context of considered variables



Source: Own calculations

Therefore, the link between the considered economic indicators and credit risk *(Past_due)* is rather fragmentary and cannot be fully generalized over all of the regions. It can also be noted that there are of course many factors influencing the level of *Past_due*, especially on the individual (client) level. Therefore, it cannot be expected that economic variables to a large extent explain this issue. However, some of the obtained results are quite straightforward and evidential (especially regarding cluster Δ) and the next section investigates whether the region-specific data can be utilized in credit scoring modelling or possibly other applications.

DISCUSSION

In this section, utilization possibilities of region-specific data will be discussed, especially in the context of credit scoring. For this purpose, additional calculations were performed. Firstly, a simple credit scoring model using dummy explanatory variables for each individual region was estimated, with the binary default variable as the dependent variable. Secondly, the findings from Section 2 were utilized. Therefore, the regions were clustered into three groups for which dummy variables were created.[7] The first region/cluster was treated as a reference category. The concept of the model was otherwise the same as in the first case – the credit scoring model had the form of the standard logistic regression and was estimated by the maximum likelihood method. The real data (as of 2014) from a small bank operating in the Czech Republic is used – specifically, a sample of nearly 90 thousand clients (private individuals).

In the first case (with dummy variables for individual regions), only 2 out of 14 regression coefficients were statistically significant at the 5% level. Moreover, one of them was a constant. In the second case (with dummy variables for clusters), all regression coefficients were statistically significant (3 out of 3). The performance of the models was then evaluated using the area under the receiver operating characteristic (ROC) curve – see e.g. Engelmann et al. (2003) or Fawcett (2006). The results for the two models described above were 0.55 and 0.54, respectively. Of course, the values are relatively small and close to 0.5 since the models predict the probability of default of clients only using information about the regions they come from. The main result is that for a relatively little loss of model performance, a much more "compact" model can be obtained, with statistically significant parameters. Therefore, from a statistical point of view, the second model may be preferred.

Khudnitskaya (2010) achieved a better scoring model performance with region-specific (or microenvironment-specific) information using logistic regression with a multilevel structure. Even though American data were used in her study, which are more hierarchically structured compared to the Czech data (given the size of the US), and therefore more comfortable and suitable to use in multilevel models, the multilevel modelling framework seems to be promising in this context.

Apart from Khudnitskaya (2010), this paper cannot be directly compared to the other studies mentioned in the introduction. The first set of studies deals with credit risk in the context of macroeconomy – Pesaran et al. (2006, 2007) use global vector autoregression models, Jakubík and Heřmánek (2008) estimate a vector error correction model (among others), Jakubík (2008) works with Merton-type models, Grešl et al. (2016) provide an interesting overview of the Czech National Bank's stress-testing framework (including time-series and macroeconometric techniques), Melecký et al. (2015) use an autoregressive distributed lag model with instrumental variables, and in a recent study Schwaab et al. (2016) investigate credit risk from a global perspective, using a non-linear state-space model.

On the other hand, there are studies devoted to regional disparities, but without a link to credit risk. Kutscherauer at al. (2010) analyze and evaluate regional disparities especially using integrated indicators and models of the economic power of the regions. Kahoun (2010) studies regional disparities with GDP per capita and net disposable income (using descriptive statistics) and evaluates the suitability of these measures. Svatošová and Novotná (2012) also use descriptive statistics for investigating regional disparities, considering several demographic, social, economic and ecologic variables. Procházková and Radiměřský (2013) study the economic performance of regions using a regression model with socio-demographic and industrial factors. Kvíčalová et al. (2014) identify differences between regions based on economic characteristics using correlation analysis and hierarchical cluster analysis. Tuleja and Gajdová (2015) investigate the economic potential of regions using in particular graphical methods of magic polygons.

[7] Although Prague was excluded from the cluster analysis, for this application it was included in cluster ○ due to greater similarity with regions inside this cluster. However, the results would not substantially differ in a quantitative way even if Prague was treated as a separate cluster.

CONCLUSION

In this paper the regression techniques and methods of cluster analysis were in a sense combined with the extension to credit scoring application, using region-specific data. This work therefore partially contributes to the both areas of literature (credit risk in the context of economy and regional economic disparities), and its main added value lies in the interconnection of the two.

At this point, it should be noted that a larger set of appropriate economic variables will be analyzed in further research, e.g. factor productivities – see Vltavská and Sixta (2011). Also, considering regional price levels may be an improvement – see Musil et al. (2012) and Kocourek et al. (2016).

Regarding the credit scoring application, banks often categorize the regions in an expert way based on their internal historical experience. Based on the character of the banks' population (character of clients), it may occur that e.g. using solely historical data in application credit scoring could lead to prior incorrect penalization of clients from certain regions if the population is in a sense "biased". On the other hand, the clustering approach used in this paper is independent of the bank-specific population and thus more robust in this context. Using clusters of regions also diminishes the number of parameters to be estimated compared to the case of dummy variables for each region, which may help to minimize potential numerical instability issues. As it was suggested, the clustering approach could be further used in certain types of stress-testing exercises, where augmentation of the analysis dimension by the regional level may help to improve the results (compared to the case where the Czech Republic would be considered as a whole).

To conclude, it can be summarized that this paper bridges the areas of credit risk and regional economic disparities, and investigates the relationship between the credit risk and economic indicators in the Czech Republic at the regional (NUTS 3) level. This relationship was consecutively examined using graphical and correlation analysis, regression techniques, and different types of clustering methods. Regions were then clustered into three groups according to their economic similarities and disparities. Subsequently, it was shown that region-specific information has the potential to be utilizable in credit scoring and possibly other applications. This area will be further investigated in the context of a multilevel modelling framework, which provides the possibility to improve scoring models to a larger extent.

ACKNOWLEDGEMENT

We thank the two anonymous reviewers for their helpful comments. This work was supported by the Internal Grant Agency of FBE MENDELU, the project PEF_DP_2017021.

References

ABDOU, H. AND POINTON, J. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting, Finance & Management*, 2011, 18(2–3), pp. 59–88.

ARELLANO, M. *Panel Data Econometrics*. Oxford: Oxford University Press, 2003.

BCBS (Basel Committee on Banking Supervision). *Principles for the Management of Credit Risk*. Bank for International Settlements, 2000.

BCBS (Basel Committee on Banking Supervision). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Bank for International Settlements, 2004.

CEBS (Committee of European Banking Supervisors). *CEBS Guidelines on the Management of Concentration Risk under the Supervisory Review Process (GL31)*. London, 2010.

CHLAD, M. AND KAHOUN, J. Factors Influencing the Rating of Regional Economic Performance or Reasons why Prague has Become the 6th Best Economically Performing Region of the EU. *Statistika: Statistics and Economy Journal*, 2011, 91(2), pp. 4–23.

CROOK, J. N., EDELMAN, D. B., THOMAS, L. C. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 2007, 183(3), pp. 1447–1465.

CRD. *Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC.*

CRR. *Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No. 648/2012*.

ENGELMANN, B., HAYDEN, E., TASCHE, D. Testing rating accuracy. *Credit Risk*, 2003, 16 (January), pp. 82–86.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, 27(8), pp. 861–874.

FRALEY, C. AND RAFTERY, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 1998, 41(8), pp. 578–588.

FRALEY, C. AND RAFTERY, A. E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 2002, 97(458), pp. 611–631.

FRALEY, C. AND RAFTERY, A. E. Model-based Methods of Classification: Using the mclust Software in Chemometrics. *Journal of Statistical Software*, 2007, 18(6), pp. 1–13.

FRALEY, C., RAFTERY, A. E., MURPHY, B. T., SCRUCCA, L. *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Technical Report No. 597, Department of Statistics, University of Washington, 2012.

FRALEY, C., RAFTERY, A. E., SCRUCCA, L. *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation* [online]. R package version 5.2, 2016. <https://CRAN.R-project.org/package=mclust>.

GREENE, W. H. *Econometric Analysis*. 7th Ed. New Jersey: Pearson, 2012.

GREŠL, A., JAKUBÍK, P., KONEČNÝ, T., SEIDLER, J. Dynamic Stress Testing: The Framework for Assessing the Resilience of the Banking Sector Used by the Czech National Bank. *Czech Journal of Economics and Finance*, 2013, 63(6), pp. 505–536.

HAUSMAN, J. A. Specification tests in econometrics. *Econometrica*, 1978, 46(6), pp. 1251–1271.

HEIJ, C., DE BOER, P., FRANSES, P. H., KLOEK, T., VAN DIJK, H. *Econometric Methods with Applications in Business and Economics*. New York: Oxford University Press, 2004.

HOLUB, L., NYKLÍČEK, M., SEDLÁŘ, P. *Credit Portfolio Sector Concentration and Its Implications for Capital Requirements*. Chapter Thematic Article 3 in CNB Financial Stability Report 2014/2015, 2015, pp. 131–138.

IFRS Foundation. *Financial Instruments 2015 Guide*. London: IFRS Foundation, 2015.

JAKUBÍK, P. AND HEŘMÁNEK, J. Stress testing of the Czech banking sector. *Prague Economic Papers*, 2008, 6, pp. 195–212.

JAKUBÍK, P. Credit risk and stress testing of the Czech Banking Sector. *ACTA VŠFS*, 2008, 2(1), pp. 107–123.

KAHOUN, J. *Regionální ekonomická výkonnost a disponibilní důchod domácností*. Working paper No. 15, Research Centre for Competitiveness of Czech Economy, 2010.

KHUDNITSKAYA, A. S. *Improved Credit Scoring with Multilevel Statistical Modelling*. PhD. Thesis, Technische Universität Dortmund, 2010.

KOCOUREK, A., ŠIMANOVÁ, J., ŠMÍDA, J. Estimation of Regional Price Levels in the Districts of the Czech Republic. *Statistika: Statistics and Economy Journal*, 2016, 96(4), pp. 56–70.

KVÍČALOVÁ, J., MAZALOVÁ, V., ŠIROKÝ, J. Identification of the Differences between the Regions of the Czech Republic based on the Economic Characteristics. *Procedia Economics and Finance*, 2014, 12, pp. 343–352.

KUTSCHERAUER, A. et al. *Regional disparities in regional development of the Czech Republic – their occurrence, identification and elimination*. WD-55-07-1, VŠB Technical University of Ostrava, Faculty of Economics, 2010.

LESSMANN, S., BAESENS, B., SEOW, H.-V., THOMAS, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 2015, 247(1), pp. 124–136.

LI, X.-L. AND ZHONG, Y. An Overview of Personal Credit Scoring: Techniques and Future Work. *International Journal of Intelligence Science*, 2012, 2(4), pp. 181–189.

McNICHOLAS, P. D. On Model-Based Clustering, Classification, and Discriminant Analysis. *Journal of the Iranian Statistical Society*, 2011, 10(2), pp. 181–199.

McLACHLAN, G., PEEL, D. *Finite Mixture Models*. New York: Wiley, 2000.

MELECKÝ, A., MELECKÝ, M., ŠULGANOVÁ, M. Úvěry v selhání a makroekonomika: modelování systémového kreditního rizika v České republice [Non-Performing Loans and The Macroeconomy: Modeling the Systemic Credit Risk in the Czech Republic]. *Politická ekonomie*, 2015, 8, pp. 921–947.

MUSIL, P., KRAMULOVÁ, J., ČADIL, J., MAZOUCH, P. Application of Regional Price Levels on Estimation of Regional Macro-Aggregates Per Capita in PPS. *Statistika: Statistics and Economy Journal*, 2012, 92(4), pp. 4–13.

NERLOVE, M. Further Evidence on the Estimation of Dynamic Economic Relations from a Time-Series of Cross Sections. *Econometrica*, 1971, 39(2), pp. 359–382.

OECD (Organisation for Economic Co-operation and Development). *OECD Regions at a Glance 2016*. OECD Publishing, 2016.

PESARAN, H. M., SCHUERMANN, T., TREUTLER, B.-J., WEINER, S. Macroeconomic Dynamics and Credit Risk: A Global Perspective. *Journal of Money, Credit and Banking*, 2006, 38(5), pp. 1211–1261.

PESARAN, H. M., SCHUERMANN, T., TREUTLER, B.-J. Global Business Cycles and Credit Risk. In: CAREY, M., STULZ, R. M. *The Risks of Financial Institutions*, National Bureau of Economic Research, 2007.

PROCHÁZKOVÁ, L. AND RADIMĚŘSKÝ, M. The economic performance of regions in the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 2013, 61(7), pp. 2661–2667.

R CORE TEAM. R: *A language and environment for statistical computing* [online]. R Foundation for Statistical Computing, Vienna, Austria, 2017. <http://www.R-project.org>.

RAFTERY, A. E. AND DEAN, N. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 2006, 101(473), pp. 168–178.

SCRUCCA, L., FOP, M., MURPHY, B. T., RAFTERY, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 2016, 8(1), pp. 289–317.

SCHWAAB, B., KOOPMAN, S. J., LUCAS, A. *Global credit risk: world, country and industry factors*. Working paper No. 1922, European Central Bank, 2016.

SCHWARZ, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 1978, 6(2), pp. 461–464.

SHANKAR, R. AND SHAH, A. Bridging the Economic Divide Within Countries: A Scorecard on the Performance of Regional Policies in Reducing Regional Income Disparities. *World Development*, 2003, 31(8), pp. 1421–1441.

SOLUS. *Archiv tiskových zpráv* [online]. 2016. <https://www.solus.cz/cs/archiv-tiskovych-zprav>.

SVATOŠOVÁ, L. AND NOVOTNÁ, Z. Regionální disparity a jejich vývoj v České republice v letech 1996–2010. *Acta Universitatis Bohemiae Meridionalis,* 2012, 15(1), pp. 103–110.

SWAMY, P. A. V. B. AND ARORA, S. S. The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models. *Econometrica*, 1972, 40(2), pp. 261–275.

THOMAS, L. C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 2000, 16(2), pp. 149–172.

TULEJA, P. AND GAJDOVÁ, K. Economic Potential of the Regions of the Czech Republic. *Journal of Economics, Business and Management*, 2015, 3(1), pp. 43–47.

VLTAVSKÁ, K. AND SIXTA, J. The Possibilities to Estimate Labour Productivity and Total Factor Productivity for Czech Regions. *Statistika: Statistics and Economy Journal*, 2011, 91(4), pp. 35–44.