Foreclosure Pricing Model

Jakub Chrenko, Petr Kadeřábek PhD
Česká spořitelna a.s. , Erste Group

Abstract

Real estate market is one of the well-developed price information markets. However, when real estate collateral held by a bank is forced to be sold, we observe significant difference in the ratio of auction selling price to market price across geographical locations within the country. The difference is from 10 up to 70%. We assume that this is due to the different liquidity in different parts of the country. When collateral is sold in the auction, the time offer is limited compared to the normal market transaction situation. We assume that the liquidity is correlated with how big number of people founds the region good for living. That is how many people are interested in buying a real estate in that respective location within a short period of time. When banks evaluate their pledged real estate collateral they usually look at the market price and apply certain expertly based haircuts. These haircuts do not represent the liquidity problem but rather insurance considerations in our bank (e.g. flood zone, fire protection etc.). We have come up with a statistical model to estimate a forced sale liquidity haircut that is more suited for predicting banks expected loss, which is the auction selling price, rather than the market price. We have used banks auction sales historical data and found predictor successfully describing 20% of this phenomenon; giving us better insight for the future haircut application. To estimate the liquidity haircut we look at the attractiveness of the real estate location. We have found this to be correlated with the proximity of other attractive real estate properties. Analogy may be found in Google Page Rank algorithm where attractiveness of a web page is measured by attractiveness of pages linked to it via hyperlinks.

Today we therefore present an adaptation of the Google Page Rank algorithm (Brin, S.; Page, L. 1998 "*The anatomy of a large-scale hypertextual Web search engine*" Computer Networks and ISDN Systems 30: 107–117.) constructing real estate location variable for predictive model of the liquidity haircut. In our presentation, first, the original Google Page Rank algorithm is briefly introduced. Then we present an analogy of this algorithm to our problem. Instead of ranking pages our aim is to rank the geographical locations. While Google uses hyperlinks for their ranking in our case the link is the geographical proximity of the location to another well ranked location. We have worked on the real data of banks auctions selling prices and with the official geographical division of Czech republic (52000 regions approx. $1.5km^2$ each). We present this application of the algorithm with some graphical results and the overall results of this new model for the bank.

Keywords: Real Estate, Collateral, Auction, Market Price, Selling Price, Google Page Rank

1 Motivation

In the banking the collateral evaluation is a well known problem. In this article we focus only on one of the aspects of evaluation of real estate held by bank as a collateral. We are interested in its forced sell price which takes place mainly to cover potential losses from mortgages which are not being repaid.

The evaluation of the collateral is a very well assessed area. In the bank it is done by advanced tools like pricing (whether it is done by a model or/and an expert opinion), rating systems, insurance and others, most importantly $LTV = \frac{\text{loan amount}}{\text{nominal value of collateral}}$ limits.
The mortgage market in the Czech republic is a very competitive environment which leads to origination processes using the LTV limits to its borders of 100%, even higher. In such cases we have to be extra careful because the real estate pledged to the bank can be the only thing which stands between the losses and well covered cases. The cases when the real estate is sold by bank and then the debt is settled are varying a lot. We have noticed the sell price being 10% to 95% of the normal market price. This is due to several causes. The first would be that the forced selling of such a property is done by an auction. This may be happening quite quickly compared to the real estate market where nobody invests the money on short notice and with limited access to the property. Also there could be some legal problems for the buyer of such real estate. Hence the liquidity in such cases is bad.

We make an assumption that the location has the biggest influence on liquidity. The main focus of this paper is to present a way how it is possible to quantify

location via known mathematical terms and algorithms. We show an approach we took in case of the Czech republic and a regression analysis based on historical data of Česká spořitelna a.s. which tells us how the process was succesfull. Finally we argue that this approach can help to establish better origination process when taking our results into account.

2 Ideas: How to Rank Regions?

It is crucial to say that in the case of Czech republic there is a geographical division of the whole state into 52227 small areas. We call these Micro-regions. Each Micro-region has an area of $\approx 1.5\text{km}^2$. This division is publicly accessible and it is constructed by the Land registry office based on location. It is defined by the GPS coordinates of the addresses assigned to the respective Micro-region. We consider it to be quite well defined and detailed enough to be used for such applications.

Thinking about a Micor-region scoring we have come up with some assumptions about its features.

- The regions which are populated a lot (e.g. big towns) should have a higher score which is reflecting that there is more potential buyers in such area.
- The regions which are close to these more populated areas should be influenced by the score of the neighboring region because they are within the reach from the original area.

3 Web Page Rating by Page Rank

When we were searching for some means to incorporate the above mentioned features we have come across the Page Rank algorithm used by Google. The Page Rank scores the internet pages via iterative algorithm that simulates a behavior of an internet surfer who is viewing the World Wide Web. By 'behavior' we mean either choosing a link from the page the surfer is currently viewing leading to another or choosing another page he will visit at random.

The original algorithm works as follows:
Assume a finite set of web pages

$$\mathcal{W} = \{w_1, \ldots, w_N\}, \tag{1}$$

with the set of indexes

$$\mathcal{I} = \{1, \ldots, N\}, \tag{2}$$

and a set of links references

$$\mathcal{S} = \{s(w_i, w_j)\}_{i,j\in\mathcal{I}} = \{s_{ij}\}_{i,j\in\mathcal{I}} \in \{0, 1\}, \tag{3}$$

where 0 represents no link and 1 represents a link going from page w_i to page w_j.
Assume a function f which for every $w_i \in \mathcal{W}$ assigns a value $f(w_i) \in [0, 1]$ such that

$$\sum_{i\in\mathcal{I}} f(w_i) = 1. \tag{4}$$

Assume an iterative calculation process which follows the following steps

$t = 0$ assign an initial probability distribution to $\mathcal{W}$, the usual pick is $f(w_i) = \frac{1}{N}, \forall i \in \mathcal{I}$,

$$t \to t+1 \;\; f_{t+1}(w_i) = \frac{1-d}{N} + d \sum_{j\in\mathcal{I}} \frac{f_t(w_j)s_{ji}}{\sum_{k\in\mathcal{I}} s_{ik}}, \quad d \in (0, 1).$$

The value of this function for iteration t and $\forall t \in 0, 1, \ldots$ forms a probability distribution.
The d is called the damping factor and is used to describe the likelihood of jumping to a **random** page.

The Page Rank algorithm simulates a behavior of a random internet surfer with a list of available web pages. The surfer starts at a random page and at each step he either clicks on a link leading him to another page or chooses randomly from the initial list. This behavior is represented by the iterative second step. The damping factor d is chosen typically around 0.85 so any Page Rank is constructed mostly as a sum of incoming scores.

The Page Rank of the web page is a probability to which this algorithm converges and it can be seen as a probability that a random surfer ends up looking at that particular page.

The convergence of this algorithm is guaranteed by the construction. We can see the iterations as a Markov chain with finite number of states and which is irreducible(again by construction, no absorbing states etc. can be found on the internet). Hence it is ergodic and converges to a stationary distribution. Also there will be no dependency on the starting probability distribution.

4 Algorithm Adjustments

We see the approach of the Page Rank algorithm parallel to the one we would like to use for the regions. We imagine regions as web pages we want to rank and apply the same above mentioned logic to our case. We assume that a random person who moves the house is described by similar behavior. He either stays close to his original location, follows the 'link' to the nearby area or chooses randomly between all the other regions. So the parallels we assume are:

internet surfer	=	**person moving house**
following links	=	**moving nearby**
random page visits	=	**moving at random**

If we accept these premises we need to do certain adjustments to the original algorithm.

The links between the regions are defined according to our assumption that the regions which are *close* to each other should influence the score of its neighbor.

First we define the region. As each region is defined by the set of addresses with its GPS coordinates there is a lot of possibilities how to approach and how much detail will be subject to study. For the simplicity and computational easiness we have chosen the following: Each area is represented by its center and we will perceive it(for the application of the algorithm) as one point. Hence the distance between two regions is clearly defined as the distance of these points representing centers.

We define links as

$$\begin{aligned} \text{if } \mathbf{dist}(w_i, w_j) \leq T \quad & s_{ij} = 1, \\ \text{otherwise} \quad & s_{ij} = 0, \end{aligned}$$

where T is a threshold parameter which is chosen prior to the computation.

We introduce weighing of the links as an additional parameter to reflect the quality of the link. The weight is defined based on the distance between two areas and number of addresses in the area as follows:

$$\begin{aligned} \mathcal{G} = & \quad \{g(w_i, w_j)\}_{i,j \in \mathcal{I}} = \\ = \{g_{ij}\}_{i,j \in \mathcal{I}} = & \quad \exp(-\tfrac{1}{2\sigma^2}\mathbf{dist}(w_i, w_j))^2 N_j. \end{aligned} \tag{5}$$

where N_j corresponds to the number of real estates in the region w_j. Then iterative step of the algorithm is then

$$f_{t+1}(w_i) = \frac{1-d}{N} + d \sum_{j \in \mathcal{I}} \frac{f_t(w_j) g_{ji} s_{ji}}{\sum\limits_{k \in \mathcal{I}} g_{ik} s_{ik}}, \quad d \in (0,1). \tag{6}$$

The stopping criterion we use is $\max\limits_{i \in N} \Delta f_t(w_i) = \max\limits_{i \in N}(f_t(w_i) - f_{t-1}(w_i)) < \epsilon$. So we stop the algorithm when the Page Rank does not change above defined level through the last iterations.

We call the resulting score (or index) attractiveness of the respective Micro-region because it can be seen according to the case of web pages as a probability that a random person moving house ends up living in the designated area.

5 Results

The algorithm was studied for different inputs after that the results were used to predict the actual loss we have observed. The sets of parameters we were inspecting were

- Different thresholds for the parameter T. This parameter, influencing the connection of regions were tested on an interval $(25km, 40km)$.
- Different values for the parameter σ. This parameter, influencing the rate at which the weight is influenced rather by the distance or rather by the number of addresses connection of regions were tested on an interval $(5, 20)$.
- Different damping factors d on an interval $(0.7, 0.9)$ were tested, influencing the ratio of random moves in the iteration step.

First let us have a look at the Page Rank results. We show the resulting scores of different Micro-regions displayed as a part of the map of the Czech republic. The resulting scores are grouped into several rating groups as illustrated on the Fig. [5]. The picture shows the results for the final setup of the algorithm we have used further. The black colour is used for the regions with the best score, the green colour is used for the ones with the worst score. It also that the calculated score has the features we wanted.

The best areas are comparable to big towns and high population density regions. Note the influence to the neighboring areas. I.e. the small settlements have different score when they are near big towns compared to the settlement of the same size in the low population density areas. Compare the results with the population density map Fig. [5] of the Czech republic where you can see some major towns and some density data from the Czech statistical office.

For the evaluation of the obtained results we have worked with the historical data of Česká spořitelna a.s. We estimate a simple regression model using different outputs of the PageRank algorithm and some other variables from the collaterall database we have at our disposal. The target is defined as

$$\ln\left(\frac{\text{payment\_sell\_price}}{\text{nominal\_value}}\right). \tag{7}$$

The main aim is to see which of the constructed Page Rank scores give the best results. The second is to see the significance of other variables at our disposal. The used technique is a linear regression. The reason for this is mainly the simplicity of the model and easy interpretation for the business department. Other reason is that the results gained by different robust methods we have inspected give comparable results hence we choose the most simple approach. We investigated OLS, LTS and M-estimator methods. Note that the observations were subject to data checks first and we have excluded some outliers before the main analysis.

We present the final estimation of the model which is yet to be tested further and calibrated in coordination

Figure 1: Visualization of Page Rank Results: Note the different levels of score in homogeneous areas

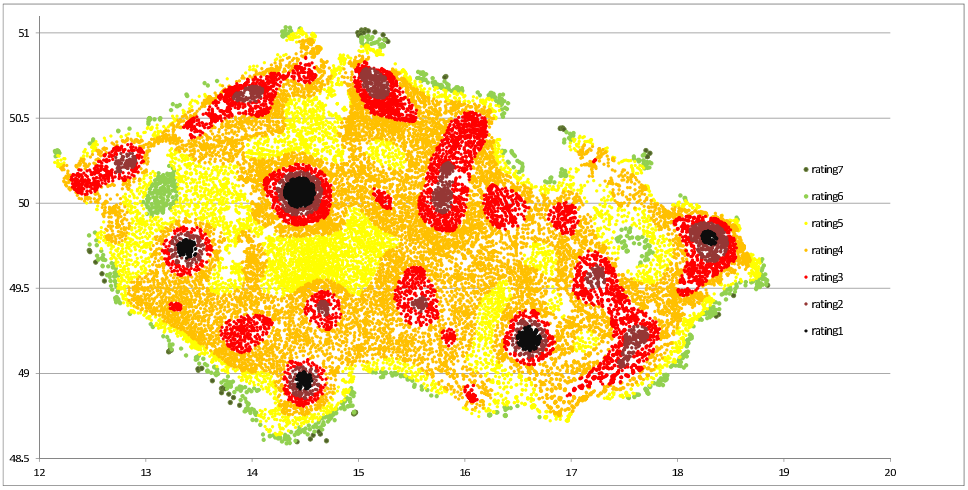
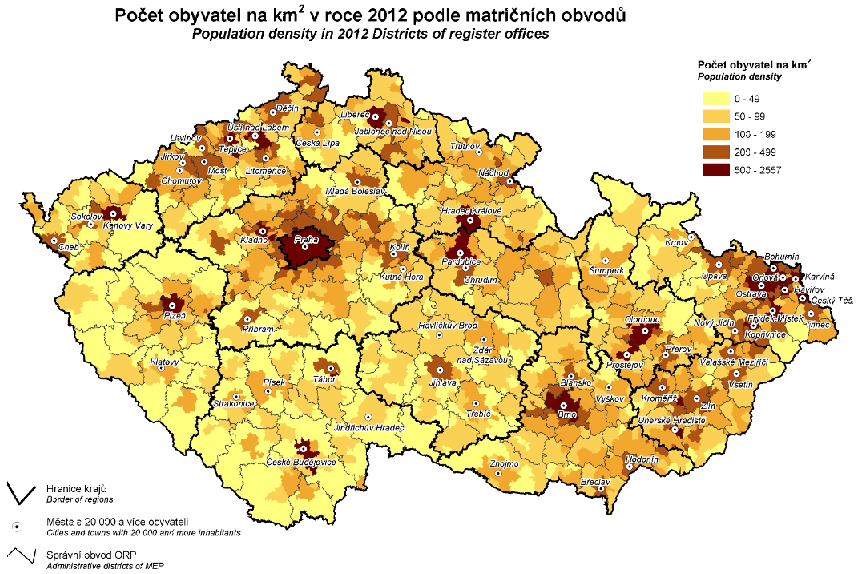


Figure 2: The Czech republic: Towns and densities



with the business department. Note that there are some transformations of the predictors which are not described in detail.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	Pr > |t|
Intercept	0.24986	0.34502	0.4692
YN_flat	0.15364	0.04500	0.0007
nominal_price	-0.08266	0.02491	0.0010
PageRank	0.28458	0.07871	0.0003

The setup of the Page Rank with the highest $R^2 = 19.85\%$ and high level of significance $p - value = 0.0003$:

Parameter	Value
T	30km
σ	10
d	0.85
iterations	101

YN_flat we have used the dummy variable to distinguish between flats and houses. The flats are easier to sell hence the loss is smaller in that case.

nominal_price the (transformed) nominal price determines the target group to which the property can be sold. The more expensive houses, flats are significantly harder to sell hence the loss is bigger in this case.

PageRank the real estate which is situated in Microregion with better Page Rank score which we describe as the attractiveness of the respective region is significantly easier to sell. The attractiveness (Page Rank) quantifies the location where the loss is more likely to appear.

6 Summary and Conclusions

We have shown that our adaptation of *Page Rank algorithm* is useful for the quantification of location of the real estates in the Czech republic. The *regression model* can be constructed giving reasonable results for the forced selling of such properties. One can gain further inside about the parameters that are used to define the Page Rank algorithm and results yielded (thresholds, damping factor,etc.) for the respective environment (city, state, country,...).

Based on the output we have seen we propose that different LTV levels are introduced based on the constructed model which will reflect the different possible losses in different areas and help to avoid some predictable losses when the forced selling of the real estate is done. This approach is yet to be tested along with the bussines department.

We are aware that the results can still be improved. One can focus on the regression analysis with different transformation of target and/or predictors. Different regression approaches can be also examined.

In the outlook we would like to incorporate exact GPS location for the Page Rank score variables construction. We also plan to examine how much added value could be gained by adding some market information from the real estate offers posted on the internet.