MS-E2177 - Seminar on Case Studies in Operations Research

Assessing Concentration Risk in a Bank's Corporate Deposit Base

Final Report

Case Group SEB

Tommi Huhtinen (Project Manager)
Marlin Jarms
Martin Laukkonen
Vilma Norja
Ville Närhi

June 15, 2024

Contents

1 Introduction

1.1 Background

In March 2023, the collapse of the American bank Silicon Valley Bank (SVB) highlighted critical vulnerabilities within the banking sector. SVB went bankrupt due to a massive bank run triggered by a loss of confidence among its depositors. The bank had extensively invested in long-term U.S. treasury bonds and mortgage-backed securities. Unfortunately, as interest rates began to rise in 2022, the market value of these assets plummeted and the bank liquidated a major part of its securities making a substantial loss of $1.8 billion. [15]

The depositor base of SVB consisted of clients mostly from the technology sector, creating a high concentration risk. This concentration meant that SVB's financial health was heavily reliant on the stability and confidence of a single industry. Also, most of the deposits were uninsured as FDIC (Federal Deposit Insurance Corporation) deposit insurance, that is meant for everyday bank customers and maxes out at $250,000. When the bank started reporting significant losses, panic among the depositors followed, leading to a swift and devastating bank run. Remarkably, this culminated in a staggering $42 billion being withdrawn in just a single day. [15]

To learn more on the topic, our client, SEB, is keen on exploring if there is a best-practice approach to effectively measure and manage concentration risk within their corporate deposit base. Concentration risk refers to the potential negative impact of having a high exposure to a particular sector, country, or variable within a portfolio. For instance high exposure to big customers volume-wise [12]. This risk is not only relevant for banks but also poses significant vulnerability to other industries. For instance, a business that relies solely on supplies from a specific country or customers from a single industry is inherently vulnerable to concentration risk.

1.2 Objectives

The primary objective of our project is to provide SEB with methodologies to accurately assess the concentration risk associated with their corporate deposits. We are conducting a thorough analysis on this topic, utilizing mathematical models and empirical research to develop robust methods for quantifying this risk. Our approach is designed to be practical and applicable in real-world banking operations. Throughout this project, we have explored various perspectives on managing concentration risk. We also familiarize ourselves with existing approaches.

Our secondary objective is critical evaluation and analysis of the different approaches. That is, comparing different approaches and finding the most suitable ones. Our approach is enriched by external insights, such as those from the paper [8], introduced by SEB. This paper examines the implications of deposit concentration among Norwegian financial institutions and offers methods that can be adapted and applied to our context. By integrating these insights with our expertise and the active collaboration with SEB, our aim is to establish a comprehensive framework that not only addresses but also captures concentration risk in the banking environment.

Lastly, deriving recommendations to the client based on the results of the project, both in terms of what should be done to measure the concentration risk efficiently and captures the risk itself is also one of the objectives.

1.3 Disclaimer

The conclusions presented in this document represent the views of the authors as students, based on a sample set of data with scrambled volumes. Due to these limitations, these conclusions should not be considered representative of real-world situations. This document should not be used as a definitive source for real-world decision-making.

2 Literature Review

When we began to carry out this project, one of our initial goals was to review literature and methodology in use at banking and other industries to assess best practices related to measuring concentration risk, as well as benchmark our solution against them. While we did find useful information in literature, it eventually turned out that the amount of literature regarding the topic was rather limited. While concentration risk itself is a thoroughly researched topic and there exists plenty of methods to measure it [7], [14], most methods were not directly related to the concentration risk of deposit portfolios, instead focusing more on concentration risk in credit portfolios. Additionally, many of the existing models were quite complex and it would most likely not have been feasible to understand and implement them in the given time frame. This forced us to think a bit more out of the box when reviewing possible solutions for the problem. The approaches we came up with are discussed in further detail in Section 3.

After internal discussions and discussions with the client, we decided to focus our literature review mostly on the approach presented by [8]. Rather than providing us with a clear-cut model to quantify concentration risk however, the article more so provided us with a deeper understanding of the problem, as well as clear ideas and methodology to quantify concentration risk within the deposit base.

Among the ideas the paper [8] provided us, the Gini coefficient and Lorenz curve stand out. While they are quite common frameworks in finance [5], the paper provided us more insight on how to apply them in the context of concentration risk. The Gini coefficient is a measure of statistical dispersion to represent inequality on a scale from 0 to 1, where a value of 0 reflects perfect equality among all individuals (i.e. all values are equal), and a value of 1 reflects maximal inequality among groups (i.e. a single individual holds everything) [6]. The Gini coefficient is explained in further detail in Section 3.1.

The Lorenz curve is a graphical representation of income or wealth, originally created to represent inequality of the wealth distribution [4], [11]. Juelsrud [8] uses the Lorenz curve to represent the distribution of deposits across the aggregate deposit distribution. Figure 1 below presents us an illustration of the Lorenz curve. In the figure, the line of equality refers to perfect equality between all groups, where the cumulative share of earned income increases at the same rate as the cumulative share of people from lowest to highest incomes. The Lorenz curve then depicts the actual distribution of wealth, where we can notice that the share of income rises exponentially as the incomes increase. The Gini coefficient can also be graphically connected to the Lorenz curve as follows: $G = \frac{A}{A+B}$, where A is the area between the line of equality and the Lorenz curve, and B is the area between the Lorenz curve.

Furthermore, the paper [8] provides us the framework for the model which yields us some of the main results of the project later on in Section 4. The model is such that we use a linear regression model of the form

$$\Delta\text{Deposits}_t = \beta \cdot \Delta\text{Deposits of Cluster}_t + e_i, \tag{1}$$

where e_t are the error terms and β is the regression coefficient.

The idea of the model is to assess how much of the total change (variation) in the full deposit portfolio is captured by the total change of the depositors within a cluster. For instance, one can select the cluster to be the largest 500 customers by size, and then evaluate how much of the total change in the deposit portfolio is explained by the total change of deposits among those 500 largest customers. The β-coefficients and the R^2 scores can then be assessed to see how well each cluster manages to capture the variation of the total deposit base. In the paper [8], it is noted that for the

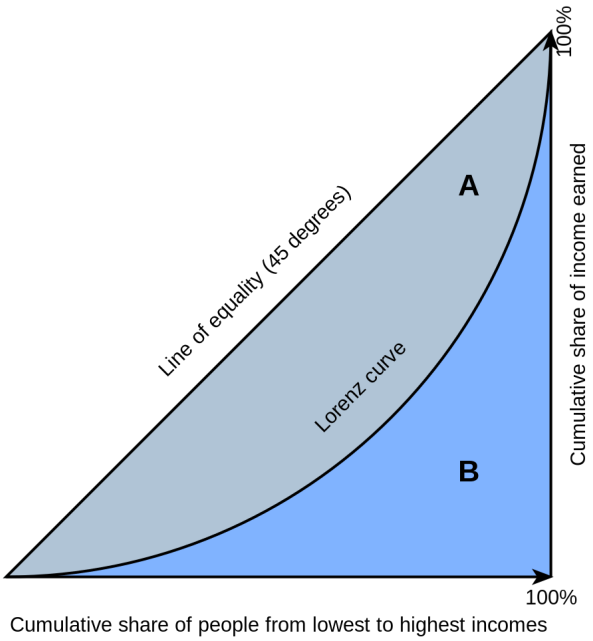

Figure 1: Illustration of Lorenz curve [17].

data at hand, the 5% largest depositors explain approximately 88% of the total flows of deposit flows for the financial institutions. Indeed, we look to accomplish similar results that would showcase how much of the total changes in the deposit portfolio is explained by a certain cluster. The clustering approach additionally allows different profiles of depositors to be explored, where some might prove more interesting than others. More precisely, depositors that are volatile/unpredictable could especially be of considerable interest.

3 Data & Methods

As a crucial component of our project, we initially received three datasets from the client, each labeled according to the date of data collection — specifically, dates within the years date1, date2, and date3, respectively. These datasets each contain identical variables and represents a sample from the bank's deposit portfolio with scrambled volumes on those specific dates.

However, we deemed this data from three dates to be insufficient for our case, and hence requested more detailed data from the client. The client then provided us with daily data, which contains a sample dataset with scrambled balances during 34 different dates from some time span.

Variable	**Description**
Date	The specific date for which the deposit portfolio was captured
Type	The type of the depositor (corporation, municipality etc.)
Product	The product involved in the deposit (different products might have different terms)
Country	The country of the depositor
Currency	The currency the deposit was made in
CustomerID	The ID used for that specific customer (worth noting that each customer can have multiple accounts)
Industry	The industry in which the depositor operates
OriginalAmount	The amount of the deposit in the original currency
ExchangedAmount	The amount of the deposit, standardized to SEK

Table 1: The variables contained in the datasets.

Each sample dataset therefore has nine different variables, which are described by Table 1. As the dataset is sampled, the data is anonymous and each observation is simply characterized by an arbitrary CustomerID. The dataset date1 has 38439 observations, date2 has 17717 observations, date3 has 19311 observations, and the daily data has 677059 observations.

In the latter parts of this section, we delve deeper into the methods used in the project and how they are used to measure concentration risk. Section 3.1 introduces the static approaches used, i.e. methods that only consider one date. Section 3.2 explains the dynamic approach we used (similar to the clustering approach of [8]) which is largely the opposite of the static approach, as in we now consider data from different dates instead of just one. In Section 3.3 we then provide intuition into how a combination of the static and dynamic approaches could provide an effective measure of concentration risk. In Section 3.4 we disembark from the approach of [8] a bit, as we discuss aspects of concentration risk that the clustering methods are not able to cover, which potentially pose greater interest with regard to concentration risk than the changes in deposit volumes. Lastly, Section 3.5 delves into scenario analysis techniques used to assess the impact of negative events on the concentration risk of the deposit portfolio.

3.1 Static Approach

To assess concentration risk we took a look at static approaches. In our case, static approach means that the methods give us insights into the concentration risk by looking into the data at one point in time. Static approaches we considered were Gini Coefficent, HH-Index (short for Herfindahl-Hirschman-Index), Bayes' Network, and Social Network. The Gini coefficient is calculated in the following way, where n is the number of clusters, i, j are indices for the clusters, and x_i is the volume of each cluster [6]:

$$G(x) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i} \tag{2}$$

To assess how concentrated the SEB sample portfolio concerning the industries of the customers is, we would use the Gini coefficient in the following way: Let n be the number of industries SEB is exposed to and let x_i be the volume of deposits coming from customers of industry i. To assess concentration risk concerning any segmentation works analogous. Possible other ways of segmentation are segmenting by country or using a discrete segmentation (every customer is its own segment). The HH-index works quite similarly to the Gini coefficient. It is computed in the following way, where n is the number of clusters, i, j are indices for the clusters, and x_i is the volume of each cluster [1]:

$$H(x) = \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i)^2} \tag{3}$$

Assessing the concentration with the HH-index works similarly to assessing concentration with the Gini coefficient. The difference between those two is that the HH-index evaluates concentration based on volume. So the HH-index is high, if there are some clusters with a high volume. The Gini coefficient measures inequality. So the Gini coefficient is high, if the volume of clusters differs by a significant amount [1], [6]. Additionally, one may consider using the normalised HH-index, it is computed in the following way:

$$H_{norm}(x) = \frac{n \cdot H(x) - 1}{n - 1}, \tag{4}$$

where $H(x)$ is the HH-index of x and x, n are defined as described at (3). The normalized HH-index has the following properties compared to the HH-index [1]: The HH-index lies between $\frac{1}{n}$ and 1, where $\frac{1}{n}$ stands for equal distributed depositor Base and 1 stands for being concentrated in one area. The normalized HH-index lies between 0 and 1. where $H_{norm}(x) = 0$ if and only if $H(x) = \frac{1}{n}$.

So overall the Gini coefficient and the HH-index are helpful tools to estimate concentration and they are easy to compute, but they have some limitations when it comes to estimating the concentration risk: The problem with these is that it does not include any information about industries or countries. For example, being concentrated in one country provides the same score as being concentrated in a different country. But it does not mean, that those two things are not necessarily equally risky, because they are not comparable. But overall the Gini coefficient has proven to be reliable and easy to compute, which makes it an appealing measure to compute concentration. There are more sophisticated approaches, for example, Bayes' network, or adapting the Gini coefficient in a way that considers the risk behind a deposit by assigning a probability of client's withdrawal. Those approaches have a disadvantage. We will need expert input from

the client or other experts in the field, as in [13], where they modelled credit risk using bayesian networks. By using Bayes' network we aim to estimate the likelihood, that a certain amount of money gets withdrawn from the customer base, and for that we need expert information on the prior. In the case where we adapt the Gini coefficient, we need expert input on how risky it is to be concentrated in a certain industry or country. Combining the expert input with the Gini coefficient can be done by analyzing the Gini coefficient. If the Gini coefficient is high we need to analyze which sectors/countries are the main driver for the high Gini coefficient. With the expert input we can analyse the main drivers.

3.2 Dynamic Approach

Dynamic approach in our case means that the data points of the deposits are from different time points, i.e., from different dates. We have a month of sample daily data of SEB's corporate deposits and we use linear regression to see how well different clusters of the depositors represent the variation of the whole deposit base. This could be analysed with the R^2 value and β estimate. In the dynamic approach we have been using the linear regression model (1) introduced in Section 2. The beta is estimated as stated in the following formula [2]

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where x_i represents the values of the independent variable. $\overline{x}$ is the is the mean (average) of the independent variable. y_i represents the values of the dependent variable. $\overline{y}$ is the mean (average) of the dependent variable. n is the number of observations. In our case the x values are the deposit changes of the cluster (ΔDeposits of Cluster$_t$) and y values represent the total deposit changes (ΔDeposits$_t$).

The regression line is fitted based on training data and the R^2 score is calculated based on testing data, so we see how well the fitted regression line is able to predict unseen data. The formula for R^2 is [2]

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$
$$SS_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$SS_{\text{tot}} = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

where SS_{res} is the sum of squares of residuals and SS_{tot} the total sum of squares. And y_i are the observed values, $\hat{y}_i$ would be the predicted values from the regression model and $\overline{y}$ is the mean of the observed values y. In our case with the training and test set the $\hat{y}_i$ is the test data point which is compared to the fitted line. In the most traditional sense the R^2 score ranges between 0 and 1. As we have split the data into training and test set, we can actually have negative values for R^2 as well. This can be interpreted so that the chosen model fits worse than a horizontal line, i.e., a constant model. By regression we can see how well some cluster, for instance larger depositors represent the changes in deposit levels of the whole deposit base. If the R^2 score is close to 1, the cluster represents changes in the deposit levels quite well and if the R^2 score is closer to 0 or negative then the cluster does not represent the changes in deposit levels.

We are using different algorithms to determine potentially most risky clusters that could be connected to the concentration risk. We are doing this to find interesting clusters that possibly cause a concentration risk for the bank. First we use clustering by size where we sort the customers in a descending order. If the depositor has multiple accounts, we sum the deposit volumes from different accounts together and only then compare the size based on the column 'TotalVolume' that represents total sum of deposits from different accounts and one-by-one, pick the largest customer as the largest customers represent tail risk of the distribution which is where the most concentration risk is. Then we have volatility-based clustering using the standard deviation of the 'TotalVolume' column, where we sort the customers in a descending order based on their respective volatility, and one-by-one, pick the customer having largest volatility. These two algorithms can also be used to find the smallest customers by sorting the customers in an ascending order. Utilizing volatility computation we can quantify the variation in deposit sizes over time. Additionally we have the size-to-volatility ratio where we pick the largest customers in terms of size-to-volatility ratio. The aim on this one is used to find larger customers with relatively low volatility. Relative to the previous we have the volatility-to-size ratio, i.e., picking the most volatile customers in terms of volatility-to-size ratio. Should give us small customers with high volatility.

In general, one important advantage in the dynamic approach is that we can see how certain customers behave over time. For instance, if they withdraw their money regularly or not. This can possibly be used to forecast the future behavior of the customers. Our model is acting as one of the dimensions of concentration risk, assuming that some cluster explains the total deposit flows well, we could argue that this cluster is concentrated. Also this approach adds more dimensions to the assessment.

This brings us to the weaknesses: as we know historical development, does not always act a good indicator for future deposit movements or account for rare, extreme events. Second, the changes in the total deposit volumes is only one dimension of concentration (risk), and not necessarily even the most interesting or riskiest, as later discussed in Section 3.4. Moreover, the dynamic approach can be harder to interpret than the static approach.

3.3 Combining the Two Approaches

There are also possibilities to get insights about the concentration and the concentration risk by utilizing static and dynamic methods together. In particular, we are interested in combining the Gini coefficient with results from dynamic approaches. Until now we used Gini to evaluate the concentration of the sample depositor base provided by SEB concerning countries and sectors. With the insights from dynamic approaches, it is possible to classify customers (see 3.4 for more details on classification). So as described in Section 3.1 we can use Gini to evaluate the concentration of the sample depositor base provided by SEB concerning the classification from the dynamic approach. See table 12 for the results.

To get more insights on the sample data provided by SEB we track the Gini coefficient over time. This also tackles the problem described in Section 3.1, that it is hard to derive how risky a certain concentration is. But in general, we can say that a more concentrated depositor base than a less concentrated depositor base is riskier, whereas both are concentrated in the same area. So by tracking the Gini coefficient we can detect if the concentration increases compared to the previous level.

Combining static and dynamic approaches gives us some advantages. First of all introducing a dynamic approach into Gini (as described above) can be interpreted as replacing the expert input

described in Section 3.1, for example through the volatility. Additionally the results we get from the described methods in this chapter are not as hard to interpret as the described time series approach in Section 3.2.

3.4 Classification

We can order the depositors based on their volatility or size, and then form clusters by taking for instance the top 1%, 5%, or 10% of the most volatile or largest depositors. As later illustrated in Section 4.2, these clusters explain the changes in the total deposit volumes relatively well. Also, the respective algorithm seems to work as expected, as by increasing the cluster size the R^2 score improves (see Table 3). In other words, we can develop some level of understanding of how the changes within the deposit are concentrated. However, it can be argued that having and knowing in particular the volatile depositors within the deposit base is in fact the edge of the bank and hence not the most interesting cluster related to the concentration risk. Namely, once the bank identifies the behavior of the volatile depositors, i.e., the pattern of when they withdraw and deposit their money, it can utilize the historical pattern in time series forecasting models. Then, by having a good knowledge of the expected future cash flows, the bank can position itself accordingly.

As our focus had been on explaining volatility by different clusters, and based on the above, we were informed there might be other aspects to consider when assessing the concentration risk. Consequently, we are also interested in finding depositors that are not only less volatile but also in relative terms larger in size. These kinds of depositors pose another type of risk to the bank, as by interpretation of non-volatility there is no historical data on they withdrawing their money, but they can still do so at any time in the future, and due the relatively larger size of such depositors, the amount of withdrawn money can be material. This creates an opportunity cost for a bank, as it can not naïvely assume that all of these depositors will remain also in the future, but it rather has to be prepared for the possible withdrawal. The key question is how conservative standpoint the bank needs to take with respect to these larger and less volatile depositors. On the other hand, the less volatile depositors that are in relative terms small in size pose only a small risk to the bank, given that the interpretation of a "small" depositor is such that even if multiple of these small depositors withdrew all of their deposited money simultaneously at an unpredictable time, such deposit outflows causes effectively no trouble for a bank.

Given all of this motivation, we next introduce the methodology to classify the depositor base into the following three categories: larger and less volatile, larger and more volatile, and smaller depositors. That is, we do not further classify the smaller depositors based on their volatility due to the inherent diversification effect from such depositors. We start the classification by first dividing the depositors base into larger and smaller depositors, because this can be done by looking at the deposit volumes on a certain date and without knowing anything about the past behavior, and from this point of view is a more fundamental property than volatility. For this purpose, we interpret the size of a depositor in an identical manner as in the algorithm to pick depositors by their size (see Section 3.2), i.e., if a depositor has multiple accounts, the size of the depositor equals the total deposit volume in all of these accounts. To the question which of the depositors are larger and which smaller, there does unfortunately not exist a clear-cut answer, as the deposit volume is a continuous variable. Instead, we seek an answer to this question by utilizing a couple of different techniques. These techniques include plotting the deposit levels of the customers both in linear and logarithmic scale, as well as plotting the cumulative share of total deposits that a set of customers corresponds to. In all of these plots, we order the depositors based on their respective sizes in a

descending order. By utilizing the input from all of these plots, we classify each depositor to be either larger or smaller.

Now that we have an approach to identify all larger customers within the deposit base, which ones of those are less volatile and which are more volatile. Again, we interpret the volatility as described in Section 3.2. As a starting point, it can be said with certainty that depositors having zero volatility are less volatile. To determine if any of the remaining larger depositors having volatility greater than zero can also be considered as less volatile, we utilize a similar approach as for determining the larger depositors. That is, we order the depositors in a descending order based on their volatility, and plot the respective volatilities both in linear and logarithmic scale. In addition, as it is a bit complicated to determine the volatility or stability of a depositor based on the numerical value only, we also plot the time series of deposit levels of a set of customers. This allows us to adjust the numerical threshold for "less volatility", as we can visually confirm whether a certain numerical value of volatility appears less volatile or not. We were also given an example of larger, less volatile depositors by SEB, so that those can be used as a reference to validate our findings.

3.5 Scenario Analysis Techniques

One of the objectives of our project is to assess the impact of running a high concentration risk in an adverse event by utilizing scenario analysis techniques. In addition, scenario analysis techniques could be leveraged to study the risk exposure of larger and less volatile customers. Since these customers are less volatile and thus do not provide information about their behavior, scenario analysis can help in evaluating the risk they pose to the bank. More precisely, we can examine how the withdrawals of these depositors affect the bank.

The scenario analysis technique that we employ is an agent-based model, wherein either individual customers or groups of customers are treated as agents that interact with each other. Agent-based models are utilized to study the behavior of a system of agents interacting with each other. These models have proven to be effective in the financial sector for modeling various cascading effects or economic events such as banking crises [9] or housing and mortgage markets [3]. Our model consists of agents whose primary action is to withdraw money. The underlying concept of the model is that customers or groups of customers in close proximity may withdraw their money simultaneously due to shared backgrounds and needs, or as a result of panic spreading among interconnected actors.

Using the model, we simulate how rapidly the bank would lose material volumes when one agent initiates a withdrawal and triggers subsequent actions by other agents. The principle of the model is as follows: one agent withdraws their money first, followed by the agent closest to them deciding to do the same. Subsequently, the next closest agent repeats this action until either there are no more proximate agents or the bank goes bankrupt, i.e., the amount of money withdrawn exceeds a predetermined threshold. If there are no more proximate agents, we move to the second closest agent and repeat the process. This continues until the bank has lost a material volume or there are no more proximate agents left for any of the agents close to the first agent. The number of iterations it takes for an agent to make the bank lose material volumes can be analyzed and compared.

Proximity is determined using two distinct strategies. The first one relies on the Hamming distance (see e.g. [16]). The Hamming distance is calculated between two depositors' attributes and is thus a static approach that requires only one point in time. The distance is calculated as follows: if two customers have different attribute values, the distance increases by one. The

attributes considered include the customer's type, country, and industry, the product they have, and the currency. Therefore, the distance between two depositors ranges from zero to five, where zero indicates identical attributes and five indicates that none of the attribute values match. This distance measures assumes that agents are similar to each other and would act similarly if they share similar attributes.

The second approach measures distance based on the Pearson correlation coefficient between two customers. This method is dynamic as it involves the computation of correlations from the time series data of the agents. Pearson correlation coefficients are computed using the formula [10]:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5}$$

where (x_i, y_i) are the deposit volumes of two customers at time point i and $\bar{x}$ and $\bar{y}$ are the averages of deposits between time points 1 and n. Correlations can be computed also between each cluster of customers. Clusters are formed based on the attributes of depositors, such that depositors with identical attributes belong to the same cluster. Pearson correlation coefficient as a distance measure assumes that agents are similar to each other and would act similarly if they have previously acted similarly.

Utilizing these two distance measures and the agent-based model, we can run simulations for each agent and observe the number of iterations needed for the bank to lose material volumes. If the bank loses a material volume rapidly, it suggests that the initiating agent is closely linked to numerous and/or relatively large agents. By comparing the iteration counts, we can investigate whether agents with certain attributes require fewer iterations than typical, indicating a potentially higher concentration risk.

4 Results & Discussion

4.1 Overview of the Deposit Base

This section provides a brief, yet comprehensive, overview of the sample data through graphs and numerical values. Namely, we are interested in any concentrations in the sample data.

Figure 2 represents the development of the amount of total deposits in the deposit base as a function of time. Perhaps most notably, we can notice a drop in the deposit levels around date 17-21. There is a quick resurgence afterwards, which is then followed by a steady downwards trend after date 25. As the graph is normalized, it illustrates the relative changes in the data clearly, especially the effect of the downwards trend.

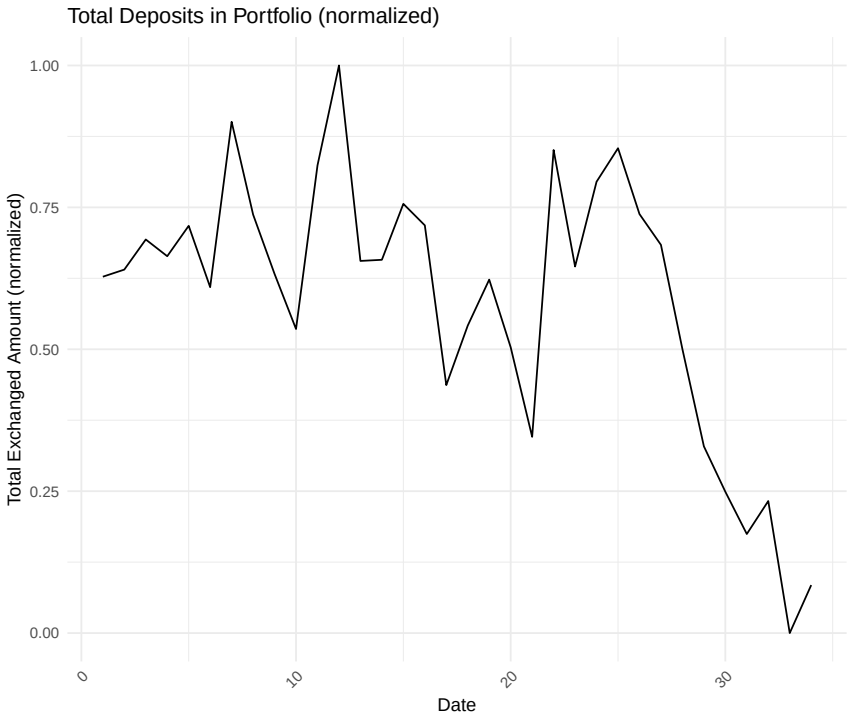

Figure 2: The development of total deposit amounts in portfolio

Table 2 showcases the Gini coefficients of different variables in the data at the last date of the daily data (date34). We see that the coefficients are quite high for some variables, such as currency, industry and CustomerID. While more context about the observations would be required to make decisive conclusions, intuitively these results make sense. For example, there are 45 unique currencies in the dataset at date34, and it is likely that most of the deposits are made in only a few currencies (e.g. SEK, EUR, USD, GBP, NOK), yet it is logical that there are some deposits made in other currencies as well due to the global presence of the client. The same idea applies for industries and countries as well; it is logical that the deposits are concentrated in a few industries but there

are deposits from less common industries as well. While the high Gini coefficient of CustomerID

Variable	**Gini Coefficient value**
Currency	0.920
Industry	0.875
CustomerID	0.862
Type	0.794
Country	0.706
Product	0.633

Table 2: Gini coefficients for the different variables on date34.

might initially seem a bit odd, we can see in Figure 3 that the sample data is very much skewed in that there are a few relatively large depositors, while the vast majority of the deposits are more evenly distributed. The scale of the graph is also quite distorted due to the few relatively large depositors; the scale makes it seem so that most the deposits are around 0 when that is not the case. All in all, the graph manages to explain the relatively high Gini coefficient of CustomerID quite well.

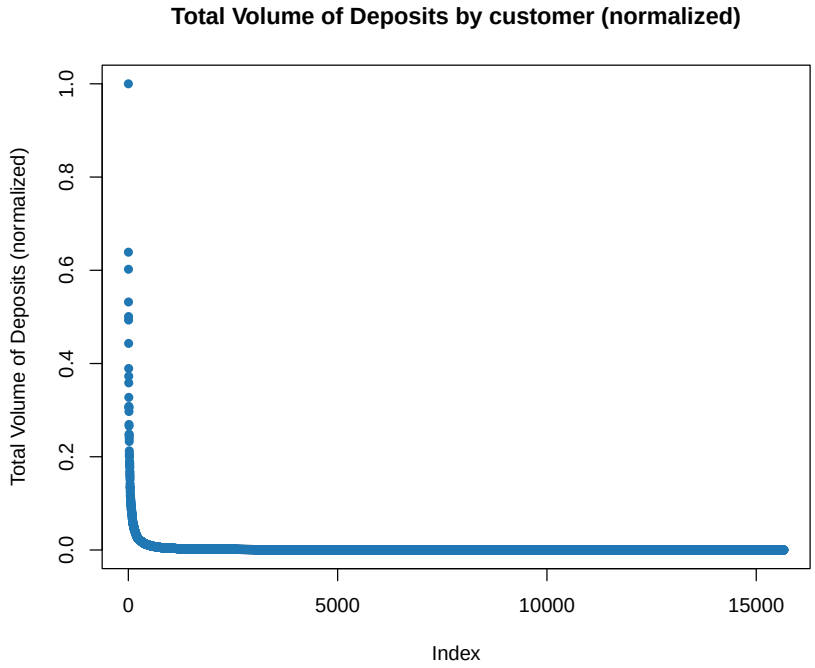

Figure 3: The distribution of deposits (highest to lowest).

Figure 4 below depicts the development of the Gini coefficient of the deposit base over time through dates 1-34. We can notice a similar trend between Figures 4 and 2. Indeed, when calculated,

the correlation between the time series present in these figures is around 0.7. When considering how the Gini coefficient is calculated (Equation 2), this correlation is to be expected. For instance, whenever the Gini coefficient increases, it usually indicates existing customers have deposited more into the deposit base and vice versa when the Gini coefficient decreases.

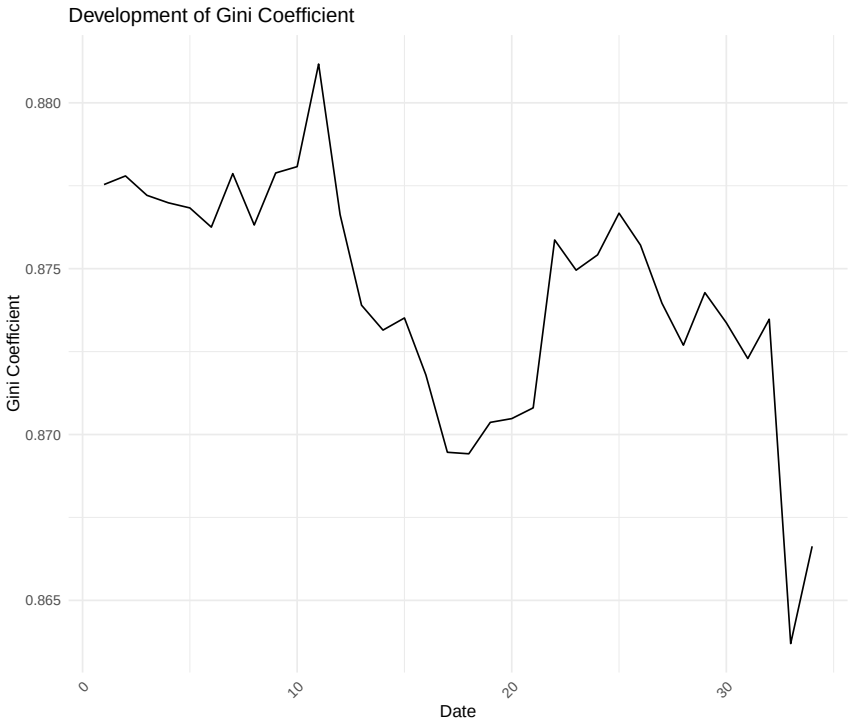

Figure 4: The development of Gini coefficient over time for the deposit base.

4.2 Explaining the Total Deposit Flows

We started clustering the data with different algorithms presented in Section 3.2. Size and volatility got us the highest R^2 values. At first we had data with only three different time points from different years and got quite high R^2 scores as even 98.6% for the top 5% and 83.3% for the top 10% of the depositors by size. Meaning that the bigger top 5% depositors seem to be the most important cluster regarding the concentration of the changes in total deposits. These results seem to be not that accurate as the top 5% has a higher R^2 than the top 10% which does not seem logical as by increasing the depositors from the total deposit base we should also be able to explain the total changes better. Also the R^2 scores are really high. The reason for these things is most likely that there are only three time points of data. That is why decided to use monthly data with multiple time points for this approach.

For the new data as daily sample data with 34 different time points. We clustered this data by size and used thresholds as the top 1% top 5% and top 10%. We concluded that the top 10%

depositors by size represents best the variation of the whole deposit base with an R^2 score of 93.6%. But the top 5% was now decreased to 83.4%.

We got better results when we used the volatility as a measure with the monthly data than with the yearly data for the top 5%. For the top 5% of most volatile customers we got an R^2 score of 93.5% compared to the value computed using the size 83.4%. This tells us that volatility seems to explain the selected cluster for us more accurately than using size. This can be seen below in Figure 5 where the picture on the left is the regression line by deposit volume and the one on the right by volatility.

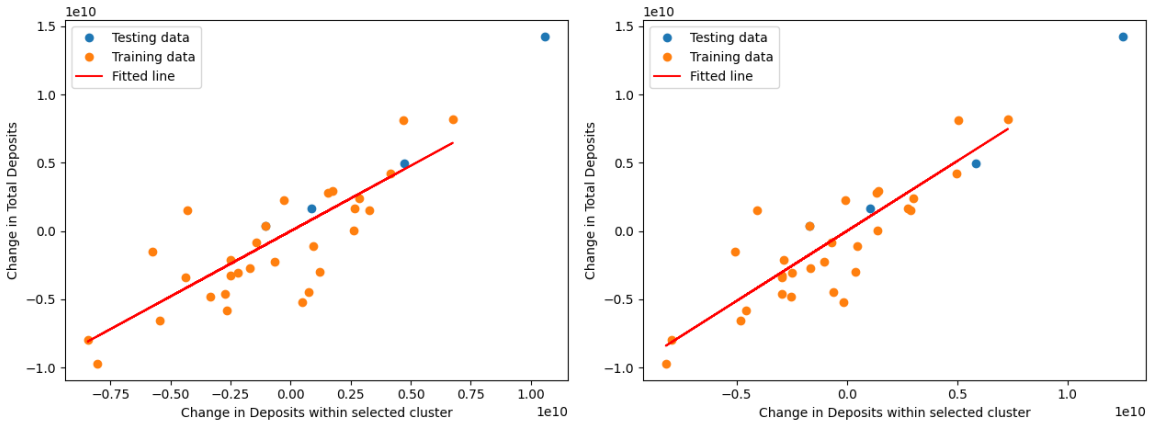

Figure 5: Regression analysis of monthly largest 5% of the customers in terms of deposit volume (left) and deposit volatility (right)

Algorithm \ Fraction of depositors	Top 10%	Top 5%	Top 1%
Size	0.936	0.834	0.663
Volatility	0.929	0.935	0.898
Higher size-to-volatility ratio	-0.997	-1.11	-0.900
Lower size-to-volatility ratio	0.945	0.890	0.631

Table 3: R^2 values with the different algorithms.

In Table 3 we have different values for R^2 computed using the different algorithms described in Section 3.2. The goal is to find different clusters and see which ones represent the variation of the total deposit base the best. For instance, high size-to-volatility ratio is being used with the intention to give us larger customers with relatively low volatility and low size-to-volatility ratio to find smaller customers with higher volatility. As we can see, larger customers with relatively lower volatility does not represent the variation of the deposit base well at all, as the R^2 is clearly negative. The reason for the negative R^2 values is explained in Section 3.2. However, lower size-to-volatility ratio represents the changes in the deposit base quite well with an R^2 score of 0.945

for the top 10% of the depositors. According to these results, we conclude that the larger, most volatile and volatile smaller customers seem to represent the variation of the whole deposit base the best.

To further validate the performance of these algorithms, we also computed the R^2 score as a function of a cluster size for cluster sizes ranging from 1 to 50. That is, we chose depositors according to the aforementioned algorithms one-by-one from a subset of the whole depositor base consisting of the largest 300 depositors, and included the chosen depositors into respective clusters. At each iteration, we also computed the corresponding R^2 score. In addition, we incorporated the so-called brute force approach to supplement the analysis. The brute force approach means that at each iteration, we searched the whole considered subset of depositors, and included into the corresponding cluster the depositor that increased the R^2 score the most. Thus, the brute force approach illustrates how good R^2 score is practically achievable with a given cluster size.

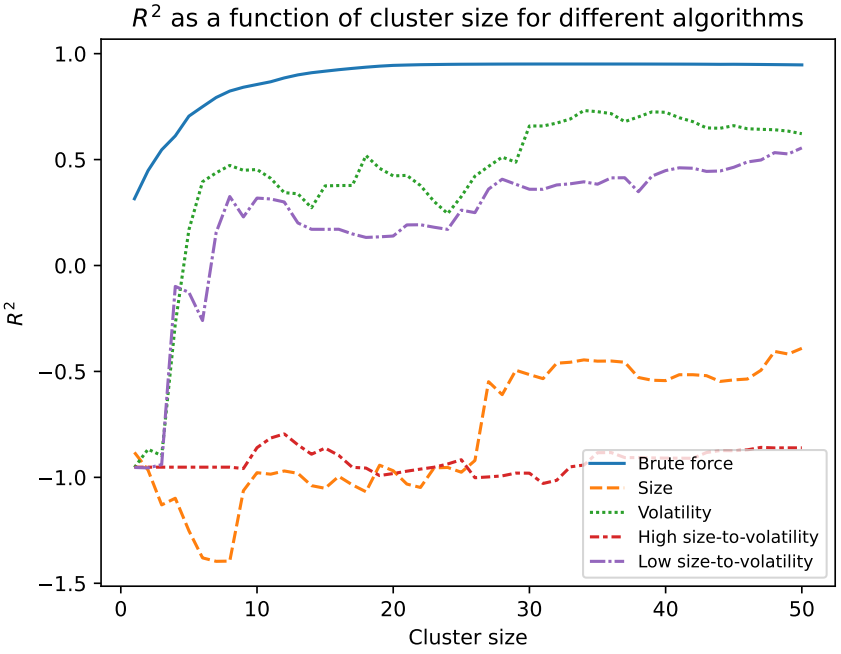

Figure 6: R^2 as a function of cluster size for the discussed algorithms.

Figure 6 visualizes the results of the process described above. Based on Table 3, it appeared that both more volatile and larger depositors explained the changes in the total deposit levels rather well, at least when top 1%, 5%, and 10% of the depositor base were considered. However, based on Figure 6 it now seems that merely larger depositors correspond to low R^2 scores, when cluster sizes between 1 and 50 are studied. This is an interesting finding, as the article from Juelsrud [8] does not discuss alternative ways to explain changes in the total deposits than larger size. Yet, our analysis suggests that high volatility is even better in terms of explaining the changes in total deposits. The reason why this conclusion was difficult to derive based on solely on the results shown by Table 3 is that as some of the most volatile depositors are also relatively large, and as we increase the cluster

size to include top 1%, 5% or 10% of the larger depositors eventually the corresponding clusters will also contain some of the most volatile depositors. Given the current evidence, it appears that these volatile depositors contribute to the good R^2 proportionally more than the merely large depositors.

At this point it should be noted that as such our analysis is comparable to the results presented in [8] only in relative but not in absolute terms. The reason is that we have fitted our regression line based on a training set and computed the R^2 based on a separate test set. To have comparable results in absolute terms, we need to leave behind the split between training and testing data, as according to our best interpretation this is the setup in [8]. Without the training and test approach we get the results in the following Figure 7 and Table 4.

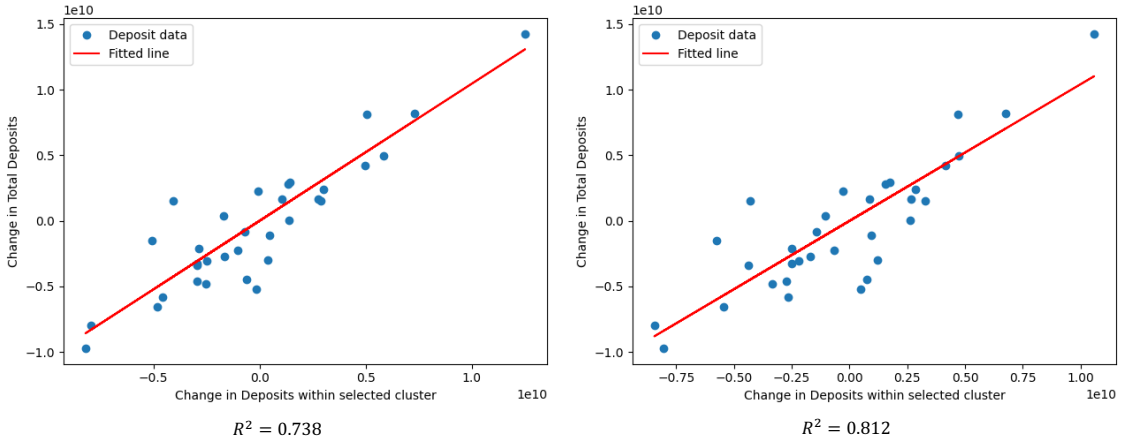

Figure 7: Regression analysis of monthly larger 5% of the customers in terms of deposit volume (left) and deposit volatility (right).

As seen in Figure 7 and Table 4 the values for R^2 are essentially lower than with the training and test split. This is because the chosen points from the test set fit in this case better on average.

Algorithm \ Fraction of depositors	Top 10%	Top 5%	Top 1%
Size	0.826	0.738	0.561
Volatility	0.823	0.812	0.728
Higher size-to-volatility ratio	0.047	0.015	0.032
Lower size-to-volatility ratio	0.790	0.734	0.450

Table 4: R^2 values with the different algorithms without the training and test approach

4.3 Identifying Larger Depositors

In Section 3.4 we explained the procedure to identify larger and less volatile depositors. Now Figure 8 shows the size distribution of depositors in the sample data both in linear and logarithmic scale on date 34 of the daily sample data. The depositors have been ordered based on their respective

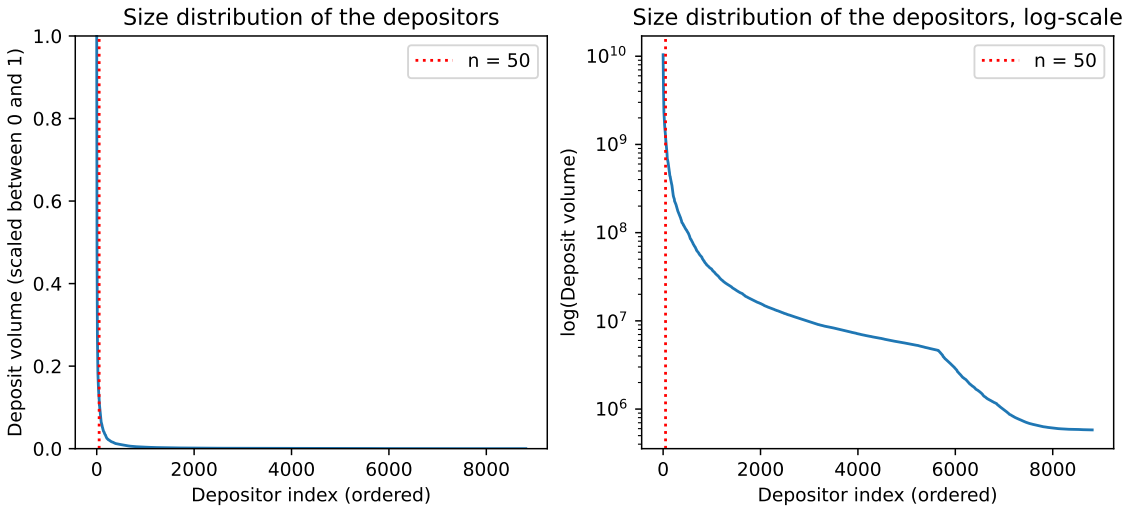

Figure 8: Size distribution of the depositors in the sample data both in linear and logarithmic scale. The y-axis in the linear scale has been scaled between 0 and 1.

sizes from largest to smallest. The y-axis in the linear scale have been scaled between 0 and 1, as the absolute deposit volumes do not affect the relative ordering. We propose that the 50 largest depositors are classified to be larger, whereas the rest of the depositors are considered as smaller.

Let us go through the arguments for this classification. First of all, by considering the 50 largest depositors as the larger ones, almost all of the depositors corresponding to the steep part of the curve illustrated by the left-hand plot in Figure 8, are considered as larger. Respectively, the flat part of the curve is considered to contain only smaller depositors. There are also the some depositors belonging to the steep part of the curve that get classified as smaller. The right-hand plot of Figure 8, i.e., the plot showing deposit volumes in logarithmic scale, helps to further understand the size distribution of the smaller depositors.

On one hand, based on Figure 8 the depositors corresponding to the steep part of the curve get classified as larger whereas the depositors corresponding to the flat part of the same curve get classified as smaller. However, on the other hand, it is much more difficult to say exactly where to place the cut-off point between larger and smaller customers. This is why in Figure 9 we have plotted the size distribution of the 200 largest depositors. Again, the left-hand plot represents the deposit volumes in linear scale and the right-hand plot in logarithmic scale. The logarithmic scale is particularly interesting in this case. There are two different parts in the curve representing the size distribution in logarithmic scale: a curved part and a linear part. Moreover, the cut-off between these two parts of the curve is roughly at index 50, indicating it could be further used as the cut-off between larger and smaller depositors.

Figure 10 illustrates the cumulative share of the total sample data deposit volume explained by the whole depositor base and the largest 200 depositors. These figures we used as the final argument for using index 50 as the cut-off point between larger and smaller customers. In particular, the right-hand plot in Figure 10 shows that the increase in the cumulative share of total volume explained

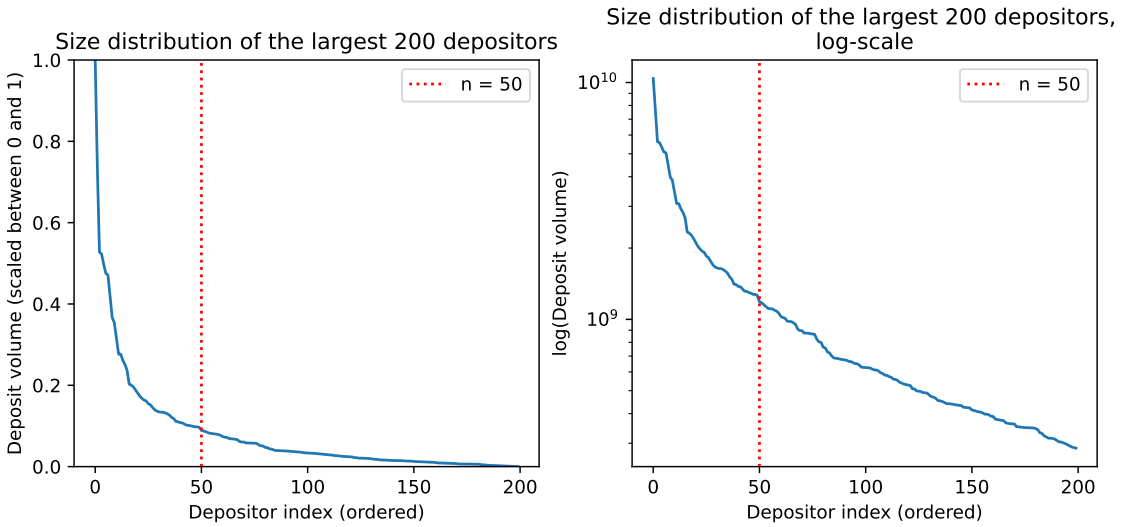

Figure 9: Size distribution of the largest 200 depositors. The y-axis in the left-hand plot has been scaled between 0 and 1, similarly as in Figure 8.

is clearly less after index 50 than it is at or before index 50. Moreover, the left-hand plot in Figure 10 shows that before and at index 50 the curve representing the cumulative share of total volume explained is virtually a vertical line, but after index 50 the curve starts to bend. These observations can be interpreted so that there is change in size of the depositors after index 50. Our interpretation in particular is that the 50 largest depositors are considered to be larger, whereas the rest of the sample data depositor base consists of smaller depositors.

As a last note we mention two things. Firstly, the index 50 as cut-off point is based on date34 only and can vary depending on the study date. Second, the index 50 as a cut-off point between larger and smaller depositors is by no means the only possible cut-off point that could be proposed in the given setup. For instance, one could argue that all of the depositors corresponding to the steep part of the left-hand plot in Figure 8 should be considered as larger depositors. This would roughly correspond to the 200 largest depositors. On the other hand, Figure 10 shows that around 60% of the total sample data deposit volume belongs to the 200 largest depositors, whereas the 50 largest depositors explain around 37% of the total deposits. Thus, by considering the 200 largest depositors, we would be considering the majority of the deposit volume of the sample data, even though we are interested in analyzing the subgroup of the depositors formed by the larger depositors. Moreover, by limiting the size of the group of larger depositors to the truly largest depositors, we are hopefully able to better illustrate the maximal potential of the adverse effects caused by the actions of the larger depositors.

4.4 Identifying Less Volatile Depositors

As described in Section 3.4, the larger depositors themselves do not necessarily pose a significant risk to the bank in a going concern scenario, but combining relatively large size with how those

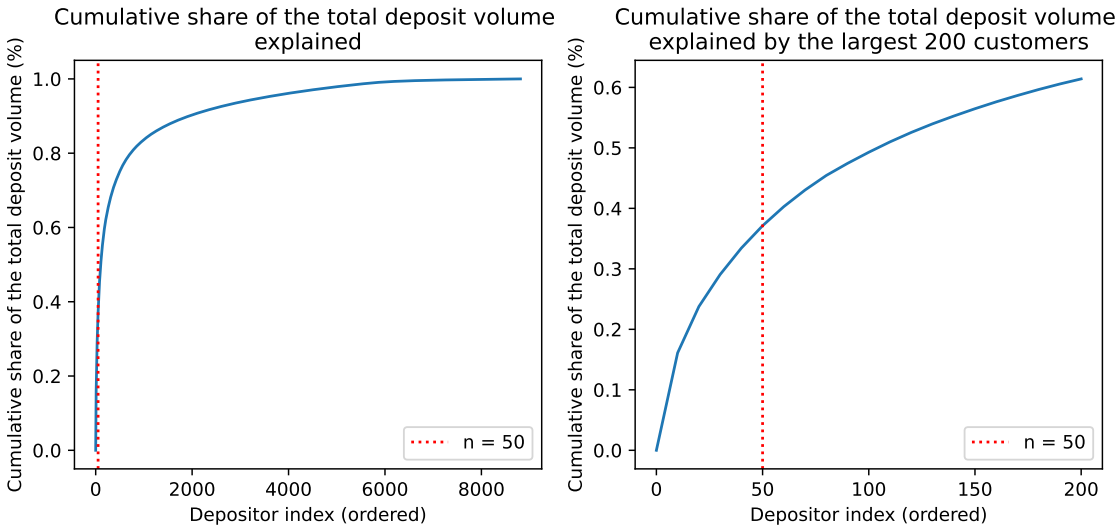

Figure 10: Cumulative share of the total sample data deposit volume explained by the whole depositor base and the largest 200 depositors.

would act in times of stress makes such depositors a source of a risk. Hence, Figure 11 shows the distribution of the volatility of the larger depositors identified in the previous Section 4.3. Here we propose that the 20 most volatile larger depositors are considered as more volatile, whereas the remaining 30 larger depositors are considered as less volatile. The first argument for that choice is that in the left-hand plot in Figure 11 the volatility decreases non-linearly as a function of depositor index before index 20, but after that the volatility decreases roughly linearly. The change in the way the volatility decreases indicates some sort of underlying change in volatility of the respective depositors, and we in particular interpret this change as a cut-off point between more volatile and less volatile depositors.

Second, Figure 12 shows the deposit volumes of the 30 least volatile larger depositors as time series that we identified to be the group of less volatile and larger depositors. Two archetypes can be identified among the less volatile and larger depositors: either the deposits levels remain roughly constant over the given time period or there is one minor bump in the time series at a certain date during the month. For instance, depositors corresponding to customerIDs 65, 109, and 210 fall into this latter archetype. After discussing with the SEB representatives, it turned out the bump is actually related to the industry in which these depositors operate, so there is nothing particularly unpredictable from the point of view of the bank. Moreover, customerID_36 is included into the set of identified less volatile and larger depositors, which is promising, as it was provided to us as a reference by SEB. Overall, there does not seem to be any outliers, so we conclude that we have not mistakenly classified some more volatile larger depositors as less volatile.

The only question that remains is whether we have included all of the less volatile larger depositors into this set of 30 larger and less volatile depositors. To answer this question we tried to include a couple more, the least volatile depositors among the more volatile depositors, to the

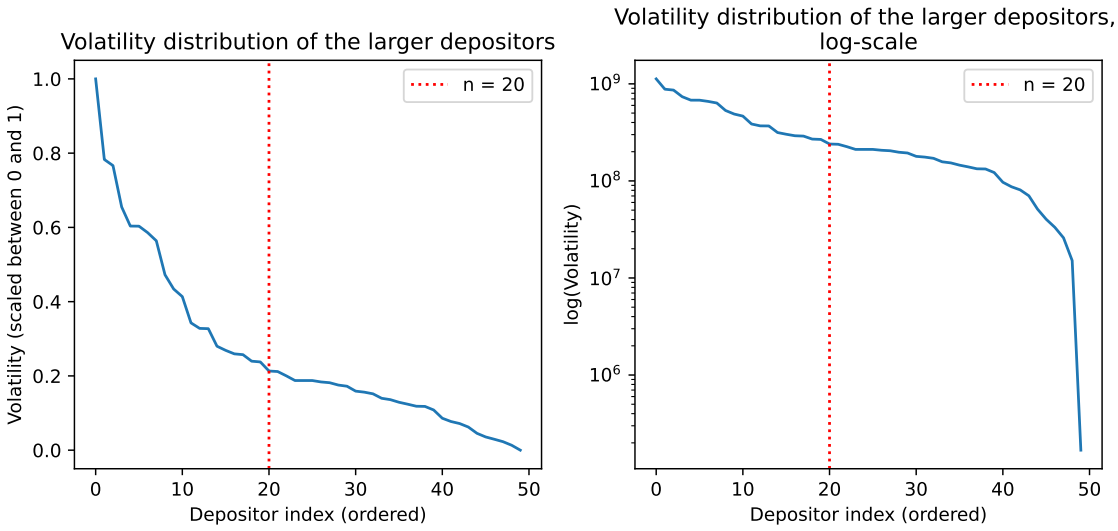

Figure 11: Volatility distribution of the larger depositors both in linear and logarithmic scale. The y-axis in the linear scale has been scaled between 0 and 1.

group of less volatile deposits and plot the deposit levels of these respective depositors as time series. However, they did not quite match either of the two archetypes described above, so those were eventually classified as more volatile. We still note that in a similar way that when identifying larger depositors, some of the depositors that we identified to be more volatile could in fact be argued to be less volatile, but at least the less volatile and larger depositors we identified represent the extreme examples of such depositors within the given sample data. These are analyzed further in Sections 4.5 and 4.6.

4.5 Analyzing Larger and Less Volatile Depositors

As described in Sections 4.3 and 4.4, we have now identified the larger and less volatile depositors in the sample data provided. Figure 12 shows the time series of these depositors, but there are also other aspects that can be analyzed. The analysis can for instance be complemented by utilizing the formed classification between larger and less volatile, larger and more volatile, and smaller depositors. Table 5 shows the results of this analysis. The values have been computed for date34 of the sample daily data.

The most obvious result demonstrated by Table 5 is the fact that in total, the larger depositors represent around 37% of the total volume of the sample data, but only 0.5% of the total number depositors. Although, by definition, the larger depositors have proportionally greater deposit volumes than the smaller depositors, in this case it appears that the larger depositors have still a somewhat large amount of the total sample data deposit volume relative to the number of these depositors. Correspondingly, 99.5% of the depositors are smaller but they represent only 62.9% of the total sample data deposit volume. Hence, it can be concluded that within the given sample deposit base there is at least a some level of concentration of larger depositors.

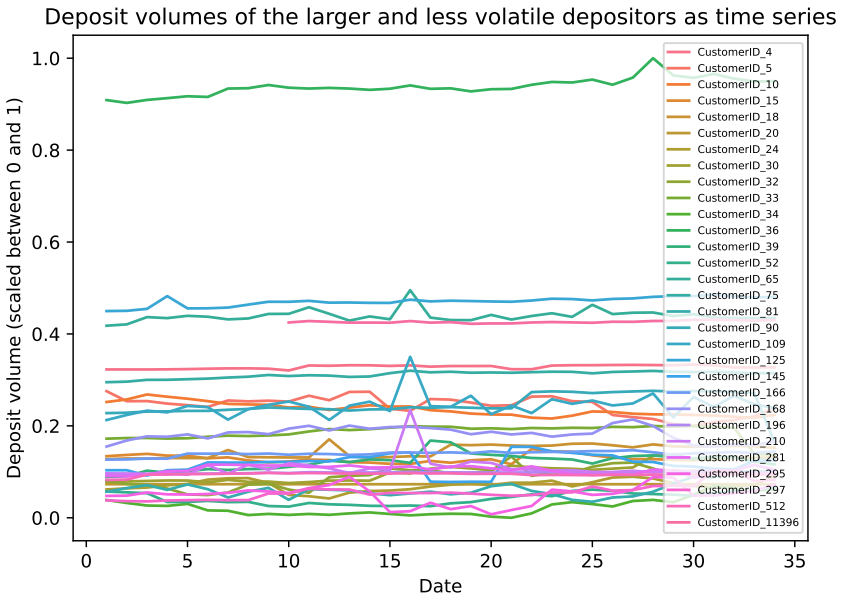

Figure 12: Scaled deposit volumes of the identified larger and less volatile depositors as time series.

To assess the riskiness of the concentration of larger depositors, one needs to consider the split between less volatile and more volatile larger depositors due to the reasons expressed in Section 3.4. First of all, it appears that the deposits are not particularly concentrated under larger and less volatile depositors, but the distribution of the deposit volume between larger and less volatile and larger and more volatile seems rather equal. In absolute terms, 22.5% of the total sample data deposit volume is obviously under larger and less volatile depositors, whereas only 14.6% of the total deposit volume corresponds to the larger and more volatile depositors. However, the ratio between the share of total deposit volume (%) and share of total depositors (%) is roughly the same in both cases, so that the difference in the shares of total deposit volume (%) is rather a by-product of the set cut-off point between the two classes and not so much of a real phenomena. This is good news for a bank, because it would not like have disproportionately high concentration of unpredictable depositors.

Second, we computed the number of distinct instances of the different attributes of the data. Note that even though the size of the client has throughout the project interpreted as the aggregated deposit volume of different accounts, here we have counted the number of distinct instances by utilizing all of the different accounts. Hence, for instance, larger and more volatile depositors can have 39 different currencies, even though we identified only 20 larger and more volatile depositors.

When it comes to interpreting these results, unsurprisingly the larger and more volatile depositors represent only a subset of the possible instances. This is particularly evident from the number of distinct industries. For larger and less volatile depositors, there are 22 of those, whereas within the whole sample depositor base there exist 639 different industries. On the other hand, as 30 larger and less volatile depositors were identified, presence in 22 different industries is a rather favorable result from the risk management point of view, even if we allowed ourselves to consider the dif-

Measure \ Depositor group	Larger and less volatile	Larger and more volatile	Smaller	Total
Gini	0.307	0.338	0.791	0.860
Share of total deposit volume (%)	22.5	14.6	62.9	100
Share of total depositors (%)	0.3	0.2	99.5	100
Number of distinct types	3	4	10	10
Number of distinct products	3	2	3	3
Number of distinct countries	11	10	11	11
Number of distinct currencies	27	39	44	45
Number of distinct industries	22	25	638	639

Table 5: Analysis of larger and less volatile depositors versus larger and more volatile depositors versus smaller depositors versus the whole sample depositor base.

ferent accounts of these depositors. Furthermore, the number of distinct countries for larger and less volatile depositors equals the number of distinct countries within the whole sample depositor base. These results imply that even if some unfavorable event occurred within a certain industry or country and the corresponding larger and less volatile depositors thus suddenly decided to withdraw their money, it would not have a material negative effect on a bank.

In addition, inspired by Section 3.3, we computed the Gini coefficient between the groups of larger and less volatile, larger and more volatile, and small depositors. That is, whereas the row "Gini" in Table 5 shows the Gini coefficient within each group, for the Gini coefficient between the groups we first summed up the total sample data deposit volume within each group, and then applied (2) to this set of three numbers. This gave us 0.323 as the Gini coefficient. The result to some extent reflects how the shares of total deposit volumes have been distributed between different groups. It is also a lot lower than the Gini coefficient corresponding to the whole sample deposit base, 0.860. This implies that within the sample deposit base there are not only few large depositors and lots of small depositors, but also, from another point of view, a noteworthy portion of larger and less volatile depositors, whose behavior under stress is unpredictable, accompanied with a set of more predictable depositors.

So far we have restricted the analysis of the larger and less volatile depositors of daily sample data. However, as we have also data from 2017, 2020, and 2022, we can equally well extend this analysis to cover the development of these largest and least volatile depositors over time. To preserve some level of comparability of the results, the 30 least volatile depositors among the 50 largest depositors are considered to form the group of the larger and less volatile depositors in each of the examined years. We note that this choice definitely makes the perception of the larger and less volatile depositors even more obscure, but at least the size as well as the formation of this group of analyzed depositors is well-defined.

Measure \ Year	2017	2020	2022
Gini	0.289	0.242	0.215
Share of total deposit volume (%)	12.5	16.3	10.2
Share of total depositors (%)	0.1	0.3	0.3
Number of distinct types	4	3	3
Number of distinct products	3	3	2
Number of distinct countries	10	11	11
Number of distinct currencies	32	37	36
Number of distinct industries	22	21	28

Table 6: A set of statistical measures for the 30 least volatile depositors among the 50 largest depositors in 2017, 2020, and 2022.

Table 6 shows the similar measures as in Table 5 for the 30 least volatile depositors among the 50 largest depositors in 2017, 2020, and 2022. The group of the 30 least volatile depositors among the 50 largest depositors is computed for each year separately. We first observe that the Gini coefficient for the analyzed group decreases over time. This implies that the size distribution of the 30 least volatile depositors among the 50 largest depositors is smoother in 2020 and 2022 than in 2017, i.e., the very largest depositors among this group are slightly smaller compared to the smallest depositors of this group in 2020 and 2022 than in 2017.

There is no clear trend how the share of total deposit volume corresponding to the analyzed group develops over time. On one hand, it is clearly greater in 2020 than in 2017, but on the other hand, in 2022 the share is even lower than in 2017. One possible explanation for the increase in 2020 is COVID-19, as the larger depositors might have had more resilience against the adverse effects of the pandemic than the smaller depositors, and therefore the latter could have been forced to withdraw at least some of their deposits. This could also explain the increase in share of total depositors, as it is 0.3% in 2020 and 2022, and 0.1% in 2017.

Lastly, one can analyze how the number of distinct instances of the different attributes develops over time. Based on the available sample data, it appears the group of the 30 least volatile depositors among the 50 largest depositors has become more diverse. For instance, the number of distinct currencies has increased from 32 to 36 and the number of distinct industries from 22 to 28, when we study the development from 2017 to 2022. This is, the group of the most unpredictable and impactful depositors is more diversified across different currencies or industries. There are no significant differences either in the number of distinct types, products, or countries.

4.6 Scenario Analysis

Scenario analysis using an agent-based model is applied to the sample of daily data. The analysis is conducted for the larger and less volatile customers identified in Section 4.4. Since these customers are less volatile, historical data cannot be used to predict their behavior or withdrawal patterns. This lack of knowledge of customers behaviour poses a risk for a bank. However, scenario analysis techniques allow us to estimate how quickly a bank would loose material volumes if one of these larger and less volatile customers withdrew their deposits and other agents acted accordingly. This information can be used to assess the risk associated with these customers and their potential actions.

The agent-based model described in Section 3.5 is based on several assumptions. First of all, agent-based models rely on agents that interact with each other. In our context, this means that when one agent withdraws deposits, similar agents would follow, as they are assumed to share the same interests, information and needs due to similar background or prior behaviour. This chain of events, causing agents to act similarly or react to the actions of other agents, can be likened to panic propagation within tightly interwoven networks or to the influence of external factors impacting an industry, country, etc. The simulation process we use is simplified, with a single starting agent initiating the withdrawals without any other influencing factors.

We assume that two agents are similar to each other and would act similarly if they either share similar attributes or have exhibited similar behavior based on historical data. While we utilized Hamming distance and Pearson correlation coefficient to measure these similarities, other metrics could also be explored. For example, Hamming distance treats all attributes equally, although some may be more interesting or important than others.

Another assumption that is made is the threshold at which the bank is considered to have lost a material volume. In our simulations, we set this threshold at 20 % of the total volume in the sample dataset, meaning that when the amount of money lost exceeds this threshold, mitigating actions are assumed to be necessary. This choice for the threshold allowed us to run the simulations efficiently while still effectively showcasing the model. However, more sophisticated thresholds incorporating expert knowledge or liquidity risk metric could be considered. In addition, we required that the correlation between two agents must be greater than 0.4 for them to be considered similar.

Finally, we also made assumptions regarding customer profiles. Since one customer can have multiple accounts with different attributes under the same ID, we simplified the situation by focusing only on the attributes of the largest account. We considered whether the different accounts of the same customer should be viewed as separate accounts or whether they should be combined under one profile. The latter approach aligns with the analysis conducted in Sections 4.3 and 4.4, where larger and less volatile customers were identified. Considering each account under the same ID as separate customers would have altered the selection of the largest and least volatile customers. In addition, customers typically have one predominant main account alongside smaller secondary accounts. These accounts often share most of the attribute values, and only certain values differ. Intuitively, combining all the accounts under one profile assumes that a customer withdraws all deposits if they withdraw deposits from their primary account. On the other hand, considering each account under the same ID as separate customers treats each account independently.

4.6.1 Scenario Analysis Using Hamming Distance

The Hamming distances between the larger and less volatile customers and all the other customers are first computed. Here we used the simplified view of the profiles by combining every account with the same ID under the same profile. The distances are between zero and five as we have five attributes. Zero means that all the attribute values are the same and five that they all differ. Figure 13 visualizes the computed Hamming distances. Note that we did not calculate the distances between all customers, but the distances to all customers were only calculated for larger and less volatile customers.

In the simulation process, we select one of the larger and less volatile customers at a time and initiate the simulation with them. Two types of simulations were conducted. In the first simulation, all agents having the smallest distance to the initiating agent withdraw their money simultaneously. This process continues until the cumulative amount withdrawn exceeds the chosen threshold. In

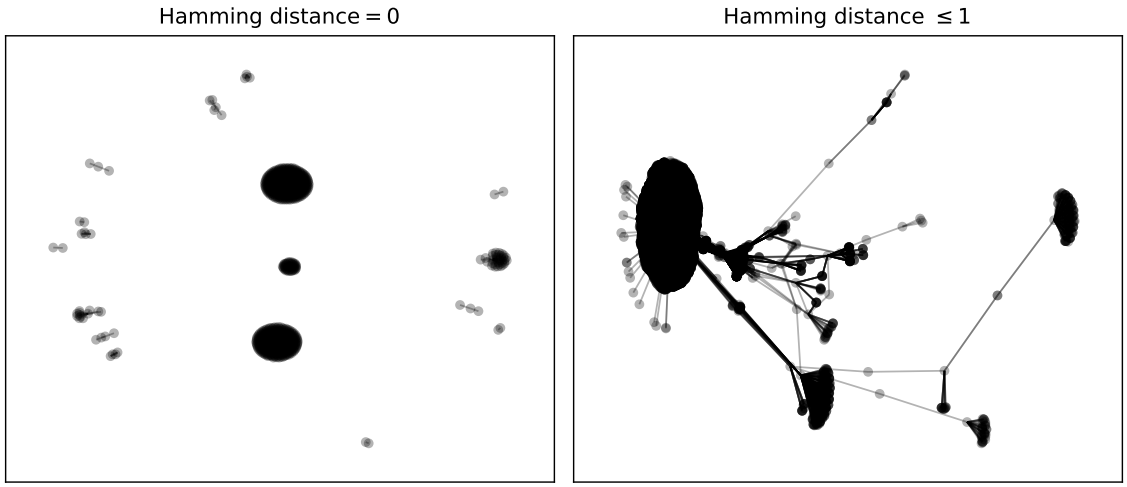

Figure 13: Visualization of Hamming distances: Nodes represent customers, with links between them indicating a distance of zero in the left image and one or less in the right image.

other words, all agents (if any) with a distance of zero from the initiating agent withdraw their money, followed by those with a distance of one, and so forth until the bank has lost material volumes. In the second simulation process, we randomly select one of the agents with the smallest distance to the initiating agent and proceed with the withdrawal process agent by agent until the bank has lost material volumes. Thus, all agents with a distance of, for example, zero are selected first before proceeding to the next closest agents with distance one to the initiating agent. The whole simulation process is repeated ten times due to the randomness in the selection of agents, and the average number of iterations is calculated. Results of the two simulation processes are shown in Table 7.

As can be seen from Table 7, the number of iterations required in the second simulation varies significantly between customers. While some of the agents required a few hundred steps, others necessitated several thousand. Intriguingly, those customers that only required two iterations in the first simulation took a large number of steps in the second simulation process. Despite having a distance of one or less to numerous agents, these customers interacted with smaller agents, leading to longer simulation periods in the second simulation process. On the other hand, customers with significantly fewer iterations in the second simulation were connected to larger, possibly slightly less proximate, agents.

When we compare the number of iterations to the total number of customers in the deposit base, we find that 300 iterations represent less than 3.5% of the depositors in the sample data. This indicates that only a small percentage of customers, who are also somewhat closely linked to each other, withdrawing their money can result in the bank losing more than 20% of its deposit base. However, as we observed from simulation one, not all customers were in such close proximity. Therefore, another perspective is that even though some customers required a larger number of iterations in simulation two, they were all in close proximity. Consequently, if a customer required only 2 steps in the first simulation, it suggests that very closely related customers are capable of

Customer ID	Simulation 1	Simulation 2	Type	Product	Country	Currency	Industry
CustomerID_75	3	287.0	1	1	2	7	31
CustomerID_11396	4	307.6	1	1	7	11	5
CustomerID_33	4	316.2	1	1	7	11	31
CustomerID_512	4	318.8	1	1	7	11	31
CustomerID_145	3	321.1	1	1	2	8	5
CustomerID_125	3	336.7	1	1	2	8	5
CustomerID_109	3	361.3	1	1	2	7	13
CustomerID_36	3	391.7	1	1	2	7	13
CustomerID_65	3	410.6	1	1	2	7	13
CustomerID_210	4	430.1	1	1	5	8	13
CustomerID_168	4	446.1	1	1	5	8	1
CustomerID_52	4	461.5	1	1	1	5	1
CustomerID_81	4	495.9	1	1	4	5	1
CustomerID_297	4	763.1	4	1	7	8	1
CustomerID_39	3	1031.3	1	1	2	3	1
CustomerID_18	3	1032.6	1	1	2	3	10
CustomerID_5	3	1136.7	1	1	2	3	1
CustomerID_30	3	1140.7	1	1	2	3	1
CustomerID_32	3	1189.6	1	1	2	3	1
CustomerID_15	3	1209.4	1	1	2	3	1
CustomerID_196	3	1506.4	2	2	8	8	85
CustomerID_281	3	1832.2	2	1	4	7	11
CustomerID_90	3	1980.0	2	1	3	7	45
CustomerID_166	3	2014.4	2	1	9	7	74
CustomerID_10	4	2492.7	2	1	3	4	3
CustomerID_295	2	2628.2	2	1	2	8	117
CustomerID_20	2	3683.3	2	1	2	3	11
CustomerID_4	2	3689.9	2	1	2	3	4
CustomerID_34	2	3697.4	2	1	2	3	19
CustomerID_24	2	3827.0	2	1	2	3	14

Table 7: Simulation results using Hamming distance.

causing the bank's deposit base to drop by 20%.

To further analyze the results, we compare the attribute values of the five lowest-ranking customers and the six highest-ranking ones from Table 7. Among the five lowest-ranked, four share common attributes: Type 2, Product 1, Country 2, and Currency 3. Customer ID 295 has otherwise the same attribute values, but instead of currency 3, it has currency 8. The value for industry is different for each of the five customers. For the six highest-ranking customers, all have the same values for attributes type and product, i.e. one for both, three had country 2 and three country 7, also three customers have currency 11, two currency 8 and one currency 7 and three have industry 31 and three industry 5. None of the five lowest-ranking customers have industry 31 or industry 5.

4.6.2 Scenario Analysis Using Pearson Correlation Coefficient

Pearson correlation coefficients were computed between all customers. In the simulation process, we select one of the larger and less volatile customers at a time and initiate the simulation with them as before. The agent closest to the initiating agent based on the correlation decides to withdraw their money too. This process continues until there are no more proximate agents, i.e. agents with correlation greater than 0.4, or the amount of money withdrawn exceeds the predetermined threshold. If there are no more proximate agents for the initiating agent, we continue the process with the second closest agent. The simulation results are shown in Figure 14.

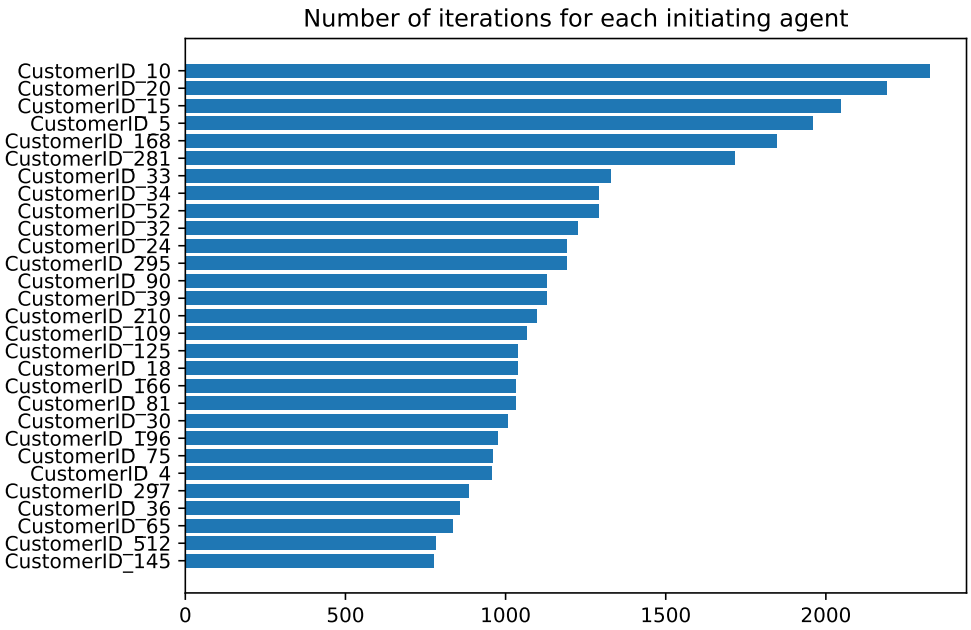

Figure 14: Simulation results using Pearson correlation coefficient. The width of the bar represents the number of iterations each agent needs to reach the threshold.

From Figure 14, we can observe differences in the number of iterations once again. The smallest number of iterations is 776, while the largest is 2323. Additionally, six customers that required the greatest number of iterations stand out from the rest.

There is not a clear set of attribute values evident in either the top-ranked or lowest-ranked customers. Among the six lowest-ranking customers, all have either type 1 or 2 and product 1, with the majority having country 1, currency 3, and industry 1 or 11. On the other hand, the six highest-ranking customers all have product 1 and either country 2 or 7. Comparing these findings to the simulation results in Table 7, we notice that many of the customers that are ranked low when using Pearson correlation coefficient are also in the lowermost part of Table 7, and similarly for the highest-ranked ones.

The variation in the number of iterations between the simulation utilizing Hamming distance and the one employing Pearson correlation coefficient is likely due to the fact that the latter used only those customers for whom data covering the entire time interval was available. Consequently,

some customers, potentially with significant deposit bases, were excluded from the analysis, leading to an increase in the minimum number of iterations from 300 to 800. This prompts the argument that, from this perspective, the first model is superior as it only necessitates a single point in time rather than data from the entire interval. However, we did not investigate whether customers with similar profiles indeed behave similarly. Thus, simulation utilizing Pearson correlation coefficient better supports the assumption of similar behavior of nearby agents. However, since the study was conducted on less volatile and relatively larger customers, for which historical data cannot be used to draw conclusions about deposit and withdrawal patterns, the Hamming distance could be more reasonable in this case.

After the scenario analyses, we have again slightly increased our understanding of larger and less volatile customers. What was particularly interesting were the fairly large differences in the prevalence of customer profiles, which can be seen in Figure 13. Some of these customers shared the same profile with numerous other customers, while some customers were quite unique in profile. This was also visible in simulation 1 in Table 7. On the other hand, it was also interesting to compare simulation 1 with simulation 2, as it revealed that a customer closely connected to other customers may still require several iterations due to the customers' smaller deposits.

In both scenario analyses, there were fairly large variations in the number of iterations, reflecting the different connections of the customers. The obtained results of the analyzes did not completely correspond to each other, although similarities were also found. Consequently, definitive conclusions about the risk posed by customers cannot be drawn. Additionally, it is important to consider the assumptions and simplifications made by the models.

5 Recommendations

In this section we discuss the recommendations regarding the management and mitigation of concentration risk that can be derived based on this project. First of all, we recommend that a bank obviously should aim for a diversified depositor base. For instance, the case of SVB showed that being reliant to a single industry can be fateful, if an incentive to withdraw deposited money arose within the industry. A synchronous occurrence of a similar incentive across different industries, on the other hand, is less likely. The question of how much of diversification is enough, depends on the preferences of the bank.

When it comes to opting for suitable preferences, we recommend the bank to take a somewhat conservative stance. The reason is that, for instance, being conservative in the assumption of how much money the larger depositors will withdraw in the near future makes the bank appear less vulnerable for possible withdrawal. This can then increase the confidence of the depositors in terms of the stability and resilience of the bank. Furthermore, as the depositors have an increased confidence that the bank is reliable, the bank might attract new depositors due to its reputation being reliable.

To complement the chosen preferences the bank also needs proper risk management tools in place. In the light of this project, the Gini coefficient appears as the single most effective tool to measure deposit concentration. As discussed earlier, its main drawback is that it tells little about the riskiness of the underlying distribution. However, if the depositors were labeled according to their riskiness, then the concentration of the riskiness within the depositor base could be also evaluated. By tracking the Gini coefficient over time, the bank receives additional information on how the deposit concentration evolves.

6 Conclusions

The aim of this project was to present SEB with diverse methodologies to accurately assess the concentration risk associated with their corporate deposits. This idea was motivated by the collapse of the American bank Silicon Valley Bank (SVB) in March 2023 which highlighted critical vulnerabilities within the banking sector.

Assessing the concentration risk turned out to be a complex problem with multiple dimensions that seems to have no straight forward solution, at least during this time frame and resources. Thus, to understand the problem multiple measures are needed. There was not much literature of this topic which made it even more challenging as it made the benchmarking of the results harder.fal

We used a static approach (looking into the data at one point in time), a dynamic approach (with multiple time points) and scenario analysis to analyse the problem. With the use of measures like the Gini coefficient and the Herfindahl-Hirschman Index, the static approach provided quick snapshots of the concentrations and made it easier to see how deposits were distributed for instance throughout deposit size, different industries and geographical areas. These measures are easy to compute but have their limitations and they consider only one point of time.

The dynamic approach, on the other hand, provided us a broader picture, using monthly time-series data to track and predict changes in the deposit volumes over time. We used linear regression on different clusters of depositors trying to find clusters that represent the changes in the deposit base the best. This was measured with the R^2 score. Volatility as a measure turned out to be an efficient method with an R^2 score of 0.935 for the top 5% most volatile customers. The dynamic approach proved helpful in identifying the depositors who represent the largest liquidity risk, particularly the larger and less volatile depositors. These depositors are interesting as there is not much historical information about them modifying their deposit patterns. This could impact a bank significantly if some of such depositors choose to withdraw their funds.

Our analysis was further enhanced by scenario analysis, which simulated the possible unfavourable event of similar customers withdrawing their money from the bank. This analysis was done by using agent-based modeling where the similarity of different profiles were measured using Hamming distances and Pearson correlation. This analysis allowed us to estimate how quickly the bank would experience significant withdrawals if one of these larger and less volatile customers withdrew their deposits and other agents acted the same way. However, definitive conclusions about the risk posed by customers cannot be drawn using our scenario analysis yet.

As mentioned in the previous section at least according to this project, the Gini coefficient appears as the single most effective tool to measure deposit concentration. It is simple to compute and interpret, and combined with an appropriate risk analysis tool, can also to some extent help to assess the riskiness of the respective concentrations.

References

[1] Ivan Brezina, J. Pekár, Z. Čičková, and M. Reiff. Herfindahl–Hirschman index level of concentration values modification and analysis of their change. *Central European Journal of Operations Research*, 24, 04 2014. doi: 10.1007/s10100-014-0350-y.

[2] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Richard Stratton, 8 edition, 2011.

[3] Einar Jon Erlingsson, Andrea Teglio, Silvano Cincotti, Hlynur Stefansson, Jon Thor Sturluson, and Marco Raberto. Housing market bubbles and business cycles in an agent-based credit economy. *Economics*, 8(1):20140008, 2014.

[4] Joseph L. Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

[5] Joseph L. Gastwirth. The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, pages 306–316, 1972.

[6] Corrado Gini. Concentration and dependency ratios. *Rivista di Politica Eonomica*, 87:769–792, 1997.

[7] Marc Gürtler, Martin Thomas Hibbeln, and Clemens Vöhringer. Measuring concentration risk for regulatory purposes. *Journal of Risk*, 12:69–104, 2010.

[8] Ragnar E. Juelsrud. Deposit concentration at financial intermediaries. *Economics Letters*, 199: 109719, 2021. doi: https://doi.org/10.1016/j.econlet.2020.109719.

[9] Peter Klimek, Sebastian Poledna, J. Doyne Farmer, and Stefan Thurner. To bail-out or to bail-in? Answers from an agent-based model. *Journal of Economic Dynamics and Control*, 50: 144–154, 2015.

[10] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[11] Max O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[12] Basel Committee on Banking Supervision. Overview of pillar 2 supervisory review practices and approaches, 2019. URL https://www.bis.org/bcbs/publ/d465.pdf. Accessed: 04.03.2024.

[13] Tatjana Pavlenko and Oleksandr Chernyak. Credit risk modeling using bayesian networks. *International Journal of Intelligent Systems Research*, 25:326–344, 04 2010. doi: 10.1002/int.20410.

[14] J. David Cabedo Semper and Jose Miguel Tirado Beltrán. Sector concentration risk: a model for estimating capital requirements. *Mathematical and Computer Modelling*, 54(7-8):1765–1772, 2011.

[15] Lai Van Vo and Huong T.T. Le. From hero to zero-the case of Silicon Valley Bank. *SSRN Electronic Journal*, 2023. doi: https://dx.doi.org/10.2139/ssrn.4394553.

[16] Bill Waggener and William N. Waggener. *Pulse code modulation techniques*. Springer Science & Business Media, 1995.

[17] Wikipedia contributors. Lorenz curve. https://en.wikipedia.org/wiki/Lorenz_curve#/media/File:Economics_Gini_coefficient2.svg, 2024.

7 Self Assessment

7.1 Implementation of the Project With Respect to the Initial Project Plan

When looking back at the initial project plan, we can notice some differences when compared to this final report. Perhaps the most notable change is the overall thinking around the goal of the project; namely focusing more on comprehensive reporting and numerical results that help the client obtain an idea of concentration risk in their deposit base, instead of developing a ready-to-use model that the client would be able to apply for quantification of concentration risk. As a result, there were some minor changes in the scheduling and risks of the project as well. Due to the limited amount of applicable literature, we primarily performed the benchmarking against the article from Juelsrud [8]. With the limited time frame of the project, we also completely scrapped the deposit stickiness assessment that was mentioned in the initial plan, since that was included as more of an optional task in the first place.

7.2 In What Regard Was the Project Successful?

Overall, we find the project to have been a success. Even though we did come to understand that the initial project plan (developing a model for quantification) was not exactly feasible in the given time frame, we adjusted accordingly and managed to follow the new approach.

In spite of the lack of relevant literature around the topic, we still found meaningful literature and recognized the weaknesses surrounding the literature. This also prompted us to come up with multiple approaches for different models that we could use to model the problem, and after the adjustment of our approach we still managed to find approaches that would provide meaningful results for the client. In addition, the selection process of different approaches throughout the project was related to the best-practice thinking of the project, i.e. solutions that we believed to be as informative as possible.

In terms of the actual results of the project, we believe we found something useful and significant related to the riskiness of a bank's deposit portfolio, and believe we have provided value to the client through this project.

7.3 In What Regard Was It Less So?

While the project was indeed a success in our eyes, there are some aspects of the project that would have made it even more successful. Namely, there was maybe a lack of a clear vision or focus at certain points during the project where we were considering different approaches. Additionally, there were some misalignments between our thinking and the client's with respect to the measurement and definition of concentration risk at points due to a lack of clear and effective communication between the parties. We also did not manage to provide a direct model as a solution for the problem as was initially suggested, but this has already been addressed with the client, so we ca not exactly label this as an unsuccessful part of the project.

7.4 What Could Have Been Done Better?

7.4.1 Project Team

Reflecting on our project journey in hindsight, it is clear that our team could have performed better in certain aspects. In the initial phase, we could have dedicated even more time and effort to the project. In particular, establishing a clear focus right from the start would have been beneficial. Initially, we approached the work with the expectation of primarily conducting a literature review, so the actual development of the models started perhaps a bit late. We could have initiated small-scale testing of the models at an earlier stage, as it could have helped us to identify the best practices for measuring the concentration risk of the bank's corporate deposits, which was the objective of the project. However, the shift from a literature review type project to a more practical one was partly necessitated by the restricted amount of available literature.

Thus, an earlier selection of models for development and a distinct separation between the literature review and model development project would have improved the efficiency of our work. An even more active interaction with the customer in the early stages could have helped here. In addition, we could have better utilized Professor Ahti's guidance.

7.4.2 Client

While reflecting on our interactions with the client, we noticed a potential ambiguity regarding the scope of the project. Clarifying whether the focus should have been on a general literature review or a specific case study using concrete models would have made it easier for us to focus our efforts on the right things. In addition, specifying the importance of each sub-objective would have helped to find the focus of the project.

The client demonstrated proactive commitment to the project, as they organized meetings whenever needed and actively inquired about the project's progress. Additionally, the client showed flexibility in adapting to changing project goals.

7.4.3 Teaching Staff

In hindsight, we believe that providing even clearer instructions for each step of the project would have been beneficial. For instance, the required contents of the project deliverables, presentations and the written feedback to the opponent team, as well as to whom it should be sent by email, could have been stated more clearly. Additionally, earlier communication of meeting dates would have been advantageous, especially as they are mandatory part of the course.