Empirical Economics (2022) 62:2893–2919
https://doi.org/10.1007/s00181-021-02120-5

A dynamic analysis of household debt using a self-organizing map

Hyun Hak Kim[1]

Received: 8 June 2020 / Accepted: 13 August 2021 / Published online: 27 August 2021

Abstract
The Korean consumer credit panel offers a well-organized set of microdata representing various characteristics of individual borrowers. To overcome the difficulty of fragmented microdata details, we construct a cluster of Korean consumers' credit, to develop a self-organizing map that visualizes individuals' characteristics along two dimensions. The result of cluster analysis reveals that most borrowers belong to one large cluster representing diligent borrowers who honor their loan payments. Conversely, several small clusters that represent borrowers with high default probability are identified, and we also found that these borrowers' characteristics vary. No significant change is found in the structure of the cluster, even when the aggregate amount of consumer credit is increased. Moreover, the expansionary monetary policy did not change the quantitative structure of household debt in Korea.

Keywords Household debt · Self-organizing map · Cluster analysis

JEL Classification G51 · C82

1 Introduction

Since the 2007–2009 global financial crisis (GFC), household debt has rapidly increased in most countries, whereas interest rates fell worldwide. In Europe and the USA, real estate prices fell following the financial crisis. In Korea, there was minimal deleveraging of asset prices during the crisis, and consequently, household debt continued to increase, even after the crisis, as shown in Fig. 1. However, if deleveraging occurs because of a fall in real estate prices, it could considerably impact both households and the Korean financial system. Thus, an accurate understanding of household debt is essential for developing financial stabilization policies. This

✉ Hyun Hak Kim
hyunhak.kim@kookmin.ac.kr

[1] Department of Economics, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea

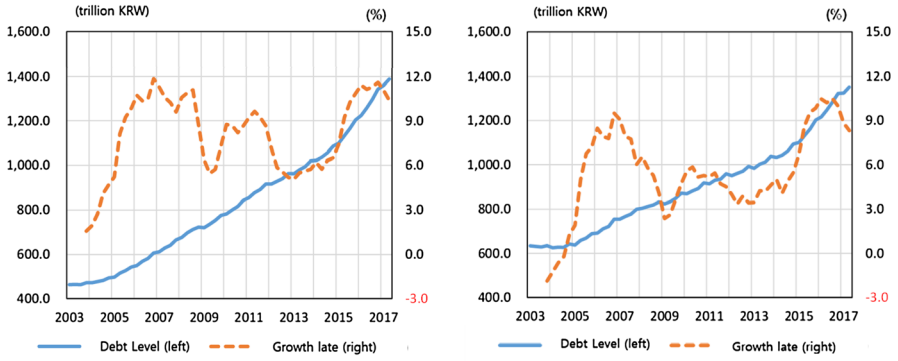


Fig. 1 Year-on-year household debt size trends and growth rate. Source: Bank of Korea Financial Stability Report

requires analyses of the microcharacteristics of household borrowers as well as macroconsiderations. The default consumer credit rate in Korea is less than 2%, which indicates that most borrowers pay back their loans on time. This is why an aggregate approach may not identify the fragile points in household debt. However, considering the small shares of subprime mortgage holders before the GFC, this small number of defaults should not be neglected, making it necessary to identify the characteristics of borrowers from various perspectives. For this purpose, this study uses a self-organizing map (SOM) to cluster borrowers on the basis of data from the Korean consumer credit panel (KCCP) and focuses on identifying both responsible and irresponsible borrowers. An SOM is useful for determining which characteristics are represented in clusters using attractive figures. This can help policymakers better understand the status of household debt.

Since one of the main triggers of the GFC was subprime mortgage lending, several studies have examined household debt from both macro- and microperspectives. The macroapproach analyzes the relationships among macroeconomic variables, such as GDP, interest rates, and aggregate household debt, along with financial institutions' ability to respond to macroshocks through stress tests. Mian et al. (2017) analyzed the role of household debt in a country-level economic system, analyzing panels for 30 countries for the period 1960–2012 and finding that household debt has a negative effect on economic growth. Mian and Sufi (2014) arrived at the same conclusion for the USA and other advanced countries. Alter et al. (2018) obtained results similar to Mian et al.'s (2017) for a wider set of countries over a longer sample period of 1950–2016. These studies emphasized the role of "debt-driven" consumption when household credit is booming by analyzing household consumption behavior. Jordà et al. (2016) examined the role of mortgages in the macroeconomy and demonstrated that housing finance plays a central role in the modern macroeconomy based on an analysis of 17 advanced countries from 1870 to contemporary times. Brunnermeier et al. (2019) studied the connection between household credit,

financial markets, and macroeconomic factors, concluding that credit and output growth were mostly positively associated.

The microeconomic approach uses household data, such as liabilities and financial position, to analyze debtors' ability to repay debt and the possibility of default, focusing more on individuals' credit. For example, Cheng et al. (2014) showed that household consumption behavior may lead to excessive borrowing when a positive credit shock occurs and an excessive drop in consumption when a negative credit shock occurs. Baron and Xiong (2017) found that the heterogeneity of a household can lead to more dynamic leverage cycles and asset price volatility. These studies highlighted the importance of consumer behavior.

Several studies have analyzed consumer credit in Korea. Kim and Kim (2010) showed that typical stress tests can underestimate the expected losses during a crisis. Park et al. (2012) analyzed the correlation between mortgage and bank loan delinquency rates using a dynamic response model. Kim and Lim (2013) used data from the mortgage accounts of large commercial banks to develop a delinquency determinant model for analyzing the loan behavior and bankruptcy factors of Korean households. In a microeconomic approach, Kim and Yoo (2013) conducted stress tests to assess household debt risk using household financial data. Hahm et al. (2010) were the first to use unit data from individual borrowers, investigating changes in borrowers' debt repayment ability. Among many others, Kim and Byun (2012) used microdata from credit rating agencies in Korea to identify fragile borrowers and explored the possibility of household debt becoming insolvent using both the micro- and macrostress tests. Analyzing data from the KCCP and real estate prices, Jung and Kim (2020) found that changes in property prices play an important role in determining a borrower's delinquency.

The SOM approach, proposed by Kohonen (1982), has been widely used for data visualization and cluster analysis in industrial engineering and statistics, whereas its use in the field of economics is relatively recent. In a leading SOM economics study, Sarlin and Peltonen (2011) used an SOM to present financial stability along a two-dimensional plane and visualize the various factors contributing to system risk to establish an early warning system for the European Central Bank. Sarlin (2013) applied an SOM for time series data to analyze the causes of structural changes in macrofinancial situations before and after the GFC. Resta (2016) applied an SOM to a value-at-risk analysis of German stock markets. Additionally, Claveria et al. (2016) applied an SOM to analyze 14 European countries, visualizing the pattern of changes in experts' expectations of economic conditions before and after the GFC. An SOM is a useful tool for clustering heterogeneous participants to enable the categorization of individual borrowers in huge consumer credit panel data for identifying common characteristics among clusters. This can help policymakers to implement more precise strategies.

In this study, we find one large cluster and several small clusters, although no significant change in borrowers' characteristics is evident during the sample period. The large cluster represents most borrowers who honor their loans. The small clusters,

which were of greater interest in this study, comprised borrowers with high debt-to-income (DTI) ratios and multiple debts, particularly the combination of personal loans and private business loans.[1] The number of borrowers with overdue loans in these groups might not have been high but could be a possible trigger in the next crisis, similar to how the subprime mortgage triggered the GFC, although they do not constitute a large share of debt. Moreover, a quantitative increase in consumer credit during the sample periods did not change the qualitative structure of household debt since the SOM results were robust during the sample periods. This is because no events were dramatic enough to cause structural changes in this study's experiments.

The next section introduces the data and methodology used in this study. Section 3 presents the results of the empirical analysis, whereas Sect. 4 provides the conclusions of this study.

2 Data and analysis

2.1 The KCCP

To identify the characteristics of borrowers, this study applied data from the KCCP, which was established by the Bank of Korea on the basis of the Korean Credit Bureau (KCB) database. The database includes information regarding the financial status of Korean residents aged 18 years and older, such as credit score, estimated income, loan amount and type, and overdue loans and their duration. The original sample size included 41 million observations of the 47 million observations listed in the KCB database. From the original sample, the Bank of Korea selected 400,000 observations using simple random selection, following the technique applied by the New York Federal Reserve Consumer Credit Panel (Lee and van der Klaauw 2010). Table 1 presents the selected variables listed in the KCCP. Following encrypted identification, quarterly data regarding individual financial transactions were obtained. Each loan was categorized on the basis of its purpose (personal or private business), type (secured/unsecured), and institution type (banking, nonbanking, or private lender.)[2] Private business loans are usually considered to be corporate credit in the literature; however, in Korea, the share of self-employed individuals is relatively higher. Referring to Jung and Kim (2020) and Kim and Jung (2019), this population exerts a contagion effect on consumer credit. Although it is not possible to obtain data from the panel regarding mortgage size, the amount of secured loans can be used as a proxy for mortgages, since most of the collateral provided by individual borrowers is real estate. In Table 1, Rows 1–7 represent personal identification information. Note that Row 7 refers to estimated, rather than true, income, and the estimated income was reported by the borrower when applying for the credit

[1] A private business loan is a loan obtained for small or self-employed businesses.

[2] Private lenders are distinguished separately because although they are not illegal, they serve as loan sharks in Korea. If an individual borrows from a private lender, the borrower is no longer considered eligible to borrow from any other financial institutions.

Table 1 List of variables selected from the KCCP

No	Short name	Description
1	YYM	Date by quarter
2	ID_LOAN	Randomly assigned personal ID
3	GU_CODE1	Postal code by residential records
4	SEX	Gender, 1 for male, 2 for female
5	AGE	Age group: 20–29, 30–39, 40–49, 50–59, 60–69, and 70 years and older
6	cbscore	Credit score, 0–1000, as reported by the credit bureau
7	INCOME	Expected income, reported by the consumer
8	P_BTC_TOT	(Personal) Total loans from all institutions
9	P_BTC_DL	(Personal) Total secured loans from all institutions
10	P_BTC_BL	(Personal) Total loans from banks
11	P_BTD_TOT	(Personal) Total overdue
12	P_NTC_TOT	(Personal) Number of total loans
13	P_NTC_DL	(Personal) Number of secured loans (private)
14	P_NTC_BL	(Personal) Number of total loans from banks
15	P_NTD_TOT	(Personal) Number of total loans overdue
16	C_BTC_TOT	(Business) Total loans from all institutions
17	C_BTC_DL	(Business) Total secured loan
18	C_BTC_BL	(Business) Total loans from banks
19	C_BTD_TOT	(Business) Total overdue
20	C_COL_TOT	(Business) Total collateral
21	USE0_A	Credit card usage out of limits, all transactions
22	USE0_S	Credit card usage out of limits, credit transaction
23	USE0_C	Credit card usage out of limits, cash advance
24	BQT0_A	Total amount of credit card usage
25	PLENDER_A	Total loans from private lenders
26	PLENDER_N	Number of loans from private lenders
27	BTC_TOT	Total loans (8 + 16 + 25)
28	BTC_DL	Total secured loans (9 + 17)
29	BTC_BL	Total loans from banks (10 + 18)
30	BTC_DR	Secured loan ratio (28/27)
31	BTC_BR	Bank loan ratio (29/27)
32	P_BTC_DR	(Personal) secured loan ratio (9/8)
33	P_BTC_BR	(Personal) bank loan ratio (10/8)
34	P_NTC_DR	(Personal) secured loan number ratio (13/12)
35	P_NTC_BR	(Personal) bank loan number ratio (14/12)
36	C_BTC_DR	(Business) secured loan number ratio (17/16)
37	C_BTC_BR	(Business) bank loan number ratio (18/16)
38	C_COL_RA	(Business) collateral ratio (20/16)

card. Rows 8–26 are original series, and Rows 27–38 are series generated from the original series. BQT0_A was set as a proxy for consumption since the use of credit and debit cards is pervasive in Korea.[3]

One drawback of KCCP data is that they do not include information regarding the collateral assets that correspond to the secured loans and affect the financial ability of borrowers. Lee et al. (2014) used the average housing price of borrowers' houses to estimate the value of collateral assets, considering that real estate is typically the biggest financial asset held by households. Subsequently, housing price (APT_M2P) and the number of house transactions (APT_NUM) were added as proxies for the collateral for secured loans, as shown in Table 2, Rows 40–41. CBGRADE indicates the discrete credit rating based on the cbscore variable in Row 6 of Table 1. INTEREST is the average interest rate, based on the amount of interest paid by the borrower and estimated using the total loan amount and total interest payment made by the borrower. Since this total interest payment is not distinguished by loan type or purpose, this study follows that of Jung and Kim (2020), who split the interest rate on the basis of the credit rate and loan type. Overdue loans are estimated separately; however, the seriousness of overdue loans was also considered, and the discrete series (P_CREDIT) was constructed to judge the seriousness of overdue loans depending on the duration of overdue status. Both the debt–service ratio (DSR) and the DTI ratio[4] were calculated; however, the DTI ratio was estimated separately since a discrete series is required for clustering. Subsequently, borrowers' income (INCOME_G) was grouped. MUL_DEBT represents the number of loans obtained by borrowers from different types of financial institutions, whereas DET_CAT represents the type of debt. Finally, borrowers' consumption preferences were represented by dividing their consumption by estimated income. Consumption was represented by the total usage of credit and debit cards.

Next, household debt owners' characteristics were dynamically identified by generating an SOM. For brevity, the SOM was generated for five periods of 2010 (Q1), 2012 (Q1), 2014 (Q1), 2015 (Q4), and 2016 (Q4). Since there were no significant differences in the structure of borrowers' characteristics quarter by quarter, SOM is applied in intervals. The results for all sampling periods are available upon request. The household debt database is the data compiled for every quarter from 2010 (Q1) to 2017 (Q2).

[3] Based on the Bank of Korea's report in 2019, credit cards and debit cards were used to pay for more than 70% of consumption in Korea, whereas the remaining 30% was attributed to cash and money transfer.

[4] Both DSR and DTI ratios are defined as the ratio of debt–service payment (principal + interest) to income; however, the object of the DTI ratio is the mortgage (or secured loan), whereas the DSR ratio includes everything, even loans obtained from private lenders.

Table 2 List of author-constructed variables

No	Short name	Description
39	CBGRADE	Credit rating, 1–10 from top to bottom, set by the author
40	APT_M2P	House price (per square meter) based on borrower's residential records
41	APT_NUM	Number of housing transactions at borrower's residence
42	INTEREST	Estimated interest based on borrower's interest payment
43	P_CREDIT	Overdue status (0 for no overdue loans, 1 for less than 90 days overdue, and 2 for more than 90 days overdue)
44	DSR	Debt–service ratio
45	DTI	Debt-to-income ratio
46	DTI_G	DTI by group (0 for DTI < 1, 1 for DTI < 1, 2 for 1 ≤ DTI < 3, and 3 for DTI ≥ 3)
47	INCOME_G	Estimated income by group (1 for low income, 2 for middle income, and 3 for high income)
48	MUL_DEBT	Number of loans (0 for no debts, 1 for one to two debts, and 2 for three or more debts)
49	DET_CAT	Type of debt (0 for no debt, 1 for personal loan, 2 for private business loan, 3 for private lender, and 4 for any type of loan other than from a private lender)
50	CON_R	Consumption preferences (consumption/estimated income)

2.2 SOM

The SOM approach is a method of autonomous learning, developed by Kohonen (1982), as an unsupervised learning technique.[5] It allows for projection and cluster analysis to be performed simultaneously. Kohonen (1982) originally used SOMs for visualizing the cerebral cortex, and the technique is often used in data mining studies because it is an uncomplicated method for describing large volumes of data. There are two approaches to reduce the amount of data or the number of dimensions required to describe large volumes of data. The first method is cluster analysis, which involves dividing an entire dataset into several clusters. An average characteristic among several clusters can be regarded as that of the entire dataset. The second method is the mapping method used by Cox and Cox (2001), which involves projecting high-level data into low-dimensional space. Although this method creates a shared information space without reducing data, it has a similar effect in reducing the volume of data. The SOM approach is basically a mapping method that also determines the characteristics of clusters on a map.

More specifically, the SOM approach starts by projecting large volumes of data into a certain data space. First, the term "node," which is frequently used in neural network analysis, meaning the junction or access point of the neural network, should be acknowledged. SOM maps quantify continuous data to grid nodes through a density function $f(x)$, wherein the location of the nodes depends on the adjacent relationship between data. The SOM then visualizes the data on a two-dimensional map, with the X and Y axes indicating the relationships among the data. For example, if a 160-cm-tall, 55-year-old woman is included in Node A on an SOM, then any 55-year-old woman, taller or shorter than 160 cm, will be located on neighboring nodes close to Node A. Similarly, the node to which a tall, heavy man belongs will be located near the node to which a tall, heavy woman belongs, rather than to one to which a short and light man belongs.

The SOM approach has advantages in that it intuitively builds profiles of members belonging to each node, visualizing these profiles on a map, and is easy to execute. Additionally, when new data that were already estimated are presented on a map, it is possible to categorize which node it belongs to. For example, if the SOM for financial stability is already estimated for G7 countries, then data from a new country, such as Korea, could be added to the existing SOM, easily enabling the evaluation of the financial stability of that country. By contrast, a drawback of the SOM approach is that it may be difficult to handle missing values or disproportionate data. Thus, data filtering must be conducted in advance, before using an SOM. Several approaches have been developed to extend the SOM approach. Vesanto and Alhoniemi (2000) presented a two-step method of reapplying an SOM to acquired nodes after executing the SOM once. Marghescu (2007) found

[5] Machine learning is largely divided into supervised and unsupervised learning. The former occurs when the outcome of the dependent variable is affected by the explanatory variable. The latter comprises only descriptive variables, without dependent variables, using cluster analysis for self-learning.

that such two-stage SOMs produced stronger results than other methods. For more general SOM approaches, Eklund et al. (2008) provided additional details about extending the SOM approach, whereas Sarlin (2013) produced a wide range of surveys regarding the SOM method.

2.3 Estimating SOM

Suppose the SOM's grid is configured with m_i $(i = 1, 2, \ldots, M)$ nodes determined by the user (theoretically, as many nodes can be configured as the number of dimensions (M) of the input data). When the nodes are created, the SOM will insert the input data into these nodes, considering the distance and weight of the data around them, while elevating the status of nodes that contain large amounts of data. The algorithms implemented by the SOM are as follows:

1. Disassemble the main components from the covariance of data Ω and extract the two main components with the greatest descriptive power to initialize the value of the node.
2. In the ith iteration, compare data x to all m_i nodes to find node m_b, which is the best match (i.e., the smallest distance from the average value of the node):

$$||x - m_b(t)|| = \min_i \left|\left|x - m_i(t)\right|\right|.$$

3. Recalculate the attribute node m_i from the mean of aggregated data:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{ic(j)}(t) x_j}{\sum_{j=1}^{N} h_{ic(j)}(t)},$$

where j is the index for the vectors of input data belonging to node b and N refers to the total number of data vectors. $h_{ic(j)}(\in (0, 1])$ is a neighborhood function and is a Gaussian function as follows:

$$h_{ic(j)} = \exp\left(-\frac{||r_j - r_i||^2}{2\sigma^2(t)}\right),$$

where $||r_j - r_i||^2$ is the square of the Euclidean distance of the coordinates of reference nodes m_j and m_i drawn on two dimensions. $\sigma(t)$ is a parameter that represents the radius of the neighborhood, a reduction function according to the repeat recovery t, which starts at half the size of the grid diagonal in the first run, that is, σ, where X and Y are the coordinate values of the center (average) of the nodes on a two-dimensional basis.
4. Repeat steps 2 and 3 for the specified number of times.
5. The nodes are clustered according to Ward (1963), using an upward method that is gradually subdivided into M clusters. Ward measured the distance between the two cluster candidate groups as follows:

$$d_{kl} = \begin{cases} \left|\frac{n_k n_l}{n_k + n_l}\right| \bullet ||c_k - c_l||^2 & \text{if } k \text{ and } l \text{ are adjacent,} \\ \infty & \text{otherwise,} \end{cases}$$

where k and l are indices representing two clusters, n_k and n_l represent the number of data belonging to clusters k and l. $||c_k - c_l||^2$ is the square of the distance of the Euclidean between the centers of colonies k and l.

Ward's (1963) clustering method was accomplished by defining all nodes as one cluster and the same cluster when adjacent nodes come within a certain distance range. In this study, only adjacent nodes were "clusterable" candidates because the distance from nonadjacent nodes was defined as infinite. In the first step of the above algorithm, the main components of the input data were extracted, and the node's initial value was used because a faster SOM converging time was expected than in the case in which the node is configured using a randomly set vector. In the SOM approach, a critical issue is that the number of nodes (size of the SOM) and the iteration criteria must be defined in advance. Normally, algorithms are performed until the distance to the center of the node is below a certain level or until the decreasing distance in the new run is below a certain level. These issues are discussed in the next section.

3 Empirical analysis

3.1 Variable screening

Although a big data environment allowed for the inclusion of more information to describe the data in this study, it also generated noise in the analysis. To overcome this drawback, the data were filtered to fit the research purpose of identifying the characteristics of clusters of borrowers. First, all the variables from KCCP and generated from the data were adjusted to fit with the SOM configuration. This process includes transforming the discrete series and obtaining the logarithm of the series. Consequently, all continuous series were transformed to become stationary. Finally, a selected 50 variables are considered. However, the inclusion of all these variables in the SOM generation would have imposed a computational burden and might have created too many clusters, making it difficult to interpret the results. Thus, two strategies were applied to choose appropriate series. One strategy involved selecting 10 variables that best described the covariance matrix of variables on the basis of Zou et al.'s (2006) sparse principal component analysis (SPCA) and the other involved selecting 10 variables that best describe household debt on the basis of (prior) theoretical knowledge. In the end, cases with more or less variables elicited similar results. Experiments conducted using other numbers of variables are available upon request.

SPCA, proposed by Zou et al. (2006), is a variable selection method that adds special constraints to a conventional principal component analysis (PCA) for selecting

Table 3 List of variables selected using the SPCA

	2010 (Q1)	2012 (Q1)	2014 (Q1)	2015 (Q4)	2016 (Q4)
1	P_BTC_TOT	INTEREST	INTEREST	INTEREST	INTEREST
2	P_BTC_DL	P_BTC_DR	P_BTC_DR	P_BTC_DR	P_BTC_DR
3	P_BTC_DR	P_BTC_BL	P_BTC_BL	P_BTC_BL	P_BTC_BL
4	P_BTC_BL	P_NTC_DL	P_NTC_DL	P_NTC_DL	P_NTC_DL
5	P_BTC_BR	P_NTC_BL	C_COL_TOT	BTC_TOT	CONSUME
6	P_NTC_DL	BTC_TOT	BTC_TOT	BTC_DL	BTC_TOT
7	P_NTC_BL	BTC_DL	BTC_DL	BTC_BL	BTC_DL
8	P_NTC_BR	BTC_BR	BTC_BR	BTC_BR	BTC_BR
9	BTC_DR	DTI_G	DTI_G	DTI_G	DTI_G
10	BTC_BR	DSR	DSR	DSR	DSR

The information in red shading refers to the variables selected at all five points of evaluation, and that in blue shading refers to variables selected at more than four points of evaluation. The order of the variable reflects the order presented in Tables 1 and 2. See Tables 1 and 2 for the description of the variables

appropriate series. According to Kim and Swanson (2018), SPCA is known to work better than PCA in forecasting literature. More specifically, the variable matrix X can be decomposed into the following form:

$$X = F\Lambda' + e,$$

where X is a matrix of $T \times N$ dimensions with N columns and F is a major factor in the $T \times r$ matrix. Λ is a matrix of coefficients in $N \times r$, which are referred to as factor loadings. Typically, the number of factors should be less than the number of columns ($r < N$) to exploit the reduction in data dimensions. For example, in the conventional PCA method, Stock and Watson (2002) estimated factors $\widehat{F}$ using singular value decomposition and then defined the principal component, P, as follows:

$$P_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{Ni}X_N,$$

where the ith factor is the principal component and P_i consists of a linear combination of variable X, where loading $a_{1i,}a_{2i}, \ldots a_{Ni}$ is not all zero in the PCA. However, this not all zero factor loading makes it difficult to identify what the principal component stands for. To address this, Zou et al. (2006) introduced SPCA, which applies a penalty function to limit some of the factor loading to zero by solving the problem below.

$$\max\left(\lambda'\left(X'X\right)\lambda\right) \quad \text{subject to} \quad \sum_{i=1}^{N}\left|\lambda_i\right| \le \varphi, \lambda'\lambda = 1,$$

where φ is the tuning parameter, which is decided by the researcher. This tuning parameter will determine the parsimoniousness of the estimated factors. Another issue in using SPCA is the number of factors. Following Bai and Ng (2002), this study applied the following information criteria:

Table 4 List of variables selected on the basis of theoretical knowledge

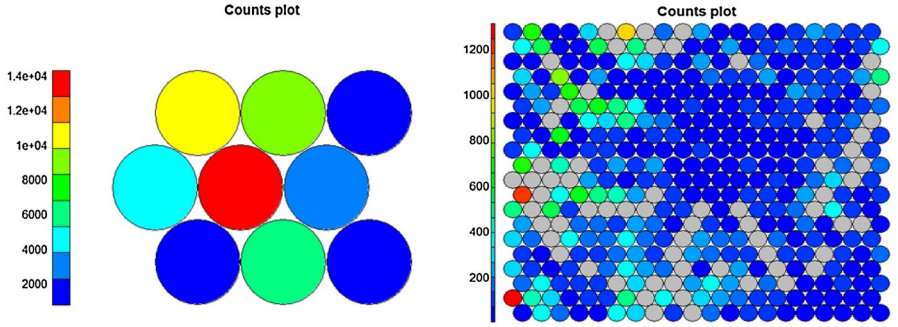
	Variable
1	INTEREST
2	P_BTC_TOT
3	P_BTC_DR
4	P_NTD_TOT
5	P_CREDIT
6	DTI_G
7	INCOME_GG
8	MUL_DEBT
9	DET_CAT
10	BTD_TOT

$$PC(r) = V\left(r, \hat{F}\right) + rh(N, T),$$

where $h(\bullet)$ is a penalty function, $V(\bullet)$ minimizes the Euclidean distance between variables in the dataset and their projection, and r is the number of factors. In the current experiment, two or three were selected as the appropriate number of factors depending on the sampling period. The first sparse principal component could explain approximately 40% of the covariance matrix of the dataset. Consequently, 10 variables without zero factor loadings in the first SPCA were chosen, as listed in Table 3. In other words, the variables noted in Table 3 represent the main explanatory components in the PCA regarding the variance of household debt in Korea.

Although the interest paid by a borrower (INTEREST) was not picked in the first sample period, it was chosen as the first main component in all other periods. The entries in red shading were picked in every sample period. The amount of collateralized loans among personal loans (P_BTC_DR), personal bank loan amount (P_BTC_BL), personal mortgage count (P_NTC_DL), and bank loan amount ratio (BTC_BR) were selected as the main variables in all five sample periods. The blue shaded entries were chosen in every period after 2010 (Q1). Considering the unstable status of 2010 in the aftermath of the GFC, the variables chosen after that period are worth highlighting. The total loan amount (BTC_TOT), total mortgage amount (BTC_DL), DTI rating (DTI_G), and the DSR were also frequently identified in this sense.

SPCA enables the identification of significant variables from a large volume of data and is helpful in explaining the overall covariance structure of the data. Considering that clustering is conducted to identify the unexposed characteristics of the panel, the variables are screened using prior knowledge about the panel. Jung and Kim (2020) analyzed the KCCP data focusing on private business loans and identified a small number of risky borrowers with greater influence on the stability of consumer credit. Although the share of subprime mortgage holders was not very large, they bore the responsibility for triggering the GFC. Jung and Kim (2020) asserted that a small number of "risky" private business loan holders could trigger the next crisis. Hence, this study considered the same KCCP variables selected by Jung and



(a) When setting up a 3 × 3 plane **(b)** When setting up a 20 × 20 plane

Fig. 2 Comparison of SOM forms based on the number of nodes. (Color figure online)

Kim (2020). Table 4 displays a list of 10 variables that directly explain households' ability to repay debt, household debt status, and household debt risk.

Borrowers' interest rate payment (INTEREST) has the most important explanatory power regarding changes in household debt since the interest rate burden on existing loans can be affected by borrowers' future borrowing behavior. Income was used less as a variable to assess borrowers' ability to repay their debt and more as an income level rating (INCOME_GG), which was classified as low, medium, and high. The total personal loan amount (P_BTC_TOT) adds to the measure of the total amount of household debt. The amount of collateralized loans among personal loans (P_BTC_DR) is also added to household debt, as mortgage loans account for most household debt and are relatively less sensitive to macroshocks, such as falling housing prices.

In the KCCP, loans from each borrower are classified into personal loans, private business loans, and personal business loans. Although the nature and vulnerability of a borrower with only personal loans and personal business loans may vary, analyzing loan types is necessary (DET_CAT). The number of accounts in arrears (P_NTD_TOT) and deferred status (P_CREDIT) indicates poor household debt. The DTI rating (DTI_G) assesses borrowers' liability repayment burden. Multiple debt status (MUL_DEBT) is also included, since the higher the number of individual debts held, the greater the burden of repayment and the lesser the ability to repay.

3.2 SOM analysis using SPCA screening variables

When generating an SOM, it is critical to determine how many nodes are required. Too few nodes can result in the allocation of excess data to a limited number of nodes, making it difficult to distinguish the typical characteristics of each cluster, whereas too many nodes can result in empty nodes with no data allocated, making it difficult to identify the contiguousness between the data.

Figure 2 illustrates how the shape of an SOM can vary with the number of nodes. If an SOM is first set to a 3×3 plane, as shown in Panel (a) of Fig. 2, count plots are obtained wherein the nodes in blue indicate fewer data observations and

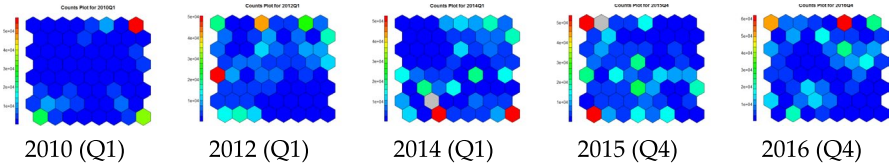


Fig. 3 Counts plot of SOM using SPCA screening variables. (Color figure online)

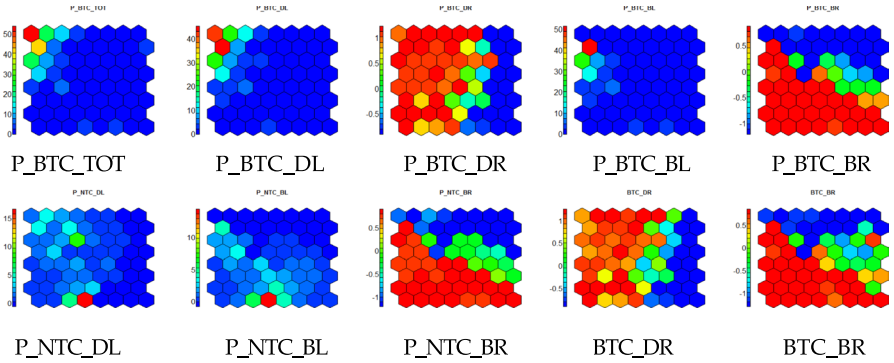


Fig. 4 Heatmap in 2010 (Q1) using SPCA screening variables. (Color figure online)

the nodes closer to red indicate a higher number of observations. In the SOM, a total of nine nodes, with data sharing heterogeneous characteristics, are contained within the same node, making it difficult to describe the characteristics of that cluster. Conversely, with an excess of nodes (20×20), as shown in Fig. 2, empty nodes (in gray) will arise, resulting in inaccurate estimates of the overall SOM. This makes identifying adjacent data on the basis of node characteristics difficult. Many SOM studies generally establish a square node; however, since the node configuration does not necessarily have to be square, such as (3×3) or (20×20), more neighbors are distributed to the same node than in a rectangle. If these nodes are not considered suitable for clustering the entire dataset, then the count plot can be assessed, and the appropriate size can be set by individually increasing or decreasing the number of nodes.

In our analysis, the number of appropriate nodes was established through trial and error for each variable chosen. Figure 3 presents the distribution of borrowers in each node along an 8×8 plane. In each panel, the left bar plot shows the temperature of each node, indicating the number of observations contained in each node. For example, red nodes contain the highest number of observations among all nodes, and gray ones indicate that there are no observations in that node.

In 2010 (Q1), a majority of the nodes comprised approximately 10,000 borrowers, whereas more than 50,000 borrowers with similar characteristics were concentrated at specific nodes on the upper right. In different periods, one or two concentrated nodes existed in different locations. This implies that even nodes in

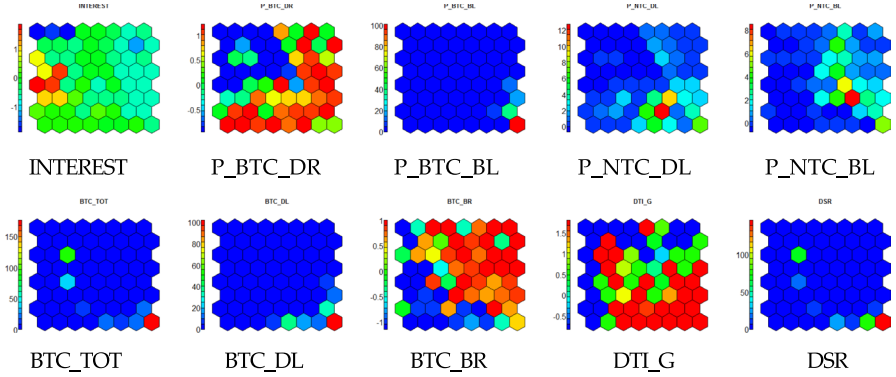


Fig. 5 Heatmap in 2012 (Q1) using SPCA screening variables. (Color figure online)

the same location on the count plot for each sample period did not indicate the same cluster. Since this study considered a dynamic sample over different periods, it was not possible to allocate a certain location for a certain cluster. Moreover, even the characteristics of the clusters were not identical over time. Technically, the SOM algorithm arbitrarily assigns an initial node value and then learns from the data. In other words, if an SOM is configured at each point in the panel data, the neighbors between each node depend on the characteristics of the data; however, the actual locations of the nodes may vary.

Heatmaps are the most prominent visualization technique for SOMs, allowing the distribution of a single variable to be visualized. More specifically, SOMs construct multiple heatmaps for each variable and compare these heatmaps to identify the nodes. Note that the individual sample positions do not change from one heatmap to another. Figure 4 presents the heatmap for individual elements in 2010 (Q1), wherein red nodes indicate the highest value and blue nodes indicate the lowest values on the map, similar to the count plot. As shown in the upper left panel of Fig. 4, a few clusters accounted for the total personal loan amount. This pattern is also revealed in the total secured loan amount (located just right of the first panel). This indicates that the cluster with a higher total personal loan amount held more personal secured loans. Many blue nodes indicate that a majority of the clusters had a lesser total amount of loans than others. A few borrowers assumed large debts, whereas the majority borrowed small amounts. By contrast, the ratio of personal secured loans to the total loan amount (P_BTC_DR) was relatively high for more than half of the nodes, and the ratio of loans from banks to total loan (P_BTC_BR) reveals a different spread of nodes. If the upper left node of a heatmap was tracked series by series, this node would have the characteristics of high total loan and secured loan amounts, and most of the borrowers in this node would have secured loans but not from banking institution; instead, they would probably obtain loans from nonbanking institutions or private money lenders.

Overall, in Fig. 4, the variables that represent household debt size, such as total personal loan amount (P_BTC_TOT), number of personal mortgages (P_BTC_DL),

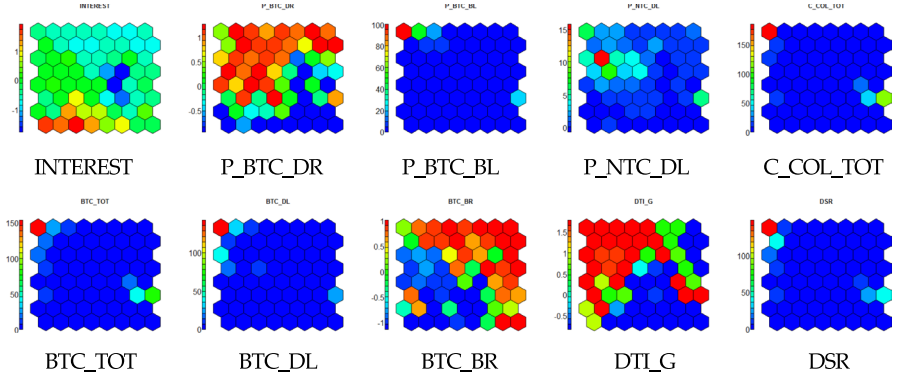


Fig. 6 Heatmap for 2014 (Q1) using SPCA screening variables. (Color figure online)

and personal bank loan amount (P_BTC_BL), showed similar heatmap patterns. This indicates that borrowers with large loans were located on the upper left-hand side. This pattern was similar to the heatmap of loan weight-related variables (BTC_DR, P_BTC_DR), but differs slightly from the heatmap of bank loan weight-related variables (BTC_BR, P_BTC_BR). Regarding personal secured loans (P_NTC_DL) and the number of personal loans from banks (P_NTC_BL), unlike heatmaps based on the loan amount, the borrowers held a large number of secured loans and loans from banks, which are located at the central bottom.

As shown in Fig. 5, the heatmap for 2012 (Q1) included new variables, such as interest rate (INTEREST), DTI rating (DTI_G), and DSR. Note that the location of nodes in 2012 (Q1) did not match with those in 2010 (Q1). However, the node with the highest personal total loan amount was identified in the upper middle panel of Fig. 5. Notably, the interest for this node was not the highest among all nodes. The node with the highest interest paid was located on the opposite side of the map, indicating that these two nodes are distinguishable by other characteristics. Regarding the total loan amount, as shown in the bottom leftmost panel, there were two prominent nodes in opposite directions of the node with the highest total loan amount, indicating that the characteristics of these nodes are also distinct. Moreover, these two nodes have higher DSR, as shown in the bottom rightmost panel of Fig. 5. Remarkably, the nodes with high INTEREST did not match nodes with high total personal loan amount (BTC_TOT) and did not have a high DTI ratio or DSR. This implies that high-interest nodes did not necessarily pose a higher risk.

In 2012 (Q1), five variables of personal secured loan ratio (P_BTC_DR), personal loans from banks (P_BTC_BL), number of personal secured loans (P_NTC_DL), number of personal loans from banks (P_NTC_BL), and loan from banks ratio (BTC_BR), revealed a 180° rotation of the 2010 (Q1) heatmap. Thus, it can be concluded that the differences between these two periods were insignificant. However, the temperature of the personal secured loan ratio (P_BTC_DR) and loan from banks ratio (BTC_BR) was lower in 2012 (Q1) than in 2010 (Q1). A lower temperature indicates that the number of red nodes decreased or that red nodes turned into orange or green ones. This can be interpreted as the personal secured loan ratio

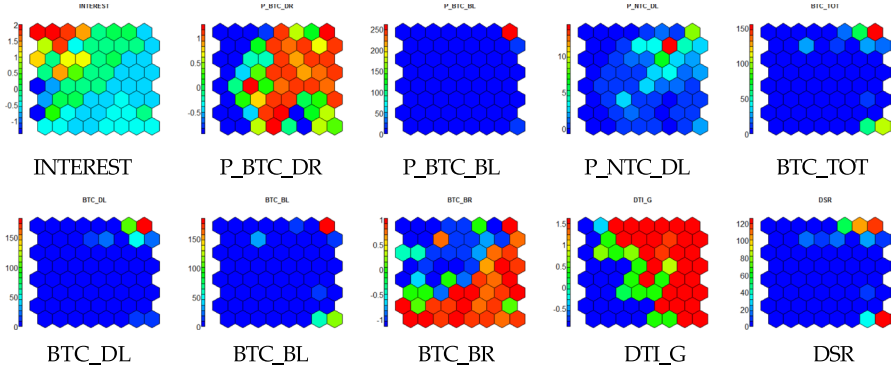


Fig. 7 Heatmap for 2015 (Q4) using SPCA screening variables. (Color figure online)

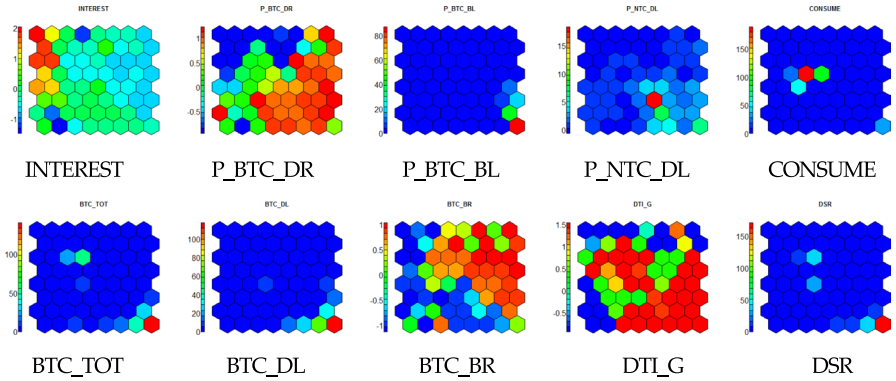


Fig. 8 Heatmap for 2016 (Q4) using SPCA screening variables. (Color figure online)

and loan from banks ratio declined from 2010 (Q1) to 2012 (Q1).[6] By contrast, the DTI_G had a similar map outlook to P_BTC_DR, whereas the DSR did not. The DSR was more similar to the BTC_TOT or the BTC_DL, implying that the DSR is more significantly affected by private business secured loans than by personal secured loans. This indicates that private business loans have an important role in household debt analysis.

Figures 6 and 7 present the heatmaps for 2014 (Q1), when consumer credit began to soar, and for 2015 (Q4), when consumer credit growth peaked. These two heatmaps do not demonstrate any dramatic changes compared with the heatmap for 2012 (Q1) but the structure of the heatmaps changed. For example, the DTI ratio in 2012 (Q1) had a disordered shape; however, in 2014 and 2015, the separation between

[6] This result is consistent with consumer credit statistics at the aggregate level. According to the Bank of Korea's consumer credit report, which presents information regarding loans issued by banking institutions, the secured loan ratio and loan from banks ratio decreased from 46.1% and 55.4% to 45.6% and 52.5%, respectively. Note that these statistics were collected from the banks.

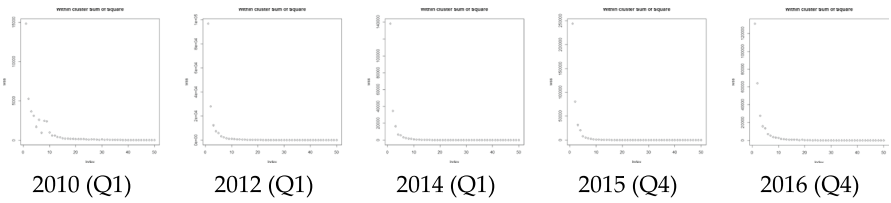


Fig. 9 Sum of squares in SOM using SPCA screening variables

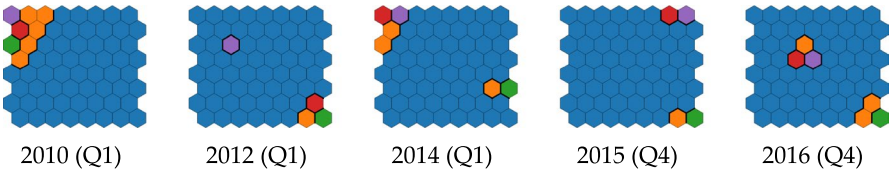


Fig. 10 Cluster analysis of SOM using SPCA screening variables. (Color figure online)

high- and low-DTI nodes became more evident. Moreover, the aggregate and quality indicators, such as total loan amount (BTC_TOT) and DSR, both became hotter nodes. This means that although the structure of household debt did not change significantly, the aggregate level and quality of household debt changed significantly.

Figure 8 presents the heatmap for 2016 (Q4), the final period in the analysis. Overall, the map structure of 2016 (Q4) had more similarity to that of 2014 (Q1) than 2015 (Q4). The main reason for this is that the credit surge in the 2014–2015 period was relatively stabilized by 2016. However, quality indicators, such as the DTI ratio, became hotter, indicating that more nodes increased the DTI ratio, although the aggregate indicator decreased in 2016.

Since the variables were screened using SPCA, the SOM analysis became easier because of the smaller number of heatmaps. When 10 variables were chosen from among 50, nine out of 10 variables were similarly selected over the different sample periods. This reveals no dramatic changes in the household debt structure debt during the various sample periods, indicating that the target variable heatmaps signify irregularities in advance. Moreover, these heatmaps allow the observation of microstructural changes in household debt at the aggregate level during consumer credit surges.

Clusters of borrowers were generated to combine information regarding the characteristics of nodes in the heatmap. A key issue in cluster analysis is determining how many clusters are appropriate. The elbow method was used to calculate the sum of squares within a cluster, by increasing the number of clusters and then selecting the appropriate number of clusters when this sum dropped rapidly. As shown in Fig. 9, the appropriate number of clusters was five on average over the sample periods. More clusters were identified in 2010 (Q1) so that a different number of clusters could be used for each period; however, this would make it difficult to obtain a unified interpretation over time.

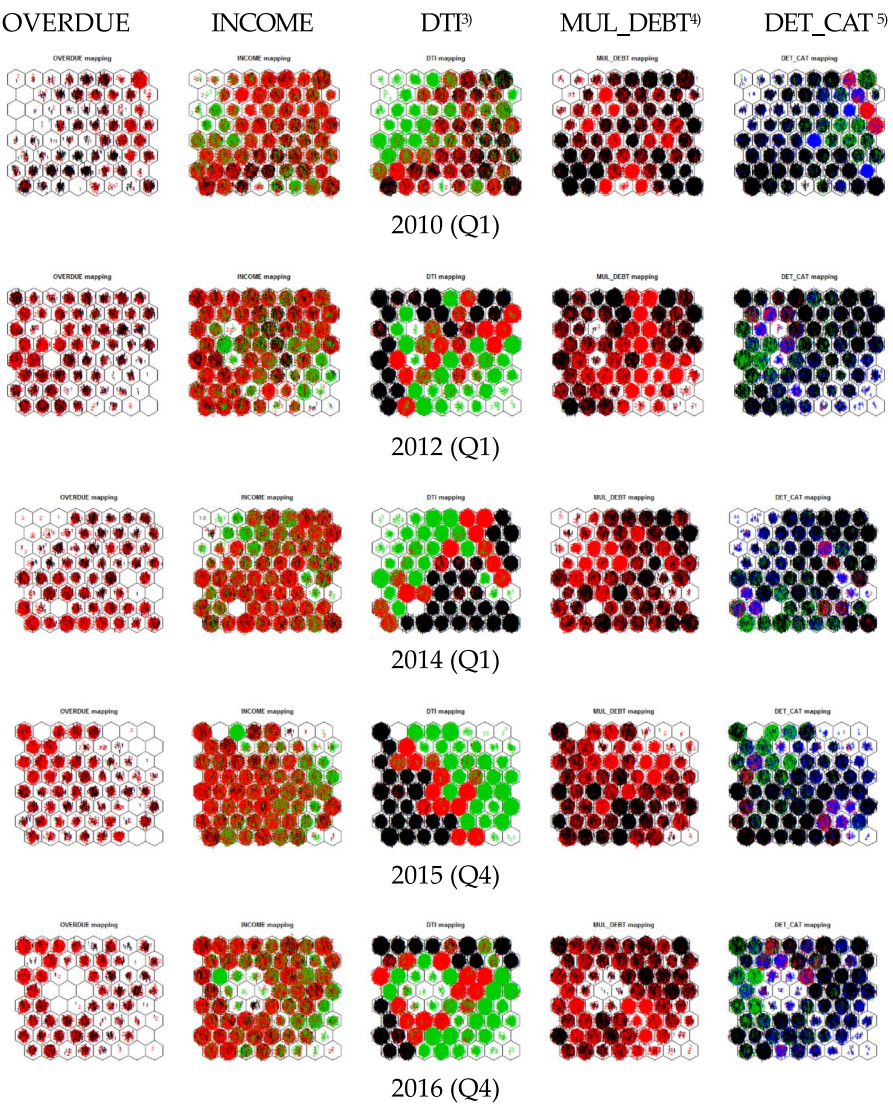


Fig. 11 Mapping results using SPCA screening variables. (1) 0 (none): no late payment, 1 (black): within 90 days of due date, 2 (red): more than 90 days in arrears. (2) 1 (black): low income, 2 (red): middle income, 3 (green): high income. (3) 1 (black): DTI ratio of below 1, 2 (red): DTI ratio of 1–3, 3 (green): DTI ratio of 3 or higher. (4) 0 (none): no loan, 1 (black): 1–2 loans, 2 (red): over 3 loans. (5) 0 (none): no loan, 1 (black): personal loans only, 2 (red): private business loans only, 3 (green): loans from private lenders, 4 (blue): both personal and private business loans. (Color figure online)

Figure 10 shows the results of a cluster analysis. In any period, there was typically one huge cluster and the rest of the clusters had relatively small numbers of nodes. Considering that most borrowers had good credit scores and were able to pay back their debts on time, the one huge cluster represents this type of

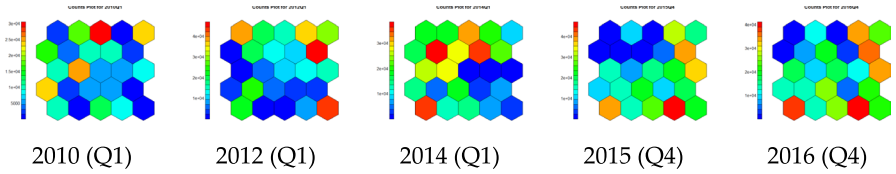


Fig. 12 Counts plot of SOM using prior knowledge screening variables. (Color figure online)

borrowers. From a policy implication perspective, more attention was given to the other clusters. Interestingly, the remaining clusters were gathered together in 2010 (Q1), indicating that they may have had several characteristics in common. However, those clusters subsequently separated.

The analysis of heatmaps is not useful for identifying the characteristics of each cluster, because they locate observations on the basis of each variable. Therefore, it was necessary to map the cluster specification and variables within the same space. However, visualizing continuous variables on a two-dimensional plot is unsuitable. Consequently, some variables were converted into discrete series, such as income, DTI ratio, and loan type. Figure 11 presents the mapping results of the observations based on the class of each variable. For example, overdue loan status was assigned values of 0 for no overdue loans, 1 for loans overdue for less than 90 days, and 2 for loans overdue for more than 90 days. Note that 0 was assigned for no overdue loan status and is not depicted in Fig. 11. The left panel of Fig. 11 shows the overdue status for each node. This map was matched with the cluster result so that the characteristics of the cluster could be identified. According to Fig. 11, borrowers in four clusters did not have long overdue loans. By contrast, there were many observations with long overdue loans on the upper right side of the map.

As shown in the second column of Fig. 11, borrowers' income levels were categorized into three, namely low, middle, and high. For example, in 2010 (Q1), high-income holders are located in the upper left and bottom right of the map. The nodes on the upper left side are categorized into minority clusters. The third column represents the DTI ratio, similar to the first two columns. The upper left side of the map represents a high DTI level, indicating that the minority clusters usually included high-income holders and high-DTI borrowers, simultaneously. The fourth and fifth columns of Fig. 11 indicate whether borrowers have multiple loans and the kind of loans they hold, where 0 represents no loan, 1 represents private loans only, 2 represents business loans only, 3 represents loans from private lenders, and 4 represents both private and business loans. Note that in the case of loans from private lenders, borrowers always hold private and business loans before obtaining loans from private lenders.

Borrowers belonging to four minority clusters from 2010 (Q1) to 2016 (Q4) appear to have held both personal and private business loans, although they had few overdue loans, good income, and relatively high DTI ratios. Meanwhile, no clear trends regarding whether borrowers held multiple debts were identified. Given their good income but high DTI ratio and concurrent business loans, it can be assumed that several self-employed borrowers with large loans but relatively good repayment

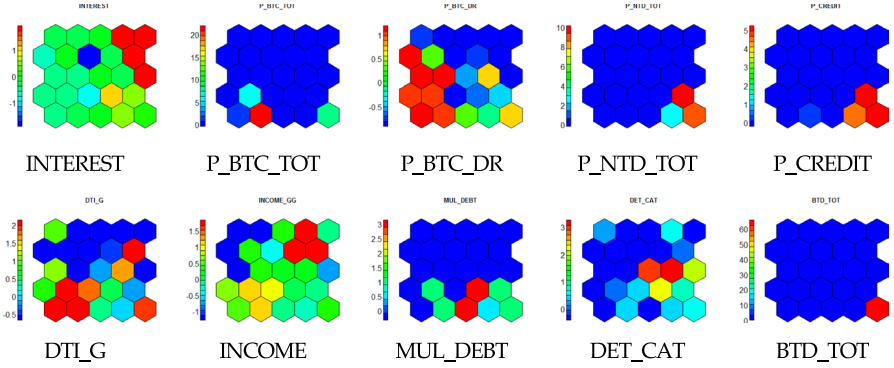


Fig. 13 Heatmap for 2010 (Q1) using prior knowledge screening variables. (Color figure online)

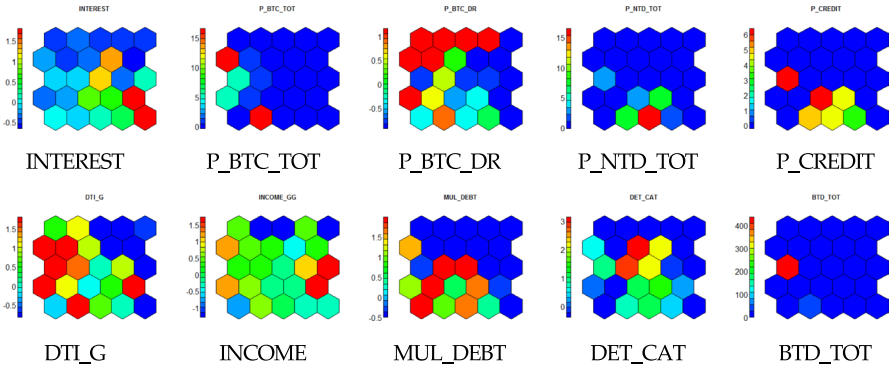


Fig. 14 Heatmap for 2012 (Q1) using prior knowledge screening variables. (Color figure online)

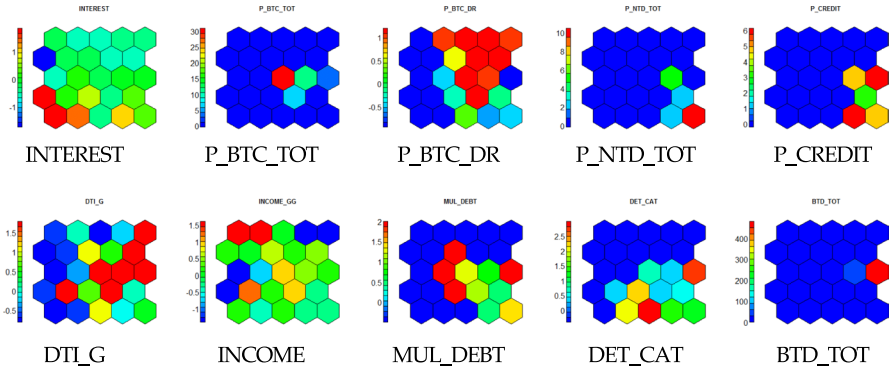


Fig. 15 Heatmap for 2014 (Q1) using prior knowledge screening variables. (Color figure online)

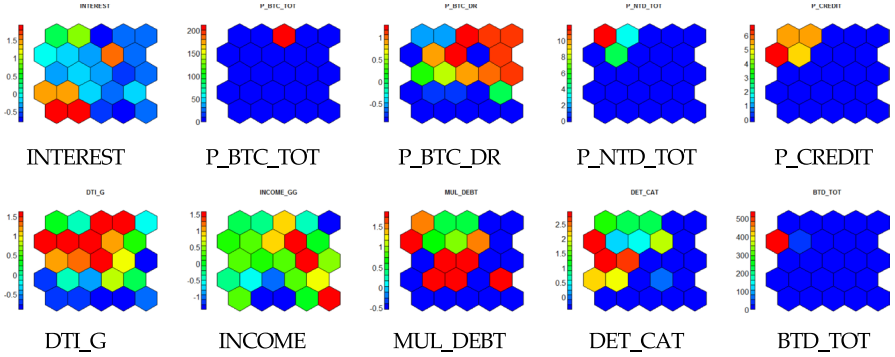


Fig. 16 Heatmap for 2015 (Q4) using prior knowledge screening variables. (Color figure online)

ability were included in these minority clusters. However, clearly describing the unique characteristics of these groups is challenging, since the main large cluster holds all the different borrowers.

3.3 SOM analysis using prior knowledge screening variables

The SOM was estimated on the basis of the selected variables shown in Table 4. Figure 12 presents the count map of KCCP on a 5×5 plane. Since the variables were selected over the sample period, a smaller number of nodes could represent all observations; thus, the map was constructed on only a 5×5 plane. Since the number of nodes was lesser than that in the previous analyses, the density of nodes became higher and no gray nodes with no observations were evident.

Figures 13, 14, 15, 16 and 17 present the heatmaps for the selected variables for five sample periods. In 2010 (Q1), as confirmed earlier, high-interest payers were not high-loan holders based on the INTEREST and BTD_TOT variables. However, high-loan holders were also likely to have overdue loans. The DTI ratio is hotter at the bottom left of the map, indicating that high-DTI borrowers had a higher personal loan amount (P_BTC_TOT) but not from the perspective of the total loan amount (BTD_TOT).

Borrowers who paid high interest rates had a higher share of secured loans and higher DTI ratios. The former result is consistent with the stereotyped fact that low creditworthiness results in higher credit-bearing borrowers holding more credit-based loans than mortgage loans. In the analysis of overdue status-related variables, such as the number of loans past due, the number of overdue days, and the ratio of overdue loans, the fragile borrowers located on the bottom right side of the SOM belonged to middle-income households and had multiple loans. Meanwhile, the DTI ratio was generally higher for borrowers in the green and yellow nodes (more small- and medium-income borrowers) than blue nodes, indicating lower income in the income heatmap. Particularly, high-income borrowers located in the right-hand center demonstrated a significantly high DTI ratio, despite their high income. This

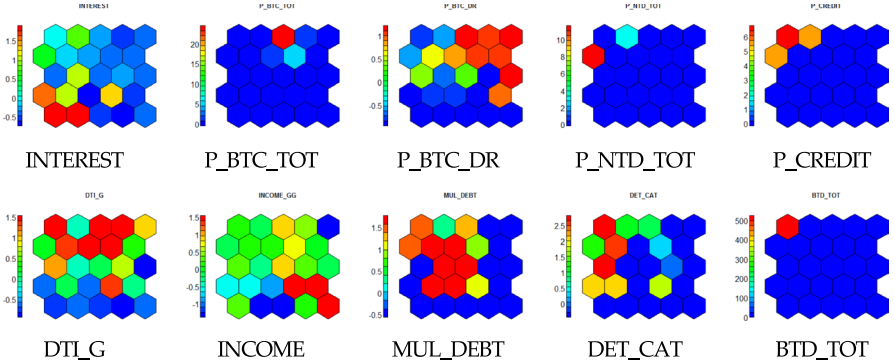


Fig. 17 Heatmap for 2016 (Q4) using prior knowledge screening variables. (Color figure online)

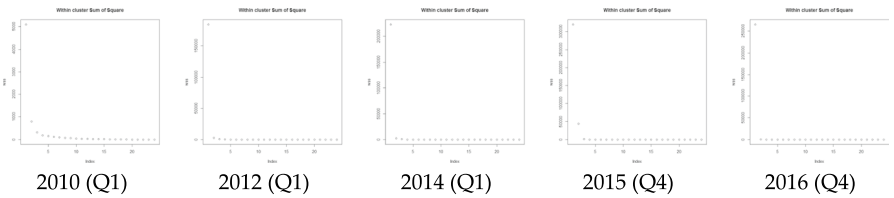


Fig. 18 Sum of squares in a cluster of SOMs using prior knowledge screening variables

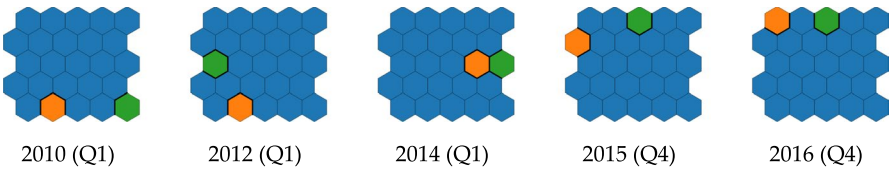


Fig. 19 Clustering analysis of the SOM using prior knowledge selection variables. (Color figure online)

is because these borrowers pay high interest rates and hold mostly credit rather than mortgage loans. Additionally, during 2010 (Q1), income level was not a key factor distinguishing the characteristics of borrowers, given the presence of red-and-blue-colored nodes, representing high-income borrowers, and blue-colored nodes, representing low-income borrowers.

The purpose of this research was to identify the characteristics of fragile borrowers. To this end, the map was analyzed focusing on borrowers with overdue status. Borrowers with overdue loans did not hold large personal loan amounts but had large total loan amounts, indicating that this type of borrower may have held private business loans or loans from private lenders and had multiple loans. Moreover, their income levels and interest payments were located in the middle. Moreover, their DTI ratio level was not in the high group but in the middle group. Thus, it

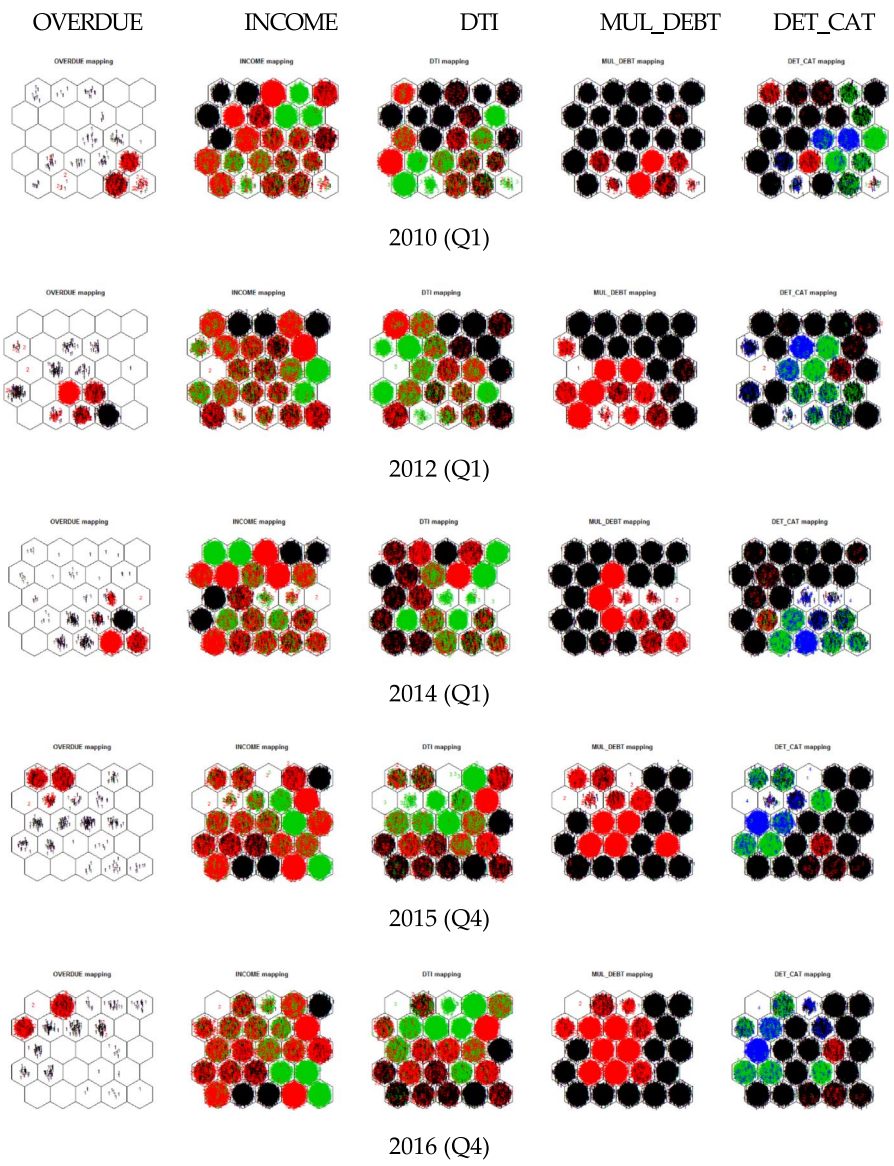


Fig. 20 Mapping results using prior knowledge screening variables. Refer to Fig. 11. (Color figure online)

can be summarized that these fragile borrowers cannot be characterized as unstable because their loan amounts and interest payment levels were not high enough to be monitored.

The number of clusters was also estimated using the sum of squares, as shown in Fig. 18. Two out of three clusters were found to be appropriate based on the

elbow method. The three clusters were applied for all sample periods. Consequently, most of the borrowers were allocated to one large cluster, whereas a small number of borrowers were separated into the remaining two clusters, as shown in Fig. 19. Although the number of clusters is less than in the case of variables selected using SPCA, there is one large common cluster. Considering that a majority of the borrowers were faithful, this result is not surprising. These two clusters were usually close to each other, indicating that the clusters share similar characteristics. Next, as discussed earlier, the characteristics of the small clusters were identified using the mapping results.

The mapping results for the selected variables are provided in Fig. 20. Unlike the SPCA results, overdue borrowers were concentrated in the bottom right of the map in 2010 (Q1). Although the node in the green cluster did not have many observations, most of them were long-term overdue borrowers, and this is also the case for the node in the orange cluster. Interestingly, many observations were clustered next to the node in the green cluster. This suggests the existence of several cases in which borrowers had very similar characteristics to those in the green cluster, indicating that these close observations are likely to move to the green cluster and could pose a potential threat if the green cluster is fragile. The income of borrowers in these two clusters is uncertain; however, these are generally middle- or high-income borrowers. Moreover, these clusters have high DTI ratios and tend to hold multiple debts; however, this is not a necessary condition because dense nodes of borrowers with multiple debts are evident. These borrowers tend to have both personal loans and private business loans but do not have loans from private lenders. This identification was effective for the rest of the sample periods.

4 Conclusions

In this study, the characteristics of clusters of consumer credit borrowers in Korea were mapped using an SOM, which is useful in big data environments and for visualizing clustering results along a two-dimensional plane. Five major points of 2010–2016 were selected using statistical techniques and prior theoretical knowledge, such as SPCA, and then using the SOM for household debt borrowers [2010 (Q1), 2012 (Q2), 2014 (Q1), 2015 (Q4), and 2016 (Q4)].

The cluster analysis found no significant change in borrower characteristics during the sample period. This finding suggests the existence of one big cluster and several small clusters. The large cluster represents a majority of the borrowers, who honored their loans expediently, whereas the small clusters, which represent a minimal number of borrowers, comprise borrowers with high DTI ratios and multiple debts, such as both personal and private business loans. The number of overdue debts in these groups was not high but could potentially trigger the next crisis. Moreover, the quantitative increase in consumer credit during the sample periods did not change the qualitative structure of household debt, since the SOM results were

robust for the sample periods. This was because no events were dramatic enough to reveal a structural change in this study's results.

The analysis had a limitation in that it was insufficient for clearly identifying the unique characteristics of distributed large clusters of heterogeneous borrowers. This is because the household debt database did not include congruent information, although the price of collateralized assets was an important variable that explained the characteristic qualities of borrowers, given that property-backed loans accounted for the bulk of household debt. Future research could collect more data to evaluate the dynamics of borrowers' characteristics in the event of micro- or macroshocks, particularly following the advent of the coronavirus disease 2019 pandemic. For this purpose, local-level data, such as the RGDP and the regional unemployment rate, should be analyzed in conjunction with consumer credit data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-021-02120-5.

Acknowledgements The author is deeply grateful for the financial support of the Bank of Korea and appreciates the valuable comments received from Dohoon Pyeon and Hosung Jung as well as the research assistance provided by Mookyung Song and Sungjoo Yoon.

Funding This study was funded by the Bank of Korea.

Declaration

Conflicts of interest The author declares that he has no conflict of interest.

References

Alter A, Feng AX, Valckx N (2018) Understanding the macro-financial effects of household debt: a global perspective. IMF Working Paper 18/76, International Monetary Fund

Bai J, Ng S (2002) Determining the number of factors in approximate factor models. Econometrica 70:191–221

Baron M, Xiong W (2017) Credit expansion and neglected crash risk. Q J Econ 132(2):713–764. https://doi.org/10.1093/qje/qjx004

Brunnermeier M, Palia D, Sastry KA, Sims CA (2019) Feedbacks: financial markets and economic activity. Working Paper 257, Princeton University, Department of Economics, Center for Economic Policy Studies

Cheng IH, Raina S, Xiong W (2014) Wall street and the housing bubble. Am Econ Rev 104:2797–2829

Claveria O, Monte E, Torra S (2016) A self-organizing map analysis of survey-based agents' expectations before impending shocks for model selection: the case of the 2008 financial crisis. Int Econ 146:40–58

Cox TF, Cox MA (2001) Multi-dimensional scaling. Chapman & Hall/CRC, Boca Raton

Eklund T, Back B, Vanharanta H, Visa A (2008) Evaluating a SOM-based financial benchmarking tool. J Emerg Technol Account 5:109–127

Hahm JH, Kim JI, Lee YS (2010) Risk analysis of household debt in Korea: using micro CB data. KDI J Econ Policy 32:1–34

Jordà Ò, Schularick M, Taylor AM (2016) The great mortgaging: Housing finance, crises and business cycles. Econ Policy 31:107–152

Jung H, Kim HH (2020) Default probability by employment status in South Korea. Asian Econ Pap 19(3):62–84

Kim YI, Byun DJ (2012) Risk assessment of Korean households' indebtedness (Evidence from CB Data). KDI Policy Research Series
Kim D, Kim K (2010) The stress test of household loan sector considering heteroscedasticity, autocorrelation and conditional loss at given default (LGD). Econ Anal 16(3):119–155
Kim HH, Jung H (2019) System risk based on network of consumer credit panel. Bank of Korea Working Paper23
Kim YS, Lim KK (2013) A study on mortgage loan borrowing behavior of Korean households: a micro perspective. Bank of Korea Working Paper 2013-9, Bank of Korea
Kim HH, Swanson NR (2018) Mining big data using parsimonious factor machine learning, variable selection, and shrinkage methods. Int J Forecast 34(2):339–354
Kim YI, Yoo JH (2013) Assessing Korean households' credit risk: stress tests with household level data. Econ Anal 19:59–95
Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69
Lee D, van der Klaauw W (2010) An introduction to the New York Fed consumer credit panel, New York Fed Staff Report Number 479, New York Fed.
Lee DG, Jun SI, Chung J, Byun DJ (2014) A study of the delinquency decision factors and vulnerability of the Korean households with debts. J Money Finance 28:137–178
Marghescu D (2007) Multi-dimensional data visualization techniques for exploring financial performance data. In Proceedings of the 13th Americas conference on information systems, Keystone, CO, USA
Mian A, Sufi A (2014) House of debt. University of Chicago Press, Chicago
Mian A, Sufi A, Verner E (2017) Household debt and business cycles worldwide. Q J Econ 132:1755–1817
Park SW, Bang DW, Park YW (2012) An analysis of the relationship between house prices and bank lending in Korea. J Money Finance 26:107–141
Resta M (2016) VaRSOM: a tool to monitor markets' stability based on value at risk and self-organizing maps. Intell Syst Account Finance Manag 23:47–64
Sarlin P (2013) Decomposing the global financial crisis: a self-organizing time map. Pattern Recognit Lett 34:1701–1709
Sarlin P, Peltonen TA (2011) Mapping the state of financial stability. ECB working paper no. 1382, European Central Bank, Frankfurt, Germany
Stock JH, Watson MW (2002) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. IEEE Trans Neural Netw Learn Syst 11:586–600
Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244
Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15:265–286