2021 대한산업공학회/한국경영과학회 춘계공동학술대회

Manifold 기반 클러스터링을 활용한 가계금융 이질성 분석

황윤태[1] · 이용재[1]

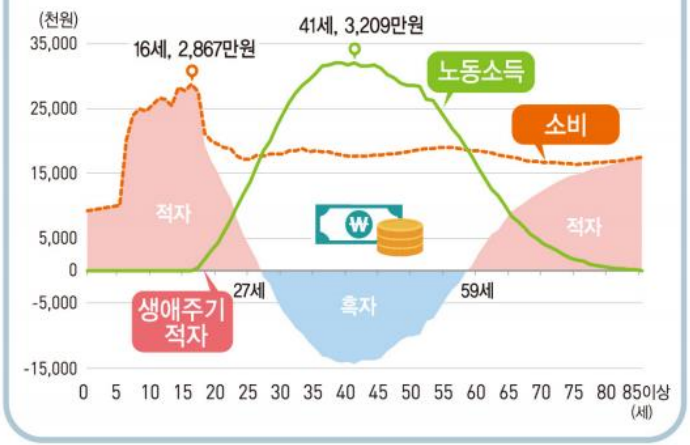
[1]Financial Engineering Lab, Department of Industrial Engineering, UNIST

[1]{yoontae, yongjaelee}@unist.ac.kr

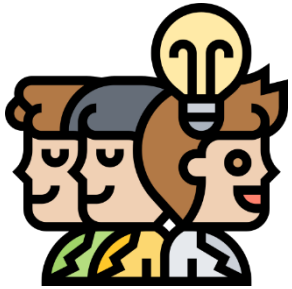
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

1-1. Introduction

Life-Cycle hypothesis



Heterogeneity behavior



- we already know different consumption and saving behaviors with similar levels of income and wealth.

- **The life cycle hypothesis** suggests that households plan their consumption and saving behavior during their lifetime.
- The life-cycle hypothesis suggests that households plan their consumption and savings behavior over their life. The underlying idea is that **all households wish to maintain their lifestyles** over the entire lifecycle

1-1. Introduction

Household finance

- Household finance has not been much researched because it is faced with difficulties in measurement and many constraints that are difficult to capture with existing models. (John Y. Campbell, 2006)
- Similar to corporate financial ratios, household financial ratios were made from household balance sheets, but heterogeneity was not considered. (Baek and DeVaney, 2004 Jacob et al., 2019)

Heterogeneity behavior

- Standard macroeconomics assumes that households are homogeneous and that they are reasonably expected in the future as representative households. (Hall, 1988) However, Lucas (1976) criticizes that this approach is not useful for policy effects even though it reflects the overall propensity to consume.
- Since it is difficult to reflect the heterogeneity of households, there are many studies that have attempted to classify households in a large frame and reflect them in various studies.
- It analyses the effects of monetary and fiscal policy by categorizing households that do not or do not participate in the financial market.(Bilbiie, 2008)To calculate the fiscal multiplier for fiscal policy by dividing households according to income level (Marcus Hagedor 2019)
- It is argued that accurate modeling of the joint distribution of household financial status through household micro data can lead to a broader understanding of household heterogeneity. (Krueger, Mitman, and Perri, 2016)

1-2. Objective

Objective

By analyzing complex patterns of high-dimensional data in a low-dimensional space by using clustering based on manifold learning through non-linear dimensional reduction, this study aims to analyze heterogeneous characteristics of households in terms of household balance sheet and socio-demographic aspects.

Necessities of study

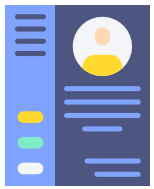
1. In a dynamic macroeconomic model, it is very difficult to calculate the optimal decision-making of an economic entity, so it does not reflect the real economy well. Therefore, there is a need for an attempt to reflect the real economy by analyzing the heterogeneous characteristics of the household.
2. Fiscal and monetary policies that disregard household heterogeneity are difficult to prepare for the impact of downside risks.

Expected benefits of study

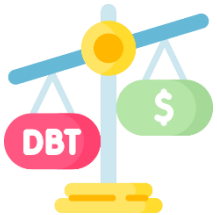
1. By reflecting heterogeneous household types, it is easy to analyze the inequality problems of income, consumption and wealth that exist in reality.
2. Selective fiscal and monetary policies according to household type are expected not only to have a greater effect, but also to increase household utility.

2-1. Materials

MDIS
MicroData Integrated Service
가계금융복지조사
공공 데이터



- Age
- Education
- Income
- Housing form
- Gender
- Job
- Capital area



- Total debt
- Mortgage loan (7 Variables)
- Credit loan (7 Variables)

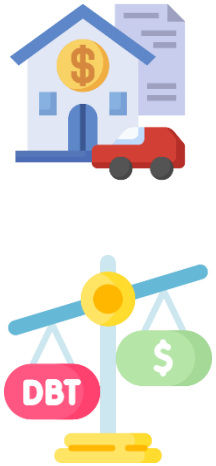
- Total assets
- Financial assets (5 Variables)
- Real estate (5 Variables)

- Total expenditure
- Health care cost
- Education cost
- communication cost
- Transportation cost
- Housing cost
- Food cost
- Other cost

- The Korean Household Finances and Living Conditions Survey (K-HLS) is collected **annually** to investigate the financial soundness of households through a survey on **assets, liabilities, and expenditures**.
- Data were used from 2017 to 2020, with 26,907 households excluding duplicates.

2-2. Data Preprocessing

Household balance sheet



Preprocessing

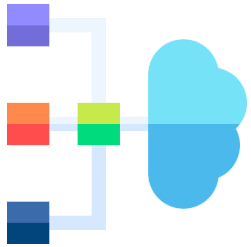
Step 1. Outlier preprocessing

Winsorization of the top and bottom 5% of each variable.

Step 2. Reflecting behavior patterns

Dividing each variable by total consumption expenditure

Preprocessed dataset



- By mitigating **the skewness of each variable**, it is possible to observe household behavior in the middle of the wealth distribution. However, because **each variable rarely informs household financial decision-making**, each variable is divided by total consumption expenditure.
- Since we want to divide the household type according to the detailed asset-liability-expenditure status, each variable is divided by total consumption expenditure to **prevent dominate on the scale** for a specific variable.

3-0. Flow chart

Part 1 Clustering | **Part 2 Cluster Evaluation** | **Part 3 Post analysis**

1. The point where the movement of the cluster change is minimized
2. The point where the fluctuation of the household balance sheet is greatest when the cluster changes

→ 1. Gini decomposition

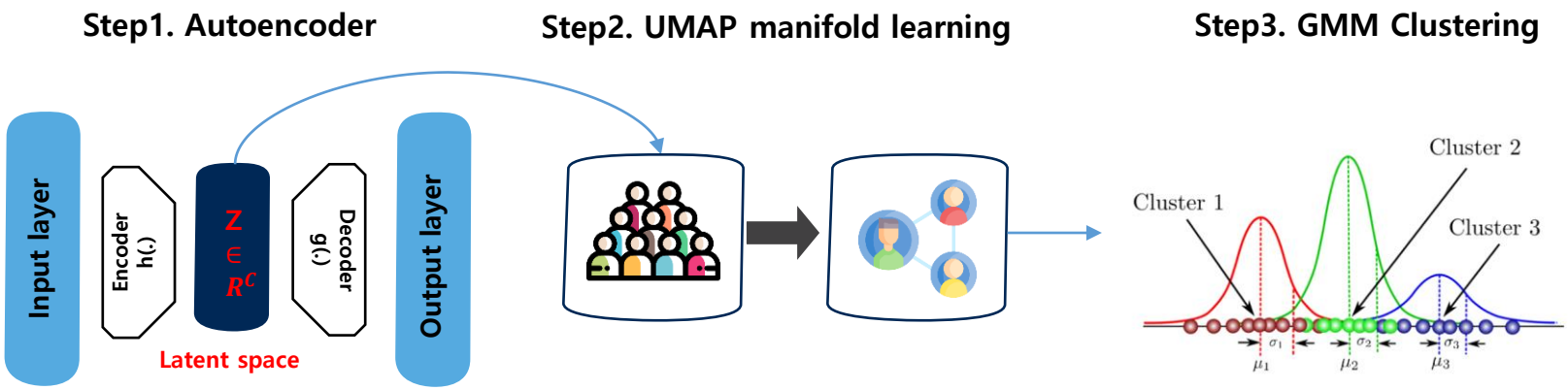
→ 1. Socio-demographic Analysis
2. Transition Matrix Analysis

Part 1	=	Non-linear dimensional reduction(Autoencoder)	+	Manifold learning (UMAP)	+	Clustering (GMM)
Part 2	=	Optimal number of clusters				
Part 3	=	Household balance sheet Analysis(Gini decomposition)	+	Socio-demographic analysis (Logit regression)	+	Cluster movement analysis (Transition matrix)

this study aims to analyze heterogeneous characteristics of households in terms of household balance sheet and socio-demographic aspects

3-1. Proposed Method

Part 1. Clustering the Local Manifold of an Autoencoded Embedding



- The purpose of the autoencoder is to learn the latent space (=manifold). The autoencoder obtains input data (X) compressed Latent space (z) through the encoder network, and output data (y) similar to the initial input data (X) through the decoder network.
- In UMAP, it is possible to find the low-dimensional representation in the manifold by finding the fuzzy toplogy structure in the latent space created by the autoencoder.
- In the last step, GMM clustering is performed with the number of dimensions of latent space $\boldsymbol{C}$ in the low-dimensional representation obtained from UMAP.
- Since local distance information is blurred in the latent space, the distance is preserved through UMAP based on manifold learning. This method performs better than clustering in the latent space.

Deep Clustering Architecture : https://arxiv.org/pdf/1908.05968.pdf

3-2. Result of measure

Part 2. How did we determine the optimal number of clusters?

Definition 4.1. We define order set $X = \{X_{t-k}, X_t\}$ where $t \in \{2018, 2019, 2020\}$ and $k \in \{1,2,3\}$. The set can be expressed in 6-combinations as follows: $\{X_{2017}, X_{2018}\}$, $\{X_{2017}, X_{2019}\}$, $\{X_{2017}, X_{2020}\}$,$\{X_{2018}, X_{2019}\}$, $\{X_{2018}, X_{2020}\}$, $\{X_{2019}, X_{2020}\}$. Then, we can count when a specific person belongs to another Cluster as the time changes. We call that the Changed cluster point, and its sum is called the Changed total cluster point(CP).

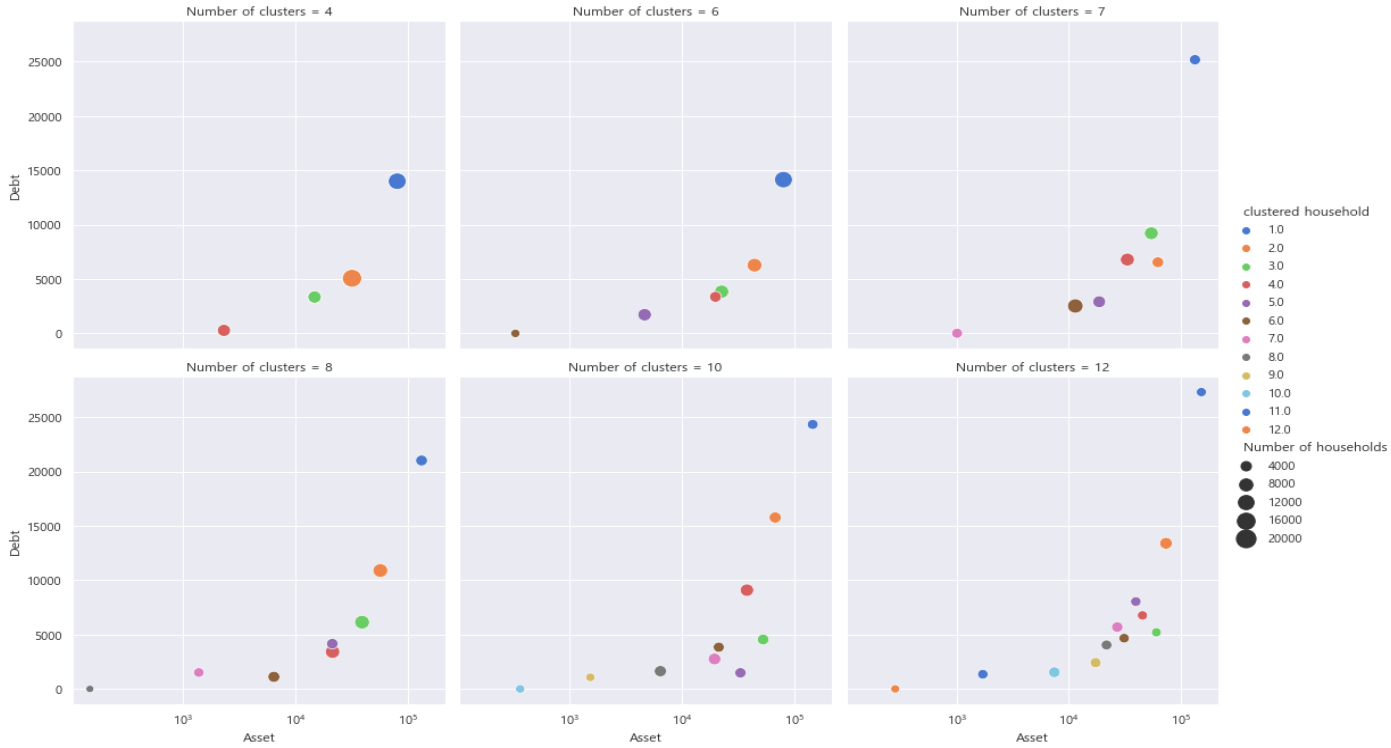
Definition 4.2. we define Cluster measure as $\frac{1}{k}\sum_{i=1}^{k} | X_{t-1}^{(i)} - X_t^{(i)} |$. Where $t \in \{2018, 2019, 2020\}$ and k is an total number of variables.

Measure / Cluster	Asset	Debt	Expenditure	Mean	CP
Cluster 4	0.434	0.734	0.291	0.486	7,473
Cluster 5	0.410	0.691	0.290	0.464	8,587
Cluster 6	0.398	0.692	0.290	0.460	9,554
Cluster 7	0.346	0.675	0.289	0.437	12,202
Cluster 8	**0.366**	**0.670**	**0.295**	**0.444**	**12,772**
Cluster 9	0.344	0.671	0.284	0.433	14,458
Cluster 10	0.343	0.674	0.289	0.435	15,767
Cluster 12	0.336	0.657	0.288	0.427	16,548

- ***Definition 4.1.*** **is an indicator of how often the cluster type changes over time for 26,907 households surveyed from 2017 to 2020.**
- ***Definition 4.2.*** **shows how much the balance sheet has changed on average when the household cluster type has changed.**

3-2. Measure

Part 2. How did we determine the optimal number of clusters?



- As the number of clusters increases, it can be seen that household types are being differentiated at the point where assets are from 100 million won to 1 billion won and liabilities are less than 100 million won.
- In particular, it can be seen that not only the overall size of asset-liability but also detailed asset-liability allocation status is changing, centering on mid-tier households.

4-1. Analysis of cluster type

Part 3. What are the important asset variables that distinguish the type of household?

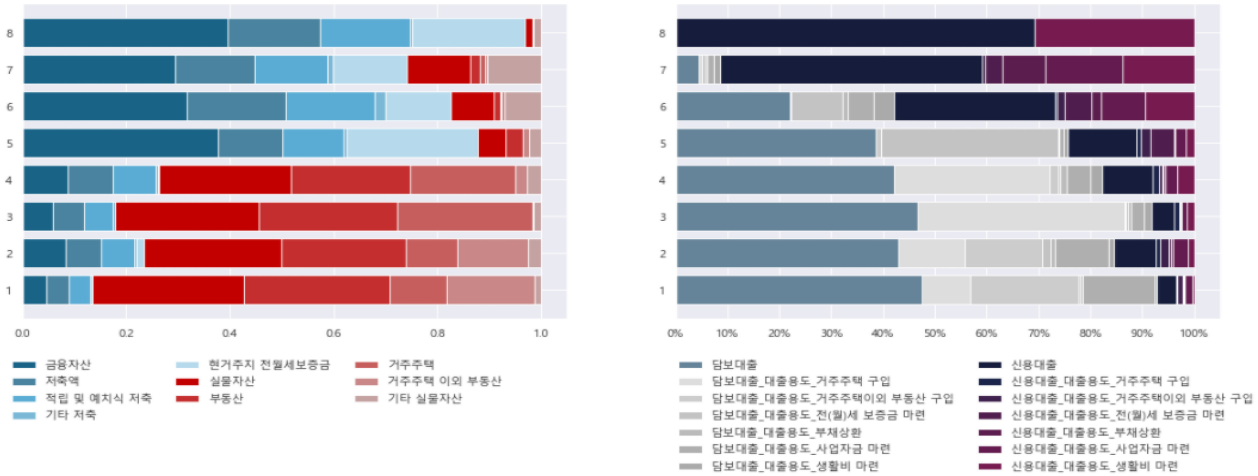
- **We will examine the characteristics of asset allocation of household types classified as deep clustering through the Gini decomposition.**
- **Inequality factor decomposition is examined by decomposing the Gini coefficient into Within-group inequality(G_W), Between-group inequality(G_B) and overlab inequality(G_O).**

Variable	Gini coefficient		Household type Within-group inequality		Household type Between-group inequality		Overlab	
	coefficient	Ratio	coefficient	Ratio	coefficient	Ratio	coefficient	Ratio
Total asset	0.5963	100%	0.0635	10.66%	0.4566	75.57%	0.0761	12.76%
Net asset	0.6101	100%	0.0663	10.86%	0.4584	75.13%	0.0853	13.99%
Financial assets	0.6472	100%	0.0874	13.51%	0.3218	49.72%	0.2378	36.75%
Real estate	0.6742	100%	**0.0647**	9.59%	**0.5533**	82.06%	0.0562	8.33%
Residential house	0.6566	100%	0.0734	11.18%	0.4665	70.97%	0.1171	17.83%
Non-Residential house	0.8744	100%	**0.0819**	9.37%	**0.7558**	86.43%	0.0366	4.19%

- **the meaning that inequality between groups is greater than inequality within groups can be said that each household type is classified well.**

4-1. Analysis of cluster type

Part 3. What are the important variables that distinguish the type of household?



- **Looking at the detailed variables of assets and liabilities according to household type, the following results were obtained.**
- **It was found that the closer to type 1, the more real estate assets were in terms of assets, and more secured loans in terms of liabilities.**
- **The balance sheet of Korean households is known that real estate and financial assets have a ratio of 7:3, but this characteristic was not observed in the case of cluster type 5 or higher.**

4-1. Analysis of cluster type

Part 3. What are the important socio-demographic variables that distinguish the type of household?

Cluster type	Type 1		Type 2		Type 3		Type 4		Type 5		Type 6		Type 7		Type 8	
	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio	Coef	Odds ratio
Constant	-5.224***	0.005	-3.972***	0.019	-2.833***	0.059	-3.025***	0.049	-1.402***	0.246	-0.367***	0.693	**-0.088**	0.916	-0.107***	0.899
Capital area	0.642***	1.900	-0.646***	0.524	0.622***	1.863	-0.977***	0.376	0.794***	2.212	**0.013**	1.013	**0.021**	1.022	-0.205**	0.815
Gender (male)	0.224***	1.251	0.474***	1.607	-0.109***	0.897	**-0.014**	0.986	-0.411***	0.663	-0.225***	0.799	**0.005**	1.005	0.137**	1.147
Number of families	-0.349***	0.705	0.062***	1.064	-0.047***	0.954	0.338***	1.402	-0.098***	0.906	0.004	1.004	-0.068***	0.934	-0.525***	0.592
Education (Under Middle School)																
High School	0.246***	1.279	-0.064*	0.938	**-0.053**	0.949	-0.353***	0.702	**-0.050**	0.951	0.204***	1.227	**0.017**	1.017	-0.599***	0.549
Upper University	0.773***	2.166	0.129***	1.138	0.134***	1.143	-0.751***	0.472	0.310***	1.363	-0.188***	0.829	-0.435***	0.647	-1.330***	0.264
Home Ownership (None)																
Housing lease	0.489***	1.630	0.597***	1.816	**0.063**	1.065	-0.463***	0.629	2.197***	8.995	-0.979***	0.376	-2.421***	0.089	-2.414***	0.089
homeowner	1.486***	4.421	0.656***	1.927	2.806***	16.547	1.648***	5.197	-1.674***	0.187	-2.840***	0.058	-2.911***	0.054	-2.758***	0.063
Age (under 39)																
40~49	0.665***	1.945	0.596***	1.815	-0.304***	0.738	0.385***	1.470	-0.578***	0.561	-0.170***	0.843	-0.177***	0.838	**-0.139**	0.870
50~59	1.330***	3.780	1.000***	2.718	-0.584***	0.558	0.164***	1.178	-0.850***	0.427	-0.368***	0.692	-0.251***	0.778	**-0.019**	0.981
Upper 60	2.539***	12.66	1.235***	3.440	-0.472***	0.624	-0.485***	0.616	-0.957***	0.384	-0.823***	0.439	-0.810***	0.445	**-0.168**	0.845
Income (Low Class)																
Middle Class	0.338***	1.403	0.356***	1.428	-0.119***	0.888	-0.031***	0.734	**-0.043**	0.958	-0.326***	0.722	-0.855***	0.425	-1.753***	0.173
High Class	0.812***	2.252	0.799***	2.224	-0.358***	0.699	**-0.257**	0.773	-0.331***	0.718	-1.018***	0.361	-2.040***	0.130	-2.299***	0.100
Working (work)	**0.012**	1.012	0.628***	1.875	-0.491***	0.612	0.193***	1.213	-0.031	0.970	0.328***	1.388	-0.262***	0.769	-1.091***	0.336
Num of households	5,937		10,644		10,699		10,001		5,614		6,204		4,223		1,598	

*p < .05, **p < .01, ***p < .001

- **The households that are likely to be classified as Type 1 are male, live in the metropolitan area, have high age, academic level, and income level, and have few family members.**
- **On the other hand, households belonging to Type 8 showed opposite results from the socio-demographic characteristics of Type 1 except for the number of family members.**

4-1. Analysis of cluster type

Part 3. What are the important socio-demographic variables that distinguish the type of household?

Type 1

프로파일링

- 수도권 거주
- 정규직 남성
- 60대 이상
- 학사 이상 학력
- 자가 주택 보유
- 고소득
- 2~3인 가족 구성원

Type 2

프로파일링

- 지방 거주
- 정규직 남성
- 40대 이상
- 학사 이상 학력
- 자가 주택 보유
- 고소득
- 2~3인 가족 구성원

Type 3

프로파일링

- 수도권 거주
- 무직 여성
- 39세 이하
- 학사 이상 학력
- 자가 주택 보유
- 저소득
- 2~3인 가족 구성원

Type 4

프로파일링

- 수도권 거주
- 정규직 여성
- 50세 이상
- 중졸 이하 학력
- 자가 주택 보유
- 중간 소득
- 3~4인 가족 구성원

Type 5

프로파일링

- 수도권 거주
- 무직 남성
- 50대 이하
- 학사 이상 학력
- 월세/전세 거주
- 저소득
- 2~3인 가족 구성원

Type 6

프로파일링

- 수도권 거주
- 정규직 여성
- 39세 이하
- 고졸 이하 학력
- 자가 주택 없음
- 저소득
- 2~3인 가족 구성원

Type 7

프로파일링

- 수도권 거주
- 무직 여성
- 39세 이하
- 고졸 이하 학력
- 자가 주택 없음
- 저소득
- 2~3인 가족 구성원

Type 8

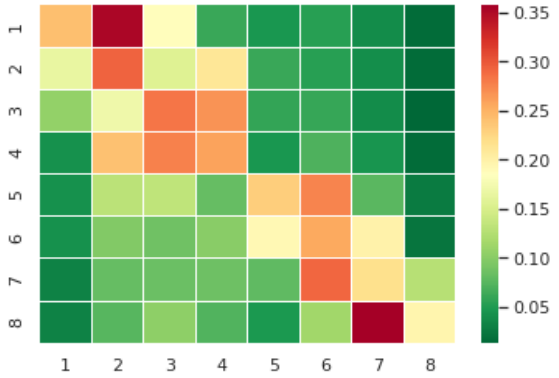
프로파일링

- 지방 거주
- 무직 남성
- 39세 이하
- 중졸 이하 학력
- 자가 주택 없음
- 저소득
- 2인 이하 가족 구성원

4-2. Analysis of cluster type

Part 3. How often do household types change?

type	1	2	3	4	5	6	7	8
1	**0.243**	0.353	0.187	0.062	0.047	0.054	0.038	0.015
2	0.166	**0.295**	0.157	0.212	0.061	0.054	0.040	0.015
3	0.106	0.171	**0.285**	0.268	0.058	0.060	0.039	0.012
4	0.043	0.242	0.277	**0.261**	0.046	0.071	0.045	0.015
5	0.042	0.131	0.133	0.082	**0.232**	0.277	0.076	0.026
6	0.043	0.097	0.088	0.101	0.193	**0.257**	0.199	0.022
7	0.031	0.081	0.085	0.086	0.080	0.292	**0.218**	0.127
8	0.032	0.075	0.104	0.073	0.048	0.114	0.358	**0.197**



- **Mobility measurement :** $\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} \mid i-j \mid p_{ij}$
- **As a result of using the Mobility measurement to quantitatively measure the change between household types, it was found to be 0.262, and the Mobility measurement also showed a low value of 0.299 or less in the change over time.**
- **This means that the movement of relatively distant household types did not appear much, which means that the clustered household types were well classified.**

5. Conclusion

Conclusion

- Our study analyzed household heterogeneity by categorizing household types.
- Basically, household types were classified according to the size of assets and liabilities, but in detail, it was found that real estate in terms of assets and credit loans in terms of liabilities showed a great influence.
- As for socio-demographic factors, it was found that housing ownership and education level had the greatest impact.

Future work

- The classification of household types is largely related to the macroeconomic model.
- As a result of explaining household behavior based on the "reasonable expectation model" without considering the heterogeneity of households, the existing research faced a limitation in not properly reflecting the real economy.
- Our research will create a dynamic macroeconomy model considering the heterogeneity of household characteristics and the financial system to which the household belongs.

Thank you for listening!