Classifying the Elderly based on their Pattern of Labor Market Transitions: The Case of Korea

Inhyuk Choi[*]

This paper aims to classify the elderly in Korea based on the (dis)similarity of their transition patterns in the labor market. To assign the elderly to groups, I apply the *K*-means clustering algorithm to the history of employment transitions observed in the Korean Labor and Income Panel Study. The clustering outcome indicates that older people can be classified into two groups: group A (91.1%) versus group B (8.9%). The transition pattern of group A individuals is characterized by short-term unemployment and long-term employment. Those assigned to group B, in contrast, tend to have experienced long-term unemployment and short-term employment. It is also documented that whether an individual belongs to one group or the other cannot be fully predicted by basic demographics, region, industry, and occupation. Lastly, I provide some empirical evidence that the heterogeneity with respect to employment transitions may be more relevant in predicting one's economic conditions after retirement, and understanding the volatility and persistence of unemployment.

Keywords : *K*-means clustering, Labor market transitions, Elderly

Ⅰ. Introduction

This paper aims to classify the elderly in Korea based on their pattern of employment transitions and study its policy implications. Understanding differences in employment transitions of the elderly is important for two reasons. First, it can be examined whether this type of heterogeneity is an important factor in explaining or predicting the economic situation in old age. The elderly's economic condition can be explained or predicted to some extent by basic demographic variables such as sex, age, and education that are directly observed in the data. However, those variables are not the only factors that affect, for example, homeownership

[*] Associate Research Fellow, Korea Institute of Public Finance, Email: ichoi@kipf.re.kr. The author would like to thank the Korea Labor Institute for providing data used for this study. This paper does not reflect the views of the Korea Institute of Public Finance. The author retains responsibility for any errors and the views expressed.

in old age, and I argue that the pattern of employment transitions when actively participating in the labor market can be a key factor in determining one's economic condition after retirement.

Classifying patterns of labor market transitions is also meaningful in that it allows to investigate whether there is any type of workers that plays a crucial role in explaining the observed volatility and persistence of unemployment. It is well-known that unemployment is sensitive to changes in productivity but its recovery is relatively slow. Although this might be viewed as a common pattern of employment transitions among homogeneous individuals, recent studies find that representative agent models cannot successfully reproduce those stylized facts of the labor market. This implies that the heterogeneity with respect to employment transitions may be related to the volatility and persistence of unemployment, and I argue that this is the case at least in Korea. Moreover, I provide some empirical evidence showing that other forms of heterogeneity (in particular, education level) do not lead to the same conclusion.

In order to assign the elderly to groups (types) based on the (dis)similarity of their transition patterns in the labor market, I apply the K-means clustering algorithm to the history of employment transitions observed in the Korean Labor and Income Panel Study (KLIPS). The K-means clustering algorithm minimizes the total sum of squared distances between the center point of each cluster and the observations within that cluster to classify all observations into K groups. The algorithm, providing relatively little room for arbitrary judgment, allows to classify based on high-dimensional data. Despite these advantages, however, it has not been widely used in the literature studying patterns of labor market transitions in Korea. Thus, the current paper is distinguished from, e.g., Min and Lee (2018) who used the group-based trajectory model to classify patterns of labor force participation among middle-aged and elderly individuals, or Son (2022) who applied sequence analysis to categorize employment histories of the self-employed.

The rest of the paper is organized as follows. Section 2 describes how to implement the K-means clustering algorithm to classify the elderly, and the cross-validation procedure to determine the number of clusters, K I present details of the data used for analysis in Section 3, and report main results in Section 4. A summary and concluding remarks are provided in Section 5.

Ⅱ. Methodology

In the dictionary, "classify" is defined as "arrange (a group of people or things) in classes or categories according to shared qualities or characteristics." From a statistical perspective, the definition can be rephrased as "assign observations to clusters based on the similarity of variables observed and considered." The *K*-means clustering algorithm, whose goal coincides with this definition, is widely used in the recent literature as a tool to categorize workers or firms into different groups (Bonhomme *et al.*, 2019, 2022; Gregory *et al.*, 2022). This paper also uses the algorithm to classify the elderly based on their pattern of employment transitions.

K in "*K*-means clustering" is a parameter representing the number of clusters. It must be chosen by the researcher before the algorithm is implemented. Obviously, the number of clusters significantly impacts the results of clustering. Thus, it is generally considered undesirable for researchers to arbitrarily choose the number of clusters without clear criteria. For this reason, various methods to systematically set the value of *K* have been proposed in the literature, and following Gregory et al. (2022), I use the cross-validation method proposed by Wang (2010).

1. *K*-means clustering

Clustering is a function $\tilde{k}:\{1,2,...,I\}\rightarrow\{1,2,...,K\}$ that assigns each individual $i\in\{1,2,...,I\}$, whose characteristics $\{s_{ij}\}_{j=1}^{J}$ with $s_{ij}\in R$ is observable to an econometrician, to a cluster $k\in\{1,2,...,K\}$.[1] *K*-means clustering is also a function that assigns I individuals to K clusters in a way that minimizes the squared distance between the characteristics of individual i and the mean characteristics of all individuals within cluster $\tilde{k}(i)$ to which individual i belongs. Mathematically, *K*-means clustering can be described as follows:

$$(1) \qquad \min_{\tilde{k}(i)}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}1\left[\tilde{k}(i)=k\right]\left(s_{ij}-s_{kj}^{*}\right)^{2},$$

where $1[\cdot]$ is the indicator function, s_{ij} represents the value of characteristic j for individual i, and s_{kj}^{*} is the average of s_{ij} across all individuals belonging to cluster k, that is,

[1] The following description of the K-means clustering algorithm is based on Hastie *et al.* (2009) and Gregory *et al.* (2022).

(2)
$$s_{kj}^{*} = \left(\sum_{i=1}^{I} \mathbf{1}\left[\tilde{k}(i) = k\right] s_{ij} \right) \times \left(\sum_{i=1}^{I} \mathbf{1}\left[\tilde{k}(i) = k\right] \right)^{-1}.$$

The optimization problem presented in (1) can be solved in an iterative way (see Table 1). First, arbitrarily set K initial centroids, and form initial clusters such that each observation belongs to the cluster whose centroid is closest to it (Step 1). Second, based on the current clustering result, recalculate the centroid of each cluster, and reassign observations accordingly (Step 2). Lastly, repeat Step 2 until the convergence condition[2] is satisfied (Step 3).

<Table 1> The procedure of the *K*-means clustering algorithm

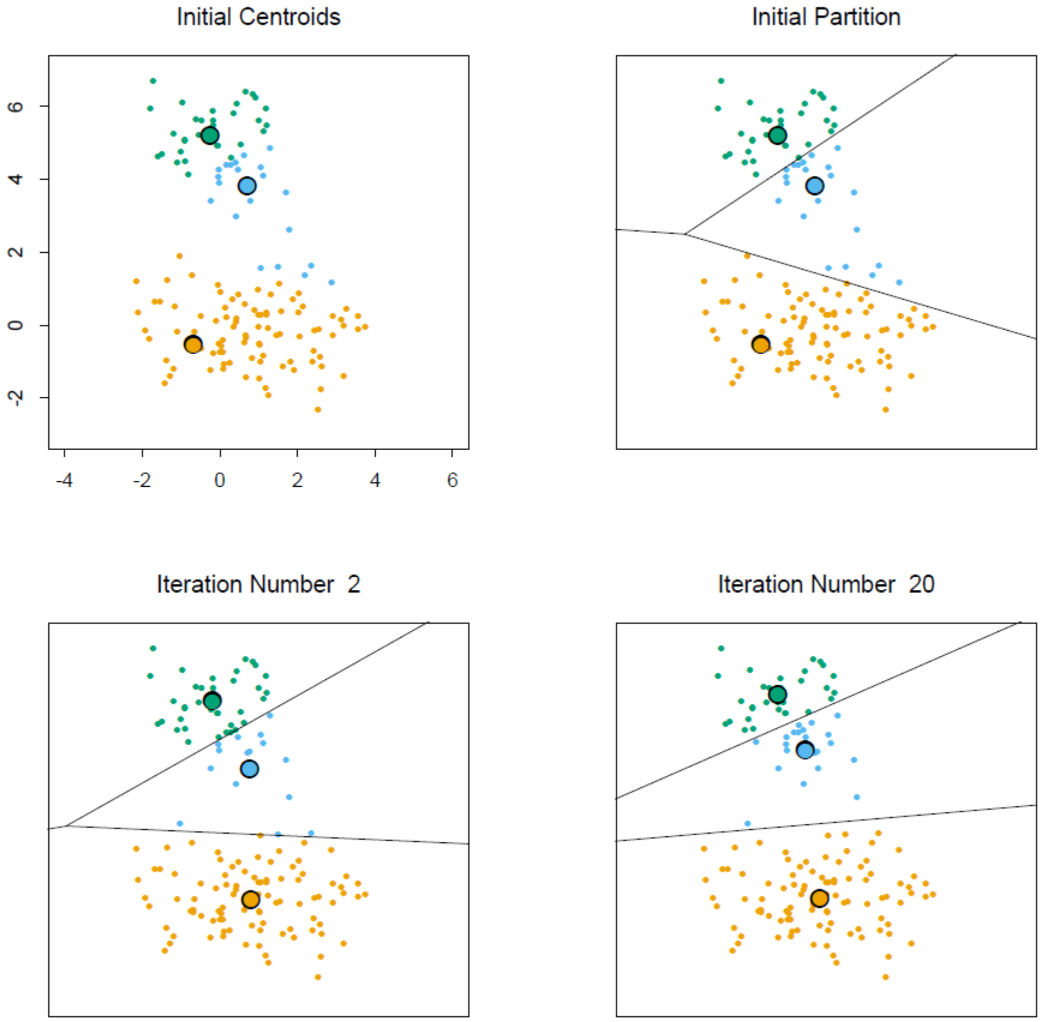
Step	Description
Step 1	Arbitrarily set K initial centroids, and accordingly form initial clusters.
Step 2	Recalculate the centroid of each cluster, and reassign observations if needed.
Step 3	Repeat Step 2 until the convergence is reached.

Notes: The number of clusters K is assumed to be predetermined.

The procedure of the *K*-means clustering algorithm is illustrated in Figure 1, where the clustering process appears to converge after 20 iterations. The *K*-means clustering algorithm is known for converging quickly, even when dealing with high-dimensional data. It should be noted that, however, the clustering results obtained from the algorithm can be sensitive to the initial centroids chosen in the first step. In other words, there is a possibility that the solution obtained from the algorithm is a local optimum, rather than the global optimum. Thus, it is necessary to check the robustness of the clustering results by examining, for example, whether the algorithm produces the same outcome regardless of the choice of initial centroids, although it cannot be guaranteed whether the global optimum is attained or not even after going through such a validation step.

[2] The condition that the clustering results no longer change is typically imposed.

[Figure 1] An example of implementing the *K*-means clustering algorithm



Notes: Figure 14.6 in Hastie *et al.* (2009). The circles with borders represent the centroids of clusters, and the lines represent the partitions resulting from the algorithm.

2. Cross-validation

Before implementing the *K*-means clustering algorithm, one needs to set the value of *K*. As mentioned earlier, the final clustering result inevitably depends on the number of clusters. Therefore, except the cases where K is predetermined for a specific purpose, the value of K should be chosen in a systematic way.

As a naive approach to choose *K*, one might implement the algorithm under various values of $K \geq 2$, and then compare the outcomes based on (1) to choose the optimal value of *K*.[3]

[3] For instance, if the value of (1) is smaller when $K=2$ than when $K=3$, the number of clusters is set

However, this approach is not desirable because the value of (1) tends to decrease as the value of K increases,[4] implying that a more sophisticated approach is needed to determine the value of K Accordingly, I follow Gregory *et al.* (2022) to use the cross-validation approach proposed by Wang (2010}.[5]

The procedure of Wang's (2010} cross-validation method can be summarized as follows (see Table 2). First, divide the entire sample into three subsamples (Step 1). Specifically, randomly allocate 25% of the entire sample to each of two subsamples, S_1 and S_2, and allocate the remaining 50% to subsample S_0. Next, for each $\widetilde{K} \in \{2, 3, \ldots, \overline{K}\}$, classify subsample S_1 to calculate mean characteristic value, denoted by $s^1_{kj}(\widetilde{K})$, for each (k, j) (Step 2.1). Similarly, for each $\widetilde{K} \in \{2, 3, \ldots, \overline{K}\}$, classify subsample S_2 to calculate mean characteristic value, denoted by $s^2_{kj}(\widetilde{K})$, for each (k, j) (Step 2.2). Then, with $\overset{*}{s}_{kj}$ in (1) replaced with $s^1_{kj}(\widetilde{K})$ obtained from Step 2.1, classify subsample S_0 (Step 3.1); and parallelly, with $\overset{*}{s}_{kj}$ in (1) replaced with $s^2_{kj}(\widetilde{K})$ obtained from Step 2.2, classify subsample S_0 (Step 3.2). Lastly, for each $\widetilde{K} \in \{2, 3, \ldots, \overline{K}\}$, calculate the difference between the clustering results obtained from Step 3.1 and Step 3.2, and set the value of $\widetilde{K}$ that minimizes the difference as the number of clusters for the K-means clustering algorithm (Step 4). To sum up, Wang's (2010} cross-validation method can be translated into solving the following optimization problem:

(3)
$$\min_{\widetilde{K} \in \{2, 3, \ldots, \overline{K}\}} \sum_{i=1}^{I_0} 1\left[\tilde{k}_1(i\,;\widetilde{K}) \neq \tilde{k}_2(i\,;\widetilde{K})\right],$$

where I_0 is the number of individuals in S_0, $\tilde{k}_1(i\,;\widetilde{K})$ denotes the cluster that individual i belongs to based on the clustering result from Step 3.1, and $\tilde{k}_2(i\,;\widetilde{K})$ is similarly defined.

to 2.

[4] To be specific, the value of (1) becomes 0 for $K = I$, which forces the approach to set $K = I$, a meaningless result.

[5] The following explanation of how to implement the cross-validation method proposed by Wang (2010} is based on Gregory *et al.* (2022).

<Table 2> The procedure of Wang's (2010) cross-validation method

Step	Description
Step 1	Divide the entire sample into three subsamples: S_0 (50%), S_1 (25%), and S_2 (25%).
Step 2.1	Classify subsample S_1 to calculate the mean characteristic value $s_{kj}^1(\widetilde{K})$ for each (k, j).
Step 2.2	Classify subsample S_2 to calculate the mean characteristic value $s_{kj}^2(\widetilde{K})$ for each (k, j).
Step 3.1	Classify subsample S_0 with $\overset{*}{s}_{kj}$ in (1) replaced with $s_{kj}^1(\widetilde{K})$.
Step 3.2	Classify subsample S_0 with $\overset{*}{s}_{kj}$ in (1) replaced with $s_{kj}^2(\widetilde{K})$.
Step 4	Choose $\widetilde{K}$ that minimizes the difference between the results from Steps 3.1 and 3.2.

The validity of Wang's (2010} method can be intuitively explained. On the one hand, if the value of K is set to be too small compared to the actual (unknown) value, individuals who should be classified into one or more additional clusters get dispersed into K clusters, meaning that each cluster ends up containing a mixture of heterogeneous individuals. As a result, the likelihood of $s_{kj}^1(\widetilde{K})$ and $s_{kj}^2(\widetilde{K})$ being different increases, leading to an increase in the value of (3). If the value of K is set to be too large compared to the actual (unknown) value, on the other hand, individuals who should be classified into one single cluster are forced to form multiple clusters. As a result, again, the likelihood of $s_{kj}^1(\widetilde{K})$ and $s_{kj}^2(\widetilde{K})$ being different increases, leading to an increase in the value of (3). To summarize, the value of (3) cannot be minimized if K is set smaller or larger than the actual one, which allows for finding the true value of K

There are several practical issues that arise when implementing the cross-validation method. First, since considering all positive integers greater than 1 as potential values for K is impractical, it is necessary to set an upper limit (denoted by $\overline{K}$) on the possible values that K can take. To the best of my knowledge, however, there is no specific way recommended in the literature, so I set $\overline{K}=5$ considering Gregory *et al.* (2022) report $K=3$ for the U.S. labor market. Second, when comparing the clustering results obtained from Steps 3.1 and 3.2, a certain criterion is required to align the clusters with each other. Again, there is no specific recommendation in the literature to the best of my knowledge. Thus, based on the observation that the size distribution of clusters formed in Step 3.1 is similar to one from Step 3.2, I correspond the n-th largest cluster formed in Step 3.1 to the n-th largest cluster formed in Step 3.2. Lastly, one needs to take into account the sensitivity of the K-means clustering algorithm to the initial centroids. In order to mitigate this issue, for each $\widetilde{K} \in \{2, 3, \ldots, \overline{K}\}$, I repeat 200 times Steps 2.1 to 3.2 with different initial centroids, and use the average across these repetitions to calculate the value of (3).

Ⅲ. Data

In order to classify the elderly based on their pattern of employment transitions, data containing detailed individual employment histories over a relatively long period are needed. In particular, it is essential to have information on job duration, and the incidence or frequency of short-term and long-term unemployment. Thus, data that can provide rich and accurate information about the start and end dates of each job can be regarded as the most suitable for this study, and the Korean Longitudinal Study of Ageing (KLoSA) can be considered as a possible choice. Indeed, the KLoSA, developed and maintained by the Korea Employment Information Service, surveys individuals aged 45 and above (excluding Jeju Island residents), and provides comprehensive work history information throughout the lifetime for each individual. However, only the start and end *years* of each job are recorded in the KLoSA, making it challenging to obtain detailed information about (un)employment spells. For this reason, this study utilizes the Korean Labor and Income Panel Study (KLIPS) for analysis. Although the KLIPS is not specifically focused on the elderly population, it provides a special dataset (called Job History) containing detailed information on employment history at the individual level. Thus, in Section 3.1, I provide a brief introduction to the dataset which has a different structure compared to general longitudinal surveys. Then in Section 3.2, I describe how to construct variables needed for the *K*-means clustering algorithm, and report summary statistics.

1. KLIPS

The Job History data is constructed at the (individual's) job level, so it includes information about all the jobs that an individual has experienced since his/her initial entry into the labor market.[6] The most recently released data provide information about 85,949 jobs that 29,293 individuals have held or currently hold, meaning that 2.93 jobs per individual are observed. For each job, one can observe its start and end dates, industry, occupation, working hours, wages, etc.[7]

Among the various job-related variables contained in the data, the start and end dates of the job are most important for the purpose of this paper. As illustrated in Table 3, the start and end dates of a job are surveyed up to the daily level. However, the date variables are missing

[6] The following description about the Job History data is based on Jang *et al.* (2023), pp. 98-105.

[7] The data prior to the first survey were collected in a retrospective way, so the start and end dates of the job are missing in many cases.

or unknown in many cases, so I calculate (un)employment spells on a monthly basis in what follows. Accordingly, for individual 111 in Table 3, for example, the duration of the first job is recorded as 120 months. Meanwhile, individual 111 is observed to have quit the first job in March 1998 and started the second one in June 1999. I regard him/her as having a 15-month unemployment spell, considering the practical difficulty of determining unemployment status accurately from the data. Of course, this does not coincide with the general definition of unemployment, so readers should keep this difference in mind when interpreting the results of this paper.

<Table 3> A snapshot of the KLIPS

PID	Job #	Job wave	Job start date			Job end date		
			year	month	date	year	month	date
11111	1	1	1988	3	-1	1998	3	-1
11111	2	1	1999	6	-1	2002	11	-1
11111	3	1	2003	1	15	·	·	·
11111	3	2	2003	1	15	·	·	·
11111	3	3	2003	1	15	·	·	·
…	…	…	…	…	…	…	…	…

Notes: -1 corresponds to "unknown" or "not responded", and · denotes missing values.

2. Summary statistics

Before applying the *K*-means clustering algorithm, I limit the sample to those who actively participated in the labor market from 1991 to 2010 by excluding individuals who have a non-employment ("unemployment") spell lasted more than 24 months. In addition, while employment and unemployment spells including January 1991 are kept, employment and unemployment spells that had completely ended before 1991 are dropped from the individual's employment history. Similarly, employment and unemployment spells including December 2010 are kept, but employment and unemployment spells that began after 2010 were excluded from the individual's employment history. Lastly, individuals whose employment history started after 1993 or whose employment history is no longer observed after 2009 are excluded from the sample.

The final individual-level job history sample is used to generate new variables that jointly summarize each individual's employment history (see Table 4). First, I divide the sum of all unemployment spells by the sum of all (un)employment spells to calculate U, the fraction of time spent in unemployment. Second, to understand the distribution of unemployment spells for each individual, I count the number of unemployment spells lasting 6 months or less, lasting 7 to 12 months, and lasting 13 to 24 months, and divide them by the number of (un)employment spells to obtain U_1, U_2, and U_3, respectively.[8] Similarly, to understand the distribution of employment spells for each individual, I count the number of employment spells lasting 24 months or less, lasting 25 to 60 months, lasting 61 to 120 months, and lasting more than 120 months, and divide them by the number of (un)employment spells to obtain E_1, E_2, E_3, and E_4, respectively.[9] Lastly, I divide the number of all employment spells by the total time spent in the labor market (converted in years) to calculate N, the average number of jobs per unit time.

<Table 4> The list of variables used for the *K*-means clustering algorithm

Notation	Meaning
U	Fraction of time spent in unemployment
U_1	# unemployment spells lasting 6 months or less*
U_2	# unemployment spells lasting 7-12 months*
U_3	# unemployment spells lasting 13-24 months*
E_1	# employment spells lasting 24 months or less*
E_2	# employment spells lasting 25-60 months*
E_3	# employment spells lasting 61-120 months*
E_4	# employment spells lasting more than 120 months*
N	# employment spells divided by the total time spent in the labor market**

Notes: * Divided by # (un)employment spells. ** Converted in years.

Table 5 provides summary statistics for 1,964 individuals and their 3,326 jobs from 1991 to 2010. In the final sample to be used for analysis, there are 1,439 males (73.3%) and 525 females (26.7%), with an average birth year of 1956. The time spent in unemployment is 3.84 months on average, with males spending an average of 4.23 months, which is 1.47 months longer than the 2.76 months spent by females. A gender difference is also observed in the time spent in

[8] One might argue that it would be more appropriate to divide by the number of unemployment spells as in Gregory *et al.*(2022), rather than the number of (un)employment spells, for the purpose of understanding the distribution of unemployment spells. Although this argument makes sense, I need to stick with the definition given in the text because the sample includes a significant number of individuals with no unemployment spells.

[9] Note that I divide by the number of (un)employment spells, rather than the number of employment spells, for consistency with the definitions of U_1, U_2, and U_3.

employment: females were in employment for an average of 36.24 years, which is 2.03 years longer than the 34.21 years experienced by males. Thus, one can conclude that males spent more time in unemployment and less time in employment, and this (combined with the fact that males held 0.29 more jobs on average) suggests the possibility that males encountered job losses or job changes more frequently during the period considered.

The summary statistics of the variables to be used in the analysis are also presented in Table 5, where the differences by gender are evident as above. Specifically, the relative proportion of unemployment spells lasting 6 months or less (U_1) is 1.81 times higher for males compared to females, while the relative proportion of employment spells exceeding 120 months (E_4) is 1.13 times higher for females compared to males. Hence, one might think that the clustering results could be explained to some extent by demographic variables such as sex and age, and I will revisit this issue when discussing the clustering results in Section 4.

<Table 5> Summary statistics

Variable	All		
		Male	Female
Birth year	1956	1957	1955
Sum of unemployment spells (m)	3.84	4.23	2.76
Sum of employment spells (y)	34.76	34.21	36.24
# jobs	1.69	1.77	1.48
U	0.011	0.012	0.008
U_1	0.099	0.112	0.062
U_2	0.021	0.024	0.014
U_3	0.024	0.026	0.018
E_1	0.020	0.022	0.015
E_2	0.027	0.029	0.020
E_3	0.040	0.042	0.033
E_4	0.770	0.745	0.839
N	0.053	0.056	0.046
# obs.	1,964	1,439	525

Ⅳ. Results

In this Section, I present the results of applying the K-means clustering algorithm to the Job History data of the KLIPS to classify the elderly based on their pattern of employment transitions. As previously discussed, it is necessary to determine the number of clusters before implementing the algorithm, so the process of choosing it through the cross-validation method is first discussed in Section 4.1. Then I present the clustering outcome, and examine the differences in the pattern of labor market transitions among the clusters (types) formed. Further, I investigate whether basic demographic variables, with region, industry, and occupation, can predict to which cluster an individual has been assigned.

Meanwhile, the differences between types can be considered as one aspect of heterogeneity that has not been focused on in previous studies, except that of Gregory *et al.* (2022). Thus, examining its potential role in understanding labor market issues would be interesting and meaningful. Motivated by this, I examine in Section 4.2 whether (and, if so, how much) the likelihood of Earned Income Tax Credit (EITC) receipt, the likelihood of homeownership, or the subjective economic well-being is affected by the type. In addition, I document the trend in employment rates[10] before and after the Korean financial crisis of 1997 differed by the type determined by the pattern of employment transitions, an observation that cannot be made with other forms of heterogeneity such as education.

1. Clustering outcome

The results of applying the cross-validation method are summarized in Table 6. As previously described in Section 2.2, for each value of $\widetilde{K} \in \{2,3,4,5\}$, I calculated the value of (3) using the mean of 200 repetitions of Steps 2.1 to 3.2 in Table 2 with different initial centroids. The results indicate that inconsistent classifications of the same observation are least likely to happen when $\widetilde{K}=2$, regardless of whether the entire sample or gender-based subsamples were processed.[11] This is an interesting finding in that Gregory *et al.*(2022), who analyzed the U.S. labor market, reported the existence of three types therein.

[10] Note that employment rates are defined in a slightly unusual way; see Section 4.2.

[11] In the case of females, however, it is difficult to completely rule out the possibility that $\widetilde{K}=3$ is optimal.

<Table 6> Results from the cross-validation

$\tilde{K}$	All		Male		Female	
	Mean	SD	Mean	SD	Mean	SD
2	0.03	0.15	0.05	0.21	0.05	0.20
3	0.10	0.28	0.19	0.32	0.06	0.21
4	0.08	0.22	0.10	0.24	0.16	0.27
5	0.16	0.21	0.37	0.28	0.25	0.28
# obs.	982		719		262	

Notes: For each $\tilde{K} \in \{2, 3, 4, 5\}$, the mean and standard deviation (SD) are calculated across 200 repetitions of Steps 2.1 to 3.2 in Table 2 with different initial centroids.

Table 7 presents the results of the *K*-means clustering algorithm with the number of clusters set to 2. Among the two clusters formed, the cluster with a larger number of observations is named "Type A," and the cluster with fewer observations is named "Type B." Then it is clear that Type A forms the majority while Type B forms minority: the individuals assigned to Types A and B account for 91.1% and 8.9%, respectively, of the total sample. The difference in the pattern of employment transitions by type is evident. Individuals in Type A, for instance, spent a shorter time in unemployment than those in Type B during the period considered (0.45% vs. 7.48%). Moreover, most of unemployment spells held by Type-A individuals lasted 6 months or less while most of their employment spells lasted more than 120 months, making them distinguished from Type-B individuals who exhibited higher relative frequencies of U_3 and E_1. To summarize, short job search duration, long job tenure, and infrequent transitions between unemployment and employment characterize individuals assigned to Type A, while long job search duration, short job tenure, and frequent transitions across employment status characterize Type B.[12] Accordingly, I conclude that two groups clearly differing in their patterns of employment transitions coexisted in the Korean labor market between 1991 and 2010. Note, however, that this is not a unique feature of the Korean labor market since the coexistence of heterogeneous groups within the labor market is commonly observed in other countries.[13]

[12] Thus, one can say that Types A and B in this study correspond to Types α and γ, respectively, in Gregory *et al.* (2022) who analyzed the U.S. labor market.

[13] See, e.g., Gregory *et al.* (2022) for the U.S., Spinella (2021) for Italy, and Darougheh and Lundgren (2022) for Denmark.

<Table 7> Clustering outcomes of the K-means clustering algorithm

Variable	All		Male		Female	
	Type A	Type B	Type A	Type B	Type A	Type B
U	0.45	7.48	0.49	7.40	0.28	7.44
U_1	8.49	23.79	9.90	22.86	4.75	24.61
U_2	1.60	7.67	1.84	7.82	0.92	6.98
U_3	1.42	12.53	1.45	13.60	0.95	11.85
E_1	0.75	14.77	0.87	14.12	0.43	14.87
E_2	1.63	13.24	1.83	12.74	1.10	13.38
E_3	3.18	11.97	3.42	11.48	2.41	14.05
E_4	82.93	16.03	80.71	17.39	89.44	14.26
N	4.14	17.66	4.36	16.97	3.53	18.38
Share, %	91.09	8.91	90.13	9.87	92.57	7.43

Notes: Reported values are the means of variables (multiplied by 100).

<Table 8> Probit coefficients for demographics on individual types

Variable	All			Male		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	−0.014	0.011	0.012	-	-	-
HS grad	0.012	0.002	0.013	0.036*	0.026	0.026
Some college	−0.043**	−0.052**	0.003	−0.021	−0.027	−0.004
College grad	−0.025	0.006	0.024*	−0.017	0.007	0.023
Master's or above	0.020	0.106*	0.045*	0.009	0.087	0.035
Birth year	0.006***	0.005***	−0.001	0.006***	0.005***	−0.000
U_1	-	-	−0.022	-	-	−0.059
U_2			0.104***			0.183***
U_3			0.140***			0.244***
E_1			0.017			0.029
E_2			0.023*			0.031*
E_3			0.010			0.003
N			2.643***			4.393***
Reg, Ind, Occ	N	Y	Y	N	Y	Y
# obs.	1,964	1,470	1,470	1,439	1,170	1,170
Pseudo R^2	0.064	0.331	0.913	0.061	0.301	0.943

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01.

As briefly mentioned in Section 3.2, one may argue that the clustering results can be predicted to some extent by demographic variables such as gender, education level, age, etc. To test this claim, I run a probit model where the dependent variable is a dummy equal to 1 if an individual belongs to Type B. The regression results are reported in Table 8, where it is observed that the coefficients for demographic variables are either statistically insignificant or negligible in magnitude (column (1)). Furthermore, the pseudo R-squared of the corresponding model is reported as 0.064, which indicates a poor fit to the data. Interestingly, when region, industry, and occupation are added to the model (column (2)), the pseudo R-squared increases to 0.331, meaning that an individual's type can be partially predicted by these factors.[14] The model fit, however, increases to 0.913 when the variables about labor market transitions (used for clustering) are additionally included (column (3)). This implies that "Type" determined by differences in employment transitions can be regarded as a new form of heterogeneity that is largely orthogonal to typically-observed variables.

2. Type as a form of heterogeneity

In order to investigate how the pattern of employment transitions acts as a form of heterogeneity, I first examine whether "Type" is a meaningful predictor of the economic situation in old age. To be specific, I run probit models where the dependent variables are (1) a dummy variable which takes a value of 1 if an individual is an EITC recipient in 2020, (2) a dummy variable which takes a value of 1 if an individual is not a homeowner in 2020, and (3) a dummy variable which takes a value of 1 if an individual is reported as being in the lowest level of subjective economic well-being (SEW) in 2020. The regression results presented in Table 9 indicate that individuals in group B have a 6.5% higher probability of receiving EITC benefits, a 8.8% higher probability of residing in rental housing, and a 7.1% higher probability of perceiving their economic situation as very bad, compared to their group A counterparts. In addition, while these estimates are statistically significant at the 5-10% level, the coefficient estimates for gender, education level, and age are either statistically insignificant or smaller than that of "Type." Thus, one can conclude that the type determined by the pattern of labor market transitions during prime ages may be more relevant in predicting one's economic conditions after retirement.

[14] In Gregory *et al.* (2022), the pseudo R-squared is reported as 0.034 when industry is included as an explanatory variable. Note that a significant number of public sector employees such as civil servants and teachers are included in the sample of this study, and I guess this might explain the difference.

<Table 9> Probit coefficients for demographics on economic situations in old age

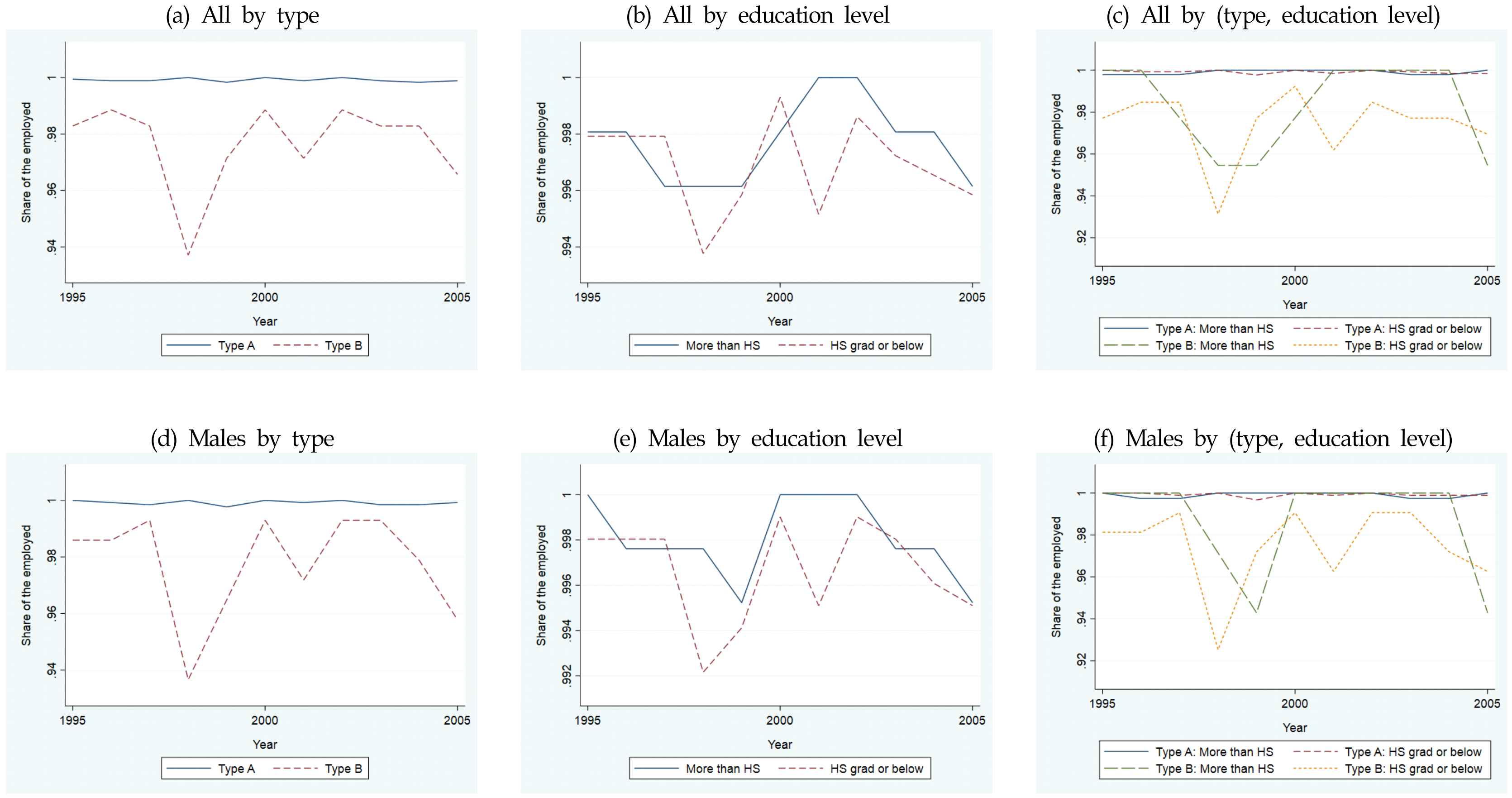
Variable	EITC recipient		Non homeownership		Lowest SEW	
	(1)	(2)	(1)	(2)	(1)	(2)
Female	0.014	0.012	0.075***	0.075***	0.013	0.013
HS grad or below	0.019	0.021*	0.016	0.013	0.014	0.010
Birth year	−0.001	−0.001	0.002**	0.002*	0.001	0.000
Type B		0.065*		0.088**		0.071**
Reg, Ind, Occ	Y	Y	Y	Y	Y	Y
# obs.	1,194	1,194	1,918	1,918	1,828	1,828
Pseudo R^2	0.156	0.175	0.121	0.125	0.098	0.109

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01.

Meanwhile, Figure 2.(a) displays the trend in employment rates[15] from 1995 to 2005 by type (determined by clustering analysis). In the figure, the solid line represents the employment rate trend for Type A, and the dashed line corresponds to the trend for Type B. Then it is clear that the employment rate for Type A remained consistently close to 100% throughout the analysis period, while the employment rate for Type B not only remained lower than Type A but also exhibited significant fluctuations throughout the same period. This suggests some evidence that the overall fluctuations in the (un)employment rate may be largely driven by Type B. On the other hand, when the entire sample is divided into two groups by education (high school graduate or below vs. more than high school), one cannot observe a clear distinction between them in terms of employment rates (see Figure 2.(b)). This implies that the importance of individual-level heterogeneity in understanding (un)employment fluctuations could be underestimated if only demographic variables such as education are considered as a form of heterogeneity.

[15] In Figure 2, if an individual had one or more jobs during a given year, then s/he is considered as “employed” for that year, regardless of the timing or duration of that employment. Therefore, it should be noted that the rates could be higher than typically reported.

[Figure 2] Employment rates by type or education level (1995-2005)



(a) All by type

(b) All by education level

(c) All by (type, education level)

(d) Males by type

(e) Males by education level

(f) Males by (type, education level)

Notes: If an individual had one or more jobs during a given year, then s/he is regarded as "employed" for that year, regardless of the timing or duration of that employment relationship.

Ⅴ. Concluding Remarks

I have documented that older people can be classified into two groups according to their pattern of employment transitions before retirement. Those classified into group A constitute the majority (91.1%), and their transition pattern in the labor market is characterized by short-term unemployment and long-term employment. Those assigned to group B, in contrast, form the minority (8.9%), and they tend to have experienced long-term unemployment and short-term employment. It has been also shown that whether an individual belongs to one group or the other cannot be fully explained or predicted by basic demographics, region, industry, and occupation. Furthermore, I have provided some empirical evidence that the pattern of employment transitions is a form of heterogeneity that needs to be considered (and further explored) at both micro and macro levels.

The findings of this paper suggest that, at the individual level, the frequency and speed of transitions between employment and unemployment should be actively monitored by the authorities. In addition, they indicate the need to selectively provide support for job search and incentive for longer job tenure to those who belong to group B, thereby improving their economic conditions in old age. However, the average birth year of the individuals focused on in this study is 1956, meaning that further research is required to determine whether the policy implications of this study can be extended to other generations who might exhibit different characteristics and patterns. Moreover, whether the individual type related to the pattern of employment transitions could change over time, and what determines differences in employment transitions that are not sufficiently explained by generally-observed variables, also need to be answered in future work.

References

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A distributional framework for matched employer employee data. *Econometrica 87*(3), 699 - 739.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625 - 643.

Darougheh, S. and G. Lundgren (2022). Worker protection and screening. Manuscript.

Gregory, V., G. Menzio, and D. Wiczer (2022). The alpha beta gamma of the labor market. *Federal Reserve Bank of St. Louis Working Paper 2021-003*.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction (Second Edition)*. Springer.

Jang, I., J. Lee, S. Shin, H. Jeong, and I. Kwon (2023). *KLIPS Waves 1-24 User's Guide*. Korea Labor Institute.

Min, H. and S.-G. Lee (2018). The patterns of the labor force participation among the middle old aged in South Korea: The application of group-based trajectory model. *Ewha Journal of Social Sciences 34*(2), 169 - 194.

Son, Y. J. (2022). A study on the types of work history for those who have experience in self-employment. *Social Welfare Policy 49*(1), 37 - 60.

Spinella, S. (2021). Discretizing workers' heterogeneity: The alpha beta gamma of the Italian labor market. Università Bocconi.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika 97*(4), 893 - 904.