Characterization of the Chilean Pension Fund Knowledge using Knowledge Discovery in Data Techniques

Abstract

This study investigates the existence of clusters and the main determinants of knowledge in the Chilean pension system. Financial literate people make better financial decisions. We use a unique database that is a panel of affiliates for the years 2006 and 2009 (The Social Protection Survey). The results show that there are two cluster of knowledge. High level of pension knowledge is associated with educated people, higher financial literacy and higher wages.

Keywords: Financial Literacy, Pension Knowledge, Pension Funds.

Characterization of the Chilean Pension Fund Knowledge using Knowledge Discovery in Data Techniques

1. Introduction

Chile had a pay as you go system before 1980. The system was collapsing because of differences in requirements for demanding a pension and the increase in the national budget for funding the pension benefits. In general, this kind of social security system reduces incentives to save, to invest in financial literacy, and to invest in risky assets (Jappelli and Padula, 2015). The pension system was replaced to a define contribution system managed by Pension Fund Administrators (PFA) in charge of collecting, investing and paying pension benefits in 1981. This new pension system transfers to affiliates some financial decisions that affect their pension accumulation wealth. This new system requires people able to decide if they need to save more, to invest their assets in a multifund scheme system, to decide between phase withdrawal and purchase an annuity for the decumulation phase, and to compare returns and costs among PFA. Therefore, insufficient savings during their working lives or bad financial decision-making in an increasing complex financial world can negatively affect the final pension at the old age. A lack of financial knowledge has been one of the main reasons to explain inadequate financial decisions. One of the consequences is the evidence that people fail to plan for retirement (Lusardi and Mitchell, 2006).

Several papers report a positive correlation between financial literacy and making good financial decisions according to a literature review showed in Lusardi and Mitchell (2014). They consider financial knowledge as a form of investment in human capital, and as such generates important implications for welfare. Some papers show that financial literacy increase stock market participation (Van Rooij et al., 2011) and increase total net wealth (Behrman et al., 2012). In addition, High-literacy investors are better at timing the market (Guiso and Viviano, 2015) and into the Chilean context could be very useful for the fund choice. Therefore, finding the determinants of the pension knowledge is important for making policy that could contribute to the social welfare. The main determinants of having a higher financial literacy are education (Agarwal et al., 2015), being married (Agarwal et al., 2015) and being a male (Lusardi and Mitchell, 2008; Agarwal et al., 2015; Almenberg and Dreber, 2015). Behrman et al. (2012) analyse the financial literacy in Chile using the Social Protection Survey in 2006 (EPS for its spanish Acronym). This data source is comparable to the Health Retirement Survey (HRS) for United States. They measure financial literacy using 12 questions and the PRIDIT methodology. They find that financial literacy have positive effects on wealth accumulation.

The main objective of this paper is to describe the characteristics of the Chilean population in terms of their level of understanding of the main features of the pension fund system that could have important effects in the pension wealth. The specific objectives of this paper is to identify groups of individuals who share similar levels of knowledge and to describe the characteristics of such individuals in terms of their level of financial and mathematical knowledge, general education, and other socio-demographic variables. We investigate the evolution of such characteristics between a panel of people followed by the EPS for the year 2006 and 2009 using a novel methodology, i.e. knowledge discovery in databases (Piatesky and Frawley, 1991).

The contribution of this paper is to show the existence of two cluster of knowledge in Chile. There is a group that have high level of knowledge and the other has a very low level of knowledge. In addition, to have a panel of information for affiliates for the years 2006 and 2009 allow us to analyze movements in groups of knowledge after the financial crisis.

The work was organized in two main stages: i) exploration of data patterns using data mining techniques and ii) confirmatory analysis through traditional statistical and econometrical tools. Stage one was implemented with the guidelines of the CRISP Data Mining reference model (Chapman et al., 2000; Shearer, 2000), and it was adapted to consider the application of clustering algorithms (Jain, Murty, & Flynn, 1999) and decision trees (Kohavi & Quinlan, 2002)(Quinlan, 1990) in sequential and complementary sub-stages.

The remainder of the article is organized as follows. Section 2 shows the Methodology. Section 3 gives the results of the estimates. Finally, Section 4 offers our conclusions and suggestions for future lines of research.

2. Methodology

The first objective in stage one is the identification of groups of survey-respondents which answered questions from "Module E" in the EPS datasets in a similar way. This task was implemented using clustering techniques (Jain et al., 1999). Questions from "Module E" measure knowledge and awareness of several characteristics of the Chilean retirement and pension fund system, and thus, survey-respondents which answered following a similar pattern can be considered part of a group with a certain homogeneous level of knowledge and awareness. The probability of membership to one of these spontaneously available groups was later use to create a dependent variable. The latter takes the form of a high score when the subject belongs to a group with high levels of knowledge, and low score when the opposite is true. A similar strategy has been used in previous studies with the important difference that the dependent variable has been constructed using some kind of linear combination of the survey answers. In these previous work it was up to the researchers to define how this score was calculated and thus, these strategies could be considered subjective.

The second objective of stage one is to explore, discover and describe non-linear patterns linking the level of knowledge and awareness of the Chilean pension fund system with the level of financial and mathematical knowledge of survey respondents. In order to do so, previously discover membership scores were used as a dependent variable, and several questions from "Module K: Financial literacy and non-cognitive skills" were used as explanatory variables. Information coming from "Module A: General information about survey-respondent" was also used as further control and descriptive variables. This task was implemented using decision trees algorithms (Kohavi & Quinlan, 2002)(Quinlan, 1990), which represent non-linear patterns available in the datasets as a collection of "if-then-rules" that are easy to understand by humans. Thus, they can provide rich insights with respect to the characteristics of the relationships under study.

In order to validate the patterns discovered in the previous stages, several steps of validation and cross-validation were undertaken as recommended by the CRISP-DM methodology (Chapman et al., 2000;

Shearer, 2000). In particular, for the first sub-stage of stage one, two different types of competing clustering algorithms were used, mainly: K-Means Algorithm, wherein the researcher chooses the *k*-number of clusters to be found, and X-Means, also known as two-step Algorithm, in which the machine automatically selects the number of groups spontaneously found in the datasets. Cluster quality was evaluated using the "Silhouette" technique (Kaufmann & Rousseeuw, 1990), and thus, alternative clustering results were ranked according to such measure. Once the best cluster algorithm and algorithm configuration was selected, groups that were discovered in the dataset were compared against each other using further statistical analysis. In particular, one-way ANOVA for independent samples was used to compare the means or "centroids" of each cluster, to reassure with statistical certainty that groups were indeed different to each other.

In relation to sub-stage two of stage one, an *unbalanced classification problem* was configured given the fact that groups in the data sets can have very different sizes (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The training datasets were artificially balanced following an oversampling approach. This with the aim of securing an equal representation of available groups, and thus, facilitate the discrimination between classes of survey-respondents. Four competing Decision Trees algorithms (C5.0, CRTree, CHAID, QUEST) were tested. Their results were evaluated in terms of class discrimination accuracy: precision and recall measures; easiness of interpretability; and the richness of information they could provide to domain experts.

The dataset was correspondingly divide into two sub-sets: one containing 50% of the observations labelled the *training sample* was used to create different decision trees with the competing algorithms, and the remaining 50% of the observations, labelled *testing sample*, was used to cross-validate the models that performed best in the training sample. Data from both the 2006 and 2009 EPS survey was analysed separately, and moreover, to further check the validity of the patterns, the decision tree obtained in the 2006 sample was used to forecast the cluster membership in 2009. The forecasted membership was later compared with the realized membership as indicated by the clustering analysis performed also in the 2009 sample. The agreement between these two metrics is also reported.

The proposed methodology possess certain salient features an advantages. Firstly, the discover of naturally occurring groups of survey respondents permits a better description of the characteristics of the population, as opposed to relying only on a strategy that creates a subjective score that follows the researcher intuition. Nonetheless, it is important to notice that in fact both strategies are complementary, and instead of abandoning the researcher intuition, this can be complemented and confirmed by the results of the clustering techniques. Secondly, to the best of our knowledge, this is the first attempt to use decision trees as tool to present and discover non-linear patterns linking the level of knowledge and awareness of the Chilean pension fund system with the level of financial and mathematical knowledge of survey respondents.

The decision trees algorithms also present some very useful characteristics: They include a feature-selection process that automatically and efficiently searches a potentially very big space of candidate explanatory variables, leaving only the ones which better describe the differences between the available groups. Furthermore, decision trees can be considered a non-parametric modelling technique that do not need or make any assumptions regarding the underlying properties of the distribution of the variables. Thus efficiently works with both continuous and categorical explanatory variables, as well as, modelling datasets with missing, extreme and outlier observations. Decision trees are usually consider an "open-

box" classifier data mining technique, as it represent the discovered patterns in a flow-chart-like tree structure which is easy to interpret by domain experts, helping to identify key explanatory variables, and main linear patterns in the datasets, but also, to uncover hidden non-linear patterns that may apply to certain cases that could be very difficult to find with traditional statistical and econometrical tools.

Finally, in stage two several logistic and ordinal regression models were used to confirm the discoveries arising from stage one. Robustness checks and post-hoc tests are reported, and final results and implications are discussed in subsequent sections of this paper. The following section will discuss further details of the implementation of the methodology, as well as, data description and results.

3. Results

Data understanding

This step started with the collection of data from the Social Protection Survey (Encuesta de Protección Social in Spanish, EPS) for years 2006 and 2009. Data dictionaries and metadata description was available via the EPS survey manual. This dataset covers information coming mainly from 3 modules in the "Subjects database" ("Base Entrevistado"): Module A "General Information about Subject"; Module E "Pension Funds Knowledge"; and Module K "Financial literacy and Non-cognitive Skills". Tables 1 and 2 present detail description of the variables in use.

Table 3 presents the summary statistics of the variables in use. 56% of the sample are male and the average age increase from 41 to 44 between surveys. The wage does not change between 2006 to 2009 survey, however the standard deviation increases. There is a little decrease in the level of the pension knowledge from 5.29 to 5.04 (over 12 questions in total). Also, the financial literacy decreases from 3.51 to 3.45 (over 7 questions in total).

Data preparation

The second step relates to the preparation of the data, in which the variables are transformed and modified from the initial raw data for use with the modelling step. The data preparation tasks are normally performed at different times during the analysis. This is an iterative process that helps with the improvement of the analysis (modelling phase) and can involve adding or removing variables and modifications or transformation of the original variables in the dataset.

For the purposes of the application of the proposed methodology, a number of data preparation steps have been carried out. In particular, for the clustering analysis fourteen "yes or no" questions from "Module E" were selected as inputs. The list and wording of such variables and questions can be found in Tables 1 and 2. For the decision tree analysis, a new independent variable "nota" (score") was created which summarized the "financial and non-cognitive skills" level of the subject coming from "Module K". This score takes a value from min=1 to max=7. Also, another variable reflecting the educational level of the subjects was created. This variable "ESC_BIN" categorized the Education variable available in the original dataset in three values ranging from 1=low to 3=high level of education considering the number of years of education of each subject. Other step of data preparation considered the selection of subjects which are part of the pension fund system. This excluded subjects which were members of the old pension

fund systems, and other types of pension fund systems, such as, the army, air forces, police forces and military men.

Modelling and Evaluation: Clustering analysis

Both k-means and x-means (two-step) algorithms identified the presence of two groups or clusters (see graphs 1 and 2), one with "high" pension fund knowledge and one with "low", in both 2006 and 2009 datasets. Cluster quality was deemed "fair" by the Silhouette measurement (S=0.4) (Kaufmann & Rousseeuw, 1990). The difference in means (centroids) for clusters is statistically significant for all variables in analysis at p-value=0.000. The most important determinants of the pension knowledge are answering rightly the questions: number of multifunds, the riskier fund, own pension fund investments, the fund with higher return in the last 10 years and voluntary saving accounts (see tables 4 and 5). (One-way ANOVA independent sample t-tests. Smaller F-test was 274.802 with df =1, 10594).

The confusion matrix reported in tables 6 and 7 shows that there is a good discrimination between belonging a low pension knowledge and a high pension knowledge. The results are more effective for low pension knowledge in both years of analysis.

Confirmatory Analysis: A test of Robustness

In order to check the results, two confirmatory analysis are developed in this paper. The first analysis are logit regressions using the EPS for year 2006 and 2009. The results show that the probability to have a low level of pension knowledge decreases as: the schooling years of the affiliate increases, the labor income increases and the nota (variable that captures math and financial knowledge) increases. The second analysis is a multi probit regression for changing the pension knowledge. The results show that having children keep the level of pension knowledge, decreases in wages reduce the pension knowledge. Change in education level increases the pension knowledge but also reduces. The last result is ambiguous and we interpret that not all new education level increases the pension knowledge.

4. Conclusions

This paper shows the existence of two cluster of pension knowledge in Chile. There is a group that have high level of knowledge and the other has a very low level of knowledge. People who know more about the multifund schemes (number of funds, riskier fund, returns) and about the voluntary savings accounts. The later results show the importance to develop financial policies to increase investment education in the population. Finally, the analysis of the panel of information for affiliates for the years 2006 and 2009 show that there are movements in groups of knowledge after the financial crisis mainly for those who increase education and for those who reduce wages. Future research should be oriented to decompose the education and explore which kind of education have a larger impact on the pension knowledge.

References

Agarwal, S., Amromin, G., Ben-David, I., & Chomsisengphet, S. (2015). Financial literacy and financial planning: Evidence from India. Journal of Housing Economics, 27, 4-21.

Almenberg, J., Dreber, A. (2015). Gender, stock market participation and financial literacy. Economics Letters, http://dx.doi.org/10.1016/j.econlet.2015.10.009

Behrman, J., Mitchell, O. S., Soo, C., Bravo, D. (2012). The effects of financial education and financial literacy. How financial literacy affects household wealth accumulation. American Economic Review, 102(3), 300-304.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Retrieved from http://www.crisp-dm.org/CRISPWP-0800.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Guiso, L. & Viviano, E. (2015). How much can financial literacy help?. Review of Finance, 19, 1347-1382.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*. doi:10.1145/331499.331504

Jappelli, T. & Padula, M. (2015). Investment in financial literacy, social security, and portfolio choice. Journal of Pension Economics and Finance. Doi: 10.1017/s1474747214000377.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis [Hardcover]*. New York: Wiley-Interscience; 99 edition. Retrieved from http://www.amazon.com/Finding-Groups-Data-Introduction-Analysis/dp/0471878766

Kohavi, R., & Quinlan, J. R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery* (pp. 267–276).

Lusardi, A. & Mitchell, O. S. (2006). Financial literacy and planning: Implications for retirement well-being. Pension Research Council, Working Paper WP 2006-01.

Lusardi, A. & Mitchell, O. S. (2008). Planning and financial literacy: How do women fare?. American Economic Review, 98(2), 413-417.

Lusardi, A. & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. Journal of Economic Literature, 52(1), 5-44.

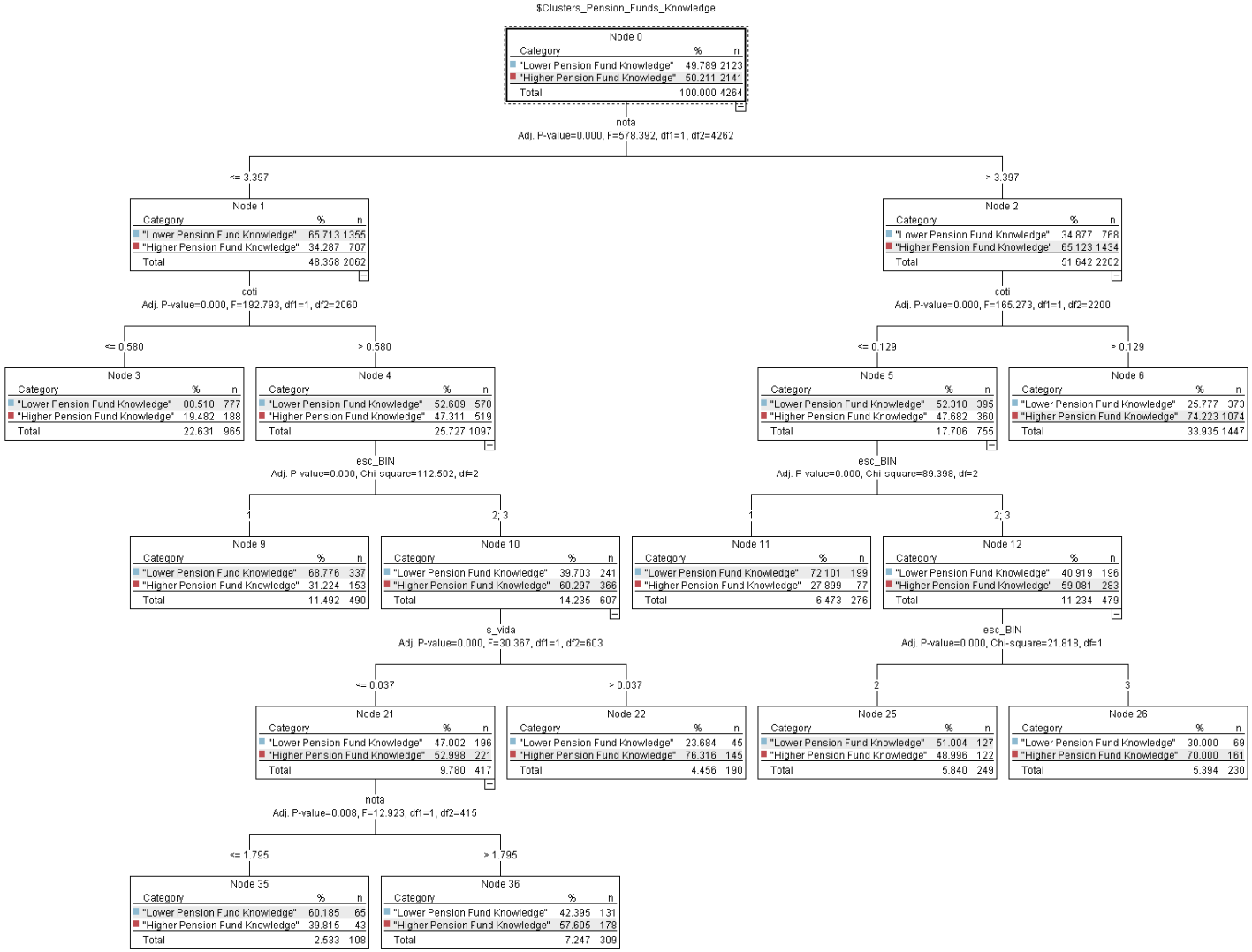
Piateski, G. & Frawley, W. (1991). Knowledge Discovery in Databases. MIT Press, Cambridge, MA, USA.

Quinlan, J. R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems, Man and Cybernetics*, *20*(2), 339–346. doi:10.1109/21.52545

Shearer, C. (2000). The CRISP-DM Model: The new blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Van Rooij, M., Lusardi, A. & Rob, A. (2011). Financial literacy and stock market participation. Journal of Financial Economics, 101, 449-472.

Graph 1: Decision Tree Analysis. **Exploratory Analysis: Cluster Description Results EPS 2006- Best Tree C&RTree**



Graph 2: Decision Tree Analysis. **Exploratory Analysis: Cluster Description Results EPS 2009- Best Tree C&RTree**

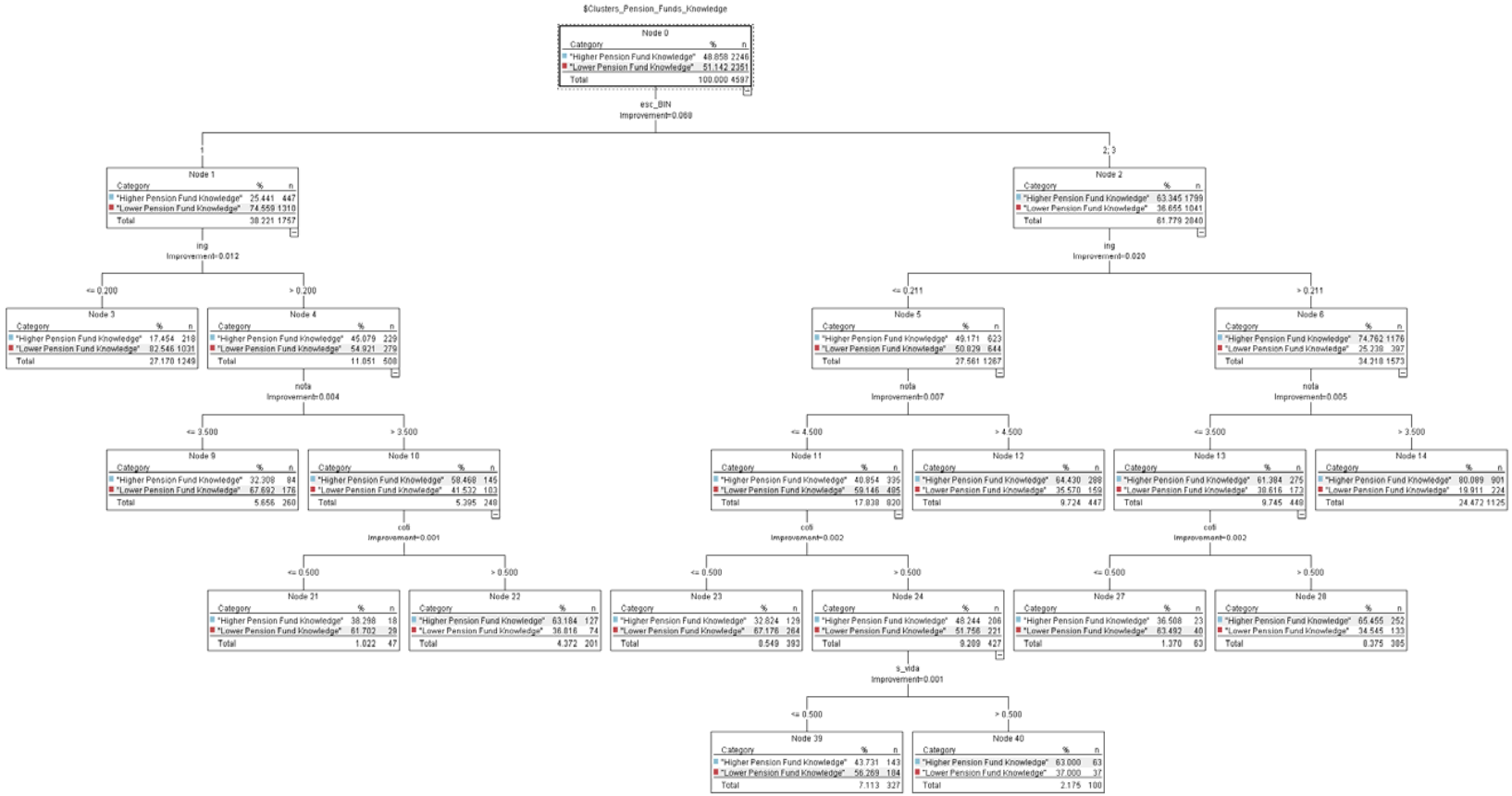


Table 1: List of questions included in the Clustering Analysis.

Label Variable	Question
ejer1	Answer correctly the men retirement age (yes=1)
ejer2	Answer correctly the women retirement age (yes=1)
ejer3	Answer correctly how the pension is determined (yes=1)
Ejer4	Knowledge about the % rate of contribution (yes=1)
Ejer5	Knowledge about the balance in his/her account (yes=1)
Ejer6	Knowledge about multifunds (yes=1)
Ejer7	Knowledge about number of multifunds (yes=1)
Ejer8	Knowledge about the riskier multifund (yes=1)
Ejer9	Knowledge about the fund higher in return for 10 years (yes=1)
Ejer10	Knowledge about Voluntary Saving Account (yes=1)
Ejer11	Knowledge about their own pension fund investment (yes=1)
Ejer12	Knowledge about pension modalities (yes=1)

Table 2: Construction of variable "nota"

Nota: Variable that capture math and financial skills (lower scale 1 to higher scale 7).

auxnota =1 if interviewee do not answer at least 1 question in math.

Variable	Question
Pre1 (=1 if answer is right)	If the probability is 10%, how many people get sick in a population of 1000?
Pre2 (=1 if answer is right)	5 people share a prize of $2 million. How much receive each one?
Pre3 (=1 id answer is right)	AFP A returns 15% and AFP B returns 20% last year. Which one get a higher return?
Pre4	You have $200 in a saving account and the annual return is 10%. How you get in 2 years?
Pre5	You have $100 and the annual return is 2%. How you get in 5 years?
Pre6	You have $100 in a saving account, the annual return is 1% and the inflation rate is 2%. After a year, you can purchase more or less?

Table 3: Summary Statistics.

PERIOD	2006			2009		
STATISTICS	Mediana	Mean	SD*	Mediana	Mean	SD*
Gender [Male = 1]	1.0	0.56	0.50	1.0	0.56	0.50
Age [Year]	40.0	41.35	11.48	43.0	43.74	11.48
Education [Year]	12.0	10.62	3.55	12.0	11.09	3.67
Marital status [Partner present = 1]	1.0	0.51	0.50	1.0	0.64	0.48
Children [Yes = 1]	1.0	0.72	0.45	1.0	0.78	0.41
Wage [M$]	0.2	0.33	0.57	0.2	0.33	1.00
Wealth [MM$]	7.2	14.69	52.95	6.9	14.23	37.82
Knowledge pension system [1 to 12]	5.0	5.29	2.80	5.0	5.04	2.97
Financial literacy [1 to 7]	3.0	3.51	1.75	4.0	3.45	1.65
(Ejer4) Meet discount [Yes = 1]	0.0	0.46	0.50	0.0	0.35	0.48
(Ejer5) Meet accumulated [Yes = 1]	1.0	0.52	0.50	0.0	0.47	0.50
(Ejer10) Meet pensions	1.0	0.64	0.48	0.0	0.48	0.50
(Ejer11) Meet pensions funds [Yes = 1]	0.0	0.32	0.47	0.0	0.39	0.49
Housing savings [Yes = 1]	0.0	0.12	0.33	0.0	0.14	0.35
Pension savings [Yes = 1]	0.0	0.12	0.33	0.0	0.14	0.35
Bank savings [Yes = 1]	0.0	0.12	0.33	0.0	0.14	0.35
Pay contributions pensions system [Yes = 1]	1.0	0.64	0.48	1.0	0.61	0.49
Health [Poor = 1 to Excellent = 6]	4.0	3.95	0.90	4.0	3.87	0.87

Source: Original calculations based on the EPS 2006 and 2009. Observations in both periods: 5.778

Table 4: Clustering for year 2006

Field	statistic	cluster-1 (High Knowledge)	cluster-2 (Low Knowledge)	F-Test	df	p-value
conpen	mean	8,104	2,809	26657,575	1, 9792	0,00%
	std. Dev	1,606	1,508			significant
	std. Error	0,027	0,019			
	count	3567	6227			
ejer7	mean	0,817	0,053	15410,889	1, 9792	0,00%
	std. Dev	0,387	0,223			significant
	std. Error	0,006	0,003			
	count	3567	6227			
ejer11	mean	0,831	0,068	14399,307	1, 9792	0,00%
	std. Dev	0,375	0,252			significant
	std. Error	0,006	0,003			
	count	3567	6227			
ejer8	mean	0,883	0,116	12934,48	1, 9792	0,00%
	std. Dev	0,322	0,321			significant
	std. Error	0,005	0,004			
	count	3567	6227			
ejer6	mean	0,857	0,163	8288,964	1, 9792	100,00%
	std. Dev	0,35	0,37			Important
	std. Error	0,006	0,005			
	count	3567	6227			
ejer9	mean	0,609	0,063	5412,406	1, 9792	100,00%
	std. Dev	0,488	0,244			Important
	std. Error	0,008	0,003			
	count	3567	6227			
ejer10	mean	0,741	0,27	2576,856	1, 9792	0,00%
	std. Dev	0,438	0,444			significant
	std. Error	0,007	0,006			
	count	3567	6227			
ejer4	mean	0,569	0,176	1924,431	1, 9792	0,00%
	std. Dev	0,495	0,381			significant
	std. Error	0,008	0,005			
	count	3567	6227			
ejer5	mean	0,652	0,303	1269,981	1, 9792	0,00%
	std. Dev	0,476	0,46			significant
	std. Error	0,008	0,006			
	count	3567	6227			
ejer3	mean	0,228	0,072	520,96	1, 9792	0,00%
	std. Dev	0,42	0,258			significant
	std. Error	0,007	0,003			
	count	3567	6227			
ejer2	mean	0,854	0,672	403,414	1, 9792	0,00%
	std. Dev	0,353	0,47			significant
	std. Error	0,006	0,006			
	count	3567	6227			
ejer12	mean	0,126	0,028	384,692	1, 9792	0,00%
	std. Dev	0,332	0,164			significant
	std. Error	0,006	0,002			
	count	3567	6227			
ejer1	mean	0,938	0,824	259,884	1, 9792	0,00%
	std. Dev	0,241	0,38			significant
	std. Error	0,004	0,005			
	count	3567	6227			

Table 5: Clustering for year 2009

Field	statistic	cluster-1*	cluster-2*	F-Test	df	p-value
conpen	mean	2,756	8,015	26677,661	1, 9792	0,00%
	std. Dev	1,472	1,656			significant
	std. Error	0,019	0,027			
	count	6104	3690			
ejer7	mean	0,046	0,802	15015,955	1, 9792	0,00%
	std. Dev	0,21	0,399			significant
	std. Error	0,003	0,007			
	count	6104	3690			
ejer8	mean	0,1	0,884	14867,323	1, 9792	0,00%
	std. Dev	0,3	0,321			significant
	std. Error	0,004	0,005			
	count	6104	3690			
ejer11	mean	0,062	0,815	14052,169	1, 9792	0,00%
	std. Dev	0,241	0,388			significant
	std. Error	0,003	0,006			
	count	6104	3690			
ejer6	mean	0,158	0,842	8076,06	1, 9792	0,00%
	std. Dev	0,365	0,365			significant
	std. Error	0,005	0,006			
	count	6104	3690			
ejer9	mean	0,055	0,605	5705,459	1, 9792	0,00%
	std. Dev	0,227	0,489			significant
	std. Error	0,003	0,008			
	count	6104	3690			
ejer10	mean	0,267	0,73	2500,439	1, 9792	0,00%
	std. Dev	0,443	0,444			significant
	std. Error	0,006	0,007			
	count	6104	3690			
ejer4	mean	0,176	0,555	1801,763	1, 9792	0,00%
	std. Dev	0,381	0,497			significant
	std. Error	0,005	0,008			
	count	6104	3690			
ejer5	mean	0,303	0,641	1204,665	1, 9792	0,00%
	std. Dev	0,459	0,48			significant
	std. Error	0,006	0,008			
	count	6104	3690			
ejer3	mean	0,07	0,227	529,141	1, 9792	0,00%
	std. Dev	0,255	0,419			significant
	std. Error	0,003	0,007			
	count	6104	3690			
ejer2	mean	0,67	0,851	410,18	1, 9792	0,00%
	std. Dev	0,47	0,356			significant
	std. Error	0,006	0,006			
	count	6104	3690			
ejer12	mean	0,027	0,124	383,358	1, 9792	0,00%
	std. Dev	0,161	0,33			significant
	std. Error	0,002	0,005			
	count	6104	3690			
ejer1	mean	0,822	0,938	275,907	1, 9792	0,00%
	std. Dev	0,382	0,24			significant
	std. Error	0,005	0,004			
	count	6104	3690			

Table 6: Confusion Matrix for year 2006

'Partition' = 1_Training	Higher Pension Fund Knowledge (% in total)	Lower Pension Fund Knowledge (% in total)
Higher Pension Fund Knowledge	2484 (67.7%)	563 (23.0%)
Lower Pension Fund Knowledge	1186 (22.3%)	1886 (77.0%)
Total	**3670**	**2449**

'Partition' = 2_Testing	Higher Pension Fund Knowledge	Lower Pension Fund Knowledge
Higher Pension Fund Knowledge	2513 (67.5%)	560 (22.3%)
Lower Pension Fund Knowledge	1208 (32.5%)	1947 (77.7%)
Total	**3721**	**2507**

Table 7: Confusion Matrix for year 2009.

'Partition' = 1_Training	Higher Pension Fund Knowledge (% in total)	Lower Pension Fund Knowledge (% in total)
Higher Pension Fund Knowledge	2263 (73.8%)	786 (26.5%)
Lower Pension Fund Knowledge	803 (26.2%)	2175 (73.5%)
Total	**3266**	**2961**
'Partition' = 2_Testing	Higher Pension Fund Knowledge	Lower Pension Fund Knowledge
Higher Pension Fund Knowledge	2291 (72.1%)	736 (24.7%)
Lower Pension Fund Knowledge	885 (28.9%)	2241 (75.3%)
Total	**3176**	**2977**

Confirmatory Analysis: Logit Regression for year 2006

Case Processing Summary

		N	Marginal Percentage
$Clusters_Pension_Funds_Knowledge	"Higher Pension Fund Knowledge"	3861	36.90%
	"Lower Pension Fund Knowledge"	6603	63.10%
s_vida	0	7932	75.80%
	1	2532	24.20%
coti	0	4062	38.80%
	1	6402	61.20%
Valid		10464	100.00%
Missing		132	
Total		10596	
Subpopulation		7858(a)	

a. The dependent variable has only one value observed in 7444 (94.7%) subpopulations.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	12105.122			
Final	9283.127	2821.995	5	0

Pseudo R-Square

Cox and Snell	0.236
Nagelkerke	0.323
McFadden	0.205

Parameter Estimates

$Clusters_Pension_Funds_Knowledge(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
"Lower Pension Fund Knowledge"	Intercept	3.289	0.111	875.37	1	0			
	esc	-0.196	0.008	610.27	1	0	0.822	0.809	0.835
	ing	-0.485	0.068	51.121	1	0	0.616	0.539	0.703
	nota	-0.259	0.014	329.26	1	0	0.772	0.75	0.794
	[s_vida=.000]	0.374	0.054	48.668	1	0	1.454	1.309	1.615
	[s_vida=1.000]	0(b)	.	.	0	.	.	.	.
	[coti=.000]	0.861	0.052	276.13	1	0	2.366	2.137	2.619
	[coti=1.000]	0(b)	.	.	0	.	.	.	.

a. The reference category is: "Higher Pension Fund Knowledge".

b. This parameter is set to zero because it is redundant.

The variables are: esc (schooling years), ing (labor income), nota (variable that captures math and financial skills, see table 2), svida (to have life insurance) and coti (to make contributions into the system).

Confirmatory Analysis: Logit Regression for year 2009

Case Processing Summary

		N	Marginal Percentage
$Clusters_Pension_Funds_Knowledge	"Higher Pension Fund Knowledge"	3640	37.80%
	"Lower Pension Fund Knowledge"	5993	62.20%
s_vida	0	7788	80.80%
	1	1845	19.20%
coti	0	4203	43.60%
	1	5430	56.40%
Valid		9633	100.00%
Missing		161	
Total		9794	
Subpopulation		472(a)	

a. The dependent variable has only one value observed in 162 (34.3%) subpopulations.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	4411.668			
Final	1403.692	3007.977	4	0

Pseudo R-Square

Cox and Snell	0.268
Nagelkerke	0.365
McFadden	0.235

Parameter Estimates

$Clusters_Pension_Funds_Knowledge(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval	
								Lower Bound	Upper Bound
"Lower Pension Fund Knowledge"	Intercept	3.204	0.122	685.19	1	0			
	esc	-0.229	0.008	723.22	1	0	0.796	0.783	0.809
	nota	-0.302	0.016	349.05	1	0	0.74	0.717	0.763
	[s_vida=.000]	0.697	0.062	127.56	1	0	2.007	1.778	2.265
	[s_vida=1.000]	0(b)	.	.	0	.	.	.	.
	[coti=.000]	0.976	0.053	345.21	1	0	2.655	2.395	2.943
	[coti=1.000]	0(b)	.	.	0	.	.	.	.

a. The reference category is: "Higher Pension Fund Knowledge".

b. This parameter is set to zero because it is redundant.

The variables are: esc (schooling years), ing (labor income), nota (variable that captures math and financial skills, see table 2), svida (to have life insurance) and coti (to make contributions into the system).

Confirmatory Analysis: Multi Probit

Marginal effects in changing knowledge on pensions (5700 data)

	Worse (-1)	Same (0)	Progress (1)
Gender [Male = 1]	-0.002	-0.020	0.022
	(0.014)	(0.022)	(0.018)
Age [Years]	0.000	0.000	-0.000
	(0.001)	(0.001)	(0.001)
Education [Years]	0.005***	-0.016***	0.010***
	(0.002)	(0.003)	(0.003)
Health (2006)	0.007	-0.011	0.003
	(0.009)	(0.013)	(0.011)
Health (+)	-0.013	-0.012	0.025
	(0.016)	(0.028)	(0.026)
Health (-)	-0.011	0.001	0.009
	(0.016)	(0.026)	(0.022)
Children [Yes = 1]	-0.009	0.014	-0.005
	(0.018)	(0.027)	(0.021)
Children (+)	-0.032	0.067**	-0.035
	(0.022)	(0.034)	(0.026)
Wage [MM$/2006]	-0.003	0.033*	-0.030*
	(0.010)	(0.017)	(0.018)
Wage (+) [Yes = 1]	0.003	-0.026	0.023
	(0.015)	(0.027)	(0.025)
Wage (-) [yes = 1]	0.084***	-0.051**	-0.033
	(0.019)	(0.025)	(0.021)
Wealth [MM$/2006]	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
Wealth (+) [Yes = 1]	-0.011	-0.010	0.021
	(0.019)	(0.027)	(0.022)
Wealth (+) [Yes = 1]	-0.004	0.012	-0.008
	(0.017)	(0.026)	(0.022)

* Significant at 10%; ** significant at 5%; *** significant at 1%.