Social Security Research Centre

SSRC Working Paper Series
No. 2016-1

Employees Provident Fund Data for Evidence-based Social Protection Policies in Malaysia

Robert Holzmann
Noor Ismawati Mohd Jaafar
Noor Azina Ismail
Halimah Awang

March 2016

About Social Security Research Centre

The Social Security Research Centre (SSRC) was established in March 2011 at the Faculty of Economics and Administration (FEA), University of Malaya to initiate and carry out research, teaching and dissemination of evidence-based knowledge in the area of social security, including old age financial protection in order to enhance the understanding of this critical topic to promote economic development and social cohesion in Malaysia.

To support the research in social security in general and old-age financial protection in particular the Employees Provident Fund (EPF) of Malaysia has graciously provided an endowment fund to create the nation's first endowed Chair in Old Age Financial Protection (OAFPC) at University of Malaya. OAFPC has the over-riding objectives to help formulate policies to promote better social security and improve old age financial protection, and to help formulate policies to promote economic growth in an aging society for consideration by the Government of Malaysia.

The interest in social security and old-age financial protection is ever growing in view of an ageing population. Malaysia is also subjected to rising life expectancy and falling fertility rates, the perceived inadequacy of current social security provisions, coupled with the added fear that simply more expenditure may not be conducive to the development and growth objectives of the society. This calls for innovative policy solutions that may be inspired by international experience based on an empirical grounding in national data and analysis.

Employees Provident Fund Data for Evidence-based Social Protection Policies in Malaysia

Abstract

This paper offers an overview on the sample data drawn from Employees Provident Fund (EPF) – from the raw data to the cleaned and recoded data set ready for research exploration. The original sample of 30,000 EPF members with financial transactions for the years 2002 to 2012 was drawn from the EPF database in 2013. The paper describes rationale and approaches of data cleaning and recoding after which the original sample was reduced to 22,560 cases and made available for analysis which consists of 21,478 Malaysian members and 1,082 non-Malaysian members.

The sample data contains socio-demographic information of the members as well as their transaction records over the 11 year period. Important socio-economic information of the members such as age, gender, nationality and ethnicity is fully available in the cleaned sample; other useful information such as US economic sector and home state of the contributor are highly incomplete and thus suppressed. The information about financial transactions of the sample members is rich and presented for the key financial flows – contributions, dividends, withdrawals, others - as well as for some sub-sets for each of the 11 years as well as the balance for the end of each year. This information should promote a much better understanding of the contribution and dissaving behavior of EPF members along main socio-economic characteristics.

1. Motivation

Good policies need to be guided by rigorous empirical evidence. This applies to all policy areas and in particular to social policies such as social protection: "The softer the issues is or seems, the harder the analyses have to be."[1] To create this empirical evidence of facts and behaviour about individuals with regard to social security schemes requires access to a variety of micro-data at individual level over time, i.e. not simply tabulated mezzo/macro data in variable cross section. To allow for the testing of the causality of hypotheses beyond statistical correlation requires data across time not only at one moment in time across individuals. Only then can hypotheses be more rigorously investigated and tested.

Once such a set of individual/micro data – as a sample and in anonymous form – is available, it does not only allow the exploration and application by a small circle of selected researchers. It opens the ground for a vibrant research community as master and Ph.D. students can finally move their research aspirations to the next level: to test the validity of hypotheses that were created in other countries for Malaysia, and to develop and test their own and new models and hypothesis and test them with Malaysian data. Only such a process will offer students and researchers access to international top journals while providing the Malaysian policy makers with essentially free research results. The way how such a data access has changed the research and policy thinking in other countries is visible in the output of research institutions exemplified by ARC Centre of Excellence in Population Ageing Research (CEPAR) in Sydney, the Center for Retirement Research at Boston College (CRR) in Boston, the Center for Research on Pensions and Welfare Policies (CeRP) in Turin, the Munich Center for the Economics of Aging (MEA) in Munich, formerly Mannheim, and the Social Protection and Labor Global Practice of the World Bank, Washington, DC.

The result of such scientific research is a critical element (but not the only one) of evidence-based policy design and implementation. Once the individual data is available, policy makers and policy institutions

[1] Robert Holzmann, then Sector Director for Social Protection and Labor at the World Bank at a consultation meeting on Human Development issues between the World Bank and UNICEF in New York in autumn 1997.

can also request and receive studies of other areas of their interest on short notice and of international quality.

The purpose of this paper is to describe the EPF data, the background of the request, the process of treating the sample data once received in particular the cleaning and recoding/reclassification process, first results of using the data, and the access to the cleaned and formatted data by researchers. It ends with a few thoughts on the next level of data for social protection research in Malaysia.

2. Types of Data

Data which is needed for such advanced social protection research is broadly grouped into 2 types: Survey data and administrative data. A third group of data emerges from randomized controlled trials that are undertaken to directly explore the effectiveness of government interventions including social protection programs.

The periodic survey data that can typically be used to address some research questions are Household Income and Expenditure Surveys (HIES) and Labor Force Surveys (LFS) and they are nowadays available in all countries, including in Malaysia. A sample for both surveys is accessible for local research needs and contains a lot of information about demographic and economic characteristics and some useful but limited information about relevant social security items. In addition, infrequent and discretionary surveys such as on dissaving behaviour of EPF affiliates can provide useful specific information that is not contained anywhere else.

The administrative data that have much of the requested information about individuals and their social security history resides with social security administration (for the public and private sector), in particular the Civil Service Pension Fund, EPF, LTAT, and SOCSO[2]. The collected and stored information is typically rich and at times complete as regards social security issues (contribution and benefit history) but

[2] For an overview of these institutions and related policies issues, see the website of the Social Security Research Centre, University of Malaya, Mansor *et al.* 2014, Holzmann 2014.

typically limited to a few demographic and economic characteristics of the insured or beneficiary.

For this reason much of the advanced international research works, matching different micro-data sets into one project data set to allow for a much richer investigation. Such matching happens best with the actual matching of individuals via (anonymous) personal identification numbers, else via proxy matching of individuals with identical (or at least similar) characteristics. In order to do so raises some demand on all available and used data sets.

3. Specific Data Request

EPF is the largest social security institution in Malaysia that covers 14.45 million enrolled members and the number of registered employers reaches 537,123, with nearly 6.74 million active members (as of 30 September 2015). As a provident fund it collects contributions (from employees and employers), registers temporary or permanent withdrawals, and records individual holdings of retirement assets at least on a yearly basis. Some of the data may be available across time (such as total individual savings), some only across individuals for several years such as monthly or yearly contributions. But even if free access to such data is restricted, the access to a sample of raw data allows for addressing important initial research questions such as on contribution level and density across socio economic characteristics or the profile of exit or disbursement from EPF by socio economic characteristics.

A 30,000 random sample of all enrolled members was requested as of end-2012 tracing their historical records for the past 11 years, i.e. since January 1, 2002. The sample of the EPF data was requested in anonymous form, i.e. the individual cannot be identified because address or other related information are withheld. The same applies for a personal identification number that may be drawn but withheld except for matching purposes and then again discarded.

The requested sample size was selected to allow for up to 1/3 of data loss during the cleaning process while keeping a sample size that allows for statistically significant differentiation around key characteristics. For example, if regional identification were not

attempted - only significance at national level, the sample size could have been much more reduced. The sample of some 20,000 to 30,000 individuals is about the same size as for the Household Income and Expenditure Surveys (HIES) or the Labour Force Surveys (LFS) referred to above. The original data contained 22 variables. The variables requested were demographic information such as age, gender, marital status; socio-economic data such as industry, state, education level, wage level and social security data such as contributions per month, accumulations, withdrawals, dividends (interests) and other transactions.

The data were provided in two datasets. The first set is the header record for each member containing 11 variables of essentially unchanging socio-economic characteristics for each member while the second set contains 12 variables related to transaction records for the member for 11 years from year January 1, 2002 to December 31, 2012. The variables are presented in Table 1 and 2.

Several meetings were conducted with IT Department of the EPF to understand and seek clarifications about the coding, missing values and ambiguities. For example, there were an unusually high number of transaction codes (1661), high number of missing values such as for the home state variable, ambiguities such as date of birth and unreasonable financial balances.

4. Data Exploration

The EPF provided the sample data in two files with information extracted from the EPF database as of the end of 2012. These data were extracted from a random sampling without any specific sampling framework. This approach was selected as little was known about the composition of the data population except averages produced in annual statistics without closer information. Using the national census as sampling framework would have provided a biased sample as the members of EPF are not a representative sample of the total population.

During the data exploration, several issues pertaining to the supplied sample data needed further clarification. The number of occurrences with respect to the missing and/or unknown values for each variable is contained in Table 1 and Table 2.

It can be observed from Table 1 that there are large numbers of missing values for three variables namely, home state (19,212 cases), nationality (3,885 cases) and ethnicity (race) (1,998 cases). There are also ambiguities in terms of unknown values for the variable gender (48 cases), ethnicity (3,311 cases) and nationality (5 cases), and dubious values for the date of birth (131 cases).

Table 1: Description of variables in the Profile dataset

No	Header/Variable	Description	Notes
1	SEQ NO	Sequence No starting from 1.	
2	SEX	Sex={Male, Female, Unknown}	a=0, b=48
3	RACE	Race={320 codes}	a=1998, b=3311
4	DOB	Date of Birth={YYYY-MM-DD}	c=131
5	NAT	Nationality={261 codes}	a=3 885, b=5
6	NOMINATION	Nomination indicator={4 codes}	
7	STATE CODE	Based on the current correspondence address={14 codes}	a=19 212, b=0
8	ACC 1 BALANCE	Current Balance (Account 1)=in RM	
9	ACC 2 BALANCE	Current Balance (Account 2)=in RM	
10	TOTAL BALANCE	Current Balance (Total Account 1 + Account 2)=in RM	
11	DATE REGISTER	Date Member register to EPF={YYYY-MM-DD}	

Note: a=number of cases with missing value; b= number of cases coded as Unknown; c=dubious

Table 2 refers to the financial transactions for each member in the sample period 2002-2012. This data is transaction oriented i.e. refers to individual inflows and outflows from the accounts such as monthly contributions and irregular withdrawals and the related information about the employer and his socio-economic characteristics. Table 2 indicates that there is a very large number of missing values for the auxiliary code (167,593 cases), employer code (215,076 cases), sector of the employer code (217,260 cases) and industrial code for the employer (221,368 cases).

Table 2: Description of variables in the Transactions dataset

No	Header/Variable	Description	Notes
1	SEQ NO	Sequence No starting from 1.	
2	TRANS CODE	Transaction Code={42 codes}	
3	AUXILIARY CODE	Auxiliary Code={122 codes}	a=167 593
4	TRANS DATE	Transaction Date={YYYY-MM-DD}	
5	CONT MTH	Contribution Month={YYYY-MM-DD}	
6	EMP NO	Employer No (only for Contribution)	a=215 076
7	CR/DR INDICATOR	Debit/Credit Indicator (-ve/+ve)={CR, DR}	
8	AMT	Transaction Amount=in RM	
9	EMP SHARE	Employer Share (only for contribution)=in RM	
10	MEM SHARE	Member Share (only for contribution)=in RM	
11	SECTOR	Sector of the employer (only for contribution transaction), either Government or private ={5 codes}	a=217 260
12	INDUSTRY	Industrial code for the employer (only for contribution transaction)={119 codes}	a=221 368

Note: a=number of cases with missing value

The exploration on a number of variables had to be aborted due to the high missing cases (state, nationality, ethnicity, employer number, sector of employer and industrial code of the employer) and/or the variable itself is not useful since certain information is outdated such as the home state. For industry and sector variables, although the variables were not explored further, these variables were maintained in the database. Subsequently, the variables were either recoded or deleted in the data transformation process.

As shown in Table 1, there were two variables with large number of missing/unknown values namely ethnicity (RACE) and nationality (NAT). The confusion arises from the fact that variable RACE produces the code for main ethnic groups in Malaysia including the sub-ethnic groups and the nationality for the non-Malaysian. Further investigation done via cross-tabulation shows many cases of inconsistency of RACE and NAT. However, some of the ambiguities were addressed in data cleaning and transformation stage.

5. Data Cleaning and Data Transformation

Data cleaning involves the review and elimination of records with dubious entries that cannot be corrected or explained and with missing important information. The data transformation involves the recoding of some of the variables under different headings. In a first stage all members with dubious date of birth (DOB) and date of registration (DATE REGISTER) were deleted from the database. Subsequently members with missing/unknown values for ethnicity and/or nationality were also deleted leaving the database with **25,418** cases, as shown in Table 3.

Table 3: Deleted variables

Original data	=		30,000
(-)DOB	=	131	
(-)DATE REGISTER	=	585	
(-)RACE and NAT	=	3866	
Final sample size	=		25,418

Data recoding was done for variables with a large number of codes, with details below:

1. Ethnicity (RACE)
2. Nationality (NAT)
3. Transaction Codes (TRANS CODE)

Ethnicity (Race) Variable

The data were re-coded to new variables named RACE2 and RACE3. Originally, there were 320 categories of ethnicity and after data exploration the categories were reduced to 5 types of ethnicity, namely:

1. Malay
2. Chinese
3. Indian
4. Other Bumiputera
5. Others.

Re-coding was needed as the ethnicity information in the raw data mixed ethnicity, its sub-groups and nationality (for non-Malaysian). The new classification follows the same style of reporting ethnicity as usually done by Department of Statistics Malaysia (DOSM).

Re-coding was done by assigning codes to Malaysian only, excluding the non-Malaysian. For Non-Malaysian, their information on ethnicity was treated as missing and was assigned the value of 99. For example, a Chinese who holds citizenship of China, Taiwan, USA, Singapore etc. was assigned with code 99, not 2.

Nationality (NAT) Variable

Referring to Table 1 with regard to large number of missing/unknown ethnicity in the data set and the issue of ambiguity between RACE and NAT, several situations emerged from data exploration process, such as between nationality and ethnicity (refer to Appendix).
In order to resolve the issue, the cases with missing value, unknown ethnicity and missing nationality were excluded and the NAT variable is re-coded as NAT2 (0: Malaysia, 1: non-Malaysia).

Transaction (TRANS CODE)

The raw data contained many transaction codes that reflect also special cases and historic episodes that were recoded into a small group of transactions, namely:

1. Contributions (from employee, employer, adjustments)
2. Dividends (i.e. account remuneration)
3. Withdrawals (for defined set of purposes)
4. Others.

In this process, the transaction codes (by referring to TRANS CODE) were recoded into TRANS2. As shown in Table 4 below, TRANS2 differentiate various types of withdrawals. To reduce the complexity of transaction coding, another variable namely TRANS3 was generated, as shown in Table 5.

It was found that there were 73 undefined transactions for 53 EPF members, as shown in Table 6. These transactions codes, however, do not exist in the glossary provided by the EPF.

Table 4: Description of various types of outflows and inflows coded TRANS2

Codes		Description
1	W1	Age 50 years withdrawal (including adjustments)
2	W2	Age 55 years/ Pensionable employees/ Optional retirement withdrawal/ Pension to Government/ Annuity/ Periodical Payment (including adjustments)
3	W3	Withdrawal to Reduce/ Redeem Housing Loan (including adjustments)
4	W4	Health/ Incapacitation/ Death Withdrawal (including adjustments)
5	W5	Leaving Country Withdrawal (including adjustments)
6	W6	Education Withdrawal (including adjustments)
7	W7	Withdrawal of Savings of More Than RM1 Million (including adjustments)
8	W8	Hajj Withdrawal (including adjustments)
9	C	Contribution (all types of contribution including adjustments)
10	D	Dividend (all types of dividend including adjustments and dividend on withdrawals)
11	O	Amalgamation of account

Table 5: Description of various types of withdrawals coded as TRANS3

Codes		Description
1	C	Contribution (all types of contribution including adjustments)
2	D	Dividend (all types of dividend including adjustments and dividend on withdrawals)
3	W	Withdrawals (all types of withdrawals including adjustments)
4	O	Amalgamation of account

Table 6: Undefined transaction codes

TRANSACTION CODE	AUXILIARY CODE		
	3028	3029	Total
364	53	0	53
366	0	10	10
982	0	10	10
Total	53	20	73

Note: Values indicate the number of transactions in the database.

Table 7 presents the distribution of transactions based on the original sample data provided by the EPF. The complexity of the data is due to the fact that, although there are 30,000 member cases, there were 1,369,698 transactions before the data cleaning process.

Table 7: The distribution of TRANS3

Code	Description	Credit	Debit	Whole sample	
				Total	Percentage
1	Contribution	1,178,077	4,302	1,182,379	84.7%
2	Dividend	21	33,340	33,361	2.4%
3	Withdrawal	173,267	7,656	180,923	13.0%
4	Amalgamation of account	35		35	0.0%
Total		**1,351,400**	**45,298**	**1,396,698**	**100%**

Note: Generated from original data sample
Excluding the 73 unidentified transactions
Frequency shows the number of transactions

Table 8 presents the distribution of withdrawals by purpose. Withdrawal at age 55/ Pensionable employees/ Optional retirement withdrawal/ Pension to Government/ Annuity/ Periodical Payment (W2) constitute the highest cases of withdrawals (46.6%), followed by withdrawals of savings of more than RM1 million (W7) of 23.1% and withdrawals to reduce / redeem housing loan (W3) of 16.4%. The lowest is leaving the country withdrawals (W5) with only 1.8%. There is no record for withdrawals for pilgrimage (Hajj) since this activity was only introduced to EPF members in 2013 and this sample stops at 2012.

Table 8: Distribution of withdrawals

Type of withdrawals	No. of withdrawal
Age 50 years withdrawal (W1)	2,101 (6.3%)
Age 55 years/ Pensionable Employees/ Optional Retirement Withdrawal/ Pension to Government/ Annuity/ Periodical Payment (W2)	15,549 (46.6%)
Withdrawal to Reduce/ Redeem Housing Loan (W3)	5,454 (16.4%)
Health/ Incapacitation/ Death Withdrawal (W4)	944 (2.8%)
Leaving Country Withdrawal (W5)	617 (1.9%)
Education Withdrawal (W6)	971 (2.9%)
Withdrawal of Savings of More Than RM1 million (W7)	7,704 (23.1%)
Total	**33,340**

The final data contains 22,560 individual records (21,478 Malaysian, 1,082 non-Malaysian) excluding members with no historical transaction records and those with negative outstanding balance as well as those with unidentified transactions (Table 9). While those members in the last two categories are insignificant in number and share, members with no financial transaction amounted to about 9 percent of the full sample and almost 11 percent of the reduced sample.

Table 9: Data cleaning and transformation process

Original data	=	30,000
(-)DOB	=	131
(-)DATE REGISTER	=	585
(-)RACE and NAT	=	3866
Sample size (after deletion at stage 1)	=	25,418
(-)no historical transaction records	=	2,766
(-)negative balance	=	39
(-)undefined transaction codes	=	53
FINAL SAMPLE SIZE (after deletion at stage 2)	**=**	**22,560 (21,478 Malaysian, 1,082 non-Malaysian)**

Members without financial transaction during the 11 years observation period may include individuals that joined the public sector after a few years in the private sector and kept their accounts with EPF. It may also include individuals that retired a long time ago and kept a minimum amount and left untouched. According to the information received, member accounts are so far never closed, even in case of death which is not communicated to EPF and thus recorded.

6. Data Available for Analysis

The data available for analysis consist of 22,560 cases, 21,478 Malaysians and 1,082 non-Malaysians, and 1,658 variables as shown in Table 10. For purposes of analysis, cases refer to EPF members in this sample data are used interchangeably throughout this report.

Table 10: List of variables

No.	Variable	Descriptions	Values
1	idCode	Identification Code	Number from 1 to 22,560
2	SEX	Gender	F= Female, M= Male
3	RACE3	Ethnicity (Definition 3)	1= Malay, 2= Chinese, 3= Indian, 4= Other Bumiputera, 5= Others
4	DATE OF BIRTH	Date of birth	Year-month-day
5	AGE	= 2012 minus year of birth	Full years
6	NATIONALITY	Nationality	The actual country
7	NATIONALITY 2	Nationality (Malaysian and Non-Malaysian)	M= Malaysian, N= Non-Malaysian
8	NOMINATION	Nomination status of right to bequest	E= Has nomination but information cannot be determined, N= No nomination, X= Has nomination but information incomplete, Y= Has nomination
9	STATE CODE 2	State (based on latest reported address)	1= Johor, 2= Kedah, 3= Kelantan, 4= Melaka, 5= Negeri Sembilan, 6= Pahang, 7= Pulau Pinang, 8= Perak, 9= Perlis, 10= Selangor, 11= Terengganu, 12= Sabah, 13= Sarawak, 14= W.P Kuala Lumpur, 15= W.P Labuan, 16= W.P Putrajaya
10	ACCOUNT 1	Balance in ACCOUNT 1 (as of end of 2012)	In RM
11	ACCOUNT 2	Balance in ACCOUNT 2 (as of end of 2012)	In RM
12	TOTAL BALANCE	Total Balance (ACT1 + ACT 2) (as of end of 2012)	In RM
13	DATE OF REGISTRATION	Date of EPF account registration	Day-month-year

No.	Variable	Descriptions	Values
14	MEMBERSHIP YEARS	Number of membership years	Full years (until end of 2012)
15	ACODE	Activeness status (in 2012 only)	0= Not active, 1= Active
16	B200i	Total balance for the year 200i, where i=2,3,4…12	In RM
17	C200i	Total cash flow in the year 200i, where i=2,3,4…12	In RM
18	C200iYC	The total net amount of contribution made in the year 200i, where i=2,3,4…12	In RM
19	C200iYD	The total net amount of dividend received in the year 200i, where i=2,3,4…12	In RM
20	C200iYW	The total net amount of withdrawal made in the year 200i, where i=2,3,4…12	In RM
21	C200iYO	The total net amount transferred to other account (Amalgamation of account) in the year 200i, where i=2,3,4…12	In RM
22	C200iMjWk	The total net amount of withdrawal type k made in month j year 200i, where i=2,3,4…12, j=1,2,3…12, k=1,2,3,4,5,6,7.	In RM
23	C200iMjC	The total net amount of contribution made in month j year 200i, where i=2,3,4…12, j=1,2,3…12.	In RM
24	C200iMjD	The total net amount of dividend received made in month j year 200i, where i=2,3,4…12, j=1,2,3…12.	In RM
	C200iMjO	The total net amount transferred to other account (Amalgamation of account) in month j year 200i, where i=2,3,4…12, j=1,2,3…12.	In RM

Note: Total balance is calculated using reverse method starting in total balance in end-2012 minus the total cash flow during each year. Total cash flow = total net cash inflow – total net cash outflow.

7. A First Descriptive Analysis

An initial descriptive analysis was performed on the cleaned and recoded 22,560 individual records (Table 11). Males constitute 54% of the total sample. Among Non-Malaysians, majority are Indonesians (69%), followed by Bangladeshis (19%). Other nationalities which accounted for the remaining 12 percent include Filipinos (3%) and Indians (2%).

Table 11: Socio-demographic information of EPF members 2002-2012

Variable	Frequency	Percentage
Gender		
Female	10,375	46.0%
Male	12,185	54.0%
Total	22,560	100.0%
Nationality		
Malaysia	21,478	95.2%
Non-Malaysian	1,082	4.8%
Total	22,560	100.0%
Other Nationality		
Australia	2	0.0%
Bangladesh	205	18.9%
Britain	2	0.2%
China	4	0.4%
Cambodia	1	0.1%
Canada	3	0.3%
Denmark	1	0.1%
Indonesia	751	69.4%
India	25	2.3%
Japan	6	0.6%
New Zealand	1	0.1%
Others	18	1.7%
Philippines	31	2.9%
Pakistan	1	0.1%
Singapore	13	1.2%
Thailand	12	1.1%
Taiwan	1	0.1%
United Kingdom	4	0.4%
United States of America	1	0.1%
Total	1,082	100.0%

Source: EPF Sample Data Profile

The initial descriptive analysis also reported the socio-demographic of the EPF active members as shown in Table 12. An EPF member is considered active if there is at least one contribution made at any point of time during the year, regardless of the amount contributed. The total active members in 2012 derived from the total sample was 10,005 (44.3%). Males constitute 56.3% of the total active members. Among Non-Malaysians, majority are Indonesians (45.2%), followed by Indians (12.9%).

Table 12: Socio-demographic information of the active members in 2012

Variable	Frequency	Percentage
Gender		
Female	4,373	43.7%
Male	5,632	56.3%
Total	10,005	100.0%
Nationality		
Malaysia	9,974	99.7%
Non-Malaysian	31	0.3%
Total	10,005	100.0%
Other Nationality		
Britain	2	6.5%
China	2	6.5%
Indonesia	14	45.2%
India	4	12.9%
Japan	3	9.7%
United Kingdom	1	3.2%
Philippines	1	3.2%
Singapore	2	6.5%
Taiwan	1	3.2%
Others	1	3.2%
Total	31	100.0%

Source: EPF Sample Data Profile

Subsequent analyses were performed on Malaysian cases only which consisted of 21,478 sample cases. The socio-demographic profile of the Malaysian sample is shown in Table 13. Males constitute 52.9% of the total Malaysian sample. More than half of the sample cases are Bumiputeras (59.1%) which comprised of Malays and other Bumiputeras, followed by Chinese (30.6%), Indians (9.2%) and other ethnic groups (1.1%). The table also shows 55.6% of cases with unknown status.

Table 13: Socio-Demographic Information of Malaysian EPF sample in the final data as of 2012

Variable	Frequency	Percentage	Variable	Frequency	Percentage
Gender			Ethnicity		
Female	10,120	47.1%	Bumiputera		
Male	11,358	52.9%	Malay	10,949	51.0%
Total	21,478	100.0%	Other Bumiputera	1745	8.1%
State			Chinese	6567	30.6%
Johor	1,261	5.9%	Indian	1980	9.2%
Kedah	609	2.8%	Others	237	1.1%
Kelantan	331	1.5%	Total	21,478	100.0%
Melaka	357	1.7%			
Negeri Sembilan	429	2.0%			
Pahang	488	2.3%			
Pulau Pinang	639	3.0%			
Perak	842	3.9%			
Perlis	56	0.3%			
Selangor	2,127	9.8%			
Terengganu	313	1.5%			
Sabah	566	2.6%			
Sarawak	647	3.0%			
W.Persekutuan (Kuala Lumpur)	812	3.8%			
W.Persekutuan (Labuan)	29	0.1%			
W.Persekutuan (Putrajaya)	33	0.2%			
Unknown	11,939	55.6%			
Total	21,478	100.0%			

Source: EPF Sample Data Profile

Of the total 21,478 Malaysian EPF members, 9,974 of them were active members (46.4%). Males constitute 56.3% of the total Malaysian active members (Table 14). More than half of the members are Bumiputeras (57.7%) followed by Chinese (32.4%), Indians (8.9%) and other ethnic groups (1.0%). The table also indicates that 37.9% of active members are with unknown status.

A distribution of the active and non-active members by age groups is provided in Figure 1. The distributions suggests a rapidly increasing active participation from the age of 16 to the age of 28 and continues well after retirement. The non-active members increase to the mid-30s,

Table 14: Socio-demographic information of Malaysian active members in 2012

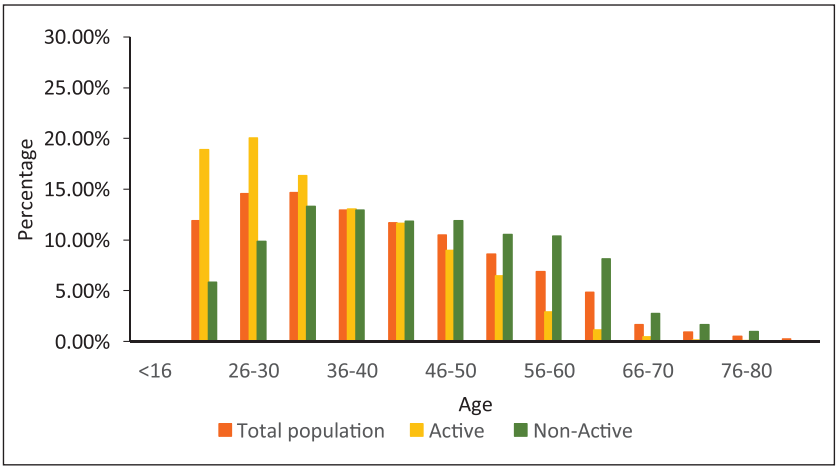
Variable	Frequency	Percentage	Variable	Frequency	Percentage
Gender			Ethnicity		
Female	4,361	43.7%	Bumiputera		
Male	5,163	56.3%	Malay	4,959	49.6%
Total	9,974	100.0%	Other Bumiputera	804	8.1%
State			Chinese	3,230	32.4%
Johor	754	7.6%	Indian	883	8.9%
Kedah	375	3.8%	Others	98	1.0%
Kelantan	219	2.2%	Total	9,974	100.0%
Melaka	230	2.3%			
Negeri Sembilan	255	2.6%			
Pahang	324	3.2%			
Pulau Pinang	431	4.3%			
Perak	481	4.8%			
Perlis	33	0.3%			
Selangor	1,470	14.7%			
Terengganu	201	2.0%			
Sabah	377	3.8%			
Sarawak	432	4.3%			
W.Persekutuan (Kuala Lumpur)	574	5.8%			
W.Persekutuan (Labuan)	23	0.2%			
W.Persekutuan (Putrajaya)	18	0.2%			
Unknown	3777	37.9%			
Total	9,974	100.0%			

Source: EPF Sample Data Profile

remain almost constant till the mid-60s, and decline rapidly thereafter. This creates a front-loaded picture for both active groups in the early ages and almost linear and steep reduction till the mid-60s. This picture offers first indication where and what to investigate are when to explore the low financial balances for most of the EPF members.

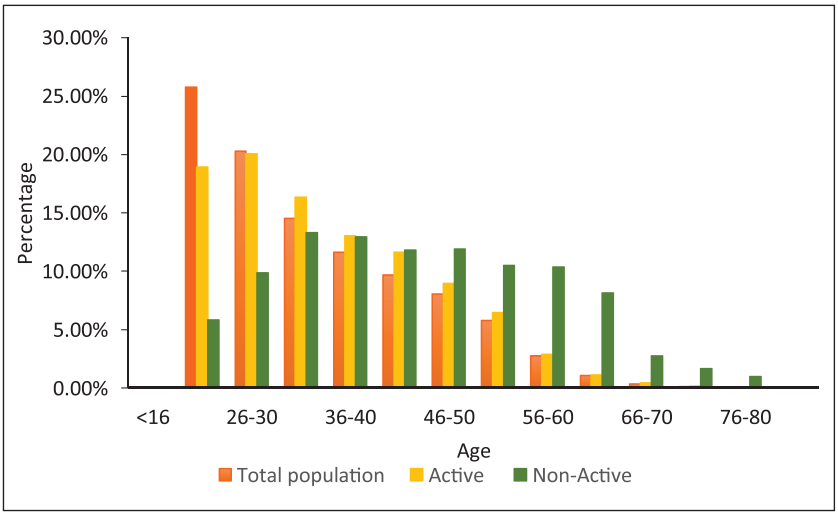
Figure 2 explores the distribution of active and non-active members as well as the population by age groups taken from the EPF data of 2012. The comparison with Figure 1 suggests that the sample data is broadly representative of the total EPF data except for the 16-25 age group where a slight under-polling may have taken place.

Figure 1: Distribution of the total Malaysian EPF members (based on sample data) with active and non-active status as of year 2012



Source: EPF Sample Data Profile

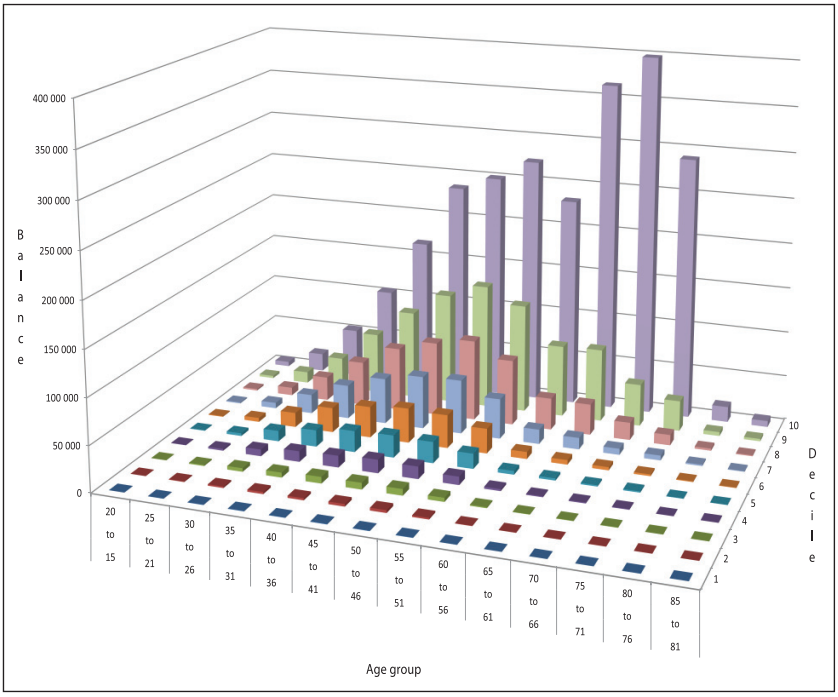
Figure 2: Distribution of the total EPF members' population (based on EPF Report 2012)



Source: EPF Report 2012

In order to offer a glimpse of the potential power of the data set for policy analysis we close with a presentation of the financial balances by deciles and age groups. Figure 3 presents the total mean balance (i.e., account 1 and 2) of those individuals with a positive balance by end-2012, by age group and by decile of balance value in RM. The balances are very small for almost two-thirds of those age groups that should have the highest balance: those aged 46-50, before account 1 can be fully accessed; and those aged 50-55, before account 2 can be fully withdrawn. Only the highest three deciles have accumulation of some significance at retirement. The development of lower deciles suggests major issues with contribution effort/contribution density at younger ages, and main withdrawals soon after reaching retirement age; the highest decile suggests no contribution issues at younger ages and the use of the EPF as an investment vehicle into older ages by rich retirees.

Figure 3: Total Mean Balance of Malaysians' EPF Accounts (as of end-2012, by age group and decile, in RM)



Source: EPF Sample Data Profile

8. Summary, Access to EPF data and outlook

This paper offers an overview on the EPF sample data – from the raw data to the cleaned and recoded data set ready for research exploration. The original sample of 30,000 EPF members with financial transactions for the years 2002 to 2012 was drawn from the EPF database in 2013. After various stages of data cleaning and recoding, the sample was reduced to 22,560 cases made available for analysis which consists of 21,478 Malaysian members and 1,082 non-Malaysian members.

The sample data contains socio-demographic information of the members as well as their transaction records over the 11 year period. Important socio-economic information of the members such as age, gender, nationality and ethnicity is fully available in the cleaned sample; other useful information such as us economic sector and home state of the contributor are highly incomplete and thus suppressed. The information about financial transitions of the sample members is rich and presented for the key financial flows – contributions, dividends, withdrawals, others – as well as for some sub-sets for each of the 11 years as well as the balance for the end of each year. This information should provide for better understanding of the contribution and dissaving behavior of EPF members along main socio-economic characteristics.

Access to the cleaned data set in Stata and other formats is provided for researchers through the Social Security Research Centre of the University of Malaya. This requires a nonprofit research proposal, confirmation of project intention by a Malaysian research institution, and the commitment to share the results with SSRC and publication of the main results in SSRC working paper series.

Next planned steps of EPF data use and enhancement include the linking of the data set with other sample data – in particular Household Income and Expenditure Survey and Labor Force Survey – as well as other administrative sample data from SOCSO and other institutions. Further down the road SSRC hopes to engage with other research institutions in Malaysia in the development and implementation of longitudinal micro-data surveys such as SHARE (Survey of Health, Ageing and Retirement in Europe) and similar efforts in Asia (China, India, Indonesia and Japan).

Acknowledgement

The authors are Professor of Economics and former chair holder of Old-Age Financial Protection, Lecturer of Statistics, Professor of Statistics, and Senior Research Fellow at SSRC, respectively. They have participants of two workshops to thank where the first results were presented and pertinent questions and comments were raised, Mr. Wong Theen Chuan (EPF) and Ms. Nur Intan Shaffinas binti Salehud-din (SSRC) for their assistance in producing the working paper, and Professor Norma Mansor, the Director of SSRC, for her continued crucial support and encouragement. The authors and all future researchers using the data are eternally indebted to the EPF leadership for providing access to this sample data.

References

Employees Provident Fund. (2012). Annual Yearbook 2012. Kuala Lumpur: EPF.

Holzmann, R. (2014). Old-Age Financial Protection in Malaysia: Challenges and Options. *SSRC Working Paper Series No. 2014-2*. Kuala Lumpur: University of Malaya.

Mansor, N., Salleh, S.N.S., Tan, L.Y., Koutronas, E. and Aikanathan, S. (2014). Social Security in Malaysia: Stock-take on Players, Available Products and Databases. *SSRC Working Paper Series No. 2014-3*. Kuala Lumpur: University of Malaya.

About the authors

Professor Dr. Robert Holzmann

Robert Holzmann, professor of economics, currently is the Distinguished Research Fellow in SSRC. Formerly, he was the Chair of Old Age Financial Protection (OAFPC) at the Faculty of Economics and Administration, University of Malaya since 2012. He is inter alia Honorary Chair, Centre of Excellence in Population Ageing Research (CEPAR), University of New South Wales and Research Fellow of Institute for the Study of Labor (IZA), Bonn and CESifo Munich. He also serves as consultant to the World Bank on financial literacy & education, migration, and pension issues. Before his return to academia, he was the Research Director of the Labor Mobility Program (Marseille Center for Mediterranean Integration), Senior Advisor of the Financial Literacy & Education Program (Russia Trust Fund), and for 12 years Sector Director and Head of the Social Protection & Labor Department leading, inter alia, the strategic and conceptual work on pensions and labor at the World Bank. Before joining the World Bank he was professor of economics and director to the European Institute at the University of Saarland, Germany, professor of economics at the University of Vienna, Austria, and senior economist at IMF and OECD. He was also Visiting Professor at various universities in Japan, Chile and Austria, and lectured at Harvard University (USA) and Oxford University (UK). His research and operational involvement extends to all regions of the world, and he has published 34 books and over 150 articles on social, fiscal and financial policy issues. His strength is strategic thinking, research organization, and innovative research. He has a broad interest in economic issues covering social, fiscal and financial issues. His life-long specialization is pensions where he is considered as one of the world's leading experts. His most recent and ongoing research and country' consultations cover the areas of financial literacy and education, the economics of aging, migration and labor markets.

Noor Ismawati Mohd Jaafar

Noor Ismawati Mohd Jaafar is currently served as a lecturer at Department of Applied Statistics, Faculty of Economics and Administration, University of Malaya. She has a Master of Science in Mathematics (Actuarial Science) from University of Connecticut (UConn), a Bachelor of Science and a Diploma in Actuarial Science from Universiti Teknologi Mara (UiTM), Malaysia. Her research interest includes consumption studies among elderly, mortality modelling and social statistics. She has also undertaken research with academicians from various areas of studies and be part of many consultation teams handing sampling and data management which includes data collection, processing and analyses to both public and private organization in Malaysia.

Professor Dr. Noor Azina Ismail

Noor Azina Ismail is a professor at the Department of Applied Statistics and currently Dean at the Faculty of Economics and Administration, a multidisciplinary faculty at the University of Malaya. Besides her duty as Dean of the faculty, she is also a Vice-President of Malaysian Economics Association. Previous positions held by her include Head, Department of Applied Statistics and Deputy Dean (Undergraduate). She obtained her degrees in Bachelor of Science (Hons.) majoring in Statistics and Master of Statistics from the University of New South Wales, Australia. She joined University of Malaya in 1991 and a few years later, went back to Australia to continue her study. In 2000, she received the doctoral degree from the Queensland University of Technology, Brisbane, Australia. She is the first Malaysian academic to gain PhD in Medical Statistics. Noor Azina is an established academician and has to-date published more than 100 publications, including journal articles, conference proceedings, books, chapter in books, monographs, working papers and research reports for both private and government agencies. She has presented papers in various seminars and conferences, both locally and internationally. Her specific methodological interests are in hierarchical modelling, structural equation modelling, Bayesian statistics and mixture models. Her applied interests are in public health, biostatistics, mathematics of education and social statistics.

Dr. Halimah Awang

Halimah Awang is a senior research fellow at the Social Security Research Centre, University of Malaya. Prior to this appointment she was a lecturer and Associate Professor at the Faculty of Economics and Administration for almost 30 years. She also served as deputy dean and head of the Department of Administrative Studies and Politics. Halimah holds a PhD in Applied Statistics from Macquarie University, Australia. Her research interests include social protection, sexual and reproductive health, women, youth, ageing and poverty. She has published extensively in both local and international refereed journals as well as book chapters. Her involvement in research and consultancy projects include those commissioned by National Population and Family Development Board, The World Bank, PLUS Expressways Berhad, Ministry of Women, Family and Community Development, Social Institute Malaysia, Cherie Blair Foundation and Ministry of Higher Education.

Recent Publications

No. 2014-1 : Social Security: Challenges and Issues

No. 2014-2 : Social Security in Malaysia: Stock-take on Players, Available Products and Databases

No. 2014-3 : Old-Age Financial Protection in Malaysia: Challenges and Options

No. 2015-1 : Framing Social Protection Analysis in Malaysia: Issues For Consideration

Social Security Research Centre (SSRC)
Faculty Economics and Administration
University of Malaya
50603 Kuala Lumpur, Malaysia.
Tel: 03- 7967 3615 / 3774
Email: ssrc@um.edu.my
Website: http://ssrc.um.edu.my