ESTIMATION OF THE NUMBER OF ACCOUNT DUPLICATES GENERATED DURING A PERIOD OF TIME IN A NATIONAL PENSION SYSTEM

Juan José Fernández-Durán [1] and Rafael Gamboa Hirales [2]

[1] *Departamento de Estadística, Instituto Tecnológico Autónomo de México*
[2] *Departamento de Computación, Instituto Tecnológico Autónomo de México*

ABSTRACT

Consider a national pension system with the following characteristics. At the end of every two-month period the employers paid contributions to the pension system for the employees that have worked during that period. These contributions are deposited in a bank account. Each employee has one account number. There is a regulatory office which main tasks are to assign new account numbers for the employees which work for the very first time to an employer registered in the pension system and to manage the account database. Every time an employee moves to a new job the new employer is responsible to notify the regulatory office. In the case that the employee is working for the very first time, a new account number is assigned to the employee otherwise, just the name of the new employer is recorded. Since the employer is responsable of notifying the regulatory office this produces errors in the employee information. In doubt, the regulatory office always assign a new account number. This procedure generates account duplicates i.e. employees with more than one account number, increasing the management costs of the pension system and producing identification problems when the employee reaches his/her retirement age. In this paper we present a methodology to estimate the total number of account duplicates that have been generated in the pension system during a determined period of time (a year) since the starting date of the pension system. The methodology is based in the comparison of two Markovian system. One of the systems is derived from national labour statistics and the other is derived from the contribution histories of all the accounts in the pension system. Also, we present techniques to determine the characteristics of the employees with a high probability of having account duplicates.

1. Introduction: Characteristics of the National Pension System

In this paper we restrict attention to national pension systems with the following characteristics:

a)Employers paid contributions at the end of every two-month period (end of February, end of April, etc) for the employees that have worked during that period. The contribution is deposited in a bank account. Every employee in the system has an account number.

b)There exists an independent regulatory office. The main tasks of the regulatory office are to assign new account numbers to each employee that works for the very first time with an employer registered in the pension system and to manage the account database. The account database has additional employee information such as name, sex, date of birth, place of birth, time at which the employee is registered in the pension system for the very first time and the contribution history of the employee. The contribution history can be reduced to a sequence of zeros and ones where a one denotes that a contribution has been made to the employee during a particular two-month period and a zero denotes that no contribution has made to the employee during a particular two-month period. Then an annual contribution history consists of a sequence of six zeros and ones.

c)The employer is responsible of notifying and giving the employee information to the regulatory office every time he/she hires a new employee. If with the information provided by the employer the regulatory office identifies the employee as one which already has an account number then only the name of the new employer is register but, if the employee cannot be identified then a new account number is assigned to the employee.

This procedure implemented by the regulatory office guarantees that a contribution made for a particular employee will not be wrongly assigned to another employee. Then, in case of any doubt about the identity of the employee the regulatory office prefers to assign a new account number. On the other side, this procedure generates account duplicates, that is, employees with more than one account number. There are many reasons for the generation of account duplicates. The two main reasons are the following:

1)employees with the same name and very similar additional information and,
2)errors in the information provided by the employer to the regulatory office.

This paper presents a methodology for the estimation of the number of accounts duplicated during a predetermined period of time (a year) from the starting date of the national pension system. The paper is divided into five sections. In the first section we present the possible states that an employee can occupy at the end of a two-month period (when a contribution to the pension sytem could be made by the employer). In section 2 the Markovian system derived from official labour statistics is developed. In section 3 we derive the initial distribution for the employees in the pension system and outside the pension system and the corresponding observed final distribution at the end of the year.

This observed initial and final distributions are derived from the contributions history of the empoyees in the pension system and some official labour statistics.
In section 3 the estimation procedure is presented. In Section 4 some ideas for the identification of the accounts with a high probability of being duplicates are developed. Finally, the conclusions and ideas for future research are included in Section 5.

2. Labour – Contribution States

The following diagram shows the possible states and transitions that an employee (not necessarily in the pension system) can experiment and are relevant to the present study.

FIGURE 1: States and Possible Transitions

At the end of every two-month period an employee should only be in one and only one of the possible three 1, 2 or 3:

a)State 1: An employee (account) who has received a contribution at the end of the two-month period.
b)State 2: An employee (account) who has not received a contribution at the end of the two-month period.
c)State 3: An employee who has never received an account in the pension system.

An employee with an account in the pension system should only be in states 1 or 2 at the end of a two-month period.

The transitions that can occur from the end of the two-month period k to the end of the two month period k+5 (a year) are as follows:

a)T(1,1): An employee who at the end of the two-month period k has received a contribution and at the end of the two-month period (k+5) also receives a contribution.

b)T(1,2): An employee who at the end of the two-month period k has received a contribution and at the end of the two-month period (k+5) does not receive a contribution.

c)T(1,3): An employee who at the end of the two-month period k has received a contribution and at the end of the two-month period (k+5) does not have an account in the pension system. This transition is impossible to occur.

d)T(2,1): An employee who at the end of the two-month period k has not received a contribution and at the end of the two-month period (k+5) receives a contribution.

e)T(2,2): An employee who at the end of the two-month period k has not received a contribution and at the end of the next two-month period (k+1) does not receive a contribution.

f)T(2,3): An employee who at the end of the two-month period k has not received a contribution and at the end of the two-month period (k+5) does not have an account in the pension system. This transition is impossible to occur.

g)T(3,1): An employee who at the end of the two-month period k does not have and account in the pension system (outside the national pension system) and at the end of the two-month period (k+5) has an account in the pension system and has received a contribution at the end of the two-month period (k+5).

h)T(3,2): An employee who at the end of the two-month period k does not have and account in the pension system (outside the national pension system) and at the end of the two-month period (k+5) has an account in the pension system and has not received a contribution at the end of the two-month period (k+5).

i)T(3,3): An employee who at the end of the two-month period k does not have and account in the pension system (outside the national pension system) and at the end of the two-month period (k+5) does not have an account in the pension system (remains outside the pension system).

Thus, new accounts for the pension system can only proceed from employees in state 3. The problem with account duplicity is that employees in states 1 or 3 can also be assigned new account numbers.

3. Markovian System Derived from Official Labour Statistics

Let S(k) denote the state in which a particular employee is at the end of the two-month period k.

The diagram depicted in Figure 1 determines an annual Markovian system (Ross (2000)) with a 3 by 3 matrix of annual transition probabilities with element (i,j), i=1,2,3 and j=1,2,3 denoted by

$$p_{ij}=\mathrm{Prob}\{S(k+5)=j \mid S(k)=i\}$$

For the estimation of the transition probabilities we assume that there exists a longitudinal employment survey (generally carried by the goverment) in which every three months (end of March, end of June, ..., etc) a panel of employees are interviewed and the state occupied 1, 2 or 3 is determined.

By considering a panel of employees that were interviewed during four consecutive three-month periods the annual transition matrix of the Markovian system can be estimated. Depending of the stratification implemented in the sampling design of the longitudinal employment survey it is possible to obtain annual transition matrices for different groups of employees, for example, for groups determined by variables such as sex, age, place of birth, etc.

Let P denote the annual transition matrix estimated from the data obtained from the longitudinal employment survey. The estimates are obtained by maximum likelihood from the individual labour-contribution histories of the employees interviewed in the longitudinal employment survey.

4. Estimation of the total number of account duplicates

At the end of the first two-month period of the national pension system every employee in the nation was in state 1 or state 3 since only for those employees with a contribution a new account was assigned. Then, at the end of the first two-month period we have observed:

$N_1(1)$: The total number of employees in state 1 at the end of the first two-month period.

$N_2(1)$: The total number of employees in state 2 at the end of the first two-month period. This is equal to zero.

$N_3(1)$: The total number of employees in state 3 at the end of the first two-month period. This number is obtained from national labour statistics as the total labour force minus $N_1(1)$.

Then, we have observed the vector

$$\underline{N}(1)=(N_1(1), N_2(1), N_3(1))$$

and we want an estimate of the vector

$$\underline{N}(6)=(N_1(6), N_2(6), N_3(6))$$

that is, the total number of employess in the different states at the end of the sixth two-month period (after one year of the starting date of the national pension system). It is important to note that the increment in the total national labour force due to persons who reach the minimum legal age to work is considered negligible. The estimate of $\underline{N}(6)$, $\underline{E}(6)$, is obtained by applying the annual transition matrix, P, estimated with data from the longitudinal employment survey, to $\underline{N}(1)$ as follows:

$$\underline{E}(6) = \underline{N}(1)P$$

Also, at the end of the sixth two-month period we have observed the total number of accounts (number of employees plus number of account duplicates) in each of the different three states

$$\underline{N^*}(6)=(N^*_1(6), N^*_2(6), N^*_3(6))$$

The vector $\underline{N}^*(6)$ is different from the vector $\underline{N}(6)$ since it includes the duplicated accounts. The estimated number of account duplicates generated during the first year of the national pension system, denoted by D(1,6), is calculated as follows

$$D(1,6)= N^*_1(6) + N^*_2(6) - N_1(6) - N_2(6)$$

Depending on the stratification of the longitudinal employment survey and the additional information that the pension system has about the employees, the estimate of the total number of account duplicates can be obtained for different groups. The groups can be determined by considering variables such as sex, date of birth, place of birth, geographic area, etc.

5. Identification of the characteristics of account duplicates

A problem related with the estimation of the number of account duplicates is the identification of the characteristics (sex, age, place of birth, geographic area, and any additional variable in the system) of the accounts suspected of being duplicates. The main assumptions about the behavior of the contribution history of accounts suspected of being duplicates are:

1)Accounts with irregular contribution patterns are suspected of being duplicates and,
2)Accounts for which a contribution has not been made since many two-month periods are suspected of being duplicates.

Based on these assumptions the following measure, which is easy to implement in a huge database (Pyle (1999)) and we call it volatility of the contributions, is developed for the identification of accounts suspected of being duplicates:

Volatility=number of two-month periods without a contribution/Total number of two-month periods since the account was created

The volatility is a number between zero and one. Obviously, accounts with a volatility equal to zero are those which every two-month period have had a contribution then this type of accounts are not suspected of being duplicates (we are not suspecting of accounts opened since few two-month periods). Accounts with volatility near to one are highly suspected.

Again, this analysis can be done for different groups depending on variables such as sex, age, geographic region, etc. Once this groups have been formed it is possible to obtain a distribution of the volatility values in each of these groups and to determine the suspicious groups as those which present high percentage of low volatility values.

6. Conclusions

We have presented a procedure for the estimation of the total number of account duplicates generated during a determined period of time in a national pension system.

The procedure is based on the construction of a matrix of transition probabilities from the data obtained from a longitudinal employment survey. Then, this matrix of transition probabilities is applied to the initial composition of the pension system to obtain the expected composition at the end of the determined time period. The difference between the expected numbers and the observed numbers is an estimate of the total number of account duplicates. This estimate is useful in order to determine the magnitude of the problem of account duplicates and to decide if it is reasonable to change the procedure for the assignment of new account numbers.

We also present a measure, which we call volatility, that is applied to the contribution history of each account in the system. Once the volatility is obtained for each account in the system it is possible to identify the groups wiht high values of the measure and to determine their characteristics. This information is useful to implement a strategy for the reduction of account duplicates in certain groups.

7. Bibliography

Ross, S. (1996) *Stochastic Processes,* 2nd edition, John Wiley & Sons
Pyle, D. (1999) *Data Preparation for Data Mining*, Morgan Kaufmann Publishers

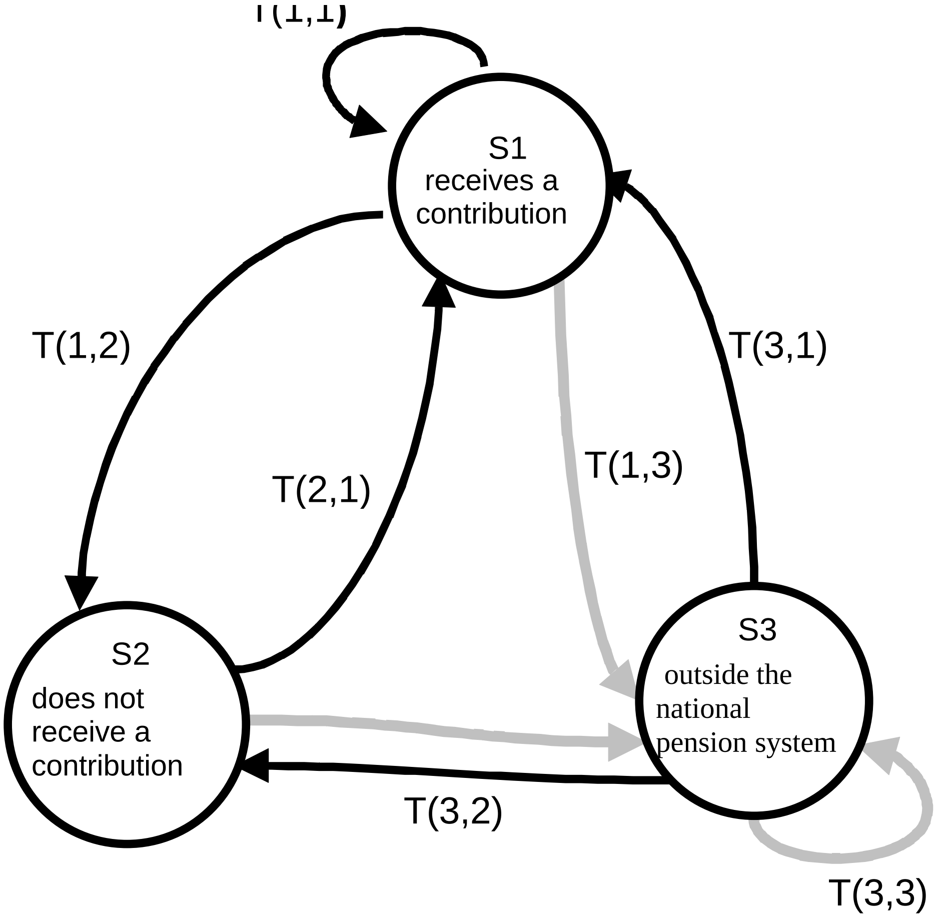

FIGURE 1. Labour- Contribution Markovian system to estimate the total number of account duplicates. States and possible transitions.