

Deterministic record linkage: a case study by using demographic data

Francesca Rinesi¹, Sabrina Prati², Claudia Iaccarino²,

¹ Department of Social, Economic, Actuarial and Demographic Studies,
“Sapienza” University of Rome, Italy;

² Italian National Institute of Statistics (ISTAT), Italy.

e-mail: francesca.rinesi@uniroma1.it; prati@istat.it; iaccarin@istat.it.

Abstract: One of the aims of record linkage techniques is to join information that refers to the same individual but which is stored in two or more data sources. Here we present the preliminary results of a case study by applying this data integration procedure to demographic data. The research question that drives our analysis was the evaluation of the predictive power of fertility intentions in anticipating the subsequent behaviour. In order to hit achieve this goal we linked the information stored in a survey with a national administrative register and quality data measures were provided. Here we have reported the first results of our analysis as well as possible future developments.

Keywords: deterministic record linkage, demographic data, fertility intentions.

1. Introduction

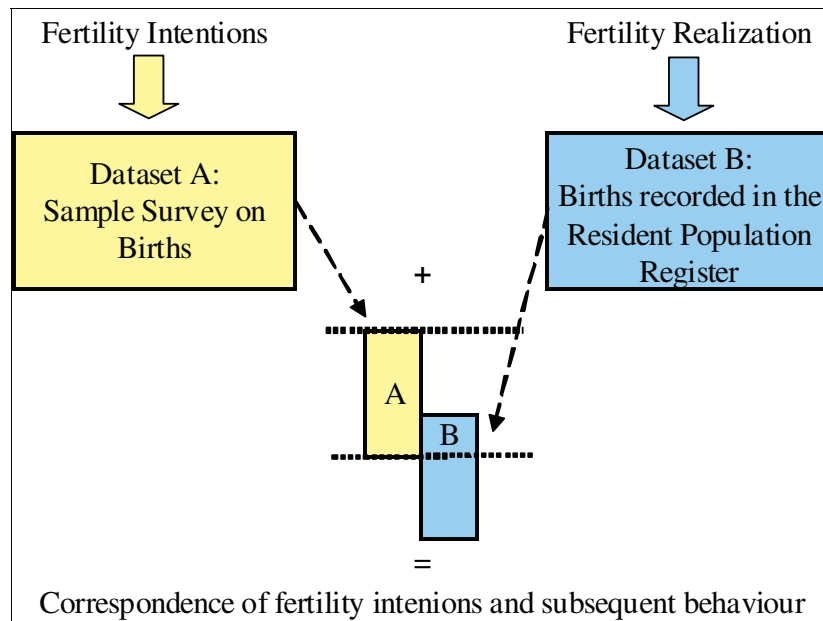
The aim of record linkage is to reconcile in a single dataset information that refers to the same entity but that is stored in different data sources. Data integration allows a better exploitation of existing data since it is by far a thrifty alternative to running a new survey. In Italy panel data on demographic topic is extremely rare because of their expensiveness and the high attrition rate that may occur across waves. At the same time, the Italian National Institute of Statistics (Istat) runs several high quality surveys and numerous national administrative lists are available.

The case study we present here shows how record linkage procedures can help researchers in studying demographic relations that cannot be investigated before because of the lack of proper data.

It is well known that Italy has an extremely low fertility rate when compared with other European countries. At the same time most women still consider two to be the number of children they expect to have during their reproductive careers. The research question we want to answer is: Which is the correspondence between fertility intentions and subsequent behaviour at individual level? This relation has been already investigated in several countries (i.e. Noack and Østby, 2002; Testa and Toulemon 2006) but there are no studies that refer to Italy because of the lack of suitable data.

Nevertheless fertility intentions were recorded by Istat in 2002 with the *Sample Survey on Births* and the effective reproductive behaviour in subsequent years can be obtained from the births recorded in the *Resident Population Register*. The combination of those two data sources made it possible to reconstruct the fertility history of the interviewed women (Figure 1).

Figure 1: *An overall research strategy*



2. Data and Methods

In order to investigate the correspondence between fertility intentions and subsequent behaviour at micro level in Italy, data integration is needed. Therefore, we had to exploit two different data sources. The first one is the *Sample Survey of Births* that was conducted by Istat in 2002. For this survey more than 47,000 women (10% of the total) were interviewed. All of them gave birth 12-18 months before the interview time. The second data source is an administrative one, and consists of the list of all the births recorded in the Resident Population Register during the subsequent four years (from the last months of 2002 to the end of 2006). Note that every year more than 530,000 births had been registered. Please note that approximately 47,000 by 2,698,000 possible records' pairs have been examined. A pair of records is considered as a link if both records agree exactly on a matching key.

The key steps in applying a deterministic record linkage procedure are the follows: (1) Selection of proper matching variables; (2) Standardization of the matching key; (3) Linkage procedure; (4) Data quality assessment.

The starting point is the choice of suitable matching variables. Obviously, a matching variable is a variable that had been recorded in both data files we want to integrate. Moreover, a variable is considered to be a good matching one if it is time-constant, if the number of missing fields or typographical errors are low and if it has a fair distribution between modalities. In the present case study women's name, surname and complete date of birth were selected. A unique identifier was generated by collapsing all those variables into a single field. Before applying the linkage procedure, both data files were standardized. Several activities were performed in order to make the two files' information comparable. Those procedures were partly driven by automatic tools and partly by manual records' inspection. The aim of standardization is both to correct

possible typographical errors and to convert the matching variables into the same format. In this stage also a de-duplication procedure was performed in order to remove those entities that were represented more than once in a single file. Note that as a consequence of de-duplication approximately 1.7% of records were removed from the original administrative list (Table1). In the next step the linkage procedure was applied. We have considered a pair of records as a match only if there was perfect agreement in the created unique identifier. This is only the case if the two records had the same name, surname and exact date of birth. The results of the merging procedure as well as some measures on data reliability are provided in the next section.

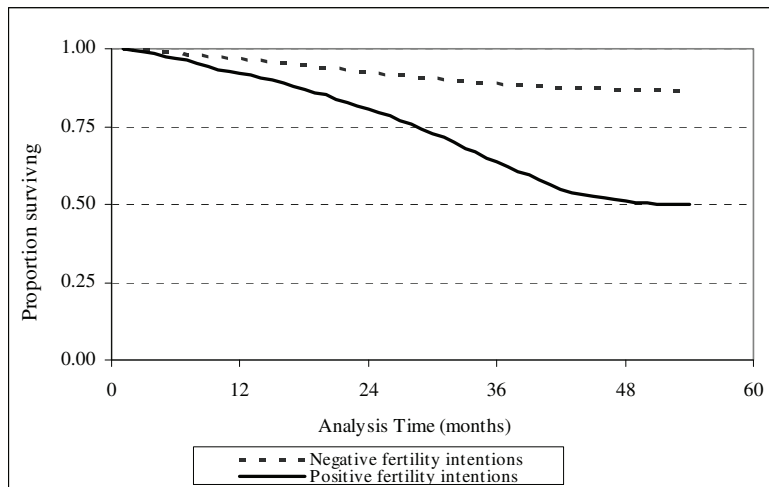
Table 1: *The de-duplication procedure: main results*

Year	Number of births occurred	Duplicated records	Number of births after de-duplication procedure	% of duplicated records
2002	530,926	9,406	521,520	1.8
2003	533,546	8,710	524,836	1.6
2004	549,757	11,794	537,963	2.1
2005	545,713	8,370	537,343	1.5
2006	538,550	8,251	530,299	1.5
Total	2,698,492	46,531	2,651,961	1.7

3. First results and future directions

Due to the data integration the reproductive histories of interviewed women have been reconstructed and it is now possible to evaluate the correspondence between fertility intentions and realization at the individual level. To begin with, In the four years following the interview more than 13,000 live births were delivered by interviewed women. Figure 2 shows the transition to the second child during the analysis time by computing Kaplan-Meier survivor curves. Since we deal with survey data, the analysis that follows are based on weighted data. We found that about 50% of one-child respondents with positive fertility intentions had an additional child within 46 months, whereas this percentage is significantly lower (13%) for women with negative intentions. Moreover, a log-rank test for the equality of the survival functions for fertility intentions indicates that reproductive plans are highly significant ($p < 0.001$) in determining the reproductive behaviours. As expected, at least in the short run, fertility intentions widely overestimate the following reproductive outcome. At the same time there is strong correspondence between negative intentions and subsequent behaviour, which is consistent with previous studies. The same analysis was made with reference to higher parities. As expected we found that the higher the number of already born children is the lower is the predictive power of fertility intentions in the short run. Further analysis (not shown here) focused on the socio-economic and demographic factors that hamper the realization of fertility expectations as well as on the reasons for the lack of correspondence between intended and actual family size. Those results are particularly relevant not only for a better understanding of present Italian fertility patterns but also in order to assess effective family policies.

Figure 2: Transition to the second child by fertility intentions. Kaplan-Meier survivor estimates, weighted data



In order to assess the quality of the linked data, we computed the percentage of births that occurred in the nine months following the interview from women that declare to be pregnant at interview time. The linkage procedure seems to slightly underestimate the number of births, as the matching between pregnant women at interview time and subsequent births occurred in 81% of the records, which is considered to be an adequate matching rate. The reasons for the misclassification are two-fold: On the one hand we have to deal with possible typographical errors in the matching variables, on the other hand we have to consider that 8% of all birth records do not include information on mothers. At the same time, some of the births expected by pregnant women may not have occurred due to miscarriage which is more frequent during the first months after conception (more than one out of four interviewed women stated being pregnant for less than four months). Lastly, women's premature mortality and migration might play a role too, even if there are good reasons to believe that it should not be relevant here. More importantly, women's socio-demographic characteristics show a fair distribution (Chi-squared test) of the unmatched cases by selected mother's characteristics (age group, citizenship, area of residence, and educational level). In order to improve the matching efficacy, we plan to perform a probabilistic record linkage procedure in the future (e.g. Fellegi and Sunter, 1969).

4. References

- Fellegi, I.P. and Sunter A.B. (1969) A Theory for Record Linkage, *Journal of the American Statistical Association*, 64 (328), 1183-1210.
- Noack, T. and Østby, L. (2002) Free to choose – but unable to stick it? Norwegian fertility expectations and subsequent behaviour in the following 20 years, in: *Dynamics of fertility and partnership in Europe – Insights and lessons from comparative research*, E. Klijzing and M. Corijn (Eds), United Nations, 103-116.
- Testa, M.R. and Toulemon, L. (2006) Family Formation in France: Individual Preferences and Subsequent Outcomes, *Vienna Yearbook of Population Research 2006*, Vienna.