rivista di statistica ufficiale

n. 1
2009

Temi trattati

Istat
Istituto nazionale di Statistica

rivista di statistica ufficiale

n. 1
2009

Temi trattati

Istat

Direttore responsabile: Patrizia Cacioli

Coordinatore scientifico: Giulio Barcaroli

per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
c/o Gilda Sonetti
Istat - Via Cesare Balbo, 16 - 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 1/2009

Periodico quadrimestrale
ISSN 1828-1982

Registrazione presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 - Roma

Stampato nel mese di agosto 2010
presso il Centro stampa dell'Istat
Via Tuscolana 1788 - Roma
Copie 350

The deflation of the annual household final consumption expenditure in Italian National Accounts: the chain-linking approach

Carolina Corea[1]

Abstract

The deflation method that Italian National Accounts (NA) has adopted since the 2000 benchmark is based on moving-base deflators, so that deflating NA aggregates now means expressing them at the precedent year prices. The moving base deflators are the first step to calculate the chain linked deflators and then to obtain the annual household final consumption expenditure expressed at a base year prices (2000). Before the 2000 benchmark, the deflation system was based on a fixed-base approach. The use of chain indices is recognised as representing a superior method to deflate NA aggregates, especially in periods of structural change and rapid movements of relative prices. The actual movements of prices (or volumes) from period to period are measured by the moving base indices, that take into account all the modifications occurred between two consecutive periods. Changes in prices (or volumes) between periods that are separated in time are then obtained by cumulating the short-term variations or, in other words, by linking the indices between consecutive periods, to form "chain indices". The main disadvantage of the chain indices is represented by the loss of additivity, as it will be discuss in the next paragraphs. This paper will focus on the particular method set up to express household final consumption expenditure at the precedent year prices.

Keywords: consumption price index, consumption transition matrix, chain indices, precedent year prices

1. The general approach via the product-purpose matrix

The general framework to deflate annual household final consumption expenditure (HFCE) is the product-purpose matrix, which is also used as a reference framework for calculating HFCE at current prices.

The product-purpose matrix, also called the consumption transition matrix (CTM), allows each expenditure heading to be simultaneously classified by sector of output (product) and purpose of consumption. In other words, this matrix makes it possible, for every good or service acquired by household, to determine the sector that produced it and the type of need it satisfies. For instance, there is a cell for the medicines, that are produced by chemical industry (product) and are used for medical treatments (purpose).

The matrix has 101 products and 56 consumption purposes and the number of the full cells is 226.

[1] Istat Ricercatrice, e-mail: corea@istat.it

The classification used for the consumption purposes is the COICOP (Classification of Individual Consumption by Purpose), while the 101 products are compatible with the Nace. Rev. 1.1 classification.

Each intersection (or cell) of the CTM, resulting from the simultaneous classification of the consumption items by product and purpose, represents the minimum level of detail at which the deflation is carried out (made). But the elementary deflators (referred to the cells) can be aggregated to calculate products or purposes deflators.

Thanks to this approach, the consumption deflators computed for the services products can also be used to construct output deflators for that part of the production which is devoted to household final consumption.

The most important source of information to calculate the deflators is the Consumer Price Index (CPI), but in some cases the basic information provided by the CPI is either not enough or not suitable to represent the price increase of the products considered in CTM, which considers a wide range of elementary products. So, there are situations in which price information from sources other than CPI (e.g. for items like rents or fuels and energy) are used to calculate specific ("ad hoc") deflators; in other cases, for example for motor vehicles, specific deflators are computed using the basic information of the CPI elementary indices with a different weighting system.

2. The linkage between different price base (CPI) and between the CPI base and the CTM

The linkage between different price base (CPI) and then between the CPI base and the CTM represents the preliminary work to calculate the deflators.

Since 1999, the Italian CPI has been a moving-base index; it means that, to reflect the evolution of consumers' expenditure pattern, the CPI basket changes every year. The updating of the CPI basket occurs every year, to eliminate those items of the CPI having a low attitude to represent the composition of the household expenditure and to replace them with new products (goods and services). But the updating of the CPI also entails changes in the items weights: the updating of the CPI basket determines not only replacements in the items of expenditure, but also modifications in the weighting system.

To detect the changes occurred in the CPI over the considered period, every year the basket has to be compared with basket of the previous year, to link the different CPI bases.

Besides that, it is necessary to classify each CPI item by product and purpose, to link the CPI information to the CTC, that is the NA framework used to estimate HFCE. In this way, the cells (the elementary items) of the CTM are linked to the CPI base. Even though it is a preliminary work to calculate the deflators, this part of the procedure is very important: not only is it a necessary practice to associate the CPI items to the CTM cells (and make the CPI elementary indices usable for the NA deflation purposes), but it also allows us to asses the CPI attitude to "cover" the CTM: in other words, by linking the CPI to the CTM, we are able to identify (to determine) the CTM cells that are either poorly represented in the CPI basket or not represented at all. In these cases it will be necessary to find other sources of information or to combine the CPI data with other information. From the NA point of view, some items of the HFCE can be not represented in a proper way by the CPI basket. First of all, it is necessary to remark that the CPI indices and the NA deflators used for expressing

HFCE at the p.y.p. can have (and they actually have) different purposes; secondly, there can be a delay between the arrival of a product on the market and the inclusion of the corresponding expenditure heading in the CPI sample. Just to give an example, the price index for the mobile telephones has been available only since 1998, although this good was on the market earlier.

This preliminary linkage between the CPI and the CTM is also interesting to observe the modifications that have occurred through the different CPI bases examined, both in the consumers' preferences (the demand side) and in the assortment of the products available on the market (the supply side).

3. The calculation of the product and the purpose deflators

In this section, a formal presentation of the deflation system is provided. In particular, we will focus the attention:

a) on the calculation of the standard deflators for the CTM cells, as a weighted average of the CPI elementary indices;
b) on the calculation of the product or the purpose deflators as aggregation of the cell deflators.

For present purposes, no distinction is made between "standard" and "ad hoc" deflators: only the standard deflation is considered, in a way that the household consumption expenditure at the precedent years prices are calculated by dividing the current prices values by suitable price indices.

Some examples of specific deflators are given in paragraph 5.

The CTM provides the reference framework for the whole deflation system devised, not only because it represents the lowest level of detail selected for deflating the household consumption expenditure, but also because, once the deflation of the CTM has been carried out, the deflated cells provide the weights to be used in aggregating the cell deflators to obtain the product or the purpose deflators.

Table 1 - Consumption transition matrix at current prices

Product	Purpose of consumption						Total
	1	2	…	j	…	m	
1	${}^tC_{11}$	${}^tC_{12}$	…	${}^tC_{1j}$	…	${}^tC_{1m}$	${}^tC_{1.}$
2	${}^tC_{21}$	${}^tC_{22}$	…	${}^tC_{2j}$	…	${}^tC_{2m}$	${}^tC_{2.}$
…	…	…	…	…	…	…	…
i	${}^tC_{i1}$	${}^tC_{i2}$	…	${}^tC_{ij}$	…	${}^tC_{im}$	${}^tC_{i.}$
…	…	…	…	…	…	…	…
n	${}^tC_{n1}$	${}^tC_{n2}$	…	${}^tC_{nj}$	…	${}^tC_{nm}$	${}^tC_{n.}$
Total	${}^tC_{.1}$	${}^tC_{.2}$	…	${}^tC_{.j}$	…	${}^tC_{.m}$	${}^tC_{..}$

The rows of the matrix are labelled with the products and allow the consumption items to be classified by sector of origin or product, whereas the columns show the breakdown of goods and services consumed by purpose of consumption. A given element ${}^tC_{ij}$ in the matrix thus indicates the value of goods or services produced by sector i and acquired by households in the year t for purpose of consumption j. The row total ${}^tC_{i.}$ represents the total

consumption corresponding to product *i*, disregarding the breakdown by purpose. The total of any given column ${}^{t}C_{.j}$ measures total spending on purpose *j*, irrespective of sector of origin.[2]

The matrix is dimensioned as follows: *i*=1, 2,…30 and *j*=1, 2,…. 41, for *t*=1980, 1981,…. 1992 and *i* =1, 2,…. 101 and *j*=1, 2,…. 56 for *t*=1993, 1994,… 2006. There are of course as many matrices as there are years in the series. A parallel series of the corresponding matrices of deflators (price indices) ${}^{t/t-1}p_{ij}$ is also available, with the same dimensions as the matrices to be deflated. The index ${}^{t/t-1}p_{ij}$, referring to the cell at the intersection of row *i* and column *j*, expresses the change in price between year *t* and the base year *t-1* for the group of products considered in cell *(i,j)*.

Table 2 - Matrix of deflators

Product	Purpose of consumption					
	1	**2**	…	**j**	…	**m**
1	${}^{t/t-1}p_{11}$	${}^{t/t-1}p_{12}$	…	${}^{t/t-1}p_{1j}$	…	${}^{t/t-1}p_{1m}$
2	${}^{t/t-1}p_{21}$	${}^{t/t-1}p_{22}$	…	${}^{t/t-1}p_{2j}$	…	${}^{t/t-1}p_{2m}$
…	…	…	…	…	…	…
i	${}^{t/t-1}p_{i1}$	${}^{t/t-1}p_{i2}$	…	${}^{t/t-1}p_{ij}$	…	${}^{t/t-1}p_{im}$
…	…	…	…	…	…	…
n	${}^{t/t-1}p_{n1}$	${}^{t/t-1}p_{n2}$	…	${}^{t/t-1}p_{nj}$	…	${}^{t/t-1}p_{nm}$

The price index in each cell is obtained by aggregating the elementary CPI indices which refer to that cell. The CPI weights are used to determine the weighting coefficients. Each elementary CPI index relating to a cell contributes to determining the index for that cell, its "weight" being obtained as the ratio of spending on that item of consumption to the total spending for the cell. These spending ratios are fixed in the base year *t-1*.

Each deflator ${}^{t/t-1}p_{ij}$, is the weighted mean of the elementary CPI indices ${}^{t/t-1}I^{(i,j)}{}_{h}$ referred to the cell *(i,j)* of the CTM:

$$ {}^{t/t-1}p_{ij}=\sum_{h=1}^{k}({}^{t/t-1}I^{(i,j)}{}_{h}*w_h), \qquad (1) $$

where ${}^{t/t-1}I^{(i,j)}{}_{h}$ is the elementary CPI index for the product *h* of the CPI classified in the cell *(i,j)* of the CTM.[3] This elementary index express the price variation in the period *(t-1,t)*. The element w_h is the weight of the product *h*, as the share of expenditure of the total consumption of the cell *(i,j)*, as it results from the CPI weights.

Deflation at the level of each cell is carried out indirectly as shown in equation (2):

$$ {}^{t}K_{ij}={}^{t}C_{ij}/{}^{t/t-1}p_{ij} \qquad (2) $$

[2] The consumption transition matrix allows a distinction between domestic production flows and import flows, giving rise to two further sub-matrices. This possibility is disregarded for present purposes.

[3] We assume *k* as the number of items (products) if the CPI basket representing the cell *(i,j)* of the CTM.

where ${}^{t}K_{ij}$ is consumption at the precedent year prices for the year *t* with reference to cell *(i,j)*. As *i* and *j* change, ${}^{t}K_{ij}$ generates the transition matrix at the precedent year prices, for the year *t*.

Table 3 - Consumption transition matrix at the precedent year prices

Product	Purpose of consumption						Total
	1	2	…	j	…	m	
1	${}^{t}K_{11}$	${}^{t}K_{12}$	…	${}^{t}K_{1j}$	…	${}^{t}K_{1m}$	${}^{t}K_{1.}$
2	${}^{t}K_{21}$	${}^{t}K_{22}$	…	${}^{t}K_{2j}$	…	${}^{t}K_{2m}$	${}^{t}K_{2.}$
…	…	…	…	…	…	…	…
i	${}^{t}K_{i1}$	${}^{t}K_{i2}$	…	${}^{t}K_{ij}$	…	${}^{t}K_{im}$	${}^{t}K_{i.}$
…	…	…	…	…	…	…	…
n	${}^{t}K_{n1}$	${}^{t}K_{n2}$	…	${}^{t}K_{nj}$	…	${}^{t}K_{nm}$	${}^{t}K_{n.}$
Total	${}^{t}K_{.1}$	${}^{t}K_{.2}$	…	${}^{t}K_{.j}$	…	${}^{t}K_{.m}$	${}^{t}K_{..}$

The implicit product deflators are obtained as the ratio of current consumption by product to the corresponding values at the precedent year prices [3]:

$$ {}^{t/t-1}p_{i.} = {}^{t}C_{i.} / {}^{t}K_{i.} \qquad (3) $$

The implicit purpose of consumption deflators are obtained in the same way, dividing current consumption by purpose by the corresponding precedent year prices consumption [4]:

$$ {}^{t/t-1}p_{.j} = {}^{t}C_{.j} / {}^{t}K_{.j} \qquad (4) $$

As regards the system of weighting, the product deflator shown in [3] may in fact also be expressed as linear combination (weighted average) of the various cell deflators falling within the product *i*:

$$ {}^{t/t-1}p_{i} = ({}^{t}K_{i1} / {}^{t}K_{i.})\ {}^{t/t-1}p_{i1} + ({}^{t}K_{i2} / {}^{t}K_{i.})\ {}^{t/t-1}p_{i2} + \ldots\ldots\ldots + ({}^{t}K_{im} / {}^{t}K_{i.})\ {}^{t/t-1}p_{im} \qquad (5) $$

A weighted average can also be used to express the purpose deflator ${}^{t/t-1}p_{.j}$ as linear combination of the cell deflators: in this case, the cell deflators considered by column, i.e. referred to the purpose *j*.

The deflators considered in formulas [3], [4] and [5] (i.e. ${}^{t/t-1}p_{ij}$ for the cell *(i,j)*, ${}^{t/t-1}p_{i.}$ for the product *i*, ${}^{t/t-1}p_{.j}$ for the purpose *j*) measure the increase (or decrease) in prices between the year *t-1* and the year *t*.

For product *i*, the coefficients (weights) of the linear combination are obtained as the ratio of expenditure at precedent prices year in each cell within the product *i* to the corresponding row total (total consumption at the precedent year prices of the product *i*).

Considering the columns rather than the rows, the same applies to the purpose of consumption deflators.

4. From the moving base deflators to the chain linked indices: the HFCE at a base year prices

The moving base deflators represent the first step to calculate the chain linked deflators and then to obtain the HFCE expressed at a reference year prices. Let's fix the reference year at *0*.[4]

For each cell *(i,j)* of the CTM, the fixed base deflators have to be computed, by linking the moving base deflators referred to all sub-periods included in *(0,t)*:

$$^{t/0}p_{ij} = {}^{t/t-1}p_{ij} \, {}^{t-1/t-2}p_{ij} \ldots\ldots {}^{2/1}p_{ij} \, {}^{1/0}p_{ij} \qquad (6)$$

The deflator $^{t/0}p_{ij}$, referred to the group of products *(i,j)*, expresses the average price variation of this group of products between the current year *t* and the fixed reference year *0*.

So, a matrix of fixed base price indices is obtained by applying formula [6] for each cell of the CTM.

For each group of products *(i,j)*, the consumption for year *t* at the prices of a fixed years *0* is expressed as follows:

$$^{t}{}_{(0)}K_{ij} = {}^{t}C_{ij} / {}^{t/0}p_{ij} \qquad (7)$$

It is worth noticing that, differently from the fixed-base methodology, the deflator $^{t/0}p_{ij}$ is obtained by comparing the price level at time *t* with the price level at time *0*, but looking at what happens between the beginning and the end of the considered time period *(0,t)*.

The main advantage of the chain-linking method is that changes in prices between periods that are separated in time are obtained by cumulating the short-term variations, i.e. by linking the indices between consecutive periods. In other words, tanks to the annual updating of the weight structure, the deflator $^{t/0}p_{ij}$ take account of the modifications occurred in all the sub-periods of *(0,t)*. So, when an annual updating of the weights is adopted, the hypothesis of fixed weights has to be maintained only from one year to the next and not for a five-year period. In the fixed-base methodology, the price variation between *t* and *0* does not incorporate the annual updating of the weights, that are fixed to the base year for five (or more) years.

The loss of additivity represents the main disadvantage of the chain-linking system. In fact, for the HFCE expressed at the reference year *0* prices, it results that the sum of the consumption by rows does not equal the consumption by product (total of row *i*), likewise the sum of the consumption by columns does not equal the total consumption by purpose:

$$^{t}{}_{(0)}K_{i.} \neq \sum_{j=1}^{m} {}^{t}{}_{(0)}K_{ij} \qquad (8)$$

$$^{t}{}_{(0)}K_{.j} \neq \sum_{i=1}^{n} {}^{t}{}_{(0)}K_{ij} \qquad (9)$$

And the sum of the consumption by purpose and by product does not equal the total consumption:

$$^{t}{}_{(0)}K_{..} \neq \sum_{i=1}^{n} \sum_{j=1}^{m} {}^{t}{}_{(0)}K_{ij} \qquad (10)$$

[4] In the formal presentation, the reference year is set at *0*. In deflating NA aggregates, the base year is 2000.

Nevertheless, the additivity is ensured considering the estimates at the precedent year prices:

$${}^{t}K_{i.} = \sum_{j=1}^{m} {}^{t}K_{ij} \quad (11)$$

And

$${}^{t}K_{.j} = \sum_{i=1}^{n} {}^{t}K_{ij} \quad (12)$$

Finally, for the total HFCE:

$${}^{t}K_{..} \neq \sum_{i=1}^{n} \sum_{j=1}^{m} {}^{t}K_{ij} \quad (13)$$

5. The calculation of the "non standard" deflators: some examples

For the products (cells of the CTM) represented by the CPI sample, deflators are calculated by aggregating the CPI elementary price indices referred to the cell (weighted average).

The term "non standard" deflators is used to cover all the cases in which there are either no price indices or few price indices available for a CTM cell and it is thus not possible to calculate an average deflator representative of the price changes of that cell (only) on the basis of the CPI information.[5] Basically, the "non standard deflation" has proved necessary in three cases: to cover a delay in the adjustment of the CPI base to reflect changes in spending patterns, to take into account the loss of certain products that were present in the CPI basket only up to a certain year, to address the existence of the CTM cells whose products are completely excluded from the CPI sample or not adequately represented for the NA purposes.

These cases have been tackled by:

1. calculating deflators not derived from the CPI ("ad hoc" deflators);
2. using the deflator of a single item of the CPI (similar to the products included in the cell);
3. using the deflator of another similar cell;
4. using the deflator of the purpose concerned;
5. using the average national index (CPI).

Case 1

In some cases, such as **motor vehicles** and **hotel spending**, the non-standard procedure consists in adopting a weighting system specially devised by the national accountants and differing from that provided by the CPI data, in order to estimate a different price index according to the various qualitative segments of a particular category of goods. For

[5] For more details, ISTAT, "Inventario sulle fonti e i metodi di calcolo per le stime a prezzi costanti", par. 3.2.1 in *Metodi e norme*, n.19, (2004)

instance, the weights to aggregate the CPI elementary price indices referred to "Hotels and other accommodations" services are given by the overnight stays ("Tourism statistics"), broken down by the kind of accommodation: hotels (distinguishing by different categories: high, medium and low) and other accommodation services. Since the CPI sample provides elementary indices for each of these types of services, it is possible to aggregate them by using the appropriate weights (the corresponding number of overnight stays).

A similar approach is considered for motor vehicles. The registrations data are obtained from the Motor Vehicles Computer Centre and compiled at segment level by the trade association ANFIA.[6] Before the 2000 benchmark, a representative sample was chosen for each segment and producing industry, both Italian and foreign: for each model, a medium-to-low engine rating and standard of equipment were also chosen to avoid overestimation of spending, given the extremely wide range currently available. The new estimates at the precedent year prices (benchmark 2000) are obtained using all the basic available information and not only a sample of them, as it was done according to the method followed for benchmark 1992.

The case of rents and **fuels and energy** is an example of the direct deflation method, that is to say to construct directly the estimates at the precedent year prices, adopting a quantity x price approach ($Q_t * P_{t-1}$).

In accordance with the EU recommendations, **actual and imputed rents** are estimated at current prices for 32 strata of dwellings (each stratum or category of dwelling is determined on the basis of the characteristics that affect rent levels). The 2001 Population and Housing Census provides the dwelling stock data (quantity), whereas the sources of information used to obtain the price data for housing category is the Household budget survey.

To obtain a valuation at the precedent year prices a direct deflation method is also used in this case. The quantity data for each current year (t) are multiplied for the prices referred to the previous year (t-1): $Q_t * P_{t-1}$.

For each stratum, a quantity index to update the stock data referred to the base year (2001) is obtained on the basis of the annual information concerning the construction of new dwellings and the demolition of the old ones. The implicit deflator is then calculated by dividing the estimates at current prices by the corresponding estimates at the previous year prices.

As regards fuels and energy, the same method used for rents it is considered. The implicit prices are calculated by dividing the household consumption at current prices by the corresponding quantities consumed, being both the information on quantity and price derived from the national energy budget. The prices so obtained (or, more precisely, the average unit values)[7] are applied to the quantity series, to determine consumption at the precedent year prices ($Q_t * P_{t-1}$).

The deflation of the **financial services** is done separately for the Fisim (Financial services indirectly measured) part and the non Fisim part. The Fisim at current prices are deducted from the whole amount of the expenditure on financial services; the remaining part is deflated with the CPI index.

The consumption of financial services (non Fisim) at the precedent year prices are then added to the Fisim at the precedent year prices

[6] National Association Motor Vehicles Industries – Turin.

[7] For the information provided by the national energy balance the level of detail is compatible with the CTM.

Case 2

An elementary CPI price index (a single item index) is used to deflate the expenditure of a CTM cell (which in general represents a small group of products). Implicitly, we make the assumption that prices of similar product tend to exhibit almost the same trend.

Some examples are presented in the following table.

Table 4 - Examples of some CTM cells covered by the elementary CPI index of a similar product

Coordinates of the cell to be covered	Description of the cell to be covered	CPI code	CPI single item description
A.2-2	"Animals and poultry purchased live for consumption as food"	1263	"Fresh poultry"
L.1-88	"Repair of personal computers"	8504	"Repair of televisions"

Cases 3 and 4

Again, in cases 3 and 4, the underlying hypothesis is that the prices of similar products will develop in similar ways.

Below some examples are provided.

Table 5 - Examples of some CTM cells covered by the CPI index of a similar CTM cell

Coordinates of the cell to be covered	Description of the cell to be covered	Coordinates of a similar cell	Description of a similar cell
G.2-100	"Outpatient thermal bath or sea-water treatments"	G.3-94	"Hospital services"
H.2-64	"Repair of motor cycles"	H.2-65	"Repair of motor cars"
I.2-70	"Repair of telephone and telefax equipment"	F.3-70	"Repair of electric household appliances "
P.3-101	"Home help provided by people who look after children, elederly or disabled persons"	F.8-101	"Domestic help hourly paid"

Table 6 - Examples of some CTM cells covered by the deflator of the consumption purpose

Coordinates of the cell to be covered	Description of the cell to be covered	Coordinates of the consumption purpose	Description of the purpose consumption
H.2-29	"Chemical products for the cleaning and the maintenance of the car"	H.2	"Operation of personal transport equipment"
L.4-58	"Artificial flowers"	L.4	"Other recreational items and equipments"
O.2-39	"Metal watch straps"	O.2	"Personal effects N.E.C."

References

Boskin M.J., Duelberg E.R., Griliches Z., Gordon J., Jorgenson D. (1996), "Toward a More Accurate Measure of the Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index".

Corea C. (2000), "La deflazione della matrice di transizione dei consumi per il periodo 1982-98: aspetti metodologici, in Istat, "Le nuove stime dei consumi finali delle famiglie", *Metodi e Norme*, n.7, Istat, Roma.

Fabiani S., Gattulli A., Sabbatini R., Veronese G. (2005), "Consumer price setting in Italy", *Temi di discussione*, n. 556, Banca d'Italia.

Eurostat (1996), *European System of Accounts ESA* 1995, Luxembourg.

Eurostat (2001), *Handbook on price and volume measures in National Accounts*, Office for Official Publications of the European Communities, Luxembourg.

Corea C. (2004), "I consumi delle famiglie", in Istat, "Inventario sulle fonti e i metodi di calcolo per le stime a prezzi costanti", pp. 57-61, *Metodi e norme*, n.19, (2004), Istat, Roma.

Landfeld J.S., Parker R.P. (1997), "BEA'S chain indexes, time series and measures of long-term economic growth", *Survey of current Business*, May 1997.

Maresca S. (1997), "I conti nazionali calcolati mediante indici a catena: alcuni primi risultati per il caso italiano", in Atti del Convegno Cide-Istat "La misurazione delle variabili economiche e i suoi riflessi sulla modellistica econometrica", *Annali di Statistica*, *Series X*–vol.15, 1998, Sistan-Istat.

Maresca S. (2000), "L'indice a catena per le valutazioni a prezzi costanti del PIL: l'esperienza italiana", paper presented at the Meeting of National Accounts Experts, OECD, Paris, 26-29 september 2000.

Maresca S. (2006), "Le novità delle valutazioni ai prezzi dell'anno precedente: aspetti teorici e pratici", paper presented at the Meeting "La revisione generale dei conti nazionali del 2005", Rome, 21-22 June 2006.

Moulton B.R. (1993), "Basic components of the CPI: estimation of price changes", *Monthly Labour Review*, volume 116, number 12.

Statistic New Zeland (1998), "Chain volume measures in the New Zealand National Accounts", Statistics New Zealand.

Triplett J.E. (1992), "Economic theory and BEA's alternative quantity and price indexes", *Survey of current Business*.

Tuke A. (2002), "Analysing the effect of the annual chain-linking on output measure of GDP", *Economic Trends*, n. 581.

Metodologie per la stima degli affitti figurativi ed impatto sulla distribuzione del reddito[1]

Claudio Ceccarelli[2], Andrea Cutillo[3], Davide Di Laurea[4]

Sommario

Questo lavoro presenta un'analisi sull'affitto figurativo condotta sui dati dell'indagine EUSILC; questa posta figurativa, stimata per chi non vive in abitazioni in affitto a prezzi di mercato, va considerata per ottenere una nozione di reddito disponibile più accurata nelle analisi su distribuzione del reddito, disuguaglianza e povertà. Vengono messe a confronto diverse metodologie, in particolare le stime derivanti dall'autovalutazione dei rispondenti con quelle derivanti da tecniche di regressione econometrica secondo la teoria della domanda edonica. Tre sono le principali questioni affrontate: quale sia la metodologia più corretta; se lo sconto temporale che si osserva tra gli affittuari debba essere considerato anche nel calcolo dell'affitto figurativo; se esiste un processo di selezione sulla condizione abitativa che può portare a stime distorte degli affitti figurativi. Vengono infine presentati alcuni risultati relativi all'incidenza del rischio di povertà e alla concentrazione dei redditi una volta che si considera questa posta come parte del reddito disponibile.

Abstract

In the relevant literature it is well established that imputed rents – a non-monetary income source for all households whose living accommodation is not rented at a market price - has to be accounted for in order to provide a more accurate and comparable measure of the total disposable household income. A more controversial issue is about the choice among distinct estimation approaches and how to operationalise them.

This paper aims to analyse the impact of different estimation methods for imputed rents on the distribution of disposable income as well as on inequality measures, using data from EUSILC. Subjective rents – self-assessed values by the respondents - are compared with predicted values coming from econometric techniques, i.e. an hedonic regression model, being both methods coherent with the 'rental equivalence value' approach.

Three main issues are assessed: i) pros and cons of each method; ii) if the tenure discount (the length of tenure), which turns to be a relevant factor in determining the observed values of the market rents, has to be taken into explicit account in predicting imputed rents; iii) how to prevent biases on model predictions eventually arising if a selection process between tenants at a market price and households living with different accommodation occurs.

Parole chiave: Affitti figurativi, *Selection bias*, Distribuzione del reddito

---

[1] In questo lavoro sono presentati i principali risultati delle analisi svolte nell'ambito del Servizio Condizioni Economiche delle Famiglie dell'ISTAT per il calcolo degli affitti figurativi nell'ambito del progetto EUSILC, European Statistics on Income and Living Conditions. I risultati sono stati presentati in forma preliminare nel Seminario sulle Strategie Metodologiche dell'indagine campionaria sui redditi e le condizioni di vita delle famiglie, svolto in ISTAT il 5 marzo 2006. Il nostro sentito ringraziamento va ai discussant del seminario, A. Lemmi, A. Russo e N. Torelli, e all'anonimo referee della rivista secondo le indicazioni dei quali il lavoro è stato rivisto. Si ringraziano inoltre A.B. Atkinson, A. Brandolini e G. D'Alessio per gli utili consigli emersi durante la preparazione del seminario. Si ringrazia anche S. Mantegazza per i commenti ad una prima stesura. Rimane agli autori la piena responsabilità di quanto scritto.

[2] Istat, Condizioni Economiche delle Famiglie, e-mail: clceccar@istat.it.

[3] Istat, Condizioni Economiche delle Famiglie, e-mail: cutillo@istat.it.

[4] Istat, Condizioni Economiche delle Famiglie, e-mail: dilaurea@istat.it.

1. Introduzione

Secondo la letteratura, l'affitto figurativo, è il costo che deve essere imputato a coloro che occupano l'abitazione di cui sono proprietari ed equivale a quello che tali individui sosterrebbero affittando ai prezzi vigenti sul mercato immobiliare un'unità abitativa equivalente, in termini di caratteristiche, a quella in cui effettivamente vivono. In alcuni casi specifici, quando si vogliono analizzare problematiche particolari, il concetto può essere esteso anche agli inquilini con affitti agevolati, inferiori ai prezzi di mercato. Questo è il caso delle analisi sulla povertà, sulla disuguaglianza e sulla distribuzione dei redditi, in cui la stima degli affitti figurativi nasce dalla necessità di introdurre una nozione di reddito disponibile più completa ed accurata possibile.

Detto con il manuale di Canberra, che stabilisce a livello internazionale le linee guida per la definizione di reddito e ricchezza: "*In theory, imputed rent is the difference between the cost of renting one's living arrangements (in a competitive market) minus the cost actually incurred in owning the home (or renting it at a below market price)*".[5]

Si prenda in esame, a titolo esemplificativo, il caso di due famiglie identiche dal punto di vista delle entrate finanziarie e delle caratteristiche socio-demografiche ma che vivono l'una in una casa in affitto, l'altra in casa di proprietà. O ancora, il caso di due famiglie identiche che vivono entrambe in casa in affitto, con l'una che ha avuto accesso ad un affitto inferiore ai prezzi di mercato a differenza dell'altra. In questi casi portati ad esempio, la mancata considerazione della posta degli affitti figurativi porterebbe a considerarle uguali rispetto alla distribuzione dei redditi. Al fine di rendere comparabili le stime delle risorse familiari per i nuclei che stanno in affitto o in casa di proprietà è quindi necessario prendere in considerazione il flusso di servizi abitativi che si ottiene in virtù del proprio titolo di godimento.[6]

Già nel 1977 le Nazioni Unite raccomandavano la loro inclusione nelle statistiche sul reddito familiare; così pure l'ILO[7] ed il manuale di Canberra. Nel SEC 95[8] si fa riferimento a questa posta figurativa come attività di produzione per uso proprio da parte di coloro che occupano l'abitazione di cui sono proprietari; specularmente, la posta rientra nella spesa per consumi finali delle famiglie. Questo equivale ad ipotizzare che la famiglia stia, al contempo, producendo e consumando un flusso di servizi abitativi senza passare dal mercato. Seguendo questo orientamento, Eurostat ne raccomanda l'inserimento nei principali aggregati stimati sulla base delle rilevazioni armonizzate tra i diversi paesi membri sia relativamente al progetto HBS, Household Budget Survey, che al progetto EUSILC, European Statistics on Income and Living Conditions.

È il caso di evidenziare che in questo lavoro viene considerato l'affitto figurativo per le sole residenze principali: a nostro avviso, la stima di questa posta non monetaria è da collegare in maniera esclusiva ai flussi di reddito e di consumo connessi al soddisfacimento del bisogno primario dell'abitazione. È vero che le abitazioni non permanentemente occupate (case di vacanza, abitazioni tenute a disposizione o anche quelle in attesa di vendita o affitto) rappresentano un capitale che, impiegato in altro modo (ad esempio sui mercati finanziari), potrebbe essere remunerato,[9] ma queste fatto è rilevante per la tematica

[5] Cfr. Canberra Group (2001), pag. 63.

[6] È da citare un'ipotesi alternativa che consiste nel calcolare gli indicatori distributivi in base al reddito "spendibile", al netto sia degli affitti realmente pagati dagli inquilini sia degli affitti figurativi delle case abitate per tutti coloro che non pagano un affitto a prezzi di mercato.

[7] Cfr. ILO (2003).

[8] Sistema europeo dei conti, 1995.

[9] Si tratta dell'ipotesi adottata correntemente dal fisco, che – tranne rari casi – attribuisce un reddito imputato alle abitazioni a disposizione.

della ricchezza piuttosto che del reddito e della sua distribuzione. Questa impostazione è peraltro confermata da Eurostat: "*The dwellings that should be included for imputation of rental in ESA, are those used entirely or primarily as residences. ... For secondary dwellings (week-end and holiday residences) imputation is not required*".[10]

2. Metodi di stima

2.1 User cost of capital

Le metodologie più largamente utilizzate sono quelle del *user cost of capital* e del *rental equivalence*.

Nel primo caso si tratta di stimare il tasso di rendimento implicito del capitale investito considerando il tasso di apprezzamento futuro sul mercato immobiliare oppure, sotto opportune ipotesi, il tasso di ritorno di un investimento sul mercato finanziario, con l'idea che il ritorno sul mercato immobiliare sia lo stesso di un investimento sul mercato finanziario. Il metodo stima in pratica il ritorno di un investimento pari al valore dell'abitazione posseduta, e pone non pochi problemi metodologici: anzitutto la determinazione del valore dell'abitazione; secondariamente, la stima del tasso di apprezzamento futuro del bene, ossia la predizione dell'andamento futuro del mercato immobiliare o del mercato finanziario; infine, l'estrema volatilità che il valore ottenuto mostra nel tempo, anche in funzione del tasso di interesse che viene usato per l'operazione di sconto. Peraltro, la stessa Eurostat (2005) sconsiglia l'utilizzo di questo metodo:

"Although this method [user cost] may solve the problem which arises when the sub-sample of privately rented dwellings is too small, it has some important drawbacks for its implementation in the context of EUSILC and HBS. The most important one is the macro-economic approach of some calculations which prevents from getting accurate imputations at micro-level".[11]

2.2 La stima del rental equivalence value

Seguendo questo approccio, per stimare l'affitto figurativo si deve valutare il flusso di servizi di un'unità abitativa il cui godimento non comporta transazioni di mercato, alla stessa stregua dei flussi corrispondenti osservabili sul mercato degli affitti.

Questo può avvenire tramite tecniche differenti: *i)* autovalutazione dei rispondenti (il cosiddetto affitto soggettivo); *ii)* metodo della stratificazione; *iii)* stima econometrica, in particolare tramite il metodo della regressione edonica, applicando o meno tecniche di correzione per la distorsione della selezione del campione (procedura di Heckman per la correzione del *sample selection bias*).

2.2.1 L'autovalutazione

Il metodo dell'autovalutazione (*self-assessment method*) consiste nella stima, da parte dei rispondenti ad indagine campionaria o censuaria, del potenziale valore sul mercato degli affitti della casa nella quale vivono con titolo di godimento diverso dall'affitto a prezzi di mercato (proprietà, usufrutto, uso gratuito o, infine, affitto con canone inferiore ai prezzi di mercato).

[10] Cfr Eurostat (2006), pag. 20.
[11] Cfr. Eurostat (2005), pag.10.

L'affitto figurativo stimato tramite autovalutazione comporta problemi proprio a causa della natura soggettiva della stima. I rispondenti potrebbero non conoscere in maniera adeguata le reali condizioni del mercato, la qual cosa può portare tanto ad una sottostima quanto ad una sovrastima dei "veri" affitti potenzialmente da pagare (ossia del "vero" flusso di servizi abitativi goduto). Arevalo (2001) riporta, utilizzando i dati dell'HBS spagnola, una sottostima dell'autovalutazione rispetto alle stime da modello econometrico del 13% nel 1980 ed una sovrastima del 27% nel 1990; tali variazioni vengono spiegate proprio con la diversa conoscenza tra i rispondenti del mercato immobiliare tra le due date prese in considerazione.

2.2.2 La stratificazione

Con il metodo della stratificazione, dal punto di vista operativo, si tratterebbe di: *i)* definire gruppi di famiglie omogenei rispetto alle caratteristiche della casa e della zona di abitazione; *ii)* calcolare l'affitto medio/mediano per il sottoinsieme di ogni strato che vive in affitto a prezzi di mercato; *iii)* imputare questo valore a tutti coloro che, nello stesso gruppo, non pagano un affitto a prezzi di mercato.

Questo metodo viene utilizzato per la stima degli affitti figurativi elaborata dalla Contabilità Nazionale dell'Istat[12] a partire dal 1996, conformemente alla Decisione della Commissione delle Comunità Europee[13] che prevede la seguente articolazione:

- per calcolare la produzione dei servizi di abitazione si deve applicare il metodo della stratificazione a partire dagli affitti effettivi, procedendo per estrapolazione diretta o per regressione econometrica. Per quanto riguarda le abitazioni occupate dal proprietario, ciò implica un'imputazione basata sugli affitti effettivi di alloggi simili dati in affitto;
- per quanto riguarda la stratificazione si deve tenere conto di caratteristiche importanti dell'abitazione. In particolare, si possono considerare le caratteristiche dell'abitazione o dell'edificio o le caratteristiche ambientali e la localizzazione o, ancora, fattori socioeconomici;
- l'analisi tabellare o tecniche statistiche sono alla base della stratificazione. E` opportuno, infatti, far riferimento alle dimensioni e alla localizzazione dell'immobile e ad almeno un'altra caratteristica importante dell'abitazione. Le tipologie da individuare devono essere di almeno 30 unità, tenendo conto di almeno tre classi per quanto riguarda la dimensione e due per quanto attiene alla localizzazione;
- per quanto riguarda gli affitti effettivi, si deve far riferimento al diritto di uso di abitazioni non ammobiliate. Per determinare il valore dell'affitto si deve far riferimento al settore privato;
- per calcolare gli affitti, riferiti all'anno base, si deve tenere conto del più recente Censimento degli immobili per garantire l'esaustività delle stime. Un valore nullo dovrà essere applicato alle abitazioni vuote e disponibili alla vendita o all'affitto.

Il metodo della stratificazione applicato ad un'indagine campionaria, e non all'intero universo di riferimento, può, tuttavia, avere delle serie controindicazioni in situazioni caratterizzate da un'alta incidenza di famiglie che vivono in case di proprietà o non vivono in case in affitto a prezzi di mercato, come è nel caso del mercato italiano. Difatti, il numero di caratteristiche rilevanti per la definizione degli strati porterebbe verosimilmente

[12] Istat (1997), La revisione della contabilità nazionale annuale, Metodi e norme - n.1, Roma capitolo II: "La revisione dei fitti".

[13] Decisione della Commissione delle Comunità Europee sui Principi per la valutazione dei servizi di abitazione. (95/309/CE, EURATOM), Gazzetta Ufficiale delle Comunità Europee N° L 186/59 del 5.8.95.

a frequenze di cella molto basse e, sicuramente, inferiori a 30,[14] oppure, alternativamente, le celle dovrebbero essere formate in maniera tale da risultare poco informative delle caratteristiche della casa.

2.2.3 Regressione edonica

La regressione edonica (Rosen, 1974) si basa sull'assunzione - standard nella teoria micro-economica neoclassica - che il prezzo di un bene indivisibile, in condizione di mercati perfettamente concorrenziali, è dato dall'intersezione tra la domanda e l'offerta delle caratteristiche del bene. La domanda e l'offerta delle caratteristiche di un'abitazione non sono direttamente osservabili, ma si può considerare l'affitto come funzione di esse. In altri termini, il prezzo che emerge dal mercato degli affitti ha il contenuto informativo necessario per un'equa valutazione delle caratteristiche rilevanti dell'unità abitativa.

In pratica, il bene in questione è rappresentato dalle sue caratteristiche per le quali si assume che ognuna abbia un proprio mercato e che esista una sorta di effetto sostituzione tra una caratteristica e l'altra. Il modello può essere stimato tramite una regressione dei minimi quadrati ordinari (*Ordinary Least Squares*, OLS) dove la variabile dipendente è il logaritmo dell'affitto di mercato e i parametri stimati per le caratteristiche del bene, che rappresentano l'elasticità al prezzo di ciascuno di essi, costituiscono il punto di incontro della domanda e offerta (non osservabili) di ciascuna caratteristica.

È importante notare che in questa maniera il prezzo dell'affitto è determinato solamente dalle caratteristiche del bene, e non dalle caratteristiche della famiglia in affitto: una famiglia benestante pagherà mediamente di più perché affitta una casa dotata di migliori caratteristiche, e non perché ha maggiori disponibilità. Inoltre, questa tecnica trascura l'uso dei prezzi come meccanismo di selezione della domanda, anche a parità di caratteristiche e di zona di ubicazione. Ossia, si esclude la possibilità che un locatario possa chiedere un prezzo superiore ai prezzi di mercato vigenti per una casa con determinate caratteristiche, anche a rischio di tenere la casa sfitta per qualche tempo, per discriminare la domanda potenziale, aumentando la probabilità di avere degli affittuari che soddisfano le sue aspettative.

C'è da constatare che l'utilizzo di questa metodologia sarebbe alquanto problematico nel caso in cui il mercato immobiliare risultasse segmentato e la distribuzione delle caratteristiche delle case in affitto fosse solo parzialmente sovrapponibile a quella delle case di proprietà; in questo caso le stime ottenute sulla base degli affitti realmente pagati potrebbe introdurre una distorsione quando usate per "predire" gli affitti figurativi.

In ogni caso, tanto il metodo della stratificazione quanto la regressione edonica sono *information intensive*, ossia molto esigenti in termini di caratteristiche da rilevare.

2.2.4 La procedura di Heckman

La procedura a due stadi di Heckman (1979) permette di incorporare nel processo di stima tramite OLS la possibile distorsione da *sample selection*: l'equazione degli affitti è infatti stimata solamente su quanti sono in affitto ai prezzi di mercato, e questi potrebbero differire in maniera sistematica dai proprietari di immobili. La procedura consiste nella stima di un'equazione che stabilisce le determinanti dell'essere in affitto a prezzi di mercato o meno calcolando allo stesso tempo un parametro (λ), pari all'inverso del rapporto di Mill,

[14] La soglia di 30 unità è anche quella minima suggerita da Eurostat per il rilascio di informazioni campionarie dalle indagini di EUSILC.

per ogni famiglia da inserire successivamente nell'equazione degli affitti come nuova variabile esplicativa; questo parametro cattura l'effetto di caratteristiche non osservabili che influenzano sia la decisione di stare in affitto sia il prezzo che si è disposti a pagare per un'abitazione con le caratteristiche di quella che si occupa.

2.2.5 Il problema del tenure discount

Un ulteriore aspetto critico, e trasversale rispetto alla scelta della tecnica più appropriata, è introdotto dall'esistenza del *tenure discount*. Succede difatti che, anche a parità di caratteristiche abitative, i prezzi degli affitti reali possono essere anche significativamente diversi a seconda dell'anno in cui i contratti di locazione sono stati stipulati. In particolare, più lungo è il periodo di locazione di un'abitazione minore è il prezzo di affitto, a parità di altre condizioni. E ciò perché:

- i locatari scambiano il minore introito con la sicurezza del pagamento da parte di chi è affittuario da lungo tempo, come fosse una ricompensa implicita del rapporto fiduciario;
- i lavori di ristrutturazione vengono spesso fatti prima di affittare la casa, quindi al momento di ingresso la casa è in buone condizioni, ma queste peggiorano nel tempo.

La rilevanza del *tenure discount* è accentuata in un mercato come quello italiano che per lungo tempo è stato regolamento. Malgrado la liberalizzazione delle tipologie contrattuali e la definitiva abrogazione, nel 1995, dell'equo canone per le nuove locazioni, risulta ancora una percentuale non trascurabile di famiglie la cui locazione è governata da questo tipo di contratto.

Il *tenure discount* è considerato da molte ricerche empiriche una caratteristica rilevante che influenza l'affitto pagato; tra gli altri, Miron (1990) sostiene che il *tenure discount* deve essere tenuto in considerazione nella stima degli affitti figurativi, poiché è una precisa condizione dell'abitazione che potrebbe modificare il prezzo del servizio abitativo ricevuto dalla famiglia. Se il *tenure discount* deve essere considerato come una determinante dell'affitto figurativo, questo costituisce un secondo aspetto per cui è preferibile non utilizzare il *self-assessment method*, che difficilmente potrebbe tenere conto di questo aspetto.

Nel manuale di Canberra questa condizione non è esplicitamente dichiarata. Il manuale recita infatti:

"*Finally, the service-yielding asset can be bought and sold on real estate markets. Hence, the value to the consumer is close to the market value of the service flow, since the owner could presumably sell the housing unit and rent it back from the new owner were it more profitable to do so. [...] imputed rent can therefore be valued to consumers at its market price*".[15]

Come si vede, non è esplicitamente dichiarato se il prezzo di mercato debba incorporare lo sconto temporale o se come prezzo di mercato si intende quello che si andrebbe ad affrontare se si entrasse in quel preciso momento sul mercato degli affitti.

Una considerazione aggiuntiva riguarda il fatto che includendo il *tenure discount* tra le determinanti degli affitti figurativi viene alterata la comparabilità dei livelli di benessere: se due famiglie vivono in abitazioni identiche, quella che è proprietaria da più tempo risulterebbe meno ricca in virtù dell'effetto del maggior periodo di occupazione, anche se nei fatti questa differenza non avrebbe ragione di esistere.

Sulla questione dello sconto temporale torneremo diffusamente più avanti.

[15] Canberra Group (2001), pag. 64.

3. L'indagine EUSILC sui Redditi e le Condizioni di Vita

Una volta esaurita la discussione sul quadro teorico di riferimento, e prima di presentare i risultati ottenuti, sembra utile dare qualche informazione sulla base dati utilizzata.

Con la prima wave condotta nell'ottobre del 2004, ha avuto inizio l'Indagine su Reddito e Condizioni di Vita delle famiglie italiane, realizzata per partecipare al progetto europeo EUSILC (European Statistics on Income and Living Conditions). L'indagine è progettata con l'obiettivo di produrre due tipologie di stime annuali: trasversali, in grado di fornire, anno per anno, la condizione socio-economica in termini di povertà, esclusione sociale, distribuzione del reddito e altri aspetti delle condizioni di vita delle famiglie; longitudinali, al fine di individuare i cambiamenti delle condizioni di vita degli individui seguiti per quattro anni. Per fare ciò, il campione trasversale è composto dalla somma di quattro campioni longitudinali: ogni anno se ne sostituisce interamente uno di modo che ogni campione longitudinale sia seguito per un totale di quattro anni; tali campioni sono quindi sfasati nel tempo in modo che ogni anno ci sia la chiusura del panel che arriva alla quarta wave e l'inizio di un nuovo panel.

L'intero campione trasversale, relativo a un determinato anno, è quindi composto dall'unione dei 4 campioni longitudinali, ognuno per la sua specifica wave: in tal modo, ogni anno il campione è composto da un quarto di famiglie che partecipano per la prima volta all'indagine, un quarto di famiglie che partecipano per la seconda volta, un quarto di famiglie che partecipano per la terza volta e infine un quarto di famiglie che partecipano per la quarta volta all'indagine.[16]

Ogni campione longitudinale è a due stadi con stratificazione delle unità di primo stadio, i comuni, mentre le unità di secondo stadio, le famiglie, sono estratte dalle anagrafi dei comuni campione. La stratificazione delle unità di primo stadio, effettuata a livello regionale, è basata sulla dimensione demografica dei comuni e determina la suddivisione del territorio nazionale in 288 strati, omogenei per popolazione residente all'interno di ogni regione.

Il campione trasversale d'avvio, relativo all'anno 2004, è composto complessivamente da 31.998 famiglie, circa 8.000 per ogni campione longitudinale.

Per gli anni successivi, la numerosità campionaria del campione trasversale viene determinata come somma delle seguenti quantità:

- numero di famiglie con individui campione rispondenti nella wave precedente per i 3 campioni longitudinali che non vengono sostituiti;
- 8.000 famiglie nuove estratte appartenenti al nuovo campione longitudinale.

La popolazione di riferimento è costituita da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente all'estero. Sono escluse le famiglie residenti in Italia che vivono abitualmente all'estero e i membri permanenti delle convivenze istituzionali (ospizi, brefotrofi, istituti religiosi, caserme, ecc.).

L'unità di rilevazione è la famiglia di fatto. Questa va intesa come un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti e aventi dimora abituale nello stesso comune (anche se non residenti secondo l'anagrafe nello stesso domicilio).

Al fine di soddisfare le esigenze del progetto EUSILC, le notizie acquisite sono riferite a due periodi distinti: alcune, come ad esempio le informazioni familiari e individuali che caratterizzano la condizione di vita attuali, alla data di indagine (anno t); le altre, principalmente sul reddito, all'anno precedente (t-1).

[16] Per la prima edizione di indagine i quattro sottocampioni sono stati trattati come se fossero uno alla quarta edizione, uno alla terza, uno alla seconda ed uno alla prima. Questo ultimo è il primo a concludere un intero panel longitudinale di 4 anni nel 2007.

Un'analisi dei differenti metodi di stima dell'affitto figurativo condotta sull'indagine EUSILC è particolarmente importante: non esiste infatti ancora una soluzione condivisa su quale sia la migliore metodologia da utilizzare nel caso di indagini campionarie, tanto che la stessa Eurostat ha previsto che l'affitto figurativo non sia inserito nei redditi disponibili durante i primi anni di EUSILC per dar modo ai paesi membri di attuare le proprie sperimentazioni e di condividere i risultati per arrivare ad una soluzione unica: in questo modo sarà possibile avere redditi disponibili e informazioni sulla disuguaglianza economica comparabili tra i diversi Stati.

4. L'applicazione empirica

L'analisi è stata condotta sull'indagine EUSILC del 2005, e sono state implementate le metodologie dell'approccio del *rental equivalence value*, con l'esclusione del metodo della stratificazione. Si è già fatto notare come questa tecnica abbia delle serie controindicazioni in situazioni come quella italiana. Stando ai dati EUSILC, una quota inferiore al 15% dichiara di stare in affitto a prezzi di mercato; la numerosità è troppo bassa per evitare efficacemente il *trade-off* fra frequenze di cella troppo piccole per alcune strati ed un livello di disaggregazione troppo approssimativo per essere realmente informativo. Nel prossimo paragrafo il livello degli affitti soggettivi sarà quindi raffrontato con le stime derivanti dall'applicazione della regressione edonica, senza e con la procedura di Heckman. Le stime ottenute tramite autovalutazione, lungi dall'essere una soluzione ottimale, sono comunque presentate perché nel questionario è comunque prevista la domanda all'intervistato.[17]

Le informazioni raccolte riguardano 22.032 famiglie. Il gruppo sul quale effettuare le stime soggettive e OLS è determinato in base alle domande sul titolo di godimento dell'abitazione e, tra quanti sono in affitto, al fatto se pagano o meno un prezzo inferiore ai prezzi di mercato (domande 1.34 e 2.5).[18] Risultano come affittuari ai prezzi di mercato 2.489 famiglie.

Prima di arrivare a questo confronto, presentiamo la formalizzazione dei modelli utilizzati.

Il modello della regressione edonica può essere scritto come:

$$\ln y_i = \beta_0 + \boldsymbol{\beta}_{1k} \mathbf{X}_{ki}^{a} + \beta_2 T_i^{a} + u_i \qquad i = 1,\ldots, n_i \qquad (4.1)$$

dove y_i è l'affitto pagato, $\mathbf{X}_{ki}$ è un vettore di k caratteristiche della casa e della zona dove è situata l'abitazione, T_i è il periodo di occupazione della casa e u_i è il termine di errore. Il carattere a in apice sta ad indicare che il modello è applicato agli affittuari ai prezzi di mercato, di numerosità pari ad n_i. I parametri $\boldsymbol{\beta}$, stimati sugli affittuari ai prezzi di mercato, sono successivamente applicati alle caratteristiche del sottocampione di numerosità n_j dei proprietari[19] (apice p) per ottenere il corrispondente valore dell'affitto figurativo.

[17] La domanda posta ai proprietari di abitazione recita: "Se Lei vivesse in affitto in questa casa, quanto dovrebbe pagare al mese (escluse le spese di condominio, di riscaldamento e altre spese accessorie)?".

[18] La domanda 2.5 recita: "La sua famiglia paga un affitto inferiore al prezzo di mercato?". La domanda così posta può comunque portare ad una distorsione dovuta al fatto che la selezione è fatta in base alla valutazione soggettiva dell'intervistato.

[19] D'ora in avanti indicheremo per comodità come proprietari quanti non sono in affitto ai prezzi di mercato, indipendentemente dal reale titolo di godimento dell'abitazione.

$$y_{fig,j} = \exp(\beta_0 + \boldsymbol{\beta}_{1k} \mathbf{X}_{kj}^{p} + \beta_2 T_j^{p} + \sigma^2/2) \qquad j = 1,\ldots, n_j \qquad (4.2)$$

L'ultimo addendo dell'esponente è il cosiddetto *smearing estimator*, dove σ^2 è la varianza stimata del termine di errore u. La sua inclusione previene dall'incorrere nel cosiddetto *retransformation bias*[20] (Duan, 1983).

Il modello descritto nell'equazione (4.1) potrebbe soffrire di un problema di distorsione da *sample selection*: l'affitto è infatti osservato solamente tra quanti pagano un affitto ai prezzi di mercato e costoro potrebbero differire in maniera sistematica e non osservabile dal resto del campione. Occorre quindi risolvere il problema della mancanza di conoscenza dei fattori che determinano la condizione di affittuario. La procedura a due passi di Heckman (1979) incorpora nel processo di stima la possibile esistenza del *selection bias*. In questo contesto, la procedura consiste nello stima di un modello *probit* con la condizione di affittuario come variabile risposta ed una serie di caratteristiche familiari come esplicative.

Sia R^*_s la variabile latente che indica l'utilità indiretta di stare in affitto per la famiglia *s*.

$$R^*_s = \alpha_0 + \boldsymbol{\alpha}_1 \mathbf{Z}_s + e_s \qquad s = 1,\ldots, n; \quad n = n_i + n_j \qquad (4.3)$$

dove $\mathbf{Z}_s$ è un vettore di variabili esplicative, $\boldsymbol{\alpha}$ è il vettore di parametri da stimare ed e_s è il termine di errore. Poiché R* non è osservabile, la variabile dipendente R assume valore 1 (la famiglia è in affitto ai prezzi di mercato) se $R^* \geq 0$ e assume valore 0 (la famiglia non è in affitto ai prezzi di mercato) se $R^* < 0$. Il valore di Y è osservabile solamente per quelle famiglie per le quali $R^* \geq 0$ e, quindi, il valore atteso di Y nell'equazione (4.1) è il valore atteso di Y condizionatamente al fatto che sia osservato (R=1).

Tramite le stime del modello *probit* si ricava l'*inverse Mill's ratio,* pari a:

$$\lambda_s = \varphi(\alpha_0 + \boldsymbol{\alpha}_1 \mathbf{Z}_s)/\Phi(\alpha_0 + \boldsymbol{\alpha}_1 \mathbf{Z}_s) \qquad \text{per R=1 e} \qquad (4.4)$$

$$\lambda_s = -\varphi(\alpha_0 + \boldsymbol{\alpha}_1 \mathbf{Z}_s)/(1-\Phi(\alpha_0 + \boldsymbol{\alpha}_1 \mathbf{Z}_s)) \qquad \text{per R=0}$$

dove φ è la funzione di densità della distribuzione normale standardizzata e Φ è la distribuzione normale standardizzata cumulata. Questo termine viene successivamente inserito come ulteriore regressore nella stima dei minimi quadrati ordinari sugli affitti (4.1), che quindi diventa:

$$\ln y_i = {}^{h}\beta_0 + {}^{h}\boldsymbol{\beta}_{1k} \mathbf{X}_{ki}^{a} + {}^{h}\beta_2 T_i^{a} + {}^{h}\beta_3 \lambda_i^{a} + {}^{h}u_i \qquad (4.5)$$

e l'equazione di stima degli affitti figurativi (4.2) diventa:

$${}^{h}y_{fig,j} = \exp({}^{h}\beta_0 + {}^{h}\boldsymbol{\beta}_{1k} \mathbf{X}_{kj}^{p} + {}^{h}\beta_2 T_j^{p} + {}^{h}\beta_3 \lambda_j^{p} + {}^{h}\sigma^2/2) \qquad (4.6)$$

dove l'h sta ad indicare che i parametri sono stimati tramite la metodologia di Heckman, differenti quindi da quelli presentati nelle equazioni (4.1) e (4.2).

[20] La natura non lineare della trasformazione della variabile dipendente comporta che il valore atteso della ritrasformata tramite la funzione esponenziale non sia centrato sul valore medio osservato malgrado il fatto che il valore atteso dei residui sia nullo (Duan, 1983).

La procedura di Heckman richiede che $\mathbf{Z}_s$ includa una o più variabili, dette variabili strumentali, che influenzino la probabilità di essere in affitto ma che possano essere legittimamente escluse da $\mathbf{X}_k$ nell'equazione (4.5). Nel contesto della regressione edonica, con il costo dell'affitto determinato dalle sole caratteristiche dell'abitazione, possiamo utilizzare le caratteristiche familiari come variabili strumentali che influenzino la condizione di affittuario o meno.

4.1 La condizione di affittuario

Come variabili esplicative sono state utilizzate informazioni sulla famiglia, sull'intestatario del foglio di famiglia e sulla zona di abitazione. Come variabili relative alla famiglia sono state utilizzate il numero di componenti, il numero di percettori di reddito, il numero di minori in famiglia, il reddito disponibile familiare mensile e il suo quadrato; come variabili relative all'intestatario della scheda familiare sono state considerate l'età e il suo quadrato, il sesso, il titolo di studio in classi, lo stato civile in classi, la cittadinanza e la condizione professionale in classi; come variabili relative alla zona di abitazione sono state utilizzate la regione e il tipo di comune di residenza (è possibile vedere le modalità delle variabili direttamente dalla tavola 1).

Vediamo brevemente i risultati del modello *probit* che descrive la condizione abitativa delle famiglie. Più è alto il numero di componenti in famiglia, maggiore la probabilità di stare in affitto, mentre maggiore è il numero di percettori e minore è questa probabilità. All'aumentare del livello del reddito disponibile mensile la probabilità di stare in affitto diminuisce, così come decresce al crescere dell'età della persona di riferimento, anche se in maniera meno che proporzionale (considerando che anche il quadrato di questa variabile risulta significativamente diverso da 0). Osservando i coefficienti delle altre variabili relative alla persona di riferimento, l'essere cittadino straniero esercita un'influenza molto forte sulle probabilità di stare in affitto, così come l'essere coniugato o convivente ha una forte influenza sulla probabilità di essere in casa di proprietà. Quanti si trovano nella condizione di operaio (o figure professionali ad essa assimilabili) presentano una maggiore probabilità degli altri di essere in affitto, così come accade per quanti hanno un basso titolo di studio.

La tipologia di comune di residenza della famiglia risulta di una certa influenza, con una probabilità di essere in affitto crescente al crescere dell'ampiezza demografica e del grado di urbanizzazione, così come la regione di residenza: Piemonte, Val d'Aosta, Liguria e Campania risultano, a parità di altre condizioni, le regioni con maggiore probabilità di essere in affitto ai prezzi di mercato, mentre Toscana, Umbria, Lazio e Sardegna sono quelle con le minori probabilità.

Tavola 1 - Le determinanti della condizione di affittuario

Variabile	Parametro	Errore standard
Costante	-0.6519**	0.2066
Numero componenti	0.1109**	0.0272
Numero percettori	-0.0704*	0.0329
Numero di minori in famiglia (def. Ocse)	-0.0520	0.0370
Reddito disponibile familiare mensile	-0.0001*	0.0000
Reddito disponibile familiare mensile al quadrato	-0.0000	0.0000
Età del p.r.	-0.0267**	0.0065

*Significativo al 5%.
**Significativo all'1%.

Tavola 1 segue **- Le determinanti della condizione di affittuario**

Variabile	Parametro	Errore standard
Età del p.r.al quadrato	0.0001**	0.0000
Sesso del p.r.	0.0043	0.0441
Cittadinanza del p.r. (italiana vs non italiana)	1.0928**	0.0786
Stato civile del p.r. (Rif: coniugato/convivente)		
Celibe / nubile	0.3242**	0.0544
Separato / divorziato	0.5179**	0.0596
Vedovo / a	0.2000**	0.0618
Posizione nella professione del p.r. (Rif: Operaio o assimilato)		
Dirigente, quadro, impiegato	-0.1258*	0.0588
Autonomo	-0.1125*	0.0537
Non occupato	-0.1640**	0.0559
Titolo di studio (Rif: Nessuno / elementari)		
Medie inferiori	-0.0956*	0.0468
Medie superiori	-0.3017**	0.0495
Universitario	-0.3808**	0.0736
Tipo di comune (Rif: Comune con meno di 2.000 abitanti)		
Centro di area metropolitana	0.6158**	0.0855
Periferia di area metropolitana	0.5057**	0.0865
Più di 50.000 abitanti	0.7162**	0.0800
Abitanti in (10.001-50.000)	0.5012**	0.0767
Abitanti in (2.001-10.000)	0.2761**	0.0765
Regione (Rif: Lombardia)		
Piemonte	0.2417**	0.0648
Val d'Aosta	0.4675**	0.1032
Trentino Alto-Adige	0.1355	0.0793
Veneto	-0.0238	0.0649
Friuli Venezia-Giulia	0.0170	0.0775
Liguria	0.2408**	0.0722
Emilia Romagna	-0.0458	0.0659
Toscana	-0.2308**	0.0712
Umbria	-0.3365**	0.0874
Marche	-0.0773	0.0815
Lazio	-0.1591*	0.0732
Abruzzo	-0.0695	0.1121
Molise	-0.1767	0.1288
Campania	0.2731**	0.0688
Puglia	0.0780	0.0781
Basilicata	0.1116	0.1061
Calabria	-0.0446	0.1006
Sicilia	-0.1569	0.0919
Sardegna	-0.2735**	0.1034
N° osservazioni	22032	
Log-verosimiglianza	-7272.72	
Pseudo-R^2	0.1310	

*Significativo al 5%.
**Significativo all'1%.

4.2 L'equazione degli affitti

Come variabili esplicative nell'OLS sugli affitti sono state utilizzate le caratteristiche dell'abitazione, le caratteristiche della zona di abitazione e il periodo di occupazione dell'abitazione.

Relativamente a quest'ultimo aspetto, abbiamo già evidenziato come il considerare o meno il *tenure discount* nelle stime tira in ballo la questione di quale debba essere il prezzo di mercato di riferimento per la valutazione dei flussi di servizi abitativi. Quale che sia la posizione nei confronti dei proprietari, quello del *tenure discount* (o *seniority*) è però un dato di fatto, un effetto realmente esistente sugli affittuari e, di conseguenza, il periodo di occupazione della casa va considerato tra le variabili esplicative della OLS per non correre il rischio che gli altri parametri del modello risultino "non veritieri". Successivamente si deciderà se considerare lo sconto temporale anche per i proprietari o se, come evidenzieremo nel par. 4, sia preferibile "sterilizzarne" l'impatto nella stima dell'affitto figurativo. A margine di questa discussione, si aggiunge inoltre che la distribuzione del *tenure discount* è significativamente diversa nelle due sottopopolazioni, come reso evidente dalla tavola 2.

Tavola 2 - Principali momenti della distribuzione del numero di anni di occupazione della casa

	Media	Mediana	I quintile	II quintile	III quintile	IV quintile
Proprietà ed altri	21	19	5	14	23	35
Affitto a prezzi di mercato	11	6	2	4	9	19

Nella tavola 3, vengono riportate le stime ottenute per le due OLS, ottenute senza e con la metodologia alla Heckman, e considerando il periodo di occupazione come variabile continua; è stato inserito anche il periodo di occupazione al quadrato, al fine di cogliere eventuali effetti non lineari sul livello dell'affitto.

Tavola 3 - Le stime OLS

Variabile	OLS senza Heckman	OLS con Heckman
	Parametro (Errore standard)	Parametro (Errore standard)
Intercetta	4.8480**(0.2646)	4.8781**(0.1375)
Tipo di abitazione (Rif:Appartamento in edificio con più di 10 appartamenti)		
Villa, villino unifamiliare	-0.1222**(0.0434)	-0.0635 (0.0339)
Villa, villino plurifamiliare	-0.1502**(0.0388)	-0.1044**(0.0284)
App. in edificio con < 10 app.	-0.0617*(0.0292)	-0.0423*(0.0209)
Altro tipo	-0.0749 (0.0673)	-0.0384 (0.0551)
MQ	0.0121**(0.0011)	0.0102**(0.0009)
MQ^2	0.0000**(0.0000)	0.0000**(0.0000)
Vasca o doccia	0.0429 (0.0712)	0.0175 (0.0697)
Bagno interno	-0.0894 (0.1430)	-0.0146 (0.1139)
Acqua calda	0.1149 (0.1562)	0.1100 (0.0856)

*Significativo al 5%.
**Significativo all'1%.
In parentesi gli errori standard.

Tavola 3 segue **- Le stime OLS**

Variabile	OLS senza Heckman	OLS con Heckman
	Parametro *(Errore standard)*	*Parametro (Errore standard)*
Bagno interno	-0.0894 (0.1430)	-0.0146 (0.1139)
Acqua calda	0.1149 (0.1562)	0.1100 (0.0856)
Terrazza o balcone	0.0663*(0.0267)	0.0671**(0.0210)
Giardino	0.0504 (0.0332)	0.0497*(0.0253)
Parti danneggiate	-0.0477 (0.0379)	-0.0551*(0.0269)
Umidità in muri, tetti	-0.0043 (0.0296)	-0.0254 (0.0236)
Scarsa luminosità	-0.0932*(0.0362)	-0.0790**(0.0274)
Inquinamento	-0.0038 (0.0287)	0.0128 (0.0238)
Inquinamento acustico	0.0406 (0.0250)	0.0287 (0.0210)
Criminalità in zona	-0.0503 (0.0373)	-0.0406 (0.0278)
Tipo di comune (Rif: Comune con meno di 2.000 abitanti		
Centro di area metropolitana	0.6998**(0.0662)	0.6718**(0.0504)
Periferia di area metropolitana	0.5251**(0.0639)	0.5103**(0.0504)
Abitanti in (2.001-10.000)	0.1788**(0.0579)	0.1637**(0.0432)
Abitanti in (10.001-50.000)	0.3048**(0.0562)	0.3089**(0.0433)
Più di 50.000 abitanti	0.4686**(0.0622)	0.4425**(0.0462)
Regione (Rif: Lombardia)		
Piemonte	-0.1069*(0.0421)	-0.1235**(0.0374)
Val d'Aosta	0.1405*(0.0667)	0.0623 (0.0598)
Trentino Alto-Adige	0.1802**(0.0594)	0.1297**(0.0475)
Veneto	0.0580 (0.0463)	-0.0128 (0.0389)
Friuli Venezia-Giulia	-0.0649 (0.0605)	-0.1045*(0.0498)
Liguria	0.0002 (0.0472)	-0.0343 (0.0413)
Emilia Romagna	0.0718 (0.0444)	0.0328 (0.0394)
Toscana	0.1218*(0.0540)	0.0635 (0.0451)
Umbria	-0.1138*(0.0521)	-0.1346*(0.0548)
Marche	-0.0208 (0.0617)	-0.1131*(0.0499)
Lazio	-0.1305*(0.0575)	-0.1992**(0.0435)
Abruzzo	-0.1615 (0.0949)	-0.2540**(0.0696)
Molise	-0.5705**(0.1343)	-0.5005**(0.0903)
Campania	-0.2358**(0.0590)	-0.3109**(0.0382)
Puglia	-0.4677**(0.0518)	-0.4962**(0.0463)
Basilicata	-0.4745**(0.0653)	-0.5191**(0.0646)
Calabria	-0.5884**(0.0911)	-0.5903**(0.0637)
Sicilia	-0.5288**(0.0499)	-0.5290**(0.0474)
Sardegna	-0.3941**(0.0665)	-0.4861**(0.0678)
Seniority	-0.0204**(0.0028)	-0.0225**(0.0021)
Seniority^2	0.0002**(0.0001)	0.0002**(0.0000)
λ	-	0.0506*(0.0236)
N° osservazioni	2489	2489
R^2	0.4437	0.4441

*Significativo al 5%.
**Significativo all'1%.
In parentesi gli errori standard.

Osserviamo innanzitutto che l'ipotesi di selezione non casuale sui dati è valida: il parametro di selezione, λ, risulta statisticamente significativo. Torneremo in maniera più approfondita su questo aspetto nel par. 5, ma possiamo subito concludere che questo implica l'esistenza di un processo di selezione che, non considerato, potrebbe portare a distorsioni nella stima dei parametri delle esplicative. Commenteremo quindi i soli risultati relativi al modello con selezione, nella colonna a destra della tavola 3.[21]

Relativamente al tipo di abitazione nella quale si abita, si osserva che l'abitare in un palazzo con dieci o più appartamenti comporta, a parità di altre condizioni, un canone di locazione maggiore; è comunque da considerare che gran parte dell'effetto del tipo di abitazione è strettamente legato alla tipologia del comune di residenza e all'ampiezza della casa. Rispetto alle dotazioni della casa, l'avere un terrazzo, così come la disponibilità di un giardino, aumenta considerevolmente il valore della casa e, quindi, del canone di locazione; viceversa, avere parti della casa danneggiate e la scarsa luminosità diminuiscono questo valore. Come atteso, la metratura dell'abitazione ha un effetto positivo sul prezzo dell'affitto.

Le caratteristiche della zona di abitazione non sembrano avere un effetto rilevante, probabilmente perché le informazioni utilizzate sono estremamente correlate con la regione e la tipologia del comune di residenza, ed il loro effetto è quindi colto da queste due ultime informazioni: proprio rispetto al tipo di comune, si vede come al crescere della dimensione demografica del comune di residenza ed all'avvicinarsi ad un'area metropolitana il livello del canone di locazione aumenta in maniera considerevole.

Rispetto alla regione di residenza, si osservano prezzi in forte decremento passando dal Nord al Mezzogiorno, con Trentino Alto-Adige, Val d'Aosta, Emilia Romagna, Lombardia e Toscana che risultano le regioni più costose e Basilicata, Calabria, Sicilia e Puglia le meno dispendiose.

Quanto all'esistenza di un effetto da *tenure discount*, dalla regressione sugli affittuari emerge chiaramente l'esistenza di questo sconto che ha un impatto negativo, non linearmente decrescente, sull'affitto da pagare.

I risultati descritti sono stati ottenuti da modelli che presentano un buon grado di adattamento: l'R^2 presenta infatti un valore superiore a 0.44, molto elevato per un'applicazione a dati *cross-section*.

5. L'impatto sul valore degli affitti figurativi

Una volta applicati i parametri alle caratteristiche delle abitazioni di proprietà (o in usufrutto, uso gratuito o in affitto con prezzo inferiore ai prezzi di mercato) come descritto nelle equazioni (4.2) e (4.6), si osserva (tavola 4) come l'utilizzo dei diversi metodi porta a risultati differenti relativamente al valore delle abitazioni di proprietà sul mercato degli affitti. Rispetto alla sterilizzazione o meno dello sconto temporale, questo viene effettuato imponendo nelle equazioni (4.2) e (4.6) il periodo di occupazione per i proprietari pari a zero ($T_j^p = 0$).

Prendendo come riferimento l'affitto soggettivo medio a livello nazionale, pari ad una stima di 559 euro, si vede che ciascuna delle quattro stime econometriche è inferiore a

[21] Il lettore può comunque osservare nel confronto tra i modelli quali siano le variabili i cui parametri cambiano in maniera più significativa quando si inserisce il parametro di selezione.

quella risultante dall'autovalutazione. Considerando al contempo il processo di selezione nelle due sub-popolazioni e lo sconto temporale anche per i proprietari, le stime si abbassano considerevolmente: sul territorio nazionale, la stima ottenuta tenendo conto di entrambi i processi risulta in media pari a 339 euro, di circa il 40% inferiore rispetto alla stima soggettiva; è inferiore del 12,2% quando non si tiene conto di nessuno dei due. Questi andamenti sono confermati anche a livello ripartizionale.

Tavola 4 - Le stime degli affitti figurativi secondo le varie tecniche del rental equivalence value.

	Italia	Nord Ovest	Nord Est	Centro	Sud	Isole
Autovalutazione	559	633	636	677	377	361
Modello che considera lo sconto temporale						
OLS semplice	370	403	438	421	272	257
OLS con Heckman	339	381	397	376	247	239
Modello che *sterilizza* lo sconto temporale						
OLS semplice	491	530	578	560	367	343
OLS con Heckman	460	511	536	511	341	326

Si pone quindi il problema di stabilire quale procedura sia più corretta: relativamente all'esistenza di uno sconto temporale, è da ribadire anzitutto come il periodo di occupazione debba essere incorporato nel processo di stima dei coefficienti della regressione con il logaritmo dell'affitto come variabile dipendente, poiché è una delle determinanti più importanti del livello del canone di locazione.

Per capire se invece debba essere considerato o "sterilizzato" nel calcolo degli affitti figurativi, ricorriamo nuovamente al manuale di Canberra:

"*Hence, the value to the consumer is close to the market value of the service flow, since* ***the owner could presumably sell the housing unit and rent it back from the new owner*** *were it more profitable to do so*".

La parte che abbiamo evidenziato in grassetto sembra indicare la strada: il proprietario vende la casa e la riaffitta nello stesso momento dal nuovo proprietario in un mercato i cui prezzi sono fissati sulla base delle condizioni vigenti. In altre parole, il contratto di locazione scatterebbe al momento dell'ipotetico ingresso nella casa nella nuova condizione di affittuario, al momento quindi dell'intervista. Se questa nostra interpretazione è corretta, il periodo di occupazione della casa non va incorporato nella stima dell'affitto figurativo per i proprietari; il periodo di occupazione deve, quindi, essere posto uguale a 0 per essere sterilizzato nelle equazioni (4.2) e (4.6).

Relativamente all'utilizzo della procedura a due stadi di Heckman, è da osservare innanzitutto come, dai nostri modelli, l'esistenza di un processo di selezione sia risultato evidente. Per provare comunque a dare un'interpretazione economica di questo processo, osserviamo che il parametro di selezione dovrebbe tenere conto di variabili non osservabili che influenzano sia la decisione di stare in affitto sia il prezzo che si è disposti a pagare per quella abitazione. Interpretando il risultato in maniera analoga alla stima del salario di

riserva (ambito in cui la procedura è stata ampiamente utilizzata), l'affitto figurativo si può interpretare come una sorta di affitto di riserva: è il prezzo massimo che il proprietario sarebbe disposto a pagare per cambiare la propria condizione, ossia per accettare di diventare inquilino nella sua stessa casa. Nel diventare inquilino, il proprietario accetta di avere meno sicurezza e, quindi, chiede ragionevolmente un affitto inferiore rispetto al prezzo di mercato. In altre parole, mentre un inquilino può essere disposto ad accettare un determinato affitto se non ha alternative, per indurre un proprietario a cambiare status, cioè a rinunciare alla sicurezza della proprietà di un bene rifugio come è la casa, anche se potrebbe investire i soldi della vendita e ricavarne un flusso futuro di guadagni finanziari, è necessario offrire un canone di locazione minore.[22] La procedura applicata ai nostri dati fornisce un valore dell'affitto figurativo all'incirca del 8% inferiore rispetto a quelle ottenute con l'utilizzo della OLS semplice. A supporto di questa interpretazione, nella citazione precedente dal manuale di Canberra abbiamo evidenziato, questa volta con sottolineatura, anche un'altra parte. La frase sembra suggerire l'incorporazione del processo di selezione nel calcolo delle stime, poiché prende proprio in considerazione il fatto che debba esistere un vantaggio economico per convincere un proprietario a cambiare status.

Quanto appena detto, insieme ai diversi commenti presentati nel corso del paragrafo, ci porta quindi a ritenere preferibile nella stima dell'affitto figurativo il ricorso a stime econometriche, tramite regressione edonica corretta con la metodologia di Heckman per il *sample selection bias*, sterilizzando i risultati ottenuti per i proprietari dallo sconto temporale che gli affittuari ottengono in base al periodo di occupazione della casa.

Nella tavola 5 vengono rappresentati gli affitti soggettivi e le stime econometriche senza considerare lo sconto temporale sui proprietari. La regressione edonica è considerata invece sia con che senza la selezione di Heckman, e questo perché, ferma restando la nostra interpretazione, questo ultimo aspetto ci pare sicuramente più dubitabile rispetto alla sterilizzazione dello sconto temporale.

I valori sono stati trasformati in numeri indice per rendere visivamente più immediate le differenze riscontrate. Decomponendo per tipo di comune, si nota come i livelli imputati tramite tecniche econometriche mostrano una variabilità più contenuta rispetto agli affitti soggettivi. Ponendo come base l'autovalutazione distinta per tipo di comune (parte bassa della tavola 5), si nota che gli scostamenti relativi più ampi si registrano per i centri delle aree metropolitane.

[22] Si noti comunque che un ragionamento del genere è valido solamente per quanti sono proprietari di casa, e si troverebbero nella condizione di potere scegliere tra proprietà e affitto, mentre non è valido per quanto sono in condizioni di usufrutto, uso gratuito o affitto inferiore ai prezzi di mercato. Per questi ultimi la procedura di Heckman non andrebbe probabilmente considerata. Per semplificare la trattazione non sembra comunque il caso di complicare ulteriormente lo studio.

Tavola 5 - Le stime degli affitti figurativi sterilizzando lo sconto temporale

Tipologia del comune di residenza	METODO DI STIMA		
	Autovalutazione	Regressione edonica	Edonica con Heckman
	Base: affitto figurativo stimato con le differenti tecniche =100		
Centro di area metropolitana	145.9	131.3	128.6
Periferia di area metropolitana	109.8	113.4	113.6
Più di 50.000 abitanti	104.9	111.8	109.8
Da 10.000 a 50.000 abitanti	89.4	92.9	94.5
Meno di 10.000 abitanti	79.8	79.5	80.5
Totale	**100** ***(=559 €)***	**100** ***(=491 €)***	**100** ***(=460 €)***
	Base: autovalutazione per tipo di comune=100		
Centro di area metropolitana	**100** ***(=815 €)***	79.0	72.5
Periferia di area metropolitana	**100** ***(=613 €)***	90.7	85.2
Più di 50.000 abitanti	**100** ***(=586 €)***	93.6	86.1
Da 10.000 a 50.000 abitanti	**100** ***(=499 €)***	91.2	87.0
Meno di 10.000 abitanti	**100** ***(=446 €)***	91.2	87.4
Totale	**100** ***(=559 €)***	87.5	83.1

Non è semplice dare conto della differenza fra affitti soggettivi e stime tramite tecniche econometriche. In linea di principio la differenza può essere tanto dovuta ad una sovrastima da parte dei rispondenti nella valutazione soggettiva quanto ad una sottostima relativa alle metodologie econometriche e dovuta all'appartenenza delle case di proprietà e delle case in affitto a mercati completamente differenti relativamente alle caratteristiche delle abitazioni.

Da una parte, non v'è dubbio che la valutazione effettuata dal rispondente, se affidabile, contiene un plus informativo altrimenti non ottenibile: solo il proprietario può dare una valutazione pienamente comprensiva dei pregi e dei difetti della propria abitazione. Il problema consiste nel fatto che la valutazione può essere gravemente corrotta dalla mancata conoscenza delle reali condizioni del mercato degli affitti prevalenti nella propria zona da parte di famiglie non in affitto che non abbiano particolari motivi (quale ad esempio un cambio di casa previsto a breve) per essere adeguatamente e accuratamente informate.

La batteria di domande riguardanti il modo in cui il rispondente ha maturato la propria convinzione in merito alla autovalutazione non aiuta a discriminare la parte di informazione da considerare affidabile da quella non affidabile. E che un problema di affidabilità ci sia è dimostrato dal fatto che, tra il 2004 e il 2005, per la componente di campione soggetta a reintervista che non ha cambiato abitazione né titolo di godimento (non vive in affitto a prezzi di mercato) e non ha sostenuto lavori di riparazione – i quali potrebbero giustificare una differente valutazione del servizio abitativo- il 49,3% presenta valutazioni dell'affitto soggettivo con un scarto percentuale superiore ai 20 punti fra un anno e l'altro. Nel 16,2% dei casi le differenze sono superiori al 50%.

La tavola 6 fornisce un ulteriore riscontro: vengono riportati i coefficienti di correlazione per ciascun tipo di stima dell'affitto figurativo per gli anni 2004 e 2005. Anche in questo caso ci si concentra sul sottoinsieme di famiglie che non vive in affitto a prezzi di mercato e non ha cambiato abitazione né ha apportato migliorie, di modo da rendere efficace il confronto fra i valori della medesima stima in due diversi istanti temporali. Nel caso dell'autovalutazione la correlazione è pari a 0.66 e inferiore a quella mostrata dalle varie stime da modello.

Tavola 6 - I coefficienti di correlazione delle stime degli affitti figurativi secondo le varie tecniche del rental equivalence value per il 2004 e il 2005

	Coefficienti di correlazione
Autovalutazione	0.66
Modello che considera lo sconto temporale	
OLS semplice	0.84
OLS con Heckman	0.85
Modello che *sterilizza* lo sconto temporale	
OLS semplice	0.81
OLS con Heckman	0.82

D'altro canto, è da considerare che l'attribuzione di valori imputati tramite modelli tende a ridurre strutturalmente la varianza, poiché elimina la componente residua, e questo potrebbe avere conseguenze sulla distribuzione del reddito. È stato deciso di non introdurre una componente random *tout-court*, tipica dei modelli previsivi, per via del disegno di panel ruotato dell'indagine, con tre quarti del campione che viene reintervistato ogni anno, per evitare incoerenze a livello sia micro sia macro, fra le stime cross-section e quelle panel. Nell'estendere la metodologia qui presentata al contesto longitudinale, che è quello proprio di EUSILC, si dovrà puntare a depurare la variabilità delle stime soggettive dalla componente dovuta a errore di misura, di modo da recuperare la quota residuale.

6. L'impatto sulla distribuzione dei redditi

Resta da dare una valutazione su quale effetto abbia la scelta della tecnica sulle misure di disuguaglianza. A tal fine, abbiamo selezionato due fra gli indicatori di Laeken, cioè gli indicatori che Eurostat utilizza per fornire un quadro delle diverse realtà nazionali in termini di distribuzione del reddito e disuguaglianza: il rapporto di concentrazione di Gini e la percentuale di famiglie a rischio di povertà, sotto la soglia identificata come il 60% della mediana della distribuzione del reddito familiare equivalente.[23]

[23] Mentre per proprietari e usufruttari la stima dell'affitto figurativo è stata interamente aggiunta al reddito disponibile, per le famiglie in affitto a prezzi inferiori ai prezzi di mercato è stata aggiunta solamente la quota eccedente l'affitto effettivamente pagato. Inoltre, seguendo le prescrizioni Eurostat, dal reddito disponibile che incorpora l'affitto figurativo vengono decurtati gli interessi pagati sul mutuo contratto per l'acquisto dell'abitazione principale.

Tavola 7 - Principali indicatori di Laeken sulla distribuzione di reddito familiare senza e con affitti figurativi (secondo differenti tecniche)

Ripartizione di residenza	Senza affitto figurativo	Con affitto figurativo stimato tramite:		
		Autovalutazione	Regressione edonica	Edonica con Heckman
		Rapporto di concentrazione di Gini		
Nord Ovest	0.317	0.292	0.286	0.287
Nord Est	0.289	0.262	0.257	0.259
Centro	0.300	0.278	0.270	0.272
Sud	0.327	0.311	0.301	0.302
Isole	0.347	0.326	0.317	0.317
Italia	**0.328**	**0.309**	**0.301**	**0.302**
		Percentuale di famiglie a rischio di povertà		
Nord Ovest	10.3	8.3	7.6	7.6
Nord Est	9.4	7.1	6.5	6.6
Centro	13.3	9.9	9.6	9.8
Sud	32.1	34.4	31.5	31.5
Isole	35.6	36.4	35.2	34.6
Italia	**18.8**	**17.8**	**16.7**	**16.6**

Dalla tavola 7 emerge come l'inclusione degli affitti figurativi nel reddito familiare implichi una minore disuguaglianza, tanto a livello nazionale quanto per singola ripartizione nel paese. Se è vero che nella coda destra (i più ricchi) della distribuzione del reddito familiare si osserva una maggiore incidenza di proprietari (e, tra questi, di proprietari di abitazioni di pregio) è altrettanto vero che questa posta figurativa, comunque calcolata, aumenta in maniera meno che proporzionale rispetto al reddito. L'inclusione dell'affitto figurativo nel reddito disponibile ha quindi un doppio effetto (Istat, 2006): da una parte aumenta le differenze in valore assoluto tra il reddito dei proprietari di abitazione (nonché gli affittuari a prezzo inferiore a quello di mercato e gli usufruttari) e gli affittuari; dall'altra riduce la diseguaglianza poiché gli affitti figurativi sono distribuiti in maniera meno diseguale rispetto ai redditi. I risultati presentati suggeriscono che questo ultimo è l'effetto prevalente.

Detto che la disuguaglianza è più elevata con gli affitti soggettivi, l'utilizzo dei differenti metodi di computo degli affitti figurativi non sembra però comportare differenze rilevanti sull'indice di Gini, quanto piuttosto sulla percentuale di famiglie a rischio di povertà.

Disaggregando per i diversi titoli di godimento dell'abitazione (tavola 8), si nota che le incidenze delle famiglie sotto la linea di povertà cambiano considerevolmente, una volta inclusi nel reddito gli affitti figurativi. Non risulta sorprendente la variazione notevole di tale incidenza per le famiglie che vivono in abitazioni in affitto a prezzi di mercato, passando dal 29.6 quando non si considera la posta figurativa ad un massimo del 39.9 con gli affitti soggettivi. Nella stessa direzione (aumento dell'incidenza di povertà relativa) si muovono le famiglie in affitto a prezzi inferiori al mercato. Sostanzialmente stabile la condizione relativa dei proprietari che pagano un mutuo per l'abitazione. Viceversa, l'incidenza delle famiglie a rischio di povertà diminuisce sensibilmente per gli altri sottogruppi.

Tavola 8 - Percentuale di famiglie sotto la linea di povertà senza e con affitti figurativi (secondo differenti tecniche)

Titolo di godimento dell'abitazione	Senza affitto figurativo	Con affitto figurativo stimato tramite:		
		Autovalutazione	Regressione edonica	Edonica con Heckman
		Percentuale di famiglie a rischio di povertà		
Affitto a prezzi di mercato	29.6	39.9	39.1	38.5
Affitto non a prezzi di mercato	28.2	33.5	29.1	29.6
Proprietà con mutuo	8.8	9.1	9.0	9.3
Proprietà senza mutuo	16.3	12.7	11.6	11.5
Usufrutto	22.9	15.8	13.0	13.2
Uso gratuito	27.2	23.3	22.2	22.3
Italia	**18.8**	**17.8**	**16.7**	**16.6**

7. Conclusioni

In questo lavoro viene presentata un'analisi sul calcolo dell'affitto figurativo tramite l'utilizzo dell'indagine EUSILC; questo, stimato per proprietari, usufruttuari e famiglie che vivono a titolo gratuito nella propria abitazione, oltre che per gli affittuari a prezzi inferiori ai prezzi di mercato, deve dare conto del flusso di servizi abitativi che si ottiene in virtù del proprio titolo di godimento e deve essere considerato come parte del reddito disponibile secondo le disposizioni di Nazioni Unite, ILO e Eurostat. Per l'indagine EUSILC, Eurostat prescrive di non includere questa posta figurativa nel reddito disponibile nei primi anni dell'indagine, per dare tempo ai paesi membri di arrivare ad una soluzione condivisa tra le diverse metodologie possibili. Si è quindi stimato questo flusso di beni e servizi tramite modello di regressione secondo la teoria della domanda edonica, per la quale il valore di un'abitazione è dato dall'intersezione tra domanda e offerta delle caratteristiche della casa stessa. Questo metodo è stato implementato come alternativa all'utilizzo del metodo dell'autovalutazione da parte dei rispondenti, che potrebbe essere troppo influenzato dalla mancata o parziale conoscenza del mercato degli immobili da parte di alcuni rispondenti. Tre sono le questioni affrontate:

- anzitutto, se la metodologia dell'autovalutazione da parte dei rispondenti porti effettivamente a una non corretta valutazione del valore del flusso dei servizi abitativi. Gli affitti soggettivi sono, in media, più elevati di quelli imputati tramite tecniche econometriche. Ma questo, di per sé, non indica quale sia il livello più vicino al valore "vero".
- Relativamente all'utilizzo della regressione edonica, si pone il problema del *tenure discount*: se lo sconto temporale che si osserva tra gli affittuari debba quindi essere considerato anche nella stima del calcolo figurativo o se, al contrario, quest'ultimo non ne debba tenere conto, come se il proprietario fosse entrato nella casa nella condizione di affittuario alla data dell'intervista.
- Infine, abbiamo verificato l'eventualità dell'utilizzo della procedura di Heckman per considerare l'esistenza di un processo di selezione sulla condizione abitativa. In questa maniera si considera non già il valore di mercato dei flussi abitativi dell'abitazione, quanto piuttosto - dati i prezzi prevalenti sul mercato degli affitti - il valore intrinseco

che il proprietario assegna ai servizi abitativi goduti, ivi compreso il costo opportunità del cambiamento di status: una sorta di affitto di riserva, minore rispetto a quello di mercato, che la famiglia accetterebbe per diventare affittuaria nella sua stessa casa, con la possibilità di investire in altro modo i propri soldi e ottenere vantaggi finanziari anche nel futuro, ma scontando comunque una maggiore insicurezza.

Stando alla nostra interpretazione del manuale di Canberra, che definisce le linee guida internazionali per le metodologie sul calcolo del reddito disponibile delle famiglie, lo sconto temporale deve essere "sterilizzato" ed il processo di selezione va considerato per tenere conto del fatto che le due sottopopolazioni possono differire in maniera sistematica. Le stime di tipo econometrico sono comunque da preferire all'autovalutazione dell'affitto figurativo. Quest'ultimo sconta infatti l'errata conoscenza della reale situazione del mercato immobiliare da parte delle famiglie proprietarie che non hanno particolari motivi (quale ad esempio un cambio di casa previsto a breve) per conoscere questo mercato, soprattutto relativamente al mercato degli affitti.

In ogni caso, è importante ricordare come il metodo proposto sia molto esigente dal punto di vista informativo; allo stato attuale esiste un problema di sottospecificazione del modello dovuto proprio alle informazioni presenti nel questionario. La selezione su quanti sono in affitto ai prezzi di mercato è fatta sulla valutazione dei rispondenti, ed è vincolata quindi ad un alto grado di soggettività. È quindi stata prevista per le edizioni successive dell'indagine l'aggiunta di informazioni relative al proprietario dell'immobile (Ente, privato, società, parente…) e a ulteriori caratteristiche dell'abitazione.

Bibliografia

Arèvalo R. (2001). *El mercado della vivienda in España*, Unpublished Ph.D. dissertation, Universidad Complutense de Madrid.

Arévelo R. e J. Ruiz-Castillo (2004). *The rental equivalence approach to non rental housing in the Consumer Price Index: evidence from Spain, Universidad Carlos III de Madrid*, Departamento de Economía Working Paper Series, No 04-17.

Canberra Group (2001). Expert Group on Household Income Statistics, Final Report and Recommendations. Ottawa.

D'Ambrosio C, Gigliarano C. (2006). *The distributional impact of "imputed rent" in Italy*, AIM-AP project, serie Deliverable 1.1b.

Dougherty A. e R. van Order (1982). Inflation, Housing Costs and the Consumer Price Index, *American Economic Review*. Vol. 72, n°1: pagg. 154-164.

Duan N (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*. Vol. 78: pagg. 605-610.

Eurostat (2006). *HBS and EU-SILC imputed rent. Meeting of the Working Group on Living Condition (HBS, EU-SILC and IPSE)*. Doc. Eu-silc 162/06.

Gillingham R.F. (1983). Measuring the Cost of Shelter for Homeowners: Theoretical and Empirical Considerations", *The Review of Economics and Statistics*. Vol. 65, n° 2: 254-265.

Heckman J. (1979). Selection bias as a specification error, *Econometrica*, Vol. 47 n°1: pagg. 153-161.

Hoffmann J. e C. Kurz (2004). *Rent indices for housing in West Germany 1985 to 1998*, Studies of Economic Research Centre of the Deutsche Bundesbank Discussion Paper n° 08/2004.

ILO (2003). *Resolution concerning household income and expenditure statistics*, 17th International Conference of Labour Statisticians.

Istat (2006). Reddito e condizioni di vita. Informazioni n.31, Roma.

Miron J.R. (1990). Security of tenure, costly tenants and rent regulation, *Urban studies,* Vol. 27, n° 2: 167-184.

Rosen S. (1974). Hedonic prices and implicit markets: product differentiation in perfect competition, *Journal of Political Economy*, Vol. 82: 34-55.

How Many Households? A Comparison of Scenarios in the European Union: from Europop2004 to Europop2008

Carlo Maccheroni,[1] *Tiziana Barugola*[2]

Abstract

In this article we propose household scenarios in a selection of EU countries until 2031. Since household projections are derived from those of the population we have compared the results achieved applying the recent EUROSTAT projection: EUROPOP2004 (Baseline Scenario) and EUROPOP2008.

To these we have applied two hypothesis about the trend of the average household size: the first, the observed decreasing trend, will continue in the future; the second, the average size, has been fixed at the values of the 2001 Census. The results seem to be closer to reality when they are derived from combining the population of the Convergence Scenarios with the hypothesis of decline in household size. Nevertheless, these highlight and exalt some problematic aspects of the evolution of single people in the context of an ageing continent.

Key word: household, projection, EUROPOP2004, EUROPOP2008

1. Introduction

The household is a constantly evolving "unit/entity" in which the passing of time is perceived on two distinct levels: the first, historical, follows the evolution of society, while the second, that of individuals, follows the *life cycle*. For example, a young couple may have children and, in a period of 5 to 20 years, the household will reach its maximum size; the couple will then return to its original size and, sooner or later, will comprise only a single member (Keyfitz, 2005). This process permits qualitative observation, tracking modifications in household nuclei or modifications in the type of households,[3] and also quantitative examination, charting changes in size, i.e. in the number of members. In this work, the household will be observed from the latter viewpoint.

[1] Ordinario di Demografia (Università di Torino), e-mail: cmaccheroni@econ.unito .it.

[2] Prof.ssa a contratto (Università di Torino), e-mail: barugola@econ.unito.it

[3] As the concepts of "family" and "household" are confused in many cases, it is necessary to make them clearer in advance. According to the OECD "Glossary of Statistical Terms" the household is based on the arrangement made by persons, individually or in groups, for providing themselves with food or other essentials for living. A household may be either (a) a one-person household, that is to say, a person who makes provision for his or her own food or other essentials for living without combining with any other person to form part of a multi-person household or (b) a multi-person household, that is to say, a group of two or more persons living together who make common provision for food or other essentials for living. The persons in the group may pool their incomes and may, to a greater or lesser extent, have a common budget; they may be related or unrelated persons or constitute a combination of persons both related and unrelated. The family is defined primarily with reference to relationships which to pertain or arise from reproductive process. It may be: married-couple family (with or without children); father or mother with children

Modifications to household structure and size are also accompanied by the process of ageing of the population due to the increase in life expectancy and changes in fertility patterns.

In a recent work, Bongaarts (2001) has highlighted that households tend to be larger in countries with the highest proportion of young people. In some of their works, Jennings, Lloyd-Smith, and Ironmonger (1999, 2004) have investigated this issue in more detail, using only predefined relationships between age ranges to estimate and project the headship rate into the future. Furthermore, as relationships between age ranges are independent of the total size of the population, it is possible to observe the declining trend in household size against the more general backdrop of the process of ageing of the population.

From a theoretical point of view, various household forecasting methods have been developed in recent years, so a considerable variety exists today; the approaches adopted can be classified in different ways: generally speaking, a distinction can be made between dynamic and static methods and therefore between micro and macro methods (Zeng, Vaupel and Zhenglian,1998, O'Neill, 2004). In addition, for household forecasts with a macro method, it is necessary to project the population, generally autonomous, from which we can derive those relating to households. National Statistics Institutes and also EUROSTAT draw up population forecasts at various times and the results of the projection of "private households" obtained using the recent population forecast by EUROSTAT, EUROPOP2004 and EUROPOP2008, are compared in this study.

To begin with, in this study the macro approach based on the headship rate method (see appendix) has also been adopted (the only method suitable for the data available, without referring to the structural characteristics of the population) and taking 2031 as the time horizon of the household projections, shorter than EUROPOP2004 which continues as far as 2051 and EUROPOP2008 which reaches 2061. Although the trend in headship rate are not easy to extrapolate as the term "head" may be arbitrary and vague: such person is identified using census data that indicate the "reference person", usually the oldest member of the household, without taking into account possible changes in any case, the uncertainties inherent in the method are more significant in the case of long-term projections (Zeng et al. op. cit).

The assumptions underlying the two recent EUROSTAT projections (2004 and 2008) differ considerably from each other and, consequently, produce different results in the household projections. Here, the results of household forecasts drawn up using both the 2004 projection (only the Baseline Scenario) and the 2008 projection are compared, proposing two different assumptions regarding household size; according to the first, the current declining trend in the average number of members is expected to continue while, in the second, the average size recorded during the last census remains constant for the entire projection period.

In particular, the work is structured as follows: the first paragraph describes the characteristics of the data used and also the underlying differences of EUROSTAT population projections; the second delineates certain on-going population changes and outlines the characteristics of the households, presenting the results for a cluster of EU countries: France, the UK, Germany, Italy, Spain, Bulgaria, Poland, Latvia, and Luxembourg. The last section is dedicated to comparing the results of the forecast, combining the particular demographic situation of the nine countries examined with two different assumptions regarding household size and with two population projection scenarios, anticipating the possible total number of households and distribution according to number of members and, lastly, dedicating particular attention to single or one-person households which will represent the greatest *challenges in household support programmes* in an ageing Europe where "individual conditions of health

continue to improve but where, for structural reasons, the number of persons in need of care will increase by around 50% between 2000 and 2030" (Festy, 2008). The work ends with a methodological appendix.

2. The data

When planning the study, it was decided to use a single information system to chart the changes in households in Europe between 1991 and 2001, and also to project these until 2031. As only European countries were considered in the analysis, it was decided to use the EUROSTAT database. This source, almost complete for 27-EU countries as regards the data of the 2001 census[4] does not however include, for the 1991 census, certain information regarding the 10 countries that joined the EU in 2004 and, in particular, the two countries that joined in 2006.[5] The United Nations 1995 "Demographic Yearbook" was used to bridge this gap; other information was gathered through on-line access to the sites of the various national Statistics Institutes.

The sample countries (France, United Kingdom, Germany, Italy, Spain, Bulgaria, Poland, Latvia, and Luxembourg) are those on which attention was focused in a previous work (Maccheroni et al. 2006), identified through cluster analysis with the aim of forming like groups of countries as regards assumptions of future evolution as outlined by EUROSTAT 2004 projections. [6] Nine countries were then selected from these groups and, obviously, this selection is still considered valid as regards representing the differences and peculiarities of the future population trend of EU countries.

As indicated above, household forecasts have been taken from recent EUROSTAT population projections, for each country, known as EUROPOP2004 (EUROstat Population Projection 2004-based) and EUROPOP2008 (EUROstat Population Projection 2008-based).

In both, population projections are obtained applying the traditional and still widely used "cohort-component" method. The first set of forecasts called (in full) "EUROPOP2004 – Trend Scenario Variants" considers three different assumptions regarding the future evolution of the components of population change, defining these as: "High", "Base" and "Low". The High assumption presupposes a wider reduction in mortality, especially in countries where life expectancy at birth is lower, with greater recovery for males compared with females, and an increase in the fertility rate, country by country, higher than that envisaged in the "Low" and "Base" assumptions. The other two assumptions predict a lower reduction in mortality compared with that predicetd in the High assumption; also, according to the "Low" scenario, the decline in mortality terminates for East European countries, with Romania and Bulgaria, between 2030 and 2040, subsequently remaining stable for the last ten years of the forecast. As regards fertility rate, a lower recovery of the total fertility rate is observed in the "Base" scenario and a slight drop in the "Low" scenario. Migration perspectives are more difficult to

[4] Some of the information for Belgium and Sweden is missing.

[5] Estonia, Latvia, Lithuania, Slovenia, Slovakia, Hungary, Poland, Czech Republic, Cyprus and Malta joined the EU in 2004 and Bulgaria and Romania in 2006.

[6] These countries are representative of the following clusters respectively: 1st cluster: Denmark, Netherlands, Finland, Ireland, Belgium, Sweden, France, United Kingdom; 2nd cluster: Germany; 3rd cluster: Austria, Portugal, Greece, Slovenia, Italy, Spain; 4th cluster: Estonia, Lituania, Czech Republic, Slovakia, Romania, Hungary, Bulgaria, Poland, Latvia; 5th cluster: Cyprus, Malta, Luxembourg.

predict as liable to sudden changes compared with other components; according to the "Low" scenario, in the area of Eastern Europe, emigration will exceed immigration. On the other hand, the "Base" assumption predicts an inversion of sign of the initial negative migration balance for some countries of the Eastern area. In the "High" scenario, in all East European countries, emigration will prevail in the initial phase of the forecast and afterwards immigration (Maccheroni et al. 2006).

After defining the assumptions, EUROSTAT makes population projections on the basis of possible combinations of the assumptions: each combination is referred to as "forecast scenario variant";[7] as already mentioned, we have used the forecast based on the use of the "Baseline" scenario.

The most recent EUROPOP2008 projections known (in full) as "EUROPOP2008 – Convergence Scenario" depict a single scenario developed assuming that the socio-economic and cultural differences between the countries of the European Union (EU) will continue to fade. In the long term, this implies convergence (forecast for 2150) between current mortality and fertility levels, while it is assumed that net migration will continue to decrease, also converging to zero.[8]

The main problem as regards the methodology adopted is that of setting the values of the components of population change that trigger this process of "convergence". These latter projections, which cover the period between the 2008 and 2061, presuppose a reduction in the total fertility rate in Europe. In 2008, the range of values included, at the bottom, Slovakia with 1.25 births per woman and, at the top, France with 1.98 births per woman. At the end of the projection period, Slovakia is expected to achieve a rate of around 1.5, whereas France is expected to decrease to 1.93 births per woman. Also, at the end of the period, it will be possible to observe two groups of countries, the first, comprising a higher number of countries, will include those with the lowest fertility rate, all within a range spanning between the 1.47 of Slovakia and the 1.6 of Cyprus; the second, of reduced number, comprising France, Ireland, Sweden, the UK, Belgium, Netherlands, Luxembourg and Estonia with a range of values between the 1.66 of Estonia and the 1.93 of France. Similarly, as regards life expectancy at birth, a concurrent increase in the values for males and females and narrowing of the differences between gender and between European countries is predicted. As regards females, France and Italy will top the classification of countries as regards longest life expectancy, and Italy and Sweden for males; Bulgaria and Romania will close the classification for women and Lithuania and Latvia for men. At the same time, the "distance" between the countries will shorten: in 2008, the range of values for women was between 76.6 years in Romania and 84.2 years in France, in 2060, the minimum value will be 86.6 years (Bulgaria) and the maximum 90.1 years (France). A similar process may also occur for men for whom, in 2008, Latvia with life expectancy of just 66 years is at one end of the range of values and Sweden with 79 years at the other; it is forecast that, in 2060, the range of values will be between the minimum of 80 years of Latvia and the maximum of 85.5 years of Italy.

[7] In this way, six variants have been produced that highlight the effect on the consistency of the population and on its composition according to gender and age in relation to the possible occurrence of each of these; a seventh variant is also presented in which the contribution of migration is considered nil and which has been developed considering the assumptions on fertility and mortality of the Base variant.

[8] A predictive scenario assuming that the contribution of migrations will be nil has been inserted in this new projection

3. Comparison between countries: background

3.1 Socio-demographic changes and household modfication

The change in household structure and in the very model of the household is closely tied to population change and also to socio-cultural and economic modifications in a country. In order to understand and compare the differences between the countries of the European Union, attention must first of all be focused on certain socio-demographic indicators and on the effects of changes in these. This comparison can be made taking into account the consequences of evolution of the natural components of population change, i.e., as already mentioned in the introduction, fertility and mortality.

In 2007, the Total Fertility Rate (TFR) in most countries of the EU was between 1.3 and 1.49, therefore configuring these as countries with "very low fertility"; in the same year, only eleven countries had a TFR between 1.5 and 2, configuring these as "low fertility" countries (Billari, 2005). Among these eleven countries, France, Ireland, Denmark, Finland, Sweden and the UK had the highest values, all between 1.8 and 2 births per woman. Even if still below replacement level, TFR levels continue to increase after the minimum levels recorded in the last period of the last century.

The change in the total fertility rate coincided with a change in nuptiality patterns. Due to the increase in age at first marriage, a high proportion of young people continue to live within the family in particular in the countries of Southern Europe, while in North, Central and Western Europe, a high proportion of young adults live as singles (Alders and Manting, 2003, Aassve et al., 2002). However, the timing and rate of first birth is tied to the decline in households of young people more in the South than in the North of Europe; for example, in Sweden, more than half of the births in 2006 were outside marriage, around 50% in France and Denmark, dropping to only 5% in Greece and Cyprus, the countries that close this classification .

It must also be remembered that personal preferences are accompanied by changes in cultural models and economic restrictions. The loss of the specific function of marriage has made other types of living arrangements and co-habitation more "acceptable", even if these and also marriage are characterised by higher risks of instability (Andersson, 2003). On the one hand, more favourable economic conditions may induce young people to decide to leave home and form a new household regardless of whether they marry or decide to live on their own; on the other hand, in the recent past, greater instability and uncertainty of employment (reflected in the increased number of fixed-term contracts) has been considered the main cause of young people's delay in forming a new family (Aassve et al., 2002).

Another determinant in the evolution and transformation of household structure is the process of ageing: a process that has generated an increase in the population over 65, not as yet regularly distributed in the area of the European Union. According to recent data, life expectancy at 65 years of age in East European countries is 13 years for males and around 16 for females whereas in the other countries it is around 16 and 20 years respectively.

For the over 65s, considerable gender differences exist as regards the distribution of "marital status". It is true that the distribution of marital status amongst the elderly reflects a country's past nuptiality pattern but, today, the most important aspect is the quantity and proportion of married people compared with other categories (widowers/widows, separated and unmarried). For example, in Italy, considering the 65 to 74 and 75 to 84 age ranges, according to the data of the last census, 85% and 78% of men in the two age ranges are still

married, while 5 % to 2% of women are still married. Taking into account widowers and widows, the ratio is inverted: 6% and 15% for males, 76 and 85% for women.

The different life expectancy of the two genders is the most decisive determinant of this imbalance: a difference that places women in a position of social disadvantage, depriving them not only of emotional but also instrumental and economic support, a condition further aggravated by the fact that the most frequent type of household amongst the elderly is now the "single-person household": this is also aligned with the paradigm of the second demographic transition in which the characteristic individualism present in most countries, with its need for privacy and independence, is one of the main causes of the increase in households of singles (Van de Kaa, 1987). A change that, in abandoning old values and adopting new behaviour patterns, is spearheaded by the countries of Northern Europe, closely followed by those of the West and, with a time lag, in the countries of the South. According to this thesis, it has been possible to interpret and compare the high proportions of single elderly widows in Finland and the UK with the lowest proportions in Italy and East European countries.

The analysis of the determinant factors of "living in the family" is complex. In the future, the greater variability linked to the personal and matrimonial background of the elderly, their financial possibilities, linked in turn to the different social security systems and also their health and, last but not least, gender, will have to be taken into account: all factors that tend to introduce considerable margins of uncertainty in the projections and expand the number of possible future scenarios.

3.2 Changes in quantitative aspects

In the countries of the previous 15-EU, the period between the two last censuses was marked by an increase in the population and in the number of households. This trend, which is common not only to the countries for which the results of the analysis are presented, has triggered a reduction in the average household size.[9] In the same period, the 10 countries that joined the European Union in 2004 revealed different trends. For example, among the countries considered, Bulgaria revealed a concurrent decrease in the population and in the number of households, Latvia was characterised by two different trends (decrease in the population and increase in the number of households), while the same trend already highlighted for the countries of the 15-EU (i.e. growth of the population and in the number of households), was evident in Poland.

With the aforementioned aim of recognizing differences, it should be remembered that the decline in population in East European countries was to be ascribed mainly to the migration component of population change (migration became possible in the various countries only after the dissolution of the USSR). This factor was accompanied by the repercussions of the decline in the fertility rate which was common in all European countries in addition to the fact that, after 1991,[10] life expectancy in these countries did not increase at the same rate as observed in advanced development countries.

[9] The related variations, based on the results provided by the geometric increase rates, fluctuate between – 3.9 for thousand for Luxembourg and – 12.2 per thousand for Spain

[10] This marked the start of a period of transformation accepted by the Central and East European countries in which downgrading of living conditions represented the inevitable price to be paid for transition to the new economic system.

The reduction in average household size has generated entities with a lower number of members, but what are the most common, widely-spread forms of household according to size?

In 2001, in seven of the nine countries, the most common type of household consisted of two persons (table 3 first column of each), recently overtaken in Germany and Poland by single-person households (one person). France, Germany, and the UK reveal high percentages (around 30%) of two-person households, followed by one-person households; in some cases, such as the UK, three-person households follow at some distance (with only 15%). In Bulgaria, Italy, Latvia, Poland and Spain, around one fifth of the households consists of three persons; this percentage drops slightly for four-person households, while households with five or more persons represent a slender minority everywhere.

Returning to the single-person household, past data and research indicate that this is the most frequent type of living arrangement amongst the elderly, but how many elderly people are there in single-person households? The data show a high percentage of alone elderly especially in the countries of Southern Europe; also, in this area, Italy is the country with the highest proportion (more than 50%) of single-person households above 65, followed immediately afterwards by Spain.

4. The projections

As mentioned in the introduction, the results (fig. 1 and table 1) of the projections of the number of households presented in this study are based on EUROSTAT population projections (2004 and 2008) to which two hypotheses regarding household size are added: the first, now shared by developed and non developed countries, according to which household size will continue to decline; the second called, “constant rate”, applies the headship rate, as recorded in the 2001 census, to the forecast population on the basis of two different EUROSTAT projections, maintaining this unchanged throughout the entire projection period. These latter scenarios, where only changes at population level are involved, represent a yardstick: that is to say they answer the question, “if household size does not change, how many households may be present in the future?”.

In particular, Latvia, Poland and Bulgaria are characterized by a common decline in population in both EUROSTAT projections while, in the projections of the households, two of them are distinguished by a marked decline in the total number of households, while an increase is only observed in Poland.

If the constant rate hypothesis is considered and the forecast population in 2004, the country where the greatest difference in relation to 2001 should be recorded is Bulgaria, with a reduction of more than 20% of the total (table 2); the reduction is less consistent in the case of the constant rate forecast using the population of the 2008 projection. If, on the other hand, we consider the hypothesis of a reduction of the household and the population of the 2004 projection, a reduction in households (7%) is still observed, whereas a very slight increase (2%) is predicted in the case in which the population of the Convergence Scenario is used.

In Poland, according to both the 2004 and 2008 projections, the reduction in the population is less marked at least until 2030; consequently, a slight reduction in households is observed only if the “constant rate” hypothesis is considered (around 5% in both scenarios). On the contrary, unique amongst the countries of this group, an increase of

around 18% is forecast applying the hypothesis of a reduction in household size at the population of the two EUROSTAT projections.

In Latvia, the decline in the population will be accompanied by a reduction in households in each of our forecasts: a higher reduction of around 15% is forecast in the two scenarios produced adopting the "constant rate" hypothesis and a slight reduction in those that apply the hypothesis of a reduction in household size. At the end of the projection period, the difference between these two latter scenarios will be very small, with just over two thousand households more in the scenario linked to the population of the 2008 projection.

As opposed to the previous countries, for France and the UK, the EUROSTAT projections forecast a further increase in the population at least until 2030 (around 10% in both countries according to the 2004 projections and between 16 and 18 % in those of 2008), but the greatest differences are observed in the results of the projections of the households. If the "constant rate" hypothesis is considered, higher growth is exepected when the population of the Convergence Scenario is used (16 % in France and 18% in the UK) but combining the hypothesis of a decrease in household size with the population of the Convergence Scenario, the increase will be more than 37% in France and more than 34% in the UK; using the population forecast by the 2004 projections, the aforesaid increases will be only slightly lower (31% in France and 24% in the UK).

For Germany, EUROSTAT predicts a reduction in the population according to both the EUROPO2004 and EUROPOP2008 projections. In our projections regarding the households instead we can observe a slight decrease in both scenarios when we apply the "constant rate" hypothesis, opposite an slight increase in the others – where we adopt the hypothesis of the decline in household size.

In 2031, the difference in the total number of households – between the maximum forecast with the 2008 population, assuming a reduction in the number of household members and in relation to the "constant rate" hypothesis – will be around 5 million households.

In Italy and Spain, the predictive scenarios proposed by EUROSTAT produce different trends: a decrease in the population according to the 2004 projection, compared with an increase in the more recent one. These trends generate very different results as regards the household projections: for Italy, applying the "constant rate" hypothesis to both predictive scenarios proposed by EUROSTAT, only minor differences would be observed (at the most, an increase of 8% in 2031) whereas applying the other hypothesis, a more than 30% increase in the number of households would be observed considering the Convergence Scenario, moving therefore from the current almost 24 million to more than 29 million. In Spain, according to projections adopting the "constant rate" hypothesis, a more consistent increase in households would be observed using the population of the Convergence Scenario (29%), lower than that of 2004 (11%).

Applying the other hypothesis, even more marked differences would be observed: in the projection based on the population of the Convergence Scenario, the increase could exceed 70%, which means that it is possible to observe more than 24 million households, while in the projection according to EUROPOP2004, only over 20 million would be observed.

From this point of view, the last country of our cluster, Luxemburg, shares the same future as two "small" countries (Cyprus and Malta): an increase in the population, a reduction in household size. This means that, in all the possible forecasts, an increase in households is observed with only minor differences from scenario to scenario.

If the same trend towards reduction in household size continues in the future, how will the households be configured?

According to the results of the projections of distribution by number of members, at the end of the period and with the sole exception of Spain (table 3), around 40% of households consist of a single person.[11] This implies a considerable increase in single-person households considering both the 2004 and 2008 population projections. In countries where an increase in population is forecast, noteworthy increases in the number of single-person households may be observed due to the process of ageing and the tendency towards individualism. Among those examined here, Bulgaria is the country for which the lowest increase in single-person households is forecast, considering the population of the most recent EUROSTAT projections: in absolute value, there will be almost 900 thousand, with an increase of maximum 30%. Also in Germany, the increases fluctuate around 30% (considering both the expected population in 2004 and in 2008) but, in this country, this increase implies more than 17 million singles.

In Poland, more than 5 million single-person households may be observed with an around 70% increase in relation to 2001, regardless of the population forecast adopted. A similar observation may be made for the UK where, as in the case of Poland, there may be an increase of around 70%, but single-person households could top the 10 million mark.

In France, between 12 and 13 million singles are forecast according to whether the predictive scenario of 2004 or 2008 is used (with an increase of just over 80%).

Continuing this classification, the highest increases are found in: Italy where, applying the population of the Convergence Scenario, almost 12 million singles are obtained with an increase of 200% in relation to 2001; Luxembourg, where the increase is slightly above that of Italy (210%) but, in absolute values, single-person households would amount to just over 100 thousand; lastly, in Spain where the highest increase (between 250% and 300%), is forecast according to the population considered. These increases imply a transition from around 3 million households in 2001 to almost 6 million singles if the population of EUROPO2004 is used, or to more than 8 million considering that of EUROPOP2008.

Minor increases in two-person households are forecast in all the countries examined; this will involve a slight, constant growth in this type of household; more precisely, the increases will range from a minimum of 10% for Bulgaria to around 20% in Germany and 25% for Poland. On the other hand, Italy, France, the UK and Luxembourg are characterised by a similar percentage growth of between 40 and 50% according to the population projection used; in Spain, as observed above for single-person households, it is estimated that there will be a higher increase of around 90%, equal to 2 million households.

The percentage weight of three-person households will drop in almost all the reference countries, with the sole exception of France and Italy, even though the difference is minimum (one percentage point for France, almost three for Italy). Lastly, a reduction in larger size households is expected in each country

At the moment, the only data observed and sampled that can be compared are those referring to the EUROPOP2004 Baseline scenario. In this case, the future population will be lower in almost all the countries (in particular, in 22 of the 27 countries of the European Union), with results that vary from more than 2 million for France to a few thousand for

[11] This does not include detailed data at 1991 for Latvia and, for this reason, the data of this country are not shown.

Slovenia, Latvia and Luxemburg. For certain countries, information regarding the average number of household members and the number of households has been obtained from national Statistics Institutes. With regard to the previous result, the forecasts of the households are lower. For example, in the case of Italy, the number of households observed in 2006 was 2.6% higher than forecast; in the case of Finland, in 2007 the difference exceeded 1.15%; in Austria 1%; similar percentages are observed for Latvia. In view of the trend of the projections, the results linked to the EUROPOP2008 scenario which, however, highlight and accentuate certain particularly problematic aspects of the evolution of single-person households, would seem to be closer to reality.

5. To conclude

Longer life expectancy has generated an increase in the number of over 80s and reduced the gap between the expectation of life of men and women. For married people, this implies extension of married life (Meyers, 1986) while those who have lost a partner, through death or divorce, are faced with the prospect of a period of absence of partnership even if the probability of new marriages between widowers/widows and between divorcees is tied to gender (women have less probability of contracting new marriages than men) and age (older people have less probability of remarrying compared with younger people) (Butmpas, Sweet and Martin, 1990).

In the future (but this phenomenon is already evident), compared with previous generations, new elderly cohorts will experiment different more complex, heterogeneous "living arrangements": on the one hand, a first or second marriage, co-habitation, living "together but in different houses", life as single or as single parent whereas, on the other, households incorporating non-household members are on the decline. In this condition, it is easier to assume that the elderly will help and sustain each other (even if the number of divorces continues to grow); unfortunately, the fact that one member of the couple will be affected by difficulties as age increases must be taken into account.

Living longer, having children late in life and becoming grandparents even later means that various generations co-exist inside the household (Golini and Iaconucci 2003). Consequently, a household has a "household network" consisting of three or four generations. Although, at first sight, this may seem a positive factor, it is offset by the fact that is a thin and long network and, children may experience situations of conflict, having to divide their time between work, the household and assisting their parents (Festy, works cited). What emerges today and which goes beyond and oversteps the situation of married life is the presence of a type of support that is difficult to quantify as it is *informal between members of the same generation*: an intra-generation assistance; an assistance that involves various types of human relationships (from friendship to good neighbourliness); a form of assistance in which the elderly are still active and in good health, involved in voluntary work; a support that, due to the problems tied to increasing age, will however be ever weaker. Today and in the future, assistance may be provided by members of the immediately previous generations, i.e. between 50 and 64 years of age. Certain researchers (Rowland, 1991) prefer to consider only women in this age range as able to offer assistance. Unfortunately, the forecasts are not optimistic, as the relationships between age ranges, on an ageing continent, tend to undermine this type of support.

Any forecast has a "content of uncertainty" and predicting even just the proportion of elderly singles may be misleading if we use this method as it is based on the application of an algorithm. Although this may, in the short term, produce consistent results, many variables must be considered which require the application of other methodologies and data. However, the results deriving from projections that combine the tendency towards a reduction in household size with the population of the Convergence Scenario appear more probable in the future but these results also stress the problems related to the increase in single-person households that, in the context of the process of ageing of the population, will consist for the most part of elderly people.

Table 1 - Trends in average size of households and headship rate in selected countries.

Year	Average size of household	Headship rate	Average size of household	Headship rate	Average size of household	Headship rate
Country	Bulgaria		Germany		Spain	
2011	2.51	0.3987	2.06	0.4845	2.56	0.3899
2021	2.36	0.4244	1.97	0.5052	2.34	0.4277
2031	2.23	0.4490	1.90	0.5250	2.16	0.4631
Country	France		Italy		Latvia	
2011	2.26	0.4430	2.40	0.4172	2.78	0.3595
2021	2.14	0.4681	2.24	0.4467	2.65	0.3772
2031	2.03	0.4921	2.11	0.4747	2.54	0.3944
Country	Luxembourg		Poland		United Kingdom	
2011	2.41	0.4147	2.63	0.3804	2.25	0.4435
2021	2.32	0.4305	2.45	0.4074	2.16	0.4627
2031	2.24	0.4459	2.41	0.4332	2.08	0.4811

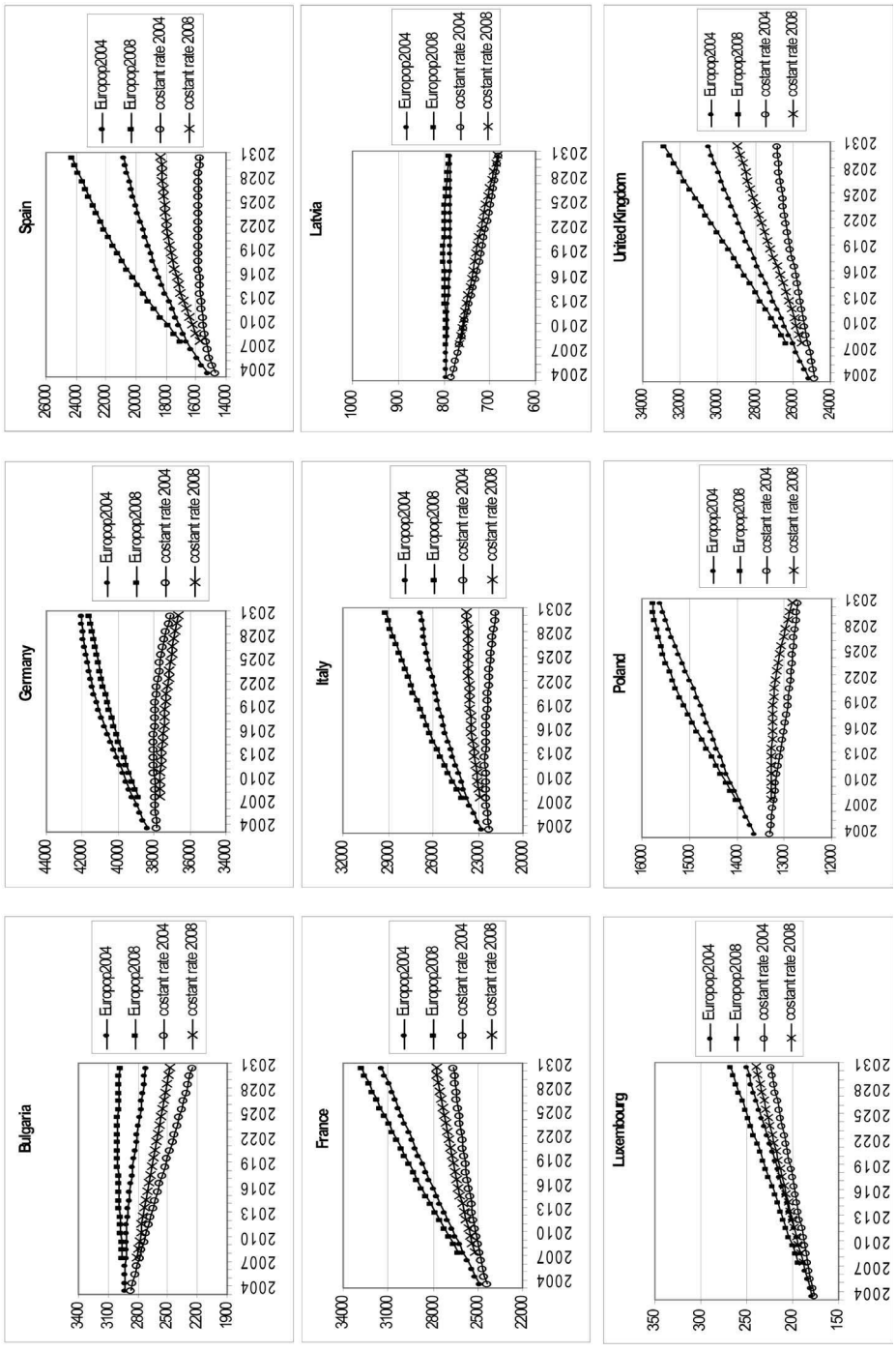


Figure 1 - Trend in number of households derived from EUROPO2004 population, EUROPOP 2008 population, constant rate applied to EUROPOP2004 population and constant rate applied to EUROPOP2008 population (values in thousand).

Table 2 - Changes in the number of the households during the periods 2001-2011, 2001-2021 and 001-2031 (I.N. based on 2001).

Year	EUROPOP 2004	EUROPOP 2008	Constant rate 2004	Constant rate 2008	EUROPOP 2004	EUROPOP 2008	Constant rate 2004	Constant rate 2008
Country	Bulgaria				Germany			
2001-2011	99.67	101.66	92.94	94.80	105.40	104.45	100.71	99.81
2001-2021	96.74	102.75	84.74	90.00	109.56	107.95	100.41	98.93
2001-2031	93.11	102.04	77.08	84.48	111.55	110.25	98.36	97.22
Country	Spain				France			
2001-2011	122.26	129.19	109.64	115.86	112.31	114.49	105.64	107.68
2001-2021	136.54	153.85	111.64	125.79	122.57	126.64	109.09	112.72
2001-2031	147.05	171.10	111.03	129.19	131.89	137.85	111.67	116.72
Country	Italy				Latvia			
2001-2011	111.41	114.40	103.11	105.87	98.79	99.21	93.79	94.18
2001-2021	118.36	125.04	102.31	108.08	97.96	99.62	88.64	90.14
2001-2031	123.02	133.78	100.06	108.81	97.95	98.27	84.76	85.04
Country	Luxembourg				Poland			
2001-2011	114.16	118.53	109.68	113.88	106.53	107.42	98.62	99.44
2001-2021	129.30	136.98	119.66	126.76	111.81	114.54	96.65	99.00
2001-2031	145.67	156.02	130.14	139.39	117.14	118.29	95.21	96.15
Country	United Kingdom							
2001-2011	108.78	110.94	103.92	105.99				
2001-2021	117.17	122.63	107.31	112.32				
2001-2031	124.44	134.22	109.60	118.22				

Table 3 - Distribution of households by number of members *(percentages)*

Number of person	**Country**							
	Bulgaria				Germany			
Year	*2001*	2011	2021	2031	*2001*	2011	2021	2031
1	*22.69*	31.9	36.7	40.7	*35.82*	37.8	39.7	41.4
2	*28.43*	28.3	29.7	30.9	*33.82*	36.6	39.2	41.6
3	*21.57*	19.9	18.6	17.1	*14.52*	12.3	10.3	8.5
4	*17.95*	15.8	13.3	10.9	*11.49*	9.7	8.1	6.7
5	*5.82*	3.2	1.3	0.1	*3.28*	2.8	2.2	1.7
6+	*3.54*	0.8	0.4	0.2	*1.07*	0.8	0.5	0.2
total	*100.00*	100.0	100.0	100.0	*100.00*	100.0	100.0	100.0
	Spain				France			
year	*2001*	2011	2021	2031	*2001*	2011	2021	2031
1	*20.28*	26.2	31.1	34.9	*31.00*	34.6	37.9	40.94
2	*25.25*	26.7	27.7	28.0	*31.14*	32.6	33.9	35.19
3	*21.18*	21.3	21.0	20.1	*16.17*	14.6	13.2	11.76
4	*21.49*	19.6	17.3	14.6	*13.76*	12.0	10.4	8.89
5	*7.75*	4.6	1.9	2.0	*5.55*	4.5	3.6	0.72
6+	*4.06*	1.6	1.0	0.3	*2.38*	1.7	1.0	0.48
total	*100.00*	100.0	100.0	100.0	*100.00*	100.0	100.0	100.00
	Italy				Luxembourg			
year	*2001*	2011	2021	2031	*2001*	2011	2021	2031
1	*24.89*	31.9	36.7	40.7	*29.30*	32.9	36.3	39.4
2	*27.08*	28.3	29.7	30.9	*28.25*	28.2	28.1	28.0
3	*21.58*	19.9	18.6	17.1	*17.01*	14.7	12.6	10.8
4	*18.96*	15.8	13.3	10.9	*16.45*	15.6	14.8	14.0
5	*5.80*	3.2	1.3	0.1	*6.36*	6.3	6.2	6.1
6+	*1.69*	0.8	0.4	0.2	*2.63*	2.3	2.0	1.6
total	*100.00*	100.0	100.0	100.0	*100.00*	100.0	100.0	100.0
	Poland				United Kingdom			
year	*2001*	2011	2021	2031	*2001*	2011	2021	2031
1	*24.79*	28.0	32.4	36.4	*30.21*	33.5	36.6	39.5
2	*23.22*	23.6	24.2	24.7	*33.92*	34.4	34.9	35.3
3	*19.90*	19.6	19.2	18.8	*15.54*	14.7	13.8	13.0
4	*18.03*	16.1	13.7	11.6	*13.37*	11.4	9.7	8.1
5	*8.14*	7.3	6.2	5.2	*4.95*	4.4	3.8	3.2
6+	*5.91*	5.2	4.3	3.4	*2.01*	1.7	1.3	1.0
total	*100.00*	100.0	100.0	100.0	*100.00*	100.0	100.0	100.0

In cursive observed values

Appendix

Households have been projected applying the specific headship rates (by age, gender) to the forecast population which is determined independently on the basis of various hypotheses regarding the components of population change.

Indicating as H the number of households, as P the population residing in the household and as *h* the headship rate, the rate at time *t* is determined as follows:

$$h(t) = \frac{H(t)}{P(t)} \tag{1}$$

the rate that can be broken down according to age and gender.

The total number of households at time *t+n,* knowing the future population, is given by:

$$H(t+n) = P(t+n) \times h(t+n) \tag{2}$$

Various methods exist for determining the headship rate, all of which assume that, in the future, this rate will continue to follow a trajectory determined according to a particular model.

One of the most frequently used methods at the moment is that developed in 1958 by the United States Bureau of the Census, known as the "modified two-point exponential".

On the basis of a time series considered, indicating the interval of time between the years as *k (0<k<t)*, the value of the rate in year *t+n*, assuming that the trend recorded in the *(n>0)* past continues also during the period of forecast, is determined in the case of a rising rate with:

$$h(t+n) = 1 + (h_{t-k} - 1) \times \left(\frac{h_t - 1}{h_{t-k} - 1} \right)^{\frac{(t+n)-(t-k)}{k}} \quad \text{if } h_t >= h_{t-k} \tag{3}$$

and with

$$h(t+n) = h_{t-k} \times \left(\frac{h_t}{h_{t-k}} \right)^{\frac{(t+n)-(t-k)}{k}} \quad \text{if } h_t < h_{t-k} \tag{4}$$

in the case of decreasing rate.

Methodologically speaking, the "modified two-point exponential" has a major characteristic that makes it preferable to other techniques, as the result of the projection produces a linear type trend with values between 0 and 1: on the contrary this result is not guaranteed for example by a generic linear type model.

Therefore, having calculated the estimated future headship rate for all the years of the period of the forecast, this is applied to the population residing in "private households", making it possible to determine the number of households year by year.

In each country, the percentage population resident in private household is more or less constant in time and, on the average, around 97%. This proportion calculated on the data of the last year available (the population at the 2001 census) remains constant in the projection period.

References

Aassve, A., Billari, F.C., Mazzucco S., Ongaro, F. (2002), "Leaving home: a comparative analysis of ECHP data*", Journal of European Social Policy*, vol. 12, No 4, pp.259-275.

Alders, M., Manting D. (2003), "Household Scenarios for the European Union, 1995-2025" in Hullen G. (eds), Living Arrangements and Households – Methods and results of Demography Projection*,* Bundesinstituut für Bevölkerungsforschung.

Becker, G. (1991), *A Treatise on the Family Cambridge*, Harvard, University Press.

Billari, F**.** (2005), "Partnership, Childbearing and Parenting: Trends of the 1990s", in Mancura M., Mac Donald A. L. And Haug W. (Eds), *The New Demographic Regime, Population Challenges and Policy Responses*, Geneva, United Nation.

Bongaarts J., (2001), "Household size and composition in the developing world in the 1990s*", Population Studies* vol 55, pp. 263-279.

Burch, T.K. (1970), "Some demographic Determinants of average household size: analytic approach", *Demography*, 4, pp. 61-69.

Butmpas L., Sweet, J. And Martin, T.C. (1990), "Changing patterns of remarriage" *Journal of Marriage and the Family*, vol 52, pp. 747-756.

Festy P. (2008), *"*Intergenerational Relations in Europe" Proceedings of the XLIV Scientific Meeting, Riunione Scientifica della Società Italiana di Statistica Università della Calabria 25-27 giugno 2008.

Golini A., Iaconucci R. (2003), "Tendenze demografiche e Rapporti fra le generazioni" *Analisi e problemi dell'Invecchiamento della Popolazione*, Collana a cura di Antonio Golini, Roma.

Jiang L. And O'neill B.C. (2004), "Toward A New Model For Probabilistic Household Forecast"International Statistical Review, vol. 72 (1),51-64.

Jennings V., Lloyd-Smith B., Ironmonger D. (1999), "Household Size And The Poisson Distribution" *Journal of the Australian Population Association,* vol.16. nos.1/2 .

Jennings V., Lloyd-Smith B., Ironmonger D. (2004), "Global Projections of Household Numbers using Age determined Ratios", Department of Economics Research Paper Number 914, The University of Melbourne.

Maccheroni C., Barugola T., Diale G., Ferraresi P.M. (2006), *Implications of demographic change in enlarged EU on patterns of saving and consumptions and in related consumer's behaviour*, Download:

http://ec.europa.eu/employment_social/social_situation/docs/walter_consumption_fin_rep_en.pdf.

OECD (2007), *Glossary of Statistical Terms,* Download: http://stats.oecd.org/glossary/download.asp.

ONU (1973)*, Methods of Projecting Households and Families*, Manual VII, New York.

Rowland D.T. (1991), *Ageing in Australia*, Longman Cheshir, Melbourne.

Van De Kaa (1987), "Europe's Second Demographic Transition", *Population Bulletin*, vol 42(1), pp. 1-59.

United Nations (1995), *Demographic Yearbook*, New York, United Nations.

Zeng Y. Vapuel J.W., Zhenglian W. (1998), "Household Projection Using Conventional demographic data", *Population and Development Review*, 24 (supplement) pp. 59-87.

Prime esperienze nel recupero di informazioni sulla mortalità neonatale mediante integrazione di dati amministrativi[1]

Cristiano Marini[2], *Alessandra Nuccitelli*[3]

Sommario

I tassi di mortalità neonatale per classe di età gestazionale e di peso alla nascita rappresentano importanti indicatori di descrizione delle condizioni di salute materno-infantile e di qualità delle cure.
Tuttavia, per effetto dell'introduzione delle leggi sulla semplificazione amministrativa e sulla privacy, questi tassi non vengono più calcolati in Italia dal 1999.
Obiettivo principale del lavoro è valutare la possibilità di recuperare informazioni sulla mortalità neonatale mediante abbinamento esatto dei records relativi ai nati vivi con i records relativi ai deceduti entro il primo mese di vita, con riferimento alle coorti dei nati in Italia negli anni 2003 e 2004.
Da un punto di vista strettamente metodologico, vengono discusse alcune criticità di un approccio spesso seguito per la scelta dei records da considerare effettivamente abbinati, quando occorra tenere conto di vincoli di compatibilità tra coppie.

Abstract

Neonatal mortality rates by gestational age and birth weight category are important indicators of maternal and child health and care quality.
However, due to new laws on administrative simplification and privacy, these specific rates have not been calculated in Italy since 1999.
The main aim of this work is to assess the possibility of retrieving information on neonatal mortality by the linkage between records related to live births and records related to infant deaths within the first month of life, with reference to 2003 and 2004 birth cohorts in Italy.
From a strict methodological point of view, some critical aspects of the most used record linkage approach are highlighted in the paper: specific problems may arise from the choice of records to be linked if there are consistency constraints between pairs.

Parole chiave: algoritmi *greedy*, età gestazionale, peso alla nascita, programmazione lineare, *record linkage*

---

[1] Il lavoro è frutto della collaborazione tra gli autori. I paragrafi 2, 3, 6 e 7 sono stati redatti da Cristiano Marini; i paragrafi 4 e 5 da Alessandra Nuccitelli; i paragrafi 1 e 8 sono stati redatti congiuntamente da Cristiano Marini e Alessandra Nuccitelli.

Gli autori ringraziano Pietro Barbieri dell'Azienda Ospedaliera di Melegnano (Milano), Cristina Mazzali e Kathleen K. Thoburn per i preziosi suggerimenti forniti nell'utilizzo del *software* Link Plus per il *record linkage* probabilistico.

[2] Titolare di assegno di ricerca (Università degli Studi di Roma 'La Sapienza'), e-mail: cristiano.marini@uniroma1.it.

[3] Ricercatore (Istituto Nazionale di Statistica), e-mail: nuccitel@istat.it.

1. Introduzione

I tassi di mortalità neonatale per classe di età gestazionale e di peso alla nascita rappresentano importanti indicatori di descrizione delle condizioni di salute materno-infantile (Zeitlin *et al.*, 2003) e di qualità delle cure (Horbar, 1999).[4]

Tuttavia, per effetto dell'introduzione delle leggi sulla semplificazione amministrativa e sulla *privacy*, questi tassi specifici non vengono più calcolati in Italia dal 1999.

Dal 2002, il flusso informativo dei *Certificati Di Assistenza al Parto* (d'ora in poi, CeDAP) – contenenti fondamentali informazioni sulla salute materno-infantile, come l'età gestazionale e il peso alla nascita – fa capo al Ministero della Salute che, annualmente, trasmette all'Istat copia dell'archivio nazionale privo di elementi identificativi diretti (CISIS, 2004; Buratta *et al.*, 2004). In tal modo, l'integrazione – a livello di individuo – delle informazioni rilevate dall'*Indagine sulle cause di morte* con quelle di interesse provenienti dai CeDAP risulta particolarmente difficoltosa.

Obiettivo principale del lavoro è valutare la possibilità di recuperare informazioni sulla mortalità neonatale mediante abbinamento esatto[5] dei *records* relativi ai nati vivi con i *records* relativi ai deceduti entro il primo mese di vita appartenenti alle coorti di nati considerate. Tale possibilità viene esplorata ricorrendo ad un abbinamento di tipo deterministico, attraverso l'utilizzo di variabili comuni alle due fonti di dati, con riferimento alle coorti dei nati vivi in Italia negli anni 2003 e 2004.

Il lavoro è strutturato come segue. Le caratteristiche principali delle fonti di dati prese in esame sono descritte sinteticamente nel paragrafo successivo; alcune valutazioni sulla qualità dei valori assunti dalle variabili comuni alle fonti e utilizzate per il *record linkage* sono riportate nel paragrafo 3. Dopo una breve introduzione ad alcuni aspetti generali delle tecniche di abbinamento esatto (paragrafo 4), nel paragrafo 5 si approfondiscono le scelte metodologiche operate nel contesto applicativo in esame. In particolare, sono discussi i risultati principali delle sperimentazioni condotte per valutare l'efficacia del criterio di abbinamento adottato (sottoparagrafo 5.2). Inoltre, vengono discusse alcune criticità di un approccio spesso seguito per la scelta dei *records* da considerare effettivamente abbinati, quando occorra tenere conto di vincoli di compatibilità tra accoppiamenti (sottoparagrafo 5.3); come alternativa, viene proposto, più coerentemente con le finalità proprie dell'abbinamento, un approccio basato su un algoritmo che in ricerca operativa è chiamato di tipo *greedy* (vorace). Nel paragrafo 6 sono riportati i risultati principali di un abbinamento eseguito, in via preliminare, a livello nazionale; alla luce di tali risultati e delle valutazioni sulla qualità dei dati, il recupero di informazioni sulla mortalità neonatale per classe di età gestazionale e di peso alla nascita viene circoscritto all'Italia settentrionale. Nel tentativo di dare un'idea dell'evoluzione del fenomeno almeno per questa ripartizione geografica, sono fornite le stime dei tassi per il 2003, primo anno utile da quando ne è stata interrotta la produzione (paragrafo 7). Infine (paragrafo 8), vengono tratte alcune considerazioni conclusive in merito alle problematiche legate alla qualità e all'utilizzazione dei risultati del processo di abbinamento nel contesto applicativo esaminato.

[4] Per 'morte neonatale' si intende la morte di un bambino che nasce vivo, quale che sia la durata della gravidanza, e che decede nel periodo neonatale (un mese dalla nascita). Il tasso di mortalità neonatale esprime il numero di morti entro il primo mese di vita ogni 1.000 nati vivi.

[5] Il termine 'abbinamento esatto' (o '*record linkage*') si riferisce all'uso di tecniche algoritmiche per identificare *records* relativi ad una stessa unità statistica contenuti in uno o più archivi (per un'introduzione si veda, ad esempio, Herzog *et al.*, 2007).

2. Le fonti di dati utilizzate

2.1 L'informazione sulle nascite in Italia

Per oltre 70 anni (dal 1926 al 1997) la più ampia fonte di informazione statistica sulle nascite in Italia è stata rappresentata dalla *Rilevazione delle nascite,* effettuata dall'Istat e realizzata mediante la compilazione di un apposito modello statistico – Istat D.1 e Istat D.2, rispettivamente, per nato di sesso maschile e nato di sesso femminile – da parte degli ufficiali di stato civile del comune di evento.

Il modello era concettualmente strutturato in tre parti:

- *notizie di stato civile* (tratte dall'atto di nascita);
- *notizie demo-sanitarie* (tratte dal CeDAP, compilato dall'ostetrica);
- *notizie socio-demografiche* (fornite dal dichiarante).

In seguito, l'entrata in vigore della Legge n° 127 del 15 maggio 1997 – detta 'Bassanini-bis' – ha modificato la procedura di rilevazione delle nascite, offrendo al cittadino la possibilità di dichiarare la nascita presso l'ufficio di stato civile del comune di residenza oppure presso il centro in cui avveniva il parto. Tuttavia, in tal modo è stato impedito il riferimento delle nascite alla popolazione presente, come invece accadeva in passato.

L'elemento che ha posto fine al processo di produzione di dati individuali di fonte stato civile, contenenti notizie demo-sanitarie sulle nascite, è contenuto nel paragrafo 2 dell'art. 8 del regolamento di attuazione "Riservatezza dei dati contenuti nei documenti acquisiti dalla pubblica amministrazione" (DPR n° 403 del 20 ottobre 1998). Questo regolamento, da un lato stabilisce il divieto ai direttori sanitari di trasmettere il CeDAP all'ufficiale di stato civile e a quest'ultimo di richiederlo, dall'altro dispone che il rilevamento a fini statistici delle nascite debba avvenire tramite l'invio dei CeDAP, resi anonimi, da parte dei direttori sanitari agli uffici competenti del SISTAN (SIstema STAtistico Nazionale), secondo le modalità stabilite dal Ministero della Salute e dall'Istat.

Dunque, l'impossibilità per l'ufficiale di stato civile di acquisire le informazioni direttamente dai CeDAP ha comportato, dal 1° gennaio 1999, la sospensione della *Rilevazione delle nascite* (Prati *et al*., 2006).

Il tentativo di colmare il vuoto informativo così creatosi ha richiesto del tempo e ancora oggi le informazioni sulle nascite desumibili indirettamente dai CeDAP non riescono a coprire l'intero territorio nazionale. La rilevazione dei CeDAP – istituita con Decreto Ministeriale n° 349 del 16 luglio 2001 – è stata avviata con carattere sperimentale da parte del Ministero della Salute solo a partire dal 2002.

Il modello utilizzato attualmente per la rilevazione dei CeDAP è strutturato in sei parti:

- *sezione generale;*
- *informazioni socio-demografiche sui genitori;*
- *informazioni sulla gravidanza;*
- *informazioni sul parto e sul neonato;*
- *informazioni sulle cause di nati-mortalità;*
- *informazioni sulla presenza di malformazioni.*

Come si evince dall'ultima riga di tavola 1, nel corso del triennio 2002-2004, la copertura della rilevazione a livello nazionale è aumentata sensibilmente, soprattutto nel 2003, secondo anno di raccolta. Tuttavia, nel 2004, il rapporto tra numero di CeDAP e numero di parti rilevati attraverso le *Schede di Dimissione Ospedaliera* (SDO) è aumentato

di appena 2,5 punti percentuali rispetto all'anno precedente, sicché è passato da meno del 70 per cento nel 2002 a 86,0 per cento nel 2004.

A livello regionale il livello di copertura ha evidenziato forti disomogeneità, con una tendenza a decrescere da Nord a Sud. Tuttavia, nel 2004 (colonne 4 e 7 di tavola 1) alcune regioni del Mezzogiorno (Sardegna, Abruzzo, Puglia, Basilicata) hanno fatto segnare un notevole incremento della copertura rispetto al primo anno di rilevazione, la Sicilia ha mostrato ancora forti criticità, mentre altre regioni (Molise e Calabria) non avevano ancora istituito un sistema informativo dedicato alla raccolta dei dati. Questa problematica ha riguardato anche la Provincia Autonoma di Bolzano, che però ha avviato la rilevazione dei CeDAP a partire dal 2005.

Tavola 1 - CeDAP pervenuti al Ministero della Salute e rapporto (in termini percentuali) tra numero di CeDAP e numero di parti rilevati attraverso le SDO, per regione e anno di nascita

Regione	CeDAP			CeDAP su n° di parti rilevati attraverso le SDO (%)		
	2002	2003	2004	2002	2003	2004
Piemonte	32.566	34.508	35.110	94,5	97,8	98,2
Valle d'Aosta/Vallée d'Aoste	1.041	1.110	1.108	99,6	98,9	100,1
Lombardia	45.933	83.031	83.481	52,4	92,5	90,6
Bolzano/Bozen (a)	….	….	….	….	….	….
Trento	*4.873*	*4.820*	*5.122*	*101,3*	*101,0*	*100,9*
Veneto	42.273	44.017	45.583	98,6	100,4	100,0
Friuli Venezia Giulia	9.492	9.887	10.071	100,4	100,6	100,9
Liguria	7.966	9.723	8.007	70,4	86,9	68,6
Emilia Romagna	30.705	33.930	35.828	86,9	94,4	95,7
Toscana	27.863	28.790	30.112	94,8	95,7	97,6
Umbria	6.752	7.060	7.620	94,6	93,7	97,5
Marche	7.683	12.091	12.557	59,3	92,2	94,1
Lazio	48.759	51.126	52.021	100,2	102,1	103,0
Abruzzo	5.396	7.017	9.567	52,9	69,5	89,0
Molise	….	….	….	….	….	….
Campania	56.888	58.220	62.710	90,2	89,5	97,0
Puglia	23.707	38.936	38.884	58,0	96,8	96,5
Basilicata	3.282	3.932	4.405	68,5	83,8	90,5
Calabria	….	….	….	….	….	….
Sicilia	12.753	18.408	20.819	26,2	36,9	40,8
Sardegna	….	6.378	11.888	….	49,4	94,0
Italia	**367.932**	**452.984**	**474.893**	**69,4**	**83,5**	**86,0**

Fonte: Ministero della Salute (2007)

(a) dal momento che i dati relativi alla Provincia Autonoma di Bolzano sono mancanti, non viene riportato il totale per la regione Trentino-Alto Adige

Come si può vedere dalla tavola, nell'ultimo anno preso in esame, diminuzioni del grado di copertura hanno interessato anche alcune regioni dell'Italia settentrionale (Lombardia e Liguria). Le maggiori criticità hanno riguardato soprattutto la Liguria, per la quale, nel 2004, il rapporto tra numero di CeDAP e numero di parti rilevati attraverso le SDO è diminuito di quasi 20 punti percentuali, rispetto al 2003, risultando inferiore anche al valore del 2002.

2.2 L'informazione sulle morti neonatali in Italia

La fonte dei dati sull'ammontare delle morti neonatali che annualmente si verificano sul territorio nazionale è costituita dall'*Indagine sulle cause di morte* condotta dall'Istat.

Le schede relative alle morti neonatali nel primo mese di vita rappresentano un sottoinsieme delle schede di morte Istat D.4 bis e Istat D.5 bis (modelli utilizzati per la rilevazione delle morti nel primo anno di vita, rispettivamente, per nato di sesso maschile e nato di sesso femminile).

Rimangono esclusi dal campo di osservazione dell'indagine i decessi di individui residenti in Italia avvenuti fuori dai confini italiani; comunque, tali decessi, nel caso di bambini con meno di un mese di vita, sono da considerarsi nulli o rari, per cui si può supporre una copertura pressoché totale degli eventi di interesse.

La statistica annuale sulle cause di morte fu avviata nel 1881 e fino al 1886 venne eseguita solo per alcuni limitati territori. Dal 1887 l'indagine fu estesa a tutti i comuni del Regno e successivamente a quelli dello Stato, nell'ambito dei confini italiani dell'epoca.

La scheda di morte nel primo anno di vita consta di due parti: la parte A, 'sanitaria', è compilata dal medico curante o dal necroscopo; la parte B, 'socio-demografica', è compilata dall'ufficiale di stato civile e le notizie rilevate riguardano soprattutto i genitori.

Il modello viene riprodotto in duplice copia. Una copia è inviata all'Azienda Sanitaria Locale in cui è avvenuto il decesso, l'altra copia è inoltrata alle prefetture e agli uffici regionali dell'Istat per un controllo quantitativo preliminare delle schede (Bruzzone, 2006). L'ufficio centrale dell'Istituto Nazionale di Statistica si occupa poi di tutte le varie fasi di lavorazione dei dati (registrazione, codifica, revisione, correzione e diffusione).

3. Qualità dei dati utilizzati per l'abbinamento

3.1 Le variabili di abbinamento

Per semplicità, siano A e B due archivi, costituiti rispettivamente da n_A e n_B *records*. Ogni *record* si riferisce ad un individuo ed è costituito da più campi o variabili.

Generalmente le procedure utilizzate per l'abbinamento esatto si basano sul confronto delle modalità (o valori) assunte da un sottoinsieme di variabili comuni – dette *variabili di abbinamento* – agli archivi da integrare.

Tali variabili, non solo possono avere un potere identificativo diverso nei confronti degli individui, ma possono risultare affette da errori in misura differente tra loro: ad esempio, un campo come il *sesso*, presentando due modalità, contribuisce poco all'individuazione univoca di una persona, rispetto ad un campo quale il *comune di nascita* che, tuttavia, può essere riportato più spesso in maniera errata.

Inoltre, nel corso del tempo alcune variabili possono assumere modalità diverse per uno stesso individuo, rendendo più arduo il riconoscimento di *records* a questo riferibili (ad esempio, si pensi al *comune di residenza*).

Le variabili comuni agli archivi relativi alle schede di morte e ai CeDAP e utilizzate per l'abbinamento sono le seguenti:

- *sesso del nato*;
- *genere del parto*;[6]
- *giorno di nascita del nato*;
- *mese di nascita del nato*;
- *comune di nascita del nato*;
- *provincia di nascita del nato*;
- *giorno di nascita della madre*;
- *mese di nascita della madre*;
- *anno di nascita della madre*;
- *comune di residenza della madre*;
- *provincia di residenza della madre*;
- *cittadinanza della madre*.

Si fa presente che, nel contesto applicativo in esame, le variabili relative alla residenza e alla cittadinanza della madre possono essere considerate pressoché immutabili nel tempo, visto il periodo di tempo che intercorre tra le rilevazioni della nascita e della morte di uno stesso individuo (al più, un mese).

Nei sottoparagrafi 3.2 e 3.3 sono riportate alcune valutazioni sulla qualità dei valori assunti dalle variabili precedenti nei due archivi da integrare, misurata in termini di incidenza delle risposte mancanti o inammissibili, rispettivamente sul totale delle schede di morte nel primo mese di vita e sul totale dei CeDAP, con riferimento alle coorti dei nati vivi nel 2003 e nel 2004.[7]

3.2 Qualità delle schede di morte nel primo mese di vita

La qualità delle schede di morte relative ai nati vivi negli anni 2003 e 2004 non risulta, nel complesso, particolarmente elevata (tavola 2).

Le variabili maggiormente problematiche, per cui si osserva anche un peggioramento della qualità nel biennio preso in esame, sono quelle relative alla data di nascita della madre (*giorno di nascita della madre*, *mese di nascita della madre*, *anno di nascita della madre*) e *genere del parto*. In particolare, a livello nazionale la percentuale dei casi in cui la data di nascita della madre è mancante o inammissibile passa da 37 per cento, per le schede di morte dei nati nel 2003, a 42 per cento, per le schede di morte dei nati nel 2004. Inoltre, se per i deceduti nati nel 2003 *genere del parto* assume soltanto due valori non validi, per i deceduti nati nel 2004 il numero dei casi non validi per la stessa variabile sale a 465 (pari a oltre il 30 per cento del totale delle schede di morte nel primo mese di vita).

Per entrambi gli anni considerati, le differenze più rilevanti tra l'Italia nel suo complesso e il Nord riguardano la data di nascita della madre: per i deceduti nati nel 2003, la frequenza dei valori mancanti o inammissibili per il Nord risulta inferiore di quasi 7 punti percentuali rispetto a quella osservata a livello nazionale, differenza che diventa ancora più accentuata per i deceduti nati nel 2004 (circa 11 punti percentuali).

[6] I valori assunti da tale variabile indicano se trattasi di parto semplice o parto plurimo.

[7] Non disponendo di un sottoinsieme di *records* per i quali si conosca il vero valore assunto dalle variabili di abbinamento, non è possibile stimare, sulla base dei dati disponibili, l'incidenza di altre tipologie di errore.

Infine, l'incremento di casi non validi per la variabile *genere del parto* riguarda il Nord in misura relativamente minore rispetto al resto del Paese.

Tavola 2 - Valori mancanti o inammissibili (frequenze assolute e percentuali sul totale delle risposte) per le variabili di abbinamento nell'archivio delle schede di morte nel primo mese di vita in Italia e in Italia settentrionale, per anno di nascita del nato

variabile di abbinamento	valori mancanti o inammissibili							
	2003				2004			
	Italia		Nord Italia		Italia		Nord Italia	
	fr. ass.	*%*	*fr. ass.*	*%*	*fr. ass.*	*%*	*fr. ass.*	*%*
sesso del nato	-	-	-	-	-	-	-	-
genere del parto	2	0,13	1	0,18	465	30,45	129	24,57
giorno di nascita del nato	-	-	-	-	-	-	-	-
mese di nascita del nato	-	-	-	-	-	-	-	-
comune di nascita del nato	23	1,51	5	0,88	-	-	-	-
provincia di nascita del nato	14	0,92	3	0,53	-	-	-	-
giorno di nascita della madre	563	36,85	171	30,00	640	41,91	163	31,04
mese di nascita della madre	563	36,85	171	30,00	640	41,91	163	31,04
anno di nascita della madre	558	36,52	169	29,65	640	41,91	163	31,04
comune di residenza della madre	92	6,02	34	5,96	2	0,13	-	-
provincia di residenza della madre	84	5,50	34	5,96	2	0,13	-	-
cittadinanza della madre	60	3,93	18	3,16	14	0,92	6	1,14

Fonte: elaborazione propria su dati Istat (Indagine sulle cause di morte)

3.3 Qualità dei Certificati Di Assistenza al Parto

Con riferimento all'archivio dei CeDAP, negli anni 2003 e 2004 l'incidenza relativa delle risposte mancanti o inammissibili appare nel complesso più contenuta, rispetto a quanto emerso per le schede di morte.

Da un confronto dei CeDAP del 2004 con quelli del 2003 (tavola 3) emerge, a livello nazionale, un certo miglioramento della qualità di alcune variabili di abbinamento. In particolare, per *genere del parto* e *cittadinanza della madre* i valori mancanti o inammissibili scendono, rispettivamente, da 12,17 e 3,25 per cento nel 2003 a 0,94 e 1,69 per cento nel 2004; per le variabili relative al luogo di nascita (*comune di nascita del nato* e *provincia di nascita del nato*) i valori mancanti o inammissibili tendono a scomparire, passando da 0,22 e 0,23 per cento nel 2003 a 0,01 per cento nel 2004.

Nell'ultimo anno di rilevazione, invece, peggiora leggermente la qualità di quasi tutte le altre variabili anagrafiche relative alla madre e, in misura maggiore, quella della variabile *mese di nascita del nato*, per la quale l'incidenza dei valori mancanti o inammissibili sul totale dei CeDAP passa da 0,16 a 2,62 per cento. Infine, rimane pressoché costante, anche se relativamente bassa, la frequenza di certificati non validi per *sesso del nato* e *giorno di nascita del nato*.

Con specifico riferimento al Nord Italia, si rileva una maggiore tendenza al peggioramento della qualità delle variabili relative alla madre, rispetto a quanto osservato a livello nazionale; l'incidenza dei valori mancanti o inammissibili sul totale delle risposte per le variabili relative al nato appare, invece, più contenuta.

Tavola 3 - Valori mancanti o inammissibili (frequenze assolute e percentuali sul totale delle risposte) per le variabili di abbinamento nei CeDAP in Italia e in Italia settentrionale, per anno di nascita del nato

variabile di abbinamento	valori mancanti o inammissibili							
	2003				2004			
	Italia		Nord Italia		Italia		Nord Italia	
	fr. ass.	*%*	*fr. ass.*	*%*	*fr. ass.*	*%*	*fr. ass.*	*%*
sesso del nato	598	0,13	263	0,12	530	0,11	143	0,06
genere del parto	55.138	12,17	39	0,02	4.447	0,94	324	0,14
giorno di nascita del nato	558	0,12	153	0,07	313	0,07	147	0,07
mese di nascita del nato	738	0,16	157	0,07	12.464	2,62	147	0,07
comune di nascita del nato	1.019	0,22	36	0,02	44	0,01	21	0,01
provincia di nascita del nato	1.022	0,23	39	0,02	44	0,01	20	0,01
giorno di nascita della madre	2.466	0,54	1.450	0,66	2.663	0,56	1.997	0,89
mese di nascita della madre	2.743	0,61	1.484	0,67	2.892	0,61	2.015	0,90
anno di nascita della madre	3.157	0,70	1.629	0,74	6.602	1,39	2.177	0,97
comune di residenza della madre	4.925	1,09	1.586	0,72	7.077	1,49	3.565	1,59
provincia di residenza della madre	6.203	1,37	2.227	1,01	8.128	1,71	4.040	1,80
cittadinanza della madre	14.710	3,25	2.812	1,27	8.002	1,69	4.790	2,14

Fonte: elaborazione propria su dati Ministero della Salute (Rilevazione dei CeDAP)

4. Aspetti generali delle tecniche di abbinamento esatto

Per introdurre in modo generale il problema dell'abbinamento esatto, siano X_1, X_2, …, X_k le k variabili di abbinamento considerate nei due archivi A e B.

Lo spazio prodotto $A \times B = \{(a,b); a \in A, b \in B\}$, costituito da tutte le $N = n_A \times n_B$ possibili coppie di *records* originate dal confronto tra i due archivi, è l'unione dei seguenti insiemi disgiunti:

- l'insieme M delle coppie i cui elementi si riferiscono allo stesso individuo (insieme dei *matches*);
- l'insieme U delle coppie i cui elementi si riferiscono ad individui differenti (insieme dei *non-matches*).

In una procedura di *record linkage* i risultati del confronto tra i valori assunti dalle variabili di abbinamento nei due archivi sono utilizzati per classificare ogni coppia di *records* come appartenente a M o U.

Nel seguito, sarà utilizzata la denominazione *link*/*non-link* per riferirsi all'etichetta assegnata definitivamente ad una coppia alla fine di una qualsiasi procedura di abbinamento e la denominazione *match*/*non-match* per indicare la vera (ma incognita) condizione di appartenenza di una coppia all'insieme M o U. In questo contesto, si possono commettere i seguenti due tipi di errore:

1) coppie appartenenti all'insieme M possono essere erroneamente etichettate come *non-links*;
2) coppie appartenenti all'insieme U possono essere erroneamente etichettate come *links*.

La scelta di un metodo di abbinamento dovrebbe essere legata alla valutazione della gravità relativa che si attribuisce ai due tipi di errore conseguenti al processo di classificazione delle coppie.

Il confronto tra i valori assunti dalle variabili di abbinamento X_1, X_2, …, X_k per la coppia (a,b), con $a \in A$ e $b \in B$, può essere espresso nel modo seguente:

$$\mathbf{y}_{ab} = f\left(x^A_{a,1},\ x^A_{a,2},\ \ldots,\ x^A_{a,k};\ x^B_{b,1},\ x^B_{b,2},\ \ldots,\ x^B_{b,k}\right). \tag{4.1}$$

Questa funzione misura la diversità tra i valori delle variabili di abbinamento nei due individui posti a confronto. In linea di principio, le coppie in *M* dovrebbero presentare bassi livelli di diversità, mentre livelli di diversità più elevati dovrebbero essere associati alle coppie in *U*. La capacità discriminante di questa funzione rappresenta, quindi, un elemento importante, che può influire anche notevolmente sulla qualità dei risultati di una procedura di abbinamento.

Il modo più semplice e utilizzato di definire la funzione (4.1) consiste nel verificare esclusivamente se i *records* di ogni coppia presentano la stessa modalità di una variabile di abbinamento oppure no. Formalmente, in questo caso il risultato del confronto per la coppia (a,b) è un vettore di *k* elementi 0 o 1:

$$\mathbf{y}_{ab} = \left(y_{ab,1},\ y_{ab,2},\ \ldots,\ y_{ab,k}\right) \tag{4.2}$$

dove

$$y_{ab,j} = \begin{cases} 1 & \text{se } x^A_{a,j} = x^B_{b,j} \\ 0 & \text{altrimenti} \end{cases} \qquad j = 1,\ 2,\ \ldots,\ k\ .$$

Con la parola 'altrimenti' si intende sia il caso in cui $x^A_{a,j} \neq x^B_{b,j}$, sia il caso in cui almeno uno dei due valori $x^A_{a,j}$ e $x^B_{b,j}$ sia mancante.

Una volta definito il vettore di confronto, rimane da stabilire come questo possa essere utilizzato per la classificazione delle coppie in *M* o *U*.

Una possibilità consiste nell'assegnare al vettore $\mathbf{y}_{ab}$ un 'peso' w_{ab}, sul cui valore basare il processo decisionale per la coppia (a,b):

a) se $w_{ab} \geq t_u$	la coppia (a,b) viene etichettata come *link*	
b) se $t_l \leq w_{ab} < t_u$	si procede ad un'ispezione manuale dei *records* per decidere se la coppia (a,b) è un *link* o un *non-link*	(4.3)
c) se $w_{ab} < t_l$	la coppia (a,b) viene etichettata come *non-link*.	

La stima dei pesi w_{ab}, $(a,b) \in A \times B$, e la scelta dei valori di t_l e t_u $(t_l \leq t_u)$ sono cruciali nella definizione di una procedura di *record linkage*. Nel caso più semplice, si può utilizzare un criterio deterministico, fissando a priori i valori dei pesi e delle soglie in funzione degli obiettivi specifici dell'abbinamento. L'intervallo tra t_l e t_u non deve essere,

inoltre, troppo ampio, in quanto la scelta di rinviare la decisione, essendo associata ad un controllo manuale delle coppie di *records*, presenta solitamente costi elevati.

Un semplice esempio di criterio deterministico consiste nell'associare i pesi al numero di concordanze[8] osservate tra valori corrispondenti delle variabili di abbinamento; la decisione di abbinare può riguardare, ad esempio, tutte le coppie di *records* con al più una discordanza, con una soglia unica implicita pari a $k-1$ nel processo decisionale (4.3).

Nella regola di decisione formulata da Fellegi e Sunter (1969), i pesi sono definiti in modo probabilistico, mediante il rapporto tra le due verosimiglianze del vettore dei confronti, rispettivamente nel caso di coppie relative ad uno stesso individuo e nel caso di coppie relative ad individui differenti:

$$w_{ab} = ln\left(\frac{P\left(\mathbf{y}_{ab}\middle|(a,b)\in M\right)}{P\left(\mathbf{y}_{ab}\middle|(a,b)\in U\right)}\right) \qquad (a,b)\in A\times B\,. \qquad (4.4)$$

Il problema principale nell'applicazione della regola di decisione proposta da Fellegi e Sunter è rappresentato dal fatto che le probabilità al numeratore e al denominatore del rapporto (4.4) non sono note e, pertanto, è necessario ricorrere ad una loro stima.

La maggior parte dei *software* attualmente disponibili per l'abbinamento esatto con metodi probabilistici utilizzano l'algoritmo[9] *EM* (*Expectation-Maximization*) per la stima dei pesi (Jaro, 1989; Winkler, 2000); i valori di t_l e t_u sono spesso determinati in modo empirico, esaminando la distribuzione dei pesi così stimati. Inoltre, alcuni *software* offrono la possibilità di aggiustare i pesi stimati con l'algoritmo *EM* utilizzando le frequenze con cui le modalità delle variabili di abbinamento si presentano, al fine di tenere conto del diverso potere identificativo delle modalità stesse.[10]

Un ulteriore elemento da prendere in considerazione in una procedura di abbinamento riguarda l'opportunità di ridurre il numero di coppie di *records* da analizzare, che può risultare molto elevato, anche per archivi di dimensione non eccessivamente grande.

In molte situazioni, una buona strategia consiste nel limitare i confronti all'interno di blocchi di *records* che presentano concordanza tra modalità corrispondenti di una o più variabili di abbinamento. In questo contesto, le variabili di abbinamento sono chiamate *variabili di formazione dei blocchi* o, più sinteticamente, *variabili di bloccaggio*.

Le coppie per cui non viene osservata alcuna concordanza tra modalità corrispondenti

[8] Occorre precisare che viene osservata una concordanza quando i valori (o modalità) di una variabile di abbinamento nei due *records* posti a confronto risultano coincidenti e non mancanti.

[9] In generale, l'algoritmo *EM* consente di ottenere la stima di massima verosimiglianza dei parametri incogniti di una distribuzione, in presenza di dati incompleti, secondo il seguente schema: si imputano i dati mancanti in base ad una stima iniziale dei parametri incogniti (passo di *Expectation*) e si procede ad una stima dei parametri incogniti basata sia sui dati effettivamente osservati che su quelli appena imputati (passo di *Maximization*), iterando il procedimento fino a quando le stime non subiscono più cambiamenti significativi (per maggiori dettagli si rinvia a Dempster *et al.*, 1977). Nel contesto della stima dei pesi (4.4), i valori $\mathbf{y}_{ab}$ costituiscono i dati effettivamente osservati, mentre il ruolo di dati mancanti è svolto dall'informazione relativa alla vera condizione di appartenenza di ogni coppia all'insieme *M* o *U*.

[10] Ad esempio, con riferimento ad una variabile di abbinamento quale il *comune di nascita del nato*, una concordanza osservata su un comune di piccola ampiezza demografica consente una più facile individuazione dei *records* relativi ad uno stesso individuo, rispetto ad una concordanza osservata su un comune di ampiezza maggiore.

delle variabili di bloccaggio sono, quindi, implicitamente etichettate come *non-links*. Se le variabili di bloccaggio prese in considerazione non sono affette da errori, tale strategia consente di ottenere una notevole riduzione del carico computazionale e, allo stesso tempo, una forte protezione contro gli errati abbinamenti.

4.1 Il problema del rispetto dei vincoli di compatibilità tra coppie di *records*

Occorre tener conto del fatto che il processo di decisione (4.3), riguardando una coppia alla volta, può condurre a risultati incompatibili, se gli abbinamenti consentiti dal contesto applicativo specifico sono del tipo 'uno a uno' – ovvero quando ad ogni individuo corrisponde al più un solo *record* in ciascuno degli archivi da abbinare – oppure del tipo 'uno a molti' (o 'molti a uno') – quando ad ogni individuo possono corrispondere più *records* in uno degli archivi da abbinare.

Più precisamente, con riferimento ad un contesto applicativo che consenta soltanto abbinamenti del tipo 'uno a uno', un *record* in A può essere accoppiato con al più un *record* in B (e viceversa); tuttavia, ad esempio, può accadere che entrambe le coppie (a,b) e (a,b') presentino un peso superiore al valore soglia t_u, per cui il *record* $a \in A$ risulterebbe abbinato sia a $b \in B$ che a $b' \in B$, violando i vincoli imposti dal contesto applicativo.

Per ovviare a questo inconveniente, solitamente il problema della scelta delle coppie da etichettare effettivamente come *links* viene ricondotto ad un problema di programmazione lineare in cui lo schema ottimo di assegnazione dei *records* è quello per cui la somma dei pesi è massima (Jaro, 1989).

Formalmente, nel caso di abbinamenti del tipo 'uno a uno', viene risolto il problema seguente:[11]

$$\text{massimizzare} \quad \sum_{a=1}^{n_A}\sum_{b=1}^{n_B} w_{ab} z_{ab}$$

$$\text{sotto i vincoli:} \quad \sum_{a=1}^{n_A} z_{ab} \le 1 \qquad b = 1,\ 2,\ \ldots,\ n_B \qquad (4.5)$$

$$\sum_{b=1}^{n_B} z_{ab} \le 1 \qquad a = 1,\ 2,\ \ldots,\ n_A ,$$

dove

[11] Con riferimento ad un contesto applicativo che consenta abbinamenti del tipo 'uno a molti' (un *record* in A può essere accoppiato con più *records* in B, ma un *record* in B può essere accoppiato al più con un *record* in A), diminuiscono i vincoli del problema:

$$\text{massimizzare} \quad \sum_{a=1}^{n_A}\sum_{b=1}^{n_B} w_{ab} z_{ab}$$

$$\text{sotto i vincoli:} \quad \sum_{a=1}^{n_A} z_{ab} \le 1 \qquad b = 1,\ 2,\ \ldots,\ n_B .$$

$$z_{ab} = \begin{cases} 1 & \text{se il record } a \text{ è assegnato al record } b \\ 0 & \text{altrimenti} \end{cases} \qquad (a,b) \in A \times B \,.$$

Secondo la proposta di Jaro, una volta ottenuto lo schema ottimo di assegnazione dei *records*, solo le coppie con peso maggiore o uguale al valore soglia t_u facenti parte dello schema vanno poi etichettate come *links*.

5. Il processo di integrazione tra le fonti di dati

5.1 La strategia adottata per l'abbinamento

La possibilità di recuperare informazioni sulla mortalità neonatale secondo caratteristiche specifiche (età gestazionale e peso alla nascita) viene esplorata ricorrendo ad un abbinamento di tipo deterministico.

La base informativa su cui opera il processo di integrazione dei dati è costituita da quattro archivi distinti:

1. i *records* relativi alle schede di morte nel primo mese di vita in Italia – ad esclusione della Provincia Autonoma di Bolzano[12] – con riferimento alla coorte dei nati vivi nel 2003 (1.528);
2. i *records* relativi alle schede di morte nel primo mese di vita in Italia – ad esclusione della Provincia Autonoma di Bolzano – con riferimento alla coorte dei nati vivi nel 2004 (1.527);
3. i *records* relativi ai CeDAP dei nati vivi in Italia nel 2003 (452.984);
4. i *records* relativi ai CeDAP dei nati vivi in Italia nel 2004 (474.893).

Dal momento che solo una piccola percentuale dei nati vivi muore entro il primo mese di vita, gli archivi relativi a CeDAP e schede di morte risultano di dimensione molto diversa e il riconoscimento delle informazioni relative ad uno stesso individuo si rivela particolarmente difficoltoso.

L'abbinamento viene effettuato distintamente tra gli archivi 1 e 3 e tra gli archivi 2 e 4; pertanto, l'*anno di nascita del nato* (che può essere considerato esente da errori) costituisce implicitamente una variabile di bloccaggio.

Per ciascuna coppia di archivi – 1 e 3, 2 e 4 – l'abbinamento viene eseguito utilizzando come variabile di bloccaggio la *regione di nascita del nato*.[13] Per ciascuna coppia di blocchi di *records* per i quali la *regione di nascita del nato* assume la stessa modalità, vengono eseguiti in sequenza i sette passi descritti qui di seguito.

Nel prosieguo sarà indicato con *A* (o *B*) il generico blocco di *records* relativo alle schede di morte (o ai CeDAP) e con *k* il numero di variabili di abbinamento considerate (in questo contesto $k = 12$).

Passo 1. Abbinamento deterministico provvisorio

Un abbinamento provvisorio viene effettuato utilizzando per ogni coppia (a,b):

[12] Come già accennato nel sottoparagrafo 2.1, la Provincia Autonoma di Bolzano ha avviato la rilevazione dei CeDAP soltanto a partire dal 2005.

[13] La variabile *regione di nascita del nato* può essere ritenuta esente da errori.

- il vettore di confronto (4.2) costituito da *k* elementi 0 o 1, come funzione per misurare la diversità tra le modalità delle variabili di abbinamento in $a \in A$ e $b \in B$;
- il numero di concordanze osservate tra modalità corrispondenti delle variabili di abbinamento in $a \in A$ e $b \in B$, come peso w_{ab} associato al vettore di confronto.

La decisione di abbinare (provvisoriamente) riguarda tutte le coppie di *records* senza alcuna discordanza, ovvero con peso superiore al valore soglia $t_l = t_u = k$ nel processo decisionale (4.3).

Passo 2. Calcolo di frequenze

Vengono calcolate le frequenze con cui le varie modalità di ciascuna variabile di abbinamento si presentano nel blocco relativo ai CeDAP.

Nel seguito sarà denotata con

$$fr_{md(x_{b,j}^{B})} \qquad j = 1, \quad 2, \quad \ldots, \quad k \qquad (5.1)$$

la frequenza con cui la modalità assunta dalla variabile di abbinamento X_j nel *record* $b \in B$ si presenta nel blocco *B*.

Passo 3. Individuazione del caso di link 'perfetto più raro'

L'individuazione di questo particolare accoppiamento avviene sulla base dei risultati ottenuti al Passo 1 e delle frequenze (5.1). In questo contesto, il *link* 'perfetto più raro' è rappresentato dalla coppia di *records* per cui è minima la somma delle frequenze (5.1) rispetto alle *k* variabili di abbinamento, tra tutti gli accoppiamenti provvisori ottenuti al Passo 1.

Anche se il contesto applicativo in esame ammette soltanto abbinamenti del tipo 'uno a uno', accoppiamenti del tipo 'uno a molti' di uguale peso w_{ab}, riconducibili a nati vivi gemelli[14] di cui uno solo morto nel primo mese di vita, sono inevitabili. Infatti, per bambini nati da un unico parto, le informazioni utilizzate per l'abbinamento provenienti dai CeDAP solitamente coincidono.

Nel caso in cui il *link* 'perfetto più raro' sia del tipo appena descritto, la sua risoluzione avviene etichettando come *link* soltanto l'accoppiamento per cui il *record* relativo ai CeDAP si riferisce al nato con *peso alla nascita* minore, sfruttando i risultati di recenti studi sulla mortalità neonatale (Branum e Schoendorf, 2003).

Passo 4. Abbinamento deterministico - Definizione dei pesi e scelta delle soglie t_l e t_u

L'abbinamento vero e proprio avviene utilizzando, per ogni coppia (a,b):

- il vettore di confronto (4.2) costituito da *k* elementi 0 o 1, come funzione per misurare la diversità tra le modalità delle variabili di abbinamento in $a \in A$ e $b \in B$;
- una misura di similarità tra $a \in A$ e $b \in B$, basata sulle frequenze con cui le varie

[14] Considerando che per ogni bambino nato vivo viene compilato un CeDAP, l'informazione relativa al fatto che più bambini siano nati da un unico parto è deducibile dal codice identificativo dei CeDAP a loro relativi che, in questo caso, coincide a meno dell'ultima cifra.

modalità di ciascuna variabile di abbinamento si presentano nel blocco relativo ai CeDAP, come peso w_{ab} associato al vettore di confronto.

Più precisamente, il peso per la coppia (a,b) è dato da

$$w_{ab} = \frac{\sum_{j=1}^{k}\left[y_{ab,j} \times \left(n_B - fr_{md(x_{b,j}^B)} \right)\right]}{k \times n_B - \sum_{j=1}^{k} fr_{md(x_{r,j}^B)}}, \qquad (5.2)$$

avendo indicato con:

- n_B il numero di *records* nel blocco *B*;
- $fr_{md(x_{r,j}^B)}$ la frequenza con cui la modalità assunta dalla variabile di abbinamento X_j nel *record* $r \in B$, relativo al caso di *link* 'perfetto più raro' individuato al Passo 3, si presenta nel blocco *B*.

I pesi così definiti assumono generalmente[15] valori compresi tra 0 e 1. In particolare, il valore 0 viene assunto in corrispondenza di tutte le coppie senza alcuna concordanza tra modalità corrispondenti delle variabili di abbinamento, per cui la similarità tra i *records* è minima; invece, il valore massimo è raggiunto di solito nel caso di *link* 'perfetto più raro'.

Le soglie t_l e t_u sono determinate in modo empirico mediante ispezione della distribuzione dei pesi (5.2), individuando un valore al di sopra del quale una coppia di *records* possa essere ritenuta verosimilmente riferibile ad uno stesso individuo. Nel contesto in esame, i valori per t_l e t_u risultano scelti, rispettivamente, negli intervalli 0,45-0,55 e 0,55-0,65, con $t_u - t_l \leq 0{,}1$.

Nel sottoparagrafo 5.2 sono discussi i risultati principali delle sperimentazioni effettuate per valutare l'efficacia del criterio di definizione dei pesi da assegnare a $\mathbf{y}_{ab}$.

Passo 5. Abbinamento deterministico - Scelta delle coppie da etichettare come links

Dal momento che eventuali accoppiamenti del tipo 'uno a molti' di uguale peso possono essere risolti facilmente in un secondo momento (Passo 7), a questo stadio del processo di integrazione si consente che essi possano essere candidati ad essere etichettati come *links*.

La strategia adottata per la scelta delle coppie da etichettare effettivamente come *links*, garantendo il rispetto dei vincoli di compatibilità, è basata su un algoritmo di tipo *greedy*.[16]

L'algoritmo opera, dapprima, ordinando in senso non crescente l'insieme dei pesi w_{ab}, $(a,b) \in A \times B$. Gli accoppiamenti aventi in comune il *record* in *B* (cioè gli abbinamenti del tipo 'molti a uno'), se caratterizzati da uguale peso w_{ab}, sono esclusi dalle successive

[15] Per ogni coppia di *records* analizzati nella presente applicazione, i pesi (5.2) risultano compresi tra 0 e 1. Tuttavia, a seconda delle frequenze associate alle varie modalità di ciascuna variabile di abbinamento, potrebbero anche verificarsi casi di coppie (con qualche discordanza) i cui pesi (5.2) risultano maggiori di 1.

[16] L'idea alla base di un algoritmo di tipo *greedy* è di costruire iterativamente la soluzione seguendo un semplice criterio di espansione, che consiste nell'effettuare, ad ogni iterazione, la scelta più conveniente compatibilmente con i vincoli del problema. Qualunque scelta, una volta effettuata, non viene mai rimessa in discussione (per maggiori dettagli si rimanda a Edmonds, 1971).

elaborazioni; l'insieme di tali accoppiamenti, è denotato, d'ora in poi, con *D*. Quindi, le coppie da etichettare come *links* sono scelte una per volta – secondo l'ordine non crescente del peso a loro associato – finché $w_{ab} \geq t_l$, $(a,b) \in A \times B - D$, rigettando una coppia se si verifica una delle seguenti circostanze:

- il *record* in *A* ad essa relativo è elemento di qualche coppia già scelta, caratterizzata da un peso superiore;
- il *record* in *A* ad essa relativo è elemento di qualche coppia appartenente all'insieme *D*, caratterizzata da un peso uguale o superiore;
- il *record* in *B* ad essa relativo è elemento di una coppia già scelta;
- il *record* in *B* ad essa relativo è elemento di qualche coppia appartenente all'insieme *D*, caratterizzata da un peso superiore.

In altre parole, si escludono dalla scelta tutti quegli accoppiamenti che non rappresentano candidati 'migliori' ad essere etichettati come *links* rispetto a coppie già scelte o escluse, tenendo conto dei vincoli imposti dal contesto applicativo. Il motivo alla base della preferenza per questo tipo di algoritmo è dettato dal fatto che l'approccio solitamente adottato per la scelta delle coppie da etichettare come *links* (sottoparagrafo 4.1) può condurre a risultati non plausibili, a prescindere dal contesto applicativo specifico. Ciò è mostrato attraverso un semplice esempio nel sottoparagrafo 5.3.

Passo 6. Controllo delle coppie selezionate con pesi compresi tra t_l e t_u

Le coppie selezionate al Passo 5 con pesi compresi tra t_l e t_u subiscono un controllo accurato, ricorrendo anche ad informazioni supplementari, prima di essere confermate come *links*.

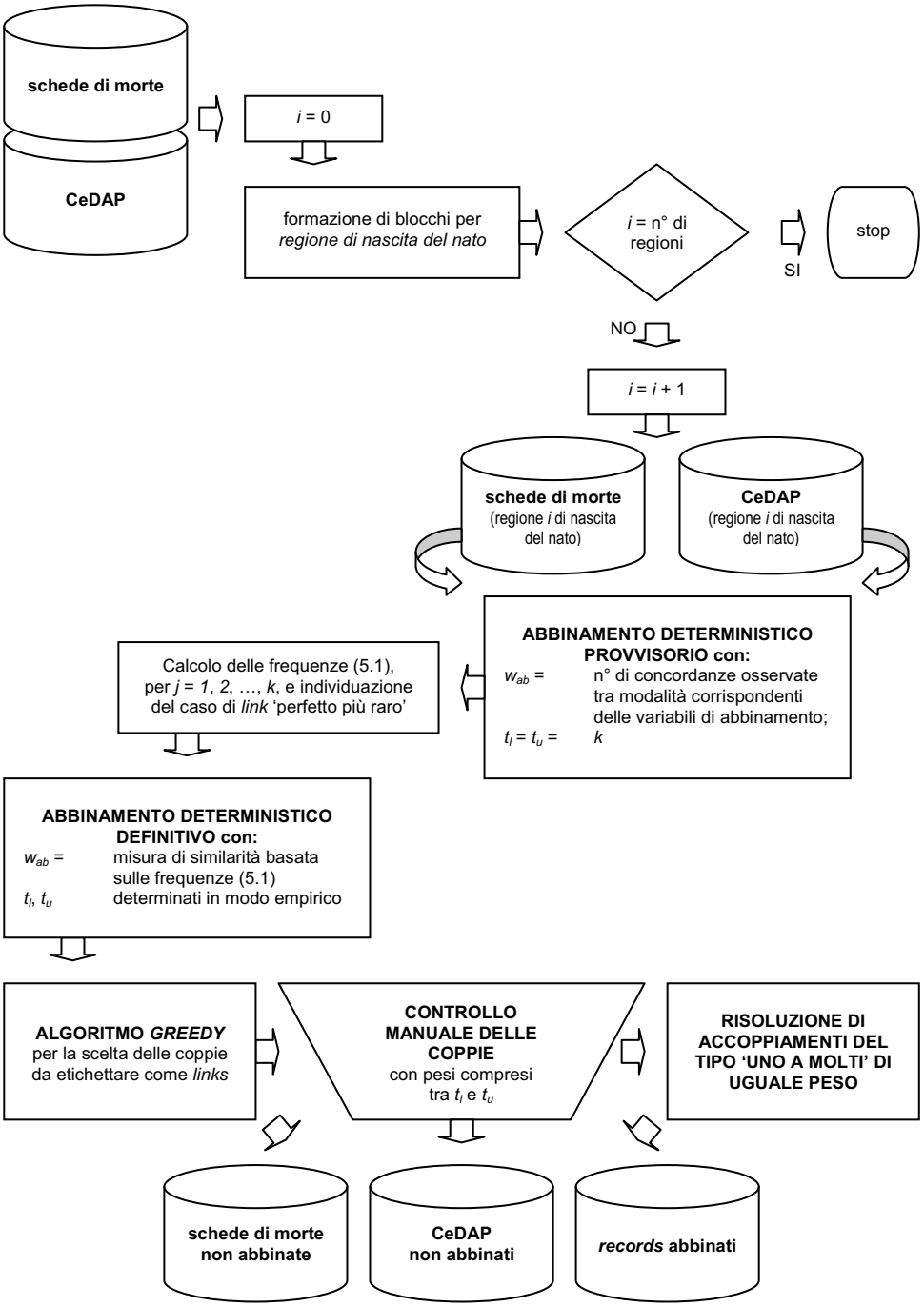
Passo 7. Ricerca di eventuali abbinamenti multipli del tipo 'uno a molti' (di uguale peso) e loro risoluzione

Come per il caso di *link* 'perfetto più raro' (Passo 3), la risoluzione di abbinamenti del tipo 'uno a molti' di uguale peso w_{ab}, riconducibili a nati da un unico parto (di cui uno solo morto), avviene etichettando come *link* soltanto l'accoppiamento il cui *record* relativo ai CeDAP si riferisce al nato con *peso alla nascita* minore (Branum e Schoendorf, 2003).

Altri eventuali (e molto rari) abbinamenti multipli di uguale peso, non riconducibili a nati da un unico parto, sono invece etichettati come *non-links*.

L'ordine di esecuzione dei vari passi in cui è articolato il processo di abbinamento è rappresentato schematicamente nel diagramma di flusso di figura 1.

Figura 1 - Il processo di abbinamento tra schede di morte nel primo mese di vita e CeDAP relativi ai nati vivi di un certo anno



5.2 Alcuni risultati delle sperimentazioni condotte per valutare l'efficacia del criterio adottato per la definizione dei pesi

In questo sottoparagrafo sono discussi i risultati principali delle sperimentazioni condotte su alcune coppie di blocchi di *records* al fine di scegliere il criterio di definizione dei pesi da assegnare al vettore $\mathbf{y}_{ab}$.

In particolare, il criterio deterministico proposto nel Passo 4 (sottoparagrafo 5.1) per la definizione dei pesi è posto a confronto con il criterio in cui i pesi sono definiti in modo probabilistico (secondo l'impostazione di Fellegi e Sunter) e stimati con l'algoritmo EM, nell'implementazione fornita dal *software* Link Plus (versione 2.0).

Tale *software* – sviluppato[17] per l'ambiente Microsoft Windows a 32 bit dalla *Division of Cancer Prevention and Control* presso i *Centers for Disease Control and Prevention* negli Stati Uniti – sebbene sia stato originariamente concepito per aggiornare le informazioni contenute nei registri sui tumori, può essere utilizzato anche in altri contesti applicativi. Il *software* presenta molte funzionalità; tra queste, vi è la possibilità di aggiustare i pesi, stimati con l'algoritmo EM, utilizzando le frequenze con cui si presentano le modalità delle variabili di abbinamento. Nell'ambito dell'analisi di tipo comparativo tra i due criteri di definizione dei pesi, si preferisce optare per questa possibilità, al fine di tenere conto del diverso potere identificativo delle modalità, e ottenere, quindi, risultati più accurati.

Una volta determinati i pesi per i vari accoppiamenti di *records* in base ai due criteri posti a confronto, la scelta delle coppie da etichettare come *links* è effettuata applicando l'algoritmo di tipo *greedy*, descritto al Passo 5 del sottoparagrafo 5.1.

Per quanto riguarda Link Plus, sono sottoposte a questo algoritmo tutte le coppie caratterizzate da peso (4.4) positivo; in altre parole, la soglia t_l è posta uguale a 0, in modo da non escludere a priori dall'essere etichettata come *link* alcuna coppia per cui nel rapporto della formula (4.4) il numeratore sia superiore al denominatore. Il valore di t_u è determinato successivamente in modo empirico, ispezionando la distribuzione dei pesi stimati con l'algoritmo EM.

Per entrambe le procedure poste a confronto, il controllo manuale delle coppie selezionate con peso compreso tra t_l e t_u e la risoluzione di eventuali abbinamenti multipli del tipo 'uno a molti' di uguale peso avviene poi secondo quanto delineato nei Passi 6 e 7 del sottoparagrafo 5.1.

Sono riportati i risultati ottenuti per le coppie di blocchi di *records* relativi all'*anno di nascita del nato* '2003', per i quali la variabile *regione di nascita del nato* assume le seguenti modalità:

- 'Friuli Venezia Giulia' e 'Veneto', caratterizzate da una copertura molto elevata della rilevazione dei CeDAP;[18]

[17] Il *software* Link Plus per il *record linkage* probabilistico è scaricabile gratuitamente all'indirizzo: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm.

I requisiti hardware necessari per l'esecuzione del *software* sono gli stessi di quelli richiesti per il sistema operativo Microsoft Windows. Tra i requisiti aggiuntivi di sistema rientrano:

- sistema operativo: Microsoft Windows XP (o più recente);
- memoria di sistema: almeno 512 MB;
- spazio libero su disco: almeno 1 GB.

[18] Come si può vedere da tavola 1, per queste due regioni il rapporto tra numero di CeDAP e numero di parti rilevati attraverso le SDO risulta superiore a 100.

- 'Liguria', caratterizzata dalla copertura meno elevata della rilevazione dei CeDAP nell'ambito dell'Italia settentrionale.

Con riferimento a queste tre regioni, le coppie etichettate automaticamente come *links* (in quanto caratterizzate da un peso superiore a t_u) risultanti dall'applicazione della strategia di abbinamento proposta nel sottoparagrafo 5.1 coincidono con quelle ottenute dall'utilizzo combinato di Link Plus e dei Passi 5 e 7, con una scelta opportuna dei valori t_u. I *links* così individuati sono 25, 77 e 34, rispettivamente[19] per 'Friuli Venezia Giulia', 'Veneto' e 'Liguria'.

Per quanto riguarda le coppie etichettate manualmente come *links*:

- per la regione 'Friuli Venezia Giulia', le schede di morte risultano tutte già abbinate in modo automatico con il corrispondente certificato di nascita;
- per la regione 'Veneto', nessuna delle coppie con peso compreso tra t_l e t_u, proposte in *output* dalle due procedure, è confermata come *link*;
- per la regione 'Liguria', le coppie confermate manualmente come *links*, tra quelle con peso compreso tra t_l e t_u proposte in *output* dalle due procedure, sono le stesse.

Qualche differenza tra le procedure messe a confronto si riscontra (ovviamente) per le coppie sottoposte a controllo manuale con pesi non particolarmente elevati (anche se compresi tra t_l e t_u) e non confermate di fatto come *links*, trattandosi di accoppiamenti ritenuti non accettabili.

Occorre osservare che, in base alle sperimentazioni condotte utilizzando il criterio probabilistico, se non si ricorre all'aggiustamento successivo dei pesi con le frequenze osservate delle modalità delle variabili di abbinamento, i risultati cui si perviene sono leggermente diversi e meno attendibili rispetto a quelli appena presentati.

Vista la finalità esplorativa del presente lavoro e la sostanziale equivalenza dei risultati ottenuti con le due procedure messe a confronto, il criterio deterministico proposto per la definizione dei pesi, seppur poco raffinato, viene preferito a quello probabilistico per la sua maggiore immediatezza, focalizzando l'attenzione su altri aspetti, ritenuti di maggiore impatto sull'accuratezza dei risultati finali nel contesto applicativo specifico.

5.3 Rispetto di vincoli di compatibilità tra coppie di *records*: criticità dello schema di assegnazione per cui la somma dei pesi è massima

Il motivo alla base della preferenza per una strategia di tipo *greedy* nella scelta dei *records* da considerare effettivamente abbinati, quando occorra garantire il rispetto di vincoli di compatibilità tra coppie, è dettato dal fatto che l'approccio solitamente adottato – secondo il quale lo schema migliore di assegnazione dei *records* è quello per cui la somma dei pesi è massima – può condurre a risultati non plausibili.

Già Armstrong e Saleh (2000) hanno riscontrato questo inconveniente nell'esperienza pratica, senza fornirne, tuttavia, una spiegazione di carattere generale.

A tal fine, si considerino i *records* $a, a' \in A$ e $b, b' \in B$ e le possibili coppie (a,b), (a,b'), (a',b), (a',b') con le seguenti relazioni tra i pesi:

[19] Si pone l'attenzione sul fatto che i risultati qui riportati, relativi al numero di coppie etichettate automaticamente come *links* per Friuli Venezia Giulia, Veneto e Liguria, non corrispondono a quelli riportati successivamente nelle tavole 4 e 5, in quanto questi ultimi sono presentati per *regione di decesso del nato*, anziché per *regione di nascita del nato*.

$$w_{ab} + w_{a'b'} > w_{ab'} + w_{a'b},$$

$$w_{ab'} > w_{ab}, w_{a'b'}, w_{a'b}.$$

Ipotizzando che tutti i pesi siano maggiori o uguali al valore soglia t_u e che il contesto applicativo ammetta soltanto abbinamenti del tipo 'uno a uno', nello schema di assegnazione dei *records* per cui la somma dei pesi risulta massima, la coppia (a,b') non sarebbe etichettata come *link*, pur essendo caratterizzata dal peso più elevato. Tale decisione risulta poco sensata, soprattutto se si pensa che la coppia (a,b') potrebbe corrispondere ad un caso in cui le modalità corrispondenti di tutte le variabili di abbinamento risultano coincidenti e non mancanti nei *records* $a \in A$ e $b' \in B$, o addirittura rappresentare – adottando la terminologia introdotta nel Passo 3 – un caso di *link* 'perfetto più raro'.[20]

Va sottolineato che lo schema di assegnazione dei *records* ottenuto come soluzione ottima del problema (4.5) dipende dal valore numerico dei pesi stessi, e non semplicemente dal loro ordinamento; in tal modo, risulta anche meno robusto dello schema risultante dall'applicazione di una strategia *greedy*, rispetto ad eventuali distorsioni presenti nelle stime dei pesi.

Infatti, si considerino, ad esempio, i *records* $a, a' \in A$ e $b, b' \in B$ e le possibili coppie (a,b), (a,b'), (a',b), (a',b') con i seguenti pesi:

$$w_{ab} = 15 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8.$$

Se si ipotizza che tutti i pesi siano maggiori o uguali al valore soglia t_u e che il contesto applicativo consenta soltanto abbinamenti del tipo 'uno a uno', massimizzando la somma dei pesi o seguendo un approccio *greedy*, sarebbero etichettate come *links* le coppie (a,b) e (a',b'). Tuttavia, variando il peso corrispondente alla coppia (a,b) senza alterare l'ordinamento dei pesi, come nel modo seguente:

$$w_{ab} = 13 > w_{ab'} = 12 > w_{a'b} = 10 > w_{a'b'} = 8,$$

lo schema di assegnazione dei *records* che si ottiene massimizzando la somma dei pesi cambia, essendo costituito dalle coppie (a,b') e (a',b), mentre quello risultante dall'adozione della strategia *greedy* rimane invariato.

Come ovvia conseguenza di quanto appena affermato, nel caso di pesi definiti secondo la formula (4.4), lo schema di assegnazione che si ottiene come soluzione ottima del

[20] Secondo Jaro (1989), la verifica delle relazioni (4.3) – al fine di etichettare le coppie come *links* – dovrebbe avvenire solo dopo aver individuato lo schema di assegnazione ottimo; ciò è discutibile in quanto, così facendo, si rischia di inserire in tale schema anche qualche coppia il cui peso è inferiore a t_u, a scapito di eventuali altri accoppiamenti che potrebbero costituire dei candidati 'migliori' ad essere etichettati come *links*.

problema (4.5) potrebbe differire da quello che si otterrebbe qualora si facesse a meno della trasformazione logaritmica.

Lo schema di assegnazione rappresentato dalla soluzione ottima per il problema (4.5) può non presentare inconvenienti in quei contesti applicativi in cui le variabili di abbinamento siano caratterizzate da un'elevata qualità e, allo stesso tempo, da un notevole potere identificativo nei confronti degli individui (ad esempio, le variabili *cognome* o *indirizzo del domicilio*). Un contesto simile è costituito proprio dal caso considerato da Jaro (1989), relativo alla stima del grado di copertura del censimento della popolazione. Tuttavia, in una situazione come quella presa in esame, caratterizzata da variabili di abbinamento meno identificative[21] e dalla qualità non particolarmente elevata, diventa fondamentale garantire il rispetto dei vincoli di compatibilità tra coppie di *records* con un approccio più adeguato al problema.

6. Risultati dell'abbinamento

6.1 Risultati preliminari a livello nazionale

In questo sottoparagrafo sono illustrati brevemente i risultati principali di un abbinamento effettuato in via preliminare sui dati relativi all'intero territorio nazionale.

Il procedimento utilizzato è rappresentato essenzialmente dal Passo 1 della strategia di abbinamento descritta nel sottoparagrafo 5.1, ovvero sono etichettate come *links* tutte le coppie di *records* senza alcuna discordanza tra modalità corrispondenti delle variabili di abbinamento. Inoltre, la risoluzione di eventuali accoppiamenti del tipo 'uno a molti' di uguale peso w_{ab}, riconducibili a nati da un unico parto (di cui uno solo morto), avviene etichettando come *link* soltanto l'accoppiamento per cui il *record* relativo ai CeDAP si riferisce al nato con *peso alla nascita* minore.

Come si può vedere dai risultati riportati per regione di decesso e anno di nascita del nato (tavola 4), il passo di abbinamento deterministico provvisorio non dà luogo ad un elevato numero di *links*. Difatti, a livello nazionale la percentuale di abbinamento – calcolata rispetto al numero di morti – è pari a 29,2 per i nati nel 2003 e scende a 23,6 per i nati nel 2004.

I *links* provenienti da abbinamenti 'uno a molti' di uguale peso w_{ab} – riconducibili a nati da un unico parto (per motivi di spazio, i dati non sono riportati nella tavola 4) – sono 22 e 20, rispettivamente per gli anni 2003 e 2004.

Scendendo nel dettaglio delle ripartizioni geografiche, la percentuale di abbinamento relativa ai nati nel 2004 risulta minore di quella relativa ai nati dell'anno precedente, soprattutto per l'Italia centrale che, nel secondo anno, presenta anche il valore più basso (si passa dal 25,0 al 10,3 per cento). Per quanto riguarda il Mezzogiorno, in entrambi gli anni presi in esame, per circa una scheda di morte su cinque viene individuato il certificato di nascita corrispondente.

L'Italia settentrionale risulta la ripartizione geografica con la percentuale più elevata di *links*, nonostante per il 2004 si osservi una riduzione rispetto al 2003 (dal 42,5 al 37,5 per

[21] Soprattutto per i *records* per i quali la data di nascita della madre è mancante o inammissibile, come avviene per circa un terzo delle schede di morte, le altre variabili di abbinamento hanno un potere identificativo limitato.

cento). Rispetto a quanto osservato per le altre ripartizioni, questo risultato è strettamente legato sia alla maggiore copertura della rilevazione dei CeDAP, sia alla minore incidenza di valori mancanti o inammissibili per le variabili di abbinamento (a tale proposito si rimanda al paragrafo 3).

Tavola 4 - Risultati dell'abbinamento deterministico provvisorio (Passo 1) tra *records* relativi alle schede di morte nel primo mese di vita e *records* relativi ai CeDAP, per regione di decesso e anno di nascita del nato

regione di decesso del nato	2003			2004		
	n° di morti	n° di *links*	n° di *links*/ n° di morti (%)	n° di morti	n° di *links*	n° di *links*/ n° di morti (%)
Piemonte	94	52	55,3	75	49	65,3
Valle d'Aosta/Vallée d'Aoste	1	1	100,0	1	1	100,0
Lombardia	222	72	32,4	204	49	24,0
Trento (a)	*11*	*10*	*90,9*	*10*	*7*	*70,0*
Veneto	86	39	45,3	97	36	37,1
Friuli Venezia Giulia	26	9	34,6	15	4	26,7
Liguria	40	21	52,5	36	11	30,6
Emilia Romagna	90	38	42,2	87	40	46,0
Nord	**570**	**242**	**42,5**	**525**	**197**	**37,5**
Toscana	51	10	19,6	90	18	20,0
Umbria	24	7	29,2	17	6	35,3
Marche	34	13	38,2	23	-	-
Lazio	179	42	23,5	172	7	4,1
Centro	**288**	**72**	**25,0**	**302**	**31**	**10,3**
Abruzzo	31	8	25,8	36	15	41,7
Molise	3	….	….	3	….	….
Campania	206	54	26,2	219	49	22,4
Puglia	137	42	30,7	151	34	22,5
Basilicata	5	1	20,0	10	3	30,0
Calabria	61	….	….	64	….	….
Sicilia	199	25	12,6	195	26	13,3
Sardegna	28	5	17,9	22	8	36,4
Sud e Isole	**670**	**135**	**20,1**	**700**	**135**	**19,3**
Italia	**1.528**	**449**	**29,2**	**1.527**	**363**	**23,6**

Fonte: elaborazione propria su dati Istat (Indagine sulle cause di morte) e Ministero della Salute (Rilevazione dei CeDAP)

(a) dal momento che i dati relativi alla Provincia Autonoma di Bolzano sono mancanti, non viene riportato il totale per la regione Trentino-Alto Adige

A livello regionale, il decremento maggiore nella percentuale di *links* si riscontra per la regione Marche, per la quale nessuna delle 23 schede di morte relativa ai nati nel 2004 risulta abbinata. Per quanto riguarda il Sud e le Isole, si registra, invece, un lieve incremento in quasi tutte le regioni, ad eccezione di Campania e Puglia che, essendo tra le

più popolose, determinano la diminuzione osservata al livello della ripartizione di appartenenza.

Escludendo la regione Valle D'Aosta – per cui l'unica scheda di morte viene abbinata con il certificato di nascita corrispondente – la Provincia Autonoma di Trento è l'area geografica per cui il passo di abbinamento preliminare conduce alla quota più elevata di *links* (90,9 e 70,0 per cento, rispettivamente per gli anni 2003 e 2004); nell'ambito dell'Italia settentrionale, la regione Lombardia – dove, almeno per i nati nel 2003, si osserva il numero più elevato di morti neonatali – fa registrare, invece, la quota minima (32,4 e 24,0 per cento, rispettivamente per gli anni 2003 e 2004).

La bassa percentuale di *links* raggiunta con il passo di abbinamento deterministico provvisorio indirizza verso l'adozione di una strategia più raffinata, quale quella descritta nel sottoparagrafo 5.1. Inoltre, alla luce delle valutazioni effettuate sulla qualità dei dati utilizzati per l'abbinamento e sulla copertura della rilevazione dei CeDAP, il recupero di informazioni sulla mortalità neonatale viene circoscritto alla sola Italia settentrionale.

6.2 Risultati dell'abbinamento per l'Italia settentrionale

L'applicazione della procedura di abbinamento proposta ai dati dell'Italia settentrionale consente di ottenere un numero elevato di *links* (tavola 5); in alcune regioni, per tutte le schede di morte vengono individuati i certificati di nascita corrispondenti. Tuttavia, a livello di ripartizione geografica, la percentuale di abbinamento passa da 92,3 per i nati nel 2003 a 87,6 per i nati nel 2004.

Scendendo nel dettaglio regionale, nei due anni presi in esame l'aumento più rilevante nella quota di *links* si osserva per il Veneto (da 91,9 a 96,9 per cento). Valle d'Aosta, Provincia Autonoma di Trento e Piemonte mantengono percentuali di abbinamento molto elevate; in particolare, nei primi due casi, che sono anche quelli caratterizzati dal minor numero di decessi in termini assoluti, tutte le schede di morte risultano abbinate.

Le diminuzioni nella percentuale di *links* riguardano Friuli Venezia Giulia, Liguria, Lombardia ed Emilia Romagna. Sono le prime tre regioni a far registrare le maggiori riduzioni; per Liguria e Lombardia, tali riduzioni sono probabilmente imputabili anche al decremento del grado di copertura della rilevazione dei CeDAP (tavola 1).

I *links* provenienti da abbinamenti 'uno a molti' di uguale peso w_{ab}, tutti riconducibili a nati dello stesso sesso da un unico parto (i dati non sono riportati nella tavola 5, per motivi di spazio), rappresentano circa il 4,8 e il 5,0 per cento del totale dei *links*, rispettivamente per gli anni 2003 e 2004.

Come si può dedurre dai risultati riportati nella tavola 5, il 91,4 e il 95,7 per cento dei *links*, rispettivamente per gli anni 2003 e 2004, sono ottenuti in modo automatico, presentando pesi superiori a t_u; l'8,6 e il 4,3 per cento dei *links* sono etichettati, invece, mediante conferma manuale, essendo caratterizzati da pesi compresi tra t_l e t_u.

Tavola 5 - Risultati della procedura di abbinamento deterministico definitivo (Passi 1-7) tra *records* relativi alle schede di morte nel primo mese di vita e *records* relativi ai CeDAP, per regione di decesso e anno di nascita del nato - Italia settentrionale

regione di decesso del nato	n° di morti	n° di *links*	n° di *links* con peso ≥ t_u	n° di *links* con peso compreso tra t_l e t_u	n° di *links*/ n° di morti (%)
			2003		
Piemonte	94	91	84	7	96,8
Valle d'Aosta/Vallée d'Aoste	1	1	1	-	100,0
Lombardia	222	192	161	31	86,5
Trento (a)	*11*	*11*	*11*	-	*100,0*
Veneto	86	79	79	-	91,9
Friuli Venezia Giulia	26	26	26	-	100,0
Liguria	40	37	34	3	92,5
Emilia Romagna	90	89	85	4	98,9
Nord	**570**	**526**	**481**	**45**	**92,3**
			2004		
Piemonte	75	72	69	3	96,0
Valle d'Aosta/Vallée d'Aoste	1	1	1	-	100,0
Lombardia	204	160	147	13	78,4
Trento	*10*	*10*	*10*	-	*100,0*
Veneto	97	94	91	3	96,9
Friuli Venezia Giulia	15	13	13	-	86,7
Liguria	36	29	29	-	80,6
Emilia Romagna	87	81	80	1	93,1
Nord	**525**	**460**	**440**	**20**	**87,6**

Fonte: elaborazione propria su dati Istat (Indagine sulle cause di morte) e Ministero della Salute (Rilevazione dei CeDAP)

(a) dal momento che i dati relativi alla Provincia Autonoma di Bolzano sono mancanti, non viene riportato il totale per la regione Trentino-Alto Adige

L'elevata percentuale di *links* per cui la totalità o la gran parte delle variabili di abbinamento presenta modalità concordanti – per oltre il 96 per cento dei *links* si verificano almeno 8 concordanze (tavola 6) – non esclude, tuttavia, la presenza di errori (falsi *matches* o mancati *matches*) nei risultati, soprattutto per alcune criticità relative ai dati utilizzati: copertura della rilevazione dei CeDAP non ancora soddisfacente in alcune regioni (Lombardia e Liguria); variabili di abbinamento non particolarmente identificative; elevata incidenza di valori mancanti o inammissibili nell'archivio delle schede di morte per i campi relativi alla data di nascita della madre.

D'altra parte, risulta difficile ottenere una stima dei tassi di errore di abbinamento, non disponendo di un sottoinsieme di coppie per le quali possa essere stabilita con certezza l'appartenenza a *M* o *U*; per di più, una stima basata su modello, seguendo altri approcci proposti in letteratura (ad esempio, in Armstrong e Mayda, 1993), potrebbe risultare non sufficientemente accurata a causa delle criticità cui si è accennato sopra.

Tavola 6 - Distribuzione percentuale degli individui abbinati, per numero di modalità concordanti delle variabili di abbinamento, per anno di nascita del nato - Italia settentrionale

numero di modalità concordanti delle variabili di abbinamento	individui abbinati	
	2003	2004
12	46,01	42,83
11	12,17	18,26
10	8,17	6,30
9	19,96	15,22
8	9,70	14,57
7	2,47	2,61
meno di 7	1,52	0,22
totale	**100,00**	**100,00**

Fonte: elaborazione propria su dati Istat (Indagine sulle cause di morte) e Ministero della Salute (Rilevazione dei CeDAP)

7. Tassi di mortalità per classe di età gestazionale e di peso alla nascita

Come già evidenziato nel paragrafo 1, il 1998 rappresenta l'ultimo anno in cui in Italia è stato possibile calcolare tassi di mortalità neonatale per classe di età gestazionale e di peso alla nascita. Sulla base dei risultati del *record linkage* per l'Italia settentrionale viene fornita una stima di questi indicatori limitatamente al 2003, primo anno utile per una ricostruzione della serie storica e, soprattutto, anno per cui la qualità delle informazioni utilizzate per l'abbinamento (in particolare, quelle relative alla madre) risulta più elevata.

I tassi di mortalità neonatale per l'anno 2003 riportati in questo lavoro sono costruiti come segue:

- al denominatore viene utilizzata la distribuzione dei nati vivi in Italia settentrionale nel 2003, per classe di età gestazionale e di peso alla nascita, risultante dai dati CeDAP;
- al numeratore viene utilizzata la distribuzione dei deceduti entro il primo mese di vita in Italia settentrionale appartenenti alla coorte di nati vivi nel 2003,[22] per classe di età gestazionale e di peso alla nascita, risultante dall'abbinamento precedentemente descritto.

Per le 570 morti neonatali totali è ipotizzata la distribuzione effettiva per classe di età gestazionale e di peso alla nascita dei 481 *links* con peso non inferiore a t_u .[23]

I risultati ottenuti (figure 2 e 3) sembrano essere in linea sia con altri studi sulla sopravvivenza dei nati pretermine e di basso peso alla nascita condotti in altre nazioni (Demissie *et al*., 2001; Horbar *et al*., 2002), sia con le tendenze degli anni novanta del

[22] Rispetto agli anni precedenti della serie storica, viene calcolato un tasso per generazione e non per contemporanei; comunque, ipotizzando una distribuzione uniforme delle morti neonatali nei vari mesi dell'anno, questi tassi risultano essere del tutto comparabili.

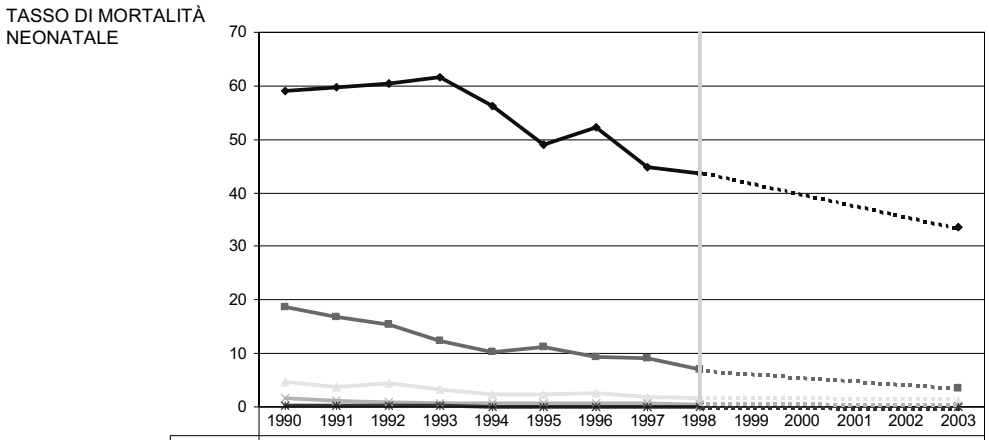
[23] Al fine di contenere l'eventuale effetto distorsivo dovuto alla possibile presenza di falsi *matches*, viene utilizzata la distribuzione dei 481 *links* che presentano un peso più elevato (colonna 4 della tavola 5) e non la distribuzione dei totali 526 *links* (colonna 3 della tavola 5).

fenomeno oggetto di studio, nel senso di un progressivo e consistente contenimento della mortalità neonatale, in modo particolare per i nati fortemente pretermine e di più basso peso alla nascita. Rispetto al 1998, ultimo anno per cui sono stati calcolati i tassi, nel 2003 si assiste ad una consistente riduzione assoluta e relativa della mortalità neonatale per le prime classi di età gestazionale e di peso alla nascita, dove peraltro il valore del tasso è decisamente più elevato:

- di circa 10 (da 43,7 a 33,6 per cento nati vivi) e 6 punti (da 36,7 a 30,8 per cento nati vivi) per il tasso di mortalità al di sotto delle 28 settimane di gestazione e dei 1.000 grammi di peso alla nascita, rispettivamente;
- di circa il 50 per cento del tasso di mortalità sia tra le 28 e le 31 settimane di gestazione (da 7,0 a 3,4 per cento nati vivi), sia tra i 1.000 e i 1.499 grammi di peso alla nascita (da 6,1 a 3,1 per cento nati vivi).

Inoltre, rispetto al passato, dai risultati ottenuti per il 2003 emerge una maggiore concentrazione delle morti neonatali nelle classi di età gestazionale e di peso alla nascita meno elevate: il 42 e il 55 per cento di queste morti riguardano bambini nati da una gestazione inferiore, rispettivamente, alle 28 e alle 32 settimane; il 40 e il 48 per cento delle morti neonatali riguardano bambini nati con peso al di sotto, rispettivamente, dei 1.000 e dei 1.500 grammi.

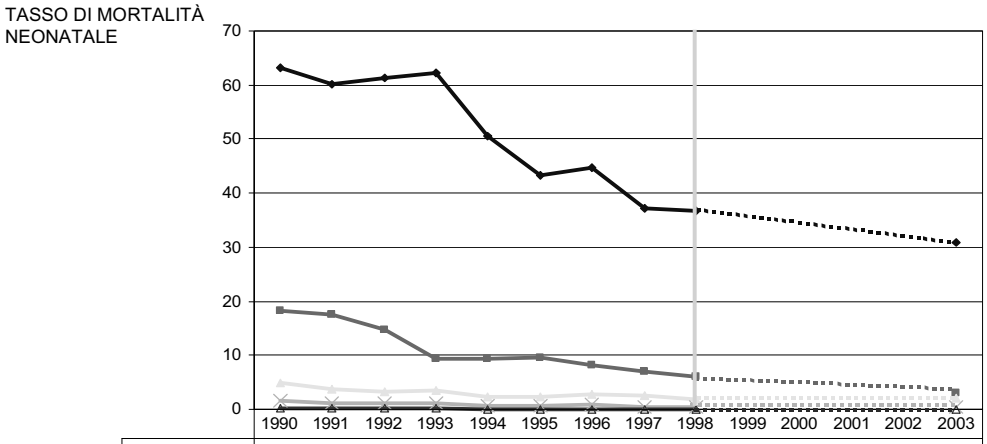
Figura 2 - Tassi di mortalità neonatale per classe di età gestazionale e anno di nascita del nato - Italia settentrionale, anni 1990-98 e 2003 (valori per 100 nati vivi della stessa classe di età gestazionale)



CLASSE DI ETÀ GESTAZIONALE (in settimane)	1990	1991	1992	1993	1994	1995	1996	1997	1998	2003
<28	59,1	59,8	60,4	61,5	56,3	49,1	52,3	44,9	43,7	33,6
28-31	18,6	16,7	15,5	12,4	10,3	11,2	9,3	9,0	7,0	3,4
32-34	4,8	3,7	4,3	3,2	2,4	2,3	2,5	1,8	1,6	1,1
35-36	1,5	1,2	1,0	0,8	0,8	0,6	0,8	0,6	0,5	0,3
>=37	0,2	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1

Fonte: per gli anni 1990-98, elaborazione propria su dati Istat (Rilevazione delle nascite, Indagine sulle cause di morte); per l'anno 2003, elaborazione propria su dati Istat (Indagine sulle cause di morte) e Ministero della Salute (Rilevazione dei CeDAP)

Figura 3 - Tassi di mortalità neonatale per classe di peso alla nascita e anno di nascita del nato - Italia settentrionale, anni 1990-98 e 2003 (valori per 100 nati vivi della stessa classe di peso alla nascita)



CLASSE DI PESO ALLA NASCITA (in grammi)	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
<1000	63,1	60,2	61,3	62,3	50,5	43,3	44,7	37,2	36,7					30,8
1000-1499	18,2	17,6	14,7	9,4	9,3	9,7	8,3	7,0	6,1					3,1
1500-1999	4,8	3,8	3,3	3,5	2,3	2,4	2,8	2,6	1,8					2,2
2000-2499	1,6	1,2	1,1	1,1	0,8	0,7	0,9	0,6	0,6					0,4
>=2500	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1					0,1

Fonte: per gli anni 1990-98, elaborazione propria su dati Istat (Rilevazione delle nascite, Indagine sulle cause di morte); per l'anno 2003, elaborazione propria su dati Istat (Indagine sulle cause di morte) e Ministero della Salute (Rilevazione dei CeDAP)

8. Considerazioni conclusive

Alla luce delle valutazioni effettuate sulla copertura della rilevazione dei CeDAP e sulla qualità dei valori assunti dalle variabili di abbinamento, si è ritenuto opportuno circoscrivere il recupero di informazioni sulla mortalità neonatale alla sola Italia settentrionale. Allo stato attuale delle fonti di dati disponibili, la bassa percentuale di casi per cui tale recupero è possibile, verificando soltanto la concordanza perfetta delle informazioni comuni ai due archivi, indirizza senz'altro verso l'adozione di una strategia di abbinamento più raffinata, di tipo deterministico o probabilistico.

Nel presente lavoro la possibilità di recupero di informazioni sulla mortalità neonatale è stata investigata ricorrendo ad una strategia di abbinamento di tipo deterministico. Il criterio di definizione dei pesi adottato, sul cui valore è basato il processo di classificazione di ogni coppia di *records* negli insiemi M o U, tiene conto delle frequenze con cui le varie modalità di ciascuna variabile di abbinamento si presentano nell'archivio dei CeDAP, di dimensione maggiore. In base alle sperimentazioni effettuate sui dati di alcune regioni (caratterizzate da una diversa copertura della rilevazione dei CeDAP), tale criterio, seppur non estremamente raffinato, ha consentito di raggiungere sostanzialmente la stessa classificazione delle coppie di *records* ottenuta con un criterio di tipo probabilistico.

Il contesto applicativo specifico, richiedendo che una scheda di morte potesse essere associata ad un solo certificato di nascita, ha offerto lo spunto per una revisione critica dell'approccio comunemente seguito per risolvere il problema del rispetto dei vincoli di

compatibilità tra abbinamenti, qualora il processo di classificazione riguardi una coppia alla volta. Lo schema di assegnazione dei *records* per cui la somma dei pesi è massima, secondo tale approccio, può condurre, infatti, a risultati non plausibili, dal momento che dipende dal valore numerico dei pesi. Si ritiene che un approccio di tipo *greedy*, dipendendo semplicemente dall'ordinamento dei pesi, sia più appropriato, a prescindere dal contesto applicativo specifico.

L'alta percentuale di casi per cui il recupero di informazioni sulla mortalità neonatale è stato possibile ricorrendo alla strategia proposta, non esclude tuttavia la presenza di errori di abbinamento. Dal momento che tali errori potrebbero essere anche di entità non trascurabile, è richiesta cautela nell'interpretazione dei tassi di mortalità per classi di età gestazionale e di peso alla nascita ottenuti, per la possibile presenza di effetti distorsivi.

Nel recupero di informazioni sulla mortalità neonatale per il calcolo di questo tipo di tassi si ritiene che ulteriori sforzi vadano rivolti verso l'obiettivo di controllare eventuali errori di abbinamento. I risultati ottenuti, seppur in linea con le tendenze in Italia degli anni passati e con le evidenze delle casistiche di singoli centri ospedalieri (Corchia *et al.*, 2003), necessitano ovviamente di verifiche più approfondite che non possono però realizzarsi senza un miglioramento generale dei processi di raccolta dei dati, sia in termini di copertura della popolazione oggetto di studio che di completezza dell'informazione su di essa rilevata.

Riferimenti bibliografici

Armstrong J. B., Mayda J. E. (1993), "Model-based estimation of record linkage error rates", *Survey Methodology*, 19, 137-147.

Armstrong J., Saleh M. (2000), "Weight estimation for large scale record linkage applications", *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 1-10.

Branum A. M., Schoendorf K. C. (2003), "The effect of birth weight discordance on twin neonatal mortality", *Obstetrics & Gynecology*, 101, 570-574.

Bruzzone S. (2006), "Mortalità infantile e neonatale: fonti statistiche e indicatori", in: Caselli G., Loghi M., Pierannunzio D. (a cura di), *Comportamenti riproduttivi ed esiti sfavorevoli delle gravidanze. La Sardegna come caso paradigmatico*, *Atti del seminario "Determinanti biodemografiche della 'fitness' nella popolazione italiana. Sardegna: caso paradigmatico di longevità riproduttiva?"*, Roma, 15 novembre 2005.

Buratta V., Prati S., Burgio A., Loghi M., Lo Conte M. (2004), "L'informazione sulle nascite in Italia", in: Osservatorio Regionale della Patologia in Età Pediatrica, Regione Veneto, (a cura di), *La nascita: dall'informazione all'intervento*, CLEUP scarl, Padova.

CISIS - Centro Interregionale per i Sistemi Informatici, geografici e Statistici (2004), "La rilevazione dei dati del Certificato di assistenza al parto: stato di attuazione ed esperienze a confronto", *Atti dell'incontro tecnico Ministero della Salute - Istat - Regioni*, Roma, 29 ottobre 2003.

Corchia C., Gualtieri R., Stronati M. (2003), "Epidemiologia dei VLBW in Italia: analisi territoriale dei centri di assistenza e della mortalità", *Atti IX Congresso Nazionale Società Italiana di Neonatologia*, Napoli, 21-24 maggio 2003.

Demissie K., Rhoads G. G., Ananth C. V., Alexander G. R., Kramer M. S., Kogan M. D., Joseph K. S. (2001), "Trends in preterm birth and neonatal mortality among blacks and whites in the United States from 1989 to 1997", *American Journal of Epidemiology*, 154, 307-315.

Dempster A. P., Laird N. M., Rubin D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, B, 39, 1-38.

Edmonds J. (1971), "Matroids and the greedy algorithm", *Mathematical Programming*, 1, 127-136.

Fellegi I. P., Sunter A. B. (1969), "A theory for record linkage", *Journal of the American Statistical Association*, 64, 1183-1210.

Herzog T. N., Scheuren F. J., Winkler W. E. (2007), *Data quality and record linkage techniques*, Springer, New York.

Horbar J. D. (1999), "The Vermont Oxford network: evidence-based quality improvement for neonatology", *Pediatrics*, 103, 350-359.

Horbar J. D., Badgers G. J., Carpenter J. H., Fanaroff A. A., Kilpatrick S., LaCorte M., Phibbs R., Soll R. F. (2002), "Trends in mortality and morbidity for very low birth weight infants 1991-1999", *Pediatrics*, 110, 143-151.

Jaro M. A. (1989), "Advances in record-linkage methodology as applied to matching the Census of Tampa, Florida", *Journal of the American Statistical Association*, 84, 414-420.

Ministero della Salute (2007), *Certificato di assistenza al parto (CeDAP). Analisi dell'evento nascita - Anno 2004*, Dipartimento della Qualità, Direzione Generale del Sistema Informativo, Ufficio di Direzione Statistica.

Prati S., Caropreso M., Loghi M. (2006), "Esiti dei concepimenti: fonti statistiche e indicatori", in: Caselli G., Loghi M., Pierannunzio D. (a cura di), *Comportamenti riproduttivi ed esiti sfavorevoli delle gravidanze. La Sardegna come caso paradigmatico*, *Atti del seminario "Determinanti biodemografiche della 'fitness' nella popolazione italiana. Sardegna: caso paradigmatico di longevità riproduttiva?"*, Roma, 15 novembre 2005.

Winkler W. E. (2000), "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage", *Statistical Research Report Series* RR2000/05, U. S. Bureau of the Census, Washington.

Zeitlin J., Wildman K., Bréart G., Alexander S., Barros H., Blondel B., Buitendijk S., Gissler M., Macfarlane A. (2003), "Selecting an indicator set for monitoring and evaluating perinatal health in Europe: criteria, methods and results from the PERISTAT project", *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 111, 5-14.

Norme redazionali

La Rivista di Statistica Ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche Istat corredati, a parte, da una nota informativa dell'Autore contenente: appartenenza ad istituzioni, attività prevalente, qualifica, indirizzo, casella di posta elettronica, recapito telefonico e l'autorizzazione alla pubblicazione firmata dagli Autori. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di un referente scelto tra gli esperti dei diversi temi affrontati. Gli originali, anche se non pubblicati, non si restituiscono.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file Template.doc disponibile on line o su richiesta. In base a tali standard la lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 30-35 pagine.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 12 righe); quelli in italiano dovranno prevedere anche un *Abstract* in inglese. La bibliografia, in ordine alfabetico per autore, deve essere riportata in elenco a parte alla fine dell'articolo. Quando nel testo si fa riferimento ad una pubblicazione citata nell'elenco, si metta in parentesi tonda il nome dell'autore, l'anno di pubblicazione ed eventualmente la pagina citata. Ad esempio (Bianchi, 1987, Rossi, 1988, p. 55). Quando l'autore compare più volte nello stesso anno l'ordine verrà dato dall'aggiunta di una lettera minuscola accanto all'anno di pubblicazione. Ad esempio (Bianchi, 1987a, 1987b).

Nella bibliografia le citazioni di libri e articoli vanno indicate nel seguente modo. Per i libri: cognome dell'autore seguito dall'iniziale in maiuscolo del nome, il titolo in corsivo dell'opera, l'editore, il luogo di edizione e l'anno di pubblicazione. Per gli articoli: dopo l'indicazione dell'autore si riporta il titolo tra virgolette, il titolo completo in corsivo della rivista, il numero del fascicolo e l'anno di pubblicazione. Nei riferimenti bibliografici non si devono usare abbreviazioni.

Nel testo dovrà essere di norma utilizzato il corsivo per le parole in lingua straniera e il corsivo o grassetto per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare il Comitato di redazione delle pubblicazioni scientifiche Istat e per inviare lavori: rivista@istat.it. Oppure scrivere a:

Comitato di redazione delle pubblicazioni scientifiche
C/O Gilda Sonetti
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012010001000000

ISSN 1828-1982

€ 10,00