Use of administrative records in official statistics: a practical view

Amelia Fresneda Pacheco and Marina Pérez Julián
S. G. de Estadísticas del Mercado Laboral
Instituto Nacional de Estadística
amelia.fresneda.pacheco@ine.es, marina.perez.julian@ine.es

Abstract

Nowadays, the information technologies potential and vastness of available information (censuses, statistical frames) enable the development of procedures of standardisation and record linkage for its use in official statistics.

In this article a process of both standardisation and optimisation of administrative records is presented in order to ultimately transform them in valuable statistical data. Its practical application is demonstrated with the aid of the statistics on disability and labor market of the National Statistics Office of Spain that are annually executed by the combination of administrative records on disability (State Database on Persons with Disabilities, SDPD) and already consolidated labor market statistics (Spanish Labour Force Survey and Spanish Structure of Earnings Survey).

Keywords: Identifier, Record linkage, Name matching, Duplicate detection, Standardisation, Distance function, Selection algorithm, Disability statistics

AMS Subject classifications: 62D99, 62P25

1. Introduction

The utility of administrative records with statistical purposes is a reality today.

The possibilities of enrichment have been largely recognised as the vastness of the information gathered in official registers enables to investigate peculiar aspects or phenomena in our societies. It is being largely applied in the Nordic countries for years and in the last years is acquiring special relevancy and protagonism in the rest of countries of the European Union (see UN/ECE (2007) and Tnder (2008)).
Earlier paper form registers were replaced by data files with the introduction of computers in public administration. The amount of existing administrative registers was gradually increasing and a growing number of administrative data was susceptible of being used to create statistical systems of registers where many of them could be linked through identity numbers.

Nowadays the use of administrative data is a common practice. It must not be exclusively understood from the perspective of statistical production of using registers and not surveys, but as an additional tool that completes the statistical system and allows to carry out mixed processes in which both, consolidated surveys and administrative registers take part. With their mutual combination the statistical production budget considerably decreases, statistical burden to the informants is lessen, the lack of response can be solved and we are able to provide trustworthy, updated and periodic information.

It must be underlined that record linkage strongly depends on data volume and the purpose of linking administrative registers. Therefore, when linking two big registers (for example, Censuses and the Municipal Register of Inhabitants) it complicates considerably the process so the optimisation of processes becomes necessary.

However, when linking a population sample with a record with the aim of completing a survey, the objective is to link as many units as possible because the lack of matching units equals lack of response and it results in estimation problems.

The focus of this article is on record linkage with subsets of the whole population with samples and is illustrated with the particular statistical operation EPD. Likewise, the focus presented below can also be applied in the case of linking big data sets with subsets of the whole population that present identification problems.

Some examples of existing statistics that use administrative registers in their execution are:

- National Spanish Census of 2011, where Municipal Register of Inhabitants data and information from the Ministry of the Interior, from the Social Security and from de Tax Office, among others, has been used as a basic element of its structure. Full information about this operation can be found at INE (2011).

- The Spanish Annual Structure of Earnings Survey which since 2004 takes advantage of the infrastructure of the Spanish Quarterly Labour Cost Survey (QLCS) that uses the information of the Social Security and Tax Office (INE, 2012).

- The wages deciles of the main job incorporated to the Spanish Labour Force Survey (LFS) from 2010 with the aid of the information coming from the Social Security and the Tax Office (INE, 2013a).
Spanish Living Conditions Survey (EU-SILC) will complete in 2014 its information about incomes with administrative data since 2014, in order to face lack of response problems related with this sensitive variable (INE, 2013b).

The statistics "The employment of persons with disabilities” and "The wages of persons with disabilities” (EPD and SPD respectively) which since 2010 provide information about disabilities and labour market thanks to the combination of administrative data proceeding from Elderly and Social Services and Social Security with traditional labour market statistics (Labour Force Survey and Structure of Earnings Survey). See INE (2013c, 2013d).

A generalised vision about the optimisation process of administrative data and its possible combination with other information so as to obtain new statistics is presented below. These mentioned statistics about disability and labour market at the same time produced by the National Statistics Institute (INE) illustrate the use of the procedures described below.

2. Basic requirements about administrative registers

Any administrative register susceptible to be used with statistical purposes should verify the following characteristics:

a) It must include information about the phenomena in study referred to the whole population willing to analyse.

b) This information must be also updated to the date of reference of the study.

c) The record must have some identifier that enables the identification of each item (person or company) of the administrative act that they reflect, enabling the crossing with other sources so all necessary information for statistical production is obtained.

d) Total absence of duplicates and invalid records (for instance: empty units, deceases) is compulsory since they distort the reference population frame and may produce biased estimations.

The first requisite about availability of updated information of interest is fulfilled in most cases, as the records are created to witness at all time a certain administrative act that generally is the one we want to exploit statistically (for example: amount of residents in the Municipal Register of Inhabitants, income tax revenues in the Tax Office Database, discharges in Social Security Database, benefits in the Database of Social Public Benefits, degree of disability in the SDPD . . . ).
But lamentably, in terms of concepts, administrative definitions are not always equal to statistical ones. While producing regulated statistics this may be a serious obstacle for their use. In addition to this, sudden changes in legal regulations which permit modifications may risk the comparability of concepts considered in the register, but unfortunately statisticians can not do anything about this issue.

In general, statistical producers should weigh the decision of whether having data restricted to administrative definitions or lacking any kind of information derived from them. Their decision must be made by evaluating purposes, demands and needs of information. In any case, it will be a task of statistical producers to specify clearly the definitions adopted to avoid an inadequate use of the information.

In the particular case of statistics of disability and labour market in INE, the adopted concept of disability is the legal one which is more restricted than the one recognised by the World Health Organisation (WHO) that is internationally accepted. This limitation was discussed inside a technical group, which benefited from the collaboration of organisms, institutional speakers and representatives of persons with disability. In spite of the limitations inherent to the legal concept, there was a general consent in accepting the formal one as the best alternative as it enables a panoramic view of the labour situation of this collective and ultimately allows its analysis.

Related to the second premise of records quality, namely being up to date, generally the problem lies on the territorial and departmental division of the Public Administration which may generate problems or delays on the general update of the whole content of the record. Nevertheless, it is a question of the system structure whose solution resides in an adequate management policy.

Once the register has been updated and it has been decided whether the administrative concept will be assumed for its statistical use or not, the rest of incidents that may result from its use are of a technical nature and can be solved with the use of procedures and algorithms which let the producers:

1. Standardise and optimise (control of errors and duplicates) the content of the administrative register.

2. Link and combine files.

3. Obtain statistical quality data

The main technical aspects to consider about these three points are described below.
3. Standardisation and normalisation of administrative registers

To combine administrative registers it would be enough the existence of a unique identifier for every entity (person, company, household) precisely used to capture every relationship with the citizens and the Administrations.

But unfortunately, when putting this idea into practice many difficulties are found since not in all the cases this identifier is available –for example, for little children– and in many cases, though it is available it is not properly standardised –for example passports– (see Teijeiro Alfonsín, 2006). Due to its fiscal and contributory ends, traditionally in economical records (those of the Social Security and Tax Office) the identifiers present optimum quality which is not always the case of those of social data bases.

Although in the brand new records created today standardisation rules are already adopted so the identification is always guaranteed, in other cases we do not have at our disposal an adequate identifier. This happens often with registers related to old administrative acts that originally were registered in paper and afterwards digitalised.

Hence, harmonising the identification fields of the administrative records to join is an essential task. Some essential phases that enable to realise this task successfully are detailed in the following paragraphs:

3.1. Order number definition

Before the accomplishment of any task it is really convenient to assign an order number (NORDEN) to each of the components (rows or observations) of the file to treat. This order number will act as a univocal identifier of every observation of the record and will facilitate the detection and following treatment of incidents, especially duplicities. A good practice will be to construct an order number by assigning consecutive numeric codes and independent for each province or region.

<table>
<thead>
<tr>
<th>IDENTIFIER</th>
<th>PROVINCE</th>
<th>NORDEN</th>
<th>IDENTIFIER</th>
<th>PROVINCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NX011655-R</td>
<td>01</td>
<td>0100000001</td>
<td>NX011655-R</td>
<td>01</td>
</tr>
<tr>
<td>POX-014697</td>
<td>01</td>
<td>0100000002</td>
<td>POX-014697</td>
<td>01</td>
</tr>
<tr>
<td>X00241/86</td>
<td>02</td>
<td>0200000004</td>
<td>X00241/86</td>
<td>02</td>
</tr>
<tr>
<td>X00327/86</td>
<td>02</td>
<td>0200000005</td>
<td>X00327/86</td>
<td>02</td>
</tr>
<tr>
<td>521382901</td>
<td>03</td>
<td>0300000001</td>
<td>521382901</td>
<td>03</td>
</tr>
</tbody>
</table>

Figure 1: Assigning order numbers

3.2. Identifier standardisation

Codes that are officially in use for identifying persons or companies, are usually constituted by an alphanumeric string with a certain length and format.
There are some identifiers exclusively formed with numeric strings of a determined length, as occurs with the "Social Security number" for persons, or the "Social Security Account Code" for companies.

Some identifiers are composed by a numeric code together with some letters derived from an algorithm depending on the previous numbers. This is the case, for example, of the National Identity Card (ID CARD), the Spanish "Foreigners Identification Number" or the Spanish "Fiscal Identification Code" for companies. Finally, in some other cases, the identifier may be an alphanumeric string that does not follow any standard, as it happens with Passports and Resident Cards.

But in some registers the identifier code may have errors, and then a transformation is needed before working with it. If this phase is not carefully executed, every work that requires crossing files will result affected as no match could be performed for a same record existing in two files but coded differently, or on the other side wrong information would be incorporated for two elements with different identifiers but the same transcription (for instance produced by zeros in the left side, confusions between the number zero and the vowel "O", among others).

It is vital to establish an adapted procedure of standardisation to avoid this kind of errors.

In addition, there are cases in which it is preferable to have many candidates even if not all of them are valid than missing a good one. This is particularly useful in many cases. For example when working with registers with incomplete identification variables, registers with non available or not optimal numeric identifier or even those registers with good numeric identifiers but that that could be preferable linked by another different linking variable. In all these cases it is recommended to consider multiple candidates and afterwards use a selection algorithm to get the ideal one. Hereby one assures that if the individual in question exists in both files the linkage of both will be achieved. A procedure of standardisation of any alphanumeric string that follows the work scheme mentioned may consist of the following steps:

1. Firstly, every alphanumeric character is transformed into capital letters, any strange character is cleaned out from the string and the whole string is separated in as many particles as observed changes from numeric blocks to alphanumeric (or vice versa). This is discussed in Amón and Jiménez (2010) and Elmargarmid and Ipeirotis (2007).

2. In second place, identifiers can be assigned to each individual. It is recommended to consider as many identifiers as possible combinations with the numeric blocks extracted in the previous step.

---

1This is regulated in: ORDEN INT/2058/2008, de 14 de julio, por la que se modifica la Orden del Ministro del Interior de 7 de febrero de 1997, por la que se regula la Tarjeta de Extranjero.
3. If any alphanumeric dividers between two numeric blocks are an "O", one of the resultant numeric codes is obtained by replacing the vowel "O" with the number "0".

<table>
<thead>
<tr>
<th>NORDEN</th>
<th>IDEN</th>
<th>Numeric blocks</th>
<th>Numerical identifiers standardised</th>
</tr>
</thead>
<tbody>
<tr>
<td>1100000001</td>
<td>NOX9011655-R</td>
<td>0 9011655</td>
<td>9011655 - - -</td>
</tr>
<tr>
<td>1200000004</td>
<td>X000241/86</td>
<td>241 86</td>
<td>241 86 24186 -</td>
</tr>
<tr>
<td>1200000005</td>
<td>X000327-86</td>
<td>327 86</td>
<td>327 86 32786 -</td>
</tr>
<tr>
<td>1300000001</td>
<td>D07816O423</td>
<td>07815 423</td>
<td>07816 423 7816423 781604238</td>
</tr>
</tbody>
</table>

Figure 2: Standardisation: extraction of numeric standardised codes

Therefore every observation remains identified univocally with the order number (NORDEN) but with more than one numeric standardised code. These numeric codes correspond with the numeric part of the official identifiers and they will be used for joining with other records in which they count with standardised identifiers. The alphabetical part of code should be reserved to take decisions about the validity of the crossings when more than one candidate appears or to solve possible duplicities.

It should be noted that in the particular linkage case with two big registers are linked (for example, a Census and a Municipality register) the assignation of multiple numeric codes can strongly complicate the process.

3.3. Name standardisation

Although the most common case is the one that refers to names and surnames, this procedure can be applied to any alphabetical string, for example company names, addresses or street names.

To have at our disposal standardised names turns out to be indispensable in the use of administrative registers. The simple incorporation by mistake of a blank at the beginning of a name is an obstacle to identification. While for the human eye the identification is direct, any computer will treat both strings as different values.

In spite of the fact that this kind of disadvantages impede many comparisons, the application of a series of very simple rules clearly improves the comparability between alphabetical strings. The simplest example is the mere conversion to capital letters of all the letters of the field name. Besides this trick it is also convenient to:

1. Analyse the content of strings converting odd symbols that systematically are confused with certain letters. For example, letters "é" or "è" are systematically replaced by symbols and it is also frequent the systematical change of the "O" vowel to number zero "0".

2. Clean the string, substituting accents, dieresis (ÁÉÍÓ ÑÆÇÛ ÆÉÎÜ ÁÉÍÓ) and erasing non alphabetical symbols: (, < ( + &$ )-;::@=?)
as well as unnecessary spaces –double space– (see Amón and Jiménez, 2010; Elmargarmid and Ipeirotis, 2007).

3 Extract those particles that appear in the string but are not names in their selves (articles, prepositions, contractions). For addresses it is relevant to extract from their names all those particles referring to the type of route (St., Sq.) and store them in an independent field for possible unions with other files. Likewise, in the case of treating businesses it is pertinent to separate particles referring to the type of company (Plc, Ltd ...).

4 Have available dictionaries that enable transformations or codifications of abbreviations, initials, nicknames, diminutives, compound names and any possible lexical variations or composition of the considered names. The ideal situation is the one in which already exists a dictionary constituted with directories or censuses susceptible of being used as reference. But if we lack any, the ad-hoc manufacture of a basic one that considers those most frequent equivalences will serve to standardise a big volume of elements. With such little effort we ameliorate the comparability of fields in different files. Usual examples of these situations are abbreviations of the name "María" or types of road such as "Street", "Square" or "Boulevard".

4. Union and combination of files (record linkage)

Once the main fields that enable identification have been standardised (names and codes) the register is ready to be linked to other standardised files. If the registers to join have a high quality identifier, the mere union of them through this identifier is sufficient to combine them. Nevertheless, there are some cases in which the identifier does not guarantee the correct union of files (poor quality, subsets of individuals or entities that lack of an identifier) and it is necessary to adopt alternative solutions in which two steps must be taken:

- To select those variables and combinations of them to link files.
- Consider for every individual of the first file all the possible candidates in the second file that could connect with it, and afterwards select the optimal one.

4.1. Selection of the link or links of union

It is convenient to define diverse combinations and matching phases in a systematic and orderly manner depending on the reliability of the available identification variables. If we have at our disposal several types of identifier, the matching can be executed in several phases, that is, crossing in a first stage with the most frequent or reliable identifier and in successive phases with the rest of identifiers sorted in descending order of trustworthiness.
If more than one suitable name identification field is available, there exist two possibilities of linking. One of them is to perform several unions, that can be executed through those diverse available names. The other one, is to select just one or more key fields as link to join information and then use the others as auxiliary information for the algorithm that selects the optimal candidate. The most appropriate option must be based on the situation of the files and the aims of their union.

In the particular case of persons, usually name fields are made available and can be used: name, surname, second name, and others as can be the personal address. It is strongly recommended to use the conjunction of all variables name + surname (or name + primary last name + second last name for Spain and Hispanic countries) as the link variable, and leave the rest of variables (i.e.: address) for its later use in the optimal candidate choice.

When treating addresses the use of the name of the address as linking variable is highly recommended, while we suggest to leave other information about the address such as the route type for the optimisation algorithm.

As mentioned in the paragraph about standardisation, the comparison of alphabetical strings turns out to be pretty complex as the simple change or omission of a letter in a name blocks the correct IT identification. A common alternative used to save this obstacle is to use as linking variable between two files not the complete name but some of its initials.

It is noted that the use of initials is an alternative resource for the crossing of alphanumeric fields, after the failure of linking with numerical identifiers. It is applied after previous coding of names in order to obtain effective results and dismiss erroneous cases.

As a practical example, in order to join the persons of the SDPD with other files it is compulsory to treat the possible identifier. The available alphanumeric identifier can be the ID card number, the "Foreigners Identification Number" (NIE), passport code, a fictitious code or even be blank. Moreover, in this particular database there are codes corresponding to the legal tutors of some persons with disability. After a detailed process of identifier standardisation, the crossing of the file with other sources is carried out in 3 phases. In the first one, files are merged by the ID card or NIE. In a second phase, all those units that were not crossed in the first phase are crossed with the passport code as link. Finally, those units not yet crossed are merged by the alphabetic string composed with by the 3 first initials of the name plus the 3 first initials of primary and second last names.

4.2. Choice of the optimal candidate

Once the linking of files has been performed, for each individual of the first file all possible candidates in the second file are considered and then by the
A. Fresneda Pacheco, M. Pérez Julián

<table>
<thead>
<tr>
<th>Phase</th>
<th>Linking Variables Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National identity card and foreigner identity card</td>
</tr>
<tr>
<td>2</td>
<td>Passport and non standard codes</td>
</tr>
</tbody>
</table>
| 3     | 3 first name initials + 3 first surname initials  
       | or for Spain and Hispanic countries: 3 first name initials + 3 first primary last name initials + 3 first second last name initials² |

Table 1: Matching variables and union phases used in disability and labour market statistics.

selection of suitable algorithms the ideal candidate has to be chosen. For this purpose adequate distance or similarity functions are required.

Distance or similarity functions necessarily must provide a numeric value that indicates the degree of similarity between two alphanumeric strings (Amón and Jiménez, 2010; Elmargarmid and Ipeirotis, 2007).

Though in mathematics a wide range of tools can be used –for example, the Levenshtein distance (Levenshtein, 1966) or Jaro-Winkler (Winkler, 1990)– the evaluation of alphanumeric strings exceeds the mathematical area and has something to do with the lexicon. In the concrete case of administrative registers there is an extra difficulty because texts to treat are names and combinations of them so they do not follow any semantic rules.

In addition to this, the choice of a particular distance function is based on the type of variable to compare, as for example, the comparison of surnames differs from the one of dates of birth. Whereas in the first case the length, dyslexia and phonemes can be evaluated in the second one the comparison is restricted to the figures composing it and their order.

These distances or similarities can be very sophisticated, though complexity is not always a synonym of efficiency.

Such is the case of Spanish disability and labour market statistics, where to compare two names (of persons, companies, routes, . . . ) a complex distance function was constructed based on Jaro Winkler’s philosophy. That entails to compare alphanumeric strings by evaluating possible mistakes in the character transcription and weighting their similarity at the beginning (it is a common and frequent mistake to transcribe correctly the beginnings and during the transcription make mistakes, especially at the end). This handmade distance also extents the spectrum of possible transcription mistakes with the extra evaluation of words backwards written and also taking into account similarities in their pronunciation. Nevertheless, for the comparison of dates of birth a very simple distance is applied to evaluate dates similarity, just depending on the comparison of numbers and their position in the string. It is a discrete function that takes entire values from 10, in case of total similarity, up to 0 in case of total
Comparison | Description | Similarity d2
--- | --- | ---
`yyyymmdd` | Both dates fully coincide | 10
`yyyydmm` | Year coincides and month and day are swapped | 9
`Joker` | They fully coincide, but the date refers to the end/half or beginning of the year (i.e. 01/01/yyyy, 30/06/yyyy, 31/12/yyyy) | 8
`yyyymm00` | Year and month coincide | 7
`yyyy00DD` | Year and day coincide | 6
`yy00mmDD` | They only differ in the dozen unit | 5
`yyy0mmDD` | They only differ in the year unit | 4
`yyyy0000` | Only the year coincides | 3
`yyyy000000` | They differ in the unit and dozen of the year | 2
`yyyy0000000` | Only hundreds and dozens of the year coincide | 1
`000000000` | No coincidences found | 0

Table 2: Matching variables and union phases used in disability and labour market statistics.

discrepancy. In spite of its simplicity, its correct use provides a powerful tool for the correct identification of persons with disability in Spanish labour market statistics.

In this context of different nature of identifiers, it is recommended to have at disposal an adequate distance function for each identification field to be used for record linkage: one to evaluate alphanumeric strings (such as name and last names or route names), other to evaluate dates, other to evaluate codes of municipality, and so on.

Once the numeric values that indicate similarity or divergence are available, the selection algorithm for the optimal candidate is defined by establishing decision rules depending on combinations of scores obtained for each of the compared fields.

In this way, in the matching process of the SDPD before using an algorithm to obtain the ideal candidate, several phases of linkage are executed. Phase 1 and phase 2 are performed by using as linking variable the ID number or Foreigner Card and passport, respectively. Then, the algorithm uses as input the similarity value between names and surnames of the candidates and the similarity of dates of birth and municipality codes.

5. Obtaining statistical results

With the view of statistical production, the administrative record is itself a frame of reference for the investigated phenomenon. Because of this, on one hand it must have neither duplicates nor invalid units, and at the same time it can be used as source of auxiliary information for calibrating the sampling factors that shape the statistical estimators object of study.
5.1. Duplicates detection

The existence of duplicated units in an administrative register distorts the frame of reference leading to mistakes in the results. The correct treatment for this tricky situation is a big deal and needs technical skills both for their detection and resolution. A duplicate takes place when there are two individuals or entities that after the standardisation process of the identifier fields they both share the same identification.

Pure duplicates (two units with the same content for all the contemplated variable or field composing the file) are the easiest ones to treat. The solution consists on choosing only one of them, but when the element responds to a same identification and presents different values for some file fields, we talk about non pure duplicates and in this case both their detection and resolution are much more complex. If we have an identity code (ID card, Social Security number, Fiscal Identity Code...), duplicates can be detected with those units related to a same value before and after standardisation.

It goes without saying that if a same identifier presents several administrative situations in different dates, the unit recognised as valid will correspond to the one referred to the research date. But if besides having the same value for the identifier, two units present the same administrative situation in a certain reference date, to tackle the duplicate problem requires additional information. It is very useful to match one file with another reference one (for example, with an updated population directory) so the situation of the reference file determines which of those two units is the valid/good one.

If even after this crossing, still duplicate units remain, assuming the lack of any selection criterion it would be solved randomly. It should be reminded that these cases must be a residual group, so as the lack of rigour does not affect the final results.

As an example, the SDPD presents many cases of possible duplicities because a same person can be evaluated in different moments and different provinces and no united register exists for the whole territory. Neither we dispose of a disability assessment date for each register so the solution for duplicities takes as reference firstly the Municipal Register of Inhabitants and in second order the assessment dates available. Thanks to these aids only a 5.7% of duplicates remain unsolved and are treated with a random choice.

If we do not have any identifier at hand, duplicate detection must be done combining all the alternative identification variables. The most common case is to use the name (name and last names for persons, registered name for companies) together with a reference date (date of birth or creation). In addition to the use of name and reference date, a good idea is to perform an additional crossing with the reference directory so as to help in the resolution of duplicates.
Use of administrative records in official statistics: a practical view

<table>
<thead>
<tr>
<th>Case</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Municipality of residence in SDPD=</td>
<td>69.5%</td>
</tr>
<tr>
<td>Municipality of residence in the Municipal Register of Inhabitants</td>
<td>0.5%</td>
</tr>
<tr>
<td>Province of residence= SDPD =</td>
<td>19.5%</td>
</tr>
<tr>
<td>Province of residence in the Municipal Register of Inhabitants and it is the only duplicate that meets this condition.</td>
<td></td>
</tr>
<tr>
<td>Most recent disability assessment date</td>
<td>5.5%</td>
</tr>
<tr>
<td>Random solution</td>
<td>5.7%</td>
</tr>
</tbody>
</table>

Table 3: Algorithm of duplicate choice in SDPD with the aid of the Municipal Register of Inhabitants.

5.2. Unit validation

The purpose of using administrative registers is the achievement of statistical data about a certain phenomenon. After their identification fields have been standardised and before crossing it with other files we must decide which units to consider in the research depending on the reference date of the phenomenon of study and the field of study.

In the majority of cases, registers have a date field informing about the referential date when the administrative act took place, so depending on the date of study this variable enables to validate whether registers belong to the study scope or not.

With regard to the sphere of study, sometimes the administrative registers do not have variables that delimit the characteristics of the population to assess: they can lack date of decease, date of leave, date of birth or date of creation, etc. In these cases the use of a reference frame turns out to be essential as it allows to establish if a certain unit must be taken into account or be excluded instead.

In the example "The Employment of the persons with disability", the reference frame is the set of persons from 16 to 64 years old that reside in households. The administrative register used to identify those persons with disability is SDPD. This database has dates of birth, but lacks dates of casualties and a variable indicating whether the person lives in a household or a collective centre (hospital, residence...).

To obtain this information the SDPD is initially crossed with the Municipal Register of Inhabitants which is the population directory of reference and is used to check the birth of date in SDPD and to identify deceases. The SDPD is updated by erasing those persons out of age scope or that have deceased.

In addition to this source, it is also available one Directory of Centres which is annually updated and helps the standardisation of addresses and assignation of route codes. Also the Municipal Register of Inhabitants provides the total amount of persons living in a dwelling and if it is a collective or personal
A. Fresneda Pacheco, M. Pérez Julián

Non-acceptable cases | Detection
--- | ---
Deceases | With the Municipal Register of inhabitants
Age out of scope (<16 or >64) | With the date of birth existing in SDPD completed with the one listed in the Municipal Register of inhabitants
Persons residing in centres | With the Municipal Register of inhabitants in conjunction with the Directory of Centres

Table 4: Validation of observations in the official register of disability (SDPD).

household. The combination of the Directory of Centres (residences of elderly, hospitals) and the Municipal Register allows to take out every person living in collective dwellings from SDPD.

Once the units are validated depending on the frame of reference for the study, the next step is the statistical treatment to describe the phenomenon target of study.

5.3. Calibration of sampling factors

When treating with a representative sample instead of the whole population subject of study, statistical producers must use an estimator that enables the measurement of the population value willing to analyse.

The usual form of estimators is given by the expression

\[ \hat{X} = \sum_{i=1}^{n} F_i X_i, \]

where \( X_i \) is the variable value in study for element \( i \) of the sample and \( F_i \) the associated sampling factor.

Each sample element has assigned a sampling factor that equals the number of population elements it represents, so the sum of all factors must coincide with the whole population size. That is, the sample must be calibrated at least to the population.

If further information is available (that is, there are variables \( Z \) correlated with the studied phenomenon whose distribution can be approximated), calibration techniques, such as in Pérez Arriero (2008) are used so the estimator of these variables \( \hat{Z} = \sum_{i=1}^{n} F_i Z_i \) match its population value \( \hat{Z} = Z \).

It is obvious that if the process to be measured is the one that collects the administrative register, the register itself constitutes a population census enabling to know the whole of records or individuals that verify any of the study characteristics. Therefore, it can be employed as auxiliary information to calibrate sample results so that the new estimator reaches the goal of respecting the official registered counts.

For example, in the statistic "The Employment of persons with disabilities" (EPD), the objective is to count the total of persons with disabilities occupied, unemployed or inactive and to compare them with the rest of the population. It is evident that the sum of occupied, unemployed and inactive persons match
the total of them and this value is listed in the SDPD, so using it to calibrate sampling factors perfectly makes sense.

Likewise, EPD statistics is obtained from the information of the Spanish Labour Force Survey (LFS) restricted to the population from 16 to 64 years, so the total of occupied, unemployed, and actives in this age (sum of both those who have a disability and who do not) that provides LFS should be coherent with LFS’s results.

LFS it is at the same time a calibrated survey to the main population values obtained with the population figures obtained by the Spanish Statistical Office (INE). Also the total of households provided by LFS is a reference value applied as a marginal value in the calibration of other population surveys in INE.

Therefore for the construction of EPD statistics a calibration procedure is carried out taking in the following way:

- Repeating exactly the calibration of LFS, fitting the population in a NUTS3 level (regions), groups of age and sex and nationality (Spanish / foreigners).

- Additionally to the LFS calibration, marginal variables for calibration at NUTS3 level are taken so as to guarantee total coherence with LFS results: the amount of occupied, unemployed and inactive persons according to LFS detailed by sex for people in labour age guaranteeing then total coherence with LFS results.

- Also, marginal variables for calibration are added based on the total of persons with disability of the SDPD by degree and type of disability, by sex and three groups of age, so that the population with disability in labour age fits the total amount officially registered.

6. Future view

The use of administrative records in official statistics is nowadays a reality that completes information gaps with a low economic cost but entails a long complex set of processes.

The availability of bigger volume and diversity of registers will undoubtedly influence the way of producing official statistics. In many cases, it will enable to jettison many surveys and censuses so costs and burden to respondents will be significantly reduced. Although official statistics may not be done without some surveys, especially social ones as only by asking direct questions the self-perception of the individual can be known, having registers at our disposal will enable to design shorter questionnaires than current ones. In addition to this, the quality of results may improve if the questionnaires previously include registers information so the interviews are focused on different aspects according to previous data. Also, the reduction of surveys and their questionnaires length will open new research possibilities, enabling to conduct research into new variables
and topics which so far could not be deeply analysed due to lack of resources or because they were of second priority in current overburdened questionnaires. A suitable design and a correct structure of the diverse phases that compose procedures to treat administrative registers enable their use as a trustworthy source for statistic use. In order to achieve statistical applicability, mutual collaboration of the different administrative units involved in the production and management of any possibly usable register is absolutely essential.

Nowadays the bases of the future system are being settled with the implementation of procedures that unify identification variables and combine information from different sources together with statistical tools that provide accurate and consistent estimators. In this context, the new routes to obtain results have enabled a statistical approach to some phenomena that till today lacked of any structural information for those periods where priorities were others and the budget was tight.

If possible, the participation of National Statistics Offices in the construction of registers to unify standardisation criteria and even collaborate in registers management, would enable to establish a set of core variables common to all registers of the same type (persons, households, companies). This will ease the linking of diverse information and will ameliorate not only the statistical production but many administrative processes and resources of public organisms, as well as avoid many bureaucratic paperwork to all citizens. Several works in this way have already begun in Public Administration, but the greatest advantage for the future would be to become effective for the whole of National Administration organisms and also at international level so as to reach the desirable goal of making possible any comparison among different countries. The future of statistics is undoubtedly focused on the major availability of information in digital format that, adequately managed, will allow to have a complex system of interrelated information.

This data integration process not only will improve and extend all current statistics but also will allow continuous, coherent and comparable results at any level (for instance: sectorial analysis and international comparisons).

References


About the authors

Amelia Fresneda Pacheco is Head of Unit of the Employment of Persons with Disabilities Area and previously worked as Head of Unit of the Labour Cost Area in the National Statistics Office of Spain. Among her duties are the research and development of statistical techniques for the integration of structural and temporary/short-term results and for the calculation of retrospective series. During the last years her work has been centred in the investigation of record linkage tools and calibration techniques and its application for the combination of administrative data with Labour Market surveys.

Marina Pérez Julián is Senior Statistician in the Employment of Persons with Disabilities Area and previously worked in the Educational Statistics Area in the National Statistics Office of Spain. Up till now her work has been focused on data analysis, modelling and the research of record linkage tools and calibration techniques for linking administrative data of small populations with consolidated Labour Market surveys.