# Measurement errors in geographical labour mobility using data linkage: the Spanish case

**Abstract:**

This paper analyses reliability and accuracy of the relationships between migration and employment status, when they are estimated using a linked data set. The analysis will be carried out using a new source, the Labour and Geographical Mobility Statistics, which is provided by the Spanish Statistical Office. This statistic is constructed by exact matching procedure, linking the Labour Force Survey with the Official Population Register. The findings reveal that in order to study accurately geographic labour mobility, timing and geographical coherence between the two data sets must exist, if not, causality relations between the labour market and geographical mobility cannot be properly analysed. Although our analysis is referred to the microdata quality and internal coherence of that new Spanish statistic, this conclusion can be extended to any linked data set for studying geographic labour mobility, including the population censuses mainly based on administrative data.

**Keywords:** LFS, Population Register, geographic mobility, labour market, record linkage.

**JEL Classification:** J61, C81 and Y10

## 1. Introduction

According to Künn (2015), data linkage can supplement highly reliable administrative records with survey information that is crucial for statistical analysis but is usually unobserved in the administrative information. As a result, the linked data set contains a large number of variables providing an optimal data source for statistical analysis. It can also afford new insights into methodological questions regarding the validation of survey information with administrative records, *vice versa* or simultaneously[1]. Data linkage may also lead to shorter interviews, less respondent burden, and an overall reduction in survey costs. Furthermore, administrative data can potentially offer a significant increase in the number of auxiliary variables that may be used for nonresponse bias adjustment (Sakshaug *et al*. 2012, 2017).

[1] See, for example, Oberski *et al*. (2017) or Angel *et al.* (2019).

The aim of this paper is to analyse the reliability and accuracy of the relationships between migration and the employment status, when they are estimated using a linked data set. The analysis will be carried out using the *Labour and Geographical Mobility Statistics* (LGMS), which is provided by the Spanish Statistical Office (*Instituto Nacional de Estadística*, INE). The LGMS is constructed by linking the employment information of those people surveyed by the Spanish *Labour Force Survey* (LFS) with the residential mobility data drawn from the Spanish *Population Register* (PR). As a result, the relationships between geographical and labour mobility can be studied, taking advantage of the extraordinary amount of personal and employment information contained in the LFS.

Our mainly contribution to the literature is the following. In order to study properly the relationships between migration and labour market status using a linked data set, timing and geographical coherence between the two data sets must exist. If that coherence does not exist, then causality relations between the labour market and geographical mobility cannot be properly analysed. Although our analysis is referred to the microdata quality and internal coherence of a new statistic produced in Spain by merging survey and administrative data, this conclusion can be extended to any linked data set for studying geographic labour mobility, including the population census constructed exclusively through the register linkage method – as 15 National Statistical Offices have planned to do for the 2021 census round (UNECE, 2019).

The paper is structured as follows. After a brief review of methodological aspects about linked data set for studying geographic and labour mobility in Section 2, the internal coherence of the LGMS is addressed in section 3. This coherence is achieved by carrying out an exhaustive analysis of the concordance of the information that is simultaneously drawn from both sources regarding the *place of residence in the previous year*. As we will prove that

54 the concordance of information on recent migration is very low, we will try to separate the
55 theoretically coherent records in the sample from the rest, and determine whether there are
56 any significant differences between the two groups of records. Provided that the coherent
57 records faithfully represent the whole sample in which there is mobility, it is possible to use
58 only this part of the information. However, if the characteristics of the coherent individuals
59 differ from the rest, the estimated relationships between migrations and the labour market
60 will be biased. This is what Section 4 seeks to determine. The paper ends with the main
61 conclusions and recommendations.

62 **2. Methodological aspects about linked data set for studying geographic and labour mobility**

63 In Europe, since the beginning of the 2000s, both the Statistics Offices of the Nordic countries
64 (Denmark, Finland, Iceland, Norway and Sweden) and the Netherlands *Centraal Buerau voor*
65 *de Statistiek* have been gaining extensive experience in the use of administrative records to
66 produce official statistics (Kupiszewska et al., 2010). The experience has gradually extended –
67 promoted by institutions such as the UNECE (2007, 2011) and the EUROSTAT (2013)- and,
68 currently record linkage is a normal procedure in the official production of statistical
69 information.

70 In the USA, the Census Bureau operates with a congressional mandate to use
71 administrative records to improve, cut costs associated with, and reduce the respondent
72 burden on Census Bureau surveys such as the *American Community Survey*, as well as the
73 decennial census (Foster *et al*. 2018 a, b). In fact, in 2012 the linking procedures had already
74 been suggested to reduce the lack of response in the 2010 Census (Rastogi and O'Hara 2012)
75 or they had been used to generate longitudinal samples from the Current Population Survey
76 (Rivera-Drew *et al*. 2014). Additionally, throughout 2018, the interest in linking sources

77 resulted in a large number of working papers by the Center for Administrative Records
78 Research and Applications (CARRA) of the US Census Bureau.

79 Despite the increase in the production of linked data set, there is little experience of
80 linking surveys and administrative registers in order to study relations between the labour
81 market and geographical mobility. An example is Goetz (2017) for the US. He presents a novel
82 data set combining survey information from the *American Community Survey* (ACS) with
83 administrative data on employment from the *Longitudinal Employer-Household Dynamics*
84 (LEHD) database, with the aim of studying geographic labour mobility. Using a personal
85 identifier, an ACS individual is linked to all of their LEHD-covered jobs, and the crucial piece of
86 information for the purpose of measuring job migration is in the geographic location of the
87 employee's workplace establishment. Goetz states that the existing US surveys that include a
88 longitudinal component, such as the Current Population Survey, do not follow an individual
89 after they leave their residence, making it impossible to determine where the individual
90 moved to.

91 The ACS-LEHD database provides detailed information about the respondent's
92 situation only during the time before and during their move. The individual data from this
93 survey (demographic, labour, income and household information) are available when the
94 survey is carried out, whereas the data from the records – which are updated on a quarterly
95 basis – are available for the following period. When the survey respondent appears in the
96 post-compiled employment records with a job in a different place to the survey location, a
97 migration has taken place. Thus, with the ACS-LEHD the forward migration sequence may be
98 studied, provided that the respondent continues to have a job or has just found one, switching
99 their status from being unemployed to being employed. However, the relationship between

100 unemployment and mobility cannot be analysed using the ACS-LEHD, because in order to
101 appear in the administrative records, individuals have to be employed.
102 In Europe, information from the LFS does not offer this drawback, because when the
103 backward sequence is contemplated, the LFS provides the demographic, employment and
104 household information of individuals at the present moment and that of their labour aspects
105 in year before[2]. Thus, if the LFS is combined with a suitable population register, which allow
106 researchers to track previous residential mobility, all links between the labour market and
107 geographical mobility can be established, irrespective of the employment status. Therefore,
108 the relationship between unemployment and geographical mobility could be perfectly studied.
109 Not all European countries can make that combination, depending on the existence or
110 not of a suitable population register. Spain is one of the European countries whose population
111 register allows you to do that. Specifically, the *Municipal Population Register* is the
112 administrative register of inhabitants in a municipality (usual domicile) and its data constitute
113 proof of residence. The Town Councils are responsible for maintaining these registers and are
114 required to communicate monthly variations to the INE (Acts 4/1996 and 7/1985) so that this
115 institution can carry out checks to correct errors and duplicates. Based on this information,
116 since 1996 the INE has generated a centralized file, the *Continuous Population Register* (in
117 Spanish, *Padrón Continuo*). Registration of changes of residence is compulsory in Spain, but
118 there is no guarantee of complete coverage of flows since compliance depends on the
119 (dis)incentives for registration of movements. To register, the only documents required are
120 proof of identity (identity card, driving license, passport, etc.) and some kind of evidence of

[2] It must be noted that the LFS also contains some retrospective information about the previous place of residence but, in the majority of EU countries, the LFS sample design leads to a systematic underestimation of mobility, in the case of recently arrived immigrants (see Martí and Ródenas, 2007, 2012). We will not address this here because in the LGMS the administrative register determines the mobility, not the LFS information.

residence at the address (title deeds, rental agreement, utility bills, or a letter from the first adult already registered at the address). When a registration is accepted, de-registration at the previous municipality is generated automatically. A local residency certificate is required to access basic services such as public education or health care, to vote in elections, to renew identity cards, to obtain grants, public employment, parking permits or home purchasing grants. In the immigrant regularization programs, this certificate was considered as proof of residence in Spain for illegal foreign immigrants. The PR is continuously updated and annually disseminated. It covers all persons living in the national territory

Given the characteristics of the PR, the Spanish Statistical Office can merge the residential mobility data drawn from the PR with the employment information of those people surveyed by the Spanish LFS[3]. This statistic, the LGMS, is produced by *exact matching* procedure. Basically, there are two procedures to match record pairs. Both require the two data sets to contain overlapping information. The most straightforward technique is based on one or more unique personal identifiers, such as Social Security or identity card number[4]. In this ideal case, the data can be directly linked, usually with almost no error. In the literature, this procedure is known as "exact matching" (Künn, 2015). The other technique, probabilistic linkage[5], is used when there are no unique personal identifiers or when they are not reliable

[3] The Spanish LFS has been conducted by the INE since 1964 and covers the entire national territory. This quarterly survey is aimed at population living in family dwellings, and intended to collect data on the labour force and its various categories (employed, unemployed) as well as on the population outside the labour market (inactive). The quarterly sample size is taken from about 65,000 dwellings, equivalent to approximately 160,000 people. The Spanish LFS is completely harmonized with the standards established by the EUROSTAT for the UE-LFS.

[4] Other examples of unique identifiers relating to people are tax identification numbers or health insurance numbers, for example.

[5] See, for example, Zhu *et al*. (2015), Herzog *et al.* (2007) or ASPE (2002). From a theoretical point of view, the specific problems that may appear when surveys are linked with administrative records using these techniques can be found in Blom and Korbmacher (2018), Sakshaug (2018a, 2018b) or Di Consiglio and Tuoto (2018).

in the overlapping information. In that case, nonunique identifiers (like first and last name or date of birth) are used to obtain likely record pairs.

Particularly, in the Spanish LGMS, the personal identification card number (*Documento Nacional de Identidad* -DNI- for nationals or *Número de Identidad de Extranjero* –NIE- for foreign nationals) (INE, 2013, p. 12) is used in the matching procedure. According to Spanish law, this may be done without requesting consent. In this way, the INE directly links the personal data of all of the responders in the LFS over the age of 16 with the corresponding records in the PR. Therefore, in this particular case, there is no reason to expect any bias due to a lack of consent.

Non-consent errors may appear in those countries where respondents are explicitly asked for permission to link their survey information to the corresponding administrative records, and not all of them agree to the linkage. Sakshaug and Antoni (2017) and Gessendorfer *et al.* (2018) discuss about the consequences of non-consent. Its implications are twofold. First, the effective sample size is reduced and thus standard errors of estimates are inflated. And, second, systematic error in the linked-data estimates can arise if respondents who consent to the linkage are different from those who do not. In other words, precision and bias problems arise[6].

The exact or deterministic linkage procedure is the least prone to error, but some kind of error can occur. Sakshaug and Antoni (2017, p. 561) point out that "a unique identifier obtained from the survey respondent may be incomplete or recorded with error and thus may not correspond to the same unit located in the administrative database". Moreover, these identifiers may not be centralized and as Sakshaug and Antoni (2017, p. 565) said "a lack of

[6] See Abowd *et.al* (2018) or Sakshaug *et.al* (2012) for a discussion on the biases that the requisite of personal consent can generate, even with exact linkage.

coordination may result in numbers being issued more than once". In addition, "survey respondents may remember and report their originally issued identification numbers correctly, but still provide invalid numbers, because they are unaware that their individual numbers have been changed within the administrative records". This kind of problem is insignificant in the LGMS for two reasons. First, the LFS sample is directly drawn from the PR, which contains the unique identifier. Thus, the respondent does not provide the identifier. Second, the Spanish identifiers (like in Nordic countries) are issued very early in a person's life -for everyone older than 16 years-, are time-consistent and used by the whole Spanish Administration. Thus, as expected, the linkage rate between the LFS and the PR is 99.4% (INE 2018); that is, 99.4% of the people over 16 years old surveyed by the LFS have also been found in the PR.

From this methodological point of view, the linkage process in the LGMS is high quality because it is used an exact matching procedure, non-consent errors are not expected and identifiers problems are irrelevant. Therefore, this new source is a good example to prove the accuracy of the relationships between mobility and labour status when it is used the linkage procedure.

**3. The internal consistency in the LGMS: place of residence in the previous year**

The accuracy of the estimations of geographic labour mobility depends on the internal consistency in the dataset in relation to date, origin and destination of the movements.

The LGMS is carried out on a yearly basis and incorporates some specific variables derived from the PR into the sample of the LFS of the first quarter of each year. The individual records of the LGMS should contain identical information in the comparable constructs drawn from both statistical sources. This is the case of the *place of residence in the previous year*.

183 In the LFS the mobility question is "*What was your municipality of residence exactly*
184 *one year ago?*". People can respond either the same or different, in which case they must
185 specify whether it was in Spain or abroad. The PR information comes from the last individual
186 registration in the PR (place –origin and destination- and date). It is expected that, if the
187 majority of people have been interviewed in their place of registration and there are no delays
188 in the inscription, the total of those who are in the current municipality for less than one year
189 will be equal to the total of those who have been registered in the PR of their municipality for
190 less than 12 months.

191 Table 1 shows the total numbers of people by place of residence in the previous year
192 from both sources. It can be seen that every year the PR mobility more than doubles that of
193 the LFS[7]. Moreover, if the two migration variables are crossed, additional problems arise as
194 not all of the records coincide with their classification. Reading by columns and in relative
195 terms, in general only 20-24% of those who have been registered for less than one year in the
196 PR have declared in the LFS that they resided in a different place one year before. For example,
197 for 2016 only 199,700 people coincide, representing 23.6% of the 846,267 that showed
198 mobility according to the PR. In reality, the majority - between 76% and 80%- of the population
199 that have been registered for less than one year, stated in the LFS that the previous year they
200 were residing in the municipality where the survey was conducted. Reading by rows in Table
201 1, an average of around 50% declared in the LFS that they resided in a different municipality
202 the year before while being registered in the PR for more than one year in the municipality. In
203 2016, for example, 212,369 people were in that situation (51.5% of 412,069).

[7] The focus of this paper is not to assess which data set measures mobility better. A discussion of this issue can be found in Martí and Ródenas (2004, 2007, 2012) or Ródenas and Martí (2009). Here our objective is to explain the reasons for the differences in the mobility estimations provided by the PR and the LFS based on an identical set of individuals.

[Table 1 near here]

The only justification for these results is that the two initial assumptions are not confirmed. This means i) that there may be a significant lack of coherence between the place of registration and that of the survey and ii) that there may be significant delays in the inscription. To assess the first assumption, the only information provided by the INE (2018) is that the linkage rate between the LFS and the PR is 99.4% of those surveyed. This only means that the inscriptions in the PR of almost all of the people interviewed for the LFS in the first quarters of each year were found, but it is not the proportion of the LFS records that have been effectively found and surveyed in the municipality in which they are registered[8]. The observation of this assumption is fundamental. If the *spatial coherence* rate were not sufficiently high, the quality and reliability of the LGMS would be brought into question. Let us consider an example: an unemployed person, interviewed by the LFS in Madrid who does in fact reside in this city but is and has been registered since birth in his or her city of origin, for example Barcelona. Obviously, if it is not taken into account that the place of the interview and the place of registration are different, when the labour data are linked with those of the PR, it may be wrongly concluded that unemployed people have a low propensity to emigrate, when the situation is precisely the opposite.

For the population aged sixteen years or over, Table 2 contains all of the situations that may arise as a result of the crossing of the possible values of the LFS and PR mobility variables, distinguishing the place of origin of the migratory movement. In this way, those who declare in the LFS that they have changed residence with respect to the previous year are disaggregated in accordance with whether they were previously residing in Spain or abroad.

[8] Although, on more than one occasion in private conversations, the INE has confirmed a very high percentage of coherences between the municipality of the LFS survey and that of residence according to the PR, they have not published this data or provided any figures.

Also, for those who have been registered for less than one year in the municipality, the PR variable that establishes the relationship between the current municipality and the previous one has been used to distinguish those who have never moved from their place of birth, from those who have come from Spain.

[Table 2 near here]

In Table 2, the 18 coherence possibilities established are shown, which leads to three main types of records: those that are coherent, those that are not coherent but explainable, and finally, those that are incoherent. The coherence that can be guaranteed is limited to those records in which the LFS and the PR coincide both in terms of the time period that has elapsed since the movement (more or less than one year) and the place of origin (Spain or abroad). The assumption here is that when there is time and geographical coherence, there is no reason to believe that there would not also be coherence between the place of the survey and the place of registration.

Under this assumption, the coherent records (white boxes) will be made up of the 4th, 5th, 6th, 8th and 15th groups. The first three are those without mobility in either of the two sources and the second two are those with mobility declared in the LFS and confirmed geographically and in terms of time by the inscription in the PR. Meanwhile, the records that are not coherent but explainable (light grey boxes) will be divided into three sub-types. First that of non-coherent records that are easily excusable due to the existence of two forms of delay in the inscription in the PR (2nd and 3rd groups and 10th, 11th and 12th groups) and those who did not de-register before emigrating abroad and who have now returned (16th, 17th and 18th groups).

The individuals in the 2nd and 3rd groups declare that they reside in the same municipality as one year before, but they have been registered for less than 12 months. If the

place of registration and the LFS interview coincide (as stated by the INE), they would be immigrants who had arrived more than one year before but who had not registered when they arrived, but during the last 12 months; in other words, with a delay. If the place of the interview does not coincide with the place of registration, there is no other alternative than to consider that these are *false registrations*. In the 10th, 11th and 12th groups, the individuals declare that they have recently moved, but according to the PR they have been registered in their current municipalities since they were born or for more than one year. In this case, the most rational option is to think that the place of the survey and that of registration are not the same. These are recent immigrants who have still not registered in the PR because they have not had the time or need to register in their new municipality. However, if the place of the survey and registration coincides, this situation can only be explained by a high level of *circularity* (coming and going) migrations with no change in the initial inscription in the PR.

The last three groups of records -16th, 17th and 18th-, correspond to those who declare that they had been residing abroad one year before, but who have been registered for more than one year (or since their birth) in a municipality in Spain. This type of inconsistency could be justified by short round trips, which is why these people have never de-registered in Spain. In the case of foreigners, the mismatch could be explained by short temporary stays in their places of origin. For the Spaniards, the explanation may be very different. Their departure abroad and their fast return (without de-registration from the PR) may correspond to the young Spaniards who left the country due to the economic crisis, were not successful in their destination and have now returned.

Finally, the series of *incoherent* records (dark grey boxes) are made up of the 1st, 7th, 13th, 9th and 14th groups. In the first three groups, an absurd situation arises because nobody over the age of 16 can be registered in the PR since birth if the inscription was made less than

12 months before (whether they have declared mobility or not in the LFS). In the last two groups, incoherence arises because the individuals have declared that they have moved recently and have also recently registered in the PR, but the places of origin (Spain or abroad) do not coincide in the two sources.

We have estimated the volume of these eighteen groups for the period 2010-2016 using the LGMS microdata[9]. Table 2 shows their annual average distribution in terms of the total population. It should not be surprising that practically the whole population is concentrated in the groups without declared mobility in the LFS and registered in the PR for more than 12 months, given the traditionally low migration rate of the Spanish population. These three groups, 97% of the population of the LGMS, are classified coherently by the LFS and the PR, assuming the condition *the place of the survey coincides with the place of registration* is met.

However, this absolute predominance of the apparently coherent records masks the degree of discrepancy between the LFS and the PR when there are migrations. If the percentage distribution is calculated only for those individuals who had mobility (the 4th, 5th and 6th groups are omitted), the percentage of the population with coherent mobility information in both sources rises to 16.5% (the sum of the weights of the 8th and 15th groups, in brackets in Table 2).

With respect to the records considered as being *not coherent but explainable due to the existence of a possible delay in registration*, the 2nd, 3rd, 10th, 11th and 12th groups represent the majority of those classified. In particular, there is a strong concentration in the 2nd and 3rd groups (around 64.4%), referring to those who have declared being in their current residence for more than one year but who have been registered in the PR for less than

[9] The microdata files can be obtained on request at the INE.

12 months. Geographical coherence exists in these groups because the place where the LFS interview takes place coincides with that of registration. The records of the 10th, 11th and 12th groups can be reasonably explained when we accept that these individuals have just moved but have not yet registered in the PR of the new municipality; thus, the place of the interview does not coincide with that of their registration.

In the case of the other type of reasonable lack of coherence, those who declare that they had been residing abroad one year before but never de-registered from their municipality of residence (the 16th, 17th and 18th groups) account for no more than 2.4%.

Finally, the clearly incoherent records of the 9th and 14th groups together account for 0.9% in total. All of the incoherencies in the 14th (9th) group correspond to individuals who declare in the LFS that they were residing abroad (in Spain) one year before, but that in their recent inscription in the PR less than one year before they indicated that their previous residence was in Spain (abroad).

Only under the *ad-hoc* afore-mentioned assumptions, can the migratory information of the LFS and the PR be considered as being coherent for practically all (99.1%) of the LGMS sample in which there is mobility. Specifically, for 16.5% of the records (8th and 15th groups) consistence exists *per se*, but for the remaining -82.6% which are non coherent but explainable-, it would be necessary to acknowledge that there is a delay in the registration in the municipality of residence (2nd and 3rd groups, 64,4%); that the current place of residence does not always coincide with the place of registration (10th, 11th and 12th groups, 15.7%) and, finally, that movements abroad with non-registered returns in the PR have also taken place (2.5%, sum of the 16th, 17th and 18th groups).

Nevertheless, using the previous arguments to give internal coherence to the migratory information of the LGMS seriously weakens the capacity of this source to capture

the sign and intensity of the geographic labour mobility. The reason is simple: for the great majority of the individuals interviewed, the timing of the migratory information does not coincide with the timing of the labour information. This means that the accuracy of the estimated links between mobility and the labour market situation cannot be ensured.

## 4. Differences between the sub-samples of the coherent and incoherent records with mobility

This problem of internal consistency would not be serious, if the coherent records faithfully represented the whole of the sample showing mobility. However, as the characteristics of the individuals making up the coherent (16.5%) and the rest of the groups (83.5%) differ, the estimated relationships between migrations and the labour market will be biased. In this section we will focus on assessing the similarity in the distribution of some selected individual characteristics (age, relationship with the household reference person, gender, place of birth and years of residence in Spain) and some labour market related ones (employment, time working in the company, type of contract, job search time and recipients of unemployment benefits).

Figure 1 summarizes the main categories of all these characteristics. We can see that *age* and *relationship with the household reference person* are the only variables that do not show substantial differences between the two main groups. However, the similarities end here.

[Figure 1 near here]

In order to statistically determine whether the afore-mentioned characteristics are homogeneously distributed between the group of coherent records and the rest, different measurements and significance tests have been used Once it was confirmed that the two quantitative variables (*years of residence in Spain* and the *length of the relationship with the*

346 *company*) are not normally distributed between the two groups of records, a non-parametric
347 test for two independent samples, the Mann-Whitney-Wilcoxon "U" test, was performed. The
348 same test has been used for the ordinal categorical variables, *age* and *length of job search*. In
349 the four cases, significant differences have been found in the distribution of the characteristic
350 depending on whether or not the records belong to the coherent group.

351 Finally, the non-parametric test used for the rest of the seven nominal categorical
352 variables was the Pearson's $\chi^2$. For all of them, the null hypothesis that the characteristic is
353 distributed homogeneously between the two groups of records was rejected. Since, by
354 definition, the $\chi^2$ statistic tends to establish that there are differences when the samples are
355 very large, some additional measurements based on the value of the $\chi^2$ have been used. These
356 either reduce the effect of the sample size in the estimated relationship (contingency
357 coefficient, *Phi* (ϕ) and Cramer's V), or are based on the proportional reduction of the error
358 (*Lambda*, Goodman and Kruskal´s *Tau* or Theil's uncertainty coefficient). When using these
359 measurements it can be observed that the differences between the coherent records and the
360 rest are statistically significant.

361 Taking the differences and its statistically significance into account[10], it is obvious,
362 therefore, that the strictly coherent group does not represent the whole of the LGMS sample.
363 This means that if only the sub-sample of coherent records is used, significant biases will arise
364 when the relationship between mobility and employment variables is estimated.

365 **5. Conclusions**

366 The aim of this paper is to analyse the reliability and accuracy of the relationships between
367 mobility and labour market, when they are estimated using a linked data set. Although, there

[10] A level of significance (α) of 0.05 has been used and the records of the sample have been weighted for all observational units in LGMS. The p-value associated to each estimator is equal to 0 in all the proofs.

368 is little experience of linking surveys and administrative registers in order to study geographic
369 labour mobility, the new source, LGMS, is a good example to prove that because of the high
370 quality of the linkage process. Particularly, LGMS is constructed by exact matching procedure,
371 non-consent errors are not expected and identifiers problems are irrelevant.

372 The accuracy of the estimations of geographical labour mobility depends on the
373 internal consistency in the dataset. In that sense, an exhaustive analysis has been made of the
374 information regarding the recent migration.

375 In order to confirm the coherence in the *place of residence one year before*, it has been
376 necessary to assume that for 64.4% of the sample there is a delay in the registration in the
377 municipality in which the individual effectively resides, that for another 15.7% the current
378 place of residence does not coincide with the place of registration and, finally, it must be
379 accepted that for 2.5% of the records, movements have been made abroad with returns that
380 have not been registered in the PR. Therefore, for the majority of the LGMS records, the timing
381 of the migratory information not coincides with the timing of the employment information.
382 Consequently, a satisfactory level of accuracy of the estimated associations between mobility
383 and employment status cannot be guaranteed.

384 However, the problem of internal consistency timing would not be serious if the
385 coherent records faithfully represented the whole sample with mobility in the LGMS. Of the
386 ten characteristics studied, only the variables of *age* and the *relationship with the household*
387 *reference person* do not graphically reveal substantial significant differences in the distribution
388 of their main categories between the group of coherent records and the rest. But the
389 similarities end here. None of the other variables are similar enough. Moreover, the statistical
390 tests used determine that these variables are not distributed homogeneously between the
391 coherent group of records and the others. Given these results, the use of the LGMS, at least

while its methodology remains the same (that is, without confirming that place of PR residence is the place of LFS survey) is not advisable.

In general, the findings reveal that in order to study accurately geographic labour mobility using linking data set, timing and geographical coherence between the two data sets must exist, if not, causality relations between labour and geographical mobility cannot be properly analysed.

Finally, we can also extend this conclusion to the population censuses that are mainly based on administrative data. The variable "place of residence" must be absolutely reliable, if not, the correlation between mobility and employment -or any other individual features- will not be accurate. Moreover, the demographic and socioeconomic features of population, which theoretically belong to a certain area, will be completely incorrect.

## References


Abowd, J. M., Schmutte, I. M. & Vilhuber, L. (2018). *Disclosure limitation and confidentiality protection in linked data* (Working Paper No. 18-07). Retrieved from Center for Economic Studies, U.S. Census Bureau website: https://ideas.repec.org/p/cen/wpaper/18-07.html.

Angel, S., Disslbacher, F. & Humer, S. (2019) What did you really earn last year?: Explaining measurement error in survey income data. *Journal of the Royal Statistical Society*, A, 182 (part. 4), 1411-1437.

ASPE (2002). *Studies of welfare populations: data collection and research issues. Two methods of linking: probabilistic and deterministic record-linkage methods*. Retrieved from U.S. Department of Health & Human Services website: https://aspe.hhs.gov/report.

Blom, A. G. & Korbmacher, J. (2018). Linking survey data to administrative records in a comparative survey context. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research*, (pp. 267-273). Doi: 10.1007/978-3-319-54395-6_34.

Di Consiglio, L. & Tuoto, T. (2018). Population size estimation and linkage errors: the multiple lists case. *Journal of Official Statistics*, 34(4), 889-908.

EUROSTAT (2013). *Microdata linking - international sourcing*. Retrieved from Eurostat webpage: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Microdata_linking_-_international_sourcing.

Foster, T.B., Ellis, M. & Fiorio,L. (2018a). *The Opportunities and Challenges of Linked IRS Administrative and Census Survey Records in the Study of Migration*. CARRA Working Paper Series nº2018-06. Retrieved from US Census Bureau website: https://www.census.gov/library/working-papers/2018/adrm/carra-wp-2018-06.html.

Foster, T.B., Ellis, M. & Fiorio,L. (2018b). *Foreign-Born and Native-Born Migration in the U.S.: Evidence from IRS Administrative and Census Survey Records*. CARRA Working Paper Series nº2018-07. Retrieved from US Census Bureau website: https://www.census.gov/library/working-papers/2018/adrm/carra-wp-2018-07.html.

Gessendorfer, J., Beste, J., Drechsler, J. & Sakshaug, J.W. (2018). Statistical matching as a supplement to record linkage: A valuable method to tackle nonconsent Bias?. *Journal of Official Statistics*, 34(4), 909-933.

Goetz, C. F. (2017). *The Potential for Using Combined Survey and Administrative Data Sources to Study Internal Labor Migration* (Working Paper No 17-55). Retrieved from Center for Economic Studies, U.S. Census Bureau website: https://ideas.repec.org/p/cen/wpaper/17-55.html.

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

INE (2013). *Estadística de Movilidad Laboral y Geográfica. Explotación estadística de la Encuesta de Población Activa y de la Base Padronal del INE. Metodología y descripción general de la operación*. Retrieved from INE website: https://www.ine.es/metodologia/t22/t2230209.pdf.

INE (2018). Labour market and migration statistics. Standardised Methodological Report. Retrieved from INE website: https://www.ine.es/dynt3/metadatos/en/RespuestaDatos.html?oe=30209.

Kupiszewska, D., Kupiszewski, M., Martí, M. & Ródenas, C. (2010.) *Possibilities and limitations of comparative quantitative research on international migration flows* (Working Paper No4). Retrieved from European Commission webpage: https://ec.europa.eu/migrant-integration/librarydoc/prominstat-working-paper-no-04---possibilities-and-limitations-of-comparative-quantitative-research-on-international-migration-flows.

Künn, S. (2015). The challenges of linking survey and administrative data. Retrieved from *IZA World of Labor* website: https://wol.iza.org/uploads/articles/214/pdfs/challenges-of-linking-survey-and-administrative-data.pdf?v=1.

Martí, M. & Ródenas, C. (2004). Migrantes y migraciones: de nuevo la divergencia en las fuentes estadísticas [Migrants and migrations: again the divergence in the statistical sources]. *Estadística Española*, 46(156), 293-321.

Martí, M. & Ródenas, C. (2007). Migration estimation based on the Labour Force Survey: An EU-15 perspective. *International Migration Review*, 41(1), 101-126.

Martí, M. & Ródenas, C. (2012). Measuring international migration through sample surveys: some lessons from the Spanish case. *Population-E*, 67(3), 435-464.

Oberski, D.L., Kirchner, A., Eckman, S. & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, 112(520), 1477-1489.

Rastogi, S. & O'Hara, A. (2012). *2010 Census Match Study. Final Report*. US Census Bureau. Retrieved from US Census Bureau website: https://www.census.gov › 2012 › dec › 2010_cpex_247.

Rivera-Drew, J.A., Flood, S. & Warren, J.R. (2014). Making full use of the longitudinal design of the Current Population Survey: Methods for linking records across 16 months, *Journal of Economic and Social Measurement*, 39, 121–144.

Ródenas, C. & Martí, M. (2009). Estimating false migrations in Spain. Population-E, 64(2), 361-376.

Sakshaug, J.W. (2018a). Methods of linking survey data to official records. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research*, (pp. 257-261). Doi: 10.1007/978-3-319-54395-6_34.

Sakshaug, J.W. (2018b). Linking survey data to official government records. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research*, (pp. 597-606). Doi: 10.1007/978-3-319-54395-6_34.

Sakshaug, J. W. & Antoni, M. (2017). Errors in linking survey and administrative data. In Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., … West, B. T. (Eds.), *Total survey error in practice* (pp. 557-573). Hoboken, NJ: John Wiley & sons.

Sakshaug, J.W., Antoni, M. & Sauckel, R. (2017). The quality and selectivity of linking federal administrative records to respondents and nonrespondents in a general population sample survey of Germany. *Survey Research Methods*, 11(1), 63-80.

Sakshaug, J.W, Couper, M.P., Ofstedal, M.B. & Weir, D.R. (2012). Linking survey and administrative records: mechanisms of consent. *Sociological Methods & Research*, 41(4), 535–569.

UNECE (2007). *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics*. Retrieve from UNECE website: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf.

UNECE (2011). *Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices*. Retrieved from UNECE website: https://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf.

UNECE (2019). *2020 Population Census round.* Retrieved from UNECE website: https://statswiki.unece.org/display/censuses/2020+Population+Census+Round.

Zhu, Y., Matsuyama, Y., Ohashi, Y. & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56, 80–86.

**Table 1**: Population by place of residence in the previous year. 2010-2016

| Place of residence one year before (LFS): | | Years registered in the municipality (PR): | | |
|---|---|---|---|---|
| | | **Same** (more than one year) | **Other** (less than one year) | **Total** |
| **2010** | **Same** | 37,322,223 | 922,606 | *38,244,829* |
| | **Other** | 254,422 | 225,721 | *480,143* |
| | ***Total*** | *37,576,645* | *1,148,327* | *38,724,972* |
| **2011** | **Same** | 37,544,593 | 826,341 | *38,370,934* |
| | **Other** | 208,889 | 228,920 | *437,809* |
| | ***Total*** | *37,753,482* | *1,055,261* | *38,808,743* |
| **2012** | **Same** | 37,680,864 | 774,231 | *38,455,095* |
| | **Other** | 211,037 | 188,037 | *399,074* |
| | ***Total*** | *37,891,901* | *962,268* | *38,854,169* |
| **2013** | **Same** | 37,684,542 | 689,995 | *38,374,537* |
| | **Other** | 184,114 | 174,584 | *358,698* |
| | ***Total*** | *37,868,656* | *864,579* | *38,733,235* |
| **2014** | **Same** | 37,446,287 | 668,185 | *38,114,472* |
| | **Other** | 183,277 | 185,852 | *369,129* |
| | ***Total*** | *37,629,564* | *854,037* | *38,483,601* |
| **2015** | **Same** | *37,415,745* | *676,734* | *38,092,479* |
| | **Other** | *217,786* | *206,923* | *424,709* |
| | ***Total*** | *37,633,530* | *883,657* | *38,517,187* |
| **2016** | **Same** | 37,433,221 | 646,567 | 38,079,788 |
| | **Other** | 212,369 | 199,700 | 412,069 |
| | ***Total*** | *37,645,590* | *846,267* | *38,491,857* |

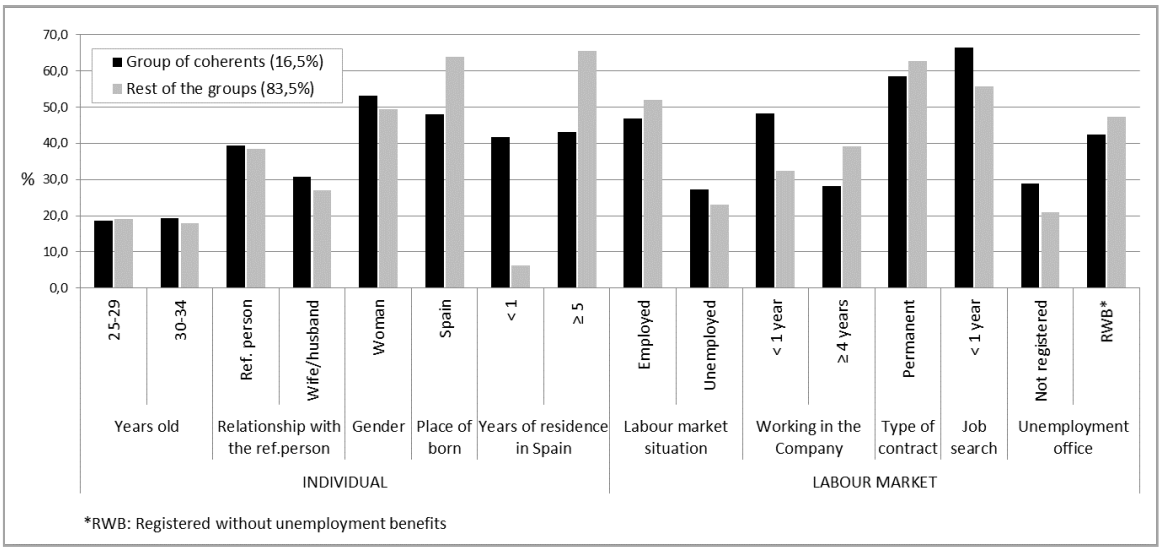
Source: LGMS and own elaboration.

**Table 2:** Population by residence in the previous year. Annual average 2010-2016.

| | | | Years registered in the municipality (PR): | | | | | | TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| | | | Less than 1 year | | | More than 1 year | | | |
| | | | Previous place of residence | | | Previous place of residence | | | |
| | | | No variation since birth | Previous residence in Spain | Previous residence abroad | No variation since birth | Previous residence in Spain | Previous residence abroad | |
| Place of residence one year before (LFS): | The same | | Group 1: 0 | Group 2: 630,328 (54.6%) | Group 3: 113,195 (9.8%) | Group 4: 14,133,325 | Group 5: 20,121,434 | Group 6: 3,249,166 | **38,247,448** |
| | Other place | Spain | Group 7: 0 | Group 8: 143,113 (12.4%) | Group 9: 3,330 (0.3%) | Group 10: 47,714 (4.1%) | Group 11: 116,139 (10.1%) | Group 12: 17,063 (1.5%) | **327,359** |
| | | Abroad | Group 13: 0 | Group 14: 7,361 (0.6%) | Group 15: 47,587 (4.1%) | Group 16: 8,060 (0.7%) | Group 17: 9,784 (0.8%) | Group 18: 11,510 (1.0%) | **84,303** |
| TOTAL | | | **0** | **780,802** | **164,112** | **14,189,099** | **20,247,357** | **3,277,739** | **38,659,109** |

Note 1: ☐ coherent, ☐ not coherent but explainable and ☐ incoherent.
Note 2: percentages calculated over people with mobility.
Source: LGMS and own elaboration.

**Figure 1**: Distribution of main categories of selected characteristics between the coherent migration group and the rest. Average 2010-2016




Source: LGMS and own elaboration.