When is administrative data enough to replace statistical information? A based on census comparison quality indicator

Anabela Delgado*, Sandra Lagarto*, Paula Paulino*, João Capelo*

**Statistics Portugal* (*INE*), *Census Unit 2021*

Abstract: Statistics Portugal is considering the use of administrative data in the 2021 Census. To face this challenge, the quality of the available administrative data sets is measured comparing administrative data with census information. The goal is to evaluate the risks of replacing part of the census collected information with information obtained from administrative sources. Record linkage methods were applied to the 2011 Census results and administrative datasets. Fifteen variables from seven administrative datasets (namely Social Security or Students register) were selected based on the administrative source contribution as a potential replacement of census collected information. For each matched record pair, information from corresponding administrative variables is compared, producing an equality rate estimate. The results show very high equality rates when comparing information from each matched pair of records, to both geographical and demographic variables (municipality of residence, sex, date of birth, legal marital status, country of birth, country of citizenship). When comparing socioeconomic variables, results are less homogeneous: identical information has less uniform distribution between Census and administrative sources (nevertheless, some data obtained by certain sources, related with labour force characteristics, also got high correspondence rates for compared record pairs). Additionally, considering that some statistics might be obtained by other sources, some Census microdata (regarding economic and educational characteristics of the population) were compared with data from national Labour Force Survey. These results converge to the general comparison results of this exercise. Finally, the results of the Post Enumeration Survey of 2011 Census were used to verify the reliability of the comparison results.

Keywords: 2021 Portugal population and housing Census, administrative data, Census microdata, linked data

1. Background

The Portuguese strategy for the 2021 Census considers the use of administrative data to provide information on some specific census topics, following the EU and UNECE countries' general trend regarding a more efficient census method, with high quality standards, but less burdensome for the respondents and less costly for the State. Statistics Portugal (INE) is currently conducting a feasibility study for the 2021 Population and Housing Censuses' new model which evaluates the usability of available administrative data for statistical purposes.

One of the steps of that study is to compare the characteristics of a set of register-based population with the respective characteristics from national results of the 2011 Population Census. This exercise will show how administrative data collected by several sources approximates to census collected data and point out discrepancies.

To support the results to some economical and educational characteristics of the population, we also compare microdata from 2011 Census to Portuguese Labour Force Survey for the 1st quarter of 2011 (LFS). Also, we use Census 2011 Consistency Index (ICG) from Post Enumeration Survey (PES) to validate results.

2. Selection of administrative sources and variables

Considering the feasibility study for the 2021 Census, the legal frame which allowed Statistics Portugal the access to administrative data was established by the Law no. 22/2008 on the National Statistical System of 13 May and the Deliberation of the National Commission for Data Protection no. 929/2014 of 11 June (numeric identifiers were encrypted and no full access to both names and addresses were allowed).

For the current exercise, 9 data sources were selected considering the potential use of administrative data for census information (see Table 1). In the selected administrative data sources, 15 target variables formerly provided by 2011 national Census were identified: 7 concerning geographical and demographic characteristics and 8 concerning economical and educational characteristics (see Table 2).

Table 1. *Administrative datasets sources to compare with the 2011 Census microdata*

Administrative sources	Year	No. records	Description	Name
Institute of Registration and Notary (IRN)	2010	11 565 714	Civil register	BDIC
Immigration and Borders Service (SEF)	2011	434 708	Foreigner register	SEF
Social Security Institution (ISS)	2011	7 209 027	Social Security register	ISS
Strategy and Planning Office (GEP)	2011	2 736 659	Employment register (Bulletin of Labour and Employment)	QP
Institute of Employment and Training (IEFP) and Regional Directorate of Statistics of Madeira (DREM)	2011	702 215	Unemployment register	IEFP
General Directorate of Education and Science statistics (DGEEC) and Regional Secretariat for Education and Human Resources of the Autonomous Region of Madeira (DRE)	2011	1 965 842	Students register	DGEEC
General Retirement Fund (CGA)	2010	1 103 980	Public administration retirement fund register	CGA

Table 2. *Selected administrative topics to compare with 2011 Census variables*

Administrative dataset	Available information on population topics
BDIC	Place of residence (municipality), sex, date of birth, legal marital status, country of birth, country of citizenship
SEF	Country of birth, country of citizenship, current activity status, occupation
ISS	Current activity status, place of work, status in employment
QP	Place of work, occupation, industry (establishment), status in employment, number of persons working in the enterprise, hours usually worked, educational attainment
IEFP	Current activity status
CGA	Current activity status
DGEEC	School attendance

3. Methodological aspects

The aim of this exercise is to compare, for each person, the exact value of the target variable on administrative datasets, which is the closest as possible with the statistical concept and definition, with 2011 Census microdata.

The population in comparison results from a previous match-key process between the 2011 Census microdata and the administrative records, selected from the several sources in a stepwise manner (using combinations of available information – sex/name/date of birth/ marital status/country of citizenship/municipality of usual residence – to link census microdata to each administrative datasets, sequentially). Data preparation (including recoding) and standardization were previously performed. There were no missing characteristics added to the registers and data was considered up to date.

It was possible to match 9 949 599 census records to administrative records from selected sources, which means 94 per cent of the resident population stock back in 2011, with a false positive rate of 6 per cent (that value represents the total number of matched census records with at least one administrative dataset).

Considering the matched records, the main purpose of this exercise is to evaluate, for a selection of variables, if we get the same information from administrative datasets on individuals as the one collected in 2011 Census. Only after the analysis of these results we could consider the use of administrative data to replace census collected statistics information.

The equality rate was estimated based on the comparison of exact information on each pairs of records that were possible to match. For those records, which represent the same person, our hypothesis is that, if equality is verified, we can rely on administrative information for statistical purposes. To support this decision, we have two additional criteria: results from 2011 Census Post Enumeration Survey ICG and also results from comparison between 2011 Census and 2011 first quarter Labour Force Survey microdata.

4. Results and discussion

Table 3 summarizes the obtained results from the comparison exercise, for the set of selected census variables with available administrative information to compare with. We show the population numbers, the number of available administrative records and the actual number of administrative records compared to census microdata (resulted from matching process). We also present the values of the Global Consistency Index (ICG) from the Post Enumeration Survey (PES) of the 2011 Census (INE, 2013).

Before presenting the results, two notes: one to categorical variables and another one to variables with different detail levels of information. In this paper, we only show results for all categories and aggregated information, but the study carried out was exhaustive and detailed in comparisons, producing a vast set of results.

The first note is to enhance that all categorical variables were also compared by groups. If we take, for instance, current activity status, the equality rate point in table 3 is about 81 per cent when we compare census microdata to individual ISS registers for all categories. In this case, within groups, comparisons may have some variations. Considering again current activity status, 92 per cent of those who respond in census questionnaire that were *employed* are registered in Portuguese Social Security system as *employed*.

The second note is to consider variables with different levels of information. If we take occupation, for instance, table 3 points to about 63 per cent of equality rate when census microdata is compared to individual QP registers. That value indicates a higher aggregation level of information, that is to say, one-digit level. The general trend, for this type of variables, is that the higher the disaggregation, the lower the equality rate estimated.

Let´s now analyse the global exercise comparison results on table 3. The comparison results on demographic variables show high equality rates – 90 to 99 per cent – on date of birth, sex, country of birth, country of citizenship and legal marital status. Also, the place of usual residence obtained an equality rate quite high: about 95 per cent of all registered pairs compared had the exact same information.

Table 3. *2011 Census microdata and administrative records comparison results*

Variable	2011 Census population to be compared	Number of administrative records to be compared to 2011 Census by source		Number of pairs compared	Equality rate on compared pairs (%)	ICG[1] (%)
Place of residence (municipality)	10 562 178	BDIC	11 565 714	9 308 384	94,6	**97,7**
Sex	10 562 178	BDIC	11 565 714	9 308 384	99,9	**99,0**
Date of birth	10 562 178	BDIC	11 565 714	9 308 384	92,6	**95,7**
Legal marital status	10 562 178	BDIC	11 565 714	9 308 384	95,3	**97,4**
Country of birth	10 562 178	BDIC	11 565 714	9 308 384	94,7	**84,0**
		SEF	434 708	107 136	91,3	**84,0**
Country of citizenship	10 562 178	BDIC	11 565 714	9 308 384	99,4	**97,8**
		SEF	434 708	107 136	90,3	**97,8**
Current activity status	8 989 849	ISS	7 066 838	4 910 073	81,2	
		SEF	379 965	107 136	27,1	
		CGA	1 103 980	716 264	92,1	
		IEFP	702 215	454 479	42,1	
Place of work (municipality)	4 361 187	ISS	4 107 425	2 788 758	56,6	**77,6**
		QP	2 736 659	2 045 476	81,6	**77,6**
Occupation	4 361 187	QP	2 736 659	2 045 476	61,9	
		SEF	124 721	171 370	52,9	
Industry	4 361 187	QP	2 736 659	2 045 476	74,1	
Status in employment	4 361 187	QP	2 736 659	2 045 476	93,0	**82,2**
		ISS	4 107 425	2 788 758	85,5	**82,2**
Number of persons working in the enterprise	4 361 187	QP	2 736 659	2 045 476	54,4	**51,6**
Hours usually worked	4 361 187	QP	2 736 659	2 045 476	56,8	
Educational attainment	10 445 093	QP	2 736 659	2 210 930	59,5	
School attendance	10 445 093	DGEEC	1 965 842	1 359 916	82,2	**69,8**

[1] ICG measures content errors; it represents the percentage of statistical units (resident population), with the same classification both in the 2011 Census and Census PES, of all common units to the two statistical operations.

As for the socioeconomic variables, the results are less homogeneous. We identify three situations:

- High equality rates for certain variables on all sources with available information; e.g.: status in employment with about 86 per cent of census correspondence via ISS and 93 per cent via QP;
- Equality rates with large variation by source: variables like profession, industry and current activity status (in this last one, about 92 per cent of correspondence via CGA, while, considering the IEFP, this value decreases to 42 per cent);
- Equality rates estimated from comparison with a single source: from 50 per cent correspondence in the number of persons working in the enterprise or hours usually worked to more than 80 per cent on school attendance.

To support the census – administrative datasets comparison results, we decided to get results from the 2011 Census PES quality indicator, ICG. Surprisingly, the estimated equality rates and ICG values are very close to most selected variables (even though, to some variables, concepts are close, but don't exactly match). This fact supports the results obtained from the comparison and increases the credibility of using administrative information.

Finally, to have an additional indicator to validate results, we also did the comparison 2011 census – 2011 first quarter LFS[2] microdata. The LFS sample size was 39 884 individuals. For this exercise, it was necessary to apply a match-key (sex/name/date of birth/ marital status/ municipality of usual residence) with census records. We obtained 17 732 pairs of records to compare with 2011 Census microdata (6 995 aged 15 years and over).

[2] The Portuguese LFS, which is conducted nationwide, is a sample survey providing quarterly results (recently monthly). Back in 2011, it collected labour market information for approximately 40,000 individuals.

Table 4 shows corresponding comparison results, census microdata vs. administrative information and census microdata vs. LFS microdata, on 8 labour force and educational variables.

Table 4. *2011 Census microdata and LFS comparison results*

Variables	**Equality Census-LFS (%)**	**% Equality Census – administrative records by selected administrative data source**	
Labour force status	**84,3**	81,2	ISS
Occupation	**67,8**	61,9	QP
Industry	**77,6**	74,1	QP
Status in employment	**86,5**	93,0	QP
Number of persons working in the enterprise	**60,6**	54,4	QP
Hours usually worked	**72,6**	56,8	QP
Educational attainment	**80,2**	59,5	QP
School attendance	**86,5**	87,4	DGEEC

For this purpose, we use the highest equality rate comparison results (from table 3), census microdata vs. administrative information, whenever several administrative sources were available for a target variable.

Except for educational qualifications, equality rates values from both comparisons, for selected variables, are similar. We consider that these results increase the overall consistency of the comparison exercise between the 2011 Census microdata and administrative records.

Last, a note on coverage issues. From table 3, it is obvious that some variables are not fully covered by the Portuguese administrative data available for the Feasibility Study of the 2021 Census. In fact, we know, from initial diagnostic information needs, that some core topics for population and housing censuses (e.g. related with household or education) are not fully or even partially covered by Portuguese administrative data.

That is not an issue for the current exercise and neither are inconsistencies between sources (a set of rules have been prepared for that matter).

5. Conclusions

The evaluation of administrative data quality for statistical purposes can be a huge task. One step in this evaluation process is – after dealing with concepts, classifications, timeliness, processing and data treatment, data linkage and matching and other issues – verify if (despite cover issues) the information that we get from administrative data sources is what we need for census statistics, that is to say, if it is valid and precise.

It is common sense that the compromise between what we have and what we need is difficult to achieve, particularly when the process involves resources that we do not detain or control, like administrative datasets. In this particular task, many countries that face transition on census model from traditional to register-based models, have the same problems of Portugal. For Statistics Portugal, this simple comparison exercise is part of a complex project which is a work in progress and should continue beyond the 2021 Census.

We consider that the results can be a base for discussion on the purpose of administrative data usage to replace or to be used in addition to census data collection. At this time, we point out some conclusions/ reflections on the obtained results:

- Results show huge consistency between administrative data and 2011 census microdata;
- We compared administrative data individuals records to seven 2011 Census demographic variables (all used in the match-key exercise). Those equality rates are very high (90 per cent of the compared pairs of records' information is exactly the same);

- We also compared labour force related and educational characteristics – from eight selected 2011 census variables, we obtained more than 80 per cent of equality for some labour market variables;

- When comparing administrative data with 2011 Census microdata, QP source was the most consistent – with, globally, the highest equality rates – across the set of variables with available information;

- Comparability indicators show inequalities only based in unequal values (the differences are not caused by impossible data conversion or missing description); so, we consider that, thought there´s an obvious under coverage issue, administrative data can be use for add or replace information collect by census;

- Time lags between datasets and some conceptual issues could explain differences on comparison results; also, data sources holders are being contacted for new data flows and we believe that some of the issues that cause inequalities can be solved with more recent incomes;

- The reliability of using administrative data for statistical purposes was confirmed by using additional quality information criteria from PES and 2011 Census vs. 2011 LFS comparison results;

- For future work, cross comparison and hierarchical rules between sources of administrative information is being studied.

6. References

INE – Instituto Nacional de Estatística (2013), *Inquérito de Qualidade dos Censos 2011 – Metodologia e resultados*, Instituto Nacional de Estatística, Lisboa.