Census and Administrative Data in Portugal: Results and Challenges

Joana Araújo[1], Cecília Cardoso[1], Paula Paulino[1], Cláudia Pina[1], Pedro Raposo[2]

[1]Statistics Portugal, Portugal

[2]Catolica Lisbon School of Business and Economics, Portugal

Abstract

Democratic governance relies on accurate census data for policymaking and resource allocation. Portugal, lacking a resident population register and a unique personal identification number, is creating a Resident Population Database (BPR) as part of Statistics Portugal National Data Infrastructure. This study leverages the unique opportunity provided by the traditionally collected 2021 Census to evaluate the methodology of the BPR exercise based on administrative data.

Our findings show promise for the BPR's success in transitioning Portugal from traditional to administrative census data. Significantly high convergence between the BPR and 2021 Census data is observed in demographics like population distribution by sex and age, and by region. Microdata linkage revealed a 92% individual match and similarity rates exceeding 97% for demographic core variables.

The paper also unveils some challenges, particularly in terms of population coverage. In this regard, we conclude that the current methodology used by BPR slight under/overestimates some population age groups. Access and analysis of new administrative data sources (e.g. Monthly Statement of Earnings and Electronic Invoices database, both from Tax Authority) will be important to improve BPR methodology.

Keywords: Resident Population Database, Census, opportunities, challenges, administrative data

1. Introduction

In the context of democratic governance, census data serves as an indispensable foundation, shaping policies, planning, funding of the municipalities and resource allocation. There are different ways to carry out a population census, that we can summarise in “traditional” (direct count of all individuals and their characteristics through questionnaires, either in paper form or electronically), “registered based” (traditional enumeration is replaced using administrative

data held in various registers) and “combined” (use of administrative data with a limited collection of data from field enumeration) (see UNECE, 2015). The UNECE 2020 census round indicates that the trend away from traditional census is rapidly in progress. Considering the benefits of a register-based census – e.g. lower costs, timelier and more frequently updated information, Statistics Portugal is working on the transformation of the traditional census model into a more efficient one using administrative data (see UNECE, 2018 and 2021).

The Administrative Census project falls within the scope of the National Data Infrastructure (IND), which embodies Statistics Portugal's strategy of integration and creation of value for society from different data sources. The creation of the Resident Population Database (BPR) is the central element of the administrative census. This is a crucial and challenging project for Portugal in a particular context related to the nonexistence of an administrative population register and the lack of a unique Personal Identification Number (PIN) to link the various administrative sources.

The 2021 Census results proves to be a unique opportunity to assess the results of the work carried out within the scope of the BPR project and constitutes an important benchmark to support the transition process between the two census models.

In this paper we compared the two datasets in order to understand the convergence of results, and also to identify potential limitations or aspects to change in the current methodology used to construct BPR. The analysis was carried out using data at aggregated level and linking individual records.

2. From the traditional to the administrative Census

Statistics Portugal has been studying and testing the contribution of available administrative databases to replace information collected each ten years through full enumeration (traditional census model) and to produce census-type statistics on an annual-basis, namely annual statistics on resident population.

A statistical resident population database - Resident Population Database (BPR) - is the central element of the project and results from the linkage of several administrative data sources at an individual record level.

The two structural pillar databases used for the construction of the BPR are:

- Civil Identification Database (BDIC): database of all Portuguese nationals, regardless of country of residence or country of birth, that contains the basic demographic characteristics (e.g., name(s) and surname(s), sex, date of birth, place of birth, civil status, postal address declared by citizens as their usual residence).
- A database of non-nationals with a valid resident permit or an EU resident card; this database, besides the basic demographic characteristics (name(s) and surname(s), sex, date of birth, place of birth, civil status, postal address declared by the foreign citizens as their residence), also contains socio-economic characteristics.

However, these two important and exhaustive databases cannot be considered as "administrative population registers of residents", meaning that neither of them corresponds, for different reasons, to a database of residents. In the case of BDIC, evaluations conclude to an overestimation of approximately 10 per cent when compared with the population enumerated in the last two Census.

Due to this constraint, to estimate the resident population in Portugal, according to the concept of usual residence, we applied a methodology called "signs of life". These are given by the presence of the person in other administrative sources, such as Income Statement, Social Security or Education, besides the presence in BDIC or in the non-nationals database.

Another limitation is the fact that, due to Constitutional Law, Portugal does not have a unique Personal Identification Number (PIN). However, three central public administration systems - civil identification, tax, and social security – have each one a specific identification number: Civil Identification Number (NIC) – only Portuguese nationals; Social Security Number (NISS) and Tax Register Number (NIF).

As most of the administrative sources have, at least, one of the three ID numbers listed above, exact determinist match of these encrypted ID numbers (encrypted by the data holders by applying a hash with a SHA256 algorithm) is used. For data sources where common ID are not available, determinist matching of attributes or combination of attributes, like name(s) and surname(s), date of birth, sex, place of birth and place of residence, are used to link records at an individual level.

3. Comparison between Census and Resident Population Database (BPR)

The 2021 Census was a unique opportunity to assess the results of the work carried out within the scope of the BPR project. It was an important benchmark to support the transition process between the two census models, especially to evaluate the BPR methodology.

In this sense, data from BPR 2020 was compared with the results of 2021 Census, at aggregated level and at a microdata level, linking individual records.

3.1 Aggregated results

The total population figures obtained by 2021 Census (10,369,900[1]) and BPR (10,384,936) are very similar, representing at national level a difference of just 0.1%.

However, if we observe the distribution of the population in more geographical detail, differences tend to increase. At the NUTS2 level, the differences ranged between -0.7% and 4.5%, with a greater difference in regions with less population, namely autonomous region of Azores and Madeira (see Table 1).

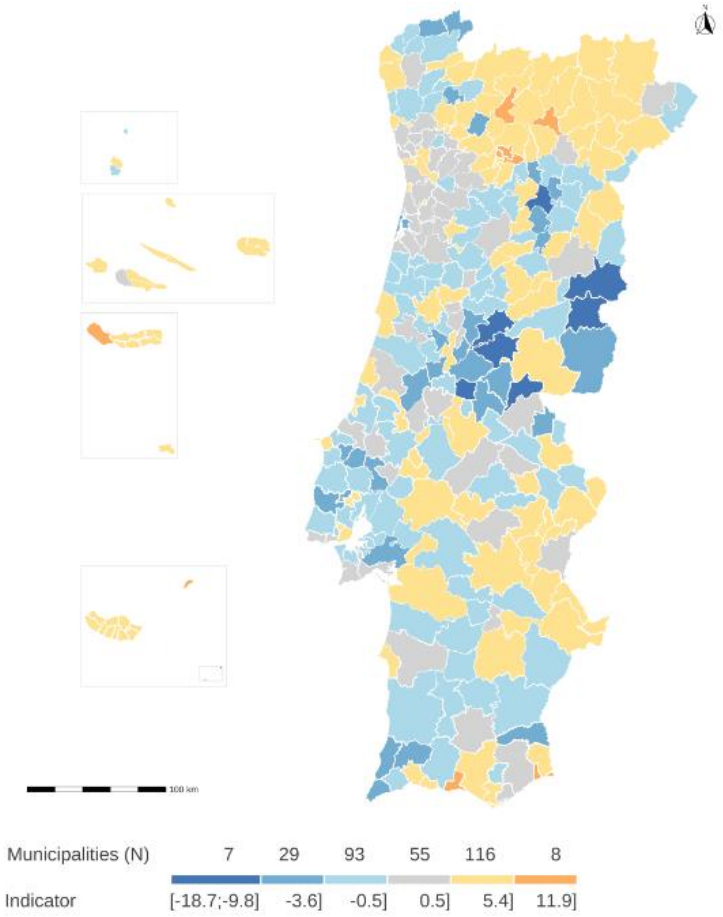
Table 1: Population difference (%) between Census and BPR, by Place of residence (NUTS2)

NUTS2	BPR	Census[1]	Difference (%)
Portugal	**10,384,936**	**10,369,900**	**0.1**
Norte	3,604,970	3,593,339	0.3
Centro	2,222,072	2,235,372	-0.7
AM Lisboa	2,866,336	2,877,213	-0.4
Alentejo	709,280	707,791	0.2
Algarve	477,507	468,335	2.0
R.A. Açores	247,210	236,620	4.5
R.A. Madeira	259,561	251,230	3.3

[1] The Census data was adjusted to reflect the population on December 31st, 2020.

At the municipality level (see Figure 1), most of the municipalities (85.7%) show a difference between -3.6% and 5.4%.

Figure 1: Population difference (%) between Census and BPR, by place of residence (Municipality)

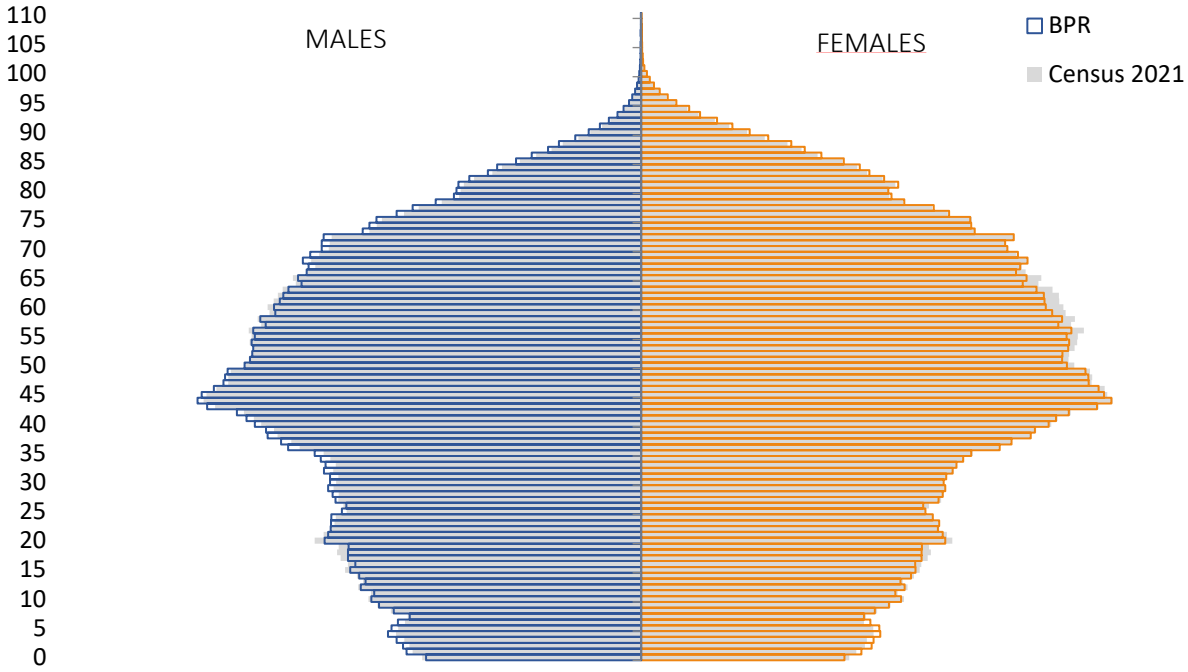


Census and BPR age pyramids show a very similar age-sex population distribution at a national level (see Figure2).

However, in relation to the Census results, it is generally possible to detect slight deviations in the BPR in some age groups: negative in both sexes at age 0 (-2.0%) and positive at ages between 1 and 6 years (3.0%); also negative between the ages of 7 and 21 for both sexes (-1.5%); positive deviation in the male population aged between 23 and 49 (1.9%); more pronounced negative deviations in the female population between the ages of 40 and 67 (-1.8%) and an overestimation in the elderly population, particularly men (3.9%).

While there might be slight differences, the core structure in population demographics appear to be well-represented in BPR data.

Figure 2: Age pyramid, Census and BPR – Portugal



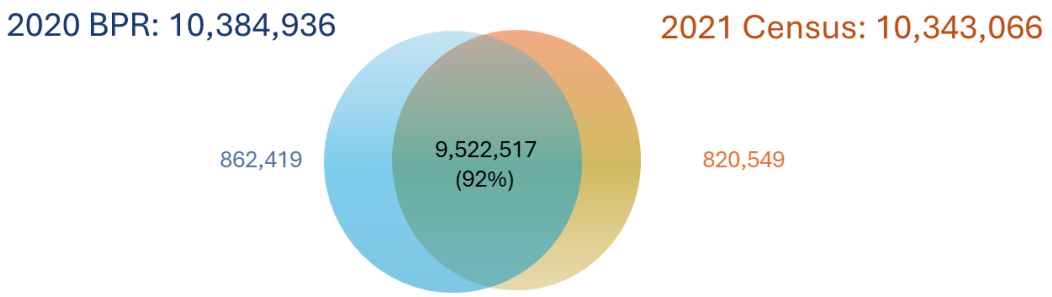
3.2 Micro data level results

3.2.1 Matched records

Measuring the degree of similarity between the characteristics of individuals included in BPR and the information reported in Census is also a relevant aspect to assess BPR results. To meet this goal an individual record data link process was applied, and similitude rates were calculated.

The BPR and Census individual records were linked by deterministic matching and similarity distance techniques (Jaro Winkler and Levenshtein), using name and birth date attributes. Through the linking process it was possible to identify 9,522,517 individuals that were present in both Census and BPR datasets, corresponding to 92% of total records (see Figure 3).

Figure 3: Matching individuals between Census and BPR



To evaluate the convergence between Census and BPR at individual record level a set of key demographic variables was considered.

The analysis revealed a high level of consistency between the two datasets for Sex, Citizenship, Age[2] and Place of birth, which achieved 100.0%, 99.5%, 99.4%, and 99.0% similarity rates, respectively. This remarkably high concordance suggests a near-identical representation of individuals across both the Census and BPR data for these demographic variables. Marital status data exhibited a very good fit, with a rate of 97.5%. While not quite as good as the previous variables, this high level of agreement still indicates a strong correspondence between characteristics of individuals on both sources (see Annex for detail results).

The variable Place of residence also demonstrated a very good fit, particularly at the municipality level (95.4%). However, residence by Local Administrative Unit (LAU)/parish showed a slightly lower similarity rate of 91.8%, suggesting some discrepancies between place of residence declared in Census (based in a usual residence concept) and the administrative residence registered in BPR.

Through the marginal effects, obtain from a Probit model (see Table 2), its possible to see that there is statistical evidence that allows to explain the differences between the parishes of residence reported in the Census and in the BPR. In this way, there is evidence that older individuals, over 85 years old, and divorced persons have a higher probability of not having the same parish by about 7.4 and 6.2 percentage points, respectively.

Foreigners have a greater probability of having a different parish in the two datasets, with Asians nationals presenting the highest value (15.5 percentage points).

[2] Considering five years age groups the similarity rate reaches 99.8%.

Table 2: Probit: Differences in the place of residence (LAU/parish) between Census and BPR

	Marginal Effect
Sex (vs Male)	
Female	0.006
Age (vs]40-85])	
[0-5]	0.041
]5-20]	0.021
]20-40]	0.046
>85	0.074
Marital status (vs Married)	
Divorced	0.062
Single	0.046
Widowed	0.037
Citizenship (vs Portuguese)	
America	0.098
Oceania	0.098
Other countries in Europe	0.039
European Union	0.045
Africa	0.105
Asia	0.155

Notes: All effects are statistically significant at a significance level of 1%. The dependent variable assumes a value of 1 if the parish in the Census is different from the parish in the BPR. Table 2 presents the marginal effects of the probit regression.

3.2.2 Unmatched records

The analysis of the records found in Census and not in the BPR (820,549) (see Figure 3) was also decisive to understand the effectiveness of the BPR's methodological approach, particularly regarding the “signs of life” rules.

There are many possible reasons for these unmatched records. On one hand, there might be incompleteness of the attributes used in the matching procedure, for instance the name. On the other hand, the reference date is not the same for both sources. Additionally, some BPR rules can explain the exclusion of certain individuals that were found in Census.

In fact, when Census individuals who were not present in the BPR were linked directly to administrative sources it was possible to identify some of them in those sources. Non-nationals and non-economically active population were more represented in the unmatched group.

Although these population groups (unmatched) need further investigation, it seems evident that small adjustments should be made in BPR to correct some minor coverage biases.

4. Conclusions

This study revealed a high degree of convergence between the BPR and the 2021 Census data in several dimensions. This is particularly evident in demographic profiles such as distribution by age and sex, and by region. Furthermore, when microdata was linked (92% matched individuals), similarity rates above 97% for demographic variables (Sex, Citizenship, Age, Place of birth, and Marital status) were found.

When Place of residence was compared specific groups, such as the elderly, divorced individuals, and non-nationals, exhibited higher probabilities of having a different parish of residence.

The paper also unveils some challenges, particularly in terms of population coverage. In this regard, it points out that the current methodology used by BPR slightly under/overestimates some population sub-groups. Access and analysis of new administrative data sources (e.g. Monthly Statement of Earnings and Electronic Invoices database, both from Tax Authority) will be important to improve BPR methodology.

The findings reveal that data gathered in BPR is deemed plausible and comprehensive, foreseeing a successful transition to an administrative census in Portugal.

References

Albertha, Winida et al. (2020), “Assessment of the Quality of Administrative Data for Use in a Population Census; Applying Census Test Results in Indonesia”. 1 Jan. 2020: 925 – 932.

Dygaszewicz, Janusz (2020), “Transition from traditional census to combined and registers-based census”, Statistical Journal of the IAOS, 36: 165–175.

Ernsten, A., McCollum, D., Feng, Z., Everington, D., & Huang, Z. (2018). “Using linked administrative and census data for migration research”. Population Studies, 72(3), 357–367. https://doi.org/10.1080/00324728.2018.1502463

Harper, G. & Mayhew, L. (2012). Using Administrative Data to Count Local Populations. Applied Spatial Analysis and Policy, 5(2), pp. 97-122. doi: 10.1007/s12061-011-9063-y

Statistics Portugal (2019), Press Release "Census with Administrative Data" https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=dia_europ_est

UNECE (2021), Guidelines for assessing the quality of administrative sources for use in censuses https://unece.org/statistics/publications/CensusAdminQuality

UNECE (2018), Guidelines on the use of registers and administrative data for population and housing censuses

https://unece.org/guidelines-use-registers-and-administrative-data-population-andhousing-censuses

UNECE (2015), Recommendations for the 2020 Censuses of Population and Housing

https://unece.org/fileadmin/DAM/stats/publications/2015/ECECES41_EN.pdf

Annexes

Table: Comparation between Census and BPR at a microdata level - Sex

BPR	Census	
	Male	**Female**
Male	4,510,675	1,297
Female	969	5,009,576
Total	**4,511,644**	**5,010,873**

Table: Comparation between Census and BPR at a microdata level - Citizenship

BPR	Census		
	Portuguese	**Foreigners**	**Stateless**
Portuguese	9,223,164	24,440	17
Foreigners	20,691	254,136	46
Unknown	7	13	3
Total	**9,243,862**	**278,589**	**66**

Table: Comparation between Census and BPR at a microdata level – Age Group

BPR	Census																				
	[0-4]	**[5-9]**	**[10-14]**	**[15-19]**	**[20-24]**	**[25-29]**	**[30-34]**	**[35-39]**	**[40-44]**	**[45-49]**	**[50-54]**	**[55-59]**	**[60-64]**	**[65-69]**	**[70-74]**	**[75-79]**	**[80-84]**	**[85-89]**	**[90-94]**	**[95-99]**	**[100 ou +]**
[0-4]	383,636	112	53	12	11	6	5	5	4												
[5-9]	195	408,234	205	69	13	2	8	3	3	4	3	1		1							
[10-14]	79	200	464,054	220	48	17	7	4	12	6	7										
[15-19]	24	30	143	491,197	246	39	7	9	7	5	3	3	2		1						
[20-24]	24	5	19	121	502,335	293	132	12	20	11	17	6	3		1	1			1	1	
[25-29]	13	7	8	42	337	476,162	283	130	23	19	11	11	6	3	1	2	1				
[30-34]	3	5	10	10	132	266	501,864	317	79	18	19	9	4		2	2	2		1		
[35-39]	4	2	5	12	19	124	398	593,805	315	85	16	10	7	9	4	3	3				
[40-44]	1	2	4	8	38	26	152	326	705,317	534	148	24	14	2	3	3	3				
[45-49]	3		3	12	20	25	23	104	567	733,412	446	187	23	21	16	6	2	4	1		
[50-54]	1	1	5	4	29	18	33	19	341	498	684,548	731	164	43	14	8	5	2	2		
[55-59]	1	1	3	9	15	31	14	32	34	362	768	675,836	740	154	25	9	4	4	1		
[60-64]	2		4	5	12	12	10	12	46	36	290	644	641,006	771	96	19	13	6	3		
[65-69]	2	2	3	6	8	23	12	16	18	56	64	274	843	607,87	509	101	10	19	6	2	
[70-74]	3		2	4	1	5	10	11	23	36	45	41	176	587	553,158	890	92	19	9	1	
[75-79]	4	2	1	5		7	5	3	17	42	22	35	50	156	893	431,273	503	85	21	7	2
[80-84]	2		2		7	3	6	11	12	23	22	19	42	31	187	459	334,814	696	95	8	1
[85-89]		2		1	2	13	7	21	5	22	11	22	19	29	37	128	674	206,603	252	43	7
[90-94]	1			2	4	2	4	3	4	7	14	5	18	8	16	14	66	155	83,087	91	10
[95-99]		1		1	2			1	1		6	3	6	5	6	5	11	21	57	18,879	6
[100 ou +]				1	2				1	1	1	1		2	2		5	3	1	6	1,985
Total	**383,998**	**408,606**	**464,524**	**491,741**	**503,281**	**477,074**	**502,98**	**594,844**	**706,849**	**735,177**	**686,461**	**677,862**	**643,123**	**609,692**	**554,971**	**432,923**	**336,208**	**207,617**	**83,537**	**19,038**	**2,011**

Table: Comparation between Census and BPR at a microdata level - Place of Birth

BPR	Census	
	Portuguese	**Foreigners**
Portuguese	8,645,679	4,511
Foreigners	93,391	778,416
Unknown	91	429
Total	**8,739,161**	**783,356**

Table: Comparation between Census and BPR at a microdata level - Marital Status

BPR	Census			
	Single	**Married**	**Widowed**	**Divorced**
Single	4,008,654	43,354	8,810	10,147
Married	11,789	3,933,213	78,128	41,566
Widowed	2,962	1,956	629,653	1,643
Divorced	22,614	10,202	8,556	699,609
Unknown	4,167	3,128	308	2,058
Total	**4,050,186**	**3,991,853**	**725,455**	**755,023**