An innovative approach to improve the quality of the household and nuclei types reconstruction in Italy

Rosa Maria Lipsi, ISTAT-Italian National Institute of Statistics, lipsi@istat.it
Anna Pezone, ISTAT-Italian National Institute of Statistics, pezone@istat.it

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

04-07 June 2024

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Outline

- Introduction
- Data and methods
- Focus on the Editing and Imputation
  - Household and family
    - Relationship item, marital status
    - Identification of potential couples
    - Household and family structure
  - “Families Procedure” (FP)
- Main results
- Concluding remarks and further developments

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly
financed by the European Union

Aim of the work

whole process to produce statistics on the household and their characteristic

BY USING → Data collection based on the integration of multiple sources

Survey
Internal Istat data
RBI-CENS2021
ANPR

IN ORDER TO HIGHLIGHT

- The main advantages of the innovative integration process
- The quality of the data

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Introduction (1/2)

Since 2018, the Italian National Institute of Statistics, as other European countries, moved from the traditional ten-year "door-to-door" census to a yearly "register-based" system (the Permanent Population and Housing Census - PPHC)

- To produce annual detailed statistics at macro-micro level
- To enrich the supply & quality of statistical information
- To reduce the statistical burden for respondents
- To reduce costs by the community

Great innovation!!!

Istat

Every 10 years

According to European regulations, EU Member States must send to Eurostat information on the main characteristics of their resident population and their social and economic conditions at national, regional and small areas levels, regardless of how they collected them.

A multisource approach, based on a combination of administrative data, registers (**RBI – Based Register of Individuals, RSBL – Statistical Base Register of Territorial Entities**) and surveys data, has been used to provides information on Italian **PPHC** for the 2021, as required by the EU regulation 2017/712.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Introduction (2/2)

The number of households and their characteristics is one of the **mandatory information for Eurostat**, but also one of the **most complex aggregates** to detect, validate and disseminate. The main problem to solve is the correct identification of household and nuclei types.

Household definition

A household is a group of persons connected by bonds of marriage, civil partnership, kinship, kinship in law, adoption, guardianship or affection, cohabitants and usually living in the same municipality (even if they are not yet registered in the municipal Population Register).

A household can also consist of just one person.

Persons who are temporarily absent, whether in another accommodation (or in an institutional household) in the same municipality or in another municipality or abroad, continue to be part of the household.

The conference is partly financed by the European Union

Data and methods (1/4)

Sample of Italian households

New census

Date	N° Households	N° Municipalities
2018, 7th October	1,400,000	2,800
2019, 6th October	1,400,000	2,800
2020 No Census	COVID-19 CORONAVIRUS	COVID-19 CORONAVIRUS
2021, 3rd October	2,400,000	4,500
2022, 2nd October	1,330,000	2,531
2023, 1st October	1,460,000	2,531

Anyway, ISTAT produced the population count using only the ***Signs of Life*** in the administrative sources

RBI 2020
31 December

AIDA
Archive of Usual Resident Population

Under-coverage

Individuals **not resident** in RBI 2020 **with** "*direct signs of life*" of at least one year in AIDA

Individuals resident in RBI 2020 **with** "*direct and indirect signs of life*" in the administrative archives

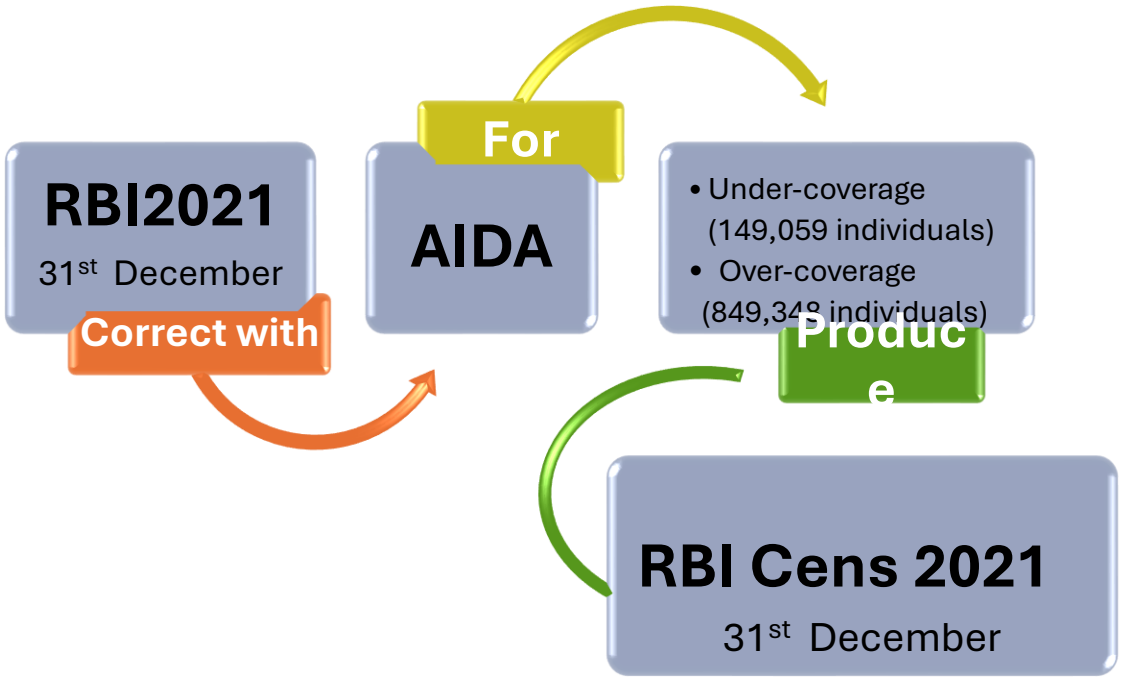
Individuals resident in RBI 2020 **without** "*direct and indirect signs of life*" in the administrative archives

Over-coverage

RBI CENS2020
31 December

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Data and methods (2/4)



Distribution of the Italian Population and Households at 31st December 2021 by Household size by number of members. Absolute and Percentage values

Household size by number of members	Number of Individuals		Number of Households	
	A.V.	%	A.V.	%
1	9,636,232	16.4	9,636,232	36.8
2	14,241,696	24.3	7,120,848	27.2
3	14,039,430	23.9	4,679,810	17.9
4	14,181,536	24.2	3,545,384	13.5
5	4,529,020	7.7	905,804	3.5
6 or more members	2,050,881	3.5	318,168	1.2
Total	**58,678,795**	**100**	**26,206,246**	**100**

Source: Our elaboration

Average Household Size: 2.24 members
St. Dev.: 1.59 members

The age, sex and citizenship of the Italian Legal Population have been fixed!

The conference is partly financed by the European Union

Data and methods (3/4)

For households reconstruction

Demographic variables

- ID Number (Individual code)
- ID HHold (Household code)
- Age
- Sex
- Citizenship
- Relationship with reference person
- Marital status
- Year of marriage or civil union
- Number of members
- Municipality of residence

+

Auxiliary variables

The Italian Base Register of Individuals

	VARIABLES								
UNITS	ID NUMBER	ID HHOLD	GENDER	DATE OF BIRTH	CITIZENSHIP	RELATIONSHIP	MARITAL STATUS	YEAR OF MARRIAGE OR CIVIL UNION	.....
	000001	000001	x11	x12	x13	x14	x15	x16	.....
	000002	000001	x21	x22	x23	x24	x25	x26	.....
	000003	000001	x31	x32	x33	x34	x35	x36	.....
	.....	000002	x..	x..	x..	x..	x..	x..	.....
	.....	000002	x..	x..	x.	?	?	x..	.....
	.....	000003	x..	x..	x..	x..	x..	x..	.....
	.....	.....	x..	x..	x..	x..	?	?	.....
	.....	.....	x..	x..	x..	?	x..	x..	.....
	.....	.....							.....
		TOTAL	X.1	X.2	X.3	X.4	X.5	X.6	.....

RBI 2021 + ANPR (National Register of Resident Population)

The conference is partly financed by the European Union

Data and methods (4/4)

- **Reclassification of the variable relationship with the reference person (RP)**

  Taking into account the compatibility rules between components

Example

"Cohabiting with adoption or emotional ties"

Two categories of the census:
- ✓ "Cohabiting in consensual union with RP"
- ✓ "Other cohabiting person without being a member of a couple or a relative"

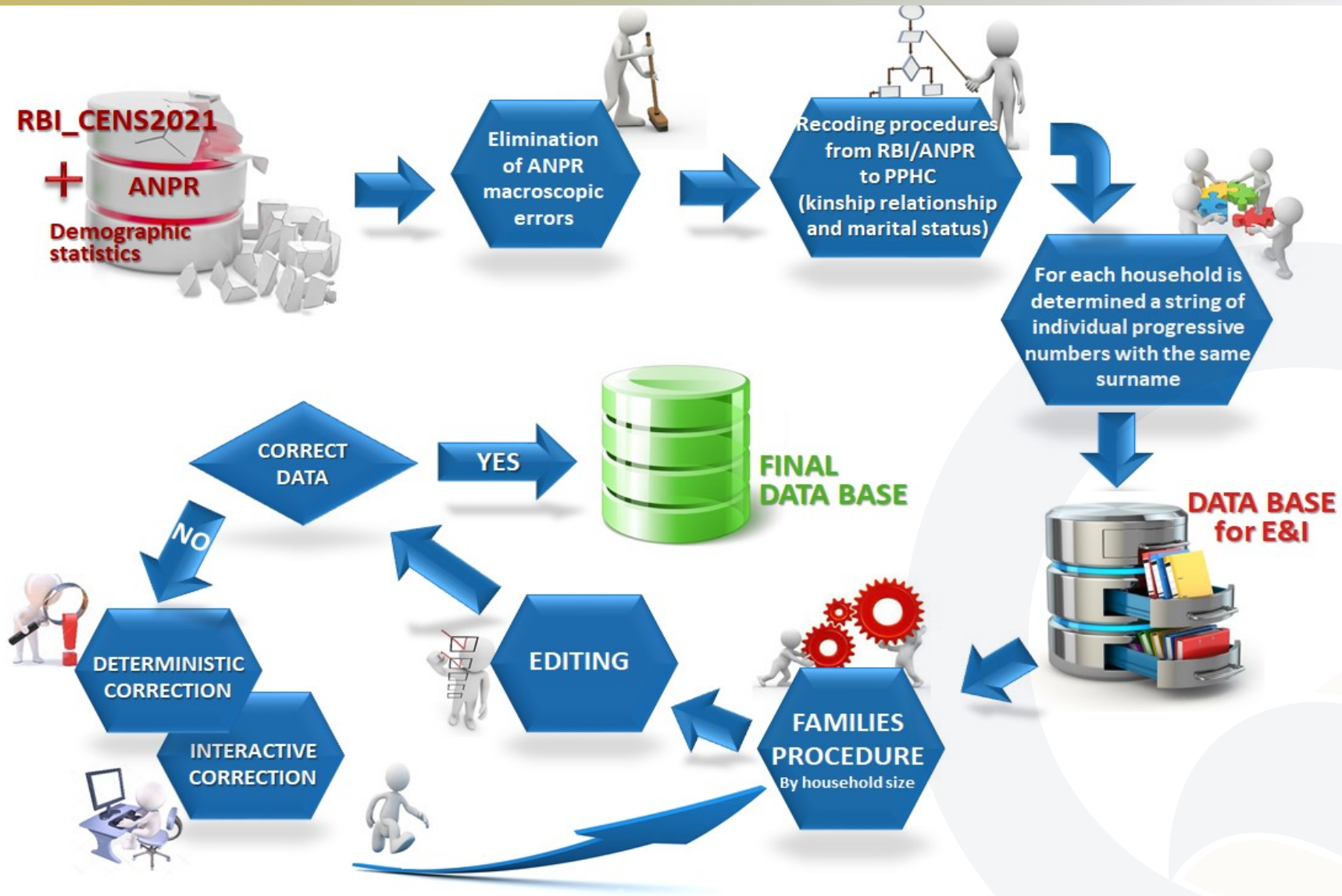
Generic categories of ANPR (30 categories)

Specific categories of the census (23 categories)

- **Calculation of auxiliary variables**
  - ✓ Identification of potential couples
  - ✓ The string of progressive numbers of individuals with the same surname within each household, respecting the anonymization process by guarantee the privacy of the individual data, according to the General Data Protection Regulation (GDPR – Regulation 2016/679)
- **Household reconstruction**

The conference is partly financed by the European Union

The E&I and household reconstruction process



The conference is partly financed by the European Union

Household reconstruction

Admin sources + Register use + Istat source + Survey

PF USE

Review of the overall Households and Nuclei reconstruction process

The conference is partly financed by the European Union

1st Level of Editing and Imputation (1/2)

Distribution of households and households with at least one individual error by household size. Absolute and percentage values.

	Households with at least one individual error			N° of Households
Household size by number of members	A.V.	%	Row %	A.V.
1	2,029,259	42.0	21.1	9,636,232
2	1,197,503	24.8	16.8	7,120,848
3	838,359	17.4	17.9	4,679,810
4	443,861	9.2	12.5	3,545,384
5	195,591	4.1	21.6	905,804
6 or more members	121,572	2.5	38.2	318,168
Total	**4,826,145**	**100**	**18.4**	**26,206,246**

Source: Our elaboration

The conference is partly financed by the European Union

1st Level of Editing and Imputation (2/2)

Distribution of missing data and the number of failures by type of edits. Absolute and percentage values.

6.5 % of Total Population

After imputation from Istat source

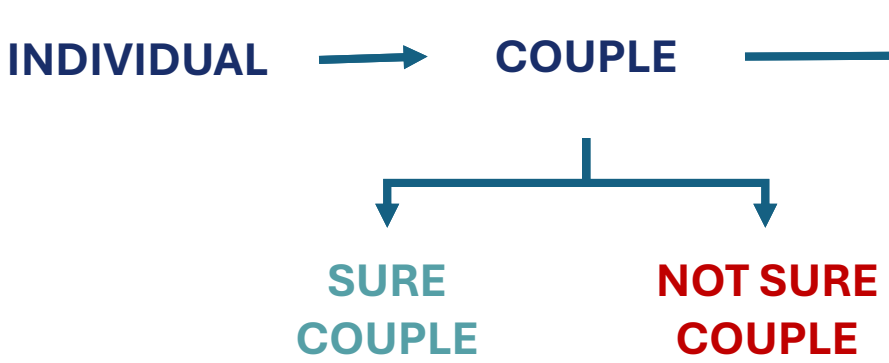
	Number of errors	
	A.V.	%
Missing data	**3,788,122**	**100**
Relationship with reference person	*174,585*	*4.6*
Marital status	*1,418,407*	*37.4*
Year of marriage or civil union	*2,195,130*	*57.9*
Individual Inconsistencies	**937,328**	**100**
Relationship with reference person	*42,174*	*4.5*
Marital status	*810,615*	*86.5*
Others	*84,539*	*9.0*
Familial Inconsistencies	**223,082**	**100**
Relationship with reference person and marital status	*114,301*	*51.2*
Age differences among children	36,942	16.6
Partners and marital status	*19,953*	*8.9*
Children and marital status	*17,999*	*8.1*
Age differences between partner's RP and children	*13,832*	*6.2*
Age differences between partner's RP and grandchildren	*8,091*	*3.6*
Others	*11,964*	*5.4*

1,4 mln foreign:
30% of total foreign
2.4% of total population

Source: Our elaboration

The conference is partly financed by the European Union

Identification of Potential couples (1/2)



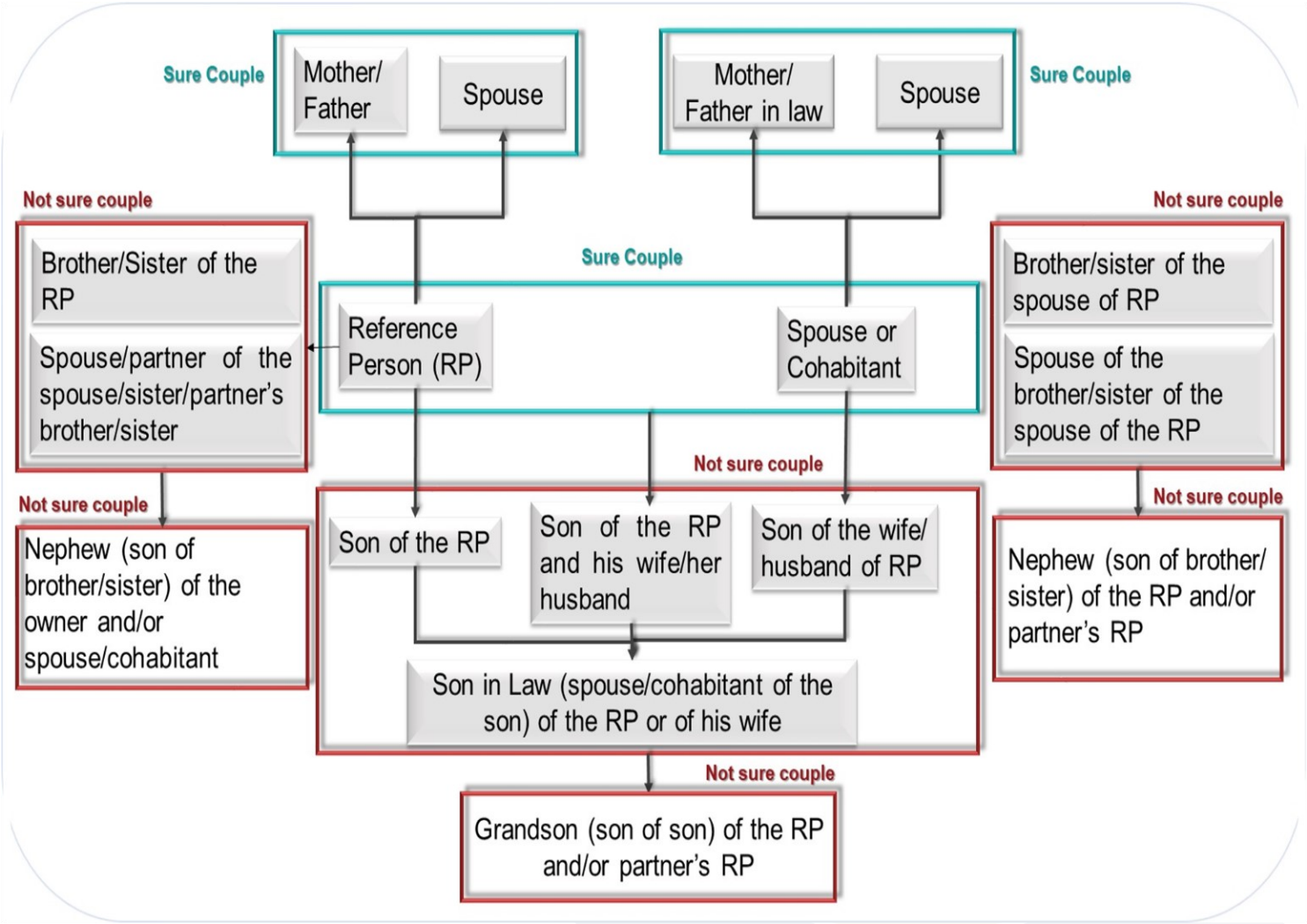
TYPE OF FAMILY NUCLEUS
Couple without children
Couple with children
Lone father with children
Lone mother with children

HOUSEHOLD TYPE	
HOUSEHOLD WITHOUT NUCLEI	
1	People living alone
2	People living with other persons
HOUSEHOLD WITH A NUCLEUS WITHOUT OTHER PERSONS	
3	Couple without children
4	Couple with children
5	Lone father with children
6	Lone mother with children
HOUSEHOLD WITH A NUCLEUS WITH OTHER PERSONS	
7	Couple without children
8	Couple with children
9	Lone father with children
10	Lone mother with children
HOUSEHOLD WITH TWO OR MORE NUCLEI	
11	Household with two or more nuclei

The conference is partly financed by the European Union

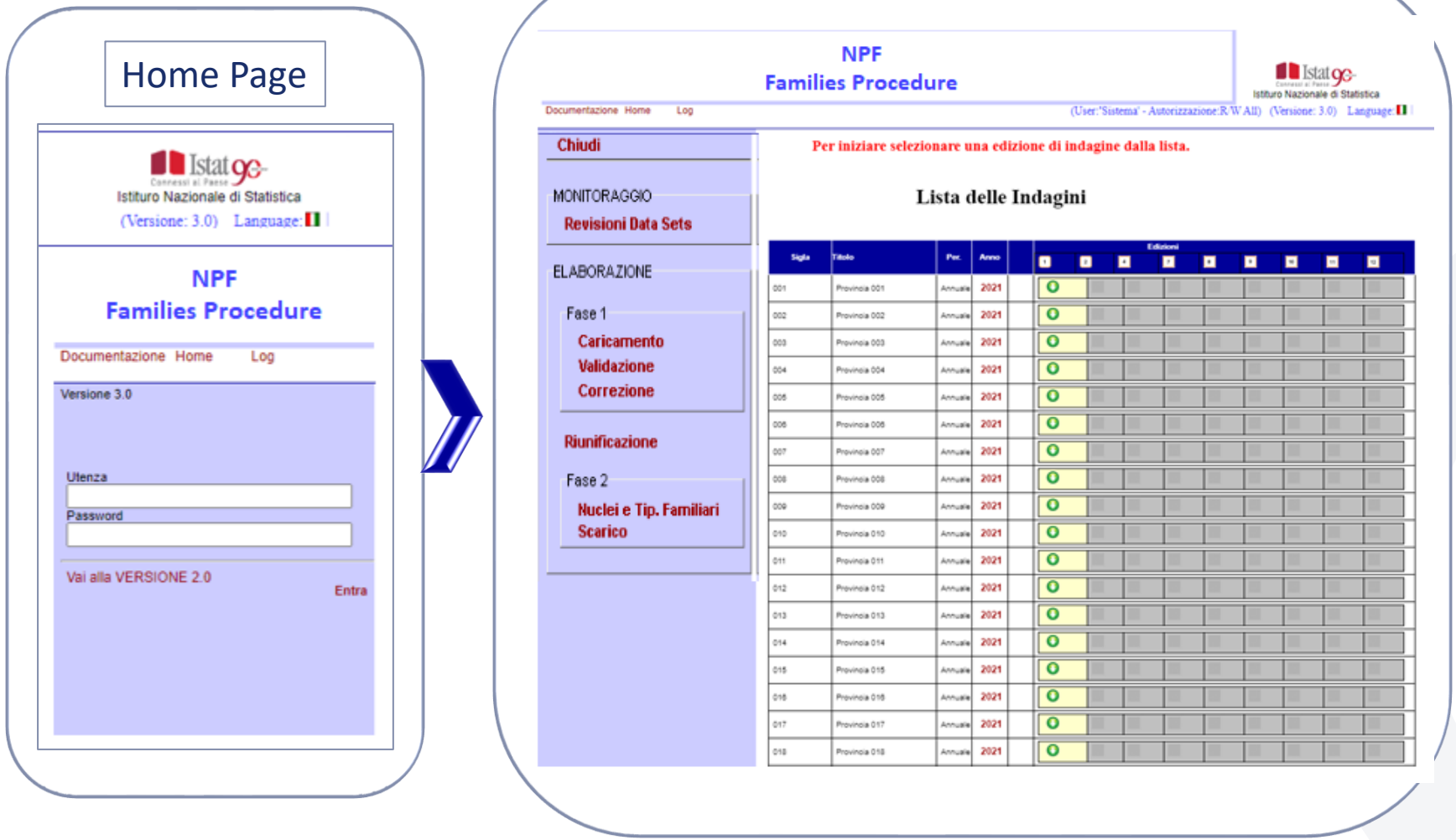
Identification of Potential couples (2/2)

- Based on optimization techniques components of couples having non unique relationship to Reference Person (RP)
- Identified prior to editing
- Score based on the responses provided to the demographic variables



The conference is partly financed by the European Union

“Families Procedure” Software



Test & Integration of the Families Procedure used for Social Survey in the E&I process of Permanent Census (2018, 2019 and 2021):

- ✓ Definition & Execution of *test case* (data, rules, server application) of the procedure on some provinces
- ✓ Metadata tables for integration with the E&I system
- ✓ Adjustment of the classifications of some variables
- ✓ Analyses of the performance and Output

Adaptation to Household and Nuclei reconstruction of RBI-CENS at 31.12.2021

- ✓ Subsetting the number of households by the household size & by grouping some provinces in order to improve the performance and reduce the execution time of the PF
- ✓ Loading data from Oracle DB
- ✓ Parallel execution by groups of provinces/editions
- ✓ Batch execution of the components of the Families Procedure

The conference is partly financed by the European Union

Main results (1/3)

Distribution of the corrections (modified or imputed values) of the variable marital status, by age groups, gender and citizenship (Italian (It) and Foreign (For)) after the E&I process. Absolute values. Percentage values in brackets.

	Women			Men			
Age groups	It	For	TotW	It	For	TotM	Total
0-16	4,645 (0.11%)	5,121 (0.13%)	9,766 (0.24)	5,084 (0.13%)	5,783 (0.14%)	10,867 (0.27%)	**20,633 (0.51%)**
17-29	24,780 (0.61%)	65,179 (1.61%)	89,959 (2.22%)	14,129 (0.35%)	82,623 (2.04%)	96,752 (2.39%)	**186,711 (4.61%)**
30-59	1,059,383 (26.18%)	265,477 (6.56%)	1,324,860 (32.74%)	947,532 (23.41%)	254,623 (6.29%)	1,202,155 (29.71%)	**2,527,015 (62.44%)**
60-84	523,396 (12.93%)	76,637 (1.89%)	600,033 (14.83%)	604,018 (14.93%)	41,4247 (1.02%)	645,442 (15.95%)	**1,245,475 (30.78%)**
85 and over	31,866 (0.79%)	2,631 (0.07%)	34,497 (0.85%)	31,495 (0.78%)	1,131 (0.03%)	32,626 (0.81%)	**67,123 (1.66%)**
Total	**1,644,070 (40.62%)**	**415,045 (10.26%)**	**2,059,115 (50.88%)**	**1,602,25 8 (39.59%)**	**385,584 (9.53%)**	**1,987,842 (49.12%)**	**4,046,957 (100%)**

Source: Our elaboration

The conference is partly financed by the European Union

Main results (2/3)

Distribution of the variable marital status before and after the E&I process (modified or imputed values). Absolute values. Row percentage values in brackets.

Before	After				
	Never married	Married or civil partnership	Divorced	Widowed	Total
Never married	0	37,067 (98.6%)	184 (0.5%)	355 (0.9%)	**37,606 (100%)**
Married or civil partnership	73,953 (1.9%)	3,801,876 (96.6%)	17,016 (0.4%)	42,352 (1.1%)	**3,935,197 (100%)**
Divorced	150 (0.8%)	19,599 (99.1%)	16 (0.1%)	16 (0.1%)	**19,781 (100%)**
Widowed	510 (13.6%)	3,067 (81.9%)	136 (3.7%)	29 (0.8%)	**3,744 (100%)**
Unknow	483,297 (59.6%)	157,205 (19.4%)	60,224 (7.4%)	109,889 (13.6%)	**810,615 (100%)**
Total	**557,910 (11.6%)**	**4,018,814 (83.6%)**	**77,578 (1.6%)**	**152,641 (3.2%)**	4,806,943 (100%)

81.9%

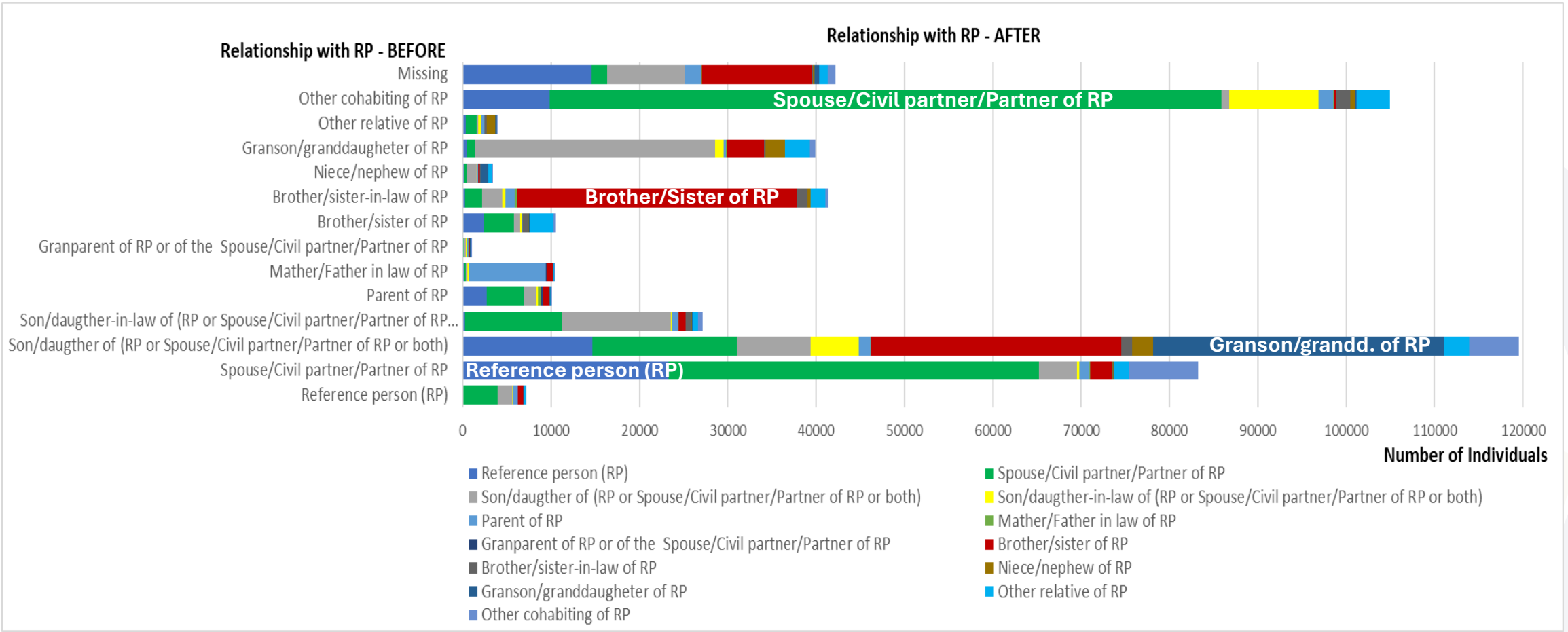
16.9%

Source: Our elaboration

The conference is partly financed by the European Union

Main results (3/3)

Distribution of the variable relationship with reference person before and after the E&I process (modified or imputed values). Bars indicate the percentage of



Source: Our elaboration

The conference is partly financed by the European Union

Household type distribution

Distribution of households type for RBI-CENS2021 and Population census at 2011. Differences between the two values. Absolute and percentage values.

	RBI-CENS2021		CENSUS 2011		DIFFERENCES	
HOUSEHOLD TYPE	**A.V.**	**%**	**A.V.**	**%**	**A.V.**	**%**
HOUSEHOLD WITHOUT NUCLEI	**10.161.306**	**38,8**	**8.319.826**	**33,8**	**1.841.480**	**22,1**
People living alone	9.636.232	36,8	7.667.305	31,2	1.968.927	25,7
People living with other persons	525.074	2	652.521	2,7	-127.447	-19,5
HOUSEHOLD WITH A NUCLEUS WITHOUT OTHER PERSONS	**14.620.430**	**55,8**	**14.879.765**	**60,5**	**-259.335**	**-1,7**
Couple without children	4.660.056	17,8	4.628.991	18,8	31.065	0,7
Couple with children	7.026.068	26,8	8.062.226	32,8	-1.036.158	-12,9
Lone father with children	564.869	17,8	374.599	1,5	190.270	50,8
Lone mother with children	2.369.437	9	1.813.949	7,4	555.488	30,6
HOUSEHOLD WITH A NUCLEUS WITH OTHER PERSONS	**1.037.178**	**4,0**	**1.061.785**	**4,3**	-24.607	**-2,3**
Couple without children	200.465	17,8	339.692	1,4	-139.227	-41,0
Couple with children	323.319	1,2	470.891	1,9	-147.572	-31,3
Lone father with children	179.189	17,8	61.454	0,2	117.735	191,6
Lone mother with children	334.205	1,3	189.748	0,8	144.457	76,1
HOUSEHOLD WITH TWO OR MORE NUCLEI	**387.332**	**1,5**	**350.390**	**1,4**	**36.942**	**10,5**
Household with two or more nucleus	387.332	17,8	350.390	1,4	36.942	10,5
TOTAL	**26.206.246**	**100**	**24.611.766**	**100**	**1.594.480**	**6,5**

Source: Our elaboration

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Concluding Remarks

- Briefly description of the process of the household and nuclei types reconstruction:
  - 2018, 2019 and 2021 census experiences
  - RBI-CENS2021 based on 2018, 2019 and 2021
- Point out the complexity linked to:
  - the integrated use of data gathered from registers, survey, administrative and Istat sources;
  - the adaptation of the PF to a huge amount of data.
- The first time that PF was used on integrated data, without never having tested it on big dataset.
- Improvement the quality of data released to Eurostat with reference to census hypercubes involving household and nuclei types.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Further developments

Methodological aspects

- Further studies, both on sources and methods, useful to reduce missing data and errors as much as possible to better **improve** the data **quality**.
- Analysing and comparing the classification of relationship with other social survey to **harmonize** the outcome.
- Use of ML or AI to improve the household reconstruction **minimizing errors**, especially for households with numerous members which internal composition is difficult to detect.

IT aspects

- Reengineering the “Families Procedure” to **optimize** the speed of its execution and the performance by reducing some anomalous household.
- Use of new generation programming languages in order to **better maintain** the application.
- Generalised solutions to **adapt** PF to specific need of other social survey.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Some references

- ANPR (2024). Anagrafe Nazionale Popolazione Residente. https://www.anagrafenazionale.interno.it
- Bianchi G, Filippini R, Lipsi RM, Pezone A, Scalfati F. (2020). An overview of the editing and imputation process of the 2018 Italian Permanent census. UNECE, online workshop on Statistical Data Editing.
- Bruni R, A Reale and R Torelli, Optimization Techniques for Edit Validation and Data Imputation (2001). In Proceedings of Statistics Canada Symposium: Achieving Data Quality in a Statistical Agency, Ottawa, Canada.
- Budano G. e P. Piergentili (2010), La Procedura Famiglie in G. Budano e S. Demofonti, La misurazione delle tipologie familiari nelle indagini di popolazione in Metodi e Norme, 2010, n. 46. Istat.
- D.P.R. 30 maggio 1989, n. 223.
- Eurostat, (2017). European Commission. Commission Regulation (EU) 2017/712 of 20 April 2017 establishing the reference year and the programme of the statistical data and metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council, OJ L 105, 21.4.2017, p. 1–11, https://eur-lex.europa.eu/eli/reg/2017/712/oj
- GDPR (2016). General Data Protection Regulation (GDPR – Regulation 2016/679).
- Istat (2022). Nota tecnica sulla produzione dei dati del Censimento Permanente: la popolazione residente per genere, età, cittadinanza e grado di istruzione al 31.12.2021. pp.14. https://www.istat.it/it/files//2022/12/Nota-metodologica-censipop-_2021.pdf

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

An innovative approach to improve the quality of the household and nuclei types reconstruction in Italy

THANKS FOR YOUR ATTENTION!

Rosa Maria Lipsi, ISTAT-Italian National Institute of Statistics, DIRM/DCME/MEB, lipsi@istat.it
Anna Pezone, ISTAT-Italian National Institute of Statistics, DIPS/DCDC/DCB, pezone@istat.it

05 June 2024

Session 7

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union