Vienna, 7-9 October 2024
UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing

THE EDITING AND IMPUTATION PROCESS OF THE 2021 HOUSEHOLD AND NUCLEI TYPES RECONSTRUCTION IN ITALY

ROSA MARIA LIPSI
Istat | DCME

ANNA PEZONE
Istat | DCDC

Outline

- **Introduction**
- **Data and methods**
- **Focus on the Editing and Imputation**
  - Phase I: Preliminary E&I activities
  - Auxiliary variables
  - Phase II: E&I process after the “Families Procedure”
- **Main results of the E&I process**
- **Final remarks and further developments**

Aim of the work

E&I process: Revision and Innovation

Taking into account

The whole process to produce statistics on the household and their characteristics

→ By using →

Data collection based on the integration of multiple sources:

- RBI-CENS2021
- ANPR
- Survey
- Internal Istat data

Introduction

CENSIMENTI PERMANENTI L'ITALIA, GIORNO DOPO GIORNO. POPOLAZIONE E ABITAZIONI

Istat

Since 2018, ISTAT, 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)

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

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 (as **RBI – Based Register of Individuals**) and surveys data, has been used to provides information on Italian Population and Housing Census for the 2021, as required by the **EU regulation 2017/712**.

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 households, as well as Nuclei and Family types.

Data and methods (1/3)

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	COVID-19
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

Sample of Italian households

RBI 2020
31 December

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

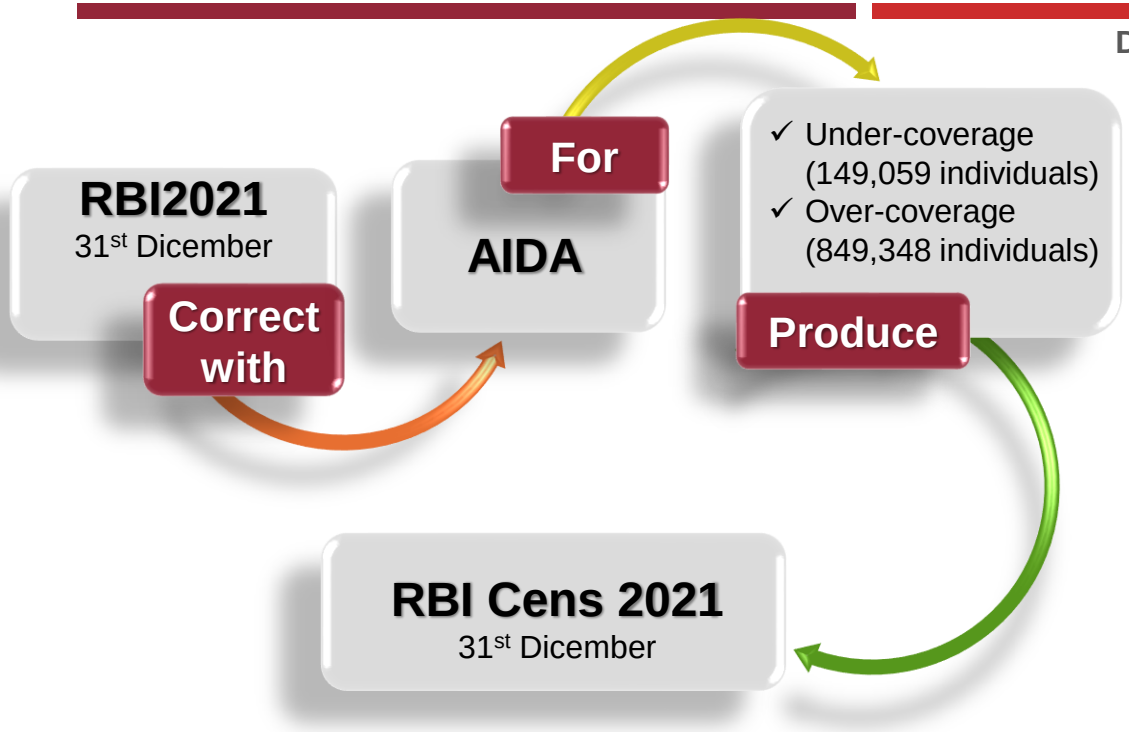
AIDA
Archive of Usual Resident Population

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

Over-coverage

RBI CENS2020
31 December

Data and methods (2/3)



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

Distribution of the Italian Population and Households by regions at 31st of December 2021. Absolute and percentage values

	Number of Individuals		Number of Households	
Regions	A.V.	%	A.V.	%
Piemonte	4,218,723	7.2	2,001,951	7.6
Valle d'Aosta	122,547	0.2	60,468	0.2
Lombardia	9,882,579	16.8	4,492,423	17.1
Trentino-Alto Adige/Südtirol	1,061,745	1.8	469,907	1.8
Veneto	4,812,583	8.2	2,109,478	8.0
Friuli Venezia Giulia	1,184,966	2.0	564,743	2.2
Liguria	1,495,874	2.5	760,931	2.9
Emilia Romagna	4,391,763	7.5	2,032,219	7.8
Toscana	3,642,200	6.2	1,662,574	6.3
Umbria	853,493	1.5	383,931	1.5
Marche	1,479,967	2.5	646,864	2.5
Lazio	5,672,202	9.7	2,630,892	10.0
Abruzzo	1,270,858	2.2	558,313	2.1
Molise	290,367	0.5	130,888	0.5
Campania	5,606,656	9.6	2,212,896	8.4
Puglia	3,910,701	6.7	1,635,899	6.2
Basilicata	538,773	0.9	237,160	0.9
Calabria	1,848,679	3.2	808,445	3.1
Sicilia	4,812,598	8.2	2,066,148	7.9
Sardegna	1,581,521	2.7	740,116	2.8
Total	**58,678,795**	**100**	**26,206,246**	**100**

Source: Our elaboration on Istat data

Data and methods (3/3)

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

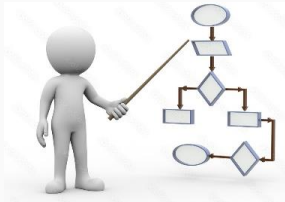
	VARIABLES								
UNITS	ID NUMBER	ID HHOLD	GENDER	DATE OF BIRTH	CITIZENSHIP	RELATIONSHIP	MARITAL STATUS	YEAR OF MARRIAGE OR CIVIL UNION	.....
	000001	000001	x11	x12	x13	x14	x15	x16	.....
	000002	000001	x21	x22	x23	x24	x25	x26	.....
	000003	000001	x31	x32	x33	x34	x35	x36	.....
	.....	000002	x..	x..	x..	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 CENS 2021 + **ANPR** (National Register of Resident Population)

Phase I: Preliminary E&I activities

Data Base RBI_Cens2021

Elimination of ANPR macroscopic errors



Recording procedure from RBI/ANPR to PPHC (Kinship Relationship –KR and Marital Status – MS)

For each household, determination, of a string of individual progressive number codes with the same surname

Data Base for E&I

String comparison process of surname

$$d(Surname_i, Surname_j) < \delta \quad \forall i,j=1,2,..,\ n° components$$

Function that measures the similarity between two strings calculated by using an internal method based on N-gram algorithm and Jaro-Wikcler distance. Smaller distances correspond to more similar strings.

Acceptability threshold (calculated taking into account observed data) The $0 \leq \delta \leq 1$ (with 0=max similarity, 1=min similarity) value

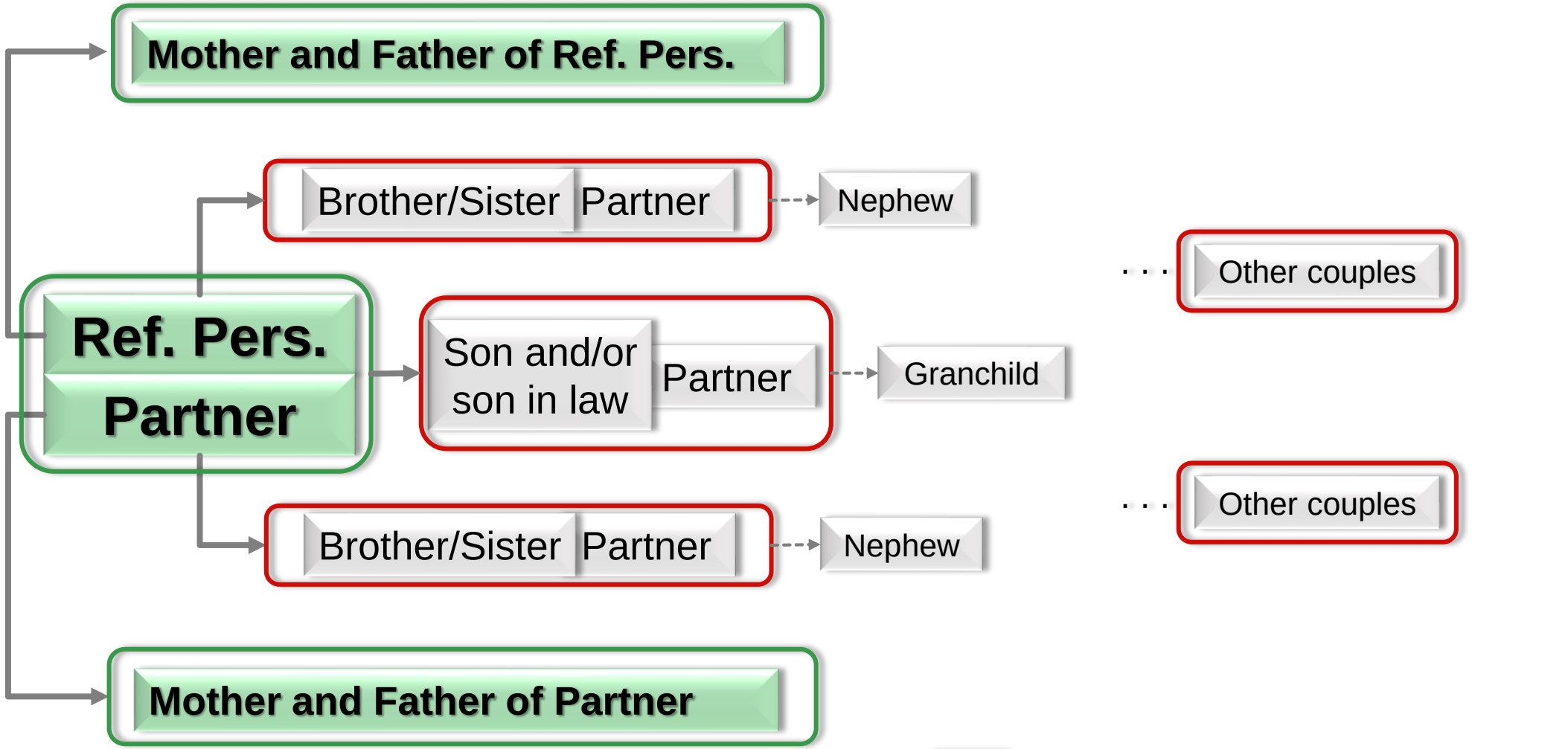
ID family	Prog. Individual	Surname	Relathionship
0000001	**1**	**Xxxxxxx**	**Ref. Person**
0000001	2	Yyyyy	Partner
0000001	**3**	**Xxxbxxxx**	**Daugther**
0000001	**4**	**Xxxxxxx**	**Son**
0000001	5	Kkkk	Other relative

A new variable is created using 'progressive individual' when there are two or more components with $d<\delta$

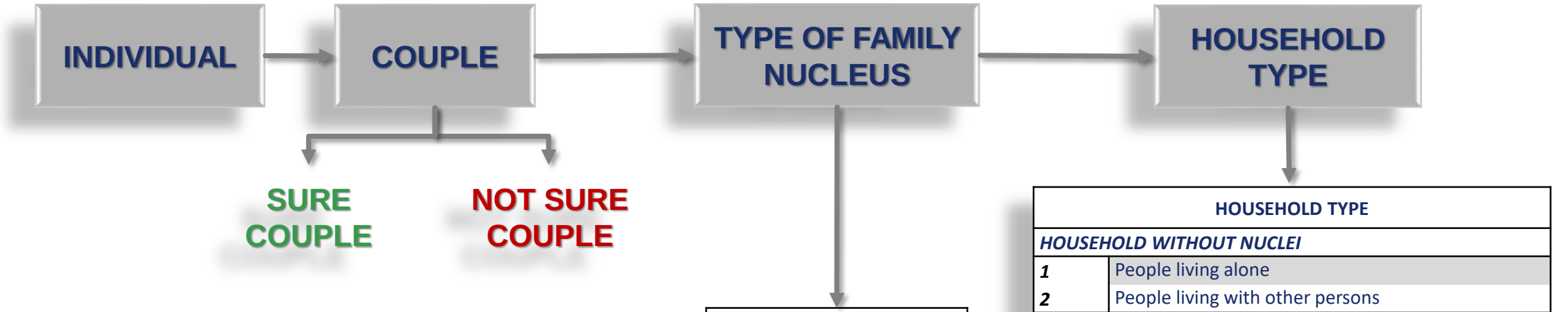
String=1-3-4

Individuals N° 1,3 and 4 have surnames with $d<\delta$

Identification of potential couples (1/2)



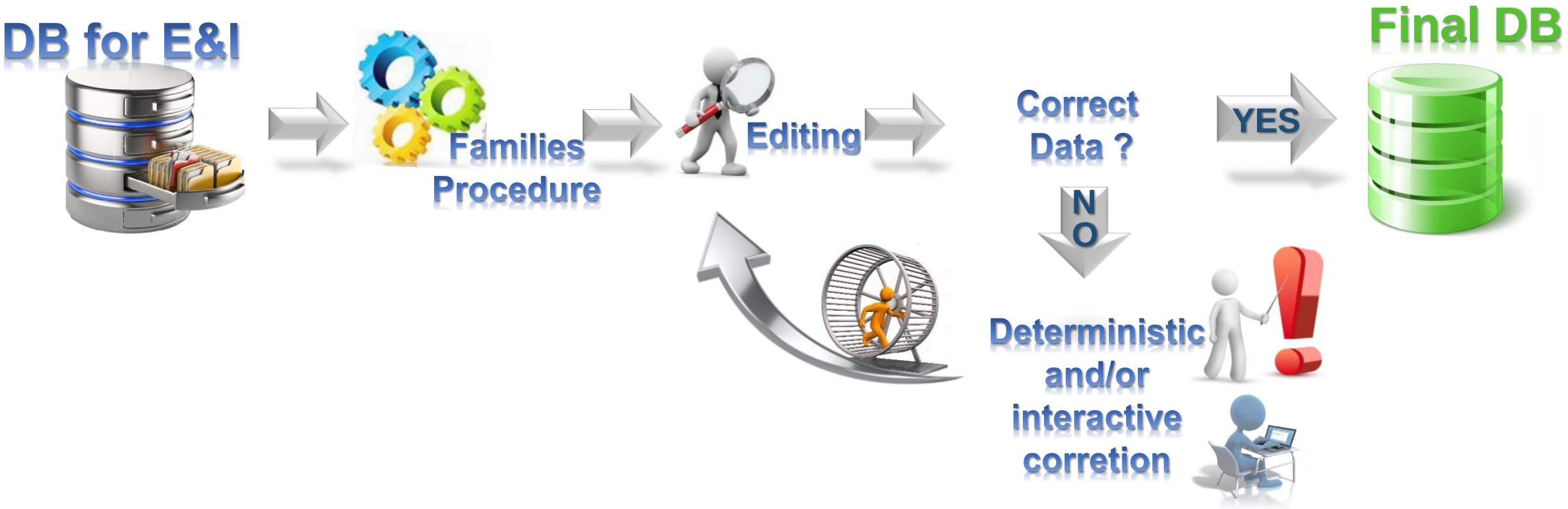
Identification of potential couples (2/2)



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

Phase II: Recursive E&I process after the FP



The main results of the E&I process (1/6)

6.5 % of Total Population

MISSING DATA	Number of errors	
	A.V.	%
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
Total	3,788,122	100

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

Distribution of missing data for marital status, year of marriage or civil union and relationship with RP by REGION. Percentage values.

Regions	Marital status	Year of marriage or civil union	Relationship with RP
Piemonte	6.38	7.03	4.75
Valle d'Aosta	0.22	0.23	0.06
Lombardia	25.45	23.19	16.98
Trentino-Alto Adige/Südtirol	2.43	1.84	1.80
Veneto	10.01	8.34	8.54
Friuli Venezia Giulia	2.81	1.94	2.09
Liguria	4.08	3.85	2.16
Emilia Romagna	11.09	9.41	7.47
Toscana	9.52	8.59	9.76
Umbria	1.49	1.39	0.94
Marche	3.26	2.62	2.20
Lazio	10.46	11.15	15.07
Abruzzo	1.40	1.86	1.53
Molise	0.15	0.28	0.97
Campania	3.56	7.30	10.78
Puglia	2.74	3.15	4.52
Basilicata	0.33	0.36	0.45
Calabria	0.91	1.61	3.07
Sicilia	2.44	4.29	5.97
Sardegna	1.25	1.57	0.88
Total	**100**	**100**	**100**

Source: Our elaboration on Istat data

The main results of the E&I process (2/6)

94 edit rules

13 individual & 81 familial

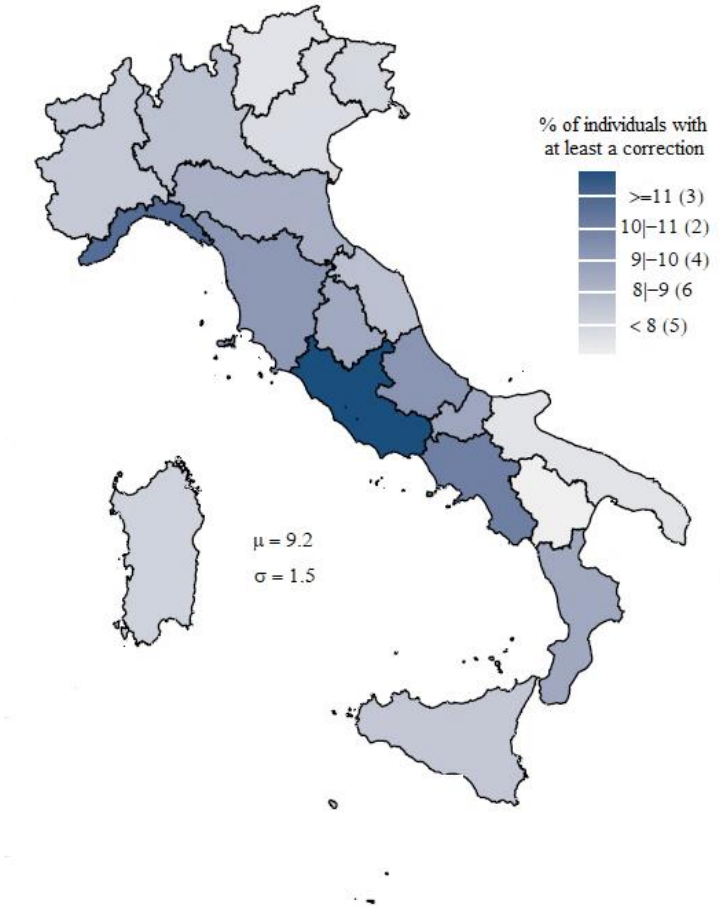
Individual inconsistencies

(937,328 failed edits (1.6%)

Familial inconsistencies

(223,082 failed edits (0.9%)

Percentage of individuals with at least a correction by REGION (in brackets the number of regions)



Source: Our elaboration on Istat data

The main results of the E&I process (3/6)

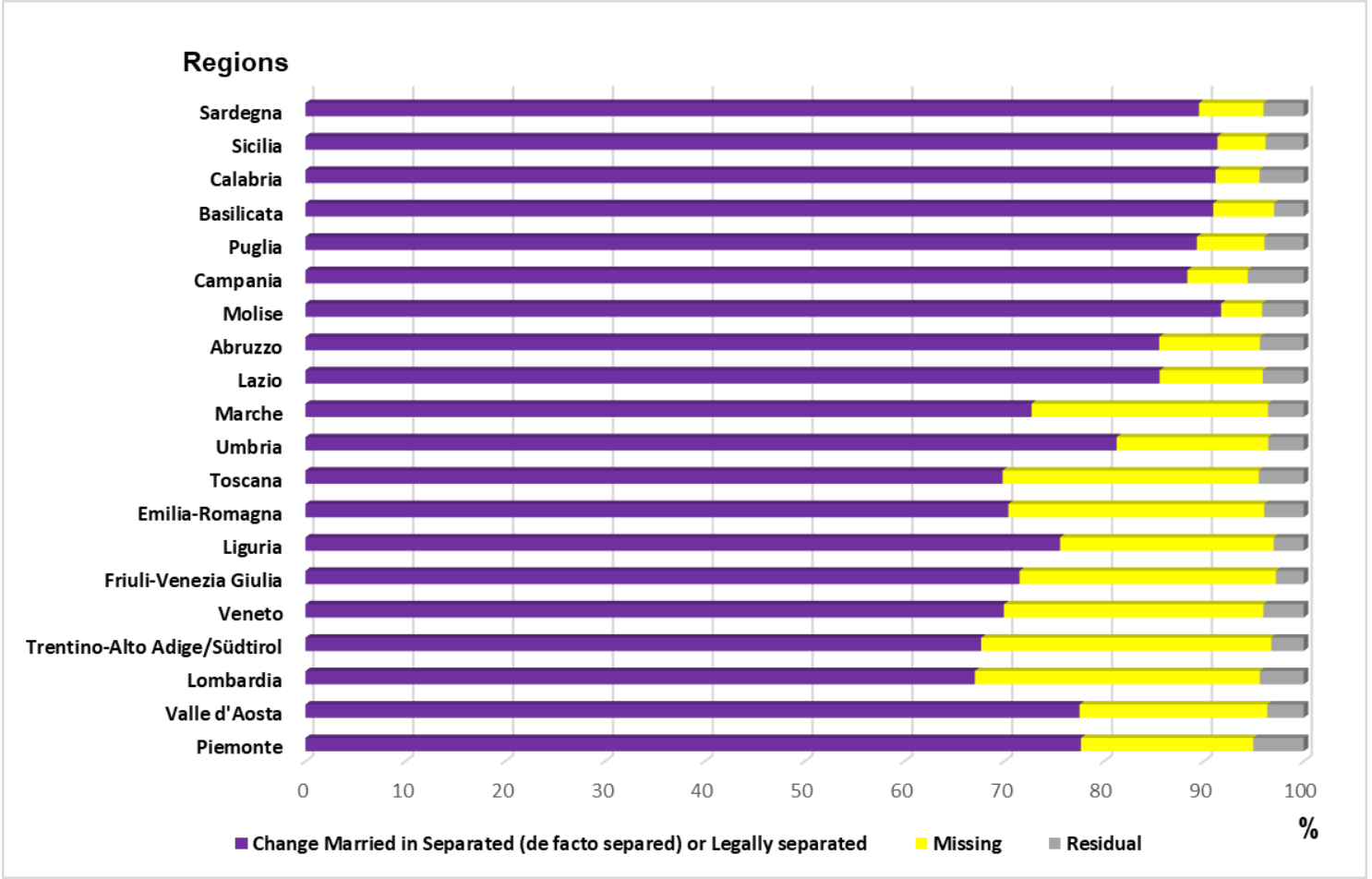
Distribution of the errors of marital status, by age groups, gender and citizenship (Italian (It) and Foreign (For))
Percentage values

	Women			Men			
Age groups	It	For	TotW	It	For	TotM	Total
0-16	0.11	0.13	0.24	0.13	0.14	0.27	0.51
17-29	0.61	1.61	2.22	0.35	2.04	2.39	4.61
30-59	26.18	6.56	32.74	23.41	6.29	29.71	62.44
60-84	12.93	1.89	14.83	14.93	1.02	15.95	30.78
85 and over	0.79	0.07	0.85	0.78	0.03	0.81	1.66
Total	40.62	10.26	50.88	39.59	9.53	49.12	100

Source: Our elaboration on Istat data

The main results of the E&I process (4/6)

Distribution of the marital status (married) before/after the E&I process. Bars are % of each category.



Source: Our elaboration on Istat data

The main results of the E&I process (5/6)

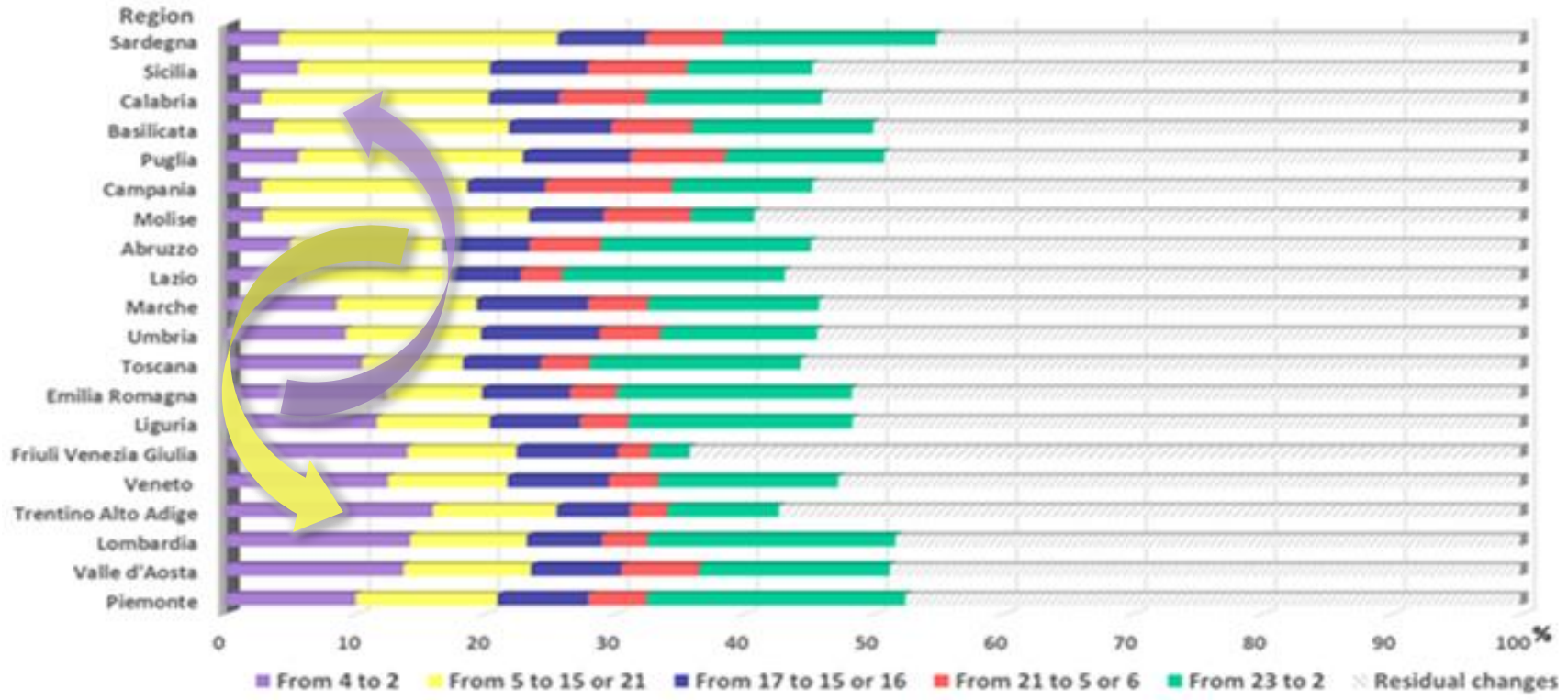
Distribution of the errors of relationship with RP, by age groups, gender and citizenship (Italian (It) and Foreign (For)). Percentage values

	Women			Men			
Age groups	It	For	TotW	It	For	TotM	Total
0-14	4.30	1.44	5.73	4.52	1.50	6.02	11.76
15-29	4.94	2.78	7.73	4.58	2.39	6.97	14.70
30-49	14.19	8.16	22.35	10.49	6.96	17.44	39.79
50-64	9.68	2.50	12.18	8.16	1.94	10.11	22.29
65-84	4.91	0.65	5.56	3.79	0.39	4.18	9.74
85 and over	1.24	0.03	1.27	0.42	0.02	0.44	1.71
Total	39.26	15.57	54.83	31.97	13.20	45.17	100

Source: Our elaboration on Istat data

The main results of the E&I process (6/6)

Distribution of some categories of the relationship with RP before/after the E&I. Bars are % of each category



Source: Our elaboration on Istat data

2 = Spouse of RP
4 = Partner of RP (consensual union)
5 = Son/daughter of RP and of the spouse/civil partner/partner
6 = Son/daughter of RP only

15 = Brother/sister of RP
16 = Brother/sister of the spouse/civil partner/partner of RP
17 = Spouse of the brother/sister of RP or of the spouse/civil partner/partner of RP
21 = Grandson/granddaughter of RP and/or of the spouse/civil partner/partner of RP
23 = Other cohabiting person without being a member of a couple, a relative, or extended family

Final 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
- Revision of the overall E&I system involving innovative generalized solution
- 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.
- First time that PF was used on integrated data, without never testing it on big dataset.

Future developments

Methodological & IT aspects

- Further studies, both on sources and methods of E&I, useful to reduce missing data and errors.
- Use of ML or AI to improve the E&I process in order to minimize errors in the household reconstruction, especially for households with numerous members which internal composition is difficult to detect.
- Reengineering the “Families Procedure” to **optimize** the speed of its execution and the performance by reducing any anomalous household.
- Further use of new auxiliary variables.
- Use of new generation programming languages in order to **better maintain** the application.

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.
- Budano G. and 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.
- Eurostat, (2017). European Commission. Commission Regulation No 763/2008 of the European Parliament and of the Council, OJ L 105, 21.4.2017, p. 1–11.
- 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.
- Lipsi R.M. and A. Pezone (2024), An innovative approach to improve the quality of the household and nuclei types reconstruction in Italy, *Q2024*, Estoril, 04-07 June 2024, pp. 10.
- Sogei (2024), Sogei - Società Generale d'Informatica S.p.A., società di Information Technology, https://www.sogei.it/it/sogei-homepage.html

THANKS FOR YOUR ATTENTION!

ROSA MARIA LIPSI | ISTAT| DCME | lipsi@istat.it

ANNA PEZONE | ISTAT| DCDC | pezone@istat.it

Istat | Istituto Nazionale di Statistica