International Network for Comparative Analysis of Social Inequalities (INCASI)

3rd ISA Forum of Sociology / Vienna

RC55: Imputation and Social Indicators: The Use of Factor Analysis for Imputing Missing Data

Imputations for Missing Data in Income Variables. Permanent Household Survey (EPH). Gran Buenos Aires, Argentina / 1990-2010

Eduardo Donza (INCASI - UBA/UCA)

The content of this conference is part of INCASI Network, a European project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie GA No 691004, coordinated by Dr. Pedro López-Roldán

7-11-2016

Research Problem and Objectives

- **The research begins with the academic concern about the critical impact of missing data in income variables regarding to living conditions and inequality.**
- **The main objective is to describe the scope of missing data in income variables and its effect for research based on these kind of variables.**
- **It also applies a methodological strategy in order to impute values to missing data, and to recalculate social indicators.**

Main Features of Income Measurement in EPH

Changes in Data Source

EPH (from 1974 to 1994) –called ‘Puntual’ –
• 2 surveys per year. • Survey questionnaire: 6 questions referred to incomes.
EPH (from 1995 to May, 2003) –called ‘Puntual - BUA’ –
• 2 surveys per year. • Survey questionnaire: 13 questions referred to incomes.
EPH (from May, 2003 until the present) –called ‘Continua’ –
• Continuous survey • Quarterly data • Survey questionnaire: 21 questions referred to incomes.

Source: author’s elaboration based on INDEC.

Possible Problems and Errors in Income Measurement

- **Representativeness bias due to sample truncation.**
- **Absence of some income questions and changes in conceptual definitions. Changes properly done should not impact in terms of comparability.**
- **Underdeclaration, no-answer or lack of information from the respondent.**

Missing Data Problems

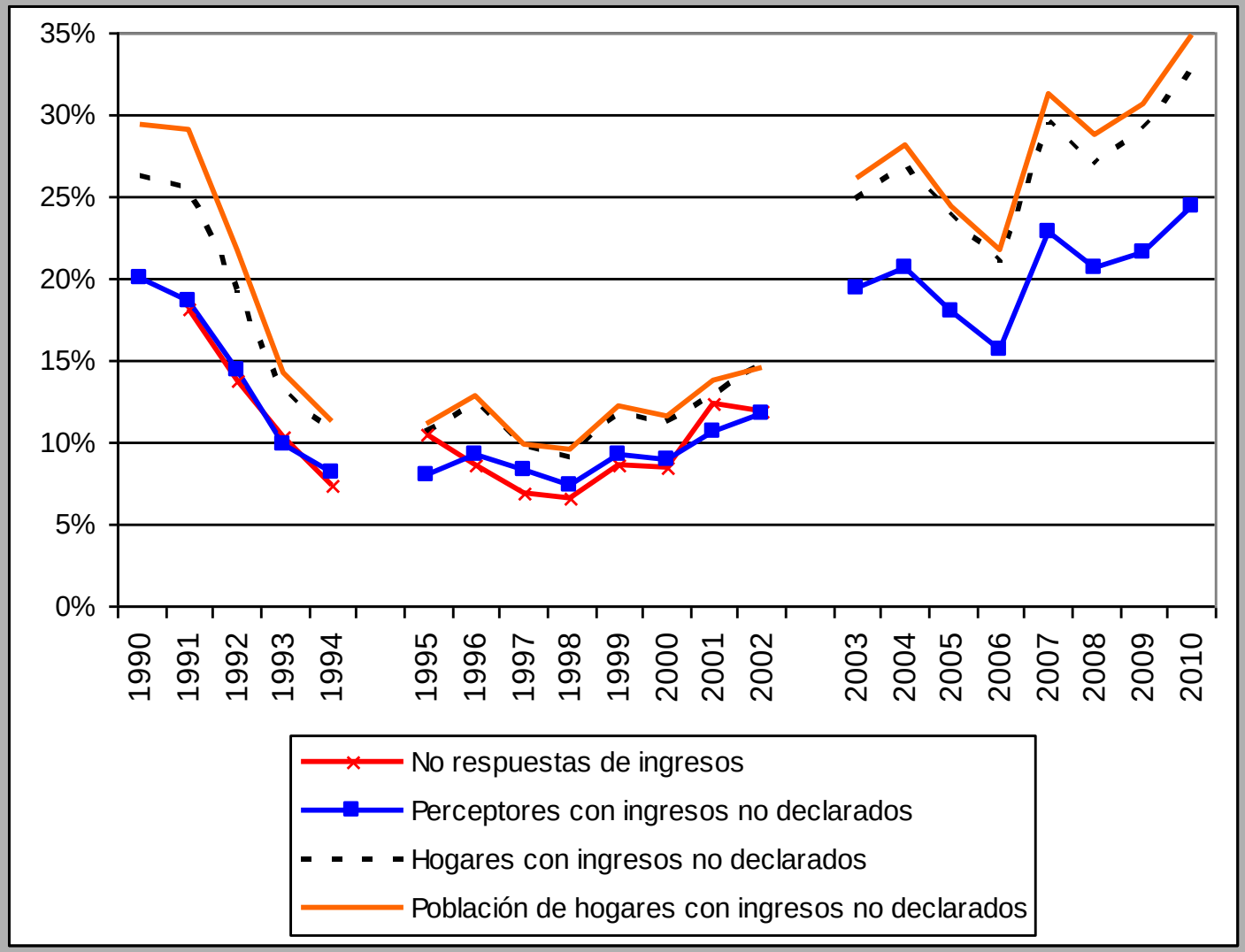
Data quality is affected both by researchers' decissions and respondents' context. Among them:

- **Conceptual framework of the questionnaire;**
- **Survey Data Collection**
- **Another phases of information.**
- **Economic, social and institutional context factors.**

Available Information and Data Analysis

Missing values in income variables, recipients, households and population affected by no-answer. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

In Percentages.



Source: author's elaboration based on EPH, INDEC.

Survey respondent with missing data in income variables by socio-economic features. Gran Buenos Aires: selected years.

Percentage of each group and significance of ANOVA Test.

Perfil Social de No	Declarantes	1990	1991	1992	.....	2008	2009	2010
Sexo	Varón	22,1	21,5	16,2		22,5	24,3	27,0
	Mujer	16,9	14,3	11,9		18,6	18,8	21,5
Edad	Hasta 24 años	21,7	16,9	15,4		19,4	20,1	21,2
	Entre 25 y 44 años	21,5	21,0	15,5		20,4	22,3	26,6
	Entre 45 y 64 años	22,6	23,1	17,6		24,7	24,4	26,4
	65 años y más	11,5	7,7	6,5		15,8	16,8	18,3
Nivel de Instrucción	Primario incompl.	15,6	12,9	10,1		14,4	12,5	14,1
	Secundario incompl.	19,3	18,1	14,8		14,7	16,5	18,6
	Secundario completo	21,2	20,5	15,3		24,1	25,3	27,0
	Sup. o univ. compl.	23,9	22,2	17,2		31,3	31,8	36,9
Posición en Hogar	Jefe	19,8	18,3	13,9		19,5	20,3	23,7
	No jefe	20,3	18,9	15,1		21,9	23,1	25,0
Condición Actividad	Ocupado	23,4	21,9	17,4		22,9	25,1	28,0
	Desocupado	18,5	25,6	18,7		16,1	14,8	20,8
	Inactivo	8,7	6,8	3,9		13,5	11,9	13,6
Categoría Ocupacional	Patrón o empleador	49,1	43,0	32,4		43,8	39,5	50,2
	Cuenta propia	35,8	35,2	29,4		27,4	29,8	33,2
	Obrero o empleado	16,4	15,6	12,2		20,5	22,6	25,4

Anova Test sig ($p < 0,01$).

Source: author's elaboration based on EPH, INDEC.

Continúa

Survey respondent with missing data in income variables by socio-economic variables. Gran Buenos Aires: selected years.

Percentage of each group and significance of ANOVA Test.

Perfil Social de No	Declarantes	1990	1991	1992	.....	2008	2009	2010
Nivel de Calificación	Profesional	24,8	22,8	22,8		34,3	38,2	39,4
	Calificado	21,5	20,2	16,7 ●		28,4 ●	28,8 ●	35,5 ●
	No calificado	24,2	23,5	15,8		16,8	16,0	19,2
Carácter de la Tarea	Producción	23,3	20,0	16,0		18,3	23,0	25,6
	Administr.–Contable	17,2	14,9	15,1		32,4	30,6	36,8
	Comercialización	33,5 ●	36,3 ●	28,2 ●		20,7 ●	24,1 ●	24,7 ●
	Transp., seg., y serv.	19,8	19,0	13,5		20,5	22,4	24,9
Cantidad de Ocupaciones	Sólo una ocupación	23,9	22,1	17,5		20,3	21,3	24,1
	Dos o más	16,7 ●	16,7	14,9		21,6	25,1	25,0
Fuente de Ingreso	Laborales asalariad.	-	15,5	12,0		21,0	23,2	26,1
	Labor. no asalariados	-	35,4	30,0		30,7	32,4	37,9
	Laborales mixtos	-	20,9	16,2		- ●	- ●	- ●
	Jubilación o pensión	-	3,3	2,3		10,3	9,0	12,1
	Otros no laborales	-	39,5	29,0		12,3	8,9	11,8
	Laborales y no labor.	-	26,9	18,6		21,4	21,6	25,4

● Anova Test sig ($p < 0,01$).

Source: author's elaboration based on EPH, INDEC.

Treatment for Missing Data in Income Variables

Imputation Method

- **It was verified that missing data are not associated with non-randomized pattern (NMAR). Little Test, Graph analysis and frequencies were used.**

- **Maximum Likelihood Estimation was used for imputation (MV) / SPSS.**

Maximum Likelihood Estimation Implementation (I)

- Imputations were done for missing data in each variable of the survey.
- Total income of the respondent is recalculated as a sum of all income sources (including those that were imputed).

Maximum Likelihood Estimation Implementation (II)

Independent variables for Labour-income variables:

- **Gender,**
- **Age,**
- **Education Level,**
- **Relationship to Household Head**
- **Current Activity Status (employed, unemployed)**
- **Employment Category (employee, employer, self-employed)**
- **Qualification**
- **Work features**
- **Number of occupations.**

Maximum Likelihood Estimation Implementation (III)

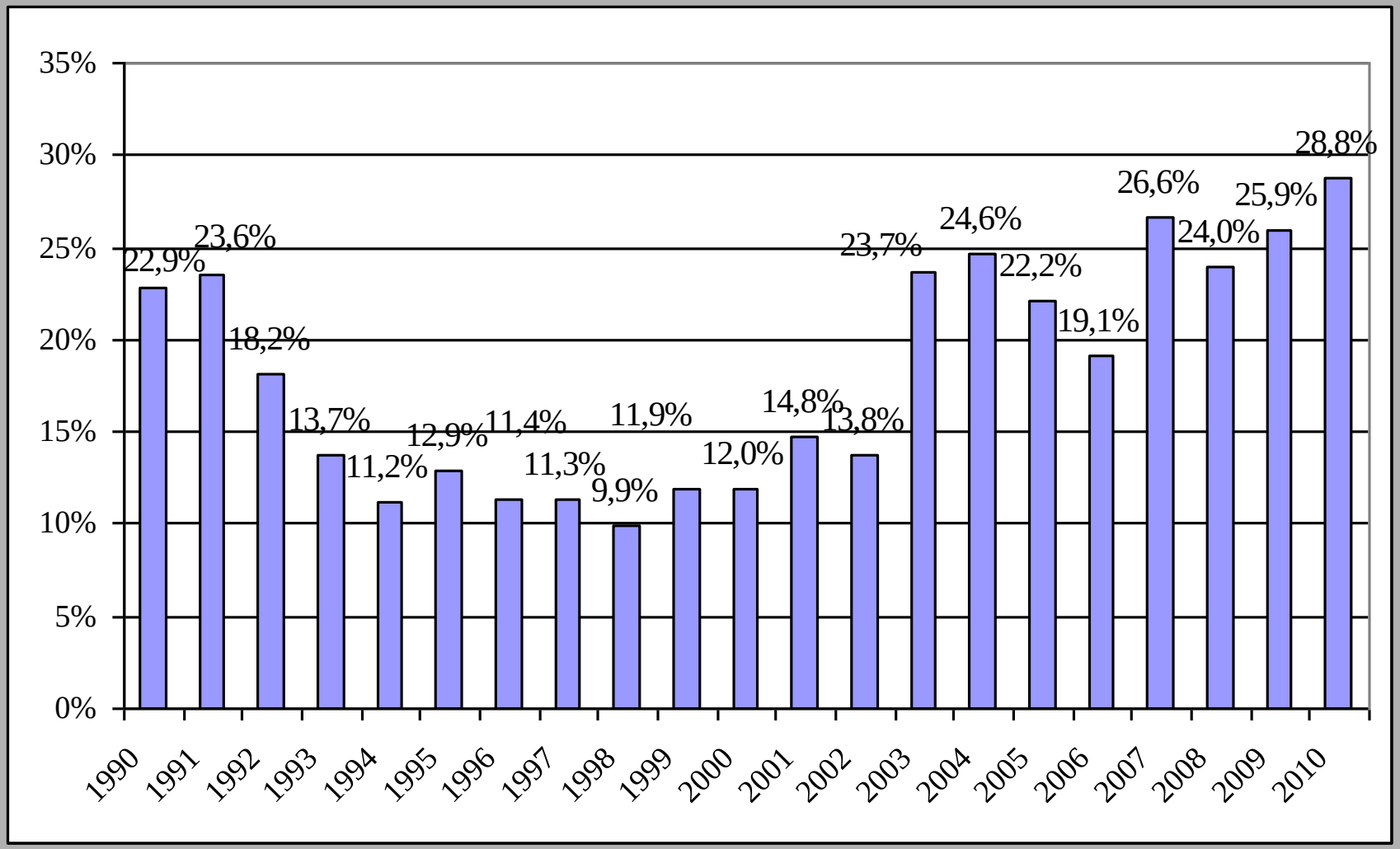
Independent variables for Non-labour income variables:

- **Gender,**
- **Age,**
- **Education Level,**
- **Relationship to Household Head**
- **Current Activity Status (employed, unemployed)**

Implementation and Results of Maximum Likelihood Estimation - Incomes before and after imputation-

Imputation Effects in Total Personal Income. Incidence of Imputed Incomes in the Total. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

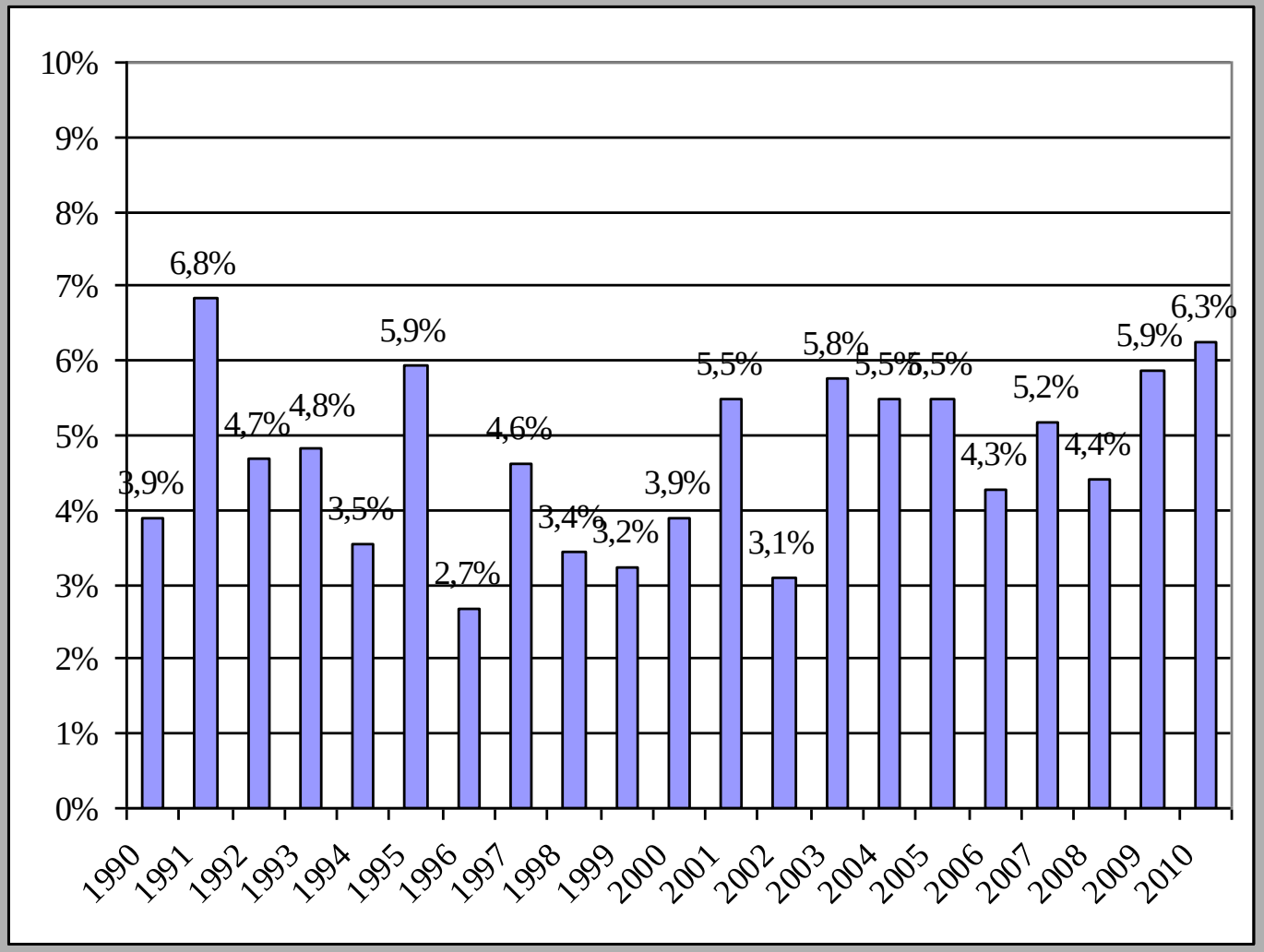
In Percentages with respect to the total of incomes of each year.



Source: author's elaboration based on EPH, INDEC.

Imputation effects in total personal income. Changes in the mean of income per recipient before and after imputation. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

From the mean



Source: author's elaboration based on EPH, INDEC.

Survey respondent with missing data in income variables by decile of income. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

As a Percentage of Incomes of each decile and significance of ANOVA Test.

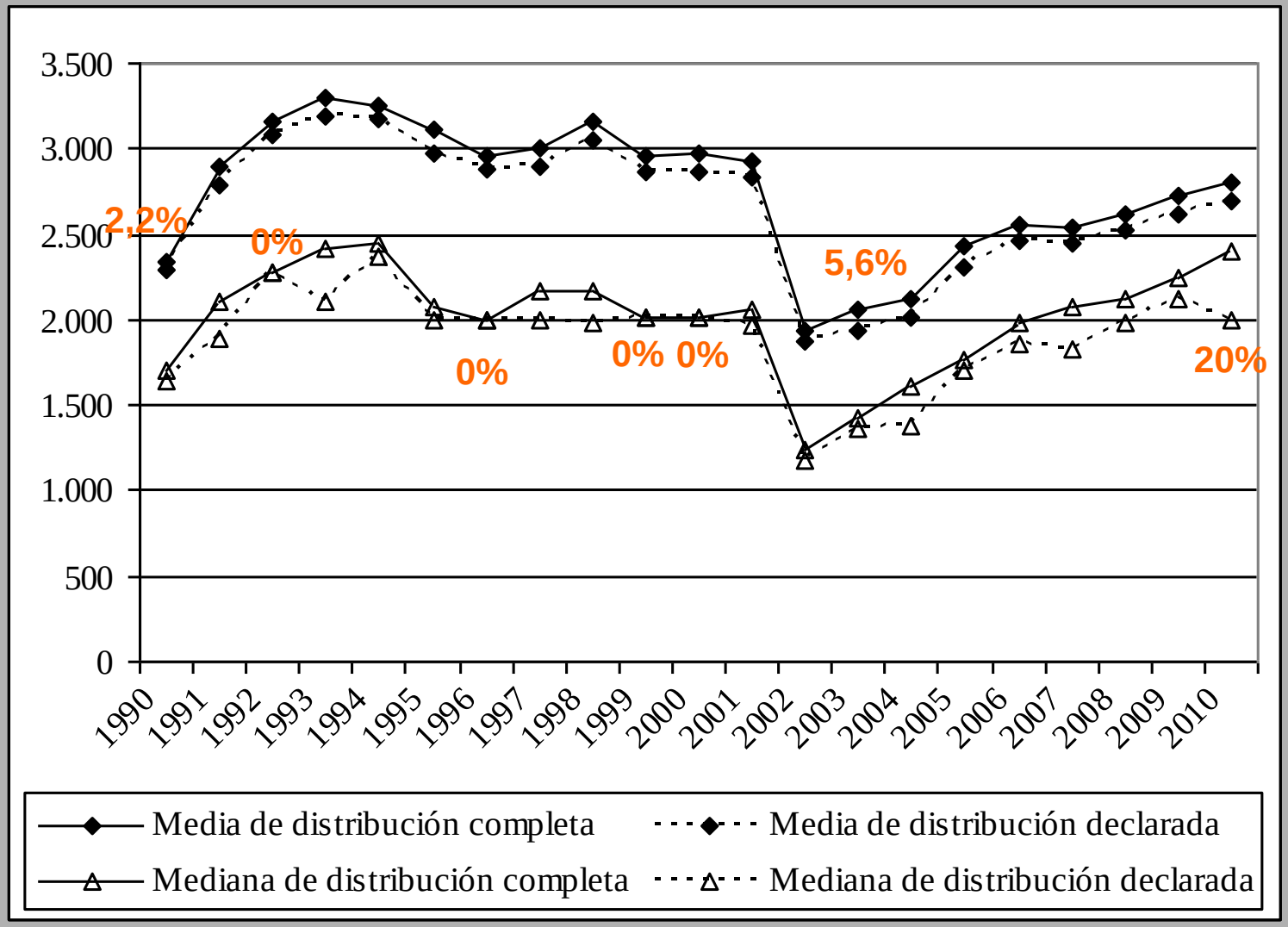
Año	Decil de ingreso total personal										Sig.
	1	2	3	4	5	6	7	8	9	10	
1990	7,4	5,4	15,5	18,2	20,9	14,1	44,9	22,4	32,7	21,9	0,000
1991	3,3	7,0	12,4	13,0	10,1	43,1	26,0	18,8	27,2	25,7	0,000
1992	1,1	3,9	14,0	8,5	12,3	12,7	27,7	17,0	31,1	17,4	0,000
1993	0,8	3,7	5,1	5,5	19,5	9,2	7,5	19,1	14,7	16,6	0,000
1994	0,5	3,8	7,0	1,8	14,7	6,9	19,0	11,5	10,9	13,4	0,000
1995	2,5	3,2	4,6	13,1	5,2	7,9	10,0	6,4	11,1	16,8	0,000
1996	2,7	4,7	5,5	14,9	8,6	7,2	14,3	7,7	12,4	15,0	0,000
1997	2,3	2,7	3,8	7,3	6,5	8,5	4,5	11,6	9,5	16,9	0,000
1998	1,8	2,2	4,7	9,5	4,9	6,5	6,7	9,7	9,7	13,5	0,000
1999	1,1	8,2	7,7	9,2	7,7	8,8	10,4	8,2	13,4	16,5	0,000
2000	3,2	3,7	4,8	11,7	3,8	10,4	12,1	7,1	16,9	14,7	0,000
2001	3,8	4,8	11,0	5,1	8,5	10,6	16,2	13,9	10,8	18,1	0,000
2002	8,3	5,4	6,0	5,1	14,0	13,9	7,0	16,9	20,3	17,9	0,000
2003	12,6	5,2	14,8	12,6	19,0	29,1	19,0	25,8	25,4	31,1	0,000
2004	5,8	13,2	14,9	16,0	20,2	27,6	21,2	27,3	29,2	30,5	0,000
2005	7,2	9,9	9,7	16,0	18,0	14,7	24,3	27,0	25,9	28,0	0,000
2006	5,4	6,7	8,5	19,6	10,8	23,7	14,7	25,8	21,8	24,1	0,000
2007	8,4	11,2	10,1	20,4	21,2	23,3	48,4	32,4	25,5	31,7	0,000
2008	7,9	9,4	13,2	16,1	20,2	26,6	31,4	26,1	29,9	26,0	0,000
2009	6,5	10,6	12,1	16,4	22,2	28,9	28,0	39,5	25,8	26,7	0,000
2010	6,6	14,1	16,4	16,6	31,4	25,2	40,7	27,0	38,9	27,6	0,000

Source: author's elaboration based on EPH, INDEC.

Implementation and Results of Maximum Likelihood Estimation - Effects on Social Indicators-

Imputation effects on mean and median of labour incomes. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

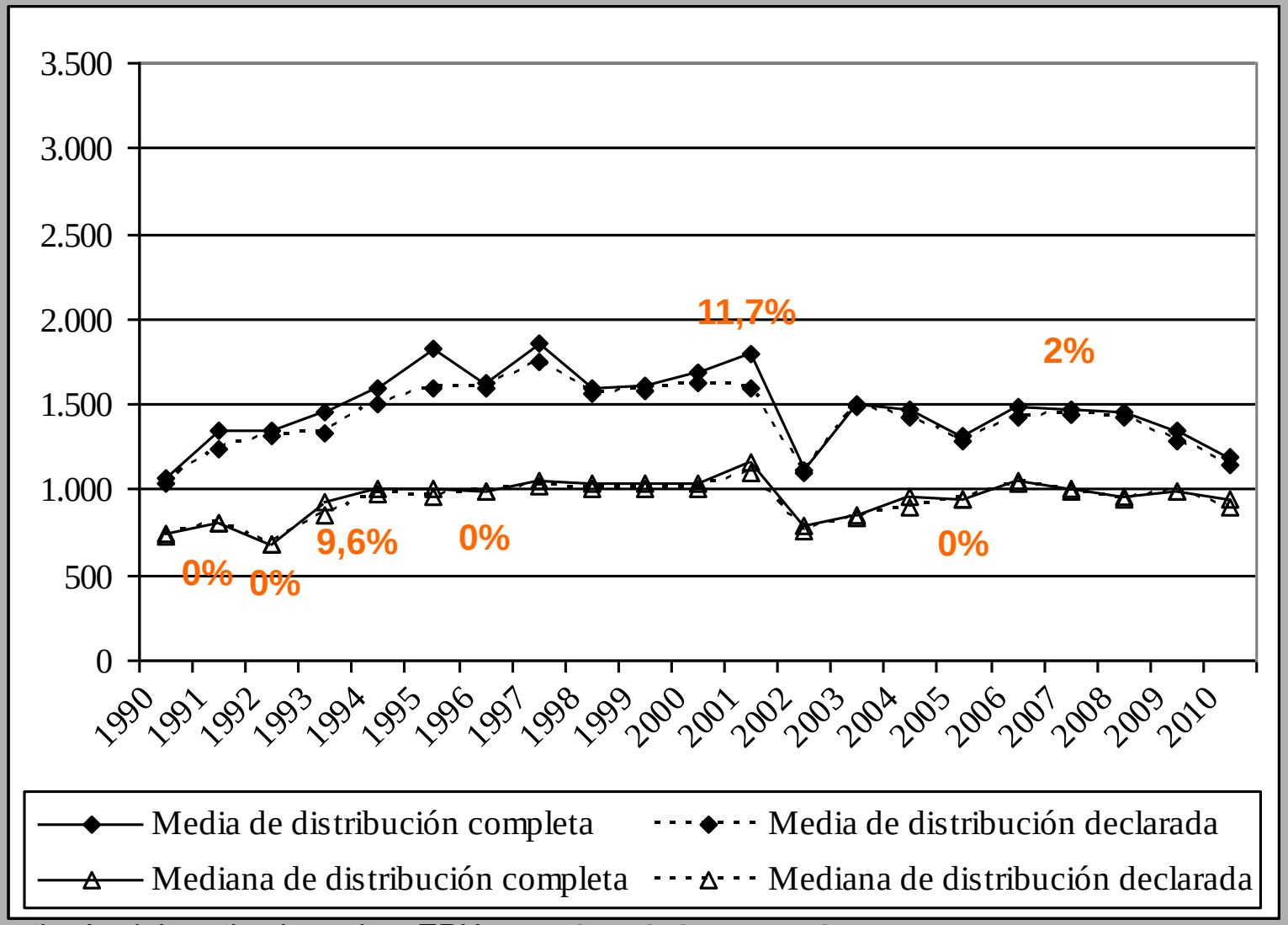
In pesos (2010).



Source: author's elaboration based on EPH-INDEC, IPC-CENDA y CENDA..

Imputation effects on mean and median of non-labour incomes. Gran Buenos Aires: 1990-2010. October 1990-2002 and 2nd Semester 2003-2010

In pesos (2010).



Source: author's elaboration based on EPH-INDEC, IPC-CENDA y CENDA..

Concluding Remarks

- **Incidence of missing data does not have a static pattern during the period under analysis.**
- **Missing data do not present a completely randomized distribution. In contrast, it is more common among highly-educated recipients, employers, independent workers, non-wage recipients or among those with mixed incomes sources.**
- **Working only with sub-samples that have no missing data would produce biased estimations and will not represent the population from which the sample was taken.**

- **It is necessary to make data imputation, in order to have a representative sample of the population under analysis.**
- **Changes in missing-data patterns are associated to methodological and technical factors, socio-economic context, and several social factors that determine some groups' propension to answer. Some institutional factors (such as credibility of the public institution that make the survey) also impact on this.**
- **Impact of missing data imputation for living conditions analysis is relevant: the mean of income per recipient increases. This is the most important effect in labour incomes.**

Based on this…

- **A new method was identified. It is based on Maximum Likelihood Estimation, and it is useful for missing data problems and to obtain new and reliable income statistics.**
- **This method allow us to make analysis based on the complete sample.**
- **We should remark the importance of making valid and realiable imputations of missing data for this kind of variables.**
- **It is recommended that institutions that are in charge of data production make efforts in order to diminish the percentage of missing data in income variables.**

Thanks for your attention

Imputations for Missing Data in Income Variables.
Permanent Household Survey (EPH).
Gran Buenos Aires, Argentina / 1990-2010

Eduardo Donza