INNOVATIONS IN THE METHODOLOGICAL DESIGN OF THE RECURRING INAPP PLUS SAMPLE SURVEY

Francesca Bergamante, Gianni Corsetti,
Francesca della Ratta-Rinaldi, Andrea Spizzichino

Abstract. The PLUS (Participation, Labor, Unemployment, Survey) is a recurring national sample survey conducted by INAPP on a sample of 45,000 individuals aged 18 to 74 years. This paper aims to analyse the methodological innovations introduced in the 2024 version of the survey (the latest), highlighting the reasons behind these changes, aiming at making the estimates more reliable and addressing the need for transparency and scientific rigor.
The innovations concern various aspects of the survey. First of all, the information about the subjective working condition coming from Istat LFS has been considered in order to redefine the key aggregates and the sample strategy. Secondly the final weights calibration has been changed, inserting between the constraints even the occupation (1° digit). Furthermore, has been improved the control and the correction processes for most of the relevant variables.

1. Introduction

This contribution presents the main innovations introduced in the most recent wave of the INAPP-Plus survey, conducted in 2024. These innovations concerned various aspects of the survey design, including both the sampling framework and the weighting procedures, as well as certain processes related to data validation and correction. The overarching objective of these enhancements was to produce more robust and reliable estimates, while improving comparability with LFS data, all the while preserving the core structure of the survey and ensuring the continuity of historical time series.

The paper is structured as follows. The first section provides an overview of the INAPP-Plus survey. The second section presents sampling design characteristics and the improvements inserted in the calibration system. The third section focuses on the improvements implemented in data control and correction procedures. The contribution concludes with final reflections.

2. The recurring INAPP Plus sample survey

PLUS - Participation, Labor, Unemployment, Survey is a biennial recurring national sample survey conducted by INAPP Labour Market Unit[1]. The first edition of the survey was conducted in 2005, and since 2006 it has been included in the National Statistical Programme (PSN).

The survey, which is representative of the entire Italian population is conducted on a sample of 45,000 individuals aged 18 to 74 years using the CATI (Computer-Assisted Telephone Interview) technique. The planning of interviews is based on quota sampling, stratified with the definition of partially overlapping study domains. The eleventh edition of the survey was conducted in 2024. The next edition of the survey (the twelfth) is scheduled to be conducted in 2026.

The survey adopts a longitudinal design, employing a non-rotating panel structure consistent with the classical longitudinal framework. This design facilitates dynamic analyses across various individual states—not limited to employment—thus enabling the reconstruction and study of detailed individual labour market trajectories.

The main objective is to analyse specific aspects of the labour market, such as young people's entry into the labour market, the extension in length of employment of the population in older age groups, participation of the female population, and job search patterns. The survey also aims to detect the labour characteristics of employed people and changes in employment status and needs as a result of the health emergency.

The questionnaire is extensive and consists of several modules, some of which are specifically dedicated to different types of employment contracts or different employment status. The questionnaire has a fixed core, but thanks to its flexible design, it is possible to include new questions or modules aimed at investigating novel or emerging aspects deemed important for producing estimates and analyses.

Each wave of the survey produces a dataset along with a methodological note, which supports the use of the dataset and provides insight into the more technical aspects of the survey design. The databases generated through the survey are freely disseminated and are widely used by the research community and academia. Plus data are extensively used within Inapp to produce institutional documents (such as Inapp Policy Briefs[2]), support parliamentary hearings, contribute to the Inapp annual

[1] https://www.inapp.gov.it/en/surveys/periodic-surveys/participation-labour-unemployment-survey-plus

[2] https://oa.inapp.gov.it/server/api/core/bitstreams/a5716ef6-5aa6-4fea-ba09-82c04d948175/content
https://oa.inapp.gov.it/server/api/core/bitstreams/aa36011a-f9be-4888-a5bf-1de340a3646a/content
https://oa.inapp.gov.it/server/api/core/bitstreams/803134bf-5b3a-4690-a38d-d2a6a4e96c3e/content
https://oa.inapp.gov.it/server/api/core/bitstreams/0aa54fab-f9f7-4826-a31f-dcbcb85d68f1/content

Report[3], and develop scientific publications as well as presentations for conferences and seminars. Furthermore, in 2022[4] and 2024[5], two specific reports were produced, providing estimates and analyses on a wide range of indicators and issues (Inapp, Bergamante, Mandrone 2022; Inapp, Bergamante, Luppi 2024).

3. The INAPP PLUS sample design

The schedule of interviews to be carried out was made on the basis of a stratified quota sampling with definition of partially overlapping domains of study (Corsetti 2002 and 2024). The choice of quota sampling was motivated by the need to greatly reduce the sample size needed to produce statistically significant estimates for small subpopulations of interest (Mandrone 2012). As an alternative we could have chosen to use strategies of classical two-stage sampling (e.g., municipalities and households) but, in addition to requiring fieldwork much more costly, would involve the surrender of a survey of the main features PLUS, one of no proxy respondents.

The sample is divided into ten key target (domain of interest):

1. Young workers, aged between 18 and 29 years;
2. Young students, aged between 18 and 29 years;
3. Young unemployed (student or inactive women), aged between 18 and 29 years;
4. Active women, aged between 18 and 39 years;
5. Inactive women, aged between 18 and 39 years;
6. Active Seniors, aged between 50 and 74 years;
7. Inactive Seniors (retired from work), aged between 50 and 74 years;
8. Unemployed aged between 15 and 74 years;
9. Employed aged between 15 and 74 years;
10. Inactive aged between 15 and 74 years.

In order to provide reliable estimates for the subgroup of these 10 domains (for example, limited to each of the Italian geographical areas) we planned a stratified

https://oa.inapp.gov.it/server/api/core/bitstreams/4324a2a6-ba16-41d1-9ead-3aa3e582775f/content
https://oa.inapp.gov.it/server/api/core/bitstreams/505215c7-aa1d-4fcc-8161-ba3be3b50ea4/content

[3] https://www.inapp.gov.it/en/publications/inapp-report/published-editions

[4] https://oa.inapp.gov.it/server/api/core/bitstreams/8e277563-6453-4460-a5a3-e1a79b839e43/content

[5] https://inapp.infoteca.it/ricerca/dettaglio/rapporto-plus-osservare-le-traiettorie-del-mercato-del-lavoro/25380#

sampling, where the layers (sex, age class, employment status, type of municipality, geographical areas) constitute a partition of the sample and (for subsets of layers) of the same domains of study.

The number of interviews to be obtained for each of the layers was determined to provide reliable estimates for the entire reference population and for particular subsets of interest, through the implementation of a procedure for multi-domain allocation, based on resolution a constrained minimization problem. More precisely, a priori minimum variance targets were set for the domains of interest listed above and for their territorial breakdowns by geographical areas and type of municipality (urban, non-metropolitan). The allocation procedure uses LFS estimates as its data source.

The next step, is related to the choice of weighting estimator constrained to be adopted for the calculation of the coefficient to the universe. In particular, it has resorted to the implementation of methods based on the use of the estimator of generalized regression (GREG estimator) (Valliant 2000 e Särndal 2005). It ensures that the estimates of the absolute frequencies of the auxiliary variables used as regressors are coincident with the known observed (LFS 2024) in the total population and imposed as calibration constraints. First, this implies that the demographic composition of the actual population and employment reference is, by construction, reproduced in the analysis.

Furthermore, it allows to correct distortions caused by factors not fully controlled in the design phase of the investigation and related instead to the detection phase, as the self-selection of the sample due to the higher average propensity to answering certain categories of persons.

3.1. Innovation in sample design

Despite the Labour Force Survey (LFS) serving as a fundamental benchmark, the INAPP survey design presents several differences in the definitions used to identify key labour market aggregates—namely, employed, unemployed, and inactive—partly due to the chosen survey technique. While the LFS employs a funnel-based approach, identifying employment status through objective and harmonized criteria (as established at the European level), the PLUS survey, conducted via CATI and based on quota sampling, necessarily begins with a filter question that captures the respondent's self-reported status (subjective definition).

This divergence in definitional criteria was not previously accounted for in the stratification and weighting system, which relied on LFS estimates based on objective definitions of employment and unemployment as calibration constraints. However, since both surveys—albeit through different question sequences—collect

information on both objective and subjective definitions, it is possible to analyse which criterion is more appropriate for determining the calibration constraints to be used in the survey.

In the 2024 revision, a differentiated approach was adopted for employment and unemployment. For employment, the discrepancy between objective and subjective definition is minimal[6], with a 99.4% concordance in the sample data, closely aligning with the 99.2% reported by LFS for the 2024 average. This high level of agreement supported us to use (as in the past) LFS objective estimate as calibration constraints for employment, also due to the need for additional employment-related information (e.g., working hours, job characteristics, occupation), which would not be available for the small subset of subjectively employed individuals in the LFS.

The situation is different for job seekers, for whom the difference between the two definitions is larger. According to the objective definition, to be considered unemployed, people must have engaged in at least one job-search activity in the four weeks preceding the interview and be available to start work within the following two weeks. The subjective definition, on the other hand, refers to an individual's perception of their own situation. This is not necessarily linked to job-search activities or immediate availability to start work and is often associated with inactivity or discouragement. ISTAT has documented the difference between the objective and subjective conditions of job seekers since 2007 (Istat 2007). Based on these considerations, it was deemed more appropriate to use the subjective response from the LFS questionnaire (QJ01) as the calibration constraints for estimating job seekers. This approach also allows for a more nuanced classification of individuals as employed, unemployed, students, retirees, or other inactive persons.

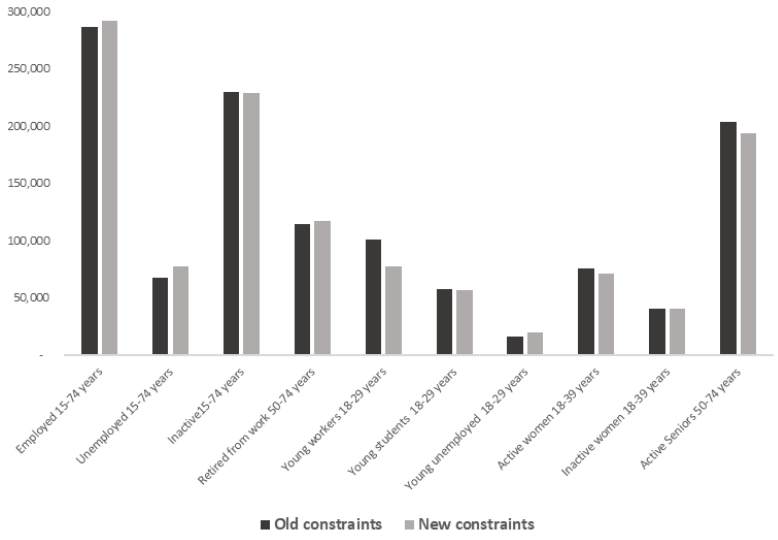
A second innovation in the calibration design concerns the auxiliary variables used for weighting. In addition to working hours and employment, the first-digit occupational classification (the nine major occupational groups) was introduced to enhance consistency between the estimates of the two surveys, particularly for variables closely linked to occupation.

To reach this decision, several tests were conducted using 2022 data, allowing for a comparison between the previous weighting system and various combinations incorporating additional variables selected from those available and highly correlating with key study variables (e.g., occupation, economic activity classification, or a mix of both). The selected solution was the one that minimized overall discrepancies with LFS estimates for certain employment-related variables (not included in the weighting system).

[6] Since 2022 the PLUS survey has inserted the indicators useful to detect employed also using the objective definition, according to the new definition inserted in 2021 in the European LFS as explained here: https://www.istat.it/notizia/forze-di-lavoro-2021-come-cambia-la-rilevazione/.

An analysis of the effects of the new calibration on the reliability of 2022 estimates shows that the systems are closely aligned and that the introduced changes did not increase estimation error—in some cases, they even led to improvements (Figure 1).

Figure 1 - *Estimated value (a.v.) above which 10 percent is exceeded in terms of Coefficient of Variation by domains of interest and type of constraints (Plus 2022).*



Source: INAPP Plus 2022

Further analysis on 2024 data should assess the validity of the choice made, with the aim of further enhancing the comparability and accuracy of labour market indicators derived from the PLUS survey.

4. Innovations in data control and correction

Data control and correction has historically represented a fundamental step in the production process of the survey and has always been characterized by the introduction of innovations that have made it possible to implement and enhance the changes made to the questionnaire.

The innovations introduced allow for a better interpretation of the changes taking place in society and, more specifically, in the labour market. In the latest edition of the survey, specific changes were made to the production process of income variables to account for developments that occurred over the past two years.

The revisions to the procedures for defining income involved several stages of the process. The main changes were related to the treatment of negative income for the self-employed, the adjustment of tax rates used to define gross income for employees, and the use of a variable indicating the number of months worked in 2024 to define a gross income weighted by the actual months of employment.

The control and correction interventions for negative income among the self-employed included the imputation of income for all individuals who declared zero income and reported a loss during the year. The imputation methodology used was the hot deck technique, already applied to other quantitative variables.

Regarding the grossing-up of employee income, it became necessary to apply the new IRPEF income tax rates introduced in 2024 to calculate total annual gross income. Finally, concerning the number of months worked in the last year, this edition of the survey reintroduced the question on the number of months worked in the current job, if it began during the current year. This allowed for the calculation of gross income based only on the months actually worked.

Table 1 presents some results on income, comparing the reported and imputed data from the last two editions of the survey. It can be observed that the mean and median income for the self-employed, employees, and coordinated and continuous collaborators are substantially the same between the observed and imputed data.

Table 1 – *Statistical indicators on observed and imputed income (years 2022 and 2024).*

CoCoCo					
Year	Inputed	Mean	Median	CV	Kurt
2022	0	1354	1100	59	2
2022	1	1447	1250	69	3
2024	0	1661	1480	52	4
2024	1	1688	1300	58	5
Employees					
Year	Inputed	Mean	Median	CV	Kurt
2022	0	1573	1500	61	236
2022	1	1503	1400	49	155
2024	0	1597	1500	40	499
2024	1	1573	1500	37	170
Self-employed					
Year	Inputed	Mean	Median	CV	Kurt
2022	0	32685	25000	119	50
2022	1	33603	25000	118	47
2024	0	39427	30000	124	67
2024	1	39454	30000	113	55

Source: INAPP Plus2022 e INAPP Plus 2024.

The comparison between the income data for 2022 and 2024, both for the observed and imputed values, shows an increase in the mean income, a decrease in the coefficient of variation (CV), whit the exception of self-employed respondents, and an increase in kurtosis. These changes suggest that the income distribution has shifted toward higher values, with less dispersion around the new mean. In other words, incomes have become more concentrated around the central value, but with a higher frequency of extreme values in the tails of the distribution, both at the high end (very high incomes) and at the low end (very low incomes).

The increase in kurtosis indicates that, compared to 2022, there are more outliers in the income data, reflecting a greater occurrence of exceptional economic events (such as extremely high or low incomes). However, the quality of these extreme values is ensured by a rigorous outlier evaluation system that precedes the imputation process. This system ensures that extreme income values are properly handled before being included in the dataset, preventing them from distorting the analysis and preserving the integrity of the data.

5. Conclusions

This paper has illustrated a series of significant innovations introduced into the methodological framework of the 2024 edition of the Inapp-Plus survey as the result of an articulated and rigorous redesign process. However, the analyses of the impacts of the new calibration on the 2024 edition of the survey have not yet been completed, and a definitive assessment will be possible once all tests have been finalized.

Given the longitudinal nature of the survey, a guiding rationale behind these changes was to achieve an appropriate balance between methodological innovation and the preservation of consistency and comparability across historical data series.

The modifications were intended to enhance the internal consistency and overall reliability of the collected data. Subsequent analysis of the final dataset will enable a comprehensive assessment of the impact and implications of these innovations.

In recent years, the increasing difficulties encountered in conducting Computer-Assisted Telephone Interviewing (CATI) surveys—a trend not limited to the Plus survey—have raised critical methodological questions. In particular, the declining response rates and coverage issues (especially for students, employed men in the 30-39 age group and employed women in the 40-49 age group) associated with CATI have prompted reflection on the potential transition to Computer-Assisted Web Interviewing (CAWI) techniques, at least for the selected variables.

At the moment, we correct distortions caused by this factor by implementing a sampling strategy based on the use of the estimator of generalized regression (GREG estimator). It ensures that the estimates of the absolute frequencies of the auxiliary

variables used as regressors are coincident with the known observed in the total population and imposed as calibration constraints.

Looking ahead to future survey cycles, consideration is being given to the use of Municipal Population Registers (Liste Anagrafiche Comunali, LAC) as the primary sampling frame and consequent transition to a probabilistic sample. This shift would offer multiple advantages:

- reduce respondent burden by eliminating the need to collect information already available from administrative sources;
- mitigate the biases introduced by non-random sampling from telephone directories that cause a self-selection of the respondents due to the higher average propensity to answering certain categories of persons.

However, the implementation of this approach would necessitate the adoption of mixed-mode data collection strategies, with implications for both design complexity and data comparability.

Overall, the methodological evolution of the Inapp-Plus survey reflects a broader transformation in the field of social survey research, wherein the integration of administrative data and the adoption of flexible, multimodal strategies are becoming increasingly central to ensuring both efficiency and data quality.

References

CORSETTI G. 2024. Metodologia dell'Indagine campionaria longitudinale PLUS. In Inapp, Bergamante F., Luppi M. (a cura di). *Rapporto Plus 2023. Osservare le traiettorie del mercato del lavoro*. Roma, Inapp.

CORSETTI G. 2022. Metodologia dell'Indagine campionaria Inapp-PLUS. In Inapp, Bergamante F., Mandrone E. (a cura di). *Rapporto Plus 2022. Comprendere la complessità del lavoro*. Roma, Inapp.

DEVILLE J.C., SÄRNDAL C.E. 1992. Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, n.87, pp.376-382.

INAPP, BERGAMANTE F., MANDRONE E. (a cura di). *Rapporto Plus 2022. Comprendere la complessità del lavoro*. Roma, Inapp.

INAPP, BERGAMANTE F., LUPPI M. (eds). 2024. *Rapporto PLUS 2023. Osservare le traiettorie del mercato del lavoro.* Roma: Inapp.

ISFOL, MANDRONE E. 2012. Labour Economics: PLUS empirical studies. *Temi&Ricerche*, n.3, Roma, Isfol.

ISTAT. 2012. Disoccupati, inattivi, sottoccupati. Indicatori complementari al tasso di disoccupazione. *Statistiche report 19 aprile 2012. (*https://www.istat.it/comunicato-stampa/disoccupati-inattivi-sottoccupati-anno-2011*/)*

SÄRNDAL C.E., LUNDSTRÖM S. 2005. *Estimation in Surveys with Nonresponse*. Chichester (UK), John Wiley & Sons.
VALLIANT R., DORFMAN A., ROYALL R.M. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York, John Wiley.

Francesca BERGAMANTE, Inapp, f.bergamante@inapp.gov.it
Gianni CORSETTI, Istat, giacorsetti@istat.it
Francesca DELLA RATTA-RINALDI, Inapp, f.dellaratta@inapp.gov.it
Andrea SPIZZICHINO, Istat, spizzich@istat.it