Non-parametric bounds on irregular workers' share[*]

Antonio Dalla Zuanna[†] Domenico Depalo[‡] Edoardo Santoni[§]

March 25, 2025

Abstract

This paper introduces a novel approach to estimate the prevalence of undeclared employment. Our method leverages partial identification, comparing aggregate data from the Labor Force Survey with administrative data to derive bounds on the proportion of undeclared workers. By imposing limited assumptions, we produce credible estimates with narrow bounds, using publicly available and timely data. Applying this method to Italian data, we find that undeclared employment remained possibly stable at 10–12% until 2020 but declined to 8–10% in the post-pandemic period. Additionally, our results show higher levels of undeclared employment in the southern regions but with a declining trend. This approach offers an alternative tool for monitoring informal employment with minimal assumptions and a shorter lag compared to official estimates.

JEL classification:
Keywords:

[*] We would like to thank ... Replication files and additional results will be available at the webpage: http://sites.google.com/site/domdepalo/. The views expressed in this paper are those of the author and do not imply any responsibility of the Bank of Italy. Corresponding address:

[†] Bank of Italy

[‡] Bank of Italy

[§] Università di Ferrara

1 Introduction

Informal employment is present in all labor markets. Although it is typically regarded as problematic due to issues such as the lack of social protection for workers and the increased fiscal burden on the formal sector (Packard et al., 2012), irregular contracts can also provide employment opportunities for low-skilled workers who might otherwise remain outside the labor force, or act as a buffer during economic downturns (Boeri and Garibaldi, 2005; Fiess et al., 2007). Its prevalence within an economy can then have a significant impact on poverty and inequality measures (Nakamura, 2013; Pham, 2022).

There are thus several reasons why obtaining a reliable and timely measure of informal (or undeclared) employment is of interest. However, recovering such a measure proves to be extremely challenging. An undeclared worker is defined as an individual who receives a salary for a job that, in violation of the law, is not reported to the authorities, meaning no taxes or social security contributions are paid on that salary. As a result, undeclared workers do not appear in official registries.[1] While this lack of formal recognition does not preclude practitioners from estimating the share of workers hired informally, exisiting estimates heavily depend on the methods and data sources employed, often more so than estimates of observable phenomena. In this paper, we introduce a novel approach for estimating the prevalence of undeclared employment, grounded in partial identification (Manski, 1990). Unlike previous methods, our approach requires fewer assumptions and utilizes aggregate data at the country level, which are generally more accessible and readily available.

We propose to quantify the prevalence of undeclared workers based on discrepancies between different statistical sources. We compare aggregate survey data from the Labor Force Survey (LFS), which measures total employment (i.e., the sum of regular and irregular workers), with administrative data from the social security institute, which records workers for whom social security contributions are paid, thus reflecting regular employment.[2] The difference between the number of overall workers and the number of regular workers corresponds to the number of irregular workers.

[1] According to the EU definition, undeclared workers are those performing "any paid activities that are lawful in nature but not declared to public authorities, considering differences in the regulatory systems of Member States."

[2] The LFS asks individuals about their working condition regardless of the contractual agreement.

However, the comparison of the numbers from these two sources is complicated by the fact that (i) there are time discrepancies in the way individual data are generally aggregated (i.e. the yearly average number of workers in the LFS may not correspond to any of the aggregates provided by the social security insitute) and (ii) the two data sources partly focus on different populations of interest and may incur in some error while retrieving the data, such as misreporting or measurement errors. In other words, we need to recover estimates of the yearly average of overall and regular workers which are credible and comparable. Using partial identification we can impose limited assumptions to solve these issues and obtain bounds—i.e., a range of admissible values for the share of irregular workers. This method enables us to progressively impose a larger or more stringent set of assumptions to derive narrower bounds. Since the contribution of each assumption is transparent, and we begin with very mild assumptions, our bounds are "credible", in the terminology of Manski (2011). Furthermore, we rely on aggregate data that are publicly available, regularly updated, and accessible in nearly every country, ensuring the timeliness and broad applicability of our approach.

We apply our method to Italian data, focusing specifically on employees. The average number of workers employed in one year is estimated by the National Statistical Institute (Istat) exploiting data from the Italian LFS. The Social Security Institute (Inps), provides aggregates such as the total number of workers with at least one day of social security contributions, the total weeks worked, and the distribution of employment contracts (e.g., full-time, part-time, permanent, temporary) which we exploit to estimate the number of regular workers. Although the average provided by Istat can in principle be biased by misreporting or measurement error, in a first step we assume it is correct and we focus on estimating the corresponding number of regular workers from Inps sources. We relax this assumption at a later point.

We begin by noting that the average number of regular workers in the year cannot be smaller than the total number of weeks worked in a year divided by the total number of weeks in a year (52). This quantity, when subtracted from the average number of workers from the LFS data, provides an upper bound for the average number of irregular workers. On the other hand, the yearly average of regular workers cannot be larger than the overall number of workers with at least one day of regular employment. With this information we can compute the lower bound for the yearly average of irregular workers. While these bounds are credible, they may be judged too wide, especially if the aim is to compare the evolution of the phenomenon over time and space. To refine them, in computing the lower bound we adjust for workers with non-full-time, non-permanent contracts by applying a rescaling coefficient to account for their lower weight in estimating the yearly average.[3] This narrows the bounds and reveals that undeclared employment in Italy remained possibly stable at 10–12% until 2020, but declined to 8–10% post-pandemic. Applying this method to different regions, we find higher undeclared employment in the south, consistent with other studies, but also a declining trend in all the areas.

The approach of comparing multiple datasets to estimate the share of undeclared workers has been employed in several countries to obtain official estimates of undeclared work (EEPO, 2017). However, these methods typically require access to individuals' non-anonymized identities, as they involve matching observations from one dataset to another (see e.g. De Gregorio and Giordano, 2015). To comply with legal and privacy protections, some countries have formal agreements between institutions to enable such data linkage. For example, this is the strategy used to compute the official numbers for irregular employment in Italy. However, this highly formalized process is not reproducible without access to sensitive data and usually requires significant time to align the two datasets at the individual level. Furthermore, statistical procedures must be implemented to ensure that mismatches are attributed to undeclared work rather than other factors (e.g., discrepancies in the timing of data collection), thus imposing parametric restrictions in modelling the mismatch probability.

Other methods proposed in the past rely on assumptions about the determinants of irregular employment or on specific types of data which may not be available or comparable across countries. Such methods vary significantly, reflecting differences in approaches (macro vs. micro), econometric techniques, and levels of analysis. Macro-level approaches typically use aggregate data, often at the country level, with the Multiple Indicators Multiple Causes (MIMIC) model being a common statistical method. The MIMIC model leverages cross-country variability to estimate the relationship between potential causes and indicators that proxy for the unobserved share of undeclared work-

[3] As explained in Section 3.1, this lower weights are reflected in the average computed by Istat because these workers have a lower probability of being employed when sampled during the LFS survey.

ers, allowing for measurement errors in capturing the relationship between proxies and the target share. Buehn and Schneider (2012) applies a MIMIC model to estimate undeclared employment in 162 countries. While the approach is useful for cross-country comparisons, offering a homogeneous methodology with harmonized data, MIMIC has important limitations including the model specification, namely what is a determinant and what is an indicator of irregular employment (see also Schneider and Buehn, 2017).[4]

Micro-level approaches focus on individual-level data and can be classified into two main categories: direct inference and deductive approaches. The direct approaches rely on surveys asking explicitly about irregular labor market contracts. With this information available, one can directly compute the proportion of undeclared workers based on survey's responses. For example, Cappariello and Zizza (2010) uses the Survey on Household Income and Wealth (SHIW) in Italy, which includes a question about the payment of social security contributions.[5] While this method can be powerful, it is subject to potential measurement errors, such as recall bias, which could distort the results. Additionally, such surveys may not be available in all countries, and, even where they exist, their comparability can be challenging, particularly since respondents might interpret questions differently based on local cultural perceptions of the underground economy (Schneider and Buehn, 2017). Deductive approaches compare expected and realized production, attributing the difference to the presence of irregular workers. For instance, Tirozzi (2022) calculates excessive production using predictions from a rich model specification of the production function based on the LASSO method. Gries et al. (2022) use a similar strategy to estimate the share of irregular employees in vineyards, exploiting an exogenous shock from the Arab Spring wave on southern Italian coasts. Deductive approaches could lead to misleading conclusions if, for example, the production function model is mis-specified, particularly if firm-specific effects are inadequately considered. Exploiting an exogenous shock delivers internally valid estimates that may be hardly generalized to the rest of the economy or to other periods.

[4] For example, "'tax immorality' should be included into the model as a determining variable. From a theoretical point of view it could be argued equally well that tax immorality is an indicator of the shadow economy or even a consequence. The same can be stated for other variables. A short work week, for instance, can be regarded as a determinant of the shadow economy instead of an indicator." (Helberger and Knepel, 1988, p.970)

[5] The question is phrased as: "Considering the lifetime work experience of (name), did he/she ever pay, or did his/her employer pay, pension contributions, even for a short period (and even if long ago)?"

It is important to clarify that the procedure we suggest does not aim in any way to replace existing official numbers. Point identification is likely necessary for policy reasons, and to reach it additional assumptions must necessarily be imposed. Our method, however, can both serve the purpose of providing an estimate of undeclared workers with as little as one year delay (as compared to the two years delay required by the official method in the Italian case) and of providing some bounds where the official number must credibly fall in. A number which significantly deviates from the bounds we propose would signal that the assumptions imposed to retrieve it are likely not valid, thus making the estimate not credible.

The paper is structured as follows. In Section 2 we specify the measure we aim at estimating and we give details of the official procedure used and its main drawbacks. Section 3 describes our partial identification strategy. In Section 4 we list the publicly available data sources we use to estimate the undeclared workers' ratio for Italy also in order to ease the process of replication for other countries. Section 5 reports the estimated bounds for the period 2014-2023, displaying the results when we impose different assumptions and also separating between different geographical areas. We also compare the estimated bounds to the official estimates. Section 6 concludes.

2 Definition of Undeclared Workers' Ratio

In this work, we focus only on undeclared employees, thus ignoring the self-employed. Self-employed likely constitute a relevant part of undergroud economy, but this often comes in the form of unreported taxable income, while our interest is in the proportion of workers who are employed but do not appear in any official registry and thus are excluded from the social security system. The share of undeclared workers among the employed individuals in year t is defined as:

$$\text{Share}_t = \frac{\text{Undeclared workers}_t}{\text{Employed individuals}_t}, \tag{1}$$

where we think of these quantities as averages throughout the year, i.e. the ratio between the average number of workers who are undeclared and those who are employed at any point in time during the year. This fraction measures the prominence of the undeclared work phenomenon in terms of headcount workers and estimating these averages is the aim of this paper. Alternative measures can take into account the intensive margin of labor supply and count the full-time equivalent share of undeclared individuals, or the share of positions which are in fact not declared. The latter two definitions would allow to measure the relevance of the so-called "grey area" of workers who work irregularly only part of their working day or for only some jobs but not others. While our methodology can be adapted to capture the grey area, in what follows we emphasize the validity of the partial identification approach in capturing the dynamics of undeclared work. Because the trends in undeclared work in Italy are similar when one focuses on either of the three indicators, our chioce does not significantly affect the conclusions we draw on the phenomenon.

$Share_t$ by definition is not observed in any official register. As a consequence, we cannot easily benchmark the estimates using the method we propose to the true number. Throughout the discussion of our results in Section 5 we compare our estimates to the official numbers estimated by Istat. In the rest of this section we highlight where the official numbers may fall short, to clarify how our method can be a helpful addition to these existing estimates.

Istat recovers the annual estimates of $Share_t$ by merging LFS with the individual level data from Inps which has information on all the job position for which social security contribution is paid. The former data are used to compute the denominator of Equation 1, the latter to isolate the number of declared workers and then compute the numerator. In practice, workers who declare themselves as employed in the LFS and do not appear in Inps are potentially considered as undeclared workers. Different adjustments are then applied to make sure that workers who are not directly matched between the LFS and the Inps data are indeed undeclared workers, by parametrically modelling the probability of wrongly classifying workers which are not matched between the data sources. While the procedure is not explained in all its details in official publications, Appendix A describes it exploiting all the available information, also providing an overview of the methods used in other countries. In addition, revisions of the methodology happen on a regular basis to make sure that the procedure captures mismatched workers correctly. In September 2024, for example, the official number has been substantially revised downward by, on average, 1.5 percentage points per year for the period 2014-2022, with the average rate dropping from 12.2% to 10.7%.[6] This is a substantial change which highlights the fact that undeclared employment has likely been overestimated in the official statistics in the past.

We note two main drawbacks of the official procedure. First, some of the information required for the adjustments when matching survey and administrative data are available with some lag, which implies that Istat's estimates appear with 2 years lag. In addition, as mentioned, changes in the procedure may lead to significant revisions. Second, the adjustments described above rely on assumptions about the distribution of undeclared workers across sectors and about the determinants of the phenomenon. For example, consider the possibility that a worker who is in the official registries may not be an actual regular worker. This may be due to misreporting in terms of starting/end date of the position, or to the fact that the individual was employed in some dates but these dates did not correspond to the date when the LFS interview happened. To solve this issue Istat assumes that the probability of mis-reporting a regular work depends on some observable characteristics of the individual and estimates a logit model to infer this probability (Istat, 2015). While this is a legitimate approach, model mis-specification might return biased predictions. Likewise, distributional assumptions might not be appropriate, especially in the tails. The partial identification method we propose does not need to impose any assumption about why the misreporting happened but allows to take it into account by computing some bounds which explicitly measure what could be the range of the admissible values of the share.

3 Empirical Strategy

All the existing methods recover point estimates of the share of undeclared workers in Equation 1 by imposing assumptions, such as functional form restrictions, and derive conclusions that are coherent with them. This approach is valid if the restrictions hold true; however, if they are violated, the results become unreliable (see the "law of decreasing credibility" Manski, 2011). To address this limitation, we employ partial identification, an innovative tool for this literature.

[6] A major point of the revision has been a change in the way workers employed for only few hours in the domestic sectors are considered (Istat, 2021c).

Partial identification combines assumptions and data to generate a set of admissible values, or bounds (Manski, 1990). The process begins with the least restrictive assumptions and incrementally adds more structure; the stronger the set of assumptions, the narrower the bounds. By introducing each assumption separately, readers can evaluate the strength of each restriction and determine their level of agreement with it. This section begins with minimally restrictive assumptions and incorporates additional restrictions to reduce the bounds' width.[7]

3.1 Least Demanding Assumptions

We utilize only publicly available data. Provided that there is at least one regular employee in the Inps registries, it follows that "Undeclared workers" < Employed individuals" and $Share < 1$ (from now on, time subscript t omitted for simplicity). By definition

$$\text{Undeclared workers} \quad = \quad \text{Employees} - \text{Declared workers}. \tag{2}$$

We obtain the average number of employees who worked in year t, encompassing both declared and undeclared workers, from the LFS (Section 2). Despite potential inaccuracies in the number of employed individuals, we assume this value is observed without error because, according to Eurostat (2022), the reliability of this indicator is high within our age range, with a 95% confidence interval smaller than 0.5% of the target value.[8] In Section 3.3, we introduce a simple method to relax the assumption that the number of employed individuals is observed without error. LFS methodology is such that families are surveyed continuously throughout the year and they are then averaged exploiting weights which allow to reconduct the surveyed population to the total. A worker who is employed only for a short period and is interviewed while employed is thus considered as an employed person during the year, with a similar person who has a similar employment spell but is

[7] The least restrictive bounds assume either that no employed individuals are irregular, i.e., "Undeclared workers"=0, or that all are irregular, i.e., "Undeclared workers"=Employed individuals". These extreme values match the numerator's domain. Substituting these into Equation 1, we obtain the so called "domain bounds", such that $Share \in [0, 1]$. These bounds are universally valid and can be calculated without observations, but they offer minimal information. For a similar property, see Manski (1990).

[8] Debates about this source's reliability often focus on distinguishing between unemployed and inactive populations (Battistin et al., 2007; Brandolini et al., 2006), while our analysis solely concerns individuals who report employment, for whom the definition is less controversial.

interviewed when non-employed is considered as non-employed. In this way, the average provided by Istat implicitely assigns a lower weight to individuals who are only employed for short periods.

To derive the upper bound of the proportion of undeclared workers, we need to identify the *smallest* possible number for the average number of regular workers in the year, consistent with the observed data.[9] As we observe the total number of weeks worked regularly (Section 2), this number cannot fall below

$$\frac{\text{Total weeks}}{\text{Maximum weeks per worker}}.$$

With a maximum of 52 weeks per year, this equates to

$$\frac{\text{Total weeks}}{52}. \tag{3}$$

For the lower bound of undeclared workers, we need the *largest* possible number of regular workers who can contribute to the yearly average. This number corresponds to the total workers recorded as having at least one day of paid regular job in Inps registries ("Observed employees").[10] Combining these pieces of information, the bounds are defined as follows:

$$\text{Share} \in \left[\frac{\text{Employees} - \text{Observed employees}}{\text{Employees}}, \frac{\text{Employees} - \frac{\text{Total weeks}}{52}}{\text{Employees}}\right]. \tag{4}$$

On top of assuming the correctness of the LFS numbers for employees, we implicitly also assume that the number of weeks spent in regular employment corresponds to the number of weeks reported to Inps. Since employers pay social security contributions based on the reported weeks of employment, there is no incentive for upward misreporting, while any downward misreporting would accurately be defined as undeclared employment. Given that 52 weeks represent a full year, under current assumptions, no improvement is possible to estimate a smaller upper bound.

[9] Formally, this follows from $\frac{\partial(\text{Employed}-\text{Declared workers})}{\partial \text{Declared workers}} < 0$.

[10] As explained in Section 4, we use the average over the 12 months of the number of workers with at least one day of paid employment in each month. This is an upper bound for the yearly average because it assigns the same weight to all workers, not considering that the contribution of workers who are only employed for few days in a month in computing such average is smaller.

Similarly, employers have no incentive to report as a regular worker an individual who do not work for them. Therefore, since we only use this headcount for estimating the lower bound, no possible improvements could increase the lower bound under current assumptions, indicating that the bounds are sharp (Molinari, 2020).

3.2 Additional Assumptions to Narrow the Bounds

To narrow the bounds, we need to impose some restrictions. In principle, this goal could be achieved by imposing arbitrary assumptions, e.g. that certain sectors employ a larger share of irregular workers. However, this approach would predetermine the answer; instead, we leverage features of the contracts, which impose credible yet powerful restrictions. Specifically, we aim to adjust the lower bound by calculating a lower number of the yearly average of regular workers (and thus a higher share of undeclared workers) based on information about the type of contracts workers hold.

The starting point is that individuals can be full-year (FY) or part-year (PY) workers, depending on the amount of time they work within the year. Similarly, workers may be full-time (FT) or part-time (PT). Consequently, we need to reweight non-FY/FT workers because, by definition, they worked only part of the unit of time (the year in our case) and therefore their sampling probability in the LFS is correspondingly lower. This approach is consistent with that from Istat (Section 3.1). We compute the average number of workers in the lower bound by rescaling the number of non-FY/FT workers based on the proportion of weeks they worked in the year.

More in details, by construction, the number of weeks worked in the country is equal to:

$$N\,\bar{W}_{Tot.} \quad = \quad N_{FY,FT}\,\bar{W}_{FY,FT} + N_{others}\,\bar{W}_{others} \tag{5}$$

where, for each subsample defined as in the subscript, N represents the number of workers and $\bar{W}$ is the average number of weeks worked in the year. We observe the total number of weeks worked in the year, but not by employment type.[11]

[11] In particular, this is due to the fact that to compute the total number of weeks worked we need to add those weeks during which individuals are in short time scheme (see Section 4), for which the information by contract type

We utilize the equation $N\,\bar{W}_{Tot.} = N_{FY,FT}\,\bar{W}_{FY,FT} + (N - N_{FY,FT})\,\bar{W}_{others} \Rightarrow \widehat{W_{others}} = \frac{N\,\bar{W}_{Tot.} - N_{FY,FT}\,\mu}{(N - N_{FY,FT})}$, where $\widehat{W_{others}}$ indicates that the quantity depends on μ, i.e. the number of weeks worked by full-year/full-time workers. We have information on the total number of weeks worked by regular workers in Italy ($N\,\bar{W}_{Tot.}$) and the total number of workers observed in the year (N). We assume that the number of weeks for full-year/full-time workers is $\mu = 52$, so that we have all the necessary information to derive the rescaled $N_{others}\,\bar{W}_{others}$ and obtain the rescaled average number of workers in the year:[12]

$$\widehat{N} = N\,\frac{\widehat{W_{Tot.}}}{W_{Tot.}} = \frac{N_{FY,FT} \times 52}{52} + \frac{N_{others}\,\widehat{W_{others}}}{52}. \tag{6}$$

If we assinged a number of weeks worked lower than 52 to the FT/FY workers, we would have obtained a lower $W_{FY,FT}$ and larger W_{others} in Equation 6. Because the total number of FY/FT worker is larger than those non-FY/FT this would have resolved in an overall lower $\widehat{N}$. However, in computing the lower bound we want to impose the assumption which gives us the largest possible average number of regular workers.

Putting everything together, the new bounds become

$$\text{Share} \in \left[\frac{\text{Employees} - \widehat{N}}{\text{Employees}}, \frac{\text{Employees} - \frac{\text{Total weeks}}{52}}{\text{Employees}}\right]. \tag{7}$$

Our objective in this paper is to describe what the data can reveal about irregular workers. In doing so, we follow (Manski, 1990, p.444) and restrict ourselves to the issue of identification and do not address the problem of statistical inference.

3.3 Considering Misreporting in the LFS

In this section, we extend our previous analysis by considering the potential for errors in the reported number of employed individuals. Until now, we have assumed that the employment figures provided by Istat, regardless of whether they pertain to regular or irregular workers, are accurate. However,

is not available.

[12] For a similar strategy, see Stoye (2010).

it is possible that these figures contain errors. We propose an extension to our bounds that accounts for this potential source of error.

To begin, it is useful to distinguish between the reported number of employed individuals (y) and the true number (y^*). Under an “invariance assumption” these two values are identical, i.e., $y^* = y$. However, Manski and Pepper (2013, 2018) introduce the assumption of “bounded variation”, under which we can assume that the true number of employed individuals differs from the reported number by a bounded amount δ, such that $y^* = y \pm \delta$.[13]

For illustrative purposes, consider the bounds presented in Section 3.1, where we impose the least demanding assumptions. By incorporating the bounded variation assumption into equation 4, we obtain the following revised bounds:

$$Share \in \left[\frac{\text{Employees} - \delta - \text{Observed employees}}{\text{Employees} - \delta} ; \frac{\text{Employees} + \delta - \text{Total weeks}/52}{\text{Employees} + \delta} \right].$$

The primary challenge in implementing these revised bounds lies in determining an appropriate value for δ. To address this, Manski and Pepper (2013, 2018) suggest using available evidence as a guide to estimate the size of the error. In the empirical application, we follow their suggestion and exploit information on potential misreporting and measurement error from different available sources.

4 Data

As discussed in Section 3, we need to recover the following quantities:

1. Total number of workers in the labor market, defined as the sum of declared and undeclared workers
2. Total number of regular workers
3. Total number of weeks worked by regular workers

[13] The concept of ”bounded variation” is used by Manski and Pepper (2013, 2018) to estimate average treatment effects under various designs, such as regression discontinuity or difference-in-differences.

4. Total number of regular workers employed full-time throughout the entire year (for the narrower version of the bounds)

To this end, we rely on two primary data sources: the LFS administered by Istat and social security contribution data from INPS. In contrast to Istat's methodology (De Gregorio and Giordano, 2015), we utilize publicly available aggregate data rather than individual-level information.

For the first quantity, namely the total number of workers in the labor market, we use the official publication from Istat, which is based on the LFS. This survey collects data over multiple waves conducted at various points throughout the year, allowing Istat to report the average number of employed workers in Italy across the entire year. As outlined in Section 3.1, we assume that the LFS data is reliable. However, the target population of the survey does not include irregular migration and domestic workers, both of which are likely significant components of irregular employment in Italy (see Appendix B). To account for these gaps, we complement the LFS data with additional sources.

Irregular migration is monitored annually by a separate institution. We focus on non-EU irregular migrants, as intra-EU migrants are not considered irregular under EU law (Directive 2004/38/EC) and are possibly surveyed in the LFS as long as they live in Italy. For the irregular component, we rely on estimates provided by the Iniziative e Studi sulla Multietnicità (ISMU) Institute and assume all these migrants are employed.[14] For data on domestic workers, we obtain direct information from Istat through a specific request.[15]

The total number of workers is thus the sum of the LFS estimate, irregular migrants, and domestic workers aged over 15.

For quantities (2)–(4), namely the total number of regular workers, total weeks worked, and the number of regular workers employed full-time throughout the entire year, we primarily rely on Inps data. These data are categorized by the social security fund to which workers contribute, allowing us to separately identify non-farm private sector employees, public sector employees, agricultural workers, and domestic workers. For each category, we have monthly data on the number of workers

[14] ISMU is an independent research organization focused on migration and integration processes, and is the source Istat used for similar estimates until 2019 (Istat, 2021b).

[15] We can share the details of our interaction with Istat upon request but not the data, which are personal.

who had at least one day of paid employment (i.e., workers for whom an employer made social security contributions, even if the employment lasted only one day). To estimate the total number of workers over the year we calculate the annual average of these monthly worker counts. Additionally, we can derive the total number of weeks worked and the average number of full-time, permanent workers each month from the same data sources.

An important consideration is that the Inps data on total weeks worked do not account for weeks in which social security contributions are paid through short-time work schemes. These schemes are widely used in Italy, not only during recessions, as their eligibility conditions are relatively generous (Carta et al., 2022). Moreover, during the Covid-19 pandemic, both the generosity of benefits and the eligibility criteria were expanded to support employment. Under certain conditions, workers receiving short-time work benefits are considered employed in the Istat data (Istat, 2021a). Therefore, in order to make an accurate comparison, we must add these weeks back into the Inps data. We achieve this by utilizing publicly available information from Inps on the total number of weeks authorized for Cassa Integrazione Guadagni (CIG), the Italian short-time work scheme. Not all authorized weeks are actually used, so we adjust this number using appropriate weights provided by Inps, which have a lag of about three months. The total number of weeks worked is therefore the sum of the weeks reported by Inps and the weeks during which workers were covered by CIG.

Additionally, for the geographical breakdown of our results, we note that all of the aforementioned data is also available separately for the North, Centre, and South of Italy. The only exception is the data on irregular migrants, which is available only at the national level. As a result, we need to impute regional data for irregular migrants using ISMU estimates. Our preferred procedure combines ISMU data with information from Istat and the Ministry of Labour. First, we categorize overall employment into five sectors (agriculture, retail, industry, construction, and services) using Istat's data, based on the ATECO 2007 classification (the Italian counterpart to NACE).[16] Next, we calculate the share of employment in each sector for each of the three areas (North, Centre, and South), thus deriving the relevance of each sector in the area. We combine this information with

[16] For instance, industry includes sectors like mining, manufacturing, electricity, and water supply, while services include sectors such as transportation, accommodation, financial services, and real estate.

data by the Ministry of Labor (Ministry of Labour, 2024), which provides the share of extra-EU workers in these sectors at the national level.[17] Using this information, we distribute the ISMU estimates for irregular migrants across sectors and then weight these estimates by the regional sectoral employment shares derived from Istat data. For further details on these calculations, refer to Appendix B.2.

Further information on the data sources and procedures described in this section can be found in Appendix B. We note that, due to the data sources utilized, we are able to provide our bounds approximately 12 months after the end of the year, well in advance of the official estimates from Istat (which have 2 years delay).

5 Results

5.1 National Estimates

Figure 1, panel (a), shows the bounds for the ratio of undeclared workers over the period 2014–2023, based on the least demanding assumptions outlined in Section 3.1.[18] Specifically, the lower bound relies on the number of workers who are declared at some point during the year, regardless of the number of weeks worked. For comparison, we also include the national estimate of undeclared employment, both before and after the revisions discussed in Section 2. Although these bounds are relatively wide, they reveal some important patterns. First, the recent revision of the official statistics appears particularly significant for the years 2014–2017, where the previous estimate exceeded our upper bound, suggesting that the earlier estimate was likely an overestimate of the actual rate of undeclared workers. Second, both the upper and lower bounds exhibit a downward trend after the pandemic years, mirroring the trend observed in the official estimates. While the wide range of these bounds does not allow us to definitively conclude that the rate of undeclared workers is decreasing (for instance, they could be consistent with a stable rate of undeclared workers around 9%), they do provide evidence that there has not been a strong increase in undeclared

[17] Specifically, the Ministry of Labor produces a report on the labor market conditions of migrants every year. We use these reports to retrieve the sectoral and geographical composition of extra-EU migrants' employment between 2014 and 2023.

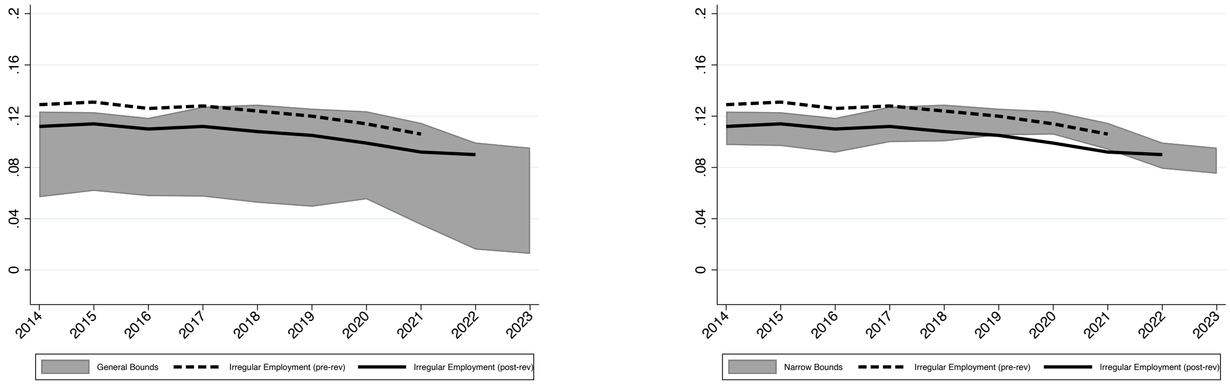
[18] The numbers corresponding to Figure 1 are reported in Appendix Table 1.

employment during this period.

In panel (b) of Figure 1, we incorporate the additional assumption that workers not employed full-time/full-year have a lower sampling probability in the LFS and should thus be rescaled when computing the lower bound (as discussed in Section 3.2). This adjustment significantly narrows the bounds, which now range from 9.8% to 12.4% in 2014 and 7.5% to 9.5% in 2023. Despite these narrower ranges, the bounds still encompass the official estimates (both pre- and post-2024 revision) for most of the years considered. This demonstrates that our partial identification strategy accounts for the uncertainty associated with the model specification in the official release. Importantly, while we include this uncertainty, our approach does not compromise the informativeness of our findings. In fact, with these revised bounds, we can conclude that the rate of undeclared jobs in the post-pandemic years has certainly declined: the upper bound for 2023 (9.5%) is below the lower bound for 2019 (10.5%).

Panel (b) also highlights that the official statistic for 2020 falls below the computed lower bound. This discrepancy is likely due to a larger decline in the number of registered workers according to the Inps data compared to the LFS data (the former shows a decline of about 700,000 workers, while the latter reports a decline of 600,000 between 2019 and 2020). While this may be attributed to misreporting in one of the data sources during the pandemic year (see Section 5.3), we argue that it is not unrealistic to observe an increase (or a stable rate), rather than a decline, in undeclared workers during the peak of the pandemic for two reasons. First, pandemic-related restrictions in 2020 affected legal employment possibly providing incentives to work without following the rules. Furthermore, undeclared employment has been proven increasing during recessions (Loyaza and Rigolini, 2006). Second, the official statistics show a smooth trend, both before and after the revisions, suggesting that some adjustments made during the definition of the official figures might prevent abrupt changes in the time series from one year to the next (and would thus prevent from estimating a sudden increase after a period of constant decline in the official rate). While this assumption of smooth transitions is reasonable in normal times, it may be less appropriate during turbulent periods, such as the COVID-19 pandemic. Indeed, for 2022, when both the upper and lower bounds show a significant decline compared to the previous year, the official statistic predicts a stable rate, which may reflect a re-adjustment process following the overly generous declining trend observed during the pandemic years.

Figure 1: Bounds for undeclared workers ratio in Italy



(a) General Bounds

(b) Narrow Bounds

5.2 Regional Estimates

As discussed in Section 4, we are able to estimate the narrower version of the bounds at the local level, differentiating between the northern, central, and southern regions of Italy.[19] This exercise illustrates the effectiveness of our bounds in making regional comparisons, as they are narrow enough to avoid overlap between regions, thereby enabling us to identify areas where undeclared work is more frequent.

Figure 2 presents the bounds for the three regions from 2014 to 2023.[20] Official data on the geographical decomposition of undeclared work after the 2024 revision are only available for the period 2021-2022 and are reported accordingly in the figure. The bounds for northern Italy tend to be narrower, which may be due to a larger proportion of workers employed full-time throughout the year, who are thus more likely to work 52 weeks annually. Additionally, the bounds for the northern region align with the official rate for the entire period before and after the revision. In contrast, similar to what was observed at the national level before the revision, in the center and

[19] The only information unavailable at the regional level pertains to irregular migration (Section 4). Appendix B.2 outlines the procedure used to allocate migrants to different areas. Since the share of irregular migrants is relatively small, the assumption made regarding their regional distribution has a minimal impact on the results.

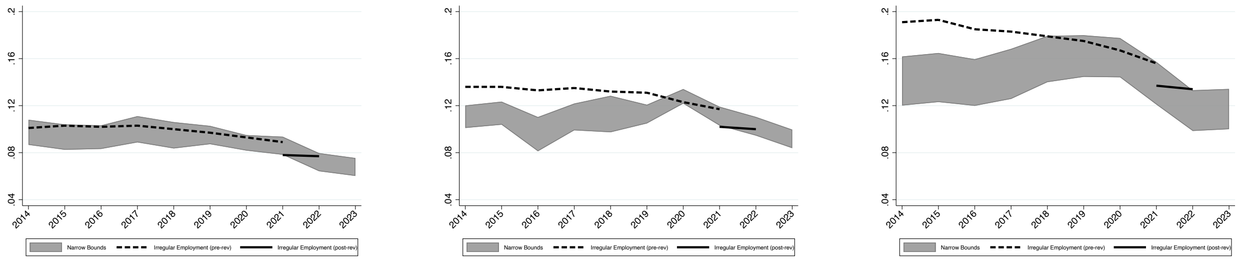
[20] The numbers corresponding to Figure 2 are reported in Appendix Table 2.

south the official rate exceeds the upper bound for the pre-2018 period before the revision, but it is within the bounds afterwards. Consequently, the rate of undeclared workers in the center and south before 2018 was likely overestimated by the official statistics. One notable strength of our approach is that it would have provided insights into this issue even before the official statistics were revised.

By comparing the three regions, we can assess where undeclared work is more prevalent and how its evolution has differed across areas. First, undeclared work is more common in the south and the center compared to the north; for instance, the lower bound in the south (which peaks at 12% in 2014 and 2016) consistently remains above the upper bound in the north (which reached a maximum of 11.1% in 2017). This pattern holds for the center as well, but only in the more recent years (in 2023, the lower bound for the center is 8.3%, while the upper bound for the north is 7.5%). Consistent with the official statistics, our non-parametric approach confirms that undeclared work is more widespread in the central and southern regions of Italy. Second, the undeclared workers' ratio has decreased in all areas following the pandemic, with the upper bound for 2022 falling below the 2020 values across all regions, and also below the 2019 values for both the north and the south. Before the pandemic, the bounds we estimated are consistent with a stable trend in undeclared employment in all regions. Third, similar to the national data, the bounds suggest an increase in undeclared work in 2020, particularly in the central region, where the lower bound for 2020 matches the upper bound for the previous year. However, any increase appears to be short-lived, with a rapid decline thereafter.

Overall, the regional analysis validates the robustness of our non-parametric approach. The trends we identify are consistent with existing evidence and the broader institutional context, with the southern regions traditionally being more exposed to undeclared work (Boeri and Garibaldi, 2005), and a recent decline in the prevalence of undeclared work. Our approach also captures subtleties that the official statistics fail to reflect, such as a plausible increase in undeclared work during the pandemic.

Figure 2: Bounds for undeclared workers ratio in different areas



(a) North

(b) Center

(c) South

5.3 Extension: Misreporting in the LFS Data

So far, we have assumed that the number of individuals employed each year, as reported in the LFS, is accurate. However, it is possible that these figures are inaccurate due to (either voluntary or involuntary) misreporting by individuals who claim to be non-employed when, in reality, they are employed (see e.g. Boeri and Garibaldi, 2005), or due to sampling errors. As discussed in Section 3.3, we can recompute the bounds by incorporating potential misreporting, provided we can make assumptions about its magnitude, denoted as δ. In this section, we proceed in two steps: first, we conduct two exercises to address 1) non-employed individuals who are in fact employed, and 2) measurement errors in the reported size of both regular and irregular employment. Second, we reverse the process and compute the value of δ that would be necessary to justify a rate of irregular employment that falls outside the bounds we have established so far.

Mis-reporting such that individuals declared as non-employed are in fact employed. We first consider the case of misreporting where individuals who are declared as non-employed are, in fact, employed.[21] To estimate the size of this misreporting, we rely on the results from Istat, which provide insights into misreporting in regular employment and use this information to project similar proportions for irregular employment. Istat estimates that around 0.5% of the inactive population may, in fact, be working (Istat, 2015). Using this estimate, we calculate the number of individuals corresponding to this fraction for each year and add them to the number of

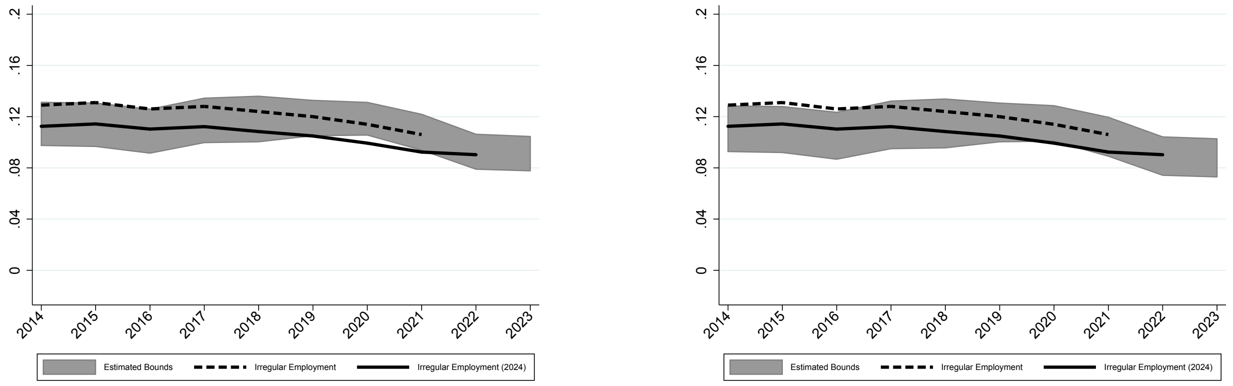
[21] Another type of misreporting concerns individuals who declared themselves as employed but in fact are not. According to the analysis by Istat (2015) this occurence is negligible.

employed individuals recovered from the LFS in building the upper bound. This adjustment for misreporting can only affect the undeclared workers' rate upward, as it increases the number of employees. Therefore, the lower bound, which estimates the lowest possible rate consistent with the data, remains unchanged. Formally, in this exercise, the misreporting adjustment, δ, ranges from 0 to 0.5% of the inactive population. Another important advantage of the bounded variation over traditional approaches is that the reasons why $\delta \neq 0$ need not be known (Manski and Pepper, 2018). As a consequence, it is important to note that this exercise also captures potential positive correlations between economic conditions and misreporting. If misreporting is more likely during times of economic downturns, linking it to the inactivity rate (which is cyclical) could help capture this phenomenon. In Figure 3, panel (a), we report the misreporting-adjusted bounds. The upper bound rises slightly, now encompassing the previous official statistics, but continues to show a downward trend in recent years, with the 2023 value just below the 2020 lower bound.

Measurement error in the size of (regular and irregular) employees. Whilst the previous exercise addresses a strategic misreporting, a different source of error is related to the sampling error from the LFS. Eurostat computes a 95% confidence interval for the estimate of the number of employed individuals from the 2020 LFS which ranges from -87,000 to +87,000 individuals. This corresponds to approximately 0.5% of the employed individuals in the LFS for that year (Eurostat, 2022). Based on this information, we adjust our upper and lower bounds for the number of regular workers by adding and subtracting 0.5% of the regular workers in every year, respectively, when computing the bounds. The results of these adjustments are shown in Figure 3, panel (b). The upper bound remains similar to what we computed in panel (a). The lower bound moves down slightly, with the 2023 lower bound aligned with the upper bound from 2020.

How much misreporting would be needed to include the official statistics in the bounds. Alternatively, we can consider the amount of misreporting in the LFS necessary to adjust the lower bound so that it includes the official undeclared worker rate observed in 2020. Specifically, the lower bound we compute for 2020 is 10.56%, while the official estimate reports 9.93%. To adjust our lower bound to match the official figure, it would require around 130,000 individuals to incorrectly

Figure 3: Adjusted Bounds



(a) Misreporting

(b) Sampling Error

declare themselves as employed when they were actually not employed. As shown in Figure 3, this adjustment would correspond to the maximum admitted sampling error. However, it could also be attributable to our assumption that all irregular migrants are indeed working. The number of misreported workers needed to reconcile the bounds is quite large—corresponding, for example, to more than 1 out of 4 irregular migrants. This suggests that, for the lower bound to align with the official estimate, a significant proportion of irregular migrants would have needed to be without paid employment.

6 Conclusion

The method proposed in this paper addresses two key limitations in previous empirical approaches for estimating the prevalence of undeclared employment in the labor market. Specifically, it does not rely on assumptions about the underlying factors or functional forms that might explain the phenomenon, and it provides credible estimates with minimal delay. We demonstrate that our bounded estimates not only offer insights into the evolution of informal employment but also highlight where official estimates may fall short in capturing the true numbers. Additionally, by clearly specifying the assumptions and data used, our method is easily adaptable to other countries and allows for extensions that relax the imposed restrictions or incorporate additional data to refine the analysis. For instance, if more detailed sector-specific data become available through targeted surveys or alternative methods, these can be integrated into the analysis for relevant sectors, while applying our approach to the rest of the economy, further narrowing the bounds.

Furthermore, our estimates can be used to explore the determinants and consequences of undeclared employment. Partial identification can be incorporated into causal analysis frameworks, such as instrumental variable techniques, with a valid (Bhattacharya et al., 2012) or invalid (Flores and Flores-Lagunes, 2013) instruments, to obtain a more valuable insights into the effects of informal employment.

References

AitBihiOuali, L. and O. Bargain (2021). Undeclared work-evidence from france. *Economie et Statistique/Economics and Statistics* (526-527), 71–92.

Battistin, E., E. Rettore, and U. Trivellato (2007). Choosing between alternative classification criteria to measure the labour force state. *Journal of the Royal Statistical Society Series A: Statistics in Society 170*(1), 5–27.

Bhattacharya, J., A. M. Shaikh, and E. J. Vytlacil (2012, jun). Treatment effect bounds: An application to Swan–Ganz catheterization. *Journal of Econometrics 168*(2), 223–243.

Boeri, T. and Garibaldi (2005). Shadow sorting. In *NBER International Seminar on Macroeconomics*, Volume 2005, pp. 125–170. The University of Chicago Press Chicago, IL.

Brandolini, A., P. Cipollone, and E. Viviano (2006). Does the ilo definition capture all unemployment? *Journal of the European Economic Association 4*(1), 153–179.

Buehn, A. and F. Schneider (2012, feb). Shadow economies around the world: novel insights, accepted knowledge, and new estimates. *International Tax and Public Finance 19*(1), 139–171.

Cappariello, R. and R. Zizza (2010, June). Dropping the Books and Working Off the Books. *LABOUR 24*(2), 139–162.

Carta, F., A. Dalla Zuanna, S. Lattanzio, and S. Lo Bello (2022). Il sistema di ammortizzatori sociali in Italia: aspetti critici nel confronto europeo (Social Shock Absorbers in Italy: A Comparison with the Main European Countries). *Bank of Italy Occasional Paper* (697). In Italian.

De Gregorio, C. and A. Giordano (2015). The heterogeneity of irregular employment in italy: some evidence from the labour force survey integrated with administrative data. istat working papers 1, Istat.

EEPO (2017). *European Platform tackling undeclared work: Member State Factsheets and Synthesis Report.* European Employment Policy Observatory, Brussels, Belgium.

Eurostat (2022). *Quality report of the European Union Labour Force Survey 2020 - Edition 2022.* Brussels: European Commission.

Fiess, N. M., M. Fugazza, and W. F. Maloney (2007). Informal labor markets and macroeconomic fluctuations.

Flores, C. A. and A. Flores-Lagunes (2013, oct). Partial Identification of Local Average Treatment Effects With an Invalid Instrument. *Journal of Business Economic Statistics 31*(4), 534–545.

Franić, J., I. A. Horodnic, and C. C. Williams (2023). *Extent of undeclared work in the European Union.* European Labour Authority, Bratislava, Slovakia.

Gries, T., L. Trapani, and M. Valente (2022). Quantifying Informal Employment From Irregular Migration Shocks. *SSRN Electronic Journal*.

Helberger, C. and H. Knepel (1988). How big is the shadow economy?: A re-analysis of the unobserved-variable approach of b.s. frey and h. weck-hannemann. *European Economic Review 32*(4), 965–976.

ISMU (2023). *Ventottesimo Rapporto sulle migrazioni.* FrancoAngeli.

Istat (2015). Soluzioni metodologiche per l'utilizzo integrato delle fonti statistiche per le stime dell'occupazione. *Istat Working Paper* (19/2015). In Italian.

Istat (2021a). Labour force 2021: what's new in the survey. `https://www.istat.it/en/news/labour-force-2021-whats-new-in-the-survey/`.

Istat (2021b). L'economia non osservata nei conti nazionali — anni 2016-2019. Technical report. In Italian.

Istat (2021c). L'economia non osservata nei conti nazionali — anni 2019-2022. Technical report. In Italian.

Koser, K. (2010). Dimensions and dynamics of irregular migration. *Population, Space and Place 16*(3), 181–193.

Loyaza, N. V. and J. Rigolini (2006). Informality trends and cycles. *Policy, Research working paper; Washington, D.C.: World Bank Group* (WPS 4078).

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review 80*(2), 319–323.

Manski, C. F. (2011, aug). Policy Analysis with Incredible Certitude. *Economic Journal 121*(554), F261–F289.

Manski, C. F. and J. V. Pepper (2013, mar). Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections. *Journal of Quantitative Criminology 29*(1), 123–141.

Manski, C. F. and J. V. Pepper (2018, may). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics 100*(2), 232–244.

Ministry of Labour (2024). *XIV RAPPORTO ANNUALE - Gli stranieri nel mercato del lavoro in Italia*. Ministry of Labour and Social Policies, Rome. In Italian.

Molinari, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics*, pp. 355–486.

Nakamura, H. (2013, dec). Wages of regular and irregular workers, the price of education, and income inequality. *The Journal of Economic Inequality 11*(4), 517–533.

Packard, T. G., J. Koettl, and C. Montenegro (2012). *In from the shadow: integrating Europe's informal labor.* World Bank Publications.

Pham, T. H. H. (2022, dec). Shadow Economy and Poverty: What Causes What? *The Journal of Economic Inequality 20*(4), 861–891.

Reyneri, E. (1998). The role of the underground economy in irregular migration to italy: cause or effect? *Journal of ethnic and migration studies 24*(2), 313–331.

Salis, E. (2012). Labour migration governance in contemporary europe. the case of italy. *Country Report for the LAB-MIG-GOV Project "Which labour migration governance for a more dynamic and inclusive Europe*.

Schneider, F. and A. Buehn (2017). Shadow economy: Estimation methods, problems, results and open questions. *Open Economics 1*(1), 1–29.

Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics 1*(2), 323–357.

Søndergaard, J. (2023). Undeclared danish labor: Using the labor input method with linked individual-level tax data to estimate undeclared work in denmark. *Journal of Economic Behavior Organization 214*, 708–730.

Tirozzi, A. (2022, dec). Searching for informal work using administrative data.

Vanderseypen, G., T. Tchipeva, J. Peschner, P. Rennoy, and C. C. Williams (2013). Undeclared work: recent developments. *Employment and social developments in Europe 2013*, 231–274.

Williams, C., P. Bejaković, D. Mikulić, J. Franic, A. Kedir, and I. A. Horodnic (2017). *An evaluation of the scale of undeclared work in the European Union and its structural determinants: estimates using the labour input method.* European Commission, Brussels.

A Official Methodologies

A.1 Istat Methodology

Istat leverages administrative and survey data to estimate labor input comprehensively, integrating microdata through different techniques. These different sources of information also allows to separate between regular and undeclared employment. Two key datasets are created annually to estimate labour input: one representing labor demand (administrative data focused on employers) and the other labor supply (LFS, focused on workers). The administrative registers used are described below following Istat (2015), and are generally provided by Inps (unless differently specified). For employees there are five data sources: Emens (EMEN12), which collects detailed information about employees in industry and services, including data on individual weeks of the year, taxable income, working hours, contracts, etc.; Inpdap (INPD), which gathers data on employees in public administrations; Enpals (ENPA), which covers employees in the entertainment industry; Colf and Badanti (DOME), which includes caregivers and domestic workers; and DMAG, which collects information on employees in agricultural enterprises. Note that for our estimates (which are focused on employees only) we use aggregate data from each of these registries (see Section 4). For self-employed there are four data sources: SILO_I (INDI), the Istat archive for self-employed workers in industry and services, which primarily compiles information from Inps, the Chamber of Commerce, and the Tax Agency; Autonomi agricoli (AUAG); the Inps separate management for parasubordinate workers (PARA); and the Inps separate management for professional collaborators (PROF). In addition to these sources, there is data from another govenrment agency (INAIL) concerning workers' insurance coverage, specifically: the DNA archive (INAD) for employees, the parasubordinate workers archive (INAP), and the temporary workers archive (INAI). The information derived from the LFS, instead, is supplemented with data about non-residents working for resident entities, using different data sources depending on their legal status (e.g., residents or undocumented workers). Furthermore, for specific sectors such as road transport and domestic services, additional integration of irregular labor positions is performed using indirect sources and ad hoc estimation methods. Finally, the irregular labor component is extended to include estimates of positions involved in illegal activities (Istat, 2021c).

Integrated data undergo consistency checks to statistically correct over-coverage in administrative archives and under-coverage in the LFS. In particular, this procedure estimates on the one hand the probability that a person who has a valid administrative record but declares being unemployed in the LFS is in fact employed and, on the other hand, the probability that a person who is employed in the LFS but not in the administrative registries is actually employed. Both methods are based on logit models estimated on subsamples of the population of interest and then applied to the whole population (Istat, 2015). The procedure ensures accuracy in classifying employment positions as regular or irregular, based on the presence or absence of valid administrative coverage signals. With this validated dataset, Istat computes undeclared employment as the discrepancy in labor input between household surveys and administrative data with the so-called Labour Input Method (LIM) developed by Istat in 1983 (Franić et al., 2023; Søndergaard, 2023).

Thus, ISTAT calculates the irregular employment ratio that is the incidence of irregular labor units relative to the total volume of labor units where irregular units are related to work performed without compliance with current labor, tax, and social security regulations, and therefore not directly observable through businesses, institutions, or administrative sources.

A.2 Methodologies in Other Countries

According to a recent report by the European Employment Policy Observatory (EEPO, 2017), there is no uniformity in the definition and measurement of undeclared work within the European Union. Some countries have legally defined undeclared work, while others rely on indirect definitions of the phenomenon. Regarding measurement, discrepancy methods similar to the LIM have been employed in countries such as France, Austria, Poland, Hungary, and Sweden (EEPO, 2017; AitBihiOuali and Bargain, 2021).

Notably, the European Platform on Tackling Undeclared Work, initiated by the European Labour Authority, regularly publishes estimates of undeclared work in European countries using LIM. This method combines data from the European Labour Force Survey (LFS), which estimates labour supply, with the Statistics on Business Survey (SBS), which estimates labour demand. Both surveys are conducted by individual member states. After harmonizing and converting the datasets into the same units (e.g., hours worked), discrepancies between the two are calculated to estimate the extent of undeclared work.

Thus, this method holds significant relevance for both policymakers and academics (Vanderseypen et al., 2013; Franić et al., 2023), particularly for its effectiveness in estimating the scale of undeclared work. Consequently, even if there is ambiguity on the measurement of undeclared work in the European Union, there is broad consensus on the utility of indirect methods that detect discrepancies within comparable sets of secondary macroeconomic data originally collected or compiled for other purposes (Williams et al., 2017).

B More information on the data

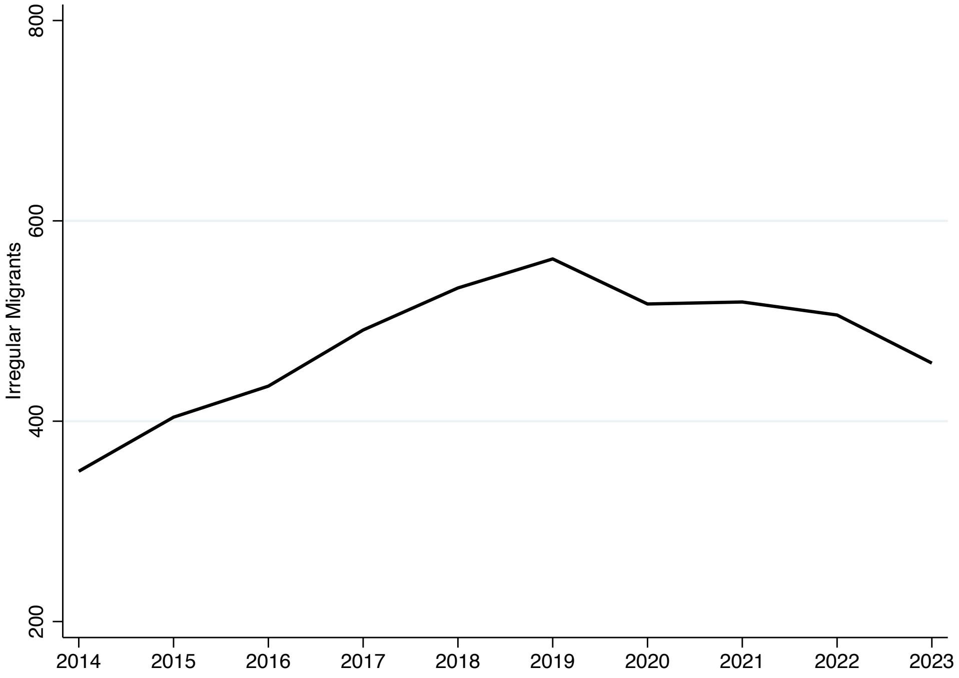
B.1 The Labor Force Survey

The Labour Force Survey (RFL) is Italy's primary source of statistical information on the labor market, providing official estimates on employment, unemployment, and key labor supply indicators such as profession, economic activity sector, hours worked, contract type and duration, and training. The survey is harmonized with EU Regulation 2019/1700, effective from January 1, 2021, to ensure standardized data collection across Europe.

Each year, over 250,000 families, amounting to approximately 600,000 individuals across 1,400 municipalities, are selected through a random sampling process using the National Resident Population Register. The individuals have to be regularly resident in the municipality considered and this excludes from the analysis irregular migrants. This sampling ensures the survey is representative of Italy's population. Families are interviewed four times over 15 months: twice in consecutive quarters, followed by a two-quarter break, and then another two consecutive quarters. Families with only inactive members aged 75 and older are excluded from re-interviews. Data collection is conducted continuously throughout the year.

The survey results are published through monthly and quarterly press releases and data is disseminated quarterly at the regional level and annually at the provincial level. Participation

Figure 4: Irregular migrants - ISMU (thousands)



in the survey is mandatory under Italian law (Legislative Decree No. 322/1989 and subsequent updates), though no penalties are imposed for non-compliance. The findings are essential for understanding labor market dynamics and feature prominently in major Istat publications such as the Annual Report and the Italian Statistical Yearbook.[22]

B.2 Irregular Migrants

As mentioned in Section 4, we collected data on irregular migrants from ISMU. Irregular migrants are more vulnerable to exploitation by organized crime, employers, and landlords compared to regular residents (Koser, 2010). This is also the case in Italy, which hosts a significant population of irregular migrants working in the informal economy (Reyneri, 1998; Salis, 2012). Therefore, it is essential to account for these factors in our analysis.

Figure 4 presents data on irregular migrants at the national level. The share of irregular migrants showed an upward trend until 2020, after which a decline occurred, partially due to the

[22] More details on the survey can be found at https://www.istat.it/en/news/general-info/.

COVID-19 pandemic and changes in Italy's regulatory framework (ISMU, 2023).

To estimate irregular migrants at the sub-national level, we imputed the ISMU data to the three main geographic areas of Italy: North, Center, and South. Our preferred approach utilizes the distribution of legal extra-EU migrants across sectors such as agriculture, industry, construction, retail, and services, combined with the distribution of total employment across these sectors in each area. As outlined in Section 4, we relied on data from ISTAT and the Ministry of Labour (MOL) for this process.

Specifically, we calculated the share of employment in each sector and area using ISTAT data. With MOL data, we estimated the share of extra-EU workers in these sectors at the national level. This allowed us to distribute ISMU's data on irregular migrants across sectors and subsequently across areas, weighted by the share of workers in each sector-area as derived from ISTAT data.

To illustrate, consider the year 2014. In that year, approximately 1.5 million extra-EU citizens residing in Italy were employed, with 5% in agriculture. ISMU reported 350,000 irregular migrants in Italy in the same year. The procedure assumes that the employment distribution of irregular migrants mirrored that of regular migrants. Thus, we estimated that around 16,000 irregular migrants were employed in agriculture. Since ISTAT reported that 38% of agricultural employment was in Northern Italy, we estimate that approximately 6,080 irregular migrants in agriculture were located in the North. Repeating this calculation for each sector provides the distribution of irregular migrants by area.

We estimate undeclared workers' ratio bounds using three alternative methods: evenly distributing migrants across the three areas; imputing migrants based on the share of employment recorded in the Labour Force Survey for each area; and using MOL data on the distribution of extra-EU workers across the areas. The main results of our analysis are confirmed irrespectively of the strategy used.

B.3 The Data from the National Security Institute (INPS)

We combine ISTAT data with administrative data from the National Social Security Institute (INPS), which maintains records of every worker who is regularly employed. INPS gathers information primarily through a form that employers must periodically submit to pay social contributions for their workers. We use INPS data to build the numerators for the shares used to construct the bounds. We have access only to aggregate data through the so-called *Osservatori INPS*. First, we measure the average number of workers with at least one day of regularly paid employment in every month provided by INPS for workers who are at least 15 years old separately for the private-non agricultrual sector, the public sector, the agricultural sector and the domestic sector. For the narrow version of our bounds, we can isolate the number of workers employed monthly working full-time and permanently. Secondly, we retrieve data on the number of weeks worked by individuals in the public or private sectors. We also obtained information from INPS on the *Cassa Integrazione Guadagni* (CIG), the Italian short-time work scheme, designed to support workers and companies during periods of economic difficulty. CIG provides temporary income support to workers whose working hours are reduced or who are temporarily laid off due to economic downturns, restructuring, or other crises affecting a company's operations. The weeks spent under the CIG regime are included in the aggregate weeks worked for a given year in our dataset, as workers are not formally laid off by their companies during the CIG period. The aggregate weeks worked by each regular employee will be used to adjust for part-time work and estimate the smallest number of regular workers to calculate the upper bound of the share in equation 1.

C Tables

Table 1: Bounds Imposing Different Assumptions and Istat Estimates

Year	**General Bounds**		**Narrower Bounds**		**Official Istat Estimates**	
	Lower Bounds	Upper Bounds	Lower Bounds	Upper Bounds	Pre-revision	Post-revision
2014	5.7	12.4	9.8	12.4	12.9	11.2
2015	6.2	12.3	9.7	12.3	13.1	11.4
2016	5.8	11.9	9.1	11.9	12.6	11.0
2017	5.7	12.7	10.0	12.7	12.8	11.2
2018	5.2	12.9	10.0	12.9	12.4	10.8
2019	4.9	12.6	10.5	12.6	12.0	10.5
2020	5.5	12.4	10.6	12.4	11.4	9.9
2021	3.5	11.5	9.4	11.5	10.6	9.2
2022	1.6	9.9	7.9	9.9		9.0
2023	1.3	9.5	7.5	9.5		

Table 2: Bounds and Istat Estimates by Region

Year	**North**				**Centre**				**South**		
	Lower	Upper	Istat		Lower	Upper	Istat		Lower	Upper	Istat
2014	8.7	10.8	10.1		10.1	12.0	13.6		12.0	16.2	19.1
2015	8.2	10.4	10.3		10.4	12.3	13.6		12.3	16.5	19.3
2016	8.3	10.3	10.2		8.1	11.0	13.3		12.0	16.0	18.5
2017	8.9	11.1	10.3		9.9	12.2	13.5		12.6	16.8	18.3
2018	8.4	10.6	10.0		9.7	12.8	13.2		14.0	17.9	17.9
2019	8.7	10.3	9.7		10.5	12.1	13.1		14.5	18.0	17.5
2020	8.2	9.5	9.3		12.2	13.4	12.3		14.4	17.8	16.7
2021	7.8	9.4	8.9		10.3	11.9	11.7		12.1	15.7	15.6
2022	6.4	8.0			9.5	11.0			9.9	13.3	
2023	6.0	7.5			8.3	10.0			10.0	13.4	