BETAM WORKING PAPER SERIES

RECOVERING PROVINCE IDENTIFIERS FROM TURKISH HOUSEHOLD LABOUR FORCE SURVEY MICRODATA

Luis Pinedo

Caro

Working

paper

#202603

Bahçeşehir University Centre for Economic and Social Research (BETAM) Osmanpaşa Mektebi Sokak No:4-6 34353 Beşiktaş, İSTANBUL, TÜRKİYE Phone: +902123810499

E-mail:

betam@eas.bau.edu.tr

11 September, 2026

Recovering Province Identifiers from Turkish Household Labour Force Survey Microdata

Luis Pinedo Caro[1]

Bahçeşehir University

Abstract

Official microdata from the Turkish Household Labour Force Survey (HLFS) provides geographic indicators at the NUTS-2 level (26 subregions). The way data is aggregated masks Türkiye's 81 provinces, putting together potentially heterogeneous local economies under the same label, thus, limiting the evaluation of policies' local impact. This paper presents a strategy to recover provincial-level identifiers in the microdata of the 2004-2013 period. A sequential identification algorithm is implemented based on the car plate provincial number, complemented with a year-specific urban-rural population re-weighting at the province level. The identification is validated internally using the logic behind existing multiprovince clusters and externally against administrative registration data from the Census, the Address Based Population Registration System (ABPRS) as well as with information from the Social Security Institution (SSI) regarding major economic activities. As an empirical proof of concept, informality rates at the province level are estimated for all 81 provinces, revealing sub-regional heterogeneity previously obscured under NUTS-2 reporting.

Keywords: Labour Force Survey, Data Reconstruction, Post-stratification, Turkish Statistical Institute (TurkStat).

J.E.L. Classification: C-81, J-21, R-12

[1] Luis Pinedo Caro is a researcher at the Bahçeşehir University Center for Socio-Economic Research (BETAM). *Email: Luis.pinedo@bau.edu.tr.*

1. Introduction

In Türkiye, social services, economic support and zone planning, among several others competencies fall on so-called metropolitan and non-metropolitan municipalities (ruling 81 provinces at the NUTS-3 level)[2] under Law 5216 and Law 6360 for the former and under Law No. 5393 for the latter. However, public-use microdata provided by the Turkish Statistical Institute[3] (TurkStat) restricts spatial granularity to 26 NUTS-2 subregions in the national Household Labour Force Survey (HLFS) and the Statistics on Income and Living Conditions (SILC), and to just 14 "regions" in the Family Structure Survey, just to mention some major, focused on social issues, surveys. These surveys are the main source of information for wages, occupations, unemployment, labour market dynamics, poverty, life satisfaction and family dynamics, just to mention a few core, relevant topics, for both local and Ankara-based policy makers. The lack of province level information regarding these topics creates, in practice, a significant analytical friction: economically distinct provinces—such as heavily industrialized manufacturing hubs and predominantly agrarian contiguous areas—are aggregated into single statistical units, obscuring localized economic shocks, hiding structural wage distortions, and rendering policy evaluations more noisy and less relevant.

This article tries to fill this information gap by identifying provinces (NUTS-3) in the HLFS for the 2004-2013 period and also by providing a post-stratification adjustment to allow for provincial survey weights. This kind of attempt at identifying hidden dimensions of public surveys is not new, though; bypassing spatial or population suppression in survey data is a practice that can be found in the international literature, increasingly relying on forensic data reconstruction and reverse-engineering techniques.

One such example is given by Enevoldsen (2023), that also addresses a critical data limitation in India's official Periodic Labour Force Survey (PLFS), where altered geographic and sampling unit codes prevented researchers from tracking urban households across consecutive annual panels. The author discovers that the identification codes in the first year's survey were systematically scrambled through a simple substitution cipher during data anonymization. An optimization algorithm was developed based on invariant household characteristics, allowing researchers to accurately link the first year's records to subsequent waves and unlocking the survey's full longitudinal potential for labor market analysis.

A different type of data recovery is shown by Helderop, Nelson, and Grubesic (2023) presenting a new method for mapping locations when exact street addresses have been hidden or altered for privacy. Standard mapping software often struggles with these blurred location data and places points inaccurately in the middle of a street or neighborhood. To fix this, the researchers developed a computer technique that uses statistical grouping to find the most accurate and realistic point within the allowed area, giving researchers a much more reliable way to map sensitive data.

---

[2] NUTS 3 is the third level of the Nomenclature of Territorial Units for Statistics. The European Union uses this system to divide regions for statistical data collection.
[3] In Turkish Türkiye İstatistik Kurumu (TÜİK).

Sometimes even the microdata itself can be reconstructed even when data is aggregated. Indeed, some literature within the computer science and statistical disclosure control have demonstrated that publicly released tabulates often contain sufficient mathematical constraints to fully reconstruct the underlying individual microdata records. A foundational example is Dinur and Nissim (2003), who provide the theoretical basis behind database reconstruction attacks. Modeling a statistical database as a simple string of bits—one per individual—they show that if enough aggregate queries (sums over subsets of records) are answered, even with some noise added to protect privacy, a fast algorithm can reconstruct almost the entire original database. In other words, they prove there is a hard mathematical limit: past a certain point, a dataset cannot be both useful for statistics and genuinely private at the same time.

This result is not merely theoretical. Garfinkel, Abowd, and Martindale (2019) show how such reconstruction attacks can be carried out in practice, treating a set of published statistical tables as a system of mathematical constraints that can be solved to recover the underlying microdata. The U.S. Census Bureau itself later confirmed the scale of this vulnerability directly: using only 34 of the 180 table sets published from the 2010 Census—out of more than 150 billion published statistics in total—an internal reconstruction exercise recovered five variables (census block, sex, age, race, and ethnicity) for individual records, and verified that all reconstructed records matched the confidential originals exactly in 70% of the country's census blocks, corresponding to roughly 97 million people (Vilhuber et al. 2023). In plain terms: publishing only aggregate tables, with no microdata released at all, still turned out to be mathematically equivalent to publishing near-complete individual-level data for most of the U.S. population. This finding was central to the Census Bureau's decision to abandon traditional disclosure limitation methods in favor of differential privacy for the 2020 Census, and it underscores the same broader lesson this paper's own methodology exploits: statistical releases can leak far more structural information than their aggregation appears to allow.

Within the Turkish applied micro-econometrics literature, examples of data recovery are notably scarce, reflecting how little attention TurkStat's suppression practices have received relative to comparable settings elsewhere. Tunalı (2023), in his methodological guide to the Household Labour Force Survey, documents well these limitations: TurkStat's rotating sample frame involves up to four visits to the same household over consecutive periods, generating a short panel structure with clear research value for studying labor market transitions—yet, as he notes, this panel dimension remains "beyond the reach of micro-data users because of TurkStat's data dissemination policy." His account catalogues a series of structural constraints imposed by the released microdata—the absence of primary sampling unit (PSU) identifiers, the suppression of the panel linkage across waves, and the non-institutional population as an awkward target population for regional estimation—but without proposing a workaround for any of them. Indeed, we observe research where some of the mentioned limitations are just simply stated, as in Bağır, Küçükbayrak, and Torun (2021) do when they note that research on wage underreporting in Turkey remains scarce precisely because the necessary disaggregated data is unavailable. In other stances, other sources are utilized to compensate for the lack of certain variables in the HLFS, see, Aydemir and Kırdar (2017), whose main wage analysis relies on the HLFS schooling variable but use the Turkish Demographic and Health Survey continuous years-of-schooling measure as an independent check. Our approach departs from this pattern precisely by exploiting the sampling mechanics of the HLFS itself and making it more informative.

Amid this general scarcity, two studies stand out as exceptions and represent the closest existing precedent to the approach developed here in the context of the Turkish HLFS. Pinedo Caro (2020) devised a residual identification strategy to isolate non-disclosed Syrian refugee sub-samples within HLFS waves: because the HLFS records place of birth but not nationality, Syrian refugees are indistinguishable from Turkey's ordinary migrant inflow. Pinedo Caro exploits the pre-war (2004–2010) arrival cohort to characterize the "standard" migration pattern, subtracts it from the 2011–2017 foreign-born population, and validates the resulting Syrian sub-sample against official age-distribution data from Turkey's Directorate General of Migration Management using a chi-squared test. Pinedo Caro (2021) similarly extracts monthly-frequency observations—finer than the aggregation TurkStat releases publicly—to nowcast labor market disruptions during the COVID-19 pandemic. Neither study, however, addresses spatial identification. Building directly upon this applied measurement tradition, this paper develops the first systematic procedure to fully reconstruct provincial indicators for 81 provinces across the entire 2004–2013 HLFS microdata archive without compromising sampling integrity.

This paper falls within the first strand of literature discussed above—exploiting deterministic sampling protocols and survey administration mechanics to recover suppressed identifiers—and makes three main contributions. First, it introduces a reproducible Stata/Mata algorithm that leverages deterministic survey sweeps and demographic distributions to recover 81 provincial identifiers from anonymized microdata, resolving the six pairs of provinces that cannot be separated deterministically through a population-based proportional assignment rule. Second, it constructs a multi-year population re-weighting matrix calibrated against official demographic censuses to correct for sample design deviations at the sub-regional level, ensuring that the reconstructed provincial panel remains macro-consistent with TurkStat's own published aggregates in every year. Third, it validates the reconstructed estimates against administrative Social Security Institution (SGK) records—demonstrating high spatial reliability across province sizes and sector classifications—and illustrates the empirical necessity of NUTS-3 granularity by uncovering intra-regional informality differences masked under standard NUTS-2 reporting.

2. Data and identification method

The primary dataset consists of harmonized microdata from the Turkish Household Labour Force Survey (HLFS) covering the 2004–2013 period.[4] During this decade, TurkStat designed the survey sampling framework to provide statistically representative estimates for the rural and urban areas of each subregion (NUTS-2 level). However, province-level (NUTS-3) information is explicitly omitted from the public microdata files, with only three exceptions: İstanbul (TR10), Ankara (TR51), and İzmir (TR31). These three metropolitan areas are identifiable solely because

[4] The sampling methodology changed in 2014, with TurkStat moving from fixed reference weeks to a continuous weekly fieldwork evenly distributed across all 52 weeks of the year. The ascending sequential spatial pattern that characterized the 2004-2013 period is masked and post-2014 province identifiers cannot be retrieved with the proposed method.

their respective NUTS-2 subregions are uniprovincial. Consequently, researchers are left with 78 provinces with no administrative identifier. In Türkiye's highly centralized administrative architecture—modeled historically after the French unitary state rather than federal structures like the German Länder or Spanish Autonomous Communities—NUTS-2 units exist strictly as artificial statistical aggregations rather than functional political or economic jurisdictions.

The problem with the HLFS's NUTS-2 level representativity lies on local labor market dynamics and administrative authority that vary substantially within these aggregate boundaries. For example, the TR41 subregion merges Bursa—a major industrial and automotive manufacturing hub—with Bilecik, a sparsely populated economy concentrated primarily in ceramics and raw material extraction. Treating TR41 as a homogeneous economic entity introduces severe aggregation bias, obscuring localized informal employment dynamics, municipal policy interventions, and targeted labor market inspections.

2.1 Identification strategy

To overcome these structural constraints and enable a more rigorous policy evaluation, I propose a deterministic identification and re-weighting methodology designed to reconstruct all 81 individual provincial identifiers. Restoring provincial resolution addresses a data gap and a public necessity, equipping researchers and policymakers with high-resolution empirical evidence grounded in actual administrative jurisdictions.

The identification strategy exploits a structural property of the raw HLFS microdata: the sampling design executes a deterministic, ascending spatial sweep within each annual wave, executed in 12 regular cycles. Rather than randomizing household level microdata observations, subregions are loaded into the final dataset in a systematic order that strictly mirrors the official numerical license-plate sequence (from 01 to 81) of Türkey's provinces. Across these multiple cycles, virtually all provinces are represented in a stable, predictable sequence. Appendix A provides the full structural verification of one complete cycle, illustrating the invariant position of the provinces as well as cases where consecutive provinces according to the plate number only appear once.[5]

Indeed, the license plate order recovers the vast majority of provinces but six specific pairs of provinces that share contiguous license-plate numbers preventing complete separation via spatial sequencing alone. To resolve these ties, I supplement the deterministic sweep with a proportional assignment rule. Following the natural plate order, initial household observations within the ambiguous block are assigned sequentially to the first appearing province up to the exact proportion that its population represents relative to the combined population of the two paired provinces.

2.2 Post-stratification adjustment and province-level survey weights

Following the identification of each province, I develop a custom post-stratification adjustment that creates survey weights at the NUTS-3 level. Benchmark population counts are derived from the 2000 General Population Census and the Address-Based Population Registration System (ABPRS) for the 2007–2014 period. For the intermediate years (2004–2006), provincial

[5] The complete cycle matrices covering all survey rounds from 2004 to 2013 is available at request.

populations are interpolated assuming a constant exponential growth rate between the 2000 and 2007 benchmarks. Official urban-rural distribution ratios are then applied to establish province- and area-specific population targets, denoted $N_{p,s,t}^{official}$, where $s \in \{U, R\}$ refers to the urban or rural stratum, t denotes the year and p the province. Household observations are then re-weighted at the province × urban/rural cell level so that sample totals match these official targets:

$$w_{i,p,t}^{temp} = w_{i,p,t} \cdot \frac{N_{p,s,t}^{official}}{\hat{N}_{p,s,t}},$$

where $w_{i,p,t}$ is the original TurkStat design weight for individual i in province p and year t, and

$$\hat{N}_{p,s,t} = \sum\nolimits_{i \in p,s,t} w_{i,p,t}$$

is the corresponding design-weighted cell population. In the rare case where a province lacks one of the two strata in a given year a province-level factor computed analogously over the full province total (see Appendix B for the provinces and years affected) is calculated.

The resulting province-level expanded population, once aggregated to the NUTS-2 level, always exceeds the corresponding NUTS-2 total as released in TurkStat's own microdata. This is because TurkStat's sampling design targets the non-institutional population rather than the total resident population. To maintain macro-consistency with official statistical releases, provincial totals are constrained to sum exactly to their respective NUTS-2 subregion population as it appears in the released microdata. Since non-institutional population figures are unobserved at the NUTS-3 level, I assume that the non-institutional population share of each NUTS-2 subregion, r_g, applies uniformly across its constituent provinces. This ratio is defined as:

$$r_{g,t} = \frac{\sum_{i \in g,t} w_{i,t}}{\sum_{p \in g} N_{p,t}^{official}},$$

where the numerator is the NUTS-2 subregion's own aggregate population as implied by TurkStat's original design weights, and $N_{g,t}^{official}$ adds up[6] the official resident population over the provinces that belong to that subregion. Each province's survey weights are then scaled by the constant NUTS-2 factor $r_{g,t}$, yielding the final post-stratification weights:

$$w_{i,p,t}^{province} = w_{i,p,t}^{temp} \cdot r_{g(p),t} \,,$$

where $g(p)$ denotes the NUTS-2 subregion containing province p. There is one exception worth mentioning, Kilis (2004) and Tunceli (2009) were not visited in the years shown in brackets but TurkStat assumed they were represented by the remaining provinces in their subregion. This assumption is not acceptable for the province-level disaggregation, and in the years where a province is not explicitly visited the total count at the subregional level deliberately excludes the population of the missing province.

[6] Note the "g" refers to the relevant subregion, hence why the sum symbol is not added.

3. Internal and External Data Validation & Application

Because the reconstructed provincial identifiers are recovered algorithmically rather than directly disclosed by TurkStat, establishing their empirical validity is a necessary prerequisite for subsequent analysis. To demonstrate that the identified panel reflects genuine administrative jurisdictions rather than a statistical artifact, I test the methodology against a dual validation process that combines an internal structural falsification test, with external administrative benchmarking against the Social Security Institution (SGK) records and the Census/ABPRS.[7]

Sequence of subregions in the dataset and plate number. Having six pairs of provinces that cannot be separated from one another through a deterministic mechanism is bad news for the accuracy of certain province-level estimates, yet this feature produces a clean falsification test with evidence in favour of plate numbers having been used by TurkStat when gathering the microdata. Suppose, instead, that provinces were randomly ordered within subregions (this would render the plate-number argument useless), there would be no reason for ambiguity to arise *specifically and exclusively* between pairs of provinces with consecutive license-plate numbers. Under a null hypothesis of "no real order," any pair of provinces within the same NUTS-2 subregion would be equally likely to generate an indistinguishable block — yet the only six ambiguous cases across the entire cycle are exactly the six pairs of provinces with consecutive plate numbers. Moreover, this is not a one-off pattern: the same six pairs, and only these six, produce an indistinguishable block consistently across all twelve sub-rounds, in every year from 2004 to 2013 — 120 repetitions of the sampling cycle in total. Under random assignment, the identity of colliding provinces would be expected to vary from one cycle to the next rather than reproduce the exact same six pairs every single time. This perfect and persistent concentration is precisely the signature predicted by the plate-ordering hypothesis, and is essentially impossible to reconcile with any alternative hypothesis of random ordering.

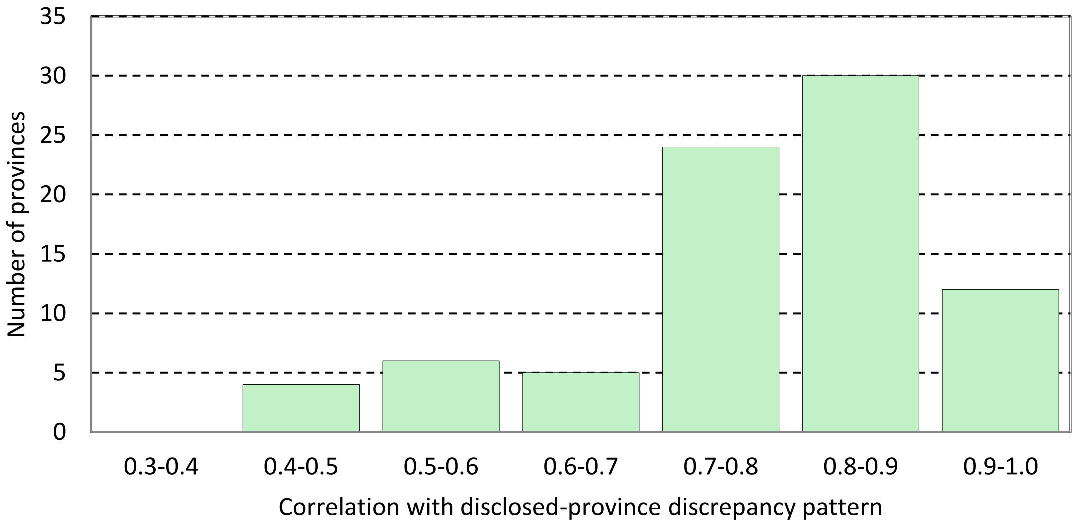
Relative importance of sample sizes against actual population shares. To assess internal validity, I compare the pooled 2004–2013 sample share of each province against its share of the actual resident population over the same period, drawn from ABPRS-based provincial population totals. Sample and population shares are highly correlated across all 81 provinces ($r = 0.964$). Restricting the comparison to the 69 provinces that are directly and deterministically identified — excluding the six dual-plate pairs, whose split is constructed by design to match population shares and therefore cannot serve as independent evidence — the correlation is, if anything, marginally higher ($r = 0.966$). Because these 69 provinces are recovered purely from their position in the license-plate number sequence, with no proportional-assignment mechanism involved, this close correspondence is not a mechanical artifact: it indicates that the recovered provincial identifiers track the real population structure closely. The average absolute gap between a province's sample share and its population share is just 0.29 percentage points, and even the largest gaps in the panel — concentrated in a handful of very large provinces such as İstanbul and Ankara, whose survey representation reflects TurkStat's stratified NUTS-2/urban-rural quota design rather than proportional national sampling — remain under one percentage point. Small provinces identified with no ambiguity track their population share particularly closely: Burdur,

[7] Address Based Population Registration System, the official registration system used in Turkey to record residents.

for instance, sits at 0.30% of the sample against 0.35% of the population. This tight, province-by-province correspondence across the full size range of Türkiye's 81 provinces is yet another evidence that the plate-based identification recovers the real provincial structure.

Employment structure and provincial heterogeneity. A striking illustration of the aggregation bias emerges directly from the reconstructed provincial indicator. Within the merged TR41 subregion, Bursa's identified employment profile is dominated by manufacture of motor vehicles and trailers (7.96% of employees) and manufacture of textiles (9.80%) — consistent with the province's role as host to Tofaş, one of Turkey's largest automotive manufacturers. Bilecik, one of its NUTS-2 partners, shows an entirely different profile: manufacture of other non-metallic mineral products alone accounts for 19.38% of employees, the single largest non-agricultural category, reflecting the province's longstanding ceramics and tile industry (Bien Seramik, VitrA Tiles etc.). No comparable concentration in either sector appears in the other province. This sharp divergence, invisible under NUTS-2 reporting, is precisely the kind of within-region heterogeneity the reconstructed provincial panel is designed to recover.

Figure 1. Sectoral discrepancy pattern, by province



Source: TurkStat HLFS 2007-2013 microdata, SGK 4a province level statistics on economic activity and authors' own calculations. ***Notes:*** For each province, I compute the ratio of HLFS (survey) to SGK (administrative) employment in seven broad sectors, excluding agriculture and mining due to small cell counts. I then average these sector-by-sector ratios across the three provinces TurkStat discloses directly (İstanbul, Ankara, İzmir) to obtain a reference discrepancy pattern. The chart shows the distribution, across all 81 provinces, of the Pearson correlation between each province's own 7-sector ratio profile and this reference pattern. A high correlation indicates that a province's HLFS-SGK discrepancy follows the same sector-by-sector structure as the directly-disclosed province.

To assess how well this pattern generalizes beyond the Bursa illustration, I gather province-level information on the head count of formal employment among employees in private companies from two sources, the HLFS and the Turkish Social Security Institute. This is done for seven sectors, manufacturing (ISIC rev.4 codes 10-33), energy generation (35-39), constructions (41-43), trade and hospitality (45-47, 55-56), transport and communication (49-53, 58-63), finance, insurance and other business activities (64-82), and all other services (84-99) for each year from 2007 to 2013. Please, note that years 2004-2006 are not available and the agriculture and mining sectors, whose formal-employee cell counts are too sparse in smaller provinces to yield stable ratios, are excluded. The point is not to compare the figures between sources, though, these are not directly comparable as the HLFS measures employment during the reference week whereas the SSI measures it at the yearly level, SSI figures are, thus, invariably higher than the HLFS ones by definition. The idea is to compare the correlation between the HLFS/SGK discrepancy profile of the 79 newly found provinces with that of the average discrepancy profile of the three provinces directly disclosed by TurkStat (İstanbul, Ankara, and İzmir). If provincial identification were unreliable, there would be no reason for a reconstructed province's discrepancy pattern to resemble that of the disclosed provinces. Instead, the median correlation across all 81 provinces is 0.81, with several provinces (Gaziantep 0.95, Trabzon 0.93, Antalya 0.92) showcasing correlations close to those of Istanbul (0.97), Ankara (0.94) and Izmir (0.93). Restricting to the 78 reconstructed provinces alone, the mean correlation is 0.78 (range 0.44–0.98). Bursa, our illustrative case, falls in line with this broader pattern at 0.75. The lowest correlations cluster among the smallest provinces by HLFS sample size (e.g., Tunceli, Osmaniye), consistent with sampling noise rather than identification failure. This distribution indicates that the sectoral discrepancy between HLFS and SGK — driven by the structural difference between household survey and administrative registration data — is a feature shared systematically across Turkey's provinces, not one confined to the provinces TurkStat happens to disclose directly.

3.3 Empirical Application: Spatial Variation of Informal Employment

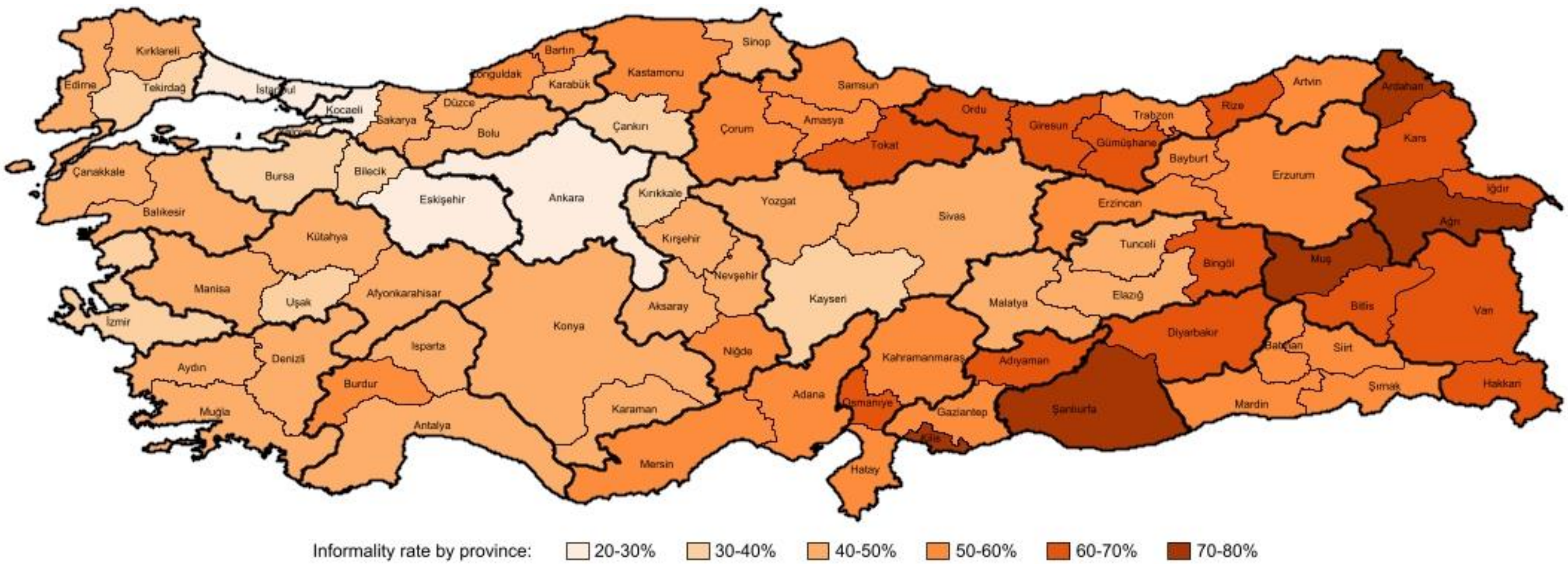
Empirical[8] literature from the 2000s and early 2010s—see, for instance, Angel-Urdinola and Tanabe (2012), and Tansel and Kan (2012)—would have benefited from high-resolution spatial microdata, as their foundational analyses of unregistered employment were constrained to coarse regional aggregates.

Had these evaluations been allowed to use the 81-province rather than the 26 NUTS-2 regions HLFS, their findings could have translated into perhaps, more actionable, targeted public policies—isolating municipal enforcement bottlenecks and localized wage distortions. Since TurkStat continues to withhold official province-level identifiers in public microdata releases, the reconstruction methodology presented here remains vital for enabling the evaluation of past local policies and covers, at least, a relevant period in Türkiye's recent history.

This section applies the identification strategy that recovers the 81-provinces in the HLFS to show within subregional informal employment rate heterogeneity, available at Figure 2, that shows a map where NUTS-2 subregional boundaries are overlayed on top of provincial borders. As it can be seen at the map, certain NUTS-2 averages merge highly industrial areas with adjacent agricultural provinces, whereas the provincial resolution developed in this article isolates localized pockets of high informality contiguous to formal manufacturing corridors.

[8] The utility of provincial-level microdata is far from a historical necessity. More recent studies, such as Del Carpio and Wagner (2015) and Tumen (2016) on the local labor market absorption of displaced populations, or Bağır, Küçükbayrak, and Torun (2021) and Baltagi and Başkaya (2022) on structural wage underreporting and spatial wage curves, demonstrate that sub-national heterogeneity remains central to Turkish labor economics but they use more recent data.

Figure 2. Informality rate, by province 2004-2013



Source: TurkStat 2004-2013 Household Labour Force Survey microdata and author's own calculations. ***Notes:*** The map shows the provincial rate of informal employment pooling the datasets from 2004 to 2013. NUTS-2 subregional borders (thicker lines) are shown for informative purposes.

This granular mapping demonstrates that labor informality in Turkey is to some degree province-specific, highlighting the necessity of NUTS-3 resolution for targeted municipal enforcement and regional policy evaluation.

An analysis of provincial data also remarks the importance of disentangling economically contrasting provinces paired under single NUTS-2 administrative aggregates. Within the TRC1 region, Gaziantep functions as a highly industrialized manufacturing powerhouse with extensive formal supply chains, whereas neighboring Adıyaman relies predominantly on smallholder agriculture and higher informal labor reliance. Similarly, despite sharing the TRC2 regional boundary, Şanlıurfa's labor market is heavily anchored in agricultural production with elevated informal employment rates, whereas Diyarbakır features a significantly higher concentration of formal public sector employment and service-oriented activities. Another example, under TR32, Muğla's economy is fundamentally driven by seasonal, service-oriented coastal tourism, whereas Denizli represents an inland textile and export-manufacturing hub with an entirely distinct formal employment structure.

3.4 Sample size limitations and confidence intervals

The sampling design of TurkStat's Household Labour Force Survey (HLFS) is officially representative at the NUTS-2 level,[9] hence yielding smaller annual sample sizes (n) for less populated NUTS-3 provinces. This is not a problem in large provinces such as Şanlıurfa, where the sample size still allows for several cuts with acceptable confidence intervals, but it is a constraint for certain types of analysis done in smaller provinces. The advice to mitigate this uncertainty is to pool microdata across several years of the 2004–2013 period. This ten-year period has the potential to substantially increase the effective sample size (N), reducing standard errors and yielding robust, dependable provincial-level estimates of informal employment suitable for empirical policy analysis. To document these precision boundaries standard errors and 95% confidence intervals are reported in Appendix C.

There is another genuine limitation, though, the Turkish HLFS public-use microdata files do not include a primary sampling unit (PSU) identifier. TurkStat's sampling design is a two-stage stratified cluster design in which the first stage selects PSUs consisting of blocks of approximately 100 households; this PSU identifier is not released to external researchers, and no survey documentation on block-level identifiers is publicly available. In its absence, I approximate the primary sampling unit using the household identifier (*famid*) in the svyset specification, with strata defined as the interaction of subregion (NUTS-2) and urban/rural status—the actual stratification variables used in TurkStat's own design. This choice captures within-household correlation but not the full within-block (PSU-level) correlation present in the true design, since multiple households within the same PSU are not distinguishable in the released data.

---

[9]The microdata used in this study originate from the Household Labour Force Survey (HLFS) provided by the Turkish Statistical Institute (TÜİK) under a standard Public-Use Microdata License (*Dağıtımında Kısıtlama Olmayan Mikro Veri Kullanım Taahhütnamesi*). The analytical procedure proposed in this paper strictly adheres to all data privacy and confidentiality principles outlined in Law No. 5429 on Turkish Statistics. The reconstructed geographical identifiers (NUTS-3 level) are compiled solely for the purpose of producing aggregate provincial statistics. At no point does this methodology attempt to identify individual households or respondents, ensuring that individual data confidentiality remains fully intact.

As a result, the standard errors reported throughout this paper should be interpreted as a conservative lower bound on the true sampling variance: they are very likely tighter than the standard errors that would be obtained under TurkStat's actual clustering structure, since neighboring households sharing an unobserved PSU are treated as independent draws. I report this limitation explicitly rather than overstate the precision of the estimates, consistent with the broader approach in this paper of documenting, rather than concealing, the boundaries of what the public microdata can support.

4. Conclusions

This article presents a systematic methodology to recover the provincial (NUTS-3) identifiers within TurkStat's Household Labour Force Survey throughout the 2004–2013 period. By exploiting the sequential order in which sub-regions appear in the raw microdata—which strictly aligns with the official Turkish vehicle license plate coding system (plaka in Turkish)—the proposed algorithm assigns block-level observations to their corresponding provincial administrative units. Furthermore, to address sampling design shifts and ensure demographic consistency, a post-stratification adjustment matrix was developed, generating province-level survey weights that preserve urban-rural representation while maintaining full coherence with TurkStat's published national expansion totals. The robustness of this spatial recovery procedure is established through internal and external validation exercises, where reconstructed labor market indicators shows a high correlation with administrative records from the Social Security Institution (SGK) across diverse province sizes and sector classifications.

Despite its demonstrated accuracy, the identification framework operates under a few structural limitations inherent to both the algorithmic strategy and TurkStat's underlying sampling design. Methodologically, while the majority of the 81 provinces are recovered deterministically through sequential sweeps, 12 smaller provinces require probabilistic assignment based on local demographic distributions. At the survey level, the sampling frame occasionally omitted specific small provinces from the data collection in select years—most notably Tunceli and Kilis—meaning spatial coverage is nearly, but not entirely, exhaustive across every wave. Additionally, in lower-population provinces, certain rural or urban strata exhibit reduced sample sizes, which naturally expands confidence intervals around highly disaggregated sub-group estimates.

These technical constraints are outweighed by the substantial analytical gains provided by the reconstructed dataset. An example of the value of this high-resolution set of labour force surveys is provided with the calculation of provincial informality rates for the first time using official Turkish microdata, exposing stark intra-regional differences that were invisible under standard NUTS-2 reporting. Ultimately, while this method rescues the provincial indicators for the 2004–2013 period, it cannot be directly applied to the post-2014 HLFS waves, as subsequent survey restructuring grouped provinces within common blocks with neither an identifiable separation nor logic. Still, rather than advocating for perpetual forensic data reconstruction, the core message of this paper is to highlight the necessity for TurkStat to officially release NUTS-3 spatial identifiers, providing researchers and policymakers with the administrative precision essential for effective locally-led, evidence-based policy design in Turkey.

References

Angel-Urdinola, D. F., & Tanabe, K. (2012). *Micro-determinants of informal employment in the Middle East and North Africa region* (Social Protection and Labor Discussion Paper No. 1201). World Bank.

Aydemir, A., & Kırdar, M. G. (2017). Low wage returns to schooling in a developing country: Evidence from a major policy reform in Turkey. *Oxford Bulletin of Economics and Statistics*, 79(6), 1046–1086.

Bağır, Y. K., Küçükbayrak, M., & Torun, H. (2021). *Declining labor market informality in Turkey: Unregistered employment and wage underreporting* (CBRT Working Paper Series No. 21/19). Central Bank of the Republic of Turkey.

Baltagi, B. H., & Başkaya, Y. S. (2022). Spatial wage curves for formal and informal workers in Turkey. *Journal of Spatial Econometrics*, 3(1), Article 1–24.

Del Carpio, X. V., & Wagner, M. (2015). *The impact of Syrian refugees on the Turkish labor market* (World Bank Policy Research Working Paper No. 7402). World Bank.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 202–210). Association for Computing Machinery.

Enevoldsen, N. (2023, February 6). *Filling gaps in India's official labor force survey: Constructing an urban panel* (EGC Perspectives). Yale University.

Garfinkel, S., Abowd, J. M., & Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62(3), 46–53.

Helderop, E., Nelson, J. R., & Grubesic, T. H. (2023). "Unmasking" masked address data: A medoid geocoding solution. *MethodsX*, 10, Article 102090.

Pinedo Caro, L. (2020). Identifying Syrian refugees in Turkish microdata. *IZA Journal of Development and Migration*, 11(1), 1–21.

Pinedo Caro, L. (2021). *The impact of COVID-19 on the Turkish labour market during 2020–2021: Results from a nowcasting model using Google mobility data* (ILO Research Brief Series). International Labour Organization.

Tansel, A., & Kan, E. Ö. (2012). *The formal/informal employment earnings gap: Evidence from Turkey* (IZA Discussion Paper No. 6556). Institute of Labor Economics (IZA).

Tumen, S. (2016). The economic impact of Syrian refugees on host countries: Quasi-experimental evidence from Turkey. *American Economic Review*, 106(5), 456–460.

Tunalı, İ. (2023). Unofficial user's guide to the Household Labor Force Survey, Turkey. *Ekonomi-tek*, 12(2).

Vilhuber, L., Abowd, J. M., Lewis, E., Goldschlag, N., Ashmead, R., Kifer, D., Leclerc, P., Rodríguez, R. A., Adams, T., Darais, D., Dey, S., Garfinkel, S. L., Moore, S., & Tadros, R. N. (2023).

A simulated reconstruction and reidentification attack on the 2010 U.S. Census (CES Working Paper 23-63R). U.S. Census Bureau.

Appendix A: HLFS Data Cycles and Sampling Patterns

Below is the order in which the variable "subregion" (NUTS2) appears in the microdata together with the *guessed* province and the plate numbering as the explanation for the logic behind the guess. Only 2004 is shown as a proof of concept due to space matters, tables exists for all other years and are available at request. No major change, just improved coverage, appears in later years thus not affecting the quality of the identification strategy.

Table 1. Sampling data pattern, plate order and identification potential

Order in cycle	Subregion (NUTS2)	Plate	Associated province	Ambiguous position?	Combined province (if ambiguous due to consecutive plate numbers)
1	TR-62 Adana	1	Adana	No	
2	TRC-1 Gaziantep	2	Adıyaman	No	
3	TR-33 Manisa	3	Afyonkarahisar	No	
4	TRA-2 Ağrı	4	Ağrı	No	
5	TR-83 Samsun	5	Amasya	No	
6	TR-51 Ankara	6	Ankara	No	
7	TR-61 Antalya	7	Antalya	No	
8	TR-90 Trabzon	8	Artvin	No	
9	TR-32 Aydın	9	Aydın	No	
10	TR-22 Balıkesir	10	Balıkesir	No	
11	TR-41 Bursa	11	Bilecik	No	
12	TRB-1 Malatya	12	Bingöl	No	
13	TRB-2 Van	13	Bitlis	No	
14	TR-42 Kocaeli	14	Bolu	No	
15	TR-61 Antalya	15	Burdur	No	
16	TR-41 Bursa	16	Bursa	No	
17	TR-22 Balıkesir	17	Çanakkale	No	
18	TR-82 Kastamonu	18	Çankırı	No	
19	TR-83 Samsun	19	Çorum	No	
20	TR-32 Aydın	20	Denizli	No	
21	TRC-2 Şanlıurfa	21	Diyarbakır	No	
22	TR-21 Tekirdağ	22	Edirne	No	
23	TRB-1 Malatya	23	Elazığ	No	
24	TRA-1 Erzurum	24	Erzincan	Yes	Erzurum (25)
25	TR-41 Bursa	26	Eskişehir	No	
26	TRC-1 Gaziantep	27	Gaziantep	No	
27	TR-90 Trabzon	28	Giresun	Yes	Gümüşhane (29)
28	TRB-2 Van	30	Hakkari	No	
29	TR-63 Hatay	31	Hatay	No	
30	TR-61 Antalya	32	Isparta	No	
31	TR-62 Adana	33	Mersin	No	
32	TR-10 Istanbul	34	İstanbul	No	
33	TR-31 İzmir	35	İzmir	No	
34	TRA-2 Ağrı	36	Kars	No	
35	TR-82 Kastamonu	37	Kastamonu	No	

36	TR-72 Kayseri	38	Kayseri	No	
37	TR-21 Tekirdağ	39	Kırklareli	No	
38	TR-71 Kırıkkale	40	Kırşehir	No	
39	TR-42 Kocaeli	41	Kocaeli	No	
40	TR-52 Konya	42	Konya	No	
41	TR-33 Manisa	43	Kütahya	No	
42	TRB-1 Malatya	44	Malatya	No	
43	TR-33 Manisa	45	Manisa	No	
44	TR-63 Hatay	46	Kahramanmaraş	No	
45	TRC-3 Mardin	47	Mardin	No	
46	TR-32 Aydın	48	Muğla	No	
47	TRB-2 Van	49	Muş	No	
48	TR-71 Kırıkkale	50	Nevşehir	Yes	Niğde (51)
49	TR-90 Trabzon	52	Ordu	Yes	Rize (53)
50	TR-42 Kocaeli	54	Sakarya	No	
51	TR-83 Samsun	55	Samsun	No	
52	TRC-3 Mardin	56	Siirt	No	
53	TR-82 Kastamonu	57	Sinop	No	
54	TR-72 Kayseri	58	Sivas	No	
55	TR-21 Tekirdağ	59	Tekirdağ	No	
56	TR-83 Samsun	60	Tokat	No	
57	TR-90 Trabzon	61	Trabzon	No	
58	TRC-2 Şanlıurfa	63	Şanlıurfa	No	
59	TR-33 Manisa	64	Uşak	No	
60	TRB-2 Van	65	Van	No	
61	TR-72 Kayseri	66	Yozgat	No	
62	TR-81 Zonguldak	67	Zonguldak	No	
63	TR-71 Kırıkkale	68	Aksaray	No	
64	TR-52 Konya	70	Karaman	No	
65	TR-71 Kırıkkale	71	Kırıkkale	No	
66	TRC-3 Mardin	72	Batman	Yes	Şırnak (73)
67	TR-81 Zonguldak	74	Bartın	No	
68	TRA-2 Ağrı	75	Ardahan	Yes	Iğdır (76)
69	TR-42 Kocaeli	77	Yalova	No	
70	TR-81 Zonguldak	78	Karabük	No	
71	TR-63 Hatay	80	Osmaniye	No	
72	TR-42 Kocaeli	81	Düzce	No	

Source: HLFS 2004 microdata from TurkStat, Turkish plate numbers from the Ministry of Interior and author's own calculations. ***Notes:*** The table shows the order of the 2004 HLFS data batches based on when the variable subregion changes as well as the plate numbers of Turkiye is ascending order together with the name of the province associated to that plate number. This table corresponds to the first of the twelve sub-round cycles in 2004; provinces absent from this particular cycle (e.g., Kilis, Bayburt, Tunceli) may still appear in other cycles within the same year — see Appendix B for full-year coverage. Combined provinces are disentangled using a proportional method where a share of observations equivalent to the population share of the respective province is assigned in ascending order.

Appendix B: Data availability matrix

Table B.1 Temporal coverage, available province and existing rural/urban strata.

Province	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Adana	F	F	F	F	F	F	F	F	F	F
Adıyaman	F	F	F	F	F	F	F	F	F	F
Afyonkarahisar	F	F	F	F	F	F	F	F	F	F
Aksaray	F	F	F	F	F	F	F	F	F	F
Amasya	F	F	F	F	F	F	F	F	F	F
Ankara	F	F	F	F	F	F	F	F	F	F
Antalya	F	F	F	F	F	F	F	F	F	F
Ardahan	F	F	F	U	U	F	F	F	F	F
Artvin	F	F	F	F	F	F	F	F	F	F
Aydın	F	F	F	F	F	F	F	F	F	F
Ağrı	F	F	F	F	F	F	F	F	F	F
Balıkesir	F	F	F	F	F	F	F	F	F	F
Bartın	F	F	F	F	F	F	F	F	F	F
Batman	F	F	F	F	F	F	F	F	F	F
Bayburt	F	U	U	U	F	F	F	F	F	F
Bilecik	F	F	F	F	F	F	F	F	F	F
Bingöl	F	F	F	F	F	F	F	F	F	F
Bitlis	F	F	F	F	F	F	F	F	F	F
Bolu	F	F	F	F	F	F	F	F	F	F
Burdur	F	F	F	F	F	F	F	F	F	F
Bursa	F	F	F	F	F	F	F	F	F	F
Denizli	F	F	F	F	F	F	F	F	F	F
Diyarbakır	F	F	F	F	F	F	F	F	F	F
Düzce	F	F	F	F	F	F	F	F	F	F
Edirne	F	F	F	F	F	F	F	F	F	F
Elazığ	F	F	F	F	F	F	F	F	F	F
Erzincan	F	F	F	F	F	F	F	F	F	F
Erzurum	F	F	F	F	F	F	F	F	F	F
Eskişehir	F	F	F	F	F	F	F	F	F	F
Gaziantep	F	F	F	F	F	F	F	F	F	F
Giresun	F	F	F	F	F	F	F	F	F	F
Gümüşhane	F	F	F	F	F	F	F	F	F	F
Hakkari	F	F	U	U	U	F	F	F	F	F
Hatay	F	F	F	F	F	F	F	F	F	F
Isparta	F	F	F	F	F	F	F	F	F	F
Iğdır	F	F	F	U	U	F	F	F	F	F
Kahramanmaraş	F	F	F	F	F	F	F	F	F	F
Karabük	U	F	F	F	F	F	F	F	F	F
Karaman	F	F	F	F	F	F	F	F	F	F
Kars	F	F	F	F	F	F	F	F	F	F
Kastamonu	F	F	F	F	F	F	F	F	F	F
Kayseri	F	F	F	F	F	F	F	F	F	F

Kilis	M	R	R	R	R	F	F	F	F	F
Kocaeli	F	F	F	F	F	F	F	F	F	F
Konya	F	F	F	F	F	F	F	F	F	F
Kütahya	F	F	F	F	F	F	F	F	F	F
Kırklareli	F	F	F	F	F	F	F	F	F	F
Kırıkkale	F	F	F	F	F	F	F	F	F	F
Kırşehir	F	F	F	F	F	F	F	F	F	F
Malatya	F	F	F	F	F	F	F	F	F	F
Manisa	F	F	F	F	F	F	F	F	F	F
Mardin	F	F	F	F	F	F	F	F	F	F
Mersin	F	F	F	F	F	F	F	F	F	F
Muğla	F	F	F	F	F	F	F	F	F	F
Muş	F	F	F	F	F	F	F	F	F	F
Nevşehir	F	F	F	F	F	F	F	F	F	F
Niğde	F	F	F	F	F	F	F	F	F	F
Ordu	F	F	F	F	F	F	F	F	F	F
Osmaniye	F	F	F	F	F	F	F	F	F	F
Rize	F	F	F	F	F	F	F	F	F	F
Sakarya	F	F	F	F	F	F	F	F	F	F
Samsun	F	F	F	F	F	F	F	F	F	F
Siirt	F	F	F	F	F	F	F	F	F	F
Sinop	F	F	F	F	F	F	F	F	F	F
Sivas	U	F	F	F	F	F	F	F	F	F
Tekirdağ	F	F	F	F	F	F	F	F	F	F
Tokat	F	F	F	F	F	F	F	F	F	F
Trabzon	F	F	F	F	F	F	F	F	F	F
Tunceli	R	F	U	U	U	M	U	F	F	F
Uşak	F	F	F	F	F	F	F	F	F	F
Van	F	F	F	F	F	F	F	F	F	F
Yalova	U	F	F	F	F	F	F	F	F	F
Yozgat	F	F	F	F	F	F	F	F	F	F
Zonguldak	F	F	F	F	F	F	F	F	F	F
Çanakkale	U	F	F	F	F	F	F	F	F	F
Çankırı	F	F	F	F	F	F	F	F	F	F
Çorum	F	F	F	F	F	F	F	F	F	F
İstanbul	F	F	F	F	F	F	F	F	F	F
İzmir	F	F	F	F	F	F	F	F	F	F
Şanlıurfa	F	F	F	F	F	F	F	F	F	F
Şırnak	F	F	F	F	F	F	F	F	F	F

Notes: Coverage: 785/810 cells FULL (97.0%). Provinces in alphabetical order. The 12 dual provinces (Erzincan/Erzurum, Giresun/Gümüşhane, Nevşehir/Niğde, Ordu/Rize, Batman/Şırnak, Ardahan/Iğdır) are marked in every year because their identification is always estimated, not directly observed.

F	FULL (urban + rural present)	M	MISSING (no observations at all)
U	ONLY URBAN (rural cell empty)		*Dual province*
R	ONLY RURAL (urban cell empty)		

Appendix C: Confidence Intervals and Precision Metrics

Table C.1. Province-level informal employment rate, Turkey, pooled 2004–2013

Plate number	Province	Informality rate	Std. error	95% CI (low)	95% CI (high)
1	Adana	0.515	0.0031	0.509	0.521
2	Adıyaman	0.649	0.0075	0.634	0.664
3	Afyonkarahisar	0.481	0.0048	0.471	0.490
4	Ağrı	0.748	0.0044	0.739	0.756
5	Amasya	0.547	0.0070	0.533	0.561
6	Ankara	0.225	0.0017	0.222	0.228
7	Antalya	0.417	0.0027	0.412	0.423
8	Artvin	0.516	0.0104	0.495	0.536
9	Aydın	0.496	0.0038	0.488	0.503
10	Balıkesir	0.472	0.0029	0.466	0.477
11	Bilecik	0.317	0.0087	0.300	0.334
12	Bingöl	0.613	0.0114	0.590	0.635
13	Bitlis	0.636	0.0089	0.618	0.653
14	Bolu	0.426	0.0089	0.409	0.443
15	Burdur	0.505	0.0102	0.485	0.525
16	Bursa	0.331	0.0023	0.326	0.335
17	Çanakkale	0.443	0.0054	0.433	0.454
18	Çankırı	0.387	0.0080	0.371	0.403
19	Çorum	0.554	0.0050	0.544	0.563
20	Denizli	0.404	0.0041	0.396	0.412
21	Diyarbakır	0.613	0.0046	0.604	0.622
22	Edirne	0.479	0.0070	0.465	0.493
23	Elazığ	0.480	0.0065	0.468	0.493
24	Erzincan	0.501	0.0078	0.486	0.516
25	Erzurum	0.566	0.0037	0.559	0.574
26	Eskişehir	0.267	0.0045	0.258	0.276
27	Gaziantep	0.527	0.0036	0.520	0.534
28	Giresun	0.605	0.0055	0.595	0.616
29	Gümüşhane	0.628	0.0102	0.608	0.648
30	Hakkari	0.618	0.0143	0.590	0.646
31	Hatay	0.576	0.0040	0.568	0.583
32	Isparta	0.410	0.0063	0.397	0.422
33	Mersin	0.550	0.0037	0.543	0.558
34	İstanbul	0.258	0.0012	0.256	0.260
35	İzmir	0.333	0.0019	0.330	0.337
36	Kars	0.653	0.0059	0.641	0.664
37	Kastamonu	0.531	0.0049	0.521	0.540
38	Kayseri	0.345	0.0042	0.337	0.354
39	Kırklareli	0.415	0.0058	0.403	0.426
40	Kırşehir	0.417	0.0100	0.397	0.436
41	Kocaeli	0.260	0.0032	0.253	0.266
42	Konya	0.478	0.0025	0.473	0.483
43	Kütahya	0.444	0.0045	0.435	0.453
44	Malatya	0.492	0.0047	0.483	0.501

Plate number	Province	Informality rate	Std. error	95% CI (low)	95% CI (high)
45	Manisa	0.465	0.0039	0.458	0.473
46	Kahramanmaraş	0.507	0.0048	0.498	0.516
47	Mardin	0.570	0.0065	0.557	0.582
48	Muğla	0.452	0.0044	0.443	0.460
49	Muş	0.764	0.0065	0.751	0.776
50	Nevşehir	0.455	0.0083	0.439	0.471
51	Niğde	0.537	0.0072	0.523	0.551
52	Ordu	0.618	0.0038	0.610	0.625
53	Rize	0.601	0.0082	0.585	0.617
54	Sakarya	0.450	0.0044	0.442	0.459
55	Samsun	0.561	0.0036	0.554	0.568
56	Siirt	0.594	0.0149	0.565	0.623
57	Sinop	0.494	0.0064	0.482	0.507
58	Sivas	0.440	0.0069	0.427	0.454
59	Tekirdağ	0.344	0.0036	0.337	0.351
60	Tokat	0.654	0.0050	0.644	0.663
61	Trabzon	0.562	0.0037	0.555	0.569
62	Tunceli	0.448	0.0235	0.402	0.494
63	Şanlıurfa	0.737	0.0041	0.729	0.745
64	Uşak	0.381	0.0059	0.370	0.393
65	Van	0.696	0.0049	0.686	0.706
66	Yozgat	0.496	0.0070	0.482	0.510
67	Zonguldak	0.591	0.0039	0.584	0.599
68	Aksaray	0.489	0.0067	0.476	0.502
69	Bayburt	0.535	0.0162	0.503	0.566
70	Karaman	0.458	0.0070	0.444	0.471
71	Kırıkkale	0.341	0.0073	0.327	0.356
72	Batman	0.577	0.0083	0.561	0.594
73	Şırnak	0.595	0.0102	0.575	0.615
74	Bartın	0.575	0.0076	0.560	0.590
75	Ardahan	0.737	0.0175	0.701	0.770
76	Iğdır	0.699	0.0138	0.671	0.725
77	Yalova	0.339	0.0105	0.319	0.360
78	Karabük	0.407	0.0083	0.391	0.424
79	Kilis	0.742	0.0174	0.707	0.775
80	Osmaniye	0.604	0.0066	0.591	0.617
81	Düzce	0.409	0.0069	0.396	0.423

Sources: TurkStat HLFS 2004-2013 microdata and author's own calculations. ***Notes:*** The table shows pooled estimates across all ten survey years (2004–2013), using the reconstructed province identifier, the re-weighted survey weight, and strata defined as the interaction of subregion (NUTS2) and urban/rural status. The primary sampling unit is approximated by the household identifier, since the true PSU identifier is not present in the public microdata (see main text for a full discussion of this limitation). Point estimates, linearized standard errors, and Wald 95% confidence intervals are reported for all 81 provinces.

Appendix D: Sample vs. population share, a province-level comparison

Table D.1 Share in sample and share in total population of each province

Province	Sample %	Population %	Province	Sample %	Population %
Adana	2.73%	2.82%	Kahramanmaraş	1.19%	1.43%
Adıyaman	0.76%	0.82%	Karabük	0.38%	0.31%
Afyonkarahisar	1.19%	0.98%	Karaman	0.50%	0.32%
Ağrı	1.33%	0.75%	Kars	0.66%	0.43%
Aksaray	0.70%	0.52%	Kastamonu	0.95%	0.50%
Amasya	0.54%	0.46%	Kayseri	1.31%	1.66%
Ankara	5.32%	6.39%	Kırıkkale	0.59%	0.40%
Antalya	2.09%	2.64%	Kırklareli	0.54%	0.46%
Ardahan	0.21%	0.15%	Kırşehir	0.52%	0.31%
Artvin	0.27%	0.23%	Kilis	0.14%	0.17%
Aydın	1.26%	1.35%	Kocaeli	2.38%	2.07%
Balıkesir	2.27%	1.57%	Konya	4.55%	2.79%
Bartın	0.35%	0.26%	Kütahya	1.32%	0.81%
Batman	0.72%	0.68%	Malatya	1.05%	1.04%
Bayburt	0.18%	0.11%	Manisa	1.98%	1.84%
Bilecik	0.25%	0.28%	Mardin	1.19%	1.04%
Bingöl	0.44%	0.35%	Mersin	2.01%	2.26%
Bitlis	0.65%	0.46%	Muğla	0.86%	1.10%
Bolu	0.40%	0.38%	Muş	0.82%	0.57%
Burdur	0.30%	0.35%	Nevşehir	0.54%	0.39%
Bursa	3.36%	3.48%	Niğde	0.61%	0.47%
Çanakkale	0.93%	0.67%	Ordu	1.20%	1.02%
Çankırı	0.52%	0.26%	Osmaniye	0.68%	0.65%
Çorum	1.00%	0.75%	Rize	0.54%	0.45%
Denizli	1.05%	1.27%	Sakarya	1.06%	1.18%
Diyarbakır	2.06%	2.08%	Samsun	1.68%	1.71%
Düzce	0.46%	0.46%	Siirt	0.49%	0.41%
Edirne	0.60%	0.55%	Sinop	0.55%	0.28%
Elazığ	0.75%	0.76%	Sivas	0.66%	0.89%
Erzincan	0.56%	0.31%	Şanlıurfa	2.32%	2.23%
Erzurum	1.99%	1.09%	Şırnak	0.61%	0.59%
Eskişehir	0.84%	1.04%	Tekirdağ	1.11%	1.07%
Gaziantep	2.18%	2.25%	Tokat	0.94%	0.88%

Province	Sample %	Population %	Province	Sample %	Population %
Giresun	0.70%	0.59%	Trabzon	1.23%	1.06%
Gümüşhane	0.20%	0.19%	Tunceli	0.09%	0.12%
Hakkari	0.36%	0.36%	Uşak	0.58%	0.46%
Hatay	1.70%	1.96%	Van	1.47%	1.38%
Iğdır	0.33%	0.25%	Yalova	0.25%	0.27%
Isparta	0.55%	0.59%	Yozgat	0.71%	0.68%
İstanbul	11.41%	17.72%	Zonguldak	1.12%	0.85%
İzmir	5.11%	5.28%			

Source: *TurkStat HLFS microdata 2004-2013, population by province 2007-2013 from the ABPRS and author's own calculations.* ***Notes:*** *The table shows provinces' shares in the total population of Türkiye from 2007 to 2013 and in the pooled sample size 2004-2013.*