Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Household-level measurement of the informal economy using unsupervised learning: The case of Turkey

Armagan Aktuna-Gunes

armagan.aktunagunes@gmail.com

ARTXDATA

Research Article

Keywords: informal economy, time use and household behavior, machine learning, statistical matching, anomaly detection, demand system estimation

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11721306/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Household-level measurement of the informal economy using unsupervised learning: The case of Turkey

Armagan Aktuna-Gunes

Lead Research Scientist at ARTXDATA, Paris, France; Ph.D. in Economics, University of Paris I Panthéon-Sorbonne, Paris, France.

Correspondence to: E-mail: armagan.aktunagunes@gmail.com

ORCID: 0009-0002-2969-8249

Statements and Declarations

Funding. The research leading to these results has received funding from the project "Analiza dobrobytu gospodarstw domowych i konsumpcji w Polsce, Francji i w Turcji z wykorzystaniem danych o wartości produkcji domowej i dochodów z pracy w szarej strefie" ["An Analysis of Household Welfare and Consumption in Poland, France and Turkey Using Data on the Value of Domestic Production and Earnings from Informal Work"], National Science Centre, Poland (Narodowe Centrum Nauki), grant agreement no. 2018/29/B/HS4/02026.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Author contributions. The author is the sole author of this work and was responsible for the study conception and design, data preparation, analysis and the writing of the manuscript.

Data availability. The microdata of the Time Use Survey and Household Budget Survey are available from the Turkish Statistical Institute (TURKSTAT) upon application.

Use of generative AI. Generative AI tools (Claude, Anthropic) were used to assist in revising the manuscript. The author reviewed and verified all content and takes full responsibility for the final text.

Abstract

This study proposes a household-level approach to measuring informal economic participation that combines microdata on time use and household expenditures. Using nationally representative data from Turkey, it makes two contributions. First, it develops a clustering-based statistical matching procedure that aligns the 2014–2015 Time Use Survey with the 2011 Household Budget Survey: principal component analysis and K-means clustering on more than thirty common variables define comparable sub-populations, so that the imputed value of unpaid domestic work can be added to monetary spending to construct full expenditure profiles. Second, it applies an ensemble of four unsupervised anomaly detection algorithms to identify households whose full expenditure is atypically high given their declared income and characteristics, an expenditure-based signal of undeclared income. The ensemble flags 61.3% of self-employed and 33.9% of wage-earning households, whereas a Complete Demand System estimated by the Generalized Method of Moments on the same survey yields 24.7% and 70.9%, respectively. The two approaches thus produce similar aggregate figures but rank the groups in opposite order. Because the detection thresholds are partly calibrated on macroeconomic income shares, the aggregate agreement is interpreted as a consistency check rather than independent validation, and the divergence as reflecting different margins of informality. The household-level labels also allow a descriptive profiling of flagged households by gender, income, age, education and marital status. The framework offers a transparent and replicable complement to structural approaches for analyzing informality with survey microdata.

Keywords: informal economy; time use and household behavior; machine learning; statistical matching; anomaly detection; demand system estimation

JEL Codes: E26, J22, C14, C38, D12

1. Introduction

The informal economy remains a persistent challenge to inclusive growth, fiscal sustainability and social protection, particularly in developing and emerging economies (Schneider and Enste 2000; La Porta and Shleifer 2014; Ulyssea 2020). Informal activity, by definition, escapes regulatory oversight and formal statistical reporting, which leads to the mismeasurement of economic output and to poorly targeted public policy. Accurate household-level estimates of informality are therefore critical for designing effective and equitable interventions.

Conventional approaches to measuring informal activity rely either on macroeconomic indicators—such as currency demand or Multiple Indicators Multiple Causes (MIMIC) models (Medina and Schneider 2018)—or on labor force surveys that define informality through social security registration (ILO 2018). However, these methods say little about the micro-level heterogeneity of household behavior. A complementary micro-econometric tradition infers undeclared income from household expenditures: households that under-report income consume more than their declared income would predict (Pissarides and Weber 1989; Lyssiotou et al. 2004; Hurst et al. 2014). Yet this tradition generally ignores time allocation and unpaid labor, although a growing body of research emphasizes the importance of domestic production in shaping consumption patterns and labor force participation (Bonke 1992;

Ironmonger 1996; Eurostat 2019). When such activity is omitted, welfare estimates and labor statistics fail to capture the full scope of economic contribution.

Microdata-based methods that combine time-use and consumption surveys address this gap by capturing behavioral dimensions that aggregate models miss. The imputed value of time spent on unpaid labor is closely related to household income, gender dynamics and broader socio-demographic characteristics (Aguiar and Hurst 2007; Frazis and Stewart 2011; Aktuna-Gunes et al. 2017). In earlier work, this logic was embedded in a Complete Demand System estimated on monetary and time expenditures (Aktuna-Gunes et al. 2014, 2017). Such structural models are, however, demanding: they require careful variable selection to satisfy identification and specification conditions, and their results are sensitive to functional-form assumptions.

Machine learning (ML) offers a complementary route. Where household production, time use, monetary transactions and socio-demographic factors intersect, ML methods can integrate heterogeneous features from several data sources and detect non-linear patterns without a pre-specified functional form (Mullainathan and Spiess 2017; Athey and Imbens 2019; Aktuna-Gunes 2021). This flexibility is valuable for the study of informality, where data fragmentation and behavioral complexity are pervasive. At the same time, ML methods do not by themselves provide the theoretical structure required for economic inference. This paper therefore combines them with an explicit identifying idea—the expenditure–income gap—and confronts their output with a structural benchmark.

The study makes two methodological contributions. First, it introduces a clustering-based statistical matching procedure for the Turkish Time Use Survey (TUS) and Household Budget Survey (HBS). Principal component analysis (PCA) reduces more than 30 variables to a small number of components, and K-means clustering groups households into behaviorally consistent sub-populations for each activity. Clusters are then linked across the two surveys through an explicit assignment rule defined on their common variables, and the time use of each TUS cluster is imputed to its HBS counterpart. Unlike regression-based matching, this approach does not impose a fixed partial-correlation structure on the variables, although it rests, like all statistical matching, on a conditional independence assumption that is discussed explicitly (D'Orazio et al. 2006; Rässler 2002). Second, households whose full expenditure is anomalous given their declared income and characteristics are identified with an ensemble of four unsupervised anomaly detection algorithms—Isolation Forest, Local Outlier Factor, One-Class Support Vector Machine and Elliptic Envelope—combined through a majority vote.

The ML-based estimates are compared with GMM estimates of a Complete Demand System (Lyssiotou et al. 2004; Fortin et al. 2009; Aktuna-Gunes et al. 2014, 2017) obtained from the same budget survey. The two approaches yield similar aggregate figures but rank self-employed and wage-earning households in opposite order: the ML ensemble flags the self-employed more often, whereas the demand system implies stronger under-reporting of wage income. Because the detection thresholds of the ML approach are partly calibrated on macroeconomic income shares, the aggregate agreement is treated as a consistency check rather than as independent validation, and the divergence between groups is interpreted as reflecting different margins of informality. In addition, the household-level labels produced by the ML framework make it possible to describe who is flagged, by gender, income, age, education and marital status.

The remainder of the paper is organized as follows. Section 2 outlines the methodological pipeline and Section 3 describes the data. Section 4 presents the statistical matching procedure, and Section 5 the anomaly-based detection of informal participation. Section 6 reports the ML estimates and compares them with the demand-system benchmark. Section 7 profiles flagged households, Section 8 discusses limitations, and Section 9 concludes.

2. Methodological pipeline of the proposed approach

The proposed pipeline estimates informal participation from survey microdata using established ML techniques. It consists of five stages:

Stage 1 – Feature construction and dimensionality reduction. A key limitation of conventional statistical matching is the restricted overlap of variables between time-use and budget surveys, typically limited to basic socio-economic indicators such as gender, age and education. The proposed approach widens the matching basis to 32 common socio-economic, asset and income variables. For each activity, PCA extracts the most informative components—termed *behavioral instruments*—separately in each survey.

Stage 2 – Cluster-based time imputation and monetization. K-means clustering groups households with similar characteristics in each survey and for each activity. TUS and HBS clusters are linked through an assignment rule defined on the common variables, and the average time allocated to the activity in each TUS cluster is imputed to the linked HBS cluster. Imputed time is converted into monetary values using the 2011 minimum wage per minute, and represents the non-market component of consumption.

Stage 3 – Full expenditure and anomaly detection. For each activity, full expenditure is the sum of monetary expenditure and monetized time. Four anomaly detection algorithms are then applied to identify households whose full expenditure is atypically high given their declared income and characteristics, and a majority vote produces a binary participation indicator.

Stage 4 – Structural benchmark. In parallel, the size of the informal economy is estimated with the structural micro-econometric model of Aktuna-Gunes et al. (2014, 2017), based on a Complete Demand System for monetary and full expenditures estimated by GMM.

Stage 5 – Comparison. The ML estimates (Stage 3) are compared with the structural estimates (Stage 4) and placed against external benchmarks.

Generative AI tools (Claude, Anthropic) were used to assist in revising the manuscript, including restructuring sections, drafting parts of the text and checking its internal consistency. The author reviewed and verified all content and takes full responsibility for the final text.

3. Data

The study uses two nationally representative surveys conducted by the Turkish Statistical Institute (TURKSTAT): the 2014–2015 Time Use Survey (TUS) and the 2011 Household Budget Survey (HBS).

The TUS covers 8,759 households. Individuals aged 15 and over completed time diaries for one weekday and one weekend day, recording activities at 10-minute intervals over 24 hours. The analysis file contains 103,968 observations. Daily time use is aggregated into eight composite activity categories that mirror the consumption categories of the HBS:

(1) Food (food preparation and related household tasks);

(2) Personal care and health (hygiene, care of household members, use of health services);

(3) Housing (cleaning, gardening, maintenance);

(4) Clothing (laundry, ironing);

(5) Education (studying and childcare);

(6) Transport (travel related to household activities);

(7) Leisure (social life, entertainment, hobbies, media use);

(8) Other (residual unpaid activities).[1]

In addition to time-use variables, the TUS contains approximately 80 socio-demographic and household characteristics, such as gender, age, marital status, education, employment status, health condition, household assets (e.g., refrigerator, vehicle, phone) and sources of income.

The 2011 HBS provides detailed monthly consumption data for about 1,104 households per month, that is, 13,248 households over the year. Its consumption categories correspond to the eight time-use activities listed above, which makes integration through statistical matching possible.

The two surveys refer to different years. The TUS is available only for 2006 and 2014–2015, and the 2011 HBS is the wave on which the demand-system benchmark (Section 6.2) is estimated, which keeps the ML and structural estimates comparable. Matching across years assumes that the relationship between household characteristics and time allocation is stable over the period, an assumption made plausible by the slow evolution of time-use patterns but not testable here.

4. Statistical matching through dimensionality reduction and clustering

4.1 Variable selection and principal components

Dimensionality reduction is a key preprocessing step for high-dimensional survey data: it lowers computational cost, limits overfitting and improves interpretability (Jolliffe and Cadima 2016). Two families of methods exist. Feature selection—through filter, wrapper or embedded methods (Guyon and Elisseeff 2003)—retains the most relevant original variables, whereas feature extraction, such as PCA, projects the data onto orthogonal dimensions that capture the largest share of variance (Jolliffe 2002). Linear projections may lose information when relationships are strongly non-linear (Van der Maaten et al. 2009), which is why both steps are combined here.

[1]Paid employment and job search are market activities, not household production. They are therefore excluded from the monetized time value used to construct full expenditure.

First, 32 variables common to the TUS and the HBS are identified. Second, for each activity, a subset of variables is selected according to its ability to capture behavioral and socio-economic variation relevant to that activity. Clustering within each survey uses the selected common variables together with survey-specific behavioral variables—time use in the TUS and expenditures in the HBS—which explains why the number of variables differs between the two surveys (Table 1). PCA is then applied to each subset, and the retained components are used for clustering.

Table 1. Number of variables used for clustering, by activity and survey.

Survey	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
TUS	15	23	18	20	19	22	18	19
HBS	25	26	26	25	26	26	26	25

Note: PCH = personal care and health. Variables are drawn from 32 common variables plus survey-specific behavioral variables (see Section 4.1).

4.2 Clustering and choice of the number of clusters

K-means clustering is performed separately for each activity and each survey. Candidate numbers of clusters were compared using the average silhouette coefficient, which measures intra-cluster cohesion relative to inter-cluster separation (Rousseeuw 1987). A common value of K = 20 was retained for all activities and both surveys, to keep the matching structure identical across activities. With K = 20, the average silhouette coefficient exceeds the conventional threshold of 0.5 in all cases except the "Other" category of the HBS (0.30) (Table 2).

Silhouette values must nevertheless be read together with cluster sizes (Tables 3 and 4). Some partitions are highly unbalanced: in the TUS Clothing category, one cluster contains 90,784 of the 103,968 observations, and in the HBS Education category one cluster contains 7,417 of the 9,304 households; conversely, several HBS clusters contain fewer than ten households. Under such imbalance, a high silhouette value (e.g., 0.92 for TUS Clothing) partly reflects the separation of a few small groups from one dominant group rather than a rich cluster structure. The estimates should therefore be read as conditional on K = 20.

Table 2. Average silhouette coefficient by activity and survey (K = 20).

Survey	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
TUS	0.57	0.63	0.66	0.92	0.58	0.60	0.53	0.73
HBS	0.50	0.53	0.51	0.54	0.75	0.58	0.50	0.30

Note: values above 0.5 are conventionally considered to indicate acceptable cluster quality. High values in very unbalanced partitions (e.g., TUS Clothing, HBS Education) should be read together with cluster sizes in Tables 3 and 4.

Table 3. Number of TUS observations in each cluster, by activity.

Cluster	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
1	11,456	9,160	11,714	90,784	7,980	1,342	7,323	3,027
2	4,682	3,557	3,195	410	5,094	12,860	8,139	2,272
3	8,041	1,824	5,556	679	1,606	4,789	7,357	45,805
4	1,022	12,959	835	13	8,359	9,962	4,054	4,608
5	11,551	56	5,825	78	8,316	9,204	2,931	1,167
6	6,464	4,982	146	1,220	100	242	1,182	7,576

Cluster	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
7	2,274	11,854	30,691	468	12,830	4,158	8,103	2,065
8	4,822	12,644	6,219	1,541	2,776	96	6,164	4,139
9	2,655	7,566	1,104	324	425	3,331	6,134	454
10	238	6	2,565	502	10,864	7,453	8,922	2,769
11	6,471	563	390	16	2,393	8,841	6,351	2,008
12	10,471	317	3,349	547	7,475	7,275	6,369	6,862
13	11,030	5,063	645	84	46	8,890	1,284	3,868
14	2,426	10,827	7,793	1,454	1,202	7,532	1,986	2,608
15	2,875	4,999	8,767	1,318	3,831	965	7,498	2,910
16	618	57	6,255	1,174	7,045	498	252	177
17	3,386	2,895	15	1,066	9,455	6,739	4,515	1,754
18	1,320	4,534	4,026	572	8,844	7,704	5,782	1,660
19	3,625	976	1,407	1,670	5,114	201	2,699	3,085
20	8,541	9,129	3,471	48	213	1,886	6,923	5,154
Total	**103,968**	**103,968**	**103,968**	**103,968**	**103,968**	**103,968**	**103,968**	**103,968**

Note: counts refer to observations in the TUS analysis file (see Section 3 on the unit of observation).

Table 4. Number of HBS households in each cluster, by activity.

Cluster	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
1	924	2,333	261	1,145	7,417	3,503	288	3,932
2	239	167	925	221	2	45	4,446	23
3	871	4	8	21	18	1	5	109
4	20	1	852	3	61	146	58	1
5	49	860	44	655	382	13	16	6
6	1,087	20	109	3,295	11	2,115	148	2
7	158	632	531	56	2	293	525	268
8	499	1,666	969	1	28	79	678	52
9	6	46	21	283	15	823	99	2
10	1	263	9	7	6	22	4	6
11	926	4	1,051	44	160	5	1,455	71
12	358	1,458	762	164	39	52	9	534
13	686	6	166	5	691	173	6	2,987
14	1,097	1	630	376	28	35	891	5
15	42	109	89	824	6	470	26	762
16	87	10	475	89	16	60	178	11
17	195	17	359	493	3	8	9	256
18	655	79	1,022	14	9	67	425	77
19	355	404	27	120	294	1,283	2	27
20	1,049	1,224	994	1,488	116	111	36	173
Total	**9,304**	**9,304**	**9,304**	**9,304**	**9,304**	**9,304**	**9,304**	**9,304**

Note: matched and trimmed sample of 9,304 households.

4.3 Linking clusters across surveys and imputing time

Because K-means labels are arbitrary, the clusters obtained separately in the two surveys must be linked through an explicit rule. For each activity a, the centroid of every TUS cluster and every HBS cluster is computed on the standardized common variables—the only space in which the two surveys are comparable. Let d_{jk} denote the Euclidean distance between the centroid of TUS cluster j and that of HBS cluster k. TUS and HBS clusters are linked one-to-one by solving the assignment problem that minimizes the total distance $\Sigma_j d_{j,\sigma(j)}$ over all permutations σ, using the Hungarian algorithm (Kuhn 1955). Each household in HBS cluster k then receives the average time devoted to activity a in the linked TUS cluster.

Like any statistical matching procedure, this imputation relies on the conditional independence assumption (CIA): given the common variables, time use and monetary expenditure are assumed to be independent (D'Orazio et al. 2006; Rässler 2002). The richer the set of common variables, the more plausible the CIA, which is the main motivation for using 32 variables rather than the handful of demographic variables used in conventional matching.

Imputed time is monetized by multiplying it by the per-minute value of the 2011 minimum wage. This opportunity-cost valuation at the minimum wage is a conservative lower bound; replacement-cost valuations based on the wages of specialized workers would yield higher values (Ironmonger 1996). Table 5 reports descriptive statistics for monetized time. Because imputation assigns the cluster mean to all members of a cluster, monetized time takes at most 20 distinct values per activity, which explains the repeated quantiles in Table 5. Consequently, imputed time contributes to differences between clusters but not within them; this is why anomaly detection (Section 5) operates on the joint vector of the eight activities, in which each household's combination of cluster memberships is specific.

Table 5. Descriptive statistics of monetized time, by activity.

Statistic	Food	PCH	Housing	Clothing	Education	Transport	Leisure	Other
Mean	126.18	48.60	49.49	1.11	122.28	59.70	303.72	154.33
Std. dev.	12.21	7.70	18.12	1.63	7.80	9.78	27.18	7.58
Min	99.39	36.23	35.86	0.00	96.86	49.48	254.54	90.46
25th pct.	112.21	43.44	40.14	0.00	125.80	49.48	288.07	154.62
Median	125.47	46.71	41.05	0.71	125.80	57.33	288.07	154.62
75th pct.	129.81	51.59	52.74	1.35	125.80	67.08	316.41	154.62
Max	152.15	148.80	119.87	8.25	142.41	96.79	353.32	200.31

Note: N = 9,304 households. Values in Turkish lira per month, at the 2011 minimum wage. Identical quantiles reflect cluster-mean imputation.

Finally, households in the bottom and top 1% of the declared income distribution are excluded to limit the influence of recording errors, leaving a matched sample of 9,304 households. Because very low declared income combined with substantial consumption is itself a typical signal of undeclared income, the trimming may also exclude some households with undeclared income.

5. Detecting informal participation

5.1 Identifying idea

The detection strategy follows the expenditure-based tradition initiated by Pissarides and Weber (1989): households that receive undeclared income spend more than their declared income and observable characteristics would predict. Hurst et al. (2014) show that this gap is particularly large for the self-employed. Here, the gap is detected without a parametric Engel curve: the feature vector of each household includes its declared income level and composition, its socio-demographic and asset characteristics, and its full expenditure for each of the eight activities.[2] A household that lies in a low-density region of this joint space—typically because its full expenditure is high relative to households with similar declared income and characteristics—is treated as anomalous. Including the value of domestic time matters because households substitute time-intensive home production for market purchases: monetary expenditure alone therefore misrepresents the resources a household actually commands (Aktuna-Gunes et al. 2014, 2017).

An anomaly is a statistical signal, not proof of informal activity. Remittances, public or private transfers, dissaving, borrowing, purchases of durable goods and measurement error can all raise expenditure relative to declared income. The resulting flags should therefore be read as indicators of likely undeclared income, and estimates for groups in which such alternative explanations are frequent—such as retirees—as upper bounds.

5.2 Algorithms and majority vote

Four established unsupervised algorithms are applied: Isolation Forest (Liu et al. 2008), Local Outlier Factor (Breunig et al. 2000), One-Class Support Vector Machine (Schölkopf et al. 2001) and Elliptic Envelope, which relies on a robust Mahalanobis distance based on the Minimum Covariance Determinant estimator (Rousseeuw and Van Driessen 1999). Their specifications and hyperparameters are given in Appendix A. Because the algorithms rely on different notions of abnormality—path length in random partitions, local density, support of the distribution and distance to a robust centre—their combination reduces dependence on any single assumption. A household is classified as an informal participant only if at least three of the four algorithms flag it.

Let $S_i \in \{-1, 1\}$ denote the label assigned by algorithm i to household h, with $S_i = -1$ if the household is flagged as anomalous and $S_i = 1$ otherwise. The total score is

$$TS_h = \Sigma_{i=1}^{4} S_{ih} \tag{1}$$

and the participation indicator is

$$D_h = 1 \text{ if } TS_h \leq -2 \text{ (flagged by at least three algorithms)}; \quad D_h = 0 \text{ otherwise.} \tag{2}$$

Note that $TS_h = -2$ when exactly three algorithms flag the household and $TS_h = 0$ when exactly two do; the threshold in equation (2) therefore implements the three-out-of-four rule.

[2]The detection feature vector comprises: employment type (5 levels), income group (5 levels), gender, age group, marital status (4 levels), education (5 levels), home ownership, number of rooms, ownership of durable goods (TV, cable TV, oven, washing machine, freezer, car, computer, refrigerator, dryer, carpet, air conditioning, camera, DVD/CD player; binary each), number of mobile phones, type of employment contract (3 levels), the shares of self-employment, wage and other income in total declared income, and full expenditure for each of the eight activities.

5.3 Calibration of detection thresholds

Each algorithm requires a contamination parameter (ν for the One-Class SVM) that sets the expected share of anomalies. Standard anomaly detection assumes anomalies are rare, whereas informality in Turkey concerns a large share of households. The contamination parameters are therefore set separately for self-employed and wage-earning households and calibrated on the contribution of these groups' incomes to national income.

This calibration has a direct implication for interpretation. Because the thresholds embed external information on income shares, the aggregate flagged shares are shaped in part by the calibration and cannot be regarded as fully independent estimates. The main informational content of the ML approach lies instead in *which* households are flagged, in the ranking of groups, and in the resulting profiles. Re-estimating the model with a common, uncalibrated contamination value would show how much the results depend on this calibration; this check is left for further work.

5.4 Estimands

The *incidence* of informal participation in group g is the weighted share of flagged households:

$$P_g = \Sigma_{h \in g} w_h D_h / \Sigma_{h \in g} w_h \quad (3)$$

where w_h is the survey weight.

Estimates are reported for self-employed and wage-earning households. Because self-employed households are a minority of working households, the incidence for all working households—the population-weighted average of the two group rates—lies below their simple average.

6. Results

6.1 ML-based estimates

Table 6 reports the incidence of informal participation produced by the anomaly detection ensemble. The share of flagged households is substantially higher among the self-employed (61.27%) than among wage earners (33.94%). This finding is consistent with evidence that self-employment provides more room for unreported or partially reported transactions, particularly where regulation and enforcement are limited (Hurst et al. 2014; Perry et al. 2007).

The simple average of the two group rates is 47.61%. However, self-employed households represent a minority of working households, so the population-weighted incidence for all working households is lower. When non-working households, including retirees, are added, the flagged share reaches 48.60%. This figure should be interpreted with particular caution: for households without labor income, high expenditure relative to declared income is more likely to reflect transfers, remittances or dissaving than informal work, so it is best read as an upper bound.

Table 6. Incidence of informal participation, ML ensemble (%).

Group	Incidence P (%)

Group	Incidence P (%)
Self-employed households	61.27
Wage-earning households	33.94
Memo: simple average of the two groups	47.61
All households, including non-working	48.60

Note: a household is flagged when at least three of the four algorithms classify it as anomalous.

6.2 Benchmark: a Complete Demand System estimated by GMM

The size of the informal economy can also be estimated with a Complete Demand System (Lyssiotou et al. 2004; Fortin et al. 2009). The approach compares budget-share equations estimated for full expenditure (monetary expenditure plus imputed time value) and for monetary expenditure only; differences in the implied resources reveal undeclared income. The method requires careful variable selection to satisfy identification and specification conditions and is sensitive to functional form (Lewbel 1990; Banks et al. 1997), which motivates the search for complementary approaches.

The theoretical and econometric specification follows Aktuna-Gunes et al. (2017), who applied it to Turkish data for 2003–2006; it is summarized in Appendix B. Here it is estimated by GMM on the 2011 HBS; the estimation sample contains 8,307 households (Table B1). The system covers eight commodity groups, of which seven budget-share equations are estimated.[3]

Table 7 reports the results. The estimated informal share is 24.72% for the self-employed and 70.87% for wage earners. These figures combine the share of each income source in declared income with its estimated under-reporting factor (equation B4); the computation is detailed in Appendix B.1. The estimated factor is larger for wages ($\theta_s = 1.68$) than for self-employment income ($\theta_r = 1.27$), so the demand system attributes more under-reporting to wage earners than to the self-employed.

Table 7. Size of the informal economy, Complete Demand System estimated by GMM, 2011 (%).

Group	Estimate (%)
Self-employed households	24.72
Wage-earning households	70.87
Memo: simple average of the two groups	47.79

Note: specification of Aktuna-Gunes et al. (2017); see Appendix B.

6.3 Comparison of the two approaches and external benchmarks

The two approaches yield similar aggregate figures: the simple averages of the two group rates are 47.6% for the ML ensemble and 47.8% for the demand system. They nevertheless rank the two groups in opposite order. The ML ensemble flags self-employed households more often than wage-earning households (61.3% vs. 33.9%), whereas the demand system implies a larger informal share for wage earners than for the self-employed (70.9% vs. 24.7%). Three considerations qualify this comparison. First, the ML detection thresholds are partly calibrated on

[3]Because of the adding-up restriction, the parameters of the omitted category ('other goods and services') are recovered residually. See Appendix B.

income shares (Section 5.3), so the aggregate agreement is a consistency check rather than an independent validation. Second, the two methods do not measure the same quantity: the ML approach measures the incidence of households with anomalous full expenditure, whereas the demand-system figures combine the share of each income source with its estimated under-reporting factor. Third, the differences stem from the identifying assumptions of each approach—the functional form and theoretical restrictions of the demand system on the one hand, the choice of features and thresholds of the ML ensemble on the other.

From a labor economics perspective, the two results can be read as capturing different margins of informality. The higher incidence of anomalous households among the self-employed is consistent with their greater time flexibility: when households face resource constraints and high unemployment, informal production becomes a substitution mechanism through which households allocate more time to time-intensive goods and services outside the formal market, and such behavior is easier for individuals with flexible schedules than for wage earners bound by formal working hours. The larger under-reporting factor for wages, in turn, is consistent with partially declared wages, whereby registered employees receive part of their pay off the record; this margin affects declared income rather than consumption patterns and may therefore be better captured by the demand system.[4]

Both sets of estimates can also be placed against external benchmarks, keeping in mind that these refer to different concepts. TURKSTAT's Household Labour Force Survey defines informal employment as employment not registered with the social security system, while macro-level MIMIC estimates express the shadow economy as a share of GDP (Medina and Schneider 2018).

7. Socio-demographic profiling of flagged households

A practical advantage of the ML framework is that it produces a participation label for every household, which allows flagged households to be described along socio-demographic dimensions. Figures 1 to 4 classify households by the gender, income group, age, education and marital status of the household reference person; "men" and "women" below therefore refer to male- and female-headed households. Three caveats apply to this section. First, the profiles are descriptive and do not identify causal effects. Second, gender, age, education, marital status and income group are among the detection features, so the profiles partly reflect the structure of the inputs. Third, the figures display numbers of households, which depend on the size of each group in the sample; the comparisons below therefore also consider the ratio of flagged to non-flagged households within each group, and a multivariate model of D_h on all characteristics jointly would be needed to establish conditional associations.

7.1 Income and gender

Figure 1 shows the number of households in each of five income groups (0 = lowest, 4 = highest), by gender of the household reference person and participation status. Among men, the number of flagged households is highest in the upper income groups (3 and 4), and the highest income group is the only one in which flagged households outnumber non-flagged ones. Among women, the number of flagged households follows a U-shaped pattern: it is

[4]Domestic production accounts for the largest share of daily time in Turkish households. According to İlkkaracan and Gündüz (2009), its value represented between 25% and 45% of GDP in 2006, depending on the type of labor activity.

highest in the lowest and the highest income groups, and in the highest group flagged female-headed households also outnumber non-flagged ones. At every income level, flagged female-headed households are far less numerous than flagged male-headed households, reflecting the small share of female-headed households in the sample.

These patterns are consistent with the dual view of informality distinguishing *exit*—a voluntary choice, for instance to avoid taxes or regulation—from *exclusion* from formal employment (Perry et al. 2007; La Porta and Shleifer 2014). The high incidence at the top of the income distribution, for both men and women, may partly reflect the former, whereas flagged female-headed households in the lowest income group are more consistent with the latter. The data used here do not allow these motives to be distinguished directly, so this interpretation remains tentative.

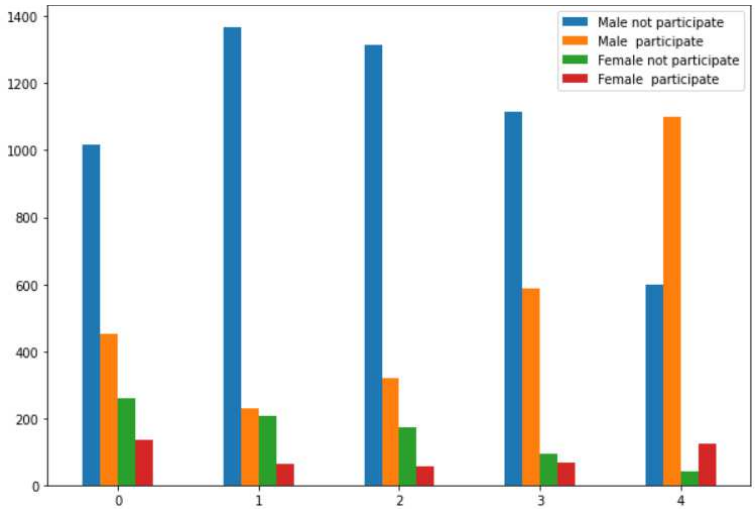


Fig. 1 Number of households by income group and gender of the household reference person, and by informal participation status

7.2 Age and gender

Figure 2 shows participation across 14 age groups (Group 1 = 18–19 years, Group 14 = 80 and over). Among men, the number of flagged households rises sharply at ages 25–29 (Group 3), peaks at ages 45–49 (Group 7), with similar levels at ages 35–44 (Groups 5 and 6), and remains high through ages 50–54 (Group 8) before declining after age 60 without disappearing. The dip at ages 30–34 (Group 4) affects flagged and non-flagged households alike and therefore reflects the age composition of the sample rather than a change in incidence. Among women, the age profile is flatter and at a much lower level, with peaks at 40–44 and 55–59 (Groups 6 and 9). After age 60, the gender gap narrows; given the alternative explanations discussed in Section 5.1, flags in this age range should not be interpreted as informal work without further evidence.

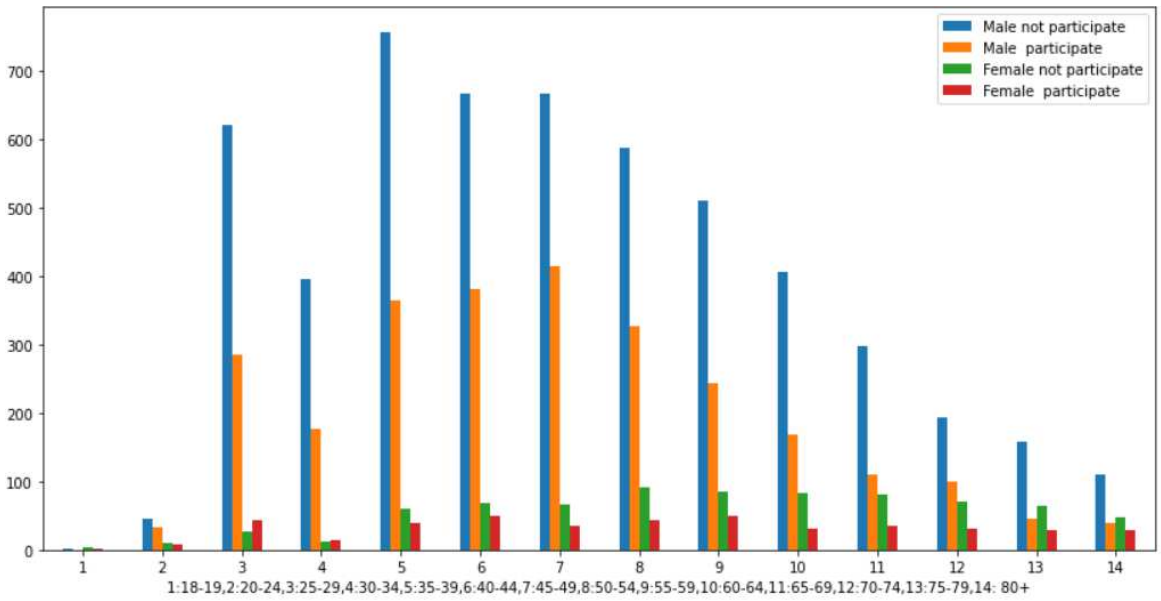


Fig. 2 Number of households by age group and gender of the household reference person, and by informal participation status

7.3 Education and gender

Figure 3 shows participation across five education levels (0 = literate without diploma, 1 = primary, 2 = lower secondary, including vocational schools, 3 = upper secondary, including technical schools, 4 = higher education). Among men, the largest number of flagged households is found at the primary level (category 1), the most populated category, but flagged men are also numerous in the upper categories, and in the highest category (4) flagged households outnumber non-flagged ones. Among women, flagged households are most numerous at the two lowest levels (categories 0 and 1), but in the highest category flagged female-headed households also outnumber non-flagged ones.

The relationship between education and informality thus appears non-monotonic for both genders. A flexible method can capture this pattern without imposing linearity, but whether it reflects high-skilled self-employment in loosely regulated professions or composition effects (e.g., income and occupation) would require conditional estimates that control for these characteristics jointly.

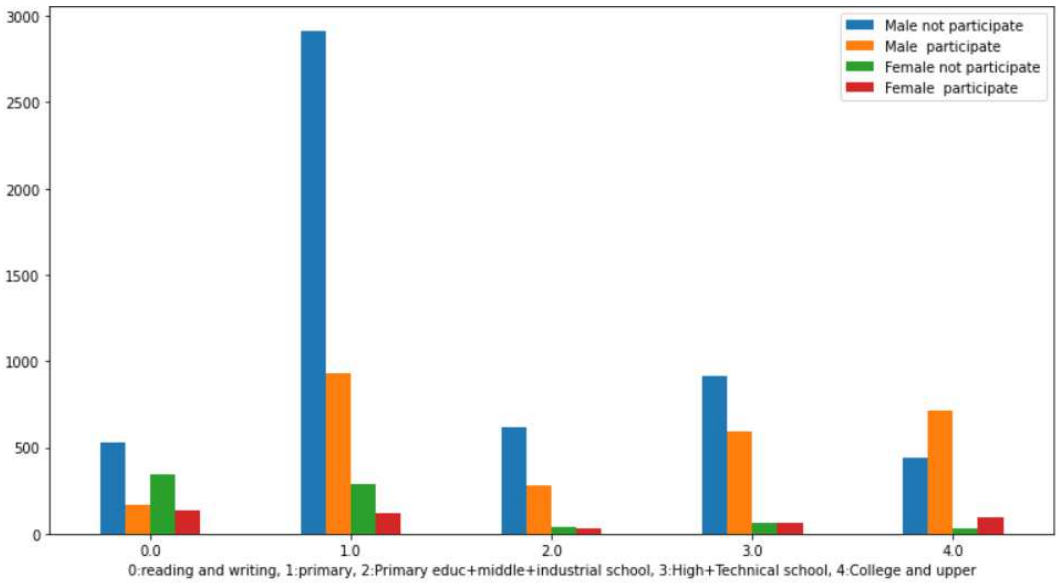


Fig. 3 Number of households by education level and gender of the household reference person, and by informal participation status

7.4 Marital status and gender

Figure 4 shows participation across four marital-status categories. Married men form by far the largest group, and roughly one in three married male household heads is flagged. Among women, the incidence is highest among widows, among whom flagged households slightly outnumber non-flagged ones, which is consistent with informal income substituting for the absence of a partner's income. Divorced women, the largest group of female household heads, show a lower incidence, with flagged households amounting to less than half of non-flagged ones. Never-married household heads are few, which makes their incidence difficult to interpret.

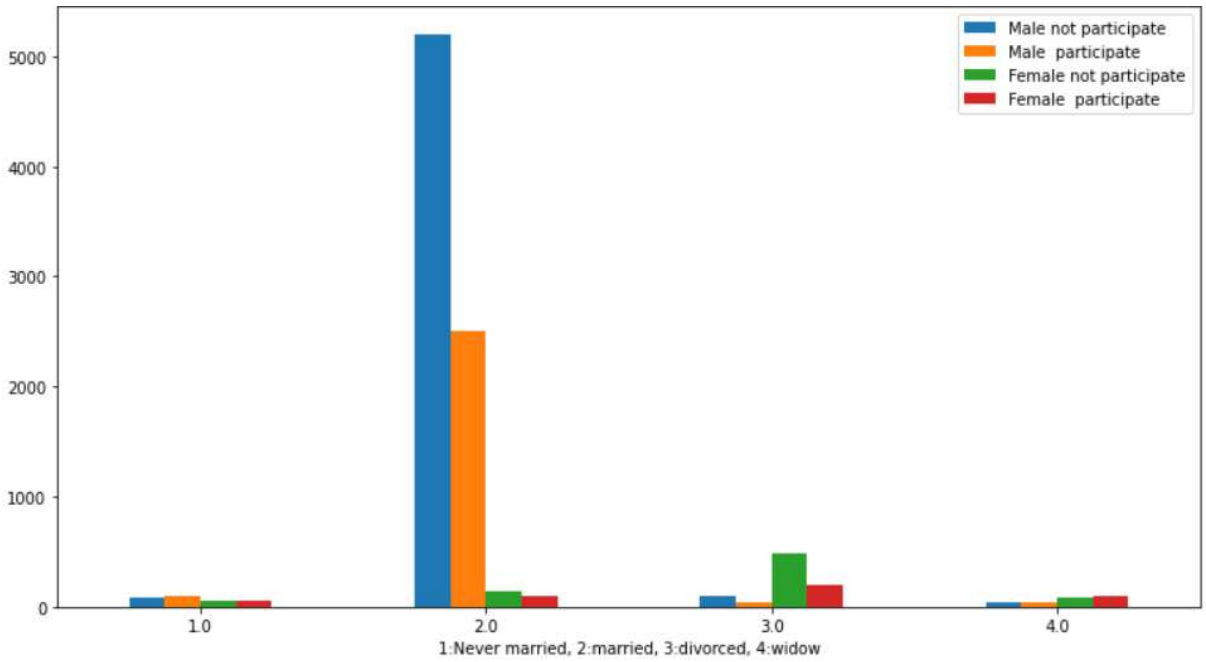


Fig. 4 Number of households by marital status and gender of the household reference person, and by informal participation status

8. Discussion and limitations

Several limitations qualify the results and point to extensions. First, statistical matching relies on the conditional independence assumption, which cannot be tested with the two surveys alone; a richer set of common variables makes it more plausible but does not guarantee it. Second, the surveys refer to different years, and the time-use structure is assumed to be stable between 2011 and 2014–2015. Third, cluster-mean imputation removes within-cluster variation in imputed time, which limits the contribution of time use to detection within clusters. Fourth, the contamination parameters are calibrated on macroeconomic income shares, so the aggregate rates are not independent of external information. Fifth, anomalous expenditure can arise from causes other than informal income, which makes the flags upper-bound indicators, especially for non-working households. Sixth, the sensitivity of the results to the main tuning choices—the number of clusters, the voting rule and the contamination parameters—has not yet been assessed. Finally, no ground-truth label of informality is available in the budget survey.

These limitations suggest a research agenda: applying the pipeline to the 2003–2006 data for which demand-system estimates already exist (Aktuna-Gunes et al. 2017), so that the two approaches can be compared on identical samples; using administrative or survey-based registration information as partial labels; and extending the framework to other countries with time-use and budget surveys, such as those covered by the project funding this research.

9. Conclusion

This paper proposes a household-level framework for measuring informal economic participation that combines time-use and budget microdata with unsupervised learning, and confronts its results with a structural demand-system benchmark. The contribution is twofold. First, the Turkish Time Use Survey (2014–2015) and Household Budget Survey (2011) are statistically matched through PCA and K-means clustering on more than 30 common variables, with an explicit rule linking clusters across surveys. This widens the matching basis beyond the few demographic variables used in conventional matching and allows the value of unpaid domestic work to be included in household full expenditure.

Second, an ensemble of four anomaly detection algorithms—Isolation Forest, Local Outlier Factor, One-Class SVM and Elliptic Envelope—identifies households whose full expenditure is anomalous given their declared income and characteristics, an expenditure-based signal of undeclared income. The resulting incidence of informal participation is 61.3% among the self-employed and 33.9% among wage earners, whereas a GMM-estimated Complete Demand System on the same survey yields 24.7% and 70.9%, respectively. The two approaches thus agree at the aggregate level—simple averages of 47.6% and 47.8%—but rank the two groups in opposite order. Because the ML thresholds are partly calibrated on macroeconomic income shares, the aggregate agreement supports the consistency of the two approaches rather than constituting an independent validation, and the divergence suggests that they capture different margins of informality: anomalous consumption behavior, which is more frequent among the self-employed, and under-reported wage income.

Beyond aggregate measurement, the household-level labels allow informality to be profiled. In Turkey in 2011, flagged male-headed households are most numerous among married men aged 35–49, and incidence is highest in the top income and education groups, where flagged households outnumber non-flagged ones. Flagged female-headed households are concentrated at the lowest income and education levels, but incidence is also high at the top of both distributions and among widows. These descriptive patterns are consistent with the coexistence of voluntary and necessity-driven informality. The framework is transparent and replicable, can be extended with additional variables, and is readily transferable to other countries that conduct time-use and budget surveys.

Appendix A. Anomaly detection algorithms

Isolation Forest (Liu et al. 2008) isolates observations by recursive random partitioning of the feature space. Anomalies require fewer splits to be isolated. The anomaly score of observation x is $s(x, n) = 2^{-E[h(x)]/c(n)}$, where $E[h(x)]$ is the average path length of x over the trees and $c(n)$ the average path length of an unsuccessful search in a binary search tree of n observations.

Local Outlier Factor (Breunig et al. 2000) compares the local reachability density of an observation with that of its k nearest neighbors. Values substantially above 1 indicate that the observation lies in a sparser region than its neighbors.

One-Class Support Vector Machine (Schölkopf et al. 2001) estimates a function that is positive on a region containing most of the data and negative elsewhere, by separating the data from the origin with maximum margin in a kernel-induced feature space. The parameter ν is an upper bound on the share of training observations treated as outliers.

Elliptic Envelope fits a robust estimate of location and covariance with the Minimum Covariance Determinant estimator (Rousseeuw and Van Driessen 1999) and flags observations with the largest robust Mahalanobis distances. It assumes an approximately elliptical distribution of the data.

All features are standardized before estimation.

Appendix B. Complete Demand System specification

The unit cost function is assumed to take the following quadratic logarithmic form (Lewbel 1990):

$$\ln C(U, p) = \ln a(p) + b(p)\, U / [1 - g(p)\, U] \tag{B1}$$

where $a(p)$, $b(p)$ and $g(p)$ are functions homogeneous in prices p, and U is the household's utility level. The Hicksian budget shares are

$$w_i = a_i(p) + b_i(p)\, U/[1 - g(p)U] + \lambda_i(p)\, \{U/[1 - g(p)U]\}^2 \tag{B2}$$

where $a_i(p) = \partial \ln a(p)/\partial \ln p_i$, $b_i(p) = \partial \ln b(p)/\partial \ln p_i$ and $\lambda_i(p) = b_i(p)\, \partial \ln g(p)/\partial \ln p_i$. To obtain the system of Engel curves, base-period prices are normalized to one ($p = r = 1$), and the subscript h denotes households. Using equation (B1), $U/(1 - g_0 U) = (\ln Y^*_h - a_0)/b_0$, where Y^*_h is total (true) income and a_0, b_0 and g_0 are the values of the corresponding functions at $p_i = r_i = 1$. Substituting into (B2) gives

$$w_{ih} = \alpha_i + \beta_i\, (\ln Y^*_h - a_0) + \delta_i\, (\ln Y^*_h - a_0)^2 \tag{B3}$$

with parameters α, β and δ. Equation (B3) is the quadratic Engel curve derived from the PIGLOG-type cost function.

True income Y^* is decomposed into three sources, denoted a, s and r, which correspond respectively to other income sources, wages and self-employment income. True income is the weighted sum of the reported amounts of these three sources:

$$Y^*_h = \theta_a\, Y_{ah} + \theta_s\, Y_{sh} + \theta_r\, Y_{rh} \tag{B4}$$

Equation (B4) states that true income equals the sum of the observed incomes (Y_a, Y_s, Y_r) multiplied by their corresponding factors (θ_a, θ_s, θ_r), where θ_r, $\theta_s \geq 1$ (under-reporting) and $\theta_a = 1$ (other incomes are correctly observed). Estimating θ_r and θ_s therefore yields the size of the underground economy and the saving tendencies associated with the under-reported part of declared income. To impose the constraints on θ_r and θ_s, Fortin et al. (2009) express each factor as $1 + e^k$, where k is a parameter estimated by the model. The true values of self-employment and wage income are thus $Y^*_r = (1 + e^k)Y_r$ and $Y^*_s = (1 + e^l)Y_s$.

Each income source is then expressed as a share of reported total income, $y_m = Y_m/Y$, where Y is the sum of wages, self-employment income and other sources such as fees and government transfers. Following Aktuna-Gunes et al. (2014), based on Banks et al. (1997), all goods and services are considered in a quadratic demand system:

$$w_{ih} = \alpha_i + \Sigma_j \alpha_{ij} Z_{jh} + \beta_{1i} [\ln Y_h + \ln(\Sigma_{m=a,s,r} \theta_m y_{mh})] + \beta_{2i} [\ln Y_h + \ln(\Sigma_{m=a,s,r} \theta_m y_{mh})]^2 + e_{ih} \quad \text{(B5)}$$

where w_{ih} is the budget share of good i, Z_h is the vector of household characteristics, which accounts for the heterogeneity of preferences, and y_{mh} are the three income components. Since $\ln Y_h + \ln(\Sigma_m \theta_m y_{mh}) = \ln Y^*_h$, equation (B5) is the empirical counterpart of (B3), with true income written in terms of reported income and the under-reporting factors. Distinguishing income sources matters because households from different social groups cannot be expected to react in the same way, in terms of consumption and saving, to different types of income, especially when these revenues are uncertain.

B.1 Computation of the size of the informal economy

The informal shares reported in Table 7 are obtained by multiplying the share of each income source in declared income (0.194 for self-employment income and 0.423 for wages) by its estimated under-reporting factor (θ_r = 1.2745 and θ_s = 1.6755; Table B1):

(1) Self-employed: $0.194 \times 1.2745 \approx 0.247$, i.e., 24.7%;

(2) Wage earners: $0.423 \times 1.6755 \approx 0.709$, i.e., 70.9%.

Table B1. GMM estimates of the full-income demand system (equation B5).

Dep. Variable:['wf1', 'wf2', 'wf3', 'wf4', 'wf5', 'wf6', 'wf7']				Hansen J:		38.68
Model:				GMMREM	Prob (Hansen J):	1.00
Method:				GMM		
Date:			Sat, 15 Jul 2023			
Time:				12:51:13		
No. Observations:				8307		
	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.0451	1.449	-3.482	0.000	-7.885	-2.205
gender	-1.8294	0.464	-3.945	0.000	-2.738	-0.920
age	0.1352	3.367	0.040	0.968	-6.464	6.734
marital	1.0630	0.528	2.012	0.044	0.027	2.099
educ	7.1980	2.063	3.489	0.000	3.155	11.241
lny_iv	-0.4889	0.146	-3.346	0.001	-0.775	-0.203
ya	-0.1863	10.132	-0.018	0.985	-20.045	19.672
ys	1.6755	2.584	2.454	0.014	1.277	11.407
yr	1.2745	2.602	1.759	0.029	-0.523	9.677
log_mm_f	0.0002	1.47e-05	15.418	0.000	0.000	0.000
log_mm_hpc	8.49e-05	5.44e-06	15.615	0.000	7.42e-05	9.56e-05
log_mm_h	5.324e-06	3.55e-07	15.007	0.000	4.63e-06	6.02e-06
log_mm_c	-1.645e-05	9.94e-07	-16.543	0.000	-1.84e-05	-1.45e-05
log_mm_e	4.589e-05	2.96e-06	15.514	0.000	4.01e-05	5.17e-05
log_mm_t	0.0002	1.04e-05	14.476	0.000	0.000	0.000
log_mm_l	0.8384	0.317	2.643	0.008	0.217	1.460

References

Aguiar, M., and E. Hurst. 2007. “Measuring Trends in Leisure: The Allocation of Time over Five Decades.” *Quarterly Journal of Economics* 122 (3): 969–1006. https://doi.org/10.1162/qjec.122.3.969.

Aktuna-Gunes, A. T. 2021. “Testing Supervised and Unsupervised Machine Learning Methods for Time Use Matching.” Paper presented at the IATUR Annual Conference.

Aktuna-Gunes, A. T., F. Gardes, and C. Starzec. 2017. “Informal Markets, Domestic Production and Demand Elasticities: A Case Study for Turkey.” *Economics Bulletin* 37 (3): 1496–1513.

Aktuna-Gunes, A. T., C. Starzec, and F. Gardes. 2014. “Une évaluation de la taille de l’économie informelle par un système complet de demande estimé sur données monétaires et temporelles.” *Revue économique* 65 (4): 567–590. https://doi.org/10.3917/reco.654.0567.

Athey, S., and G. W. Imbens. 2019. “Machine Learning Methods That Economists Should Know About.” *Annual Review of Economics* 11: 685–725.

Banks, J., R. Blundell, and A. Lewbel. 1997. “Quadratic Engel Curves and Consumer Demand.” *Review of Economics and Statistics* 79 (4): 527–539. https://doi.org/10.1162/003465397557015.

Bonke, J. 1992. “Distribution of Economic Resources: Implications of Including Household Production.” *Review of Income and Wealth* 38 (3): 281–293.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander. 2000. “LOF: Identifying Density-Based Local Outliers.” *ACM SIGMOD Record* 29 (2): 93–104. https://doi.org/10.1145/335191.335388.

D’Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester: Wiley.

Eurostat. 2019. *Guidance on the Implementation of Household Production in National Accounts*. Luxembourg: Eurostat.

Fortin, B., G. Lacroix, and D. Pinard. 2009. “Évaluation de l’économie souterraine au Québec : une approche micro-économétrique.” *Revue économique* 60 (5): 1257–1274.

Frazis, H., and J. Stewart. 2011. “How Does Household Production Affect Measured Income Inequality?” *Journal of Population Economics* 24 (1): 3–22.

Guyon, I., and A. Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research* 3: 1157–1182.

Hurst, E., G. Li, and B. Pugsley. 2014. “Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-Employed.” *Review of Economics and Statistics* 96 (1): 19–33.

ILO. 2018. *Women and Men in the Informal Economy: A Statistical Picture*. 3rd ed. Geneva: International Labour Office.

İlkkaracan, İ., and U. Gündüz. 2009. “Time-Use, the Value of Non-Market Production and Its Interactions with the Market Sector: The Case of Turkey.” Paper presented at the International Conference on Inequalities and Development in the Mediterranean Countries. Mimeo.

Ironmonger, D. 1996. “Counting Outputs, Capital Inputs and Caring Labor: Estimating Gross Household Product.” *Feminist Economics* 2 (3): 37–64.

Jolliffe, I. T. 2002. *Principal Component Analysis*. 2nd ed. New York: Springer.

Jolliffe, I. T., and J. Cadima. 2016. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions of the Royal Society A* 374 (2065): 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kuhn, H. W. 1955. “The Hungarian Method for the Assignment Problem.” *Naval Research Logistics Quarterly* 2 (1–2): 83–97.

La Porta, R., and A. Shleifer. 2014. “Informality and Development.” *Journal of Economic Perspectives* 28 (3): 109–126.

Lewbel, A. 1990. “Full Rank Demand Systems.” *International Economic Review* 31 (2): 289–300.

Liu, F. T., K. M. Ting, and Z.-H. Zhou. 2008. “Isolation Forest.” In *Proceedings of the Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17.

Lyssiotou, P., P. Pashardes, and T. Stengos. 2004. “Estimates of the Black Economy Based on Consumer Demand Approaches.” *Economic Journal* 114 (497): 622–640.

Medina, L., and F. Schneider. 2018. “Shadow Economies Around the World: What Did We Learn Over the Last 20 Years?” IMF Working Paper 18/17. Washington, DC: International Monetary Fund.

Mullainathan, S., and J. Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106.

Perry, G. E., W. F. Maloney, O. S. Arias, P. Fajnzylber, A. D. Mason, and J. Saavedra-Chanduvi. 2007. *Informality: Exit and Exclusion*. Washington, DC: World Bank.

Pissarides, C. A., and G. Weber. 1989. “An Expenditure-Based Estimate of Britain’s Black Economy.” *Journal of Public Economics* 39 (1): 17–32.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics 168. New York: Springer.

Rousseeuw, P. J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rousseeuw, P. J., and K. Van Driessen. 1999. “A Fast Algorithm for the Minimum Covariance Determinant Estimator.” *Technometrics* 41 (3): 212–223.

Schneider, F., and D. H. Enste. 2000. “Shadow Economies: Size, Causes, and Consequences.” *Journal of Economic Literature* 38 (1): 77–114.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. 2001. “Estimating the Support of a High-Dimensional Distribution.” *Neural Computation* 13 (7): 1443–1471.

Turkish Statistical Institute (TURKSTAT). 2011. *Household Budget Survey 2011*. Microdata. Ankara: TURKSTAT.

Turkish Statistical Institute (TURKSTAT). 2015. *Time Use Survey 2014–2015*. Microdata. Ankara: TURKSTAT.

Ulyssea, G. 2020. “Informality: Causes and Consequences for Development.” *Annual Review of Economics* 12: 525–546.

Van der Maaten, L., E. Postma, and J. van den Herik. 2009. “Dimensionality Reduction: A Comparative Review.” Technical Report TiCC-TR 2009-005. Tilburg: Tilburg University.