

American Journal of Economics and Business Management

Vol. 9 Issue 11 | **pp.** 685-695 | **ISSN:** 2576-5973
Available online @ https://www.globalresearchnetwork.us/index.php/ajebm

Article

Modelling and Reducing Informal Agricultural Employment in Uzbekistan: An Econometric Diagnostic Testing

Mamasoliyev Sarvar[1], Ishnazarov Akram[2]

1. Doctoral student, Tashkent State University of Economics, Tashkent, Uzbekistan
2. Associate Professor, Department of Econometrics, Tashkent State University of Economics
*Correspondence: sarvaruser007@gmail.com[1], akram.ishnazarov@tsue.uz[2]

Abstract: Work on household plots is the single most rural form of informal employment in Uzbekistan: in 2024 it engaged 1.50 million people, or 11.7% of everyone employed inside the country. This article asks what drives this form of work across districts and how much of it could realistically move into formal employment. Using the official labour resources balance for 205 districts and cities in 2022–2024, it builds a six-phase modelling framework in which every estimator is preceded by the diagnostic test that justifies it. Panel models are selected with F, Breusch–Pagan LM and Mundlak tests and checked for serial correlation, cross-sectional dependence and a structural break; the 2024 cross-section model of plot employment is tested for heteroskedasticity, functional form, normality and influential observations and then extended with quantile regression, a Hansen threshold model, an Oaxaca–Blinder decomposition and a cross-validated random forest. Districts with more formal jobs in enterprises and more registered self-employed people have markedly less plot employment: one additional firm job per 100 labour resources is associated with 0.39 percentage points less plot employment, and one additional registered self-employed person with 0.64 points less. The difference in firm employment explains 62% of the 11.9-point gap between rural districts and cities. Scenario simulations suggest that a combination of firm job creation, digital self-employment registration and a simplified social tax for plot workers could reduce the plot share from 11.7% to about 6.3%, moving roughly 695,000 people into covered work.

Keywords: Informal employment, agriculture, household plots, human capital, digital registration, panel data, diagnostic tests, quantile regression, threshold regression, Oaxaca–Blinder decomposition, random forest, Uzbekistan.

Citation: Sarvar M. Modelling and Reducing Informal Agricultural Employment in Uzbekistan: An Econometric Diagnostic Testing. American Journal of Economics and Business Management 2026, 9(11), 685-695.

Received: 10th Aug 2026
Revised: 11th Sep 2026
Accepted: 11th Oct 2026
Published: 21th Nov 2026

Copyright: © 2026 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Informal employment is concentrated in agriculture almost everywhere it has been measured. The ILO's global estimates show that the agricultural sector has the highest informality rate of any sector, above 90% of agricultural employment worldwide, and that rural workers are about twice as likely to be informal as urban workers [1]. The literature offers two broad readings of this pattern. One sees informality as a survival activity of low-productivity workers who cannot find formal jobs and who will move once modern firms expand [2]. The other stresses that firms and workers choose informality when registration, taxes and contributions cost more than the protection they buy [3]. For policy the distinction matters: under the first reading informality falls with job creation, under the second it falls when formalisation becomes cheaper and simpler.

Uzbekistan offers a useful case for testing both readings with official data. Since 2022 the country has compiled an annual balance of labour resources that splits employment

into formal and informal components down to every district and city [4], [5]. The informal sector in this balance includes four groups, and one of them, people working on household plots (*tomorqa*), is almost entirely agricultural. In 2024 it accounted for 1.50 million people. A recent joint study by the Central Bank, CERR and UNDP shows that informal workers in Uzbekistan are disproportionately rural, have lower education and lower incomes, and rarely contribute to social insurance [6]. The plot economy is also where human capital and digital tools could change outcomes most cheaply: registration as self-employed has moved online, and the barrier for a plot worker is now less a matter of distance than of information and incentives.

Existing work on Uzbekistan has mostly described these patterns rather than modelled them, and where models are estimated, the choice of estimator is rarely justified by formal tests. This article makes two contributions. First, it estimates the determinants of plot employment and of the overall informal share across 205 districts, using panel and cross-section models together with quantile, threshold, decomposition and machine-learning extensions. Second, it organises the analysis as a six-phase framework in which each estimation step is preceded and followed by the diagnostic tests that decide what comes next. The framework, shown in Figure 1, is meant to be reusable: it can be re-run every year as the new labour resources balance is released.

2. Methodology

Data. The source is the labour resources balance for January–December of 2022, 2023 and 2024, compiled by district and city. District names in the three releases were matched within each region after normalising spelling, which produced a balanced panel of 205 districts and cities over three years (615 observations). The 2024 release is more detailed: it divides formal employment into enterprise jobs, individual entrepreneurs with patents, legal dehqon farms, the public sector and registered self-employed persons, and divides the informal sector into temporary and seasonal workers, family workers, unregistered entrepreneurs and household-plot workers. This detail is used for a cross-section of 206 units. For each unit we check that formal and informal employment add up to total employment and that the components add up to each total. Samarkand region also published 2025 figures, which are used for an out-of-sample comparison.

Variables. Two dependent variables are used. *INF* is the informal-sector share of domestic employment, that is, informal employment divided by formal plus informal employment inside the country, in per cent; workers abroad are excluded because they are not part of the domestic labour market. *PLOT* is the number of household-plot workers per 100 people employed domestically and is the main measure of agricultural informality. The explanatory variables are the unemployment rate (*UR*); workers abroad per 100 economically active (*ABR*); the labour force participation rate (*LFP*); enterprise jobs (*FIRM*), registered self-employed (*SELF*) and legal dehqon farm workers (*DEHQ*), each per 100 labour resources; women on maternity leave per 100 labour resources (*MAT*), a proxy for household demographic structure; a city dummy (*CITY*); and the logarithm of population (*lnPOP*). *FIRM* measures the local demand for formal wage labour, *SELF* measures how far the simplified, digitally available form of registration has spread, and *DEHQ* captures legally registered agricultural units.

Empirical strategy. Applied panel studies in leading journals usually organise the empirical work as a sequence of pre-estimation tests, baseline estimation, post-estimation diagnostics, robustness checks and policy analysis, with each test deciding which estimator comes next [7]. The same logic is followed here in six phases, summarised in Figure 1. No estimator is chosen by convention alone: every switch from one model to another is triggered by a formal test.

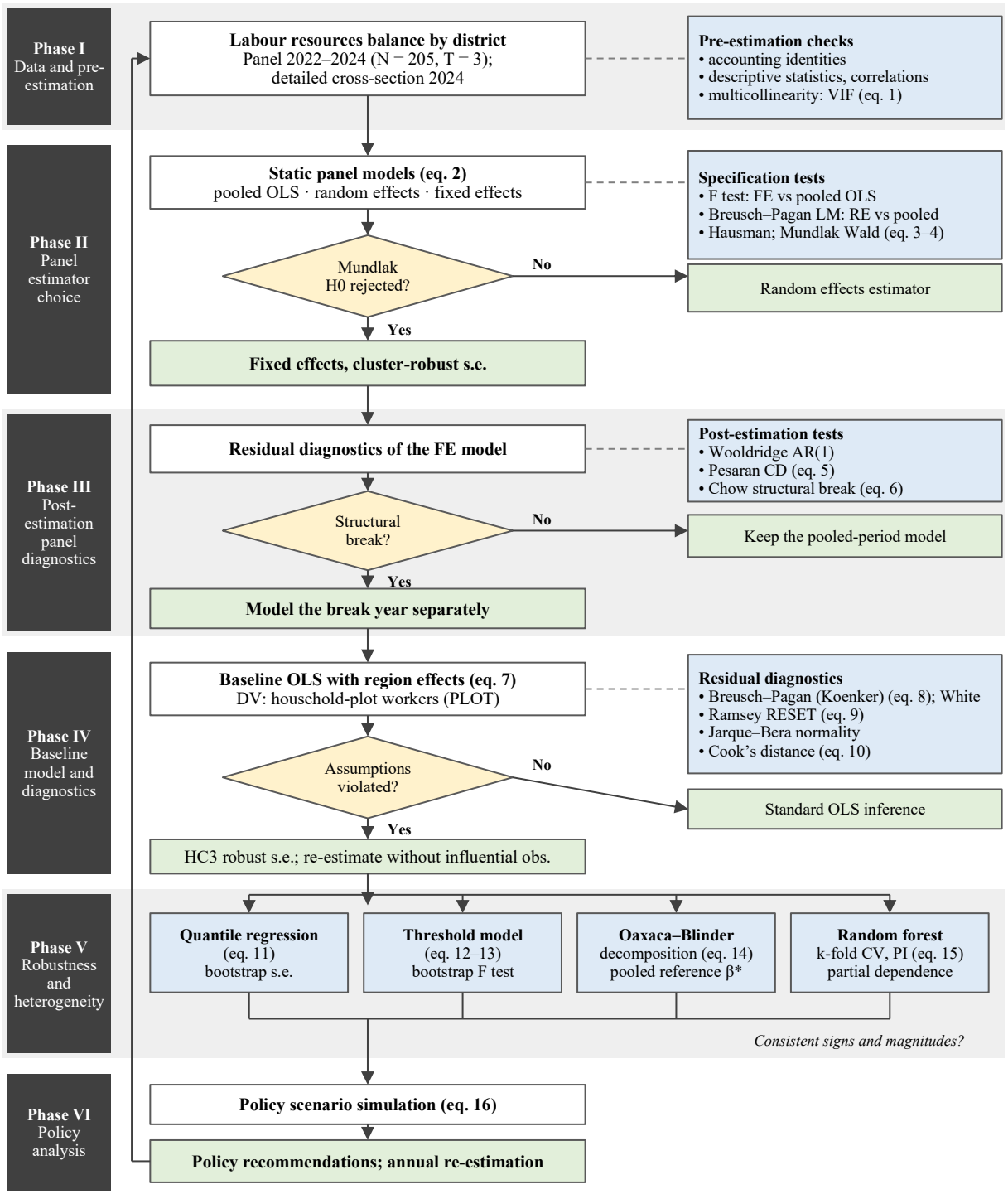

Figure 1. Econometric modelling framework for informal agricultural employment: phases, diagnostic tests and decision rules[8]

Phase I: data and pre-estimation diagnostics. Before estimation, the accounting identities of the balance are verified for every district, and the distributions and pairwise correlations of all variables are examined. Multicollinearity is assessed with variance inflation factors

$$\mathrm{VIF}_k = \frac{1}{1-R_k^2}, \qquad (1)$$

where R_k^2 is the coefficient of determination from regressing the k-th regressor on all others; values above 10 are taken as a sign of harmful collinearity [9].

Phase II: choice of the panel estimator. The district panel model is

$$INF_{it} = \alpha_i + \gamma_1 UR_{it} + \gamma_2 ABR_{it} + \gamma_3 LFP_{it} + \tau_t + u_{it}, \qquad (2)$$

where α_i is a district effect, τ_t a year effect and u_{it} the idiosyncratic error. Pooled OLS with a city dummy, the random-effects (Swamy–Arora) estimator and the fixed-effects (within) estimator are compared. An F test of the joint significance of the district effects compares FE with pooled OLS, and the Breusch–Pagan LM test compares RE with pooled OLS. The classical Hausman test [10] is also computed, but it can give a negative statistic in finite samples. The choice between FE and RE therefore rests on the Mundlak (correlated random effects) formulation [11]:

$$\alpha_i = \boldsymbol{\pi}' \bar{\mathbf{x}}_i + \eta_i, \qquad (3)$$

where $\bar{\mathbf{x}}_i$ is the vector of district means of the time-varying regressors. The hypothesis $H_0: \boldsymbol{\pi} = \mathbf{0}$ is tested with the cluster-robust Wald statistic

$$W = \hat{\boldsymbol{\pi}}' \left[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\pi}}) \right]^{-1} \hat{\boldsymbol{\pi}} \sim \chi^2_k. \qquad (4)$$

Rejection means that the district effects are correlated with the regressors, and the FE estimator is retained; otherwise RE is used. Standard errors are clustered by district throughout [12].

Phase III: post-estimation panel diagnostics. The FE residuals are tested for first-order serial correlation with the Wooldridge test and for cross-sectional dependence with the Pesaran CD statistic [13]

$$CD = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij} \sim N(0,1), \qquad (5)$$

where $\hat{\rho}_{ij}$ is the correlation between the residuals of districts i and j. Because the 2024 release changed the structure of the balance, coefficient stability is tested with the Chow test [14]

$$F = \frac{[S_R - (S_1 + S_2)]/k}{(S_1 + S_2)/(n - 2k)}, \qquad (6)$$

where S_R is the residual sum of squares of the restricted model and S_1, S_2 are those of the models estimated before and in the break year. If a break is found, the break year is modelled separately in Phase IV.

Phase IV: baseline model of plot employment. Because the components of formal and informal employment are available only for 2024, the main agricultural model is a cross-section with region effects μ_r:

$$PLOT_i = \beta_0 + \sum_{k=1}^{9} \beta_k X_{ki} + \mu_r + e_i, \qquad (7)$$

where the regressors X_{ki} are *FIRM*, *UR*, *ABR*, *LFP*, *SELF*, *DEHQ*, *MAT*, *CITY* and *lnPOP*. Four groups of diagnostics follow estimation. Heteroskedasticity is tested with the studentised Breusch–Pagan (Koenker) statistic [15] and the special White test. The Koenker statistic is

$$LM = nR_a^2 \sim \chi^2_q, \qquad (8)$$

where R_a^2 comes from the auxiliary regression of the squared residuals $\hat{e}_i^2$ on the q regressors. Functional form is tested with the Ramsey RESET test based on the auxiliary model

$$PLOT_i = \mathbf{X}_i' \boldsymbol{\beta} + \varphi_1 \hat{y}_i^2 + \varphi_2 \hat{y}_i^3 + v_i, \qquad H_0: \varphi_1 = \varphi_2 = 0. \qquad (9)$$

Normality of the residuals is tested with the Jarque–Bera test and influential observations are identified with Cook's distance

$$D_i = \frac{\hat{e}_i^2}{k\, s^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}, \qquad (10)$$

where h_{ii} is the leverage of observation i, s^2 the residual variance and k the number of parameters. Observations with $D_i > 4/n$ are treated as influential, and the model is re-estimated without them. If heteroskedasticity is present, HC3 standard errors are used; if RESET rejects the linear form, the analysis moves to Phase V.[16]

Phase V: robustness and heterogeneity. Four complementary methods are applied. First, quantile regression estimates the effects at different points of the conditional distribution of plot employment [17]:

$$\widehat{\boldsymbol{\beta}}(\tau) = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau \left(PLOT_i - \mathbf{X}_i'\boldsymbol{\beta}\right), \qquad \rho_\tau(u) = u[\tau - I(u < 0)], \qquad (11)$$

for $\tau \in \{0.25;\ 0.50;\ 0.75\}$, with standard errors from 200 bootstrap replications. Second, a threshold model tests whether the effect of firm jobs changes beyond a certain level of formal employment [18]:

$$PLOT_i = \lambda_1 FIRM_i I(FIRM_i \le \gamma) + \lambda_2 FIRM_i I(FIRM_i > \gamma) + \mathbf{Z}_i'\boldsymbol{\delta} + \mu_r + e_i, \qquad (12)$$

where $I(\cdot)$ is the indicator function and $\mathbf{Z}_i$ contains the other regressors and a dummy for agricultural districts (AGR). The threshold is estimated by least squares, and its significance is tested with

$$\hat{\gamma} = \text{argmin}_{\gamma \in \Gamma} S_n(\gamma), \qquad F = n\frac{S_0 - S_n(\hat{\gamma})}{S_n(\hat{\gamma})}, \qquad (13)$$

where S_0 is the residual sum of squares of the linear model and Γ is the central 70% of the $FIRM$ distribution. Because γ is not identified under the null hypothesis, the p-value is obtained by bootstrap. Third, the gap between rural districts (R) and cities (C) is decomposed with the Oaxaca–Blinder method [19]:

$$\overline{PLOT}_R - \overline{PLOT}_C = (\bar{\mathbf{X}}_R - \bar{\mathbf{X}}_C)'\boldsymbol{\beta}^* + [\bar{\mathbf{X}}_R'(\boldsymbol{\beta}_R - \boldsymbol{\beta}^*) + \bar{\mathbf{X}}_C'(\boldsymbol{\beta}^* - \boldsymbol{\beta}_C)], \qquad (14)$$

where $\boldsymbol{\beta}^*$ is estimated on the pooled sample; the first term is the explained part of the gap and the bracket is the unexplained part. Fourth, a random forest of 500 trees is compared with OLS under 10-fold cross-validation, and the role of each variable is measured by its permutation importance

$$PI_k = \frac{1}{K}\sum_{f=1}^{K}\left(MSE_f^{(k)} - MSE_f\right), \qquad (15)$$

where $MSE_f^{(k)}$ is the prediction error in fold f after variable k has been randomly permuted. Partial dependence functions show the shape of each relationship. The forest serves as a check rather than as the main model, because its predictions cannot be read directly as policy effects.

Phase VI: policy simulation. Scenarios change one or more regressors in every district, add the implied change to the observed plot share and aggregate the districts, with domestic employment D_i as weights:

$$PLOT^s = \frac{\sum_i \max\left(0;\ PLOT_i + \sum_k \widehat{\beta}_k \Delta x_{ik}\right) D_i}{\sum_i D_i}. \qquad (16)$$

The difference between the baseline and a scenario, multiplied by total domestic employment, gives the number of people who would move out of plot informality.[20]

3. Results and Discussion

The scale of agricultural informality. Table 1 shows how the labour market changed between 2022 and 2024. Formal employment grew by more than one million people, almost entirely in 2024, while informal employment stayed close to 5.4 million. As a result, the informal share of domestic employment fell from 45.8% to 42.0%. Behind the stable total, however, lies a large agricultural component. In 2024 household-plot workers numbered 1.50 million (27.8% of the informal sector), and temporary and seasonal workers, many of whom are hired for harvests, numbered another 1.60 million (29.5%). By contrast, legal dehqon farms formally employed only 452,500 people, and registered self-employment covered 1.12 million.

Table 1. Labour market indicators of Uzbekistan, 2022–2024[21]

Indicator	2022	2023	2024
Formal employment, thousand	6,437.6	6,503.9	7,472.5
Informal employment, thousand	5,445.0	5,466.1	5,412.7
Working abroad, thousand	1,854.7	2,047.9	1,377.7
Unemployment rate, %	8.9	6.8	5.5
Informal share of domestic employment, %	45.8	45.7	42.0
Household-plot workers, thousand			1,504.0
Household-plot workers, % of domestic employment			11.7

The plot share varies widely across regions. It is close to zero in Tashkent city and 7.4% in Navoiy region, but reaches 16.0% in Bukhara and 19.0% in Sirdaryo region; within Sirdaryo, Boyovut district has 35.2 plot workers per 100 employed. Table 2 gives the descriptive statistics of the 2024 cross-section. The strongest simple correlation is between plot employment and firm jobs ($r = -0.74$). Firm jobs are also strongly and negatively correlated with unemployment ($r = -0.70$), which already suggests that the two variables will compete in a regression.

Table 2. Descriptive statistics of the 2024 district cross-section (n = 206)[22]

Variable	Mean	Std. dev.	Min	Max
PLOT, per 100 employed	12.76	7.09	0.03	35.21
INF, %	42.95	16.97	1.97	71.08
FIRM, per 100 labour resources	25.36	12.44	10.71	66.70
SELF, per 100 labour resources	5.82	2.40	1.35	17.65
DEHQ, per 100 labour resources	2.59	2.12	0.00	10.91
UR, %	5.62	0.49	3.84	6.51
ABR, per 100 active	8.64	4.15	0.00	23.22
LFP, %	76.65	5.06	56.95	89.55
MAT, per 100 labour resources	10.74	2.23	3.45	23.06
CITY	0.15	0.36	0	1
lnPOP	11.92	0.68	9.14	13.47

Phases II–III: the panel of informal shares. Table 3 reports the three panel estimators together with the tests that choose among them. The F test rejects pooled OLS in favour of district effects ($F = 32.34$, $p < 0.001$), and the Breusch–Pagan LM test rejects pooled OLS in favour of random effects ($LM = 294.51$, $p < 0.001$). The two estimators must then be compared. In our data the classical Hausman statistic is negative, which happens when the RE covariance matrix is not smaller than the FE covariance matrix in the required sense; the test is therefore not informative. The Mundlak test is clear: the district means of the regressors are jointly significant (Wald $\chi^2 = 269.38$, df = 3, $p < 0.001$), so district effects are correlated with the regressors and the FE estimator is preferred.

Table 3. Panel estimates of the informal share of domestic employment, 2022–2024

Variable	Pooled OLS	Random effects	Fixed effects
UR	15.466*** (0.746)	4.402*** (0.631)	1.485** (0.724)
ABR	0.453*** (0.101)	−0.251*** (0.075)	−0.393*** (0.116)
LFP	0.085 (0.093)	0.364*** (0.083)	0.422* (0.220)
CITY	−13.809*** (1.335)	−20.011*** (2.337)	absorbed
Year effects	yes	yes	yes
R^2 (within for FE)	0.593	–	0.358
F test, FE vs pooled			32.34 [p < 0.001]

Breusch–Pagan LM, RE vs pooled	294.51 [p < 0.001]	
Mundlak Wald, FE vs RE		269.38 [p < 0.001]
Wooldridge AR(1), F		3.50 [p = 0.063]
Pesaran CD		1.16 [p = 0.246]
Chow test, break in 2024, F		5.05 [p = 0.002]

The comparison of columns is instructive. In pooled OLS the unemployment coefficient is very large (15.5), because it mostly reflects permanent differences between districts: places with high unemployment are also places with a large informal sector. Once those permanent differences are removed by FE, the coefficient falls to 1.49. A one-point fall in a district's unemployment rate is associated with a 1.5-point fall in its informal share over time. Labour migration changes sign. Across districts, those with more workers abroad have higher informality. Within a district over time, however, an increase in migration is associated with a lower informal share (−0.39), probably because migrants leave mainly from informal and seasonal jobs.

The residual diagnostics support the FE specification with some caveats. The Wooldridge test does not reject the absence of first-order serial correlation at the 5% level ($p = 0.063$). The Pesaran CD test does not find significant average cross-sectional dependence ($p = 0.246$), although with only three years this test has limited power. The Chow test rejects coefficient stability in 2024 ($F = 5.05$, $p = 0.002$). This break coincides with the year in which formal employment jumped by almost one million and the balance began to report the structure of informal work, and it is a reason to model 2024 separately in Phase IV.

Phase IV: what drives plot employment. Table 4 presents the cross-section model of plot employment and its full diagnostic profile. The VIF values are all below 3.6, so collinearity does not inflate the standard errors to a worrying degree, although the 0.70 correlation between *FIRM* and *UR* explains why unemployment loses its significance once firm jobs are controlled for. The Koenker test ($p = 0.048$) and the White test ($p = 0.014$) both reject homoskedasticity, so HC3 standard errors are reported. The RESET test strongly rejects the linear form ($F = 18.16$, $p < 0.001$), and the Jarque–Bera test rejects normality because of heavy tails (kurtosis 4.76, skewness close to zero). Twelve districts have Cook's distance above $4/n$; when they are excluded, the coefficients of *FIRM* (−0.364) and *CITY* (−6.64) barely change.

Table 4. Determinants of household-plot employment, 2024 cross-section[23]

Variable	Coefficient	HC3 s.e.	Without influential obs.
FIRM	−0.389***	0.055	−0.364***
SELF	−0.637***	0.115	
CITY	−6.242***	1.438	−6.640***
ABR	−0.151**	0.070	
LFP	0.201***	0.076	
UR	−0.581	0.902	−0.586
DEHQ	0.030	0.192	−0.039
MAT	0.036	0.125	
lnPOP	−0.052	0.676	
Region effects	yes		yes
R^2 / adjusted R^2	0.816 / 0.793		
VIF, range	1.25–3.54		
Breusch–Pagan (Koenker), $\chi^2(9)$	17.02 [p = 0.048]		
White (special), $\chi^2(2)$	8.57 [p = 0.014]		
RESET, F(2, 181)	18.16 [p < 0.001]		

Jarque–Bera, $\chi^2(2)$	26.71 [p < 0.001]
Cook's D > 4/n, observations	12

Two variables do most of the work. One more enterprise job per 100 labour resources is associated with 0.39 fewer plot workers per 100 employed, and one more registered self-employed person with 0.64 fewer. The self-employment effect is larger per person, which is plausible: registration as self-employed is the most direct route out of plot informality for someone who keeps the same activity but sells produce or services legally. Cities have 6.2 points less plot employment even after these controls. Legal dehqon farms are not significant, which suggests that registered farms and plot work coexist rather than replace each other. A higher participation rate is associated with more plot work, consistent with plots absorbing people, often women, who enter the labour force where there are few formal jobs.

Because RESET rejects the linear form, two corrections were tried before moving on. Adding squared terms for *FIRM* and *DEHQ* raises R^2 only to 0.821; the squared terms are not significant and RESET still rejects (F = 16.29). A model in logarithms fits better (R^2 = 0.933) but RESET still rejects (F = 8.94, p < 0.001). The misspecification is therefore not a simple curvature that a polynomial can fix. It more likely reflects heterogeneity across districts, which is what Phase V is designed to examine.

Phase V (a): heterogeneity and thresholds. The quantile regressions in Table 5 show that the main effects are not uniform. The effect of firm jobs grows slightly from −0.35 at the 25th percentile to −0.42 at the 75th, and the effect of self-employment grows from −0.28 to −0.83. In other words, both forms of formal work matter most precisely in the districts where plot employment is highest. The city effect is stable at about −4.5 to −5.4. Unemployment and dehqon farms are insignificant at all quantiles.

Table 5. Quantile regression estimates for household-plot employment, 2024[24]

Variable	τ = 0.25	τ = 0.50	τ = 0.75
FIRM	−0.349*** (0.045)	−0.395*** (0.047)	−0.422*** (0.060)
SELF	−0.283** (0.141)	−0.598*** (0.122)	−0.834*** (0.162)
CITY	−5.208*** (1.191)	−5.428*** (1.218)	−4.472*** (1.438)
UR	1.321 (0.978)	1.203 (0.880)	1.234 (1.178)
ABR	−0.089 (0.112)	−0.071 (0.094)	0.096 (0.123)
DEHQ	−0.067 (0.183)	−0.057 (0.211)	0.003 (0.250)

The threshold model searches for a level of firm employment beyond which its effect changes. The best split is at *FIRM* = 39.0 jobs per 100 labour resources, with 175 districts below it. Below the threshold the slope is −0.439 (s.e. 0.062), above it −0.292 (s.e. 0.097). The bootstrap F test, however, does not reject the linear model (F = 1.94, bootstrap p = 0.373). The difference in slopes is economically meaningful, since the first formal jobs in a poor district seem to draw people out of plots faster than additional jobs in an already formal one. Statistically, though, a single threshold cannot be distinguished from a straight line with 206 observations. The agricultural dummy *AGR* is also insignificant once other variables are controlled for (0.729, p = 0.491), even though the raw mean plot share is 15.7 in agricultural districts and 3.2 elsewhere. The difference is explained by the observed covariates, mainly firm jobs.

Phase V (b): why rural districts have more plot work. The Oaxaca–Blinder decomposition in Table 6 confirms this. The average plot share is 14.54 in rural districts and 2.68 in cities, a gap of 11.87 points. Differences in characteristics explain 8.37 points, or 70.5% of the gap, and firm employment alone accounts for 7.33 points, or 61.8% of the total gap. Participation (1.00) and dehqon farms (0.70) add smaller amounts, while self-employment slightly narrows the gap (−0.49), because self-employment is somewhat more common in rural districts. The unexplained part (3.50 points) reflects differences in how

the same characteristics translate into plot work, for example land availability and household traditions, which the balance does not measure.

Table 6. Oaxaca–Blinder decomposition of the rural–city gap in household-plot employment, 2024

Component	Points	Share of gap, %
Rural districts, mean PLOT	14.54	
Cities, mean PLOT	2.68	
Total gap	11.87	100.0
Explained by characteristics	8.37	70.5
of which FIRM	7.33	61.8
of which LFP	1.00	8.4
of which DEHQ	0.70	5.9
of which UR	0.32	2.7
of which SELF	−0.49	−4.1
of which MAT, lnPOP, ABR	−0.51	−4.3
Unexplained (coefficients)	3.50	29.5

Phase V (c): does a machine-learning model see something else? Under 10-fold cross-validation the random forest predicts plot employment only slightly better than OLS (RMSE 3.80 against 3.87). The linear model therefore captures most of the predictable variation, and its misspecification costs little in accuracy. The forest ranks the variables differently, though. Its permutation importance is highest for *DEHQ* (0.536), followed by *FIRM* (0.322), *SELF* (0.112) and *lnPOP* (0.085). The partial dependence function explains why: plot employment jumps from about 3.8 in districts with no dehqon farm employment to about 15.2 once *DEHQ* exceeds 1.3, and is almost flat after that. In other words, dehqon farm activity works as a marker that separates agricultural from non-agricultural districts, not as a continuous driver. This is exactly the kind of step that a linear term cannot represent and that RESET detects. The partial dependence of *FIRM* declines from about 14.8 to 7.2, with most of the decline between 22 and 39 jobs per 100 labour resources, which agrees with the threshold estimates. We also tested the forest on a time split, training on 2022–2023 and predicting the 2024 informal share; there it performed far worse than the linear panel model (RMSE 27.2 against 10.2) because it could not anticipate the structural break found by the Chow test. Machine learning is therefore useful here as a diagnostic tool, but less so as a forecasting tool when the policy environment changes.

Phase VI: how much could plot informality fall? Table 7 translates the estimates into scenarios. Scenario S1 uses the threshold slopes and adds five enterprise jobs per 100 labour resources in every district. Scenario S2 adds two registered self-employed persons per 100 labour resources through digital registration and information campaigns. Scenario S3 does not rely on a regression. It assumes that 30% of plot workers register under a simplified voluntary social tax, which is a purely accounting scenario included for comparison. Scenario S4 combines all three, applying S3 after S1 and S2.

Table 7. Simulated household-plot share of domestic employment under policy scenarios

Scenario	PLOT share, %	Change, p.p.	People leaving plot informality
Baseline 2024	11.67	–	–
S1: +5 enterprise jobs per 100 labour resources	9.86	−1.81	≈ 233,000
S2: +2 registered self-employed per 100 labour resources	10.67	−1.00	≈ 129,000

S3: 30% of plot workers covered by a simplified social tax	8.17	−3.50	≈ 451,000
S4: S1 + S2 + S3 combined	6.28	−5.39	≈ 695,000

No single instrument halves plot informality, but the combined scenario comes close. These figures are associations, not causal effects, and they assume that the cross-district relationships of 2024 would hold within districts over time. The panel results suggest some caution, since within-district effects are smaller than cross-district ones. Still, the ranking of instruments is informative, and the 2025 data for Samarkand region offer a partial reality check. There, formal employment grew by 18.4% in one year (from 686,600 to 813,200) and the informal share of domestic employment fell from 48.6% to 41.6%, a change consistent in direction with the firm-job coefficients.

Implications for human capital and digital transformation. The results point to a sequence rather than a single policy. First, formal jobs in enterprises remain the main factor behind the rural–urban gap, so rural industrialisation and agro-processing are the long-run solution. Second, registered self-employment has a large effect per person and is strongest where plot employment is highest. Since registration is already digital, the remaining barriers are information and skills: knowing that registration exists, understanding the tax and pension consequences and being able to use an online service. This is a human-capital task in the classic sense of investment in knowledge that raises the return to formal work and one that local employment centres and *mahalla* institutions can take on. Third, the success of S3 depends on a simplified social contribution for plot workers with a clear benefit, such as pension years. Finally, the roadmap itself is a digital-governance tool. Because the labour resources balance is now produced every year, the same sequence of tests and models can be re-run annually to see whether the coefficients, and the break found in 2024, persist.[25]

4. Conclusion

This article modelled informal agricultural employment in Uzbekistan with the district-level labour resources balance and organised the analysis as a roadmap in which diagnostic tests decide each next step. The tests changed the conclusions in concrete ways. The Mundlak test, used in place of an invalid Hausman statistic, selected fixed effects and showed that the within-district effect of unemployment is ten times smaller than the pooled estimate. The Chow test identified a break in 2024 and justified a separate cross-section model. Heteroskedasticity tests led to robust inference. The RESET rejection led to quantile, threshold and machine-learning models, which showed that the effects of formal work are strongest where plot employment is highest and that dehqon farm activity acts as a step marker rather than a continuous driver. The substantive findings are consistent across methods. Enterprise jobs and registered self-employment are the two factors most strongly associated with lower plot employment; firm jobs explain 62% of the rural–city gap; and a combination of job creation, digital self-employment registration and a simplified social tax could reduce plot informality from 11.7% to about 6.3% of domestic employment, covering around 695,000 people. The main limitations are the short panel, the cross-sectional nature of the agricultural model and the lack of district data on education, land and income, all of which keep the estimates associational. As further years of the balance and survey data on human capital become available, the roadmap can be extended to dynamic panels and to causal designs that exploit the timing of district programmes.

REFERENCES

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

[2] R. La Porta and A. Shleifer, "Informality and development," *Journal of Economic Perspectives*, vol. 28, no. 3, pp. 109–126, 2014.

[3] G. Ulyssea, "Informality: Causes and consequences for development," *Annual Review of Economics*, vol. 12, pp. 525–546, 2020.

[4] President of the Republic of Uzbekistan, "On measures to reduce the share of informal employment and to form the balance of labour resources on the basis of modern approaches," Resolution No. PQ-366, Aug. 30, 2022. [Online]. Available: https://lex.uz/docs/-6179326

[5] *Labour Resources Balance of the Republic of Uzbekistan by Districts and Cities, January–December 2022, 2023, 2024 and 2025 (Samarkand region)* [Data set]. [Source organisation to be specified].

[6] Central Bank of Uzbekistan, CERR and UNDP, *Informal Employment in Uzbekistan*. Tashkent, Uzbekistan: UNDP, 2025. [Online]. Available: https://www.undp.org/uzbekistan/publications/informal-employment-uzbekistan

[7] M. H. Pesaran, "General diagnostic tests for cross-sectional dependence in panels," *Empirical Economics*, vol. 60, no. 1, pp. 13–50, 2021.

[8] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge, MA, USA: MIT Press, 2010.

[9] T. S. Breusch and A. R. Pagan, "The Lagrange multiplier test and its applications to model specification in econometrics," *Review of Economic Studies*, vol. 47, no. 1, pp. 239–253, 1980.

[10] J. A. Hausman, "Specification tests in econometrics," *Econometrica*, vol. 46, no. 6, pp. 1251–1271, 1978.

[11] Y. Mundlak, "On the pooling of time series and cross section data," *Econometrica*, vol. 46, no. 1, pp. 69–85, 1978.

[12] A. C. Cameron and D. L. Miller, "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, vol. 50, no. 2, pp. 317–372, 2015.

[13] D. M. Drukker, "Testing for serial correlation in linear panel-data models," *Stata Journal*, vol. 3, no. 2, pp. 168–177, 2003.

[14] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960.

[15] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.

[16] R. Koenker, "A note on studentizing a test for heteroscedasticity," *Journal of Econometrics*, vol. 17, no. 1, pp. 107–112, 1981.

[17] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.

[18] J. B. Ramsey, "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society: Series B*, vol. 31, no. 2, pp. 350–371, 1969.

[19] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, vol. 6, no. 3, pp. 255–259, 1980.

[20] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[21] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[22] B. E. Hansen, "Sample splitting and threshold estimation," *Econometrica*, vol. 68, no. 3, pp. 575–603, 2000.

[23] R. Oaxaca, "Male-female wage differentials in urban labor markets," *International Economic Review*, vol. 14, no. 3, pp. 693–709, 1973.

[24] A. S. Blinder, "Wage discrimination: Reduced form and structural estimates," *Journal of Human Resources*, vol. 8, no. 4, pp. 436–455, 1973.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.