

Cash, Profit, and Selection:

Decomposing Self-Employment Income Error*

Werner Hernani-Limarino

Abstract

Self-employment income error has three components with distinct statistical properties; no single method can correct all three. The survey may get net cash wrong—stochastic measurement error accounting for 66% of the total gap. Net cash, even if correct, omits the opportunity cost of the owner’s time—a deterministic accounting gap accounting for 34%. And workers selected into self-employment because cash flow exceeds the formal wage, invalidating corrections trained on salaried workers. Using 500 informal firms observed simultaneously by a financial diary and Bolivia’s household survey, I decompose the total gap and show each layer requires a distinct tool. The central finding is that **detection and correction are different problems**. An autoencoder trained on self-employed workers achieves near-perfect detection of large cash errors ($AUC = 0.963$)—but does not recover the income ranking ($\rho = 0.056$). The autoencoder compresses toward the conditional mean, flagging anomalies without telling you how much workers actually earned. A generalized random forest bridges both layers, achieving $\rho = 0.657$ ($SD = 0.048$ across 100 splits; 5th percentile = 0.571; worst split = 0.489). A selection model assigns types probabilistically, improving Type C classification F1 from 0.251 to 0.682. The decomposition resolves a puzzle: unsupervised ML appeared to fail because the literature conflates three problems into one. The diary costs \$40,000.

JEL: C81, C55, D22, J24, J46, O17.

Keywords: measurement error, self-employment income, informal sector, autoencoders, anomaly detection, selection models, financial diaries.

*Email: wernerhl@gmail.com. The opinions expressed do not reflect the position of affiliated institutions. All errors are the author’s. Data: <https://doi.org/10.7910/DVN/KZMHGU>. Code: <https://github.com/wernerhl/ledgers-bolivia>.

1 Introduction

When Bolivia’s national household survey asks a *taquero* how much his food stall earned last month, the answer he gives is wrong. Not wrong because he is lying—he reports what he knows. Wrong because what he knows is not what the survey needs. He knows how much cash came in and how much went out. He does not know—and cannot easily compute—whether his business is economically viable once his own time is properly costed. Two companion papers ([Hernani-Limarino, 2026a,b](#)) deploy a seven-day financial diary to 500 informal firms across Bolivia’s major cities and document the scale of this gap: the survey answer diverges from true economic income by an average of 1,971 Bs/week, or 1.86 times conventional profit. This paper asks what can be done about it.

The answer is more nuanced than prior work has recognized. The gap between survey report and economic truth is not one problem but three, and these three problems have fundamentally different statistical properties. The survey may get net cash wrong: respondents forget revenue, double-count costs, or report round numbers. This is classical measurement error, stochastic and idiosyncratic, and the machine learning literature offers tools for detecting it. But even if the survey captured cash flow perfectly, cash flow is not economic profit: a worker’s own time has an opportunity cost that never appears in his mental accounting. This is not error in the usual sense—it is a deterministic accounting gap, the same for every worker in the same hours-and-sector bin. And the workers answering the survey are not a random sample of the population: they selected into self-employment, so any correction trained on salaried workers carries a selection bias into the informal sector. Conflating these three problems—as the literature does when it calls the total gap “measurement error”—produces an object that no single method can correct. Separating them, each becomes tractable.

The central finding of this paper is that **detection and correction are different problems**. An autoencoder trained on self-employed workers achieves $AUC = 0.963$ for identifying which observations have large cash errors—near-perfect classification. The same autoencoder achieves $\rho = 0.056$ for recovering the true income ranking—essentially no correction at all. The mechanism is straightforward: the autoencoder learns the conditional expectation of income given other observables and flags observations that deviate from it. This compresses the income distribution toward its conditional mean, reducing scale error without preserving rank order. If the goal is to flag suspect observations for re-interview or deletion, the autoencoder succeeds brilliantly. If the goal is to impute true income for poverty measurement or targeting, it fails entirely. No prior paper has made this distinction explicit, and the conflation has led researchers to conclude that

unsupervised machine learning “doesn’t work” for measurement error correction when the more accurate conclusion is that detection and correction require different tools.

I decompose the total gap between survey report and economic profit and find that stochastic cash error accounts for 66% of the total, with the deterministic accounting gap accounting for 34%. But this 66/34 split is an average that masks sharp sectoral variation. In Comercio (retail trade), where transactions are frequent and cash handling is complex, 75% of the gap is cash error. In Transporte (taxi and minibus drivers), where cash handling is simple but hours are long, 56% of the gap is the accounting component. A simple predictive model based on transaction complexity and hours worked explains 41% of the variance in this ratio, allowing readers to extrapolate to other settings.

Three tools address the three problems. The autoencoder, trained within the self-employed population and validated against diary-measured cash flow, detects large cash errors with $AUC = 0.963$. It improves 71% of the policy-relevant Type C firms (loss-making but cash-sufficient), though it does not recover their income ranking. A generalized random forest trained on 500 diary observations bridges both the cash error and the accounting gap simultaneously, achieving $\rho = 0.657$ on average across 100 random train-test splits. The variance across splits matters for a sample this small: the standard deviation is 0.048, the 5th percentile is $\rho = 0.571$, and the worst split achieves $\rho = 0.489$. Even the worst case substantially outperforms naive survey use ($\rho = -0.178$), but readers should expect performance in the 0.5–0.7 range rather than exactly 0.657. A selection model assigns the companion paper’s four-type typology—viable, exposed, trapped, and rationed—to the full survey population, improving Type C classification F1 from 0.251 to 0.682. Scaling to the national self-employed population suggests approximately 400,000–750,000 “trapped” workers who are economically loss-making but will decline formal job offers because their cash flow exceeds the formal wage, though this range incorporates substantial uncertainty.

The contribution is methodological: a decomposition that separates three problems previously conflated, and a demonstration that detection and correction require different tools. The diary costs \$40,000. Section 2 develops the framework and provides explicit definitions of the four firm types. Section 3 describes the methods with appropriate caveats about hyperparameter selection and identification. Section 4 describes the data. Section 5 decomposes the error and provides the predictive model for the cash error share. Section 6 presents results, with the detection/correction distinction as the organizing principle. Section 7 draws implications for poverty measurement and programme targeting. Section 8 concludes.

2 Framework

2.1 Three Objects

The framework begins with a simple observation: there are three different income concepts in play, and the gap between any two of them is a different object requiring different treatment.

Let i index a self-employed worker. The first object is y_i^{EH} , the survey response. This is what the worker tells the enumerator when asked how much his business earned. It is recalled net cash, subject to all the usual sources of survey error: revenue undercount because not every sale is memorable, cost omission because some expenses were paid in cash and forgotten, and rounding because “about a thousand” is easier to say than “nine hundred forty-seven.”

The second object is π_i^{conv} , conventional profit or true cash flow. This is what the worker would report if he had perfect recall: revenue minus explicit cash costs, nothing imputed, nothing adjusted. The financial diary measures this exactly by recording every transaction over seven days. Conventional profit is what business owners actually observe and what they use to make decisions. When the taquero decides whether to stay in business or look for a wage job, he is comparing his cash flow to the formal wage, not his economic profit.

The third object is π_i^{adj} , adjusted or economic profit. This subtracts the opportunity cost of the owner’s time: $\pi_i^{adj} = \pi_i^{conv} - w \cdot h_i^{own}$, where w is the shadow wage the owner could earn in formal employment and h_i^{own} is hours worked. Economic profit is what economists care about because it determines whether the firm is viable: a firm with positive cash flow but negative economic profit is paying its owner less than the market rate for his time. But economic profit is invisible to the owner. No cash changes hands when he “pays himself”; the cost is an opportunity foregone, not a cheque written.

The total gap between survey report and economic truth decomposes exactly:

$$\underbrace{y_i^{EH} - \pi_i^{adj}}_{\text{total gap}} = \underbrace{(y_i^{EH} - \pi_i^{conv})}_{\varepsilon_{1i}: \text{cash error}} + \underbrace{(\pi_i^{conv} - \pi_i^{adj})}_{\varepsilon_{2i}: \text{accounting gap} = wh_i^{own}} \quad (1)$$

This decomposition is the foundation of the paper. [Samphantharak and Townsend \(2009\)](#) develop the framework for constructing corporate financial statements from household data, showing that the household–enterprise boundary breaks down for the self-employed in ways that standard accounting cannot easily handle. [de Mel et al. \(2009\)](#) demonstrate that simply asking “what was your profit?” produces more accurate answers than constructing profit from revenue-minus-expenses, because cost recall is so poor that

the constructed measure adds noise. What is new here is the recognition that the two components of the gap— ε_1 and ε_2 —have fundamentally different statistical properties and therefore require fundamentally different solutions.

The cash error ε_1 is stochastic. Some firms underreport more than others, some overreport, and the magnitude varies with transaction complexity, recall period, and individual characteristics. Crucially, the error is detectable: a firm that reports high income but has low consumption and few assets is anomalous, and this anomaly can be exploited by machine learning methods.

The accounting gap ε_2 is deterministic. Every taquero working 50 hours per week at a shadow wage of Bs 13.64 per hour has an accounting gap of Bs 682 per week. This is not error in any meaningful sense—it is a structural feature of how self-employed workers think about their income. No machine learning method can detect ε_2 as an anomaly because it is structurally embedded in every observation. A taquero with high consumption and many assets working long hours is not anomalous; he is typical.

The measurement error literature has not made this distinction. From the taxonomy of [Bound et al. \(2001\)](#) through the correction toolkit of [Carroll et al. \(2006\)](#) and the identification results of [Chen et al. \(2011\)](#) and [Hu and Schennach \(2008\)](#), the literature treats the gap between survey report and truth as a single object amenable to a single correction. The empirical literature follows suit: [Hurst et al. \(2014\)](#) estimate 25% underreporting of self-employment income using Engel curves ([Pissarides and Weber, 1989](#)); [Abraham et al. \(2021\)](#) document a growing gap between tax records and survey reports; [Meyer et al. \(2015\)](#) show worsening survey quality across multiple domains; [Bollinger et al. \(2019\)](#) show that both tails of the income distribution are affected. All of these papers define error as the total gap. The decomposition in equation (1) shows why this conflation makes the error appear intractable: two problems with different statistical properties have been treated as one.

2.2 Layer 1: Cash Measurement Error

The first layer of error is cash measurement error, defined as $\varepsilon_{1i} = y_i^{EH} - \pi_i^{conv}$. This is the gap between what the worker tells the survey and what he actually earned in cash terms. The validation-study literature provides expectations about its structure: measurement error in income is typically mean-reverting ([Bound and Krueger, 1991](#)), heteroskedastic with variance increasing in true income ([Bollinger, 1998](#); [Bound et al., 1994](#)), and exhibits test–retest reliability around 0.5 ([de Mel et al., 2009](#)) with profit autocorrelations of 0.3–0.4 across survey rounds ([McKenzie, 2012](#)).

The key insight is that cash error should be detectable from internal consistency. A worker who reports high income but has low consumption, few assets, and poor housing quality is probably misreporting—either his income is lower than he claims or his consumption is higher than he admits. An autoencoder exploits this by learning the joint distribution of income and other observables and flagging observations that deviate from the conditional expectation.

Formally, an autoencoder trained within the self-employed population learns:

$$\hat{y}_i = E[y_i^{EH} \mid \mathbf{x}_{-y,i}, SE_i = 1] \quad (2)$$

where $\mathbf{x}_{-y,i}$ denotes all observables except income and the conditioning on $SE_i = 1$ restricts to self-employed workers.

The theoretical foundation is [Vincent et al. \(2010\)](#), who show that a denoising autoencoder converges to the conditional expectation $E[X \mid \tilde{X}]$ where $\tilde{X}$ is the corrupted input. [Gondara and Wang \(2018\)](#) and [Lall and Robinson \(2022\)](#) apply autoencoders to survey imputation, though their target is item nonresponse rather than measurement error. No prior work has applied autoencoders to measurement error correction in economic data. This paper fills that gap—and shows why the naive application fails.

The autoencoder approach requires one maintained hypothesis:

Hypothesis 1 (Within-population consistency). *Among self-employed workers, consumption and assets correlate positively with true cash flow: $\text{Corr}(\pi_i^{conv}, c_i \mid SE_i = 1) > 0$.*

This hypothesis is testable using the diary data, and I confirm it in [Section 6](#): the correlation is $r = 0.34$ ($p < 0.001$). The hypothesis is deliberately weaker than what would be needed to correct income using salaried workers as a reference. It requires only that self-employed workers who earn more also consume more—a maintained hypothesis throughout the poverty measurement literature ([Deaton, 2003](#); [Meyer and Sullivan, 2003](#)). It does *not* require that salaried and self-employed workers with the same demographic characteristics earn the same income. This is crucial because that stronger assumption fails badly: the correlation between salaried and self-employed conditional expectations is negative ($\rho = -0.268$), as I show below.

2.3 Layer 2: The Accounting Gap

The second layer is the accounting gap, $\varepsilon_{2i} = w \cdot h_i^{own}$. This is the opportunity cost of the owner’s time, computed as shadow wage times hours worked. Unlike cash error, the accounting gap is deterministic: given sector and hours, it can be computed exactly.

The naive approach would be to subtract the imputed labour cost from reported income: $y_i^{EH} - h_i \times 13.64$. This fails because it compounds cash error with the subtraction. If the worker overreported income by Bs 500 and worked 50 hours at shadow wage Bs 13.64, the naive correction subtracts Bs 682 from an already-inflated number, producing an estimate that is Bs 500 too high in the cash component and Bs 682 too low in the labour component. The errors do not cancel; they interact.

A better approach is to learn the mapping from observables to economic profit directly. A generalized random forest (Athey et al., 2019) trained on diary data can capture both layers simultaneously: the linear relationship between hours and economic profit (the accounting gap) and the nonlinear residual from sector-specific revenue undercount (the cash error). This parallels the logic of Chernozhukov et al. (2018): separating nuisance estimation from target estimation prevents regularisation bias from contaminating inference. Hurst et al. (2014) document the accounting gap phenomenon for U.S. self-employment income; the companion papers extend the analysis to the Bolivian informal sector with transaction-level ground truth (Hernani-Limarino, 2026a,b).

2.4 Layer 3: Selection and the Four-Type Typology

The third layer is selection. Workers do not randomly choose between self-employment and wage work; they select based on comparative advantage, and this selection invalidates corrections that treat the two populations as exchangeable.

The standard informality literature recognises heterogeneity within the informal sector—Feige (1990), La Porta and Shleifer (2014), Ulyssea (2018)—but treats measurement as secondary to the structural question of why informality exists. The experimental literature on returns to capital finds high marginal returns in the informal sector (de Mel et al., 2008; McKenzie and Woodruff, 2008; Fafchamps et al., 2014), suggesting that capital constraints rather than low ability explain small firm size. But all of these studies condition on self-reported profits whose measurement properties are unknown. If 43% of firms are loss-making but report positive income, the estimated returns to capital are biased upward by survivorship in the reporting.

The companion paper (Hernani-Limarino, 2026a) develops a theory of voluntary informality among loss-making firms. The key insight is that workers observe cash flow, not economic profit. If cash flow exceeds the formal wage ($\pi^{conv} \geq w^*$), the worker declines formal offers even when economic profit is negative ($\pi^{adj} < 0$). This creates a population of “trapped” workers: economically loss-making but behaviourally stable, invisible to standard poverty measures, and unreachable by standard active labour market

programmes.

Crossing viability ($\pi^{adj} \geq 0$) with the cash threshold ($\pi^{conv} \geq w^*$) yields a four-type typology:

- Definition 1** (Four-Type Typology). • **Type A (Viable):** $\pi^{adj} > 0$ and $\pi^{conv} \geq w^*$. These workers earn positive economic profit and receive cash flow above the formal wage threshold. They are the success stories of the informal sector: genuine entrepreneurs who have found a niche where self-employment dominates wage work on both margins. They will decline formal offers because they are better off where they are.
- **Type B (Exposed):** $\pi^{adj} > 0$ and $\pi^{conv} < w^*$. These workers are economically profitable but cash-constrained—an unusual combination that requires high depreciation, delayed payment structures, or seasonal variation. Type B is rare in practice because profitable operations typically generate cash.
 - **Type C (Trapped):** $\pi^{adj} \leq 0$ and $\pi^{conv} \geq w^*$. These are the workers this paper is most concerned with. They are loss-making in economic terms: their time is worth more in formal employment than it generates in self-employment. But their cash flow is positive and exceeds the formal wage, so they observe no signal that anything is wrong. They will decline formal job offers despite earning negative economic profit—hence “trapped.” Active labour market programmes that raise the formal offer rate are irrelevant for this population: they would decline offers even if offered.
 - **Type D (Rationed):** $\pi^{adj} \leq 0$ and $\pi^{conv} < w^*$. These workers are both loss-making and cash-constrained. They would accept formal offers if available—they are the population that standard ALMP is designed to serve. Traditional job training, matching services, and hiring subsidies can help.

In the diary sample at $w^* = 546$ Bs/week (the median formal wage): Type A = 284 (56.8%), Type B = 2 (0.4%), Type C = 150 (30.0%), Type D = 64 (12.8%). The near-absence of Type B is a testable prediction of the model: profitable operations generate cash, so the combination of positive economic profit and negative cash flow is mechanically rare.

The selection model predicts $\Pr(SE_i = 1 \mid \mathbf{x}_i)$ using a logit on education, age, sex, and household composition, then assigns types based on GRF-predicted ($\hat{\pi}^{conv}, \hat{\pi}^{adj}$) crossed with the thresholds. I emphasize that this is **descriptive prediction, not causal identification**. The same variables appear in both the selection and income equations; there is no exclusion restriction—no variable that predicts selection into self-employment but not income conditional on self-employment. This means the model cannot identify the causal effect of selection on measurement error or on income. It can, however, predict type

membership given observables, which suffices for targeting. For counterfactual policy analysis—“what would happen if we moved Type C workers into formal employment?”—the model is silent.

2.5 Why Conflation Fails

A natural question is why not simply train an autoencoder on the full population—salaried and self-employed together—and let it learn the conditional expectation of income? The answer reveals a fundamental problem with pooled-sample corrections.

The pooled autoencoder converges to:

$$\hat{y} = E[y^{EH} \mid \mathbf{x}_{-y}] = \pi_S E_S[y \mid \mathbf{x}_{-y}] + \pi_{SE} E_{SE}[y \mid \mathbf{x}_{-y}] \quad (3)$$

where $\pi_S \approx 0.75$ is the share of salaried workers in the sample. This is a weighted average of the conditional expectations for the two populations. For the reconstruction to equal $E[\pi^{adj} \mid \mathbf{x}_{-y}]$ for self-employed workers requires conditional invariance across employment types:

$$E_S[y^* \mid \mathbf{x}_{-y}] = E_{SE}[\pi^{adj} \mid \mathbf{x}_{-y}] \quad (4)$$

That is, salaried workers with the same demographics as a given self-employed worker must earn the same expected income. This assumption fails catastrophically: the correlation between salaried and self-employed conditional expectations is $\rho = -0.268$ (Section 6). Workers who would earn high wages in the formal sector are precisely those who sort out of self-employment; the self-employed who remain have lower formal-sector potential.

The decomposition resolves the puzzle. The SE-only autoencoder corrects ε_1 within the selected population without requiring comparisons to salaried workers. The accounting gap ε_2 is handled by the hours adjustment. Selection is handled by the Roy model, which assigns types within the self-employed population rather than imputing across populations.

This connects to recent evidence that machine learning does not automatically outperform traditional methods for welfare measurement. [Corral et al. \(2025\)](#) show that ML-based poverty maps can underperform traditional small-area estimation in some settings. [Echevin et al. \(2025\)](#) find that semi-supervised deep learning for poverty prediction requires careful attention to the training sample. [Meyer and Mittag \(2019\)](#) demonstrate the stakes of getting the income concept right when linking surveys to administrative data. The broader ML-for-measurement literature—[Jean et al. \(2016\)](#) on satellite imagery, [Blumenstock et al. \(2015\)](#) on mobile phone metadata, [Aiken et al. \(2022\)](#) on humanitarian

targeting—predicts welfare from non-survey data sources. The finding here is different: *correcting* a mismeasured survey variable requires structural decomposition, not algorithmic sophistication, because the error has multiple components with different statistical properties.

2.6 Relation to Existing Work

No prior paper has decomposed self-employment income error into stochastic and deterministic components, tested ML correction against diary-quality ground truth with separate validation targets for each component, or used a selection model to assign informal-sector types to a full survey population. The validation-study tradition ([Bound and Krueger, 1991](#); [Hurst et al., 2014](#); [Abraham et al., 2021](#)) matches survey to administrative data—but administrative records do not exist for informal workers and capture reported income, not economic profit. The financial diary tradition ([Collins et al., 2009](#); [Dupas and Robinson, 2013](#)) provides the measurement instrument but has applied it to household finances rather than firm-level income validation. The ML-for-survey-quality literature has applied autoencoders to imputation ([Gondara and Wang, 2018](#); [Lall and Robinson, 2022](#)) and anomaly detection ([Liu et al., 2008](#); [Breunig et al., 2000](#); [Kuriakose and Robbins, 2016](#))—but not to measurement error correction. Our framework shows why these three steps—decompose, correct each layer separately, model selection—are necessary and why conflating them produces the “failure” of unsupervised ML that would otherwise be puzzling given the strong theoretical guarantee of [Vincent et al. \(2010\)](#).

3 Methods

This section describes the three methods that address the three layers of error. Each method is matched to the statistical properties of the error component it targets. The autoencoder exploits the stochastic nature of cash error to detect anomalies. The generalized random forest learns the joint relationship between observables and economic profit, absorbing both the cash error and the accounting gap. The selection model assigns type probabilities to the full survey population.

3.1 Autoencoder for Cash Error Detection (Layer 1)

The autoencoder is a neural network that learns to compress and reconstruct its inputs. By forcing the data through a low-dimensional bottleneck, the network learns the dominant

statistical regularities in the joint distribution of income and other observables. Observations that deviate from these regularities—for example, a household that reports high income but has low consumption and few assets—receive large reconstruction errors. The reconstruction error is the signal: observations with large errors are candidates for measurement error.

The architecture follows standard practice for tabular data: an encoder that compresses the input through successively smaller hidden layers to a bottleneck, and a decoder that reconstructs the original variables. Specifically: input $\rightarrow$ 64 units (ReLU activation, dropout 0.2) $\rightarrow$ 16 units (ReLU, dropout 0.1) $\rightarrow$ 4 units (bottleneck, ReLU) $\rightarrow$ 16 units $\rightarrow$ 64 units $\rightarrow$ output (linear activation). The bottleneck dimension of 4 was chosen from a grid search; I discuss the implications of this choice below.

The loss function is weighted mean squared error, with the income variable receiving weight 5 and all other variables receiving weight 1. This asymmetry is deliberate: the goal is to detect errors in income specifically, not in the full vector of observables. Weighting income more heavily forces the network to pay particular attention to income reconstruction. The optimizer is Adam with learning rate 10^{-3} and weight decay 10^{-5} ; batch size is 256; training runs for 200 epochs with early stopping at patience 20.

Training uses 5-fold cross-validation on the self-employed subsample of the household survey ($N = 5,817$). Each of the 500 diary firms receives an out-of-fold reconstruction: the network that reconstructs firm i was never trained on firm i . This prevents overfitting to the specific observations that will be used for validation. Crucially, the validation target is π^{conv} (diary-measured cash flow), not π^{adj} (economic profit): the autoencoder is designed to detect cash error, not to correct the accounting gap.

Hyperparameter selection caveat. The bottleneck dimension (4) and income weight (5) were selected by maximising the area under the ROC curve (AUC) against the absolute value of cash error $|\varepsilon_1|$ on the diary subsample. This means the $AUC = 0.963$ reported in Section 6 is in-sample with respect to hyperparameter tuning. The diary was used twice: first to choose the architecture, then to evaluate detection performance. Out-of-sample performance on a completely new diary—one not used for any hyperparameter selection—would likely be lower, perhaps substantially so. The appropriate interpretation is: conditional on these hyperparameters being a good choice, detection performance is excellent; but the hyperparameters themselves were optimised on this sample, so the reported AUC is an upper bound on expected performance. For the generalized random forest, this problem does not arise because evaluation uses held-out splits within the diary: the training set for each split was never used to select hyperparameters for that split.

The input variables are: log income, log consumption, asset index (first principal com-

ponent of 15 durables), housing quality index (materials, crowding, services), education in years, age in years, sex, weekly hours, household size, number of children under 15, number of working-age adults, and sector dummies (one-hot encoded with one category dropped). All continuous variables are standardised to zero mean and unit variance before training.

3.2 Generalized Random Forest for Profit Prediction (Layer 2)

The generalized random forest (GRF) is a nonparametric method that estimates conditional expectations using an ensemble of trees (Athey et al., 2019). Unlike the autoencoder, which flags anomalies, the GRF directly predicts economic profit from observables. This makes it suitable for both components of the error: the nonlinear relationship between sector, hours, and economic profit (which embeds the accounting gap) and the residual variation from cash error.

The implementation uses a gradient-boosted regression forest with 200 trees, maximum depth 3, and minimum leaf size 20. These hyperparameters were chosen for stability rather than maximum fit: with only 500 training observations, deep trees would overfit. The covariates are: log reported income y^{EH} , sector code, log asset index, weekly hours, and household size. The target variable is π^{adj} (diary-measured economic profit).

Evaluation uses 100 random stratified train-test splits, with 350 observations for training and 150 for testing in each split. Stratification ensures that each sector appears in both training and test sets in proportion to its population share. All performance statistics—RMSE, MAE, Spearman’s ρ , fraction with correct sign—are computed on the 150-observation holdout set for each split. I report the mean across splits, the standard deviation, the 5th and 95th percentiles, and the worst-case performance.

This variance reporting is essential. With only 350 observations for training, random variation in which observations land in the training set produces substantial variation in performance. A reader who trains a GRF on a different 500-firm diary should not expect to replicate $\rho = 0.657$ exactly; they should expect performance somewhere in the 0.5–0.7 range. The 5th percentile ($\rho = 0.571$) represents plausible out-of-sample performance; the worst case ($\rho = 0.489$) represents a bad draw. Even the worst case substantially outperforms naive survey use ($\rho = -0.178$), so the method is robust, but the point estimate should not be overinterpreted.

For the joint prediction of cash flow and economic profit needed for type assignment in Layer 3, I train two separate GRFs on the same 500 observations with different targets: one predicting π^{conv} , one predicting π^{adj} . The two predictions are then crossed with the

thresholds to assign type probabilities.

3.3 Selection Model for Type Assignment (Layer 3)

The selection model predicts the probability of self-employment from demographic characteristics and uses this probability to weight type assignment in the full survey population. The model is a logistic regression of $SE_i \in \{0, 1\}$ on education (years), age, age squared, sex, household size, number of children under 15, number of working-age members, and department dummies. It is estimated on all urban occupied workers in the household survey ($N = 14,093$), not just the self-employed.

Standard errors are computed by bootstrap with 200 replications. The bootstrap re-samples observations with replacement and re-estimates the full model on each resample; standard errors are the standard deviation of the coefficient estimates across resamples.

Type assignment proceeds in two steps. First, for each self-employed worker in the full survey, the GRF predicts $(\hat{\pi}^{conv}, \hat{\pi}^{adj})$ from observables. Second, the predictions are crossed with the thresholds $\pi^{adj} \geq 0$ and $\pi^{conv} \geq w^*$ to assign types. The threshold $w^* = 546$ Bs/week is the median formal wage in the survey, representing the opportunity cost of self-employment time. Sensitivity analysis varies this threshold and reports how type shares change.

Identification caveat. The selection and income equations share the same covariates. There is no exclusion restriction: no variable predicts selection into self-employment but not income conditional on self-employment. This means the model cannot identify the causal effect of self-employment on income or the causal effect of selection on measurement error. The model is purely predictive: given observables, what is the probability that a worker is self-employed, and given that they are self-employed, what are the predicted values of (π^{conv}, π^{adj}) ? For type assignment, prediction suffices. For counterfactual policy analysis—“what would happen to income if we moved a Type C worker into formal employment?”—the model provides no guidance.

3.4 Simulation Validation

To confirm that the autoencoder method is sound in principle—that it can detect and partially correct measurement error when the error structure is known—I conduct a simulation exercise on salaried workers. The logic is as follows: salaried income has relatively low measurement error (test-retest reliability is higher than for self-employment income), so it provides a clean baseline. I inject known error structures into salaried income and test whether the autoencoder can detect the injected errors.

Four error structures are tested: classical (i.i.d. normal with variance calibrated to match diary-revealed ε_1 moments), mean-reverting (negatively correlated with true income, so high earners underreport and low earners overreport), heteroskedastic (variance proportional to income), and combined (all three features simultaneously). For each structure, I corrupt salaried income, train the autoencoder on the corrupted data, and measure detection performance (AUC against the injected error) and correction performance (RMSE reduction). Each simulation uses 50 replications.

Interpretation caveat. This simulation is a sanity check, not a validation of the self-employed application. The entire argument of this paper is that self-employed and salaried workers are different populations with different error structures and different selection processes. Showing that the autoencoder works on salaried workers with injected error confirms that the algorithm is correctly implemented and that the statistical theory of [Vincent et al. \(2010\)](#) holds in this setting. It does not address whether the method works for self-employed workers specifically. That question can only be answered with the diary data—and the diary has only 500 observations. I report the simulation results for completeness and to demonstrate that the algorithm is sound, but emphasise that the diary is the ground truth for the self-employed application.

4 Data

The analysis combines two data sources: Bolivia’s national household survey and a purpose-built financial diary. The household survey provides the population frame and the income measure that needs correction. The diary provides the ground truth against which correction methods are validated.

4.1 The Household Survey

Bolivia’s Encuesta de Hogares (EH) is the country’s primary household survey, administered annually by the National Statistics Institute. The analysis sample includes 14,093 urban occupied workers, of whom 5,817 are classified as self-employed and 8,276 as salaried. The self-employment sample is smaller than the 14,222 urban self-employed in the full survey because the autoencoder requires non-missing values for all input variables; I restrict to complete cases throughout. The 500 diary firms are a subset of these 5,817; sample restrictions and variable construction are detailed in the companion papers ([Hernani-Limarino, 2026a,b](#)).

The key variables are: self-employment income (y^{EH}), measured as recalled monthly

net earnings converted to weekly rates; per capita household consumption, constructed from a detailed consumption module; an asset index computed as the first principal component of 15 durable goods; a housing quality index based on construction materials, crowding, and access to services; education in completed years; age; sex; weekly hours worked in the main occupation; household composition (size, number of children under 15, number of working-age adults); and sector of activity. Sector is classified into six categories: Comercio (retail trade), Manufactura (small-scale manufacturing), Transporte (taxi, minibuses, and trucking), Gastronomía (food preparation and service), Construcción (construction), and Servicios (personal and professional services).

4.2 The Financial Diary

The financial diary covers 500 informal firms across the six sectors and nine of Bolivia's departments. The sectoral distribution is: Comercio 182 firms (36.4%), Transporte 114 (22.8%), Manufactura 67 (13.4%), Gastronomía 54 (10.8%), Construcción 50 (10.0%), and Servicios 33 (6.6%). Each firm recorded all transactions—revenue, expenses, and owner draws—over seven consecutive days using a double-entry format with pre-printed categories. The diary was administered in September–October 2024 by trained enumerators who visited each firm at the beginning and end of the recording period.

The diary produces two income measures. Conventional profit π^{conv} is total revenue minus total explicit costs: this is what the business owner would calculate if asked how much the business earned. Adjusted profit π^{adj} subtracts the imputed cost of owner labour: $\pi^{adj} = \pi^{conv} - w \cdot h^{own}$, where w is the shadow wage (set to the median formal wage in the survey, Bs 13.64/hour) and h^{own} is hours worked by the owner. The accounting identity $\pi^{conv} - \pi^{adj} = wh^{own}$ closes to within Bs 0.01 for all 500 firms—confirmation that the diary arithmetic is exact.

Each diary firm is linked to the household survey via the survey's unique household identifier (folio). Linkage is verified by matching four demographic markers: household size, number of children, education of the household head, and sector of the firm. All 500 firms match on at least three of four markers; 487 match on all four.

At the threshold $w^* = 546$ Bs/week (the median formal wage in the survey), the type distribution is: Type A (viable) = 285 firms (57.0%), Type B (exposed) = 4 firms (0.8%), Type C (trapped) = 149 firms (29.8%), Type D (rationed) = 62 firms (12.4%). The near-absence of Type B confirms the theoretical prediction that profitable operations generate cash: it is mechanically difficult to be economically profitable while receiving less cash than the formal wage.

5 The Error Decomposition

This section presents the central empirical finding: the decomposition of the total gap between survey report and economic truth into its stochastic and deterministic components.

5.1 Total Gap and Its Components

The total gap $\varepsilon = y^{EH} - \pi^{adj}$ averages 1,971 Bs/week across the 500 diary firms. The median is lower, 1,329 Bs/week, reflecting the right skew typical of income distributions. To put this in perspective: the average self-employed worker in the diary reports income that overstates economic profit by nearly two thousand bolivianos per week—more than twice the median formal wage. This is not a small discrepancy; it is a fundamental mismeasurement that would bias any poverty or inequality analysis based on the survey.

Table 1 decomposes the total gap into its components by sector.

Table 1: Decomposition: $\varepsilon = \varepsilon_1 + \varepsilon_2$ by sector

Sector	N	Mean ε	Mean ε_1	Mean ε_2	$\varepsilon_1/\varepsilon$	$\varepsilon_2/\varepsilon$
Comercio	182	2,918	2,200	718	0.75	0.25
Manufactura	67	1,513	937	577	0.62	0.38
Transporte	114	1,320	574	746	0.44	0.56
Gastronomía	54	1,465	895	570	0.61	0.39
Construcción	50	1,236	580	656	0.47	0.53
Servicios	33	1,871	1,322	550	0.71	0.29
Full sample	500	1,971	1,299	672	0.66	0.34

Note: $\varepsilon_1 = y^{EH} - \pi^{conv}$ (cash error), $\varepsilon_2 = wh_{own}$ (accounting gap). All values in Bs/week.

The headline finding is that **stochastic cash measurement error accounts for 66% of the total gap**, with the deterministic accounting gap accounting for 34%. The stochastic component is the majority—but only the majority, not the totality. A method that perfectly corrected cash error would still leave one-third of the gap unaddressed.

The sector variation is striking and substantively important. In Comercio (retail trade), 75% of the gap is cash error. Retail traders handle many small transactions daily, making revenue recall difficult; they also tend to overestimate costs because purchases are frequent and salient. The optimal correction for Comercio is Layer 1: detecting and correcting cash error.

In Transporte (taxi and minibus drivers), 56% of the gap is the accounting component. Drivers have simple cash handling—passengers pay the fare, there are few intermediate

costs—so they recall cash flow reasonably well. But they work long hours, often 60 or more per week, so the imputed cost of their time is large. The optimal correction for Transporte is Layer 2: the hours adjustment.

This heterogeneity has practical implications. A researcher applying these methods to a different country should not use the 66/34 average uncritically. A country with more retail trade and less transportation in its informal sector will have a higher cash error share. A country where informal workers put in fewer hours will have a lower accounting gap share.

5.2 Predicting the Cash Error Share

A natural question is whether the cash error share can be predicted from observable characteristics. If so, a researcher without diary data could estimate the likely decomposition for their setting and choose methods accordingly. Table 2 reports OLS coefficients predicting the cash error share from sector and hours.

Variable	Coeff.	SE
Intercept	0.812	0.042
Log transactions/week	+0.068	0.018
Weekly hours (10s)	−0.031	0.008
Comercio (ref: Transporte)	+0.284	0.031
Manufactura	+0.142	0.044
Gastronomía	+0.138	0.048
Construcción	−0.012	0.051
Servicios	+0.224	0.058
R^2	0.41	
N	500	

Note: Dependent variable is $\varepsilon_1/(\varepsilon_1 + \varepsilon_2)$. Heteroskedasticity-robust standard errors.

The model explains 41% of variance in the ratio. High transaction complexity (many daily sales, as in retail and food service) predicts more cash error; long hours (as in transport and construction) predict more accounting gap. A reader extrapolating to another setting can use these coefficients: a commerce-heavy informal sector will have $\varepsilon_1/\varepsilon \approx 0.75$; a transport-heavy sector will have $\varepsilon_1/\varepsilon \approx 0.45$.

5.3 Cash Error and the Accounting Gap

The cash error (ε_1) averages 1,299 Bs/week (median 613), positive: the survey overstates cash flow. It is heavily skewed (15.22) and varies from 574 in Transporte to 2,200 in Comercio (Figure ??). The accounting gap (ε_2) averages 672 Bs/week (median 655). Type C firms have the largest gap (median 818, from 60 hours/week) and Type A the smallest (median 546, from 40 hours). The gap is largest for the policy-relevant population.

5.4 Reliability Ratios

Table 3: Dual reliability ratios by sector and gender

Cell	N	λ (profit)	λ^{conv} (cash)
Comercio, Male	52	0.001	0.001
Comercio, Female	130	0.066	0.043
Manufactura, Male	38	0.164	0.148
Manufactura, Female	29 [†]	0.101	0.101
Transporte, Male	113	0.092	0.066
Gastronomía, Female	45	0.089	0.062
Construcción, Male	50	0.322	0.319
Servicios, Male	18 [†]	0.170	0.154
Servicios, Female	15 [†]	0.295	0.269

Note: $\lambda = \text{Var}(\pi^{adj}) / [\text{Var}(\pi^{adj}) + \text{Var}(\varepsilon)]$. $\lambda^{conv} = \text{Var}(\pi^{conv}) / [\text{Var}(\pi^{conv}) + \text{Var}(\varepsilon_1)]$. [†]Small N ; estimates imprecise.

Comercio males: $\lambda = 0.001$ —99.9% of variance is noise. The best cell (Construcción males) has $\lambda = 0.322$. Cells marked [†] have $N < 30$; their reliability estimates are imprecise and should be interpreted with caution. Cash and profit reliability are similar: the survey is bad at measuring cash *and* cash differs from profit. Both problems are severe.

6 Results

This section presents the empirical results, organised around the paper’s central finding: detection and correction are different problems requiring different methods. I begin with the headline result comparing autoencoder and GRF performance on detection versus correction tasks. I then present detailed results for each layer of the decomposition.

6.1 Detection vs. Correction: The Core Finding

The headline finding of this paper is that **detection and correction are fundamentally different problems**, and methods that excel at one may fail at the other. Table 4 summarises the distinction.

Table 4: Detection vs. Correction: The core distinction

Task	Metric	Autoencoder	GRF
Detection	AUC (large $ \varepsilon_1 $)	0.963	0.712
Correction	ρ (rank recovery)	0.056	0.657

Note: Detection = identifying which observations have large errors. Correction = recovering the true income ranking. AUC measured against observations with $|\varepsilon_1| > 1\sigma$. ρ is Spearman’s rank correlation with diary-measured π^{adj} .

The autoencoder achieves near-perfect detection (AUC = 0.963) but almost no correction ($\rho = 0.056$). The generalized random forest achieves modest detection (AUC = 0.712) but good correction ($\rho = 0.657$). The two methods solve different problems, and neither solves both.

Why does this happen? The mechanism is straightforward once understood. The autoencoder learns the conditional expectation of income given other observables: $E[y^{EH} | \mathbf{x}_{-y}]$. Observations whose reported income deviates substantially from this expectation receive large reconstruction errors. This makes the autoencoder excellent at detection: it identifies which observations are anomalous relative to the joint distribution. But the reconstruction itself is a compression toward the conditional mean. Every self-employed worker in the same demographic cell receives a similar reconstructed income, regardless of their true income within that cell. The reconstruction removes between-cell variation (good) but also within-cell variation (bad). The result is $\rho = 0.056$: the autoencoder preserves almost none of the true income ranking.

The GRF, by contrast, learns to predict economic profit from observables including reported income. Because reported income is an input, the GRF preserves some of the true within-cell variation: a worker who reports high income receives a higher predicted profit than a worker who reports low income in the same cell. The GRF achieves $\rho = 0.657$ —far from perfect, but far better than the autoencoder.

This finding has implications for the growing literature on machine learning for survey quality. If the goal is to flag suspect observations for re-interview, reweighting, or deletion, autoencoders work well: they identify anomalies with AUC near 0.96. If the goal is to impute true income for poverty measurement, welfare analysis, or targeting, autoencoders

fail: they do not recover the income ranking. The literature has not made this distinction explicit. Papers that report high detection performance are sometimes cited as evidence that ML can “correct” survey measurement error; this conflates two different tasks.

6.2 Layer 1: Detecting and Correcting Cash Error

6.2.1 Detection Performance

Table 5 reports detection performance against cash error $|\varepsilon_1|$ at three thresholds.

Method	AUC (0.5σ)	AUC (1.0σ)	AUC (1.5σ)	AUC-PR
AE (SE-only)	0.882	0.963	1.000	0.886
AE $ e_{\text{inc}} $	0.584	0.661	0.649	0.158
Isolation Forest	0.496	0.492	0.491	0.053
Linear residual	0.500	0.500	0.500	0.046

Note: AUC = area under ROC curve for binary classification of observations above threshold. AUC-PR = area under precision-recall curve. SE-only = autoencoder trained on self-employed workers only.

The SE-only autoencoder achieves $\text{AUC} = 0.963$ at the 1σ threshold and $\text{AUC} = 1.000$ at the 1.5σ threshold. This is near-perfect detection: the autoencoder identifies essentially all observations with large cash errors. Figure ?? plots the full ROC curve.

The comparison methods perform poorly. The isolation forest, a standard anomaly detection algorithm, achieves $\text{AUC} = 0.492$ —worse than random chance. This is not surprising: the isolation forest detects multivariate outliers, but cash error is not primarily about being an outlier in the joint distribution. A taquero can have perfectly typical demographics, consumption, and assets while still misreporting his cash flow. The linear residual (deviation from an OLS prediction of income) performs at exactly chance ($\text{AUC} = 0.500$): it is uninformative.

The autoencoder’s success comes from the within-population training. When trained on the full sample (salaried and self-employed together), detection performance drops to $\text{AUC} = 0.661$: the pooled autoencoder learns a mixture distribution that is poorly suited to detecting anomalies within either subpopulation. When trained only on self-employed workers, the autoencoder learns the conditional distribution specific to that population and detects deviations from it.

6.2.2 Correction Performance

Detection is only half the problem. The more important question for poverty measurement is whether the autoencoder can correct income—that is, whether the reconstructed income is closer to the truth than the reported income.

Table 6: Correction accuracy against π^{conv}

Measure	RMSE	MAE	ρ	% Improved
Reported y^{EH}	5,059	1,548	−0.044	—
AE (SE-only)	1,337	904	+0.056	59.4%
Type C: y^{EH}	4,046	1,896	+0.042	—
Type C: AE (SE)	1,245	882	+0.252	71.1%
Type D: y^{EH}	2,905	1,968	−0.143	—
Type D: AE (SE)	1,469	1,204	+0.067	66.1%

Note: $N = 500$. RMSE and MAE in Bs/week. ρ = Spearman correlation with diary cash flow. % Improved = fraction of observations where $|\hat{y} - \pi^{conv}| < |y^{EH} - \pi^{conv}|$.

The results are mixed. On the one hand, RMSE drops 74%, from 5,059 to 1,337 Bs/week. This is substantial: the reconstructed income is much closer to true cash flow on average. On the other hand, $\rho = 0.056$: the reconstructed income preserves almost none of the true rank ordering. The autoencoder reduces scale error without recovering rank order.

This distinction matters enormously for policy. Consider poverty targeting. If the goal is to set an eligibility threshold—for example, to identify all households below a certain income—then scale matters. The autoencoder’s compression toward the conditional mean reduces the false positive rate: households that appeared poor because of measurement error are correctly identified as non-poor. But if the goal is to rank households by income for progressive transfers or for analysis of inequality, then rank preservation matters. The autoencoder fails at this task: $\rho = 0.056$ means it does almost no better than random assignment.

The autoencoder is therefore a powerful diagnostic tool but a modest correction tool. It identifies which firms have large cash errors (detection) without telling you the true income ranking (correction). For Type C firms specifically, the policy-relevant category, performance is better: $\rho = 0.252$ and 71.1% of firms are improved. The autoencoder’s real contribution is to the type classification, where it moves firms across the $\pi^{conv} \geq w^*$ threshold, rather than to the income distribution itself.

6.2.3 The Cost of Ignoring Selection

What happens if we ignore selection and train the autoencoder on the full population?

Table 7 compares the pooled and SE-only autoencoders.

Table 7: AE performance: pooled vs SE-only training

Metric	AE (pooled)	AE (SE-only)
AUC-ROC (1σ)	0.962	0.963
RMSE	1,777	1,337
ρ	0.023	0.056
% Sign correct	57.8%	98.8%
Gradient recovery	62.7%	83.8%

Note: Pooled = trained on all workers. SE-only = trained on self-employed workers only.

Both autoencoders detect anomalies equally well ($\text{AUC} \approx 0.96$). The difference is in what they correct toward. The pooled AE corrects toward the salaried conditional mean; the SE-only AE toward the self-employed conditional mean. Sign accuracy jumps from 57.8% to 98.8%; gradient recovery from 62.7% to 83.8%. The conditional invariance correlation is -0.268 (Table 8)—selection, not methodology failure.

Table 8: Conditional invariance test

Metric	Value
Correlation (RF predictions)	-0.268
Mean difference (Bs/week)	1,955
NN-matched correlation	-0.162

Note: Salaried $N = 8,276$, Self-employed $N = 500$.

6.2.4 Simulation Validation

Table 9: Simulation: AE recovery of injected error in salaried income

Error type	RMSE (corrupt)	RMSE (AE)	Improv. %	AUC
Classical	390	341	12.7%	0.673
Mean-reverting	539	474	12.2%	0.684
Heteroskedastic	361	335	7.1%	0.663
Combined	522	475	8.9%	0.675

Note: $N = 8,022$ salaried workers. 50 replications.

The AE achieves AUC 0.66–0.68 and RMSE improvements of 7–13% on salaried workers with injected error. The method works when the error is anomalous (individual deviations within a homogeneous population). It achieves AUC = 0.963 on self-employed workers when targeted at ε_1 alone. It appeared to fail in preliminary analyses because it was asked to correct $\varepsilon_1 + \varepsilon_2$ while undoing a selection bias—three problems presented as one.

6.3 Layer 2: From Cash to Profit

Table 10: Correction to π^{adj} : layered methods (means)

Method	RMSE	MAE	ρ	% Sign
Naïve (y^{EH})	5,287	2,067	−0.178	57.8%
Linear ($y^{EH} - h \times 13.64$)	5,059	1,548	−0.028	60.0%
GRF (one-step)	498	346	+0.657	76.7%
AE (SE) $\rightarrow$ GRF (two-step)	503	347	+0.647	76.1%

Note: $N = 500$. GRF: 100 random 350/150 splits. All out-of-sample.

Table 11: GRF performance distribution across 100 splits

Metric	Mean	SD	5th pctl	95th pctl	Worst
ρ	0.657	0.048	0.571	0.731	0.489
RMSE	498	42	432	572	618
% Sign correct	76.7	3.1	71.3	81.3	68.0

Note: 100 random 350/150 train-test splits. All statistics computed on test set.

The linear correction barely helps (ρ : $-0.178 \rightarrow -0.028$) because it amplifies ε_1 . The GRF achieves $\rho = 0.657$ on average with RMSE 498—an order of magnitude below naïve (Figure ??).

The variance across splits matters. With only 350 observations for training, some splits perform poorly by chance. The standard deviation is 0.048; the 5th percentile is $\rho = 0.571$; the worst split achieves $\rho = 0.489$. Even the worst case substantially outperforms naïve survey use ($\rho = -0.178$), but a reader applying this to a new diary should expect performance in the 0.5–0.7 range, not exactly 0.657. The 5th percentile represents plausible out-of-sample performance; 0.571 is acceptable for type assignment but would affect poverty ranking at the margin.

The two-step (AE then GRF) performs comparably ($\rho = 0.647$): the GRF absorbs both layers internally. Its partial dependence on hours sits ~ 800 Bs/week above the linear $h \times 13.64$ line—that gap is the cash error the GRF captures on top of the accounting identity.

6.4 Layer 3: Selection and Type Assignment

6.4.1 Selection Equation

Table 12: Selection equation: $\Pr(SE = 1 \mid \mathbf{x})$

Variable	Coeff.	Marg. effect	SE
Education	−0.102	−0.025	0.001
Age	+0.098	+0.024	0.002
Male	+0.137	+0.034	0.009
Children	+0.109	+0.027	0.012
AUC	0.694	(N = 14,093)	

Note: Logistic regression. Age², HH size, department dummies included.

Education is the strongest predictor (−2.5 pp/year). Children under 15 increase SE probability (+2.7 pp), consistent with the flexibility motive.

6.4.2 Type Assignment and Typology Recovery

Table 13: Predicted types: diary vs full EH

Type	Diary (500)	EH (5,817)	Share	National
A	285	3,593	61.8%	1,087,103
C	149	1,894	32.6%	573,051
D	62	323	5.6%	97,727

Note: National = share $\times$ 1.76M (ILO). B (N = 7) omitted.

The point estimate is 573,051 trapped workers nationally. This estimate carries substantial uncertainty from three sources:

1. **GRF prediction variance.** The GRF was trained on 500 firms; its 5th-percentile $\rho = 0.571$ implies some firms are misclassified.

2. **Selection model error.** The selection equation achieves $AUC = 0.694$ —modest predictive power.
3. **Threshold sensitivity.** The type boundaries depend on $w^* = 546$ Bs/week; at $w^* = 700$, Type C share drops to 21.4%.

Incorporating these sources, a 90% confidence interval would span roughly 400,000 to 750,000 trapped workers. The point estimate is illustrative; readers should not treat 573,051 as precise.

Untargeted ALMP reaches Type D less than 15% of the time (C:D ratio = 5.9:1). Male Comercio workers are 58.3% Type C; female Manufactura workers are 8.6%.

Table 14: Layered typology recovery

Layers	Method	Precision	Recall	F1
<i>Type C (trapped)</i>				
—	Raw EH	50.0%	16.8%	0.251
1	AE (SE-only)	53.8%	28.2%	0.370
1+2	AE + GRF	61.8%	65.1%	0.634
1+2+3	Full model	68.7%	67.8%	0.682
<i>Type A (viable)</i>				
—	Raw EH	58.8%	87.0%	0.702
1+2+3	Full model	79.9%	91.9%	0.855
<i>Type D (rationed)</i>				
—	Raw EH	5.3%	1.6%	0.025
1+2+3	Full model	64.0%	25.8%	0.368

Note: $N = 500$, $w^* = 546$ Bs/week.

F1 nearly triples from 0.251 to 0.682. Each layer contributes: the AE raises recall from 16.8% to 28.2% by moving firms across the cash threshold; the GRF provides the large jump ($0.370 \rightarrow 0.634$) by bridging the accounting gap; the selection model adds the final increment.

7 Implications

7.1 Poverty: Two Distinct Shifts

Table 15: Poverty under layered corrections

Distribution	N	P_0	P_1	P_2
Full, Reported	5,817	0.292	0.089	0.040
Full, AE (SE-only)	5,817	0.278	0.078	0.033
Full, GRF $\hat{\pi}^{adj}$ [cens.]	5,817	0.939	0.759	0.668
500, Reported	500	0.296	0.113	0.063
500, π^{conv} (cash truth)	500	0.596	0.277	0.163
500, π^{adj} [cens.]	500	0.874	0.706	0.631

Note: Poverty line = 254 Bs/week per capita. [cens.] = negative income set to 0.

Two distinct shifts operate, and they correspond to different policy questions. The first shift—from reported ($P_0 = 0.296$) to cash truth ($P_0 = 0.596$)—reveals that the survey understates *consumption poverty* by 30 percentage points. This matters for social protection: many families who report adequate income are in fact cash-poor. The second shift—from cash truth ($P_0 = 0.596$) to profit truth ($P_0 = 0.874$)—reveals that most self-employed are economically loss-making once owner time is costed. This matters for firm viability and ALMP, not for consumption welfare: a taquero earning negative economic profit may still consume adequately because his cash flow is positive. The two concepts answer different questions. Conflating them—reporting $P_0 = 0.874$ as “the poverty rate”—would overstate deprivation. The cash-concept poverty rate ($P_0 = 0.596$) is the relevant welfare measure; the profit-concept rate ($P_0 = 0.874$) is a firm viability diagnostic.

Chesher and Schluter (2002) show measurement error biases FGT indices through error variance. Our decomposition goes further: ε_1 produces a stochastic bias; ε_2 produces a deterministic level shift that pushes firms with π^{conv} just above the poverty line to π^{adj} below it—amplifying the poverty shift beyond what the income gap alone would predict.

7.2 Inequality

Table 16: Inequality under layered corrections

Distribution	Gini	Theil-L	Theil-T
Full, Reported	0.362	0.230	0.228
Full, AE (SE-only)	0.344	0.206	0.202
Full, GRF $\hat{\pi}^{adj}$	0.372	0.322	0.227
500, π^{conv} (cash truth)	0.320	0.193	0.171
500, π^{adj} (profit truth)	0.468	0.461	0.374

Note: Positive incomes only.

Cash-flow inequality (Gini 0.320) is *lower* than reported (0.362): the survey overstates cash inequality through noise (Van Praag et al., 1983). Profit inequality (0.468) is higher: the accounting gap creates dispersion the survey misses entirely, with Theil-L jumping from 0.193 to 0.461 (Cowell and Victoria-Feser, 1996). Meyer and Sullivan (2023) attribute divergent U.S. income-consumption inequality trends partly to data quality; for the self-employed, the direction of the bias depends on which concept is measured.

7.3 Programme Targeting and Cost

The model predicts 97,727 Type D workers (5.6%) and 573,051 Type C (32.6%). Untargeted ALMP allocates at the population C:D ratio of 5.9:1. The F1 improvement from 0.251 to 0.682 for Type C means a model-targeted programme correctly identifies two-thirds of trapped firms (Hanna and Olken, 2018; Gazeaud, 2020; Beegle et al., 2012).

Layer 1 (AE): no diary, portable. Layer 2 (linear): no diary, computable from hours. Layer 2+ (GRF): requires diary. Layer 3 (selection): requires diary. Cost: \$40,000 for 500 diaries. Payoff: correctable survey, type-assignable population, targeting at four times the untargeted rate.

8 Conclusion

The survey asks one question. The answer is wrong in two ways. The population answering is selected.

The central finding is that detection and correction are different problems. An autoencoder achieves $AUC = 0.963$ for detecting which observations have large cash errors—

near-perfect classification. But it achieves $\rho = 0.056$ for recovering the true income ranking—almost no correction. The autoencoder compresses toward the conditional mean, reducing scale error without telling you who earns more than whom. If the goal is to flag suspect observations, it succeeds; if the goal is to impute true income, it fails. This distinction is absent from the prior literature.

Cash error (ε_1) accounts for 66% of the total gap, but the ratio varies sharply by sector (Comercio 75%; Transporte 44%). A predictive model based on transaction complexity and hours explains 41% of the variance in this ratio. The accounting gap (ε_2) is deterministic and accounts for 34%. A GRF trained on 500 diary observations achieves $\rho = 0.657$ on average (SD = 0.048; 5th percentile = 0.571; worst split = 0.489). Selection invalidates pooled-sample corrections ($\rho = -0.268$) but is handled by a Roy model; the caveat is that no exclusion restriction identifies selection from income determination.

Type C F1 rises from 0.251 to 0.682. The national extrapolation suggests 400,000–750,000 trapped workers, though this range incorporates substantial uncertainty. Three tools for three problems, one diary for all three, \$40,000.

References

- Abraham, K. G., Haltiwanger, J. C., Hou, K., Sandusky, K. and Spletzer, J. R. (2021). Reconciling survey and administrative measures of self-employment. *Journal of Labor Economics*, 39(4):825–860.
- Aiken, E., Bellue, S., Karlan, D., Udry, C. and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603:864–870.
- Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.
- Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.
- Beegle, K., De Weerd, J., Friedman, J. and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, 98(1):3–18.
- Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

- Bollinger, C. R. (1998). Measurement error in the Current Population Survey: A nonparametric look. *Journal of Labor Economics*, 16(3):576–594.
- Bollinger, C. R., Hirsch, B. T., Hokayem, C. M. and Ziliak, J. P. (2019). Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy*, 127(5):2143–2185.
- Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, 9(1):1–24.
- Bound, J., Brown, C., Duncan, G. J. and Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3):345–368.
- Bound, J., Brown, C. and Mathiowetz, N. (2001). Measurement error in survey data. In Heckman, J. J. and Leamer, E. (eds.), *Handbook of Econometrics*, 5:3705–3843. Elsevier.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T. and Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 93–104.
- Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Chapman and Hall/CRC.
- Chen, X., Hong, H. and Nekipelov, D. (2011). Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.
- Chesher, A. and Schluter, C. (2002). Welfare measurement and measurement error. *Review of Economic Studies*, 69(2):357–378.
- Collins, D., Morduch, J., Rutherford, S. and Ruthven, O. (2009). *Portfolios of the Poor: How the World's Poor Live on \$2 a Day*. Princeton University Press.
- Corral, P., Henderson, H. and Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 172:103394.
- Cowell, F. A. and Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica*, 64(1):77–101.

- de Mel, S., McKenzie, D. and Woodruff, C. (2008). Returns to capital in microenterprises: Evidence from a field experiment. *Quarterly Journal of Economics*, 123(4):1329–1372.
- de Mel, S., McKenzie, D. and Woodruff, C. (2009). Measuring microenterprise profits: Must we ask how the sausage is made? *Journal of Development Economics*, 88(1):19–31.
- Deaton, A. (2003). Household surveys, consumption, and the measurement of poverty. *Economic and Social Review*, 34(2):135–159.
- Dupas, P. and Robinson, J. (2013). Savings constraints and microenterprise development: Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1):163–192.
- Echevin, D., Fotso, A. S., Bouroubi, Y., Coulombe, H. and Li, Q. (2025). Combining survey and census data for improved poverty prediction using semi-supervised deep learning. *Journal of Development Economics*, 172:103385.
- Fafchamps, M., McKenzie, D., Quinn, S. and Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 106:211–226.
- Feige, E. L. (1990). Defining and estimating underground and informal economies. *World Development*, 18(7):989–1002.
- Gazeaud, J. (2020). Proxy means testing vulnerability to measurement errors? *Journal of Development Studies*, 56(11):2057–2078.
- Gondara, L. and Wang, K. (2018). MIDA: Multiple imputation using denoising autoencoders. *PAKDD 2018, Lecture Notes in Computer Science*, vol. 10939, pp. 260–272. Springer.
- Grosh, M. E. and Baker, J. L. (1995). Proxy means tests for targeting social programs: Simulations and speculation. World Bank LSMS Working Paper No. 118.
- Hanna, R. and Olken, B. A. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–226.
- Hernani-Limarino, W. (2026a). Trapped by cash: A theory of voluntary informality among the economically non-viable. Working paper.
- Hernani-Limarino, W. (2026b). Ledgers of the self-employed: Accounting for the invisible firm. Working paper.

- Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.
- Hurst, E., Li, G. and Pugsley, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *Review of Economics and Statistics*, 96(1):19–33.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.
- Kuriakose, N. and Robbins, M. (2016). Don’t get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, 32(3):283–291.
- La Porta, R. and Shleifer, A. (2014). Informality and development. *Journal of Economic Perspectives*, 28(3):109–126.
- Lall, R. and Robinson, T. (2022). The MIDAS touch: Accurate and scalable missing-data imputation with deep learning. *Political Analysis*, 30(2):179–196.
- Liu, F. T., Ting, K. M. and Zhou, Z.-H. (2008). Isolation forest. *Proceedings of the IEEE International Conference on Data Mining*, 413–422.
- McKenzie, D. and Woodruff, C. (2008). Experimental evidence on returns to capital and access to finance in Mexico. *World Bank Economic Review*, 22(3):457–482.
- McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2):210–221.
- Meyer, B. D. and Mittag, N. (2019). Using linked survey and administrative data to better measure income. *American Economic Journal: Applied Economics*, 11(2):176–204.
- Meyer, B. D. and Sullivan, J. X. (2003). Measuring the well-being of the poor using income and consumption. *Journal of Human Resources*, 38(S):1180–1220.
- Meyer, B. D. and Sullivan, J. X. (2023). Consumption and income inequality in the United States since the 1960s. *Journal of Political Economy*, 131(2):247–284.
- Meyer, B. D., Mok, W. K. C. and Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226.
- Pissarides, C. A. and Weber, G. (1989). An expenditure-based estimate of Britain’s black economy. *Journal of Public Economics*, 39(1):17–32.

- Samphantharak, K. and Townsend, R. M. (2009). *Households as Corporate Firms: An Analysis of Household Finance Using Integrated Household Surveys*. Cambridge University Press.
- Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from Brazil. *American Economic Review*, 108(8):2015–2047.
- Van Praag, B. M. S., Hagenars, A. J. M. and Van Eck, W. (1983). The influence of classification and observation errors on the measurement of income inequality. *Econometrica*, 51(4):1093–1108.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.
- Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.