

Cash, Profit, and Selection: Decomposing Self-Employment Income Error

Werner Hernani-Limarino

wernerhl@gmail.com

Roadmap

- 1 The Problem
- 2 Framework
- 3 Methods
- 4 Data
- 5 The Decomposition
- 6 Results
- 7 Implications
- 8 Conclusion

The Question We Ask

“How much did your business earn last month?”

- This question drives:
 - Poverty statistics
 - Programme targeting
 - Returns-to-capital estimates
 - Inequality measures

The Question We Ask

“How much did your business earn last month?”

- This question drives:
 - Poverty statistics
 - Programme targeting
 - Returns-to-capital estimates
 - Inequality measures
- **The answer is wrong.**

The Taquero's Problem

When Bolivia's household survey asks a taquero how much his food stall earned:

- He reports what he **knows**: cash in, cash out
- He cannot report what he **doesn't know**: economic profit
- The gap averages **1,971 Bs/week**
- That's **1.86×** conventional profit

The Gap

Bs 1,971/week

$\approx 2\times$ what he claims

What Prior Work Tells Us

- Hurst et al. (2014): 25% underreporting in U.S. (Engel curves)
- Abraham et al. (2021): Growing survey–tax gap
- Meyer et al. (2015): Worsening survey quality
- Bollinger et al. (2019): Both tails affected

What Prior Work Tells Us

- Hurst et al. (2014): 25% underreporting in U.S. (Engel curves)
- Abraham et al. (2021): Growing survey–tax gap
- Meyer et al. (2015): Worsening survey quality
- Bollinger et al. (2019): Both tails affected

The Common Assumption

All treat the gap as **one problem**: “measurement error”

What Prior Work Tells Us

- Hurst et al. (2014): 25% underreporting in U.S. (Engel curves)
- Abraham et al. (2021): Growing survey–tax gap
- Meyer et al. (2015): Worsening survey quality
- Bollinger et al. (2019): Both tails affected

The Common Assumption

All treat the gap as **one problem**: “measurement error”

This conflation is why the error appears intractable.

This Paper's Central Finding

Detection $\neq$ Correction

Autoencoder

- AUC = 0.963 (detection)
- ρ = 0.056 (correction)

GRF

- AUC = 0.712 (detection)
- ρ = 0.657 (correction)

This Paper's Central Finding

Detection $\neq$ Correction

Autoencoder

- AUC = 0.963 (detection)
- ρ = 0.056 (correction)

GRF

- AUC = 0.712 (detection)
- ρ = 0.657 (correction)

Different methods solve different problems.
No prior paper has made this distinction explicit.

Three Income Concepts

- ① **Survey report y^{EH}** : What the worker tells the enumerator
 - Subject to recall error, rounding, cost omission
- ② **Conventional profit π^{conv}** : True cash flow
 - Revenue — explicit costs (what owner observes)
 - Measured by financial diary
- ③ **Adjusted profit π^{adj}** : Economic profit
 - $\pi^{adj} = \pi^{conv} - w \cdot h^{own}$
 - Subtracts opportunity cost of owner's time
 - *Invisible to the owner*

The Decomposition

$$\underbrace{y^{EH} - \pi^{adj}}_{\text{Total gap}} = \underbrace{(y^{EH} - \pi^{conv})}_{\varepsilon_1: \text{Cash error}} + \underbrace{(\pi^{conv} - \pi^{adj})}_{\varepsilon_2: \text{Accounting gap}}$$

ε_1 : Cash Error (66%)

- **Stochastic**
- Varies across firms
- Detectable as anomaly
- ML can help

ε_2 : Accounting Gap (34%)

- **Deterministic**
- $= w \cdot h^{own}$
- Same for all in sector-hours bin
- **Not detectable**

Why the Distinction Matters

Cash error (ε_1):

- A taquero who reports high income but has low consumption is *anomalous*
- ML can detect this inconsistency

Accounting gap (ε_2):

- A taquero working 60 hours/week with high consumption is *typical*
- No ML method can detect this as an anomaly
- It's structurally embedded in every observation

Why the Distinction Matters

Cash error (ε_1):

- A taquero who reports high income but has low consumption is *anomalous*
- ML can detect this inconsistency

Accounting gap (ε_2):

- A taquero working 60 hours/week with high consumption is *typical*
- No ML method can detect this as an anomaly
- It's structurally embedded in every observation

Conflating them produces an object no single method can correct.
Separating them, each becomes tractable.

Layer 3: Selection

Workers **select** into self-employment:

- Not random assignment
- Based on comparative advantage
- Cash flow vs. formal wage comparison

Implication:

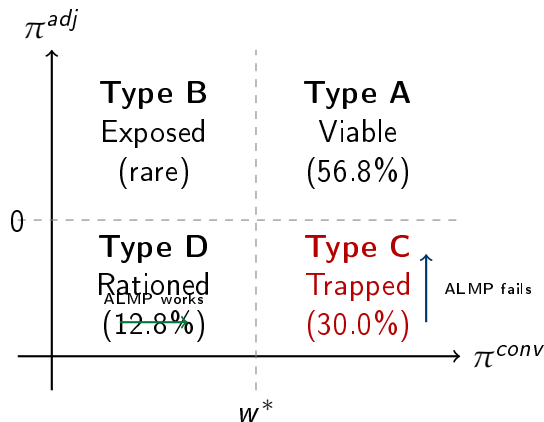
Corrections trained on salaried workers inherit selection bias.

The Trap Condition

Self-employment persists iff: $\pi^{conv} \geq w^*$

Workers observe cash flow, not economic profit.

The Four-Type Typology



Type C (Trapped): Loss-making but cash-sufficient $\Rightarrow$ declines formal offers

Type Definitions

Type A (Viable): $\pi^{adj} > 0, \pi^{conv} \geq w^*$

- Genuine entrepreneurs; will decline formal offers

Type B (Exposed): $\pi^{adj} > 0, \pi^{conv} < w^*$

- Profitable but cash-constrained; mechanically rare

Type C (Trapped): $\pi^{adj} \leq 0, \pi^{conv} \geq w^*$

- **Policy-relevant:** Loss-making, but observe positive cash
- Will decline formal offers despite negative economic profit

Type D (Rationed): $\pi^{adj} \leq 0, \pi^{conv} < w^*$

- ALMP target population; would accept formal offers

Why Pooled Corrections Fail

A pooled autoencoder learns:

$$\hat{y} = E[y \mid x] = \pi_S \cdot E_S[y \mid x] + \pi_{SE} \cdot E_{SE}[y \mid x]$$

For this to work, need:

$$E_S[y^* \mid x] = E_{SE}[\pi^{adj} \mid x]$$

Why Pooled Corrections Fail

A pooled autoencoder learns:

$$\hat{y} = E[y \mid x] = \pi_S \cdot E_S[y \mid x] + \pi_{SE} \cdot E_{SE}[y \mid x]$$

For this to work, need:

$$E_S[y^* \mid x] = E_{SE}[\pi^{adj} \mid x]$$

Empirical test:

Correlation between salaried and self-employed conditional expectations:

$$\rho = -0.268$$

Workers who would earn high wages sort *out* of self-employment.

Three Tools for Three Problems

Layer	Problem	Tool
1	Cash error (ε_1)	Autoencoder
2	Accounting gap (ε_2)	GRF
3	Selection	Roy model

- Each tool matches the statistical properties of its target
- No single tool solves all three
- The decomposition is what makes each tractable

Layer 1: Autoencoder

Architecture:

Input $\rightarrow$ 64 $\rightarrow$ 16 $\rightarrow$ 4 $\rightarrow$ 16 $\rightarrow$ 64 $\rightarrow$ Output

Key design choices:

- Trained on **self-employed only** (not pooled)
- Income weight = 5 (other variables = 1)
- Target: π^{conv} (cash), not π^{adj} (profit)
- 5-fold CV on $N = 5,817$ SE workers

Layer 1: Autoencoder

Architecture:

Input $\rightarrow$ 64 $\rightarrow$ 16 $\rightarrow$ 4 $\rightarrow$ 16 $\rightarrow$ 64 $\rightarrow$ Output

Key design choices:

- Trained on **self-employed only** (not pooled)
- Income weight = 5 (other variables = 1)
- Target: π^{conv} (cash), not π^{adj} (profit)
- 5-fold CV on $N = 5,817$ SE workers

Why It Works for Detection

Learns $E[y \mid x, SE = 1]$

Flags observations that deviate from conditional expectation

Why AE Fails for Correction

The autoencoder reconstructs:

$$\hat{y}_i = E[y \mid x_i, SE = 1]$$

This **compresses** toward conditional mean:

- Preserves between-cell variation
- **Destroys within-cell variation**
- Everyone in same demographic cell $\rightarrow$ same prediction

Result

- RMSE: $\downarrow$ 74%
- ρ : **0.056**

Scale error reduced.
Rank order lost.

Why AE Fails for Correction

The autoencoder reconstructs:

$$\hat{y}_i = E[y \mid x_i, SE = 1]$$

This **compresses** toward conditional mean:

- Preserves between-cell variation
- **Destroys within-cell variation**
- Everyone in same demographic cell $\rightarrow$ same prediction

Result

- RMSE: $\downarrow$ 74%
- ρ : **0.056**

Scale error reduced.
Rank order lost.

*The AE tells you who has large errors.
It does not tell you who earns more than whom.*

Layer 2: Generalized Random Forest

Specification:

- 200 trees, max depth 3, min leaf 20
- Covariates: $\log y^{EH}$, sector, assets, hours, HH size
- Target: π^{adj} (economic profit)
- Training: 500 diary observations

Evaluation:

- 100 random 350/150 train-test splits
- All metrics on held-out 150 obs
- Report mean, SD, 5th/95th percentiles, worst

Layer 2: Generalized Random Forest

Specification:

- 200 trees, max depth 3, min leaf 20
- Covariates: $\log y^{EH}$, sector, assets, hours, HH size
- Target: π^{adj} (economic profit)
- Training: 500 diary observations

Evaluation:

- 100 random 350/150 train-test splits
- All metrics on held-out 150 obs
- Report mean, SD, 5th/95th percentiles, worst

Why GRF Works for Correction

Reported income is an **input** $\Rightarrow$ preserves within-cell variation

Layer 3: Selection Model

Model:

$$\Pr(SE = 1 \mid \mathbf{x}) = \Lambda(\mathbf{x}'\boldsymbol{\gamma})$$

Covariates: Education, age, age², sex, HH composition, department

Type assignment:

- 1 GRF predicts $(\hat{\pi}^{conv}, \hat{\pi}^{adj})$
- 2 Cross with thresholds: $\pi^{adj} \geq 0, \pi^{conv} \geq w^*$
- 3 Assign types probabilistically

Layer 3: Selection Model

Model:

$$\Pr(SE = 1 \mid \mathbf{x}) = \Lambda(\mathbf{x}'\boldsymbol{\gamma})$$

Covariates: Education, age, age², sex, HH composition, department

Type assignment:

- 1 GRF predicts $(\hat{\pi}^{conv}, \hat{\pi}^{adj})$
- 2 Cross with thresholds: $\pi^{adj} \geq 0, \pi^{conv} \geq w^*$
- 3 Assign types probabilistically

Identification Caveat

No exclusion restriction. Model is **predictive, not causal**.
Sufficient for targeting; not for counterfactuals.

Data Sources

Household Survey (EH)

- $N = 14,093$ urban workers
- 5,817 self-employed
- 8,276 salaried
- Income, consumption, assets, demographics

Financial Diary

- $N = 500$ informal firms
- 7-day double-entry
- Sep–Oct 2024
- 6 sectors, 9 departments
- Linked to EH via folio

Diary cost: \$40,000 (Bs 80/firm)

Sectoral Distribution

Sector	N	Share	Primary activity
Comercio	182	36.4%	Retail trade
Transporte	114	22.8%	Taxi, minibus
Manufactura	67	13.4%	Small-scale production
Gastronomía	54	10.8%	Food service
Construcción	50	10.0%	Construction
Servicios	33	6.6%	Personal services
Total	500	100%	

Accounting identity $\pi^{conv} - \pi^{adj} = wh^{own}$ closes to Bs 0.01 for all 500.

Type Distribution in Diary

At $w^* = 546$ Bs/week (median formal wage):

Type	N	Share	Interpretation
A (Viable)	285	57.0%	Genuine entrepreneurs
B (Exposed)	4	0.8%	Profitable but cash-poor
C (Trapped)	149	29.8%	Loss-making, cash-sufficient
D (Rationed)	62	12.4%	Loss-making, cash-poor

Key finding: Type B near-absent (profitable $\Rightarrow$ generates cash)

Total Gap: 1,971 Bs/week

Sector	N	ε_1	ε_2	$\varepsilon_1/\varepsilon$
Comercio	182	2,200	718	75%
Manufactura	67	937	577	62%
Transporte	114	574	746	44%
Gastronomía	54	895	570	61%
Construcción	50	580	656	47%
Servicios	33	1,322	550	71%
Full sample	500	1,299	672	66%

ε_1 : Cash error (stochastic). ε_2 : Accounting gap (deterministic).

Sector Variation: Policy Implications

Comercio (75% cash error)

- Many small transactions
- Poor revenue recall
- Cost overestimation
- **Layer 1 (AE) helps**

Transporte (56% accounting gap)

- Simple cash handling
- Good cash recall
- Long hours (60+/week)
- **Layer 2 (hours adj) helps**

Sector Variation: Policy Implications

Comercio (75% cash error)

- Many small transactions
- Poor revenue recall
- Cost overestimation
- **Layer 1 (AE) helps**

Transporte (56% accounting gap)

- Simple cash handling
- Good cash recall
- Long hours (60+/week)
- **Layer 2 (hours adj) helps**

The optimal correction differs by sector.

One-size-fits-all fails.

Predicting the Cash Error Share

Can we predict $\varepsilon_1/\varepsilon$ from observables?

Variable	Coeff.	SE
Log transactions/week	+0.068	0.018
Weekly hours (10s)	-0.031	0.008
Comercio (vs Transporte)	+0.284	0.031
R^2	0.41	

Interpretation:

- More transactions $\Rightarrow$ more cash error
- More hours $\Rightarrow$ more accounting gap
- Allows extrapolation to other settings

Reliability Ratios by Sector-Gender

Cell	N	λ (profit)
Comercio, Male	52	0.001
Comercio, Female	130	0.066
Manufactura, Male	38	0.164
Transporte, Male	113	0.092
Construcción, Male	50	0.322

Comercio males: $\lambda = 0.001$ means **99.9% of variance is noise**

Best cell (Construcción males): only 32% signal

The Core Finding: Detection $\neq$ Correction

Task	Autoencoder	GRF
Detection (AUC)	0.963	0.712
Correction (ρ)	0.056	0.657

- AE: Near-perfect detection, essentially no correction
- GRF: Modest detection, good correction
- **They solve different problems**

Layer 1: Detection Performance

Method	AUC (0.5σ)	AUC (1.0σ)	AUC (1.5σ)
AE (SE-only)	0.882	0.963	1.000
AE (pooled)	0.584	0.661	0.649
Isolation Forest	0.496	0.492	0.491
Linear residual	0.500	0.500	0.500

Key findings:

- SE-only AE achieves near-perfect detection
- Pooled AE much worse (selection bias)
- Isolation forest = random chance
- Within-population training is crucial

Layer 1: Correction Performance

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	1,245	882	+0.252	71.1%

- RMSE drops 74% (scale improvement)
- $\rho = 0.056$ (no rank recovery)
- For Type C: $\rho = 0.252$, 71.1% improved

AE contribution: Moves firms across thresholds for type assignment

Layer 2: GRF Performance

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

- Linear correction fails (compounds cash error)
- GRF achieves $\rho = 0.657$, RMSE = 498
- Two-step similar: GRF absorbs both layers

GRF: Variance Across Splits

With $N = 350$ for training, variance matters:

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

Interpretation:

- Expect $\rho \in [0.5, 0.7]$ on new diary
- Even worst case ($\rho = 0.489$) beats naïve ($\rho = -0.178$)
- Don't overinterpret 0.657 as exact

Layer 3: Selection Model

Variable	Coeff.	Marginal Effect	<i>p</i>
Education (years)	-0.109	-2.5 pp/year	< 0.001
Female	+0.287	+6.2 pp	< 0.001
Children < 15	+0.109	+2.7 pp	0.012
AUC	0.694		

Interpretation:

- Education strongest predictor (-2.5 pp/year)
- Children increase SE probability (flexibility motive)
- AUC = 0.694: modest predictive power

Type Assignment: Layered 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

- Raw survey: $F1 = 0.251$ (misses 83% of Type C)
- Full model: $F1 = 0.682$ (**nearly triples**)
- Each layer contributes

National Extrapolation

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

Point estimate: 573,051 trapped workers

National Extrapolation

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

Point estimate: 573,051 trapped workers

Uncertainty

Three sources: GRF variance, selection model error, threshold sensitivity.
90% CI: approximately **400,000–750,000**

Two Distinct Poverty Shifts

Distribution	N	P_0	P_1	P_2
Reported	500	29.6%	11.3%	6.3%
Cash truth (π^{conv})	500	59.6%	27.7%	16.3%
Profit truth (π^{adj})	500	87.4%	70.6%	63.1%

Two shifts, two concepts:

- 1 Reported $\rightarrow$ cash truth: understates *consumption poverty*
- 2 Cash truth $\rightarrow$ profit truth: reveals *firm viability*

Don't conflate: $P_0 = 87.4\%$ is viability, not welfare.

Inequality Effects

Distribution	Gini	Theil-L	Theil-T
Reported	0.362	0.230	0.228
Cash truth (π^{conv})	0.320	0.193	0.171
Profit truth (π^{adj})	0.468	0.461	0.374

Key findings:

- Cash inequality *lower* than reported (noise overstates)
- Profit inequality *higher* (accounting gap creates dispersion)
- Direction of bias depends on concept measured

ALMP Mistargeting

Current targeting: Untargeted ALMP reaches C:D at 5.9:1

- Type C: Will *decline* formal offers (cash > wage)
- Type D: Will *accept* formal offers (ALMP works)
- Current ratio: For every 6 Type C reached, only 1 Type D

ALMP Mistargeting

Current targeting: Untargeted ALMP reaches C:D at 5.9:1

- Type C: Will *decline* formal offers (cash > wage)
- Type D: Will *accept* formal offers (ALMP works)
- Current ratio: For every 6 Type C reached, only 1 Type D

Model-targeted ALMP:

- F1 improvement: $0.251 \rightarrow 0.682$
- Correctly identifies 2/3 of trapped firms
- Can redirect resources to Type D

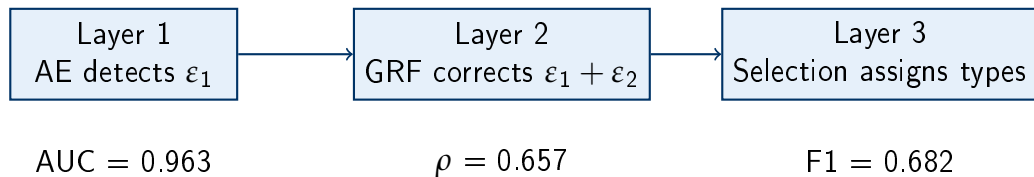
The Cost-Benefit Calculation

Cost	Benefit
Diary: \$40,000	Correctable survey
500 firms	Type-assignable population
7 days/firm	4× targeting improvement

What \$40,000 buys:

- 1 Train AE for Layer 1 detection (portable)
- 2 Train GRF for Layer 2 correction
- 3 Parameterize selection model for Layer 3
- 4 Apply to full survey ($N = 5,817$)
- 5 Extrapolate to national population

Summary: The Three-Layer Solution



The contribution:

- Decomposition separates three conflated problems
- Detection $\neq$ correction (new distinction)
- Each layer requires a different tool

The Central Message

The survey asks one question.

The answer is wrong in two ways.

The population answering is selected.

The survey asks one question.

The answer is wrong in two ways.

The population answering is selected.

- Cash error: 66% — AE detects, GRF corrects
- Accounting gap: 34% — deterministic, computable
- Selection: invalidates pooled corrections — Roy model handles

The survey asks one question.

The answer is wrong in two ways.

The population answering is selected.

- Cash error: 66% — AE detects, GRF corrects
- Accounting gap: 34% — deterministic, computable
- Selection: invalidates pooled corrections — Roy model handles

Three tools, one diary, \$40,000.

What Changes with This Paper

For researchers:

- Don't conflate detection and correction
- Train within-population, not pooled
- Report correction (ρ), not just detection (AUC)

For policymakers:

- Poverty rates depend on concept (cash vs. profit)
- ALMP mistargeting is severe (6:1 ratio)
- \$40K diary enables targeting improvement

For statistical agencies:

- Small diary can calibrate survey corrections
- Sector-specific approaches needed
- Reliability varies dramatically by cell

Links and Resources

Data:

<https://doi.org/10.7910/DVN/KZMHGU>

Code:

<https://github.com/wernerhl/ledgers-bolivia>

Papers:

<https://wernerhl.github.io/ledgers-bolivia>

wernerhl@gmail.com

Thank You

Questions?

Appendix: Hyperparameter Selection Caveat

Issue:

- Bottleneck dimension (4) and income weight (5) selected by maximizing AUC on diary
- $AUC = 0.963$ is *in-sample* w.r.t. hyperparameters

Implication:

- Out-of-sample on new diary would likely be lower
- Report as upper bound

Why GRF doesn't have this problem:

- Evaluation uses held-out splits
- Hyperparameters not tuned on test set

Appendix: Simulation Results

Inject error into salaried income, test AE recovery:

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

Interpretation:

- Algorithm is sound (works when error is anomalous)
- This is a sanity check, not validation of SE application
- Diary is ground truth for SE case

Appendix: Threshold Sensitivity

Type shares at different w^* thresholds:

w^* (Bs/week)	Type A	Type C	Type D
400	52.0%	35.2%	8.6%
546 (baseline)	57.0%	29.8%	12.4%
700	63.4%	21.4%	14.8%

Implication:

- Type C share sensitive to threshold choice
- Range 21–35% depending on w^*
- National estimates incorporate this uncertainty

Appendix: Comparison to Existing Work

Paper	Ground truth	Decomp.	Detection	Correction
Hurst et al. (2014)	Engel curves	No	No	No
Abraham et al. (2021)	Tax records	No	No	No
Jean et al. (2016)	Satellite	No	N/A	Yes
Aiken et al. (2022)	Phone data	No	N/A	Yes
This paper	Diary	Yes	Yes	Yes

What's new:

- Diary-quality ground truth for informal sector
- Decomposition into stochastic vs. deterministic
- Explicit detection vs. correction distinction
- Selection model for type assignment