Who is Bogus? Using One-Sided Labels to Identify Fraudulent Firms from Tax Returns

Aprajit Mahajan[1] **Shekhar Mittal**[2] **Ofir Reich**[3]

[1]UC Berkeley (Dept. of ARE) & CEGA

[2]UC Berkeley (iSchool)

[3]Precision Agriculture for Development

December 18, 2018

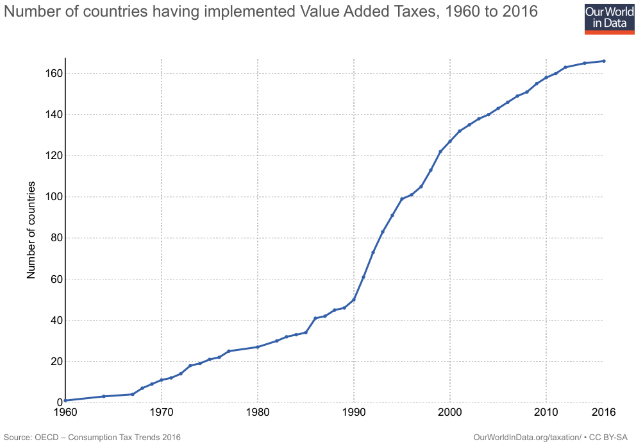
Zurich Conference on Public Finance in Developing Countries

ECONOMIC DEVELOPMENT & INSTITUTIONS

IGC International Growth Centre

FUNDED BY UKaid from the British people

Rapid Increase in Value Added Tax Adoption Since 1960



- 1 country in 1960 $\rightarrow$ 50 in 1990 $\rightarrow$ 160 in 2015.
- Tax levied at each stage of production or distribution (contra sales tax).

Evasion under VAT: Bogus Firms

- VAT requires buyer & seller to independently report each transaction.
  - Opposing incentives should reduce scope for collusion and evasion.
  - Whether this occurs, particularly in emerging economies, is an open question. (Limited) evidence.
    - In related work, we show that digitization enabled cross-checking of buyer and seller reports increased collections but only from better monitored firms.[1]
- Alternative evasion strategy – "Bogus" firms.
  - Bogus firms are shell firms created to enable firms to lower tax bills.
  - Create (fake) paper trails of transactions with genuine firms.
  - Role of ease of doing business norms.
- Precise extent and magnitudes largely unknown.
  - Media reports estimate the loss, in Delhi alone $\approx$ \$300m.[2]
- Commonly reported in many VAT systems.
  - Already documented cases in the newly launched Goods and Services Tax (India).
  - Confirmed problem in Mexico, Dominican Republic, and Zambia.

[1] Mittal and Mahajan (2017)

[2] India today article, TOI article, BS article

Detecting Bogus Firms: Current Practice

- Physical inspections gold standard, but resource intensive.
  - Audit resources limited (particularly in low-income countries).
- Key problem: How to identify firms for inspection?
  - More of them with less effort
- Officials in the central office create a list of "risky" firms.
  - Based on (limited set of) variables: low (VAT deposited/turnover), high turnover, high revisions, invalid address.
- Local inspectors sent out for inspections.
  - Firms deregistered ("cancelled") if inspection fails.

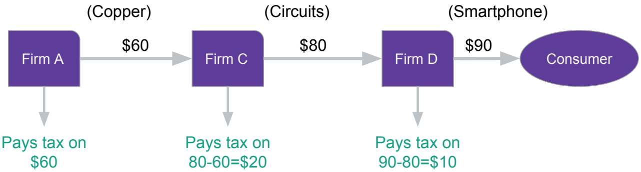
Our Work

We apply a random forest classifier to the value added tax (VAT) returns from Delhi (India) to identify bogus firms which can be further targeted for physical inspections.

Highlights

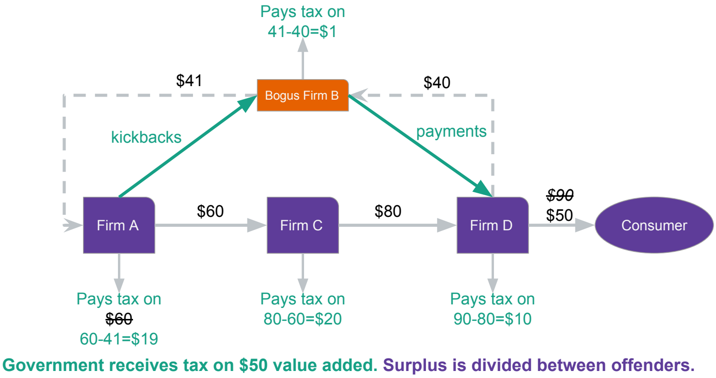
- One sided labels and in-sample predictions ⇒ Cross-validation.
- Precision, recall, F1 score not ideal ⇒ Focus on top recommendations.
- Non-RCT evaluation (for now) ⇒ Point-in-time simulation.
- Multiple firm-quarter observations but class timeless ⇒ Aggregate the predictions.

VAT: Example



Government receives tax on $90 value added.

Bogus Firms: Example



- Firms A, C and D not necessarily in the same chain.
- Bogus firm can make sales to any firm which needs input credits.
- Most tax systems suffer from much simpler mechanisms.

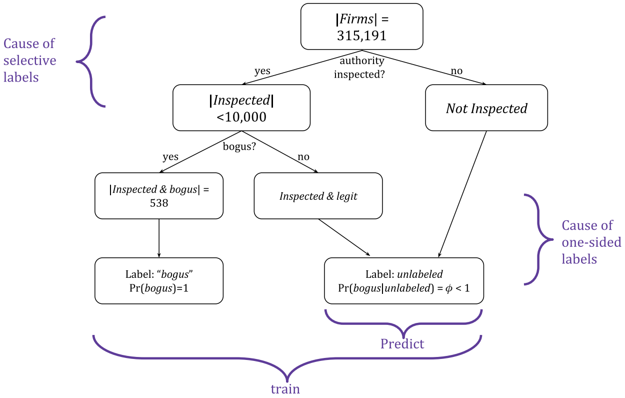
Delhi and Taxation

- Population: 16.8 million (2011 census)
- Real GSDP of NCT of Delhi for 2015-16: ₹4,560 billion (US$ 71 billion)
  - Tax to GSDP ratio: 5.7%
- VAT accounts for 52.4% of total government revenues

Data Description

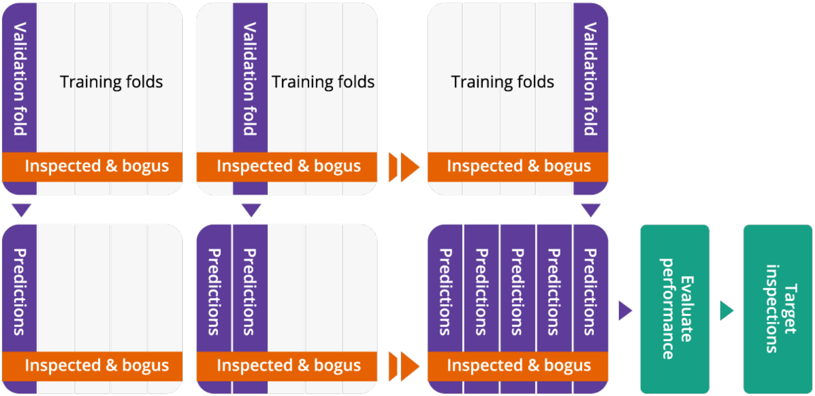
- Entire universe of registered firms from Delhi, India (315,191 firms, unbalanced).
- 3 years of quarterly VAT returns - 2012-13, 2013-14, 2014-15.
- 3 years of quarterly firm level interactions - 2012-13, 2013-14, 2014-15.
- Firm profile data.
- Bogus firm data: 531 bogus firms identified (2012-2015).

Bogus Firms: Our Challenge



- Inspection is based on the tax authority's discretion and so biased (selective labels).
- Class labels are known only for firms both inspected and found to be bogus, not for the rest (one-sided labels).
- We use all for training, but want to predict for those firms still unlabeled (in-sample predictions).

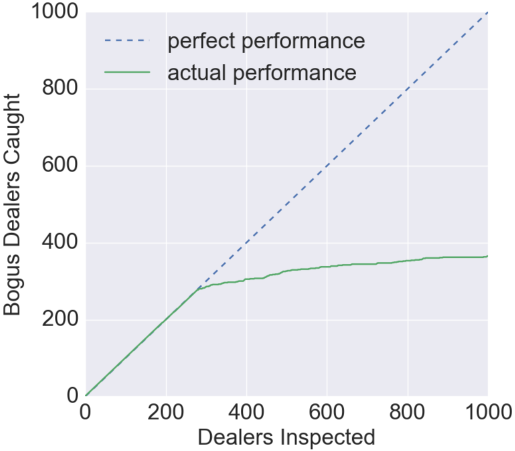
Cross-validated Prediction Procedure



Multi-Period Model

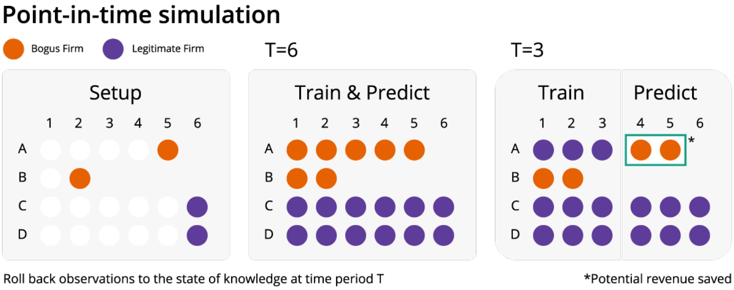
Wide Model

Random Forest Model Performance on Top 1000 Recommendations



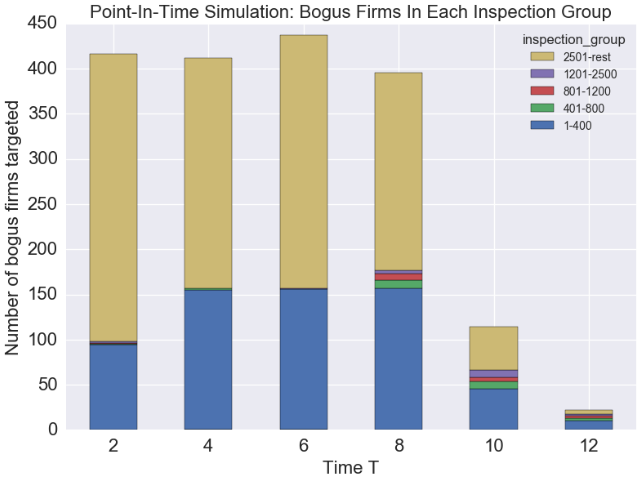
- Results similar when we control for revenue size.

Illustration of Point-in-time Simulation at T=3



- In a real-world scenario, do not have access to all returns by all firms and not required to predict retroactively.
- Blind the model to information obtained after time “T” - do not consider “future” tax returns after T.
- Only use bogus firms that were already classified by time T.

Point-in-time Simulations Performance



Point-in-time simulation performance for the 1-400 inspection group

T	*Total Bogus Firms Caught*	*Bogus Firms Caught/Inspection*	*Revenue Gained by Inspecting Entire Group (USD Millions)*	*Revenue Gained per Inspection (USD 000s)*	*Total Bogus Firms in the Sample*	*Revenue Lost from All Bogus Firms (USD Millions)*
2	94	0.24	19.44	48.60	416	49.40
4	155	0.39	43.19	107.97	412	108.38
6	156	0.39	25.48	63.70	437	63.84
8	157	0.39	9.38	23.46	395	26.43
10	46	0.11	1.70	4.24	114	4.52
12	10	0.02	0	0	22	0

Conclusions and Challenges

- Used digitized tax returns to create a ML tool to identify potentially fraudulent firms.
  - Next: Tax authority inspects most suspicious firms (create training data).
  - Finally: Compare revenue implications against current practices.
- Challenge 1: Revenue impact hard to measure.
- Challenge 2: Firms will respond to better targeting – e.g. by creating more bogus firms faster.
  - ML tool will require regular updating (more training data).
  - Real world example of adversarial ML.
- Interest from many tax authorities, potentially useful tool in the hands of high level officials.

Thanks!

- We thank GoNCTD, IGC, CEGA, EDI, and JPAL for support.
- This project was funded with UK Aid from the UK Government.

Aprajit Mahajan
aprajit@berkeley.edu
UC Berkeley & CEGA

Shekhar Mittal
shekharm@berkeley.edu
UC Berkeley (iSchool)

Ofir Reich
ofir@precisionag.org
PAD

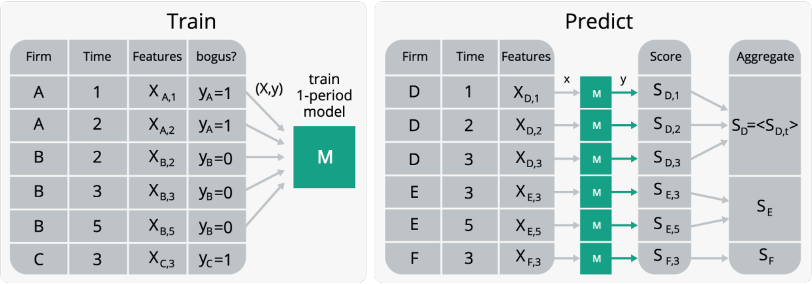
ECONOMIC DEVELOPMENT & INSTITUTIONS

FUNDED BY

UKaid from the British people

IGC International Growth Centre

Firm Level Predictions from Firm-quarter Data Points



Back

Wide Model

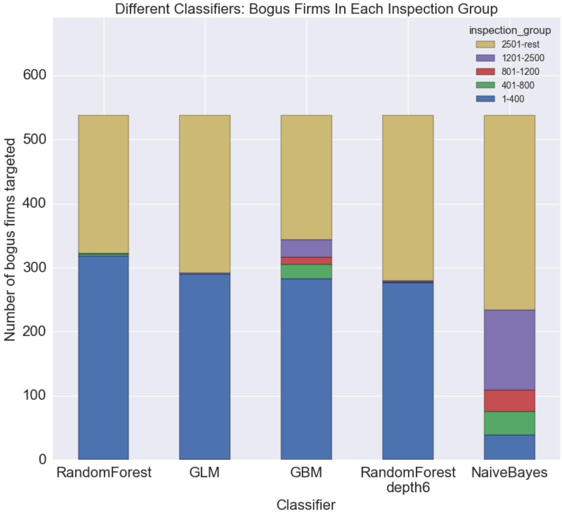
Firm	Time	Features	bogus?
A	1	$X_{A,1}$	$y_A=1$
A	2	$X_{A,2}$	$y_A=1$
B	2	$X_{B,2}$	$y_B=0$
B	3	$X_{B,3}$	$y_B=0$
B	5	$X_{B,5}$	$y_B=0$
C	3	$X_{C,3}$	$y_C=1$

Firm	Features_1	Features_2	Features_3	bogus?
A	$X_{A,1}$	$X_{A,2}$	NULL	$y_A=1$
B	NULL	$X_{B,2}$	$X_{B,3}$	$y_B=0$
C	NULL	NULL	$X_{C,3}$	$y_C=1$

- Entry and exit of firms will result in the dataset having a lot of NULL values

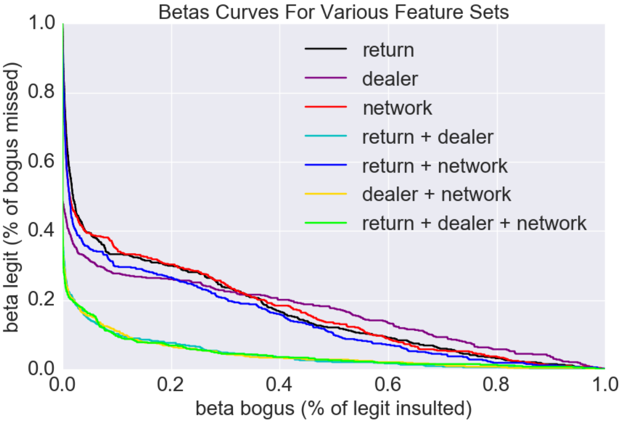
Back

Comparison of Different Classifiers



Back

Betas curves for different feature sets



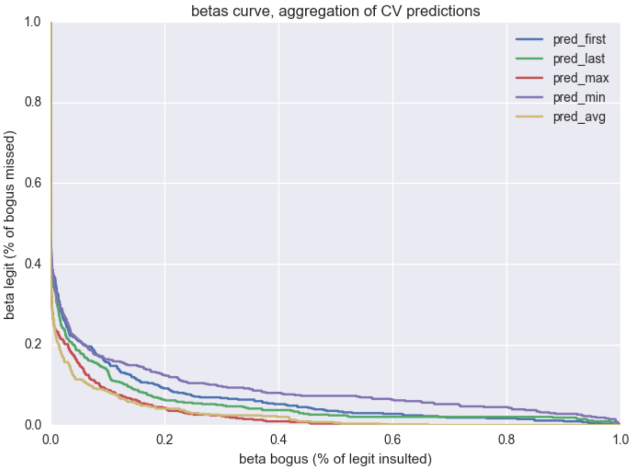
Back

Model performance on all recommendations

Inspection Group	*Firms Inspected*	*Total Bogus Firms Caught*	*Bogus Firms Caught/Inspection*
1 - 400	400	305	0.76
401 - 800	400	48	0.12
801 - 1200	400	24	0.06
1201 - 2500	1300	29	0.02
2501 - rest	313229	132	0.00

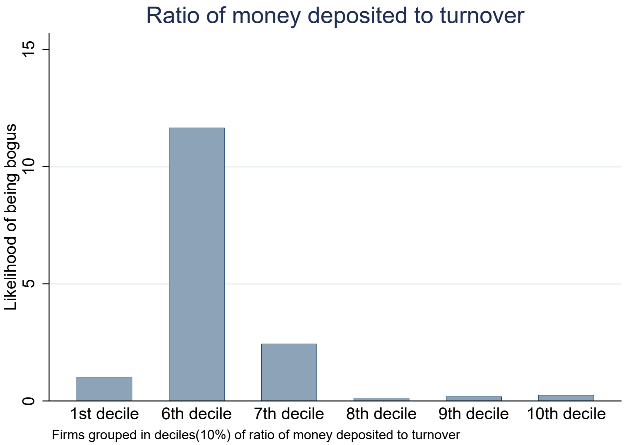
Back

Machine Learning Performance: Aggregation



Back to Results

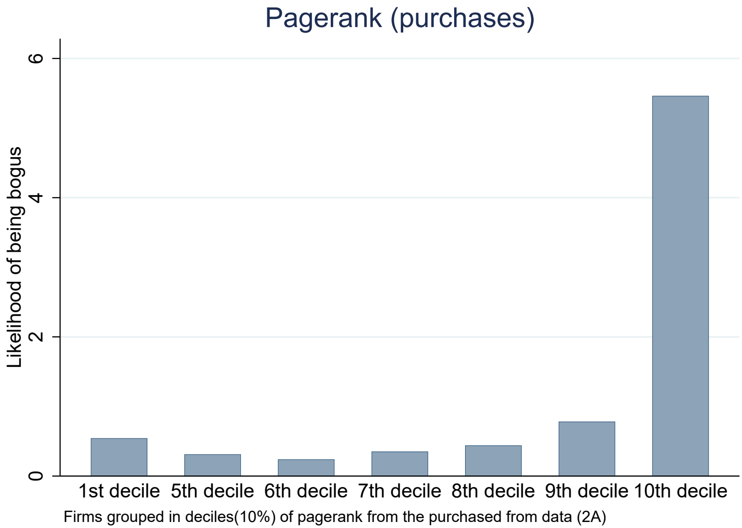
Interpreting Features: Gaming Measures



- Bogus firms likely to have ratio in middle indicates that they know tax authority monitors extreme values so they make sure they are not in extremes.

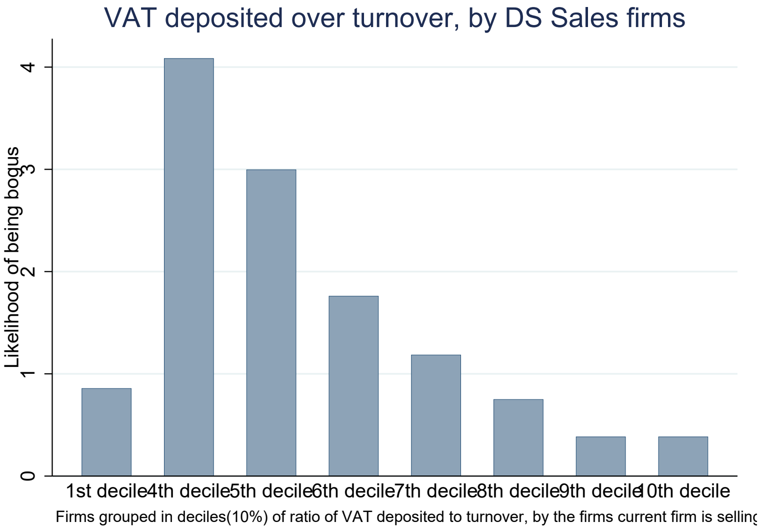
Back to Results

Network feature: Pagerank (suppliers)



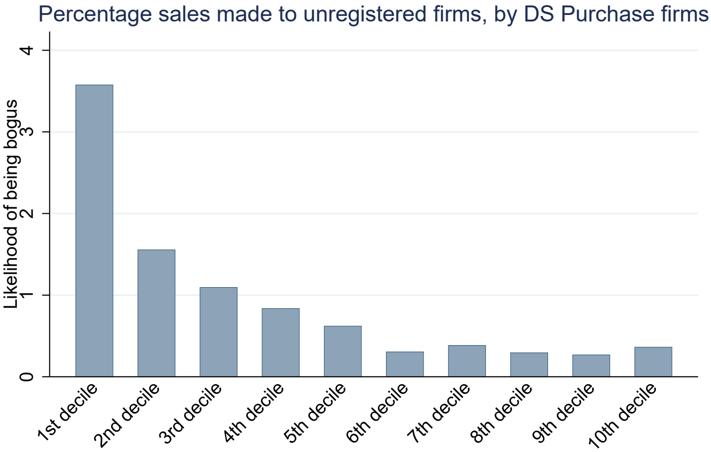
Back to Results

Network feature: VAT deposited ratio of buyers



Back to Results

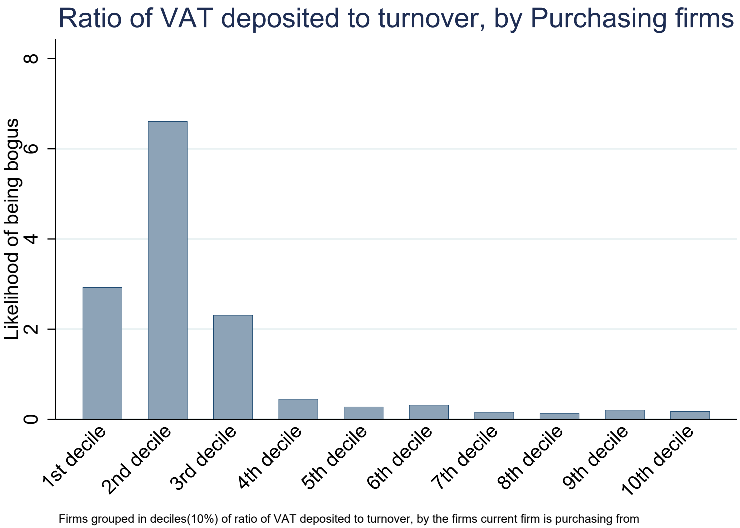
Network feature: Unregistered sales made by suppliers



Firms grouped in deciles(10%) of percentage sales made to unregistered firms,
By the firms current firm is purchasing from

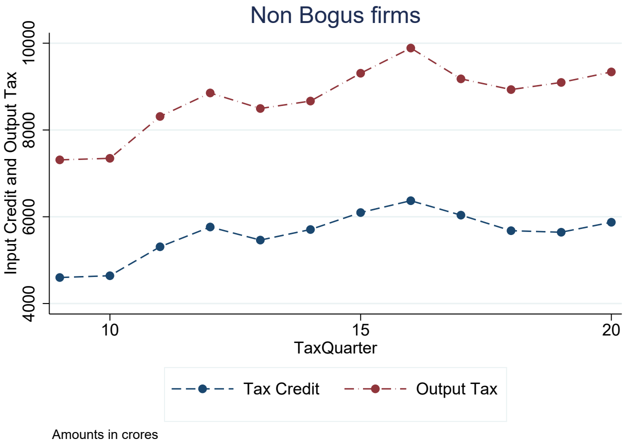
Back to Results

Network feature: VAT deposited ratio of suppliers



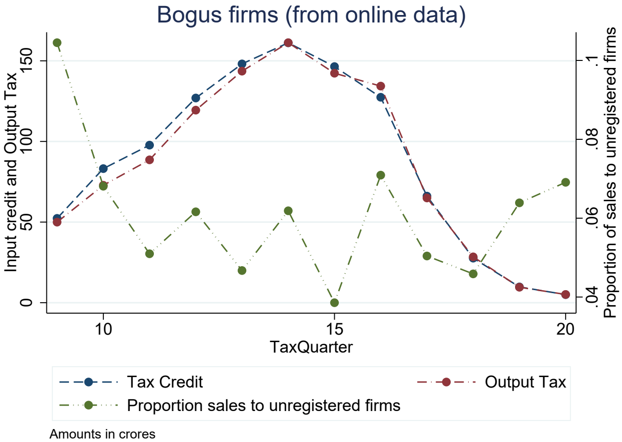
Back to Results

How genuine firms look



- Total output tax reliably larger than input tax credit.

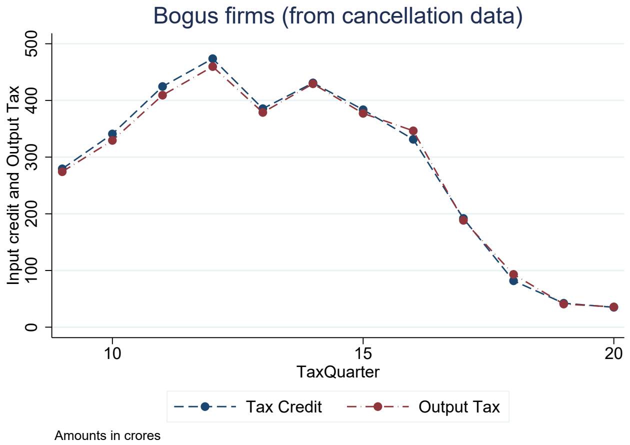
Size of problem: From explicit data



- Input credit claimed weakly greater than output tax declared
- From the limited sample, revenue loss between ₹4-6 billion, annually
- Drop in later quarters due to missing data

From cancellation records

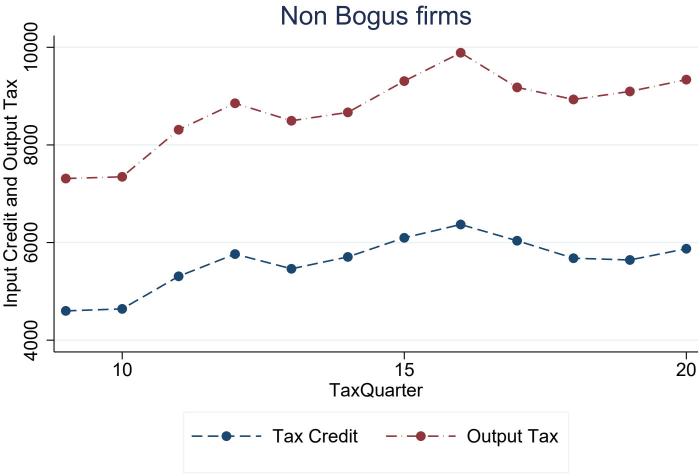
Size of problem: From cancellation records



- From the much bigger sample, revenue loss around ₹15 billion, annually
- Drop in later quarters due to missing data

Back

Revenues Non-Bogus Firms



Amounts in crores