Improving Policy using Better Data and Methods: Machine Learning and Tax Collection

Aprajit Mahajan[1] **Shekhar Mittal**[2]

[1]UC Berkeley (Dept. of ARE)

[2]UCLA Anderson (GEM)

July 21, 2017

Motivation for Machine Learning

- Explosion of (digital) data in past 20 years – (“Big Data”)
  - Mobile Data, Satellite Imagery, Social Networks
- $\implies$ Interest in techniques to analzye data – (“Machine Learning”)
- Increased data in governance contexts (e.g. Aadhar, PDS, DBT).
- $\implies$ Interest in using data to improve governance.
- Examples:
  - Measuring poverty using satellite imagery.[1]
  - Teacher attendance - mobile phone call records.
  - Agricultural yield forecasting using satellite imagery.[2]
  - Analyzing seasonal mobility patterns using cellphone data.[3]

[1] www.unglobalpulse.org/projects/measuring-poverty-machine-roof-counting

[2] www.pnas.org/content/114/9/2189.full

[3] www.unglobalpulse.org/projects/analysing-seasonal-mobility-patterns-using-mobile-phone-data

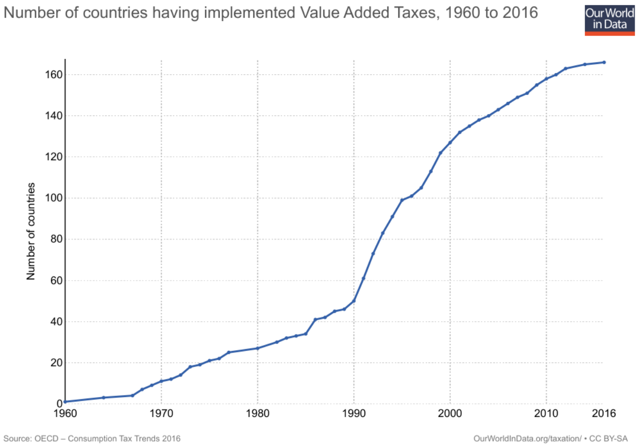
What is Machine Learning?

- Machine learning is a type of statistical analysis optimized for prediction problems with certain kinds of data.
- Key use: Predict one variable (outcome) using a range of other variables (features).
  - In our running example: find “bogus” firms (outcome) using tax return data (features).
  - Use approximately 110 features (many using network information).
- Create decision rules (algorithms) that machine (computer) can use on new data.
- Similar in spirit to linear regression as predictive tool:
  - Key Diff: Large sample size, many features ⇒ Need newer techniques and can generate better predictions.

Machine Learning: Ingredients

- Key Terms: Digital Data, Features (X), Target Variable (Y), Training Set (Estimation Sample), Model, Out of Sample Prediction (Example).
- Model:
  - Contrast with earlier approach: use experts to subjectively identify important features (usually a small number).
  - By contrast ML algorithmically chooses (many) important features (based on what generates better out of sample predictions).
- Important to understand how predictions are used.
  - Input to decision-maker (discretion) vs prediction is final policy outcome (no discretion).
- With machine learning, harder for firms to reverse engineer inspection policy and game it, but this should remain a concern.

Rapid Increase in VAT Adoption Since 1960



- 1 country in 1960 → 50 in 1990 → 160 in 2015.

Reducing Evasion: Third Party Verification

- VAT requires buyers and sellers to provide separate reports of transaction.
  - Opposing incentives should reduce scope for collusion and evasion.
- Whether this occurs, particularly in emerging economies, is an open question. Little empirical evidence.
- Focus on strategy through which firms potentially continue to reduce tax liability.

What are Bogus Firms?

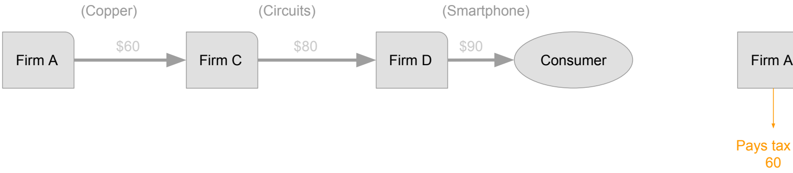
- Bogus firms are paper firms created to enable firms to lower tax bills.
  - This is done by creating fake paper trails of transactions with bogus firms.
- Commonly reported in many VAT systems.
  - Precise extent and magnitudes largely unknown.
  - Media reports estimate the loss, in Delhi alone, to be around ₹2000 crore.[4]
- Paradoxically, "ease of business" norms may ↑ bogus firm prevalence.
- Physical inspections are gold standard for detection but require considerable resources.
  - Audit resources limited.
- One difficulty: Many bogus firms may be mis-classified as return defaulter etc.

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

How are Bogus Firms Identified Currently?

- 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 are sent for verification.
  - Cancellation notice sent if firms not found at the given address.
- Maybe a different process in Tamil Nadu

Bogus Firms: Operation



- Firms A, C and D not necessarily in the same chain.
- Bogus firm can make sales to any firm which needs input credits.

Motivation for Project

- Limited resources for inspections implies that using existing resources to identify bogus firms is attractive from policy perspective.
  - Intellectually: Can we detect bogus firms from VAT returns data?
- Goal: Build and evaluate a tool that can identify bogus dealers from existing data, reliably and regularly.
- Method: Use existing VAT data & small inspection data-set to build 1st iteration of model.
  - Current inspection data-set has important limitations as training data.
- Improve model by carrying out inspections using the model predictions (remove limitations).
- Incorporate tool into the VAT/GST system to enable such predictions routinely.

Data Description

- Entire universe of registered firms (anonymized).
- 3 years of quarterly VAT returns - 2012-13, 2013-14, 2014-15.
- 3 years of quarterly dealer level interactions - 2012-13, 2013-14, 2014-15.
- Dealer profile data.
- Bogus dealer data: 538 bogus dealers identified (2010-2015).
  - **Caveat**: Do not know total list of inspected firms.
- Dealer cancellation data: Use it to identify final firm status.

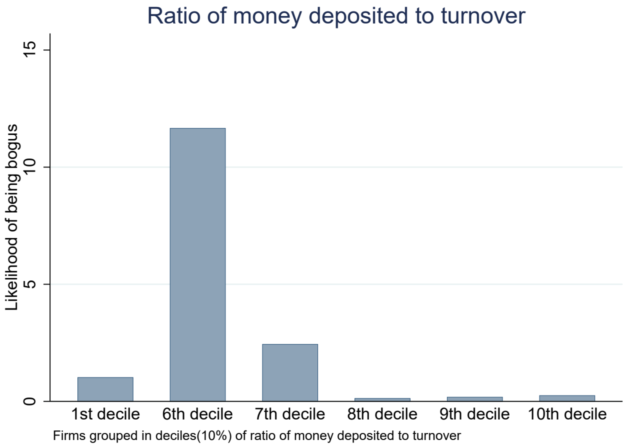
How will GST affect Bogus Firms and Evasion?

- A few types of bogus firms will not be able to operate.
- Machine learning tool still useful.
- In general, firm incentives remain the same.
    - SGST - Dealers still want to claim input credits.
    - IGST - Bogus firms can exist in a different state!
    - Tax compliance levels across states matter.
- Potential extension to incorporate inter-state data.
- Retrain the model to deploy audit resources more accurately.
- Use technology to build a rating system.

Do Tax Return Features Predict Bogus Firms?

- Hypothesis: Besides the usual indicator of low VAT/Turnover ratio.
  - Low sales to unregistered firms/final customers.
  - Purchases from retailer like firms.
  - High local sales.
- But we can do better
  - Let machine learning decide which features are important
  - Create all features we can think of and throw in everything
- Using VAT data & limited inspection data.
  - ‘Key weakness: Inspection data only reveals bogus firms, **NOT** all inspected firms.
- Next step in project: Build inspection data.
- Finally, quantify gain to department from implementing ML.
  - Need RCT to compare ML to business as usual.
  - Most credible evidence of success!

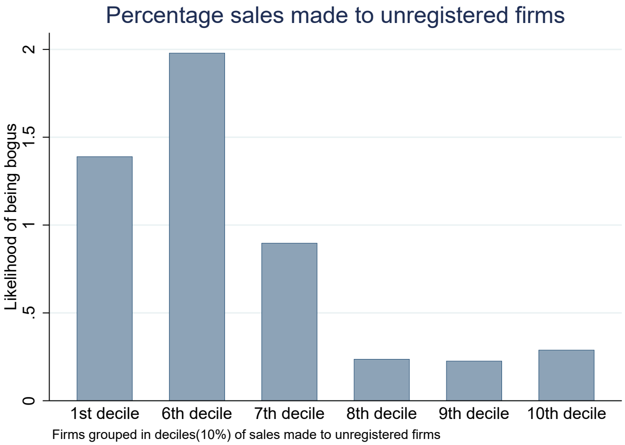
VAT Deposited Ratio (Feature)



ause

- 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.

Sales Made to Unregistered Firms (Feature)

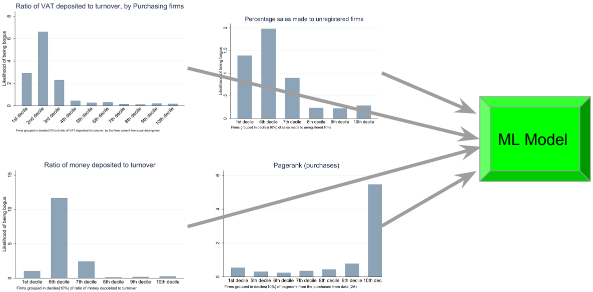


- Sales to unregistered firms is not a feature monitored by the tax officials.

Other Features

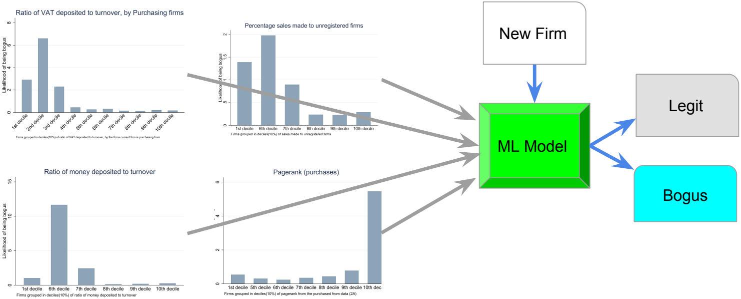
Combine All Features to Build Model

Past bogus firms -> **what is suspicious behavior -> similar behavior in present** -> target

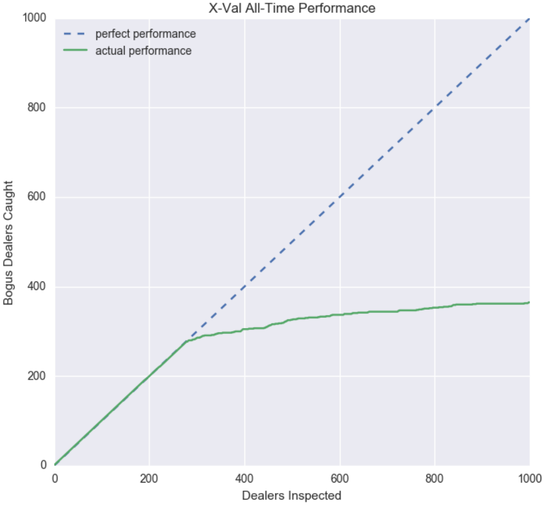


Use Model to Predict on New Returns

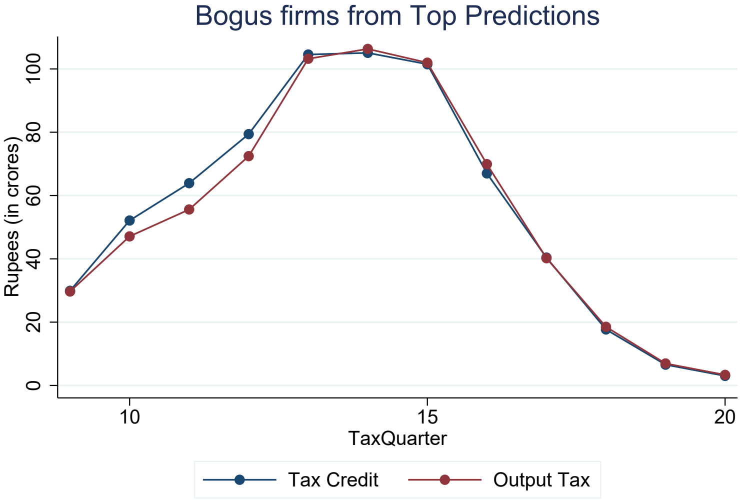
Past bogus firms -> what is suspicious behavior -> **similar behavior in present -> target**


Ratio of VAT deposited to turnover, by Purchasing firms
Likelihood of being bogus
0
2
4
6
8
1st decile
2nd decile
3rd decile
4th decile
5th decile
6th decile
7th decile
8th decile
9th decile
10th decile
Firms grouped in deciles(10%) of ratio of VAT deposited to turnover, by the firms current firm is purchasing from
Percentage sales made to unregistered firms
Likelihood of being bogus
0
.5
1
1.5
2
1st decile
6th decile
7th decile
8th decile
9th decile
10th decile
Firms grouped in deciles(10%) of sales made to unregistered firms
Ratio of money deposited to turnover
Likelihood of being bogus
0
5
10
15
1st decile
6th decile
7th decile
8th decile
9th decile
10th decile
Firms grouped in deciles(10%) of ratio of money deposited to turnover
Pagerank (purchases)
0
2
4
6
1st decile
5th decile
6th decile
7th decile
8th decile
9th decile
10th dec
Firms grouped in deciles(10%) of pagerank from the purchased from data (2A)
New Firm
ML Model
Legit
Bogus

ML Effective in Detecting Bogus Firms



Potential Tax Implications from Top Predictions



Graphs by prediction rankings. Amount in crores. From Q1,2012-13 to Q4, 2014-15

Talking Points

- Revenue recovery not trivial.
    - Not from bogus firms, but from their trading partners.
    - Need to plan deterrence on the trading partners.
    - Consider identifying bogus firms, but not cancelling them immediately. (In GST)
- Advantage of first doing it in VAT:
    - Trading partners should carry over to GST.
- Questions:
    - What are the other internal data sources we can utilize?
    - What other type of analytics can we help with?
    - What is the setup plan?

Resource Requirements

- Need list of bogus firms.
- Need expert interviews.[5]
- Codebook.
- System requirements: 64GB Ram etc.
- Software Requirements: Python, H2o, Graphlab, Stata(?)
  - Eventual integration with SAS feasible
- Engineering human capital: To quickly understand the data

[5] Have a list of questions prepared

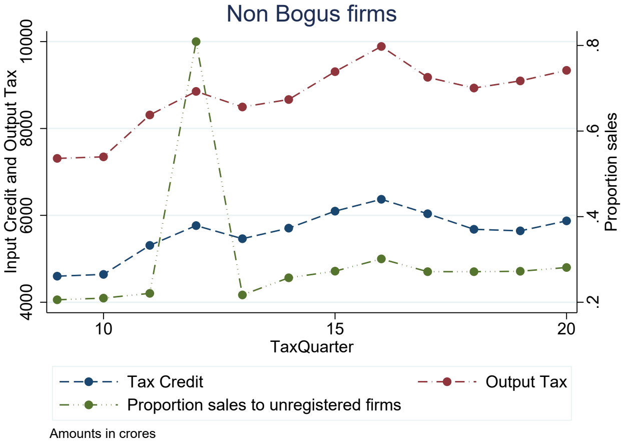
Thanks!

Aprajit Mahajan
aprajit@berkeley.edu
UC Berkeley (Dept of ARE)

Shekhar Mittal
shekhar.mittal.2017@anderson.ucla.edu
UCLA Anderson (GEM)

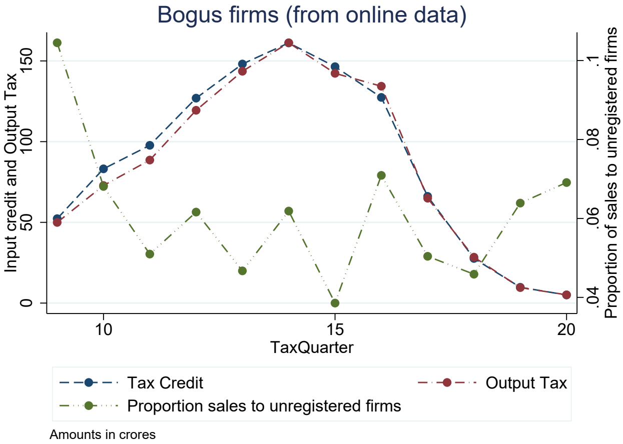
Appendix!

How genuine firms look



- Total output tax reliably larger than input tax credit.
- Proportion sales made to unregistered firms also larger.

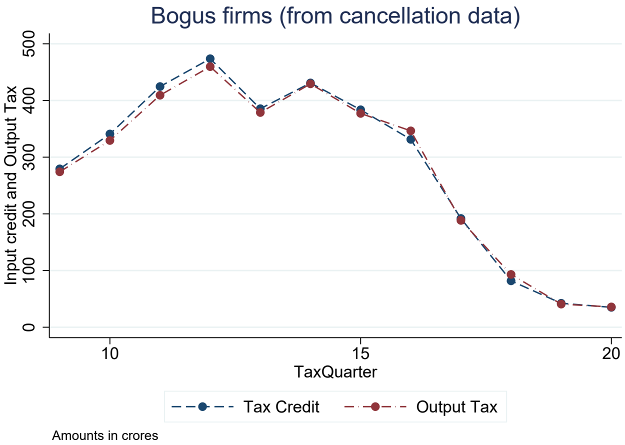
Size of problem: From explicit data



- 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

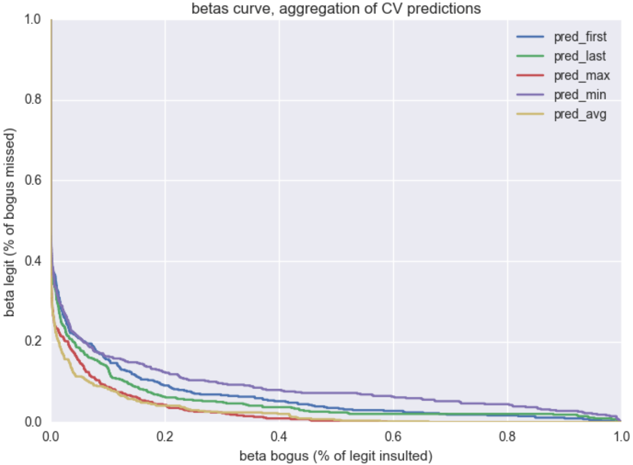
Size of problem: From cancellation records



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

Back

Machine Learning Performance

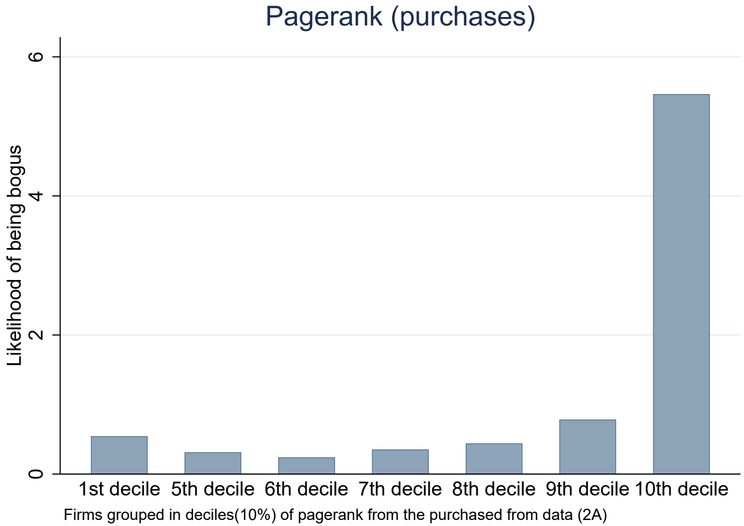


Back to Machine Learning: Ingredients
Back to Results Preview

GST: Other Ideas

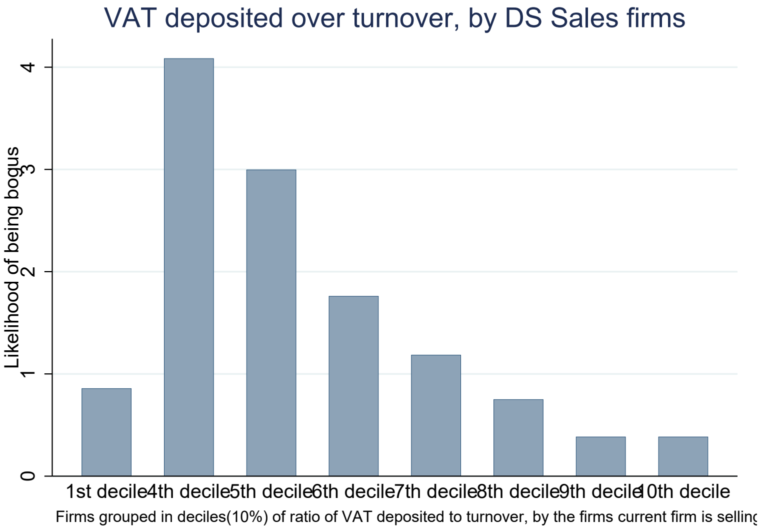
- Two important dimensions of tax reform
- **Equity**: who bears the incidence of consumption tax change? Several possibilities:
  1. Consumers through changing prices
  2. Firm owners through changing profits
  3. Workers through changing wages
- GST reform will change tax rate at the level of the state and the product. This allows us to do the following comparisons (taking prices as an example) to study the incidence of consumption taxes:
  1. Change in price of *a same product* across two states with different tax rate change due to different state level VAT pre-reform.
  2. Change in price between two products *in the same state*, where products receive different rates due to the new several tier rating system of the GST.

Network feature: Pagerank (2A)



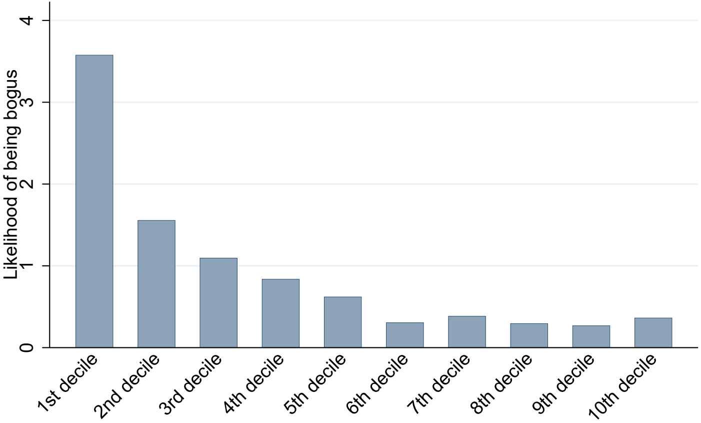
Back

Network feature: VAT deposited ratio by 2B firms



Back

Network feature: Unregistered sales made by 2A firms

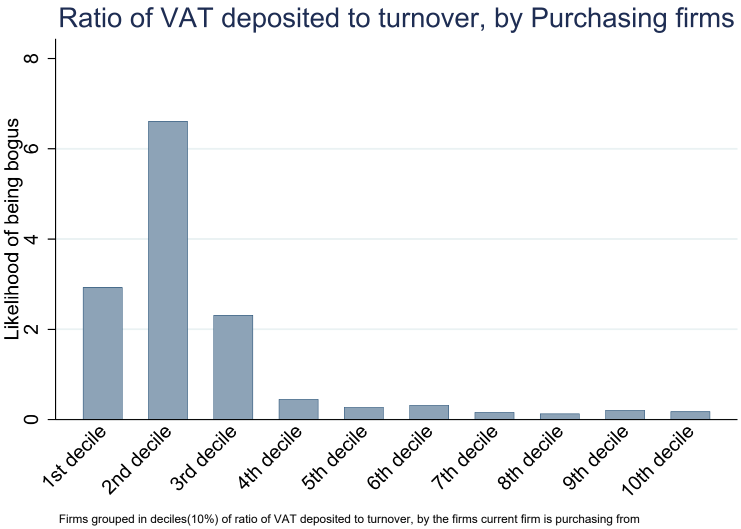


Percentage sales made to unregistered firms, by DS Purchase firms

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

Back

Network feature: VAT deposited ratio by 2A firms



Back