Manoogian Simone Research Fund

Developing alternative methodologies of tax evasion identification: improving imported goods' traceability using machine learning

Final Report

Contacts
Principal Investigator: Vardan Baghdasaryan
40 Baghramyan Ave. 0019, Yerevan, Armenia
American University of Armenia
College of Business and Economics
Tel: +37460612567
E-mail: vbaghdasaryan@aua.am

April 2022
Yerevan

Contents

Executive Summary

The second part of the project had one main and one auxiliary objective:

1. Primary objective - improve traceability of imported goods in the supply chain by utilizing the textual data contained in cash receipts and invoices
2. Secondary objective - investigate the direct and network effects of cash receipt audits using the cash receipt machine geolocation data.

Both of the objectives were mostly achieved, though the process of the analysis has revealed objective impediments which cannot be overcome at the moment. At the same time, the programming and estimation tools deployed in this report can be relatively easily replicated for future use.

The main results are:

1. NLP algorithmic solution to categorize the products in the invoice and cash receipt data.
2. Map of product realization chain from import up to final consumer via multiple transactions. In this regard it is important to underline that the approach is problematic when the good under consideration is massively produced locally as well. In particular, there was no feasible tool to differentiate between local and imported products (details are in the Part I of the report)
3. Under the assumption of 30-40% mark-up the algorithm captures considerable portion of the distribution of imported goods (if the goods were not produced in Armenia).
4. The project team is ready to run the algorithm for another set of 8-10 products, but in order to benefit from this exercise, the following conditions should be satisfied:
    - At least subset of these products should be manually checked (or expert review) by SRC specialists to verify that the algorithm is in fact performing well.
    - The product should not have parallel considerable local production.
5. The quantitative impact (in terms of change in reported revenues) of the tax receipt audit among Yerevan based turnover tax paying businesses was obtained:
    - Direct effect of audit on audited
    - Indirect effect of audit via spatial networks
    - Differentiation of the effects based on effective and ineffective (not leading to fines) audits
    - Short term and long term effects.

All the data and command files are available on the computer stationed at SRC.

PART I. Imported goods' traceability modeling

Introduction

Currently there is an issue of proper matching of customs data with tax returns data. In particular, the customs data includes detailed coding of goods which are based on standard approach (in particular, the Eurasian Economic Union Commodity Nomenclature of External Economic Activity at 10-digit level). The application of product codes in tax documents (invoices, fiscal receipts) instead is optional (or fragmented). Moreover, whenever applied, they are at 4-digit level. Given that the codes are optional, there is limited or no formal control over accuracy of the codes provided either in invoices or fiscal receipts. The only remaining option to track the goods within the supply chain is to use the textual data, which brings up a serious complication due to the high variety of possible inputs for the same item. The task is complicated also due to different metrics applied at various nodes of the supply chain. The good under consideration can undergo packaging or unpacking thus making it difficult to track it in terms of quantity and price.

The research aims to apply a combination of machine learning and analytical tools to facilitate the tracking of goods in the supply chain. The economic purpose of the tracking is in line with our overarching research theme – improving fraud detection capacities of SRC.

This applied research has a number of contributions. First of all, we apply different natural language process (NLP) approaches to classify the textual data into product categories and items. Obviously the task is complicated due to the Armenian language used. Second, the network of taxpayers involved in the trade of the given goods is constructed, which enables to attach specific imported goods to a set of taxpayers making final consumer sales.

The main milestones of this part of the projects are:

- Classify the product definitions in invoices into predefined categories.
- Construct the buyer/seller network for each product and try to track the imported product flow using invoice data.
- Check and quantify the network members' activity in retail using tax receipt data.
- Analyze the product chain, quantities and monetary value.

Methodology

In the first stage of the project a list of products was confirmed for further analysis. The list includes around potential 40 products names labeled with 4,5 or 6-digit codes. We have tried to strike a compromise between complexity and verifiability (including using expert opinions).

It is worth mentioning that all textual descriptions in the data are in Armenian which makes the analysis and the usage of state-of-the-art ML approaches complicated because of a very limited NLP research available in Armenian, thus all the textual information required for the task was translated into English using the Google Translation API. The translation not only handled the lack of techniques ready to be applied to Armenian text, but also in many cases was a good tool to handle grammatical errors in the text as the algorithm behind has auto correction functionality as well.

The main resources used to solve the task were

- ➔ the import data with labeled product codes (11 digits), product definitions
- ➔ Tax receipt data with human labeled 4-digit ADG codes (may not be 100% correct) and product definitions
- ➔ Invoice data with product descriptions only (textual data)

The main approach used for product categorization was text similarity approach based on the descriptions for imported goods. The approach consists of 2 main components: first is the keyword extraction and second the text similarity capturing.

For keyword based similarity identification unsupervised keyword extraction algorithm YAKE (Campos et al., 2018) was used. It is a frequency based automatic keyword extraction framework. The main advantage of the algorithm is that it does not use any fixed dictionary or 3rd source information for extraction and it relies on features extracted from text, making it applicable to textual documents in different languages and domains without the need of prior domain knowledge.

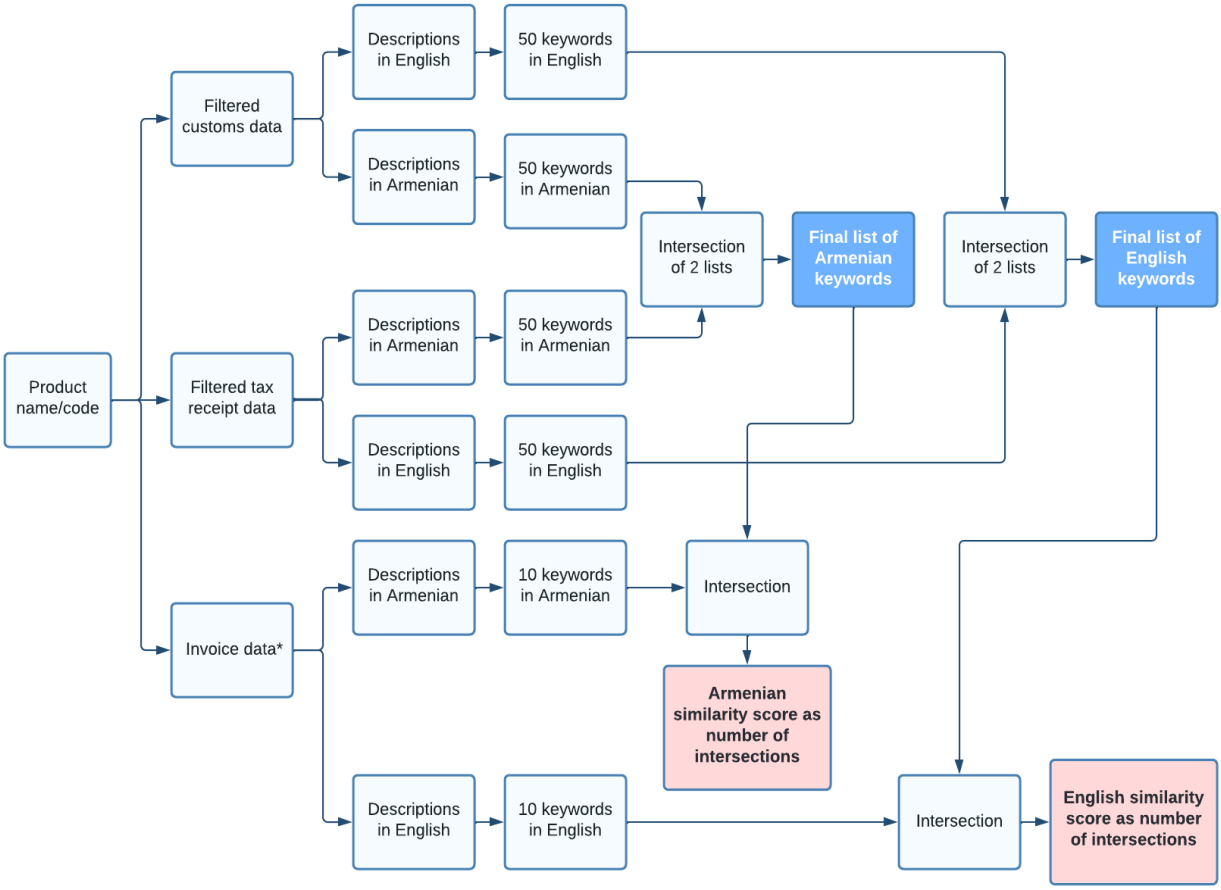
As the algorithm is language independent, it was applied to Armenian versions of product descriptions as well. We first extracted keywords from text (both from Armenian and English) and then worked with extracted words instead of full descriptions.

The list representing the intersection of top 50 keywords extracted from labeled descriptions in customs and tax receipts data was built. Then for each product definition in invoice data a

similarity score was calculated as the number of intersections or the number of keywords with similarity above threshold for Armenian and English respectively.

In order to calculate similarity between 2 text entities, the text is transformed into vectors and then cosine similarity between 2 vectors is computed.

Figure 1. Retrieving the text similarity from various transaction documents.*

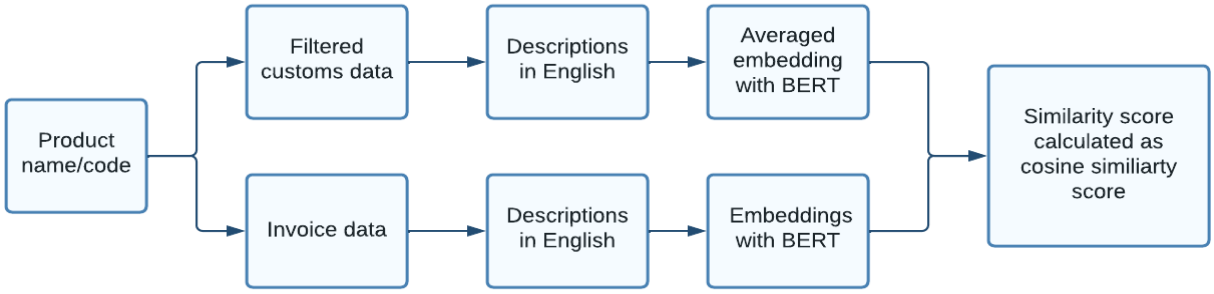


* Invoice data cannot be filtered by product code. This it is only filtered based on TINs which imported the given product as a Supplier.

The second component of the similarity approach was the full description similarity estimation. For this purpose, Bidirectional Encoder Representations from Transformers - BERT (Devlin et al., 2019) model was used. Transformer architecture is one of the latest achievements of researchers in the deep learning sphere and is widely applied in natural language processing. We use a pre-trained BERT model developed and pre-trained by Google on a huge textual corpus and designed in a way to take into account the context and to be able to learn contextualized embeddings. The other reason to choose this model for the task was the extensive use of BERT and its predecessor models for product description classification in the literature. Specifically, Jahanshahi et al. (2021)

aimed to classify the product descriptions of several top online grocery platforms in Turkey. They compared the traditional methods such a bidirectional LSTM with more advanced NLP language models (BERT, ROBERTA etc.). While none of the techniques was an absolute winner, the advanced techniques outperformed in most cases. In ProBERT: Product Data Classification with Fine-tuning BERT Model (Zahera et al., 2020) the authors fine-tuned the pre-trained BERT model for multiclass classification and used it for the end-to-end classification. The advantage of the approach is that it solves the task with a single model and ProBERT will provide the class given the description, but it is not scalable and will need repetitive training whenever new categories appear. With our similarity based approach we tried to exploit the ability of the pre-trained BERT to provide representative embeddings for product descriptions and bypass the scalability issue.

Figure 2. Using BERT to derive cosine similarity score



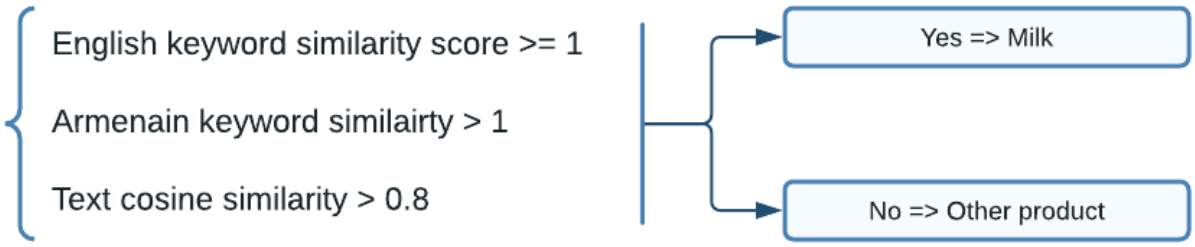
As a result, 3 similarity scores are generated, among those are:

- Intersection of keywords in Armenian. For this point only the invoice descriptions having more than 1 keyword that match the keywords in the extracted list are assigned the given product class.
- Similarity of keywords in English. The criteria of similarity for this approach is to have at least one keyword from the keywords list that was built on labeled description translated into English.
- Similarity of full description. For similarity estimation of full descriptions, cosine similarity was used as a metric and the similarity threshold equal to 0.8 was taken as the optimal one for text matching.[1]

[1] The similarity score varies between 0 and 1. The threshold of 0.8 was identified based on manual check of the description after applying the threshold and by taking into account that it is better to miss several milk products (false negatives) then to misclassify many other products as milk (false positives).

Finally, the description was categorized as a specific type of product only if the above-mentioned 3 criteria are satisfied altogether. The decision criteria are summarized in the figure below.

Figure 3. Categorizing product as milk based on 3 criteria (English similarity score, Armenian similarity score, description similarity score).



*Note: The similarity threshold can be adjusted based on some labeled examples by SRC staff.

Results

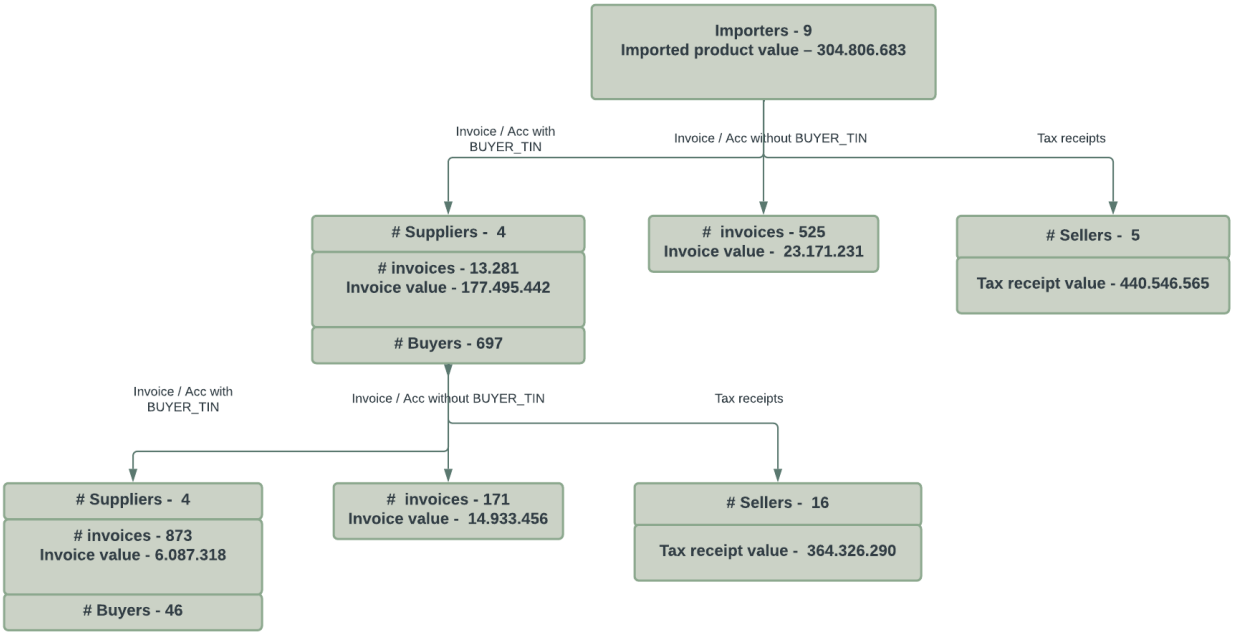
The experiment was implemented on 2 products, milk and rice. The choice was made based on several characteristics of these products. First, milk has a much shorter cycle than rice. Secondly, milk is widely produced in Armenia while rice is mostly imported. Third, there are relatively limited number of importers which makes it easier to verify the preliminary validity of the network obtained.

The network is constructed based on invoice and tax-receipt data. It is worth mentioning that in the invoices the seller has a legal opportunity to not disclose the buyer. So wherever the buyer information is provided the network goes deeper while when it is missing or when the products are sold with tax receipts it stops.

Milk is equally imported and produced locally. Given the fact that exact differentiation between imported goods and local production is not possible as there is no identifier, we apply a price filter to separate local production. The median price of the imported goods was used as threshold for filtering.

The chart below shows the supply chain of milk from import to final suppliers.

Figure 4. Chain of milk distribution: from import to final consumer.



The network is constructed from as many layers as the trade continues with invoices (i.e. the next buyer is not a final consumer). The layers end up at the point where no more sales with invoice is done by the buyers of the previous level which means that the imported goods somehow reached the final consumer.

Each layer of the network has 3 nodes.

- The first one is the main track which represents the acquisition by the buyer with proper transaction parties' information and where we are able to follow up the next steps.
- The second node shows the invoice sales with missing buyer information and here the flow is interrupted because of missing buyer data. Consequently, the network was untraceable there.
- The last node gives information on sales with tax receipts, meaning sales to the final consumer in retail stores. These nodes are the ones representing the consumption of the milk by the population and thus the end of the supply chain.

Overall, the total amount of milk consumption can be calculated as the sum of the sales with tax receipt in each layer of the network and the sales with invoice where the buyer is not identified. We should take into account that the sales invoice with missing buyer identification may not in fact be directly consumed, thus we need to add some markup there to simulate the reality. To see how successful was the identification of the given product along the network we need to compare

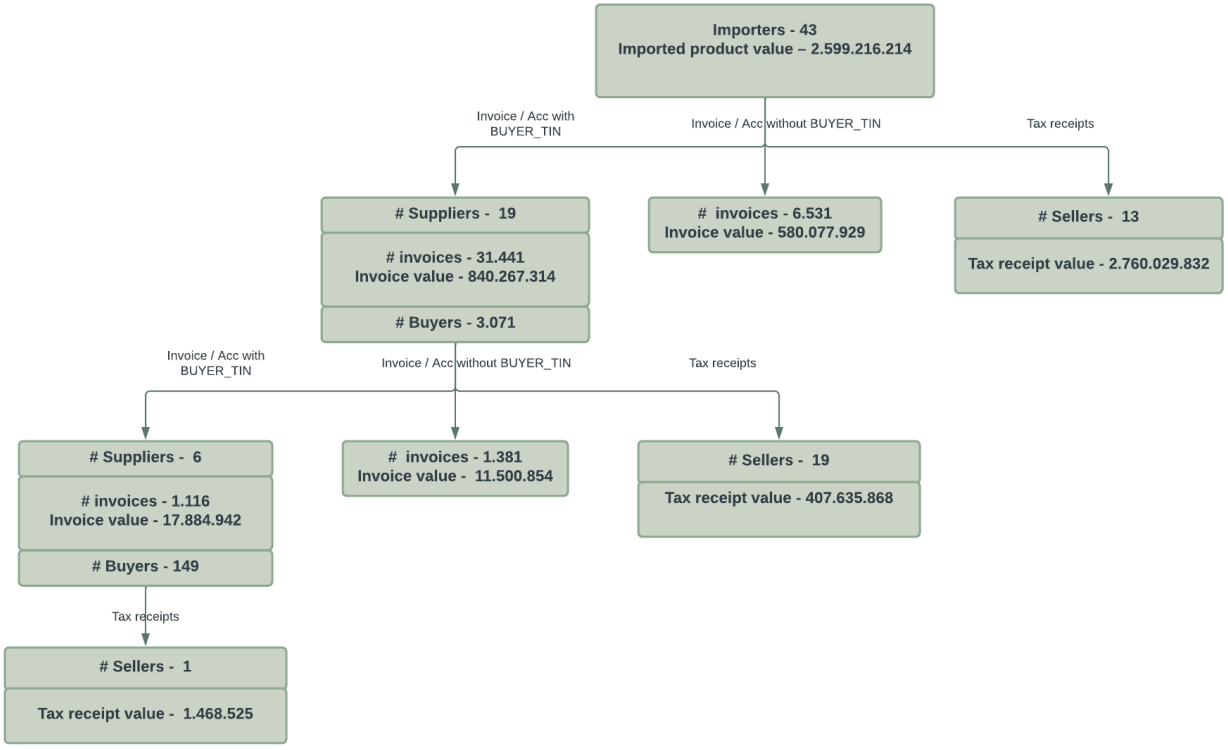
total final consumption to the imported goods and the difference should be approximately around the imposed markup, which will be considered 30% in our case.

The final consumption of milk identified in the network exceeds more than 3 times the imported amount which means that the applied unit price > 500 dram filter was not able to fully filter out local production.

To validate the above mentioned statement that the mismatch of the import and consumed amounts of milk was not because of misclassification but local production the same experiment was implemented on rice as well, given that rice is not produced locally. Consequently, no price or other filters were applied when constructing the network for rice/

As mentioned before, the final consumption can be estimated with the sum of tax receipt sales and invoice sales where the buyer is not identified in all levels.

Figure 5. Chain of rice distribution: from import to final consumer.



Overall, 2,599,216,214 armenian dram rice was imported and the amount found in the final nodes of the network representing the final consumption is around 1.1 times the imported amount. As mentioned before the calculations are done based on the assumption that the markup fin the whole chain rom import to final consumer is 30%, however the results are not too

sensitive to markup change. This shows that almost all the imported rice was identified with the approach adopted by the team and proves the assumption that the reason behind the mismatch of the milk amounts was the local production.

Potential replications for other products

To validate the findings of this, exercise the project team is ready to replicate the analysis for another 8-10 imported products. But to ensure that the exercise is meaningful and effective the following conditions should be satisfied.

- The product should not have considerable local production, otherwise it becomes technically impossible to differentiate between local and imported production
- It would be preferable to run the algorithm on at least few products that were manually/analytically analyzed for the same purposes by the SRC staff.

These two conditions would enable us to validate the results of the NLP algorithm used.

PART II. Audit spatial network effects

Introduction

The purpose of this analysis is to understand the short run (and possible mid-term) behavioral response to own or neighbors' on-site revenue audits of the small and medium sized taxpayers engaged mostly in trade and service.

Tax evasion is a crucial public finance problem in developing countries (Fuest and Riedel, 2009) and the permanent search for optimal tax administration toolset is now being enriched with data intensive approaches (Baghdasaryan et al., 2022) and behavioral interventions (see Antinyan and Asatryan 2020 for detailed review). Combined, these two approaches enable investigation of network effects - either spatial or utilizing other proximity dimensions.

Our contribution is twofold. First, we utilize high frequency data on reported sales which is aggregated on a daily level obtained from electronic cash register machines (ECRM). Combined with on-the-spot methodology of the tax audit it enables measuring the immediate response of the taxpayers concerned, minimizing the impact of possible confounding factors which become more probable if the time window is wider. Second, we utilize the natural experiment, i.e. the random event of audit happening among any of the k nearest neighbors, to understand the spatial network effects. In this regard we try to uncover the differences that might exist between various types of taxpayers - audited and not audited, fraudulent and compliant, among others.

Current work touches upon a number of strands in the literature. The analysis of taxpayers' response to their own audit is analyzed in experimental literature under two different hypotheses. First is expecting improvement of compliance due to experiencing tax audit and is explained either by availability heuristics (Tversky and Kahneman, 1974), when the mere event of audit makes it more salient and the taxpayers consider its repetition in the future more probable. Alternative explanation is the target effect, when the subjective probability of being audited again increases due to the fact of already being audited (Hashimzade et al., 2012). Second hypothesis works in the opposite direction and is dubbed in literature bomb crater effect (Mittone, 2017). It is manifested by sharp decrease in compliance after the audit and the theoretical underpinnings in the behavioral literature is the gambler's fallacy - miscalculation, or better, misperception of the event's probability (Maciejovsky et al., 2007). A study by Mittone et al. (2017) challenges this misperception hypothesis in the lab experiment. In particular, by deriving various degrees of

“bayesianity” of the subjects, they demonstrate that correct bayesian updating of audit probabilities is still associated with bomb crater effect. As an alternative explanation the authors propose the ‘duality’ in the taxpayer’s behavior by stating ‘the probability of being audited is correctly computed using all of the relevant information in the correct way, but when the agent has to decide whether and how much to comply, she is driven only by the emotions that are triggered exclusively by own audits’ (Mittone et al., 2017, p.4). In our paper we start the investigation by looking at own audit impact on subsequent behavior. Unlike majority of literature in this field, we conduct it by analyzing actual audit and reporting data from the field. In addition, in our setting the behavioral changes are immediate - daily reports of sales registered automatically every time the cash receipt is printed. We believe to be bringing in important evidence from a developing country, which are more prone to tax misreporting and evasion.

The second strand of literature we are contributing to is the one dealing with audit effects which span beyond the audited entities. There is recently an emerging literature that looks into network effects. Lediga et al. (2020) investigate spatial network spillovers of tax audits in South Africa. Using annual tax returns and panel data with fixed effects at various levels the authors measure how the audit happening among the neighbors of the taxpayer affect the reported taxes. The effects are positive, but short-run and declining in the distance of the audited neighbor. Also in this type of studies the experimental approaches prevail. Boning et al. (2020) look at how the information about the audit defusing via the network of common tax preparers affects the collection of employment taxes among non-intervened (visit by IRS or letter) entities. Drago et al. (2020) consider similar experiment – by sending out various types of messages in Austria to trigger TV license fee payments. Their main finding is the existence of central nodes, houses located in particular node of the network, that have the highest propagation among the neighbors. Our study contributes to this direction of research by unveiling immediate responses to audits happening in neighboring entities, discriminating between effective and ineffective audits, as well as considering audited and never audited entities.

Methodology

To leverage the results of our current and past efforts we have adopted a slightly different approach, namely, utilization of “natural experiment” framework. Our team has assisted the SRC to come up with comprehensive information on the location of taxpayers within the city of Yerevan. The objective of this spin-off study is to utilize the possibility of merging spatial information with tax audit information and daily reported sales. In this part of causal research, we obtain understanding of the effects specific type audits have on the taxpayer reporting behavior. The following model is estimated using daily reported turnover of companies working under turnover tax regime:

$$y_{it} = \sum_{a}^{A_j} \sum_{d=-D}^{D} 1\left(t - e_a^j \middle| d\right)\beta_d + \sum_{b}^{B_i} \sum_{d=-D}^{D} 1\left(t - e_b^i \middle| d\right)\alpha_d + \delta_t + \gamma_i + \varepsilon_{it}$$

where y_{it} is the daily turnover of taxpayer *i (*or log of turnover*)* on day *t, a* is the audit event in any neighbor taxpayer *j* and provides the calendar date of that event, *b* is the audit event in the taxpayer *i* and provides the calendar date of that event. $\delta_t$and γ_i capture the time and taxpayer fixed effects. The event window is determined with the selection of window (-D, D) and must be such that not overlap with the next event. In this framework β_d and α_d is a set of dummy variables (each) that would capture change in behavior due to neighbor and own audits respectively. Sandler and Sandler (2014) using Monte-Carlo simulations demonstrate that this specification captures the event effects in most efficient way.

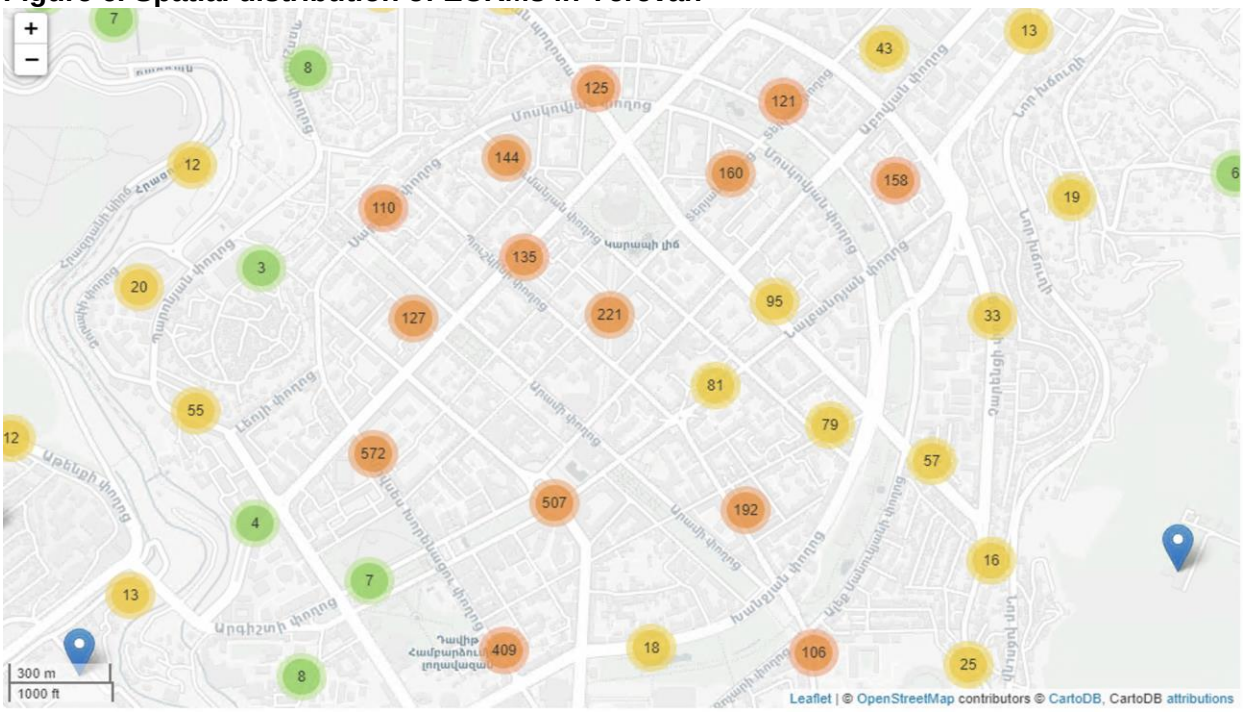
Effectively this means we are estimating the effect of not only and not as much of the own audits but of the audits happening around the company. The auditing of neighboring companies can have a significant effect on the behavior of not audited companies in terms of the incentives of reporting the true revenue. Potential communication between the closely located companies and the financial risks of being audited can lead to significant behavioral changes mostly in the short-run. In order to analyze the spatial spillover effects of auditing neighboring companies (C. Lediga et all., 2019) the specific form of data is required. Most specifically the identification of neighboring companies and the daily aggregated receipt data for evaluating the behavior of enterprises.

The identification of the nearest companies was quite challenging because of the large number of unique pairs, that's why the two-step approach was used. Within the scope of the previous project, the correction of addresses was implemented with the identification of their corresponding

latitude and longitude coordinates. Based on that data during the first phase the k-nearest neighbor algorithm was applied, which used distance of coordinates for selecting 50 closest companies for each enterprise. Afterwards the other check based on the mile distance of the enterprise and its subset of neighbors was used for selecting the nearest firms within 200-meter distance. Next, the derived neighbors' dataset was merged with the receipt data and with audit data for creating a complete dataset for further analysis of audit effects.

As a result, the total dataset includes around 12,000 companies operating in Yerevan, paying turnover tax, using electronic cash receipt machines and having exact machine location information. Moreover, these twelve thousand was obtained after eliminating from the database companies that were idle (reporting zero revenue) for more than half of the 2019. To the best of our knowledge, the remaining data is the overall population satisfying the outlined conditions.

Figure 6. Spatial distribution of ECRMs in Yerevan



Data framework

The cash receipt audits are implemented as a result of preparing a special list of susceptible taxpayers based on the analysis of the data and/or control visits-observations implemented in the fields. The objective parameters that might trigger such audits are:

a) Decrease of revenues over the reporting period
b) Absence of seasonal variations of reported revenues in the reporting period for the entities where seasonality is expected (e.g. restaurants and cafes)
c) Information received by third parties about the entities not providing cash receipts
d) Absence of variation in the revenues reported when there is inflation in the product or services offered by the entity

Based on these (or potentially other) factors the tax inspectorate opens a case which provides 10 days' time window to implement tax audit, which is implemented by conducting "control" acquisitions and verify that cash receipts are duly provided.[2] So within the audit database we have audit start date and audit end date (see footnote 2), as well as we have information about the fines payable. If not fines are payable, then the taxpayer is considered compliant.

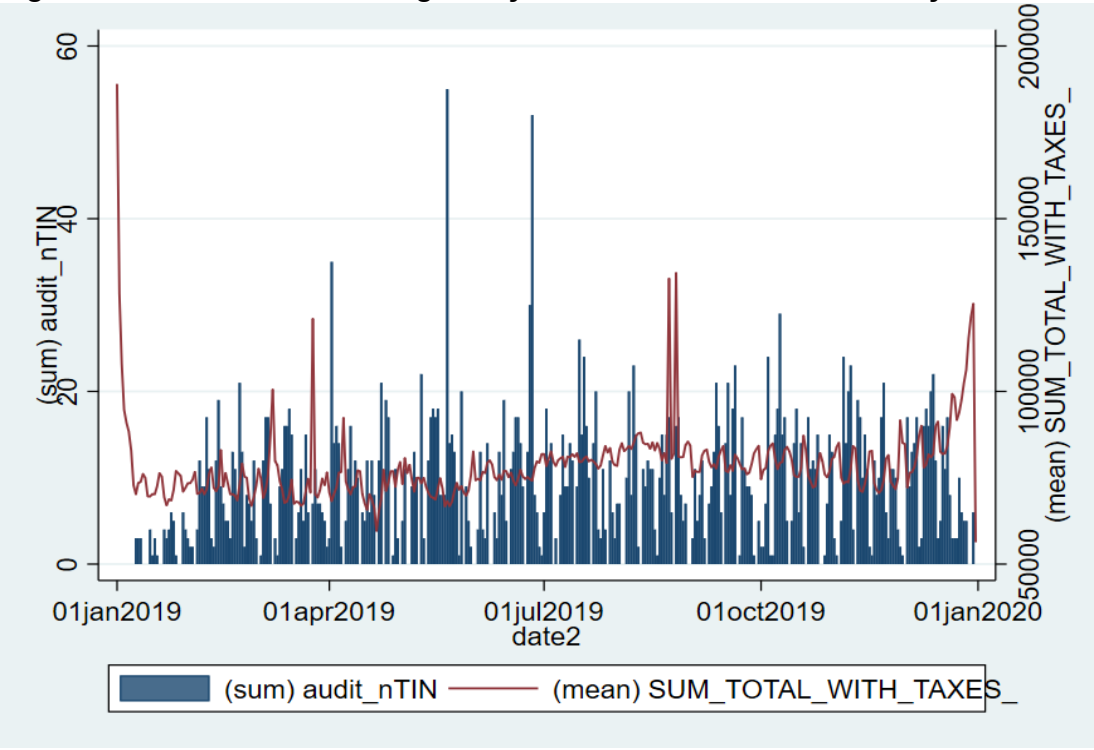
The remaining data is also daily and is composed of the population of all taxpayers operating under the Turnover tax regime in the capital city of Yerevan in 2019. More precisely, the unit of observation is the ECRM-taxpayer given that some unique taxpayers have more than one ECRMs. For each observation we have daily reported revenue as recorded by ECRM and number of cash receipts issued (number of transactions). In addition, we have information about the economic sector of activities. For each observation, following the methodology, we identify up to 10 nearest neighbors in the radius of 200 meters. In the context of daily panel data that we have this network information is utilized in the following way: a set of dummy variables is created - 10 for the event of audit (equal 1 if the audit takes place on that day in that neighbor, and zero otherwise) and 10 for event of effective audit (equal 1 if the audit resulted in fines being paid, and zero otherwise). Table 1 summarizes the basic descriptive statistics of the variables used in the analysis, whereas figure 7 presents the same data but over time.

[2] During these 10 days more than 1 control acquisitions can be implemented. If by the end of the period no irregularities are detected, then on the 10th day the taxpayer is informed about the fact of the audit and signs respective protocol. In case irregularity is detected during that period, the taxpayer learns about it on the spot, without waiting for 10 days to expire.

Table 1. Descriptive statistics of main variables used in the data, frequency of audits and their effectiveness

	Observations:	Out of which:		Audits	Out of which:	
		Trade sector (G47)	Other		Effective	Ineffective
Unique ECRM-TIN	12,338	7,630	4,708	2,674	1,765 (66%)	909
Unique TIN	10,180	6,495	3,685	2,334	1,553 (67%)	781
Based on Unique ECRM-TIN pairs for the whole period of 2019						
	Observations	Mean	St.dev	Median	Max	Min
Revenue	12,337	75,783.16	191,037.4	29,719.7	5,420,660	1
Number of receipts issued	12,337	21.7218	49.55444	6.303226	1322.553	1

Figure 7. Distribution of average daily revenues and audits over days in 2019



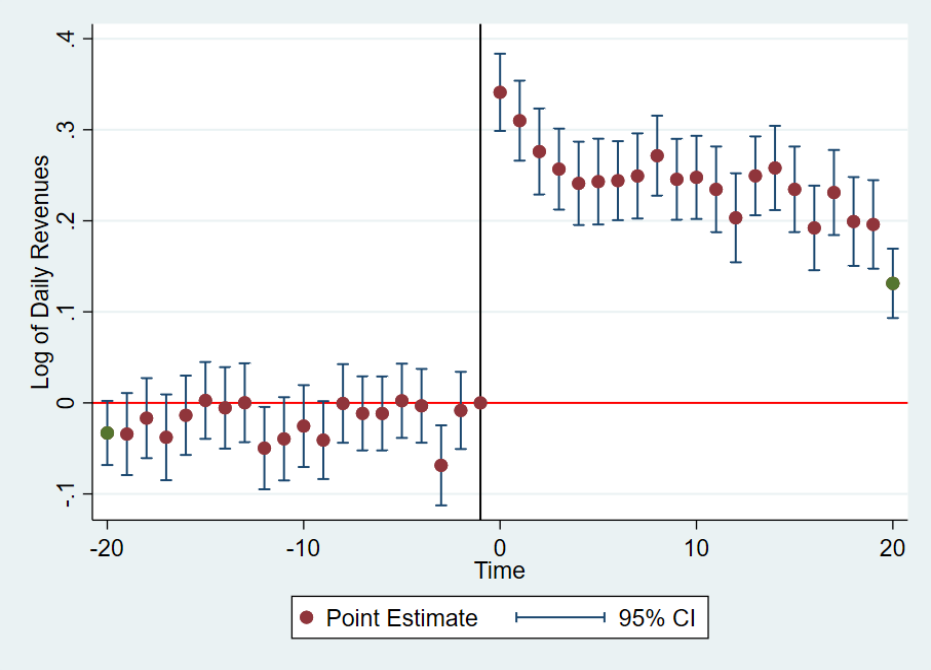
Note: The date is the audit process initiation date (number on left axis). Red line represents the average daily reported revenues of all (more than 10,000) taxpayers.

Results

a) Own audit results

We start presenting the results with the effects of taxpayer's own audits on reported turnover. The detailed regression tables are not reported in this document.[3] Figure 1 reports the effect of audit for those entities who were audited and they were found non-compliant, which gave rise to fines payable. Time 0 on the horizontal axis indicates the day of the audit (starting date), points to the left of it are 20 days after the commencement of the audit and to the right (negative) - 20 days before the audit. Log of daily reported revenues obtained via final sales and through issuing of cash receipts are on the vertical axis, hence the effect of audit can be interpreted in terms of percentage points. Thus after the audit, for the first 20 days the audited and fraudulent companies report on average 20-30 percent higher revenues, whereas the accumulated long-term effect is still positive, at around 13 percent. Very similar picture is obtained if instead of audit start date audit end date is used. We infer that overwhelming majority of the taxpayers who are not compliant are categorized as such very close to the starting date of the audit.

Figure 8. The effect of on own fraud on reported turnover.

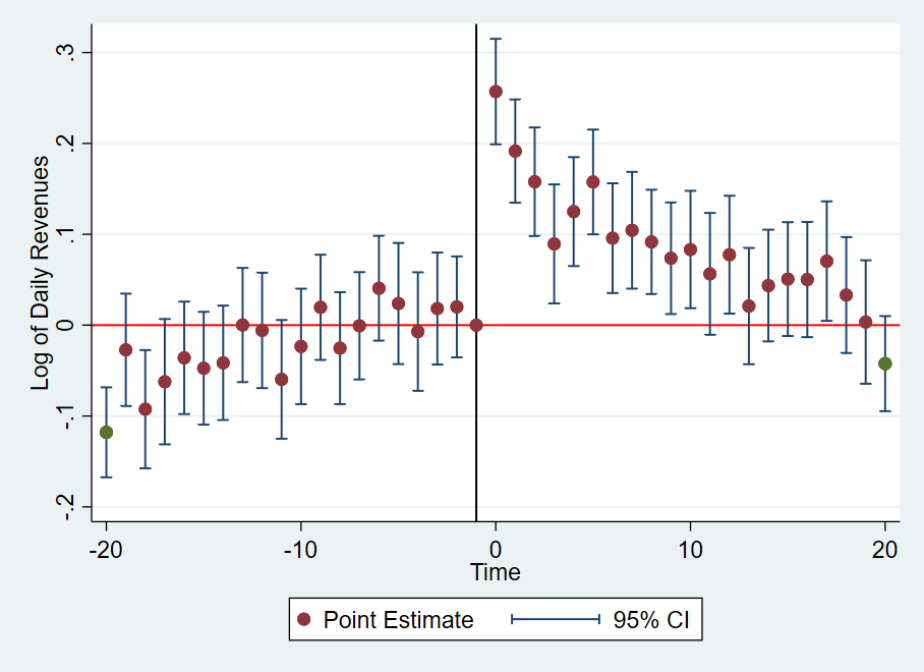


Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

[3] All the codes to replicate the analysis will be transferred to SRC and stored. SRC staff is free to use it and modify as needed.

Similar analysis is repeated for the effect of audit on those taxpayers who were found compliant. But here instead of using audit start date, we apply audit end date. We do this because if found compliant, the taxpayer should learn about the audit only at the end of it, when he is called to sign the protocol of audit. Hence her behavior change triggered by the audit should be observable after that event. Interestingly, in spite of the fact that the audit was not effective and no irregularities were detected, we observe the change in reported revenues for them as well. Though in this case the effect is weaker and is not permanent. Over the subsequent 20 days after the audit ends the taxpayers report 10 percent higher revenues, but the effect dissipates over time, as witnessed by the rightmost green point indicating the cumulative effect for the remaining periods.

Figure 9. The effect of on own audit (found compliant) on reported turnover.

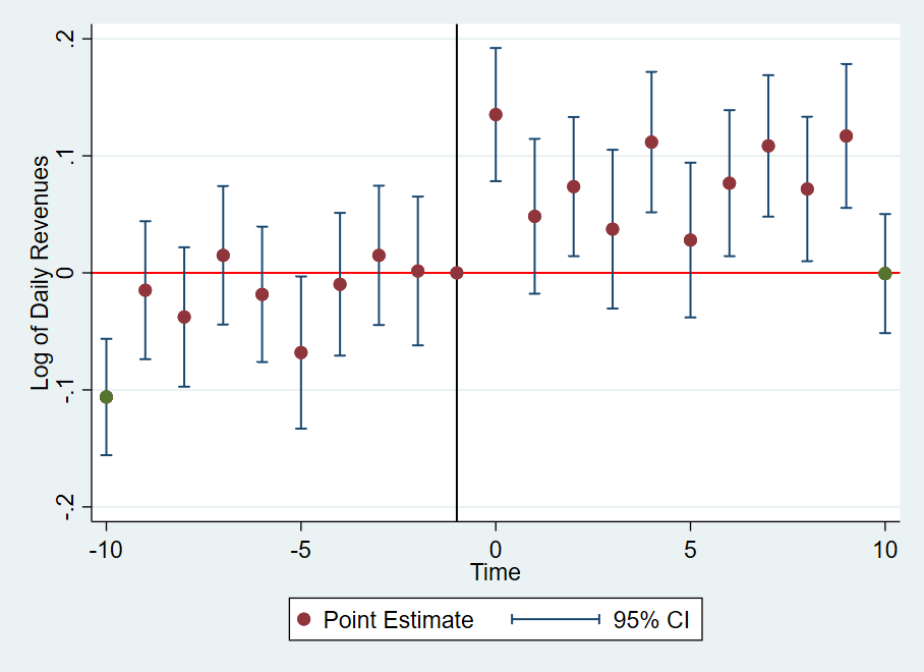


Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

In addition to these estimates, we test one additional hypothesis which deals with possible explanation of non-effective audits – information leakages about audits being conducted. Here we do not claim any intention on conspiracy, but it cannot be excluded that in certain cases the result of audit is negative not because of absence of systemic misreporting and is explained with

the fact that those audited have some information about the event taking place. To test this hypothesis, we estimate the behavior of those taxpayers who were found to be compliant within the 10 days' time window commencing from the audit start date.

Figure 10. The effect of own audit which should be unknown to the audited.



Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

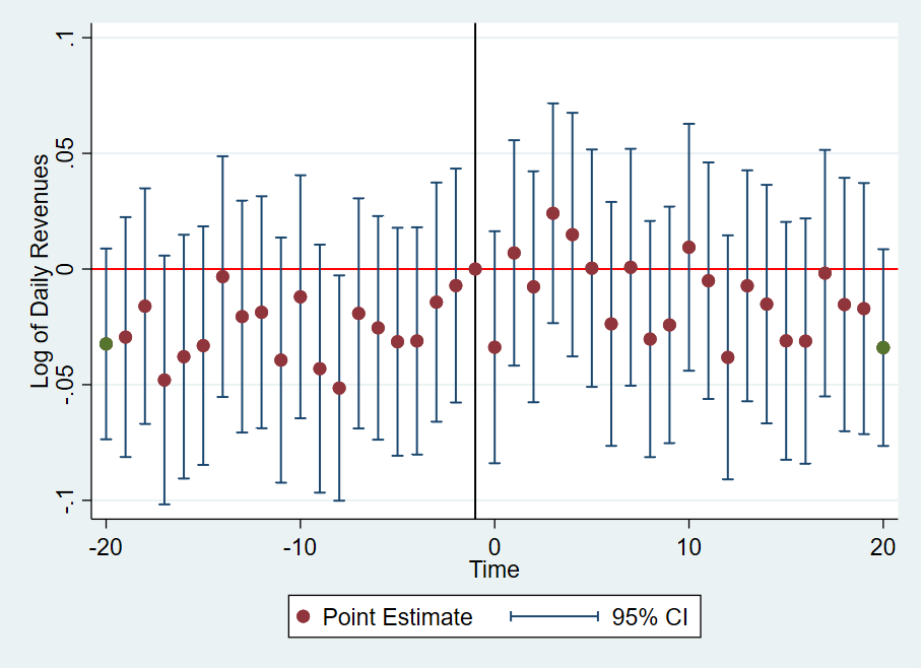
The results are not clear cut and consistent in terms of clear trend, but overall there seems to be some slight upward movement in the reported daily revenues of those taxpayers who were subject to audit and supposedly should not be aware of it.[4]

[4] One possibility is the audit happening around the taxpayer which is analyzed in the next section.

b) Spatial network externalities of tax receipt audits.

Now we turn to reporting change in behavior of the taxpayer due to tax receipt audits happening around it. In this version we report the analysis when only 3 closest neighbors of the taxpayer are considered and the "event" is considered to take place if there is an audit in any of the three neighbors.[5] Figure 3 provides the results of estimating change in revenues reported after a fact of audit among 3 nearest neighbors of the taxpayer which turned out to be compliant. Given that we focus in this analysis on compliant neighbors, it is reasonable to assume here as well that they learn about the audit only after it is concluded, hence the event is the day of communication of audit end. We don't observe any material change in the behavior of the taxpayer in response of concluding audits among the 3 closest neighbors

Figure 11. The effect of ineffective audits among 3 closest neighbors on the taxpayer's reported revenues.

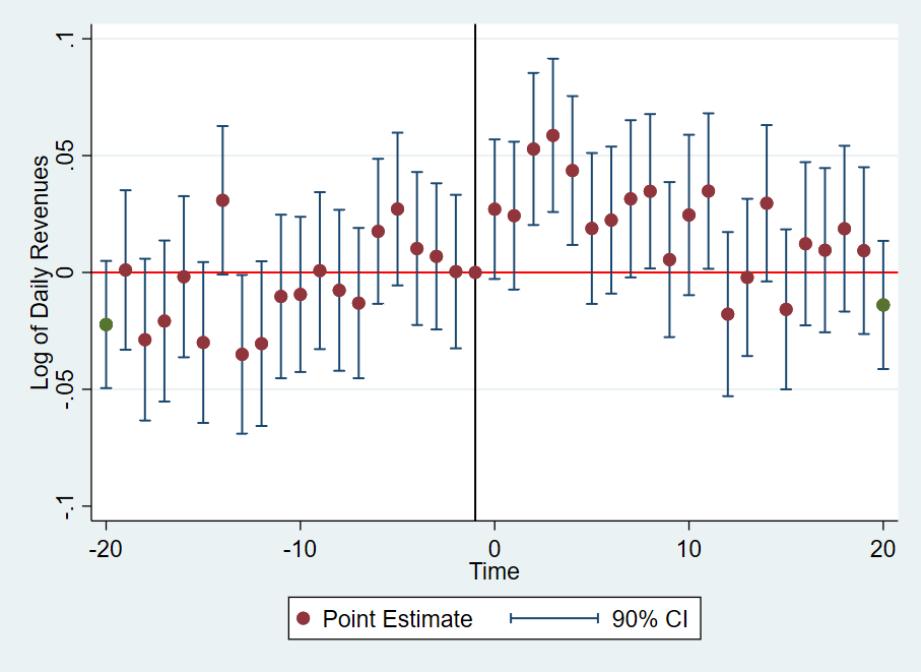


Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

[5] Potentially it can be tested whether increasing (or decreasing) the neighbors analyzed improve the results. The research team will conduct that robustness tests when developing the academic research paper based on this work and share all the finding with the SRC.

Whenever the neighbor's audit actually results in fines payable, the impact on the taxpayer is qualitatively different. Though it is not clear-cut, but there is a slight tendency of reporting higher daily revenues within the first week after the effective audit took place among the neighbors. The effect is not large, but still indicates a behavorial response.

Figure 12. The effective of effective audit among 3 closest neighbors on the taxpayer's reported revenues.



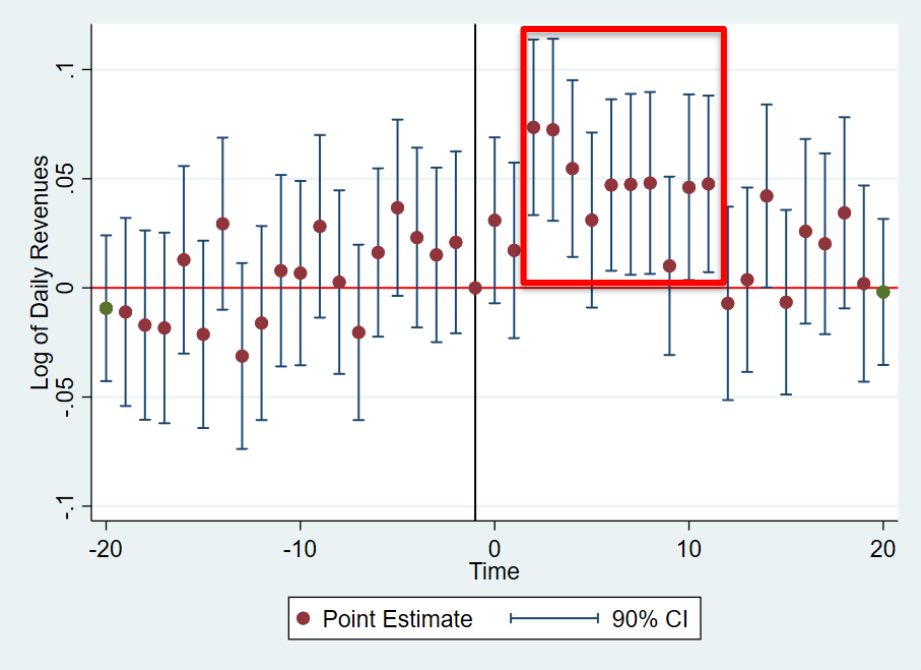
Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

The next question we ask: does the fraud among neighbors affect differently those who were audited and not audited at all during the year? Here we assume that the SRC approach to auditing is non-random, hence those audited should be potentially more prone to tax fraud. At the same time, those who are being audited during the year correct their behavior due to own audits, hence the effect might not be visible for them.

Figure 13 presents the effect of neighbor's audit on those taxpayers which were not audited during 2019, hence are not considered as potentially problematic by SRC. As it can be inferred from the results, the correction of revenues in terms of reporting highly daily turnover is observed exactly among those who were never audited during the year. This has huge policy implication as it

underlines the fact of pervasive under-reporting of daily revenues among those who were otherwise considered to be compliant.

Figure 13. The effective of effective audit among 3 closest neighbors on those taxpayer's reported revenues, who were not audited during 2019 at all



Note: Event study regression with individual level fixed effects, time trend and control for weekdays and holidays, standard errors clustered at taxpayer level.

Box 1. Estimating tax revenues forgone using the spatial analysis of tax receipt audits.

Apart from estimating behavioral responses that can be used for introducing geographic and information component in cash receipts audit planning, this analysis can be also used to estimate the lower bound of potential tax revenues forgone due to underreporting of revenues. In particular, the results can be used to derive these estimates for three different groups:

- Audited and non-compliant
- Audited and compliant
- Not audited at all

The calculations here are approximate and are done for demonstration purposes.

According to data we have these three mutually exclusive groups are composed of 1,765 audited and non-compliant, 909 audited and compliant and 9,664 non audited taxpayers. Considering that the audit is spatially random, we can consider the effect on the non-audited neighbors to be representative of all non-audited taxpayers of our sample. The general formula for the calculation is the following:

$$ATRF = N \times Treatmenteffect \times Ave.DailyRev \times Ave.Taxrate \times Ave.operationdays$$

ATRF stands for Annual Tax Revenue foregone, N is the size of the respective group we are considering, Treatment effect is the increase in percentage terms of the revenues reported for the short period of time (serving as indicator of potential increase). The rest is self-explanatory. We use 4% as average turnover tax for all the groups analyzed here:

ATRF (Audited and non-compliant) = 1,765 * 20%*76,000*4%*300 = 321,936,000

ATRF (Audited and compliant) = 909 * 10% * 76,000 *4%*300 = 82,900,800

ATRF (Non-audited) = 9,664 * 5% * 76,000 * 4% *300 = 440,678,400

Overall ARTF = 845,515,200 AMD in 2019

Part III. Auxiliary tasks

Fraud detection model

The short training was implemented for Pek staff in order to get them more familiar with the technical details and implementation of the fraud algorithm. Meanwhile, all of the code and script were organized and re-created for easier replication and usage.

Capacity Building

This project envisaged an explicit capacity building element. Two SRC employees are currently taking Business Analytics course which introduces them to main modeling approaches for building predictive models applied to business issues. One of the employees also took another introductory course on Python and Statistics during Summer 2021 semester. These courses are part of MS in Management program at the American University of Armenia.

Technical results

In line with our agreement with SRC, all the python command files and respective data files are stored on the hardware physically installed on the SRC premises. ID masking of the taxpayers is maintained at all stages which insures compliance with tax secrecy law.

References

- Antinyan, A., & Asatryan, Z. (2019). Nudging for tax compliance: A meta-analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3500744
- Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in Fraud Detection. *Applied Artificial Intelligence*, 1–23. https://doi.org/10.1080/08839514.2021.2012002
- Biases. Science, 185(4157), 1124–1131. http://www.jstor.org/stable/1738360
- Boning, W. C., Guyton, J., Hodge, R., & Slemrod, J. (2020). Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms. *Journal of Public Economics*, *190*, 104261. https://doi.org/10.1016/j.jpubeco.2020.104261
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). Yake! collection-independent automatic keyword extractor. *Lecture Notes in Computer Science*, 806–810. https://doi.org/10.1007/978-3-319-76941-7_80 https://www.semanticscholar.org/paper/YAKE!-Keyword-extraction-from-single-documents-Campos-Mangaravite/9cb32bdd43f64b36cb447ba1307869c5d8bf675c
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Proceedings of the 2019 Conference of the North*. https://doi.org/10.18653/v1/n19-1423
- Drago, F., Mengel, F., & Traxler, C. (2020). Compliance behavior in networks: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *12*(2), 96–133. https://doi.org/10.1257/app.20170690
- Fuest, C., and N. Riedel, (2009), Tax evasion, tax avoidance and tax expenditures in developing countries: A review of the literature, *Oxford University Centre for Business Taxation*
- Hashimzade, N., Myles, G. D., & Tran-Nam, B. (2012). Applications of behavioural economics to tax evasion. *Journal of Economic Surveys*. https://doi.org/10.1111/j.1467-6419.2012.00733.x

- Hashimzade, N., Myles, G. D., & Tran-Nam, B. (2012). Applications of behavioural economics to tax evasion. *Journal of Economic Surveys*. https://doi.org/10.1111/j.1467-6419.2012.00733.x

- Jahanshahi,H., Ozyegen, O., Cevik, M., Bulut,B., Yigit,D.,Gonen,F., Başar A. (2021). Text Classification for Predicting Multi-level Product Categories .https://arxiv.org/abs/2109.01084v1
- Lediga, C., Riedel, N., & Strohmaier, K. (2020). Tax enforcement spillovers – evidence from South Africa. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3616817
- Lucas, A., Schaumburg, J., & Schwaab, B. (2019). Dynamic clustering of multivariate panel data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3531721
- Maciejovsky, B., Kirchler, E., & Schwarzenberger, H. (2007). Misperception of chance and loss repair: On the dynamics of tax compliance. *Journal of Economic Psychology*, *28*(6), 678–691. https://doi.org/10.1016/j.joep.2007.02.002
- Mittone, L., Panebianco, F., & Santoro, A. (2017). The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, *61*, 225–243. https://doi.org/10.1016/j.joep.2017.04.007
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and
- Zahera, H., Mohamed, Sh. (2020). ProBERT: Product Data Classification with Fine-tuning BERT Model https://www.researchgate.net/publication/344901824_ProBERT_Product_Data_Classification_with_Fine-tuning_BERT_Model