Final report PAK-22180

September 2023

Organisational effectiveness and tax compliance in Punjab

Sher Afghan Asad
Michael C. Best
Adnan Q. Khan
Anders D. Jensen

DIRECTED BY

FUNDED BY

Final Report

Project Code: PK 21180

Project Title: Organisational Effectiveness and Tax Compliance in Punjab

Principal Investigators: Dr. Sher Afghan Asad, Dr. Michael C. Best, Dr. Adnan Q. Khan and Dr. Anders D. Jensen

Summary and Research Overview

Our research aims to generate unique insights on improving tax compliance by collaborating with the Punjab Revenue Authority (PRA), the agency responsible for collecting the Sales Tax on Services in Punjab, Pakistan.

This research will serve as an input to a full-scale randomised evaluation to understand how the tax and economic outcomes of competing firms are affected by tax leakages. Both theoretical (Acemoglu & Jackson 2017, Besley 2019) and empirical (Del Carpio 2014, Besley et al, forthcoming) literature suggests that taxpayers' compliance decisions are affected by how their peers behave. Indeed, compliant taxpayers often complain about the unfairness of having to compete against non-compliant taxpayers. These compliance decisions can have implications on the nature of enforcement strategies employed by the government.

We will experimentally vary the extent of tax compliance of each restaurant's competitors. There are two potential ways to manipulate the competition of a typical restaurant.

1. By bringing competing unregistered restaurants into the tax net.
2. By improving reporting behavior of the competing registered restaurants.

We aim to identify the competing unregistered restaurants (i.e., the restaurants that are not registered as a taxpayer with PRA although above the eligibility criteria of the annual turnover of 6 million PKR) through various third-party data sources. For this, we scrape data on restaurants from sources such as Facebook and Google Maps and local food delivery platforms such as FoodPanda and Cheetah. The scraped data consists of information on likes, reviews, ratings, location, menu prices, etc.
This data will be supplemented and verified by our baseline survey, which will capture important characteristics like the size, the number of tables,and daily foot traffic of the restaurants. This will help us maintain a directory of all the restaurants in the city and would help in our efforts to identify eligible but unregistered taxpayers. Combining these data sources should give enough markers to estimate sales and service quality. We will then create a statistical model to forecast sales of restaurants that are unregistered by the tax authority. The restaurants that are predicted to have a turnover above the eligibility threshold can then be targeted for registration using strict enforcement measures as part of our randomized evaluation. The enforcement shall be carried out through a dashboard to ensure unbroken records of all enforcement activities in both the treatment and control clusters of restaurants.

To improve the reporting behavior of the competing registered restaurants, we rely on increasing the adoption of the electronic invoice monitoring system. Baseline analysis using data on e-IMS adopters has suggested that e-IMS increases reported sales by 40 percent, tax liability by 25 percent, and value-added per unit of output by about 0.4 percent. This evidence motivates the rollout of e-IMS to non-adopter restaurants. The data used for this purpose consisted of anonymized Sales Tax Returns provided by the Punjab Revenue Authority from 2012-2021. This consisted of about 2,500 total restaurants, out of which 89 unique restaurants had adopted e- IMS.

We identify clusters of competing restaurants based on their geographic proximity. We have done one potential mapping of restaurants into 100 clusters, using k-means clustering, in the city of Lahore. This is based on partially geo-coded data of registered and unregistered restaurants. The precise number of clusters and geographical coverage will be determined based on power considerations and once all restaurants are geo-tagged.

Once the mapping of restaurants into clusters is complete, we will pursue a saturation design (Baird et al. 2018) in which we randomly assign clusters to a saturation rate of enforcement activities in the first stage of randomization and then randomly assign restaurants within clusters to enforcement activities at their assigned saturation rates. To monitor the adherence to the research design, we are building a dashboard (with support from IGC), in collaboration with PRA, that would help maintain and track the enforcement efforts of the tax officers. This dashboard will ensure that spillover effects are minimized between the treatment and control clusters, as tax collectors will have automated instructions and information on enforcement measures for different clusters.

Our primary data source is the monthly sales tax returns filed by each firm. This data includes reported sales, input costs, and tax liability. In addition, we aim to survey all the restaurants to establish a viable baseline for our sample. This survey shall first consist of a baseline survey of registered and unregistered restaurants in Lahore, capturing information factors such as service quality, foot traffic, menu prices, etc. Once our evaluation is complete, we shall conduct two end-line surveys to measure the outcomes of our treatment. For the unregistered restaurants, the survey shall measure the same characteristics as the baseline, allowing us to measure the effects of enforcement. The registered restaurants shall be subjected to a more detailed survey to capture the firm's proclivity for tax evasion, along with service quality measures. This includes a mystery shopping survey as in Eissa et al. (2014). We will also build a scraper that collects regular information from FoodPanda and Cheetah (local food delivery service providers) on the menu prices of each restaurant to assess the effect of tax competition on prices and sales.

Web Scraping

As discussed above, an integral portion of our longer-term research experiment was to identify additional service providers in the restaurant industry that were operating outside the tax net of PRA. To randomize enforcement treatment, we had to utilize third-party data sources to compile a universe of restaurants in District Lahore. This exercise was important in the context of sparse enforcement resources and technical restraints faced by the tax authority. Under the circumstances, identifying unregistered taxpayers in the field is an extremely costly and time-consuming ordeal. Previously, this process was somewhat manual and labor intensive. It consisted of field visits to multiple regions of Lahore and browsing of internet sources to discover new restaurants in the market. In our discussions with PRA, it was decided that using free resources on the web could provide actionable intel to the tax officers to begin the process of registration. This process would also set a precedent for the future as this model is replicable to other services that exist online. These services could include beauty parlors, gyms, consultant services, property dealers etc. IGC's support was essential in demonstrating the value of automated data gathering for expanding the tax net and opening the doors for future research collaborations in this field.

The research team had several brainstorming sessions with the tax authority to identify pertinent online sources for information on restaurants. The most promising resources were the following:

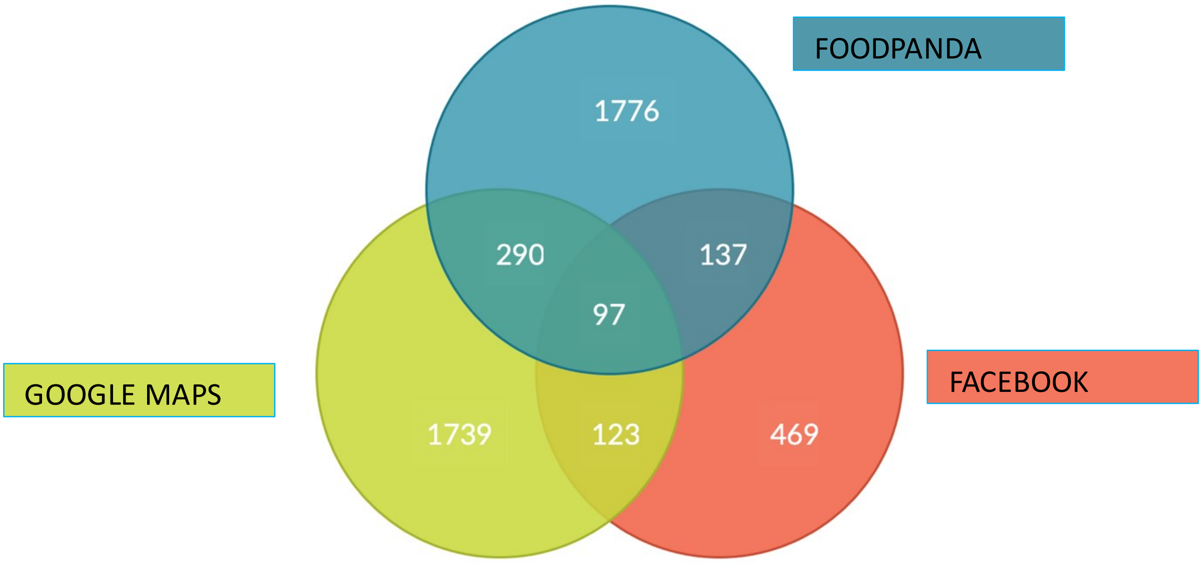
1. Google Maps
2. Facebook
3. FoodPanda

These online sources held a huge repository of information on restaurants within Lahore. We developed a scraper to gather data points on some key variables including the name, address, online website, menu, reviews, likes, followers, and price points. These variables are important for both our research and for PRA. Restaurant characteristics will enable us to cluster restaurants on a variety of factors other than geographical distance. This will be explained further in our section on clustering. Menu prices and reviews will help us to understand how changes in compliance effect the firms in question. These scrapers are being run periodically to continue to update the database on restaurants. This shall enable us to capture new entrants into the market as the turnover rate in the sector can be quite high. The database will form the basis of our sample for a proper observational survey of restaurants across Lahore.

Figure 1 displays the results of our web-scraping activities for restaurants in Lahore. All together, we have managed to capture 4631 distinct restaurants that are not presently registered with the Punjab Revenue Authority. This lends credibility to our method and model for scraping data to identify unregistered service providers in the restaurant sector. From the Venn diagram, we observe that a sizeable portion of restaurants (~76%) were found in either FoodPanda or Google Maps. There were only 97 restaurants found in all sources, highlighting the need to scrape multiple web sources for our dataset.

Another important implication of this finding is that there is a large chunk of restaurants that are currently operating outside of the tax net. PRA has 1074 restaurants registered with them, and an even smaller portion of those are actively paying taxes. These unregistered restaurants are roughly four times as many as the registered restaurants. The results are extremely encouraging for our research design. The large number of potential taxpayers might also prove to be beneficial for PRA to expand the tax base and revenue collections in the sector.

Figure 1: Web-Scraping Results



We also examine table 1 to see how the presence of the registered restaurants is distributed amongst our online sources.

Source	Frequency
PRA	398
PRA & Facebook	55
PRA & GoogleMaps	149
PRA & FoodPanda	133
PRA & All three Sources	78
PRA & More than one source	261

Table 1

Table 1 and Figure 2 show us that 63% of all restaurants registered with PRA have an online presence on at least one platform. There were only 398 restaurants that we were unable to find online. Similar to Figure 1, the pie chart also shows that Foodpanda and Google Maps had a significant proportion of registered restaurants results were derived through exhaustive matching of restaurants between sources to avoid any duplication.

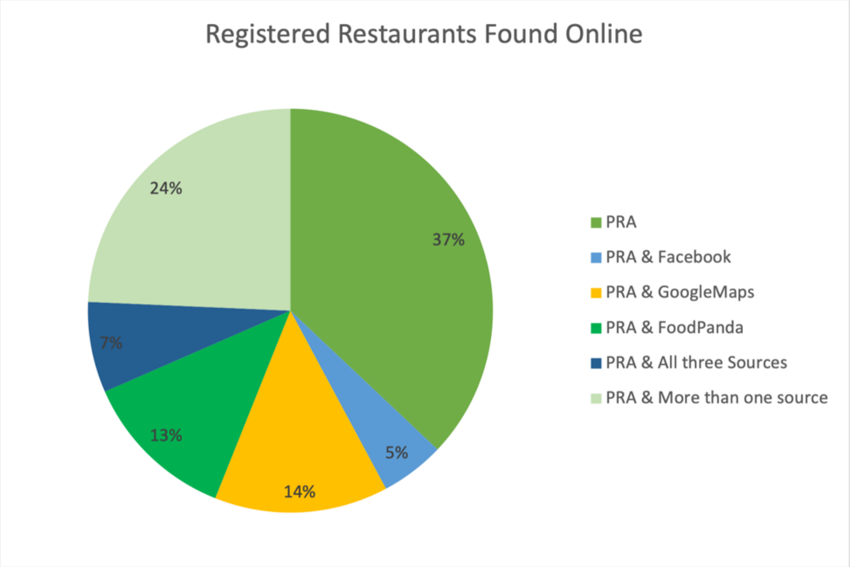


Figure 2: Distribution of Registered Restaurants found in online sources

Geo-Coding of Service Providers

For our research, having exact geo-coordinates for all restaurants in Lahore is very crucial since we plan to conduct an experiment based on varying the saturation rate of restaurant clusters to create our treatment and control groups. The geo-coordinates are also useful to our policy counterpart in assigning the taxpayer to a geographical unit assigned to a particular tax officer. We hired a third-party firm to add latitude and longitude for registered and unregistered restaurants found in the previous section. These coordinates are based on the addresses scraped from our three sources and include business addresses provided by PRA.

Figure 3 shows all the restaurants that are registered with the Punjab revenue Authority, falling in the jurisdiction of Lahore. These restaurants are super-imposed on another map layer that represents the administrative boundaries of towns used by the Local Government. These towns are divided amongst tax collectors to distribute the restaurant service providers. The map shows us that restaurants are generally clustered together around main commercial markets and roads, with the density being particularly high in the center of Lahore and gradually dipping towards the outskirts of the district. The distribution is partially due to the population density but may also indicate an enforcement bias towards central or more ‘posh’ areas in Lahore. We can understand this more by plotting the unregistered restaurants we scraped online.

Figure 4: A Map of all registered restaurants-overlay on administrative boundaries of Lahore

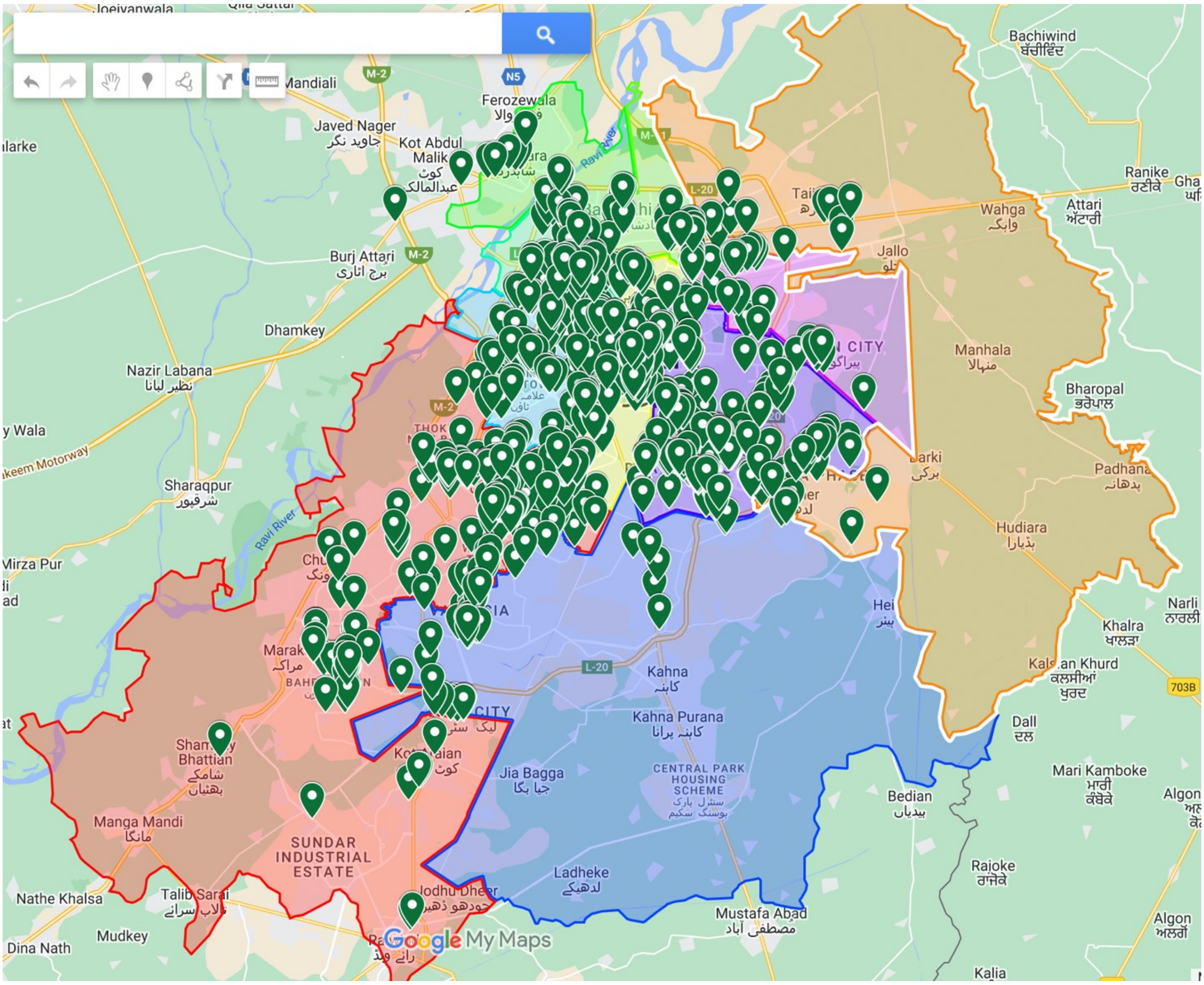
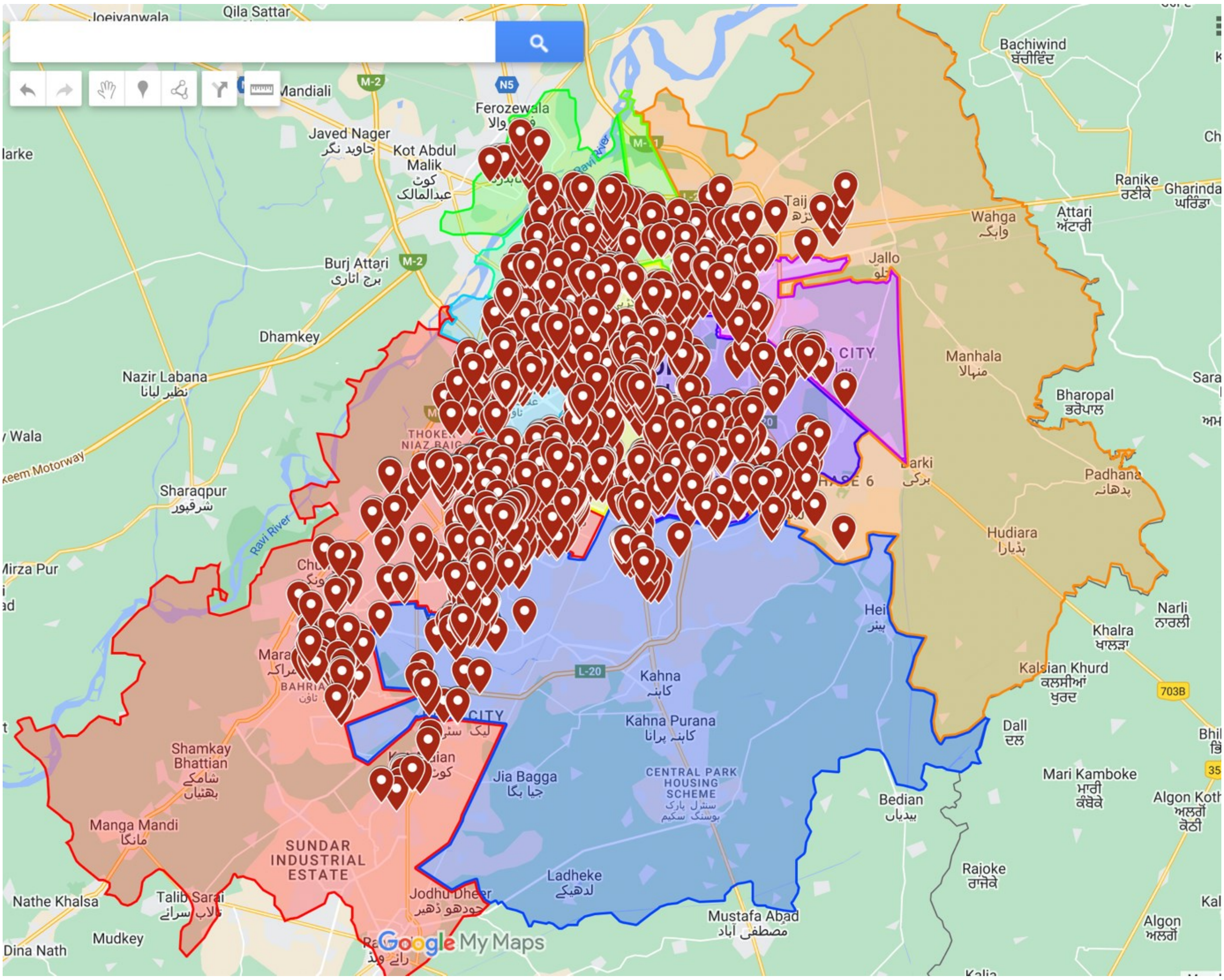


Figure 4 shows the unregistered restaurants in red across the same map layers. Comparing Figure 3 and 4, we observe that the distribution of these restaurants is similar to the registered ones. The clustering indicates that due to the distribution of economic activity and population we have certain towns with a high density of restaurants. We can also verify it by looking at the original google map view. We see that areas with no restaurants are usually green spaces or industrial areas.

Figure 4: A Map of all unregistered restaurants-overlay on administrative boundaries of Lahore



Looking closely at a few restaurant clusters, we observe that the saturation in terms of registered and unregistered restaurants varies considerably from cluster to cluster. Even in high-income and well-known restaurant markets like MM Alam or DHA we find the existence of unregistered restaurants. This is displayed in figures 5 and 6. However, we also find evidence to back the intuitive hypothesis that there is a higher ratio of registered restaurants in these central areas of activity since these areas have high-end restaurants that meet the criteria for registration.

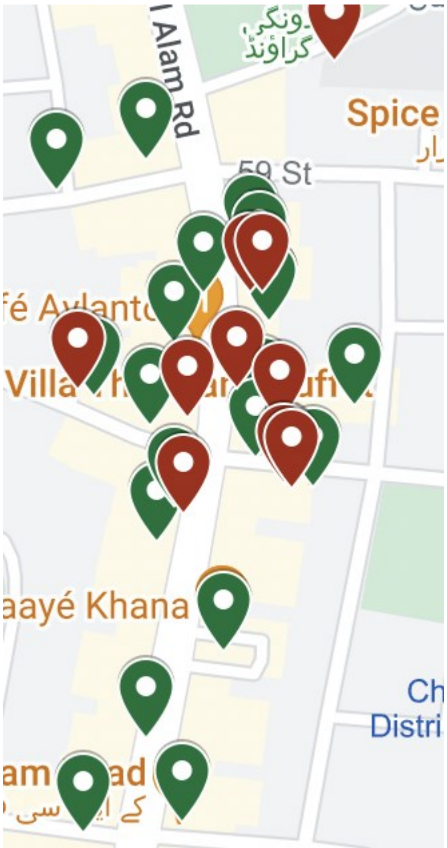
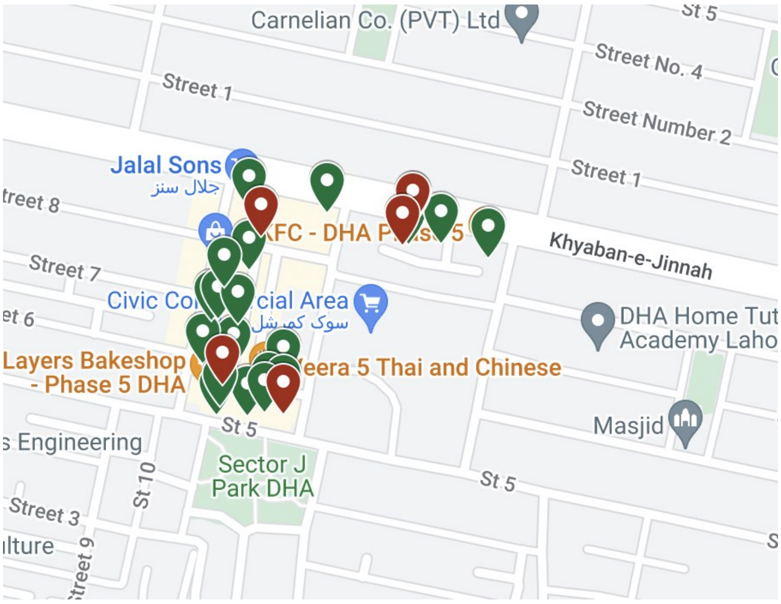


Figure 5: Close-up of MM Alam Road (a cluster with high end formal restaurants)

Figure 6: Close-up of DHA Phase 5 (a cluster with high end formal restaurants)



These findings are also interesting from a policy standpoint. Initially it was discussed that some clusters might not have any new potential taxpayers as there was more scrutiny and enforcement efforts. However, our scraping activity shows that even within the areas generating the most activity and having higher end restaurants, there are new potential taxpayers that are operating outside the ambit of the tax authority. As the experiment progresses, we shall have greater clarity on the characteristics of these restaurants and their tax liability.

We examine two more close-ups of clusters in Lahore to validate our research design. In Figure 7, we look at the areas of Wapda Town and Gulshan. These clusters show us that almost all restaurants are unregistered in certain areas but have a presence on at least one online platform.

Figure 8 is a snapshot of DHA Phase 4 that displays a roughly even split between registered and unregistered restaurants. Close inspection of different regions tells us that each neighborhood or cluster has different levels of compliant restaurants. The variance in the clusters will serve the research design of our experiment and we shall also get a better understanding of spillover effects in different types of clusters. The effect of competition in clusters with a low compliance rate versus those with a higher rate could have useful policy implications in terms of enforcement strategies employed by F

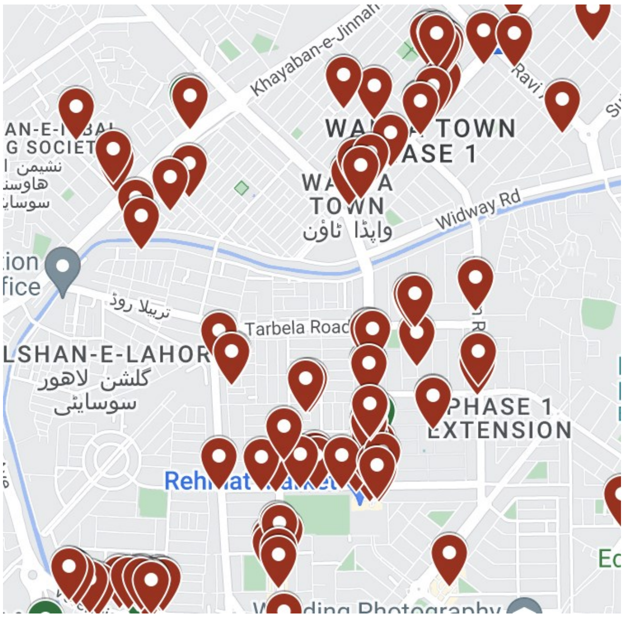
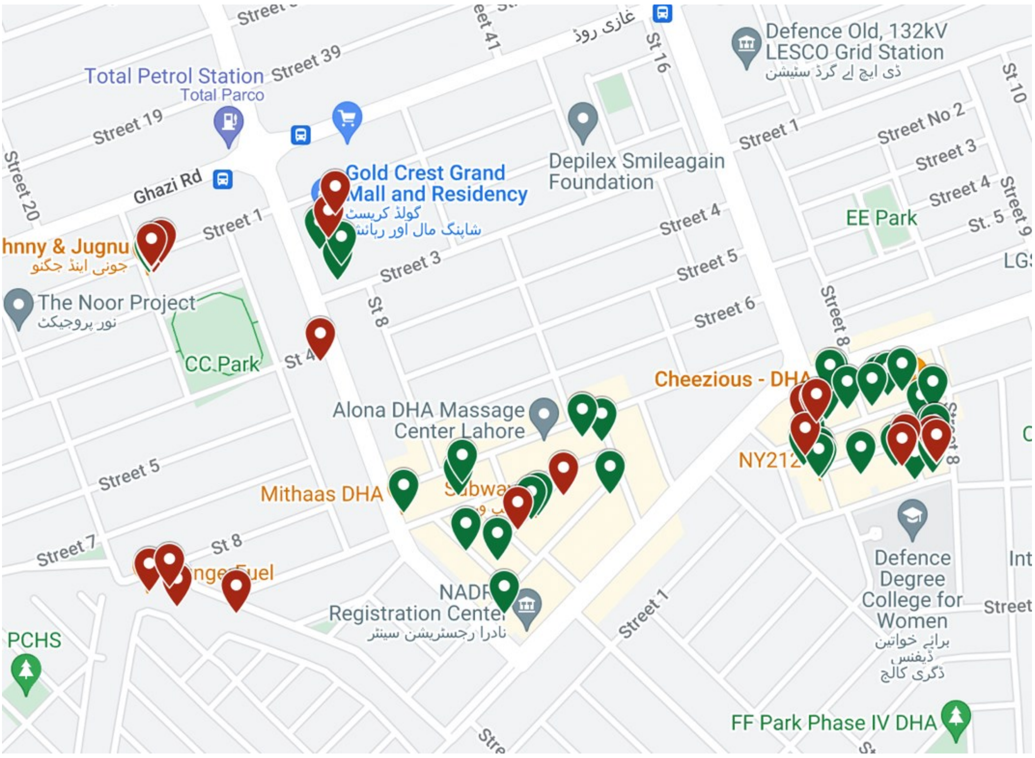


Figure 7: Restaurant cluster with mostly unregistered restaurants.

Figure 8: DHA PHASE 4 Restaurants



Baseline Survey (Census)

This project grant was also instrumental in helping the research team develop and pilot a baseline survey for all restaurants in Lahore. The purpose of the survey is to collect additional data points on the restaurants in our data base. Firstly, it will help us gather firm characteristics that we were not able to identify through web-scraping. Secondly, we want to gather as much information as possible so that PRA may initiate the registration process, which also serves as the treatment for our experiment. The survey will verify the existence of the restaurants and look for additional ones that we have missed. The goal is to create a census as accurately as possible.

The survey shall capture the following characteristics

- Type of Restaurant: This question will collect information on the classification of the restaurant i.e. fine dining, café, coffee shop etc.
- Cuisine: This question shall assign a cuisine to each restaurant. The cuisines will help us determine competitive clusters as an additional dimension to geographical proximity. For each cuisine we shall be gathering information on the price and quantity of dishes to check for spillover effects after our treatment. We are capturing
- Menu: Images will be collected for menus where possible to record the prices.
- Address: The address will be collected by the survey teams and precise geo-coordinates as well. This will help us verify and update our initial dataset.
- Footfall and Restaurant Capacity: The enumerators will capture the footfall at the time of survey and also ascertain the capacity of the restaurant based on the number of floors, chairs, and tables. These are all good indicators that can be used to group restaurants by size.
- Buying Coke/Pepsi: As part of the survey, enumerators are required to purchase either Coke or Pepsi. The price and quantity will be recorded, and we shall measure any changes in the endline survey. We will also be able to collect receipts based on the purchase which shall tell us exactly how much tax is being applied, if any. This information will also be beneficial for PRA to check if the registered restaurants are charging taxes and if there are unregistered ones collecting tax but not declaring so.

To maximise the effort and cost involved in the survey, we have also developed a protocol to search for restaurants near our sample. This is to ensure that we collect a comprehensive dataset. In our testing and pre-pilots, we found some restaurants that were not covered in scraping activities. To include these, we have established a search area of 200 meters centered around each point in our sample. Nearby restaurants will be classified as either formal, informal, or street vendors. Each classification has a different set of questions in the survey.

Figure 9 displays the survey map we have generated and the nearby search radius for additional restaurants. In a lot of the cases, because the restaurants are clustered close together, the search areas overlap. It was important to narrow down the search area due to resource constraints.

Figure 9: Survey Sample with 200m Nearby Area

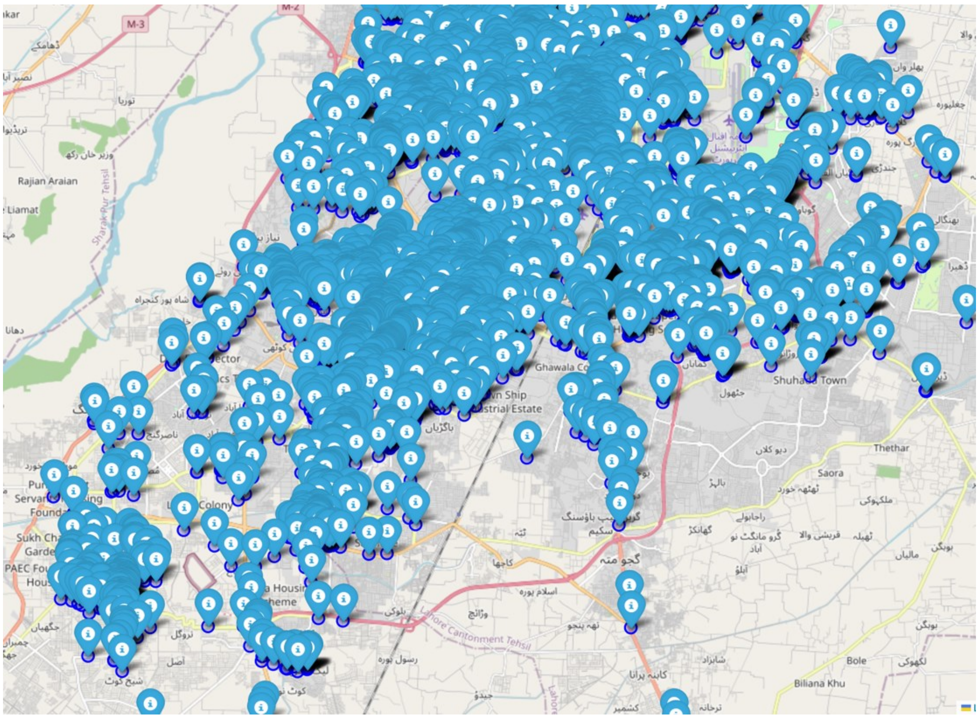
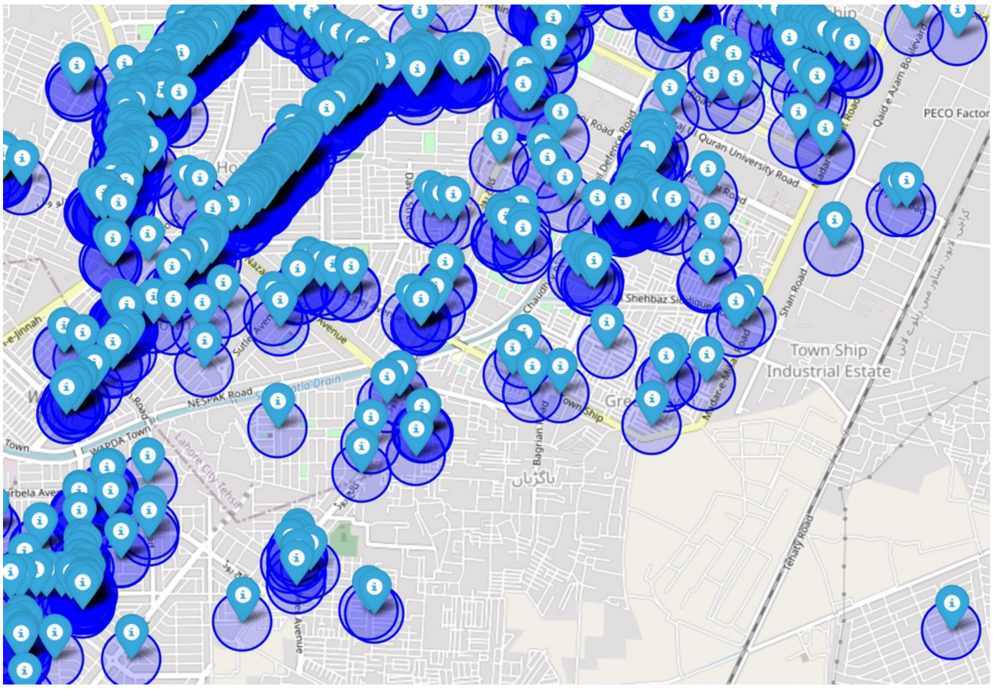


Figure 10: Close up view of the search area



Clustering

The next step in our experiment design is to create clusters based on multiple dimensions. These clusters will be assigned to either the treatment or the control groups in the beginning of the randomized trial. To create the final clusters, we require our survey to be completed. At a preliminary stage we have used a few methods to create clusters based on the dimensions we currently possess. Since we want to classify clusters on multiple dimensions we used k-means clustering approach with principle component analysis (PCA). It aims to partition the observations into a predefined number of clusters (k) in which each point belongs to the cluster with the nearest mean. It selects the centers of the initial clusters from the first observations in the data set and then assigns the other observations to the nearest cluster. When an observation is added to the cluster, k-means recalculates the mean of the cluster variables, and this mean becomes the new cluster-center. If this recalculated cluster-center changes another cluster that is closest to an observation already in the cluster, then k-means moves that observation to the closest cluster and recalculates the center of its new cluster. The process continues until the number of changes is very small.

In Figure 11, we see the result of such an analysis where k is 100 and we are also using Annual Turnover as another dimension.

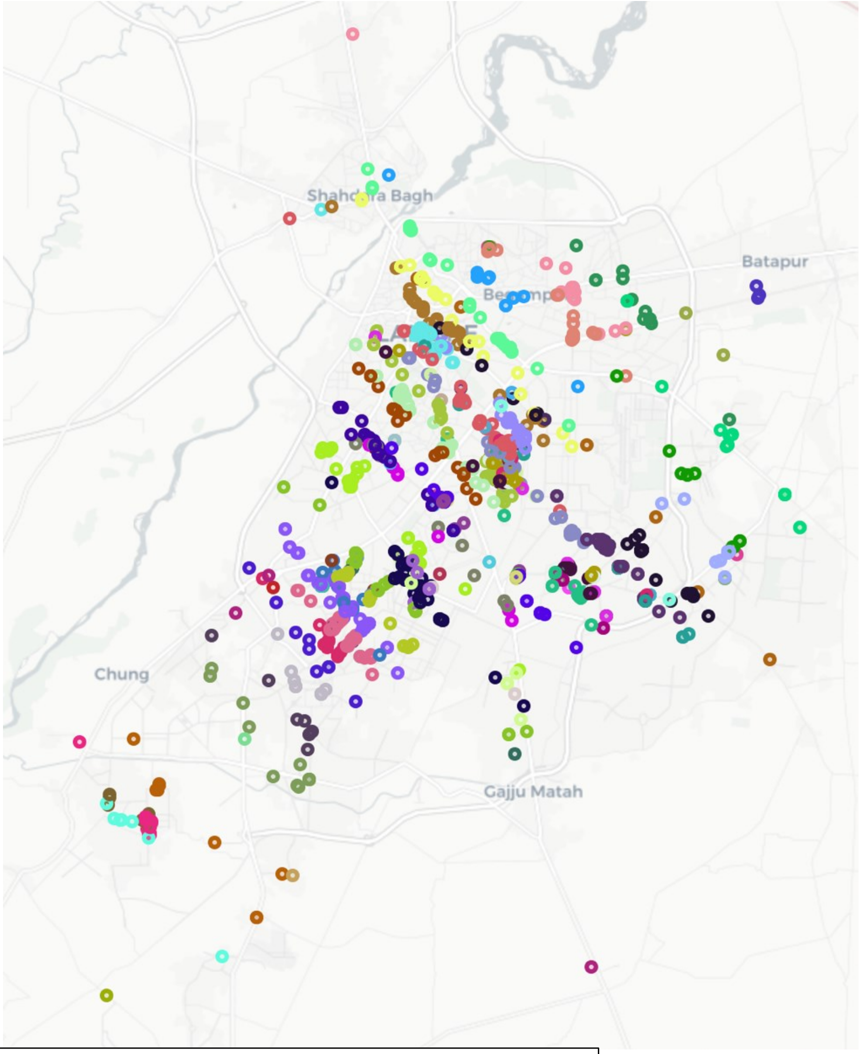


Figure 11: 100 Clusters with PCA

Dashboard Development

Purpose: An **automated web-based dashboard application** for a tax authority to streamline regular workflows of maintaining a pre-filled database, carrying out and keeping track of enforcement tasks, and recording perfomance metrics for supervisors. This dashboard shall also keep a record of all enforcement efforts and communication with the taxpayer for the purpose of our experiment. This will minimize spill-overs and loss of information.

Context: The tax authority has two work streams that require automation.

1. The first one is a list of taxpayers that are mandated to sign up for a software integration with the tax authority. These taxpayers are sent out various notices and penalties to ensure compliance with the policy. These notices are currently typed out and maintained manually (i.e. hard copies). There is structured template for sending out these notices and a chronological flow between different notices (i.e. there is a first notice sent out and bearing non compliance from the taxpayer, a second one is sent out and so on).
2. The second stream consists of registration efforts conducted by the tax authority to include businesses outside the tax net. The tax authority is planning to conduct registration efforts on around 5000 businesses. This registration process will entail a series of notices as per their legal procedure. The registration process follows a particulart order and flow, also having timelines before the next step can be taken. In the status quo all records and tracking is done through paper files and scattered excel sheets. There is a lot of data on these businesses (names, addresses, geo-coordinates, size, estimated sales etc.) that the tax authority needs to decide and carry out enforcement. Furthermore, each tax official has different businesses falling under their jurisdiction, requiring the dashboard to provide restricted access.

Modules in Development

- The Dashboard shall provide limited access to tax officials based on their designation and geographical jurisdiction. At the enforcement level, each tax officer will have access to only the relevant restaurant information.
- Supervisors and senior level Commissioner will have access to the performance metrics of all tax officers. They shall be able to look at KPIs (no. of notices, registrations etc.) and advise the tax officers accordingly. The Dashboard will ensure that taxpayer interactions are recorded and transparent.
- The Dashboard will maintain an online database for all unregistered restaurants in Lahore. The database will contain information from our survey and web-scraping. This information will be used to target the restaurants for registration. PRA requires names, addresses and images to send out notices to taxpayers. The dashboard will maintain a record of all these. The database will be continuously updated against the registrations data from PRA. Various filters based on characteristics will be available.
- The Dashboard will provide pre-filled templates and enforcement actions based on the registrations flow outlined in the Sales Tax Act. The tax officer will check a few details and edit them, if necessary, before sending out the notice.

- Each case of enforcement will have options to upload and record interactions with the potential taxpayer. This shall take into account different forms of communication such as a letter, phone call or email.
- The notices sent out via the dashboard shall record due dates for further actions and flag them to the relevant officials. This will help them keep a track of deadlines and automated reminders will be sent to the officers.
- An interactive map of restaurants in Lahore with filters based on the registration status and business characteristics.

Through IGC's support, we are well into the development of the dashboard. Developers were hired and are working with PRA to ensure an effective solution. The dashboard will be finalized once our baseline survey is completed. The dashboard is being developed to allow compatibility with PRA's own software if required. The dashboard can also be scaled to other services if required in the future.

IGC

www.theigc.org