WORLD BANK GROUP

March 2019 · Number 7

Macroeconomics, Trade & Investment
MTI Practice Notes

Working with Administrative Tax Data: A How-to-Get-Started Guide

Anne Brockmeyer

Introduction

Administrative tax data at the taxpayer or transaction level (henceforth tax data for simplicity) are a great resource for studying taxpayer behavior, assessing responses to changes in tax policy and administration, and deriving lessons for optimal tax policy design (see Pomeranz & Vila-Belda (forthcoming), and Slemrod 2017 for reviews of the emerging literature). From an operational perspective, tax data can be used to prepare technical assistance and investment projects (e.g. identify the most important tax offices from a revenue perspective and assess compliance gaps), monitor implementation of projects (e.g. record the number of declarations filed electronically in response to IT investments), and evaluate project-supported policy reforms (e.g. estimate the increase in tax revenue achieved through a tax rate change). Tax data can also be used to monitor/evaluate non-tax development projects (e.g. the effect of micro loans on business growth) and study a multitude of other questions, e.g. related to intergenerational mobility, firm production networks, or who becomes an inventor.

Governments are increasingly open to providing access to tax data and to collaborating in the analysis of these data. This goes hand in hand with an increasing desire by development organizations and researchers to make use of such data for policy analysis. This note has two objectives. First, it provides a primer on what tax data is, describing the different types of tax data, modes of accessing tax data and briefly reviewing some key upsides and downsides of working with tax data. Second, the note provides practical advice on how to get started in working with tax data – from building a first contact with the Tax Authority (TA) to pitching a project and formulating a data request. I conclude with a brief discussion on publishing studies using tax data. The note draws on my PhD research and my experience at the World Bank, where I work with tax data from various countries, including on a new pilot project between different World Bank units (the Macro, Trade and Investment Global Practice, the Research Department and the Global Tax Team) which investigates what can be learned from comparing micro tax data across countries.

The note is intended both for staff in development partner organizations (be it multilateral, bilateral or NGOs) working on tax, or desiring to use tax data for their work, as well as for graduate students and junior researchers hoping to conduct policy-relevant research using tax data. For a more academic guide to collaborating with TAs, and insights from a survey with 70 researchers working with administrative tax data, see Pomeranz and Vila-Belda (forthcoming).

1) A Primer on Tax Data

What do we mean by administrative tax data? This section provides an overview of the different types of tax data, the modes of

accessing tax data and some upsides and downsides of tax data compared to other types of data.

Different Types of Tax Data

Tax data are collected by the TA in the process of exercising its functions – collecting government revenue. I focus here on the micro-level (non-aggregated) version of this data and provide below a list of the broad categories of tax data. For an illustration with data from Costa Rica, see the context and data sections in Brockmeyer et al 2019 and Brockmeyer & Hernandez 2018.

- **Tax Register:** The register contains the list of all registered taxpayers at a point in time, with the unique identifier, name, geographic location (region or precise address), and (for firms) sector and legal form. Using registration and deregistration records (which are usually based on a specific form available on the TA's website), it is possible to reconstruct the register for previous points in time. There may be separate registers for firms and individual taxpayers.[1]
- **Taxpayer self-assessment declarations:** These are the declarations that taxpayers submit in regular intervals (e.g. monthly for the value-added tax, annually for the corporate income tax and the personal income tax, sometimes quarterly for simplified tax regimes). The self-assessment declarations contain information about the tax base (income/wealth/consumption/value-added), deductions and exemptions, the tax liability, and tax payment. Note that the tax payment is sometimes recorded on a separate payment form and not on the self-assessment declaration.
- **Informative declarations and withholding declarations:** The TA collects information from third parties (i.e. agents other the taxpayers) about the taxpayers' transactions. This third-party information may be submitted to the TA by public procurement agencies, by financial intermediaries (such as credit/debit card processing companies) that report taxpayers' financial transactions (e.g. retailers' sales through card machines), or in the form of VAT annexes in which firms need to detail their transactions with other firms. Some of these third parties are also designated as withholding agents, remitting a small fraction of the transaction amount they process as advance tax payment for the transaction partner (Velayudhan & Slemrod 2018 show that 85% of liabilities are remitted like that). The information and withholding declaration are often uniquely detailed (e.g. providing firm-firm-day transaction level information), but rarely provide information about the type of goods/services exchanged.
- **Customs data:** The customs authority maintains import and export records. The customs authority is usually an entity distinct from the TA. Although the two should collaborate closely so that the TA can cross-check taxpayers' self-assessment declarations with customs records, this is not always the case. Still, it is worthwhile checking if the TA has customs data, or whether that data can be obtained from the customs authority directly. Similarly, public procurement agencies may have data sharing agreements with the TA.
- **Process and HR data:** The TA maintains records of its internal processes, such as tax audits (e.g. list of taxpayers selected for audit, allocation to tax officers, audit outcomes). Closely related to this is information on the TA staff (spell data on work history), remuneration and bonuses.

All data except the HR data would have unique taxpayer identifiers and period identifiers (which can be month/fiscal year/quarter) that allow merging different datasets. The self-assessment, customs and informative declarations usually have monotonically increasing form numbers, which allow sorting and eliminating multiple filings per taxpayer-period. As the tax identifiers might be unique

[1] The property cadaster is a special type of tax register, usually maintained not by the central government but by local governments.

to the TA, it is not always possible to merge tax data with data from other government agencies (unless one of the two agencies can translate names and/or addresses). If a merge at the micro-level is not possible, a second-best approach is to merge semi-aggregated data at the sector and/or region level. But generally, the more disaggregated the data, the better.

Modes of Accessing Tax Data

Different countries provide access to their tax data in different ways. From least to most restrictive, these are the options I have encountered:

1. The data is available online (believe it or not, this actually happens in some Scandinavian countries; Mexico publishes anonymized data > click on SAT mas abierto > Datos Anonimizados > Declaraciones anuales de personas morales).
2. The TA extracts and hands over the data to specific individuals/institutions under a Memorandum of Understanding (this is how researchers work with data from Senegal and Pakistan).
3. The TA extracts de-identified data for specific institutions under a Memorandum of Understanding (MoU), requiring that the data be considered confidential, with restricted access in a secure computer outside of the TA premises but regulated by a data security plan which, among other provisions, requires that the computer is not connect to the internet (e.g. some state governments in Brazil have provided data access this way).
4. The TA provides remote access to the data to selected/screened individuals via a secure server (this is theoretically possible, but I have not seen an example).
5. The TA provides access to the data onsite (e.g. at the Datalab at the UK TA (HMRC) and at the Ecuador TA). In this case, external partners can either work onsite (possibly via a research assistant) or work with a TA staff who closely collaborates with the research team and runs do-files/scripts which are partly prepared remotely on simulated data.

the identifier in "its" data into the identifier in the other data, or a merge can be done based on

In any case, the data should be de-identified securely but systematically, so that the panel structure of the data is preserved, and the same de-identification algorithm applies to all datasets, so that different datasets can be merged. An MoU may specify how and for what purpose the data can be used, detail procedures for safe handling of data, and any conditions on results publication (e.g. results must be such that no data point published is based on less than X observations, results need to be discussed with TA prior to publication).

Data can be analyzed in STATA or R, depending on the mode of access and software available at the TA's data lab. Given potential unavailability of STATA at government offices and the free availability of R, junior researchers should probably invest in R.

Upsides and Downsides of Tax Data

It is good to be mindful of a few characteristics of tax data when preparing to work with them.

Upsides

- Tax data contain the universe of the formal sector. Unlike survey data, they do not suffer from selective non-reporting at the top of the income distribution. Unlike census data, they contain very detailed information.
- The data is collected at high frequency and low marginal cost.
- Most types of tax data are now collected electronically, which minimizes errors in tax filing (e.g. through internal consistency checks in tax filing software) and data processing.
- As the data is the product of actual economic processes, it measures variables with high precision, unlike survey data in which respondents provide ballpark figures as their response has no meaningful consequences for themselves.

Downsides

- The fact that the data is directly economically relevant for those people

who provide the data (mostly taxpayers or their transaction partners) also means that the data is not necessarily good in capturing real economic outcomes. Self-assessment declarations in particular capture reported outcomes. Informative declarations are more likely to capture real outcomes, as the reporting agent has less incentive to misreport and is often more tightly monitored.

- Tax data can be poor on demographic information on individuals and households unless they can be merged with other government data (studies in Denmark and Sweden have exploited the ability to merge across various types of administrative data).
- Various types of documentation are available to understand tax data (tax returns, manuals on how to file tax returns, tax laws and decrees that explain the tax system), but unlike survey/census data, tax data is not collected for primarily analytical purposes, and is thus not accompanied by researcher-friendly variable descriptions and codebooks. Understanding the data requires knowledge of the relevant language and a regular exchange with the TA on variable definitions, administrative practices and legislation.

2) How to Access Tax Data

Once a policy question that requires access to tax data is identified, some diplomacy and entrepreneurial spirit is usually necessary to make the project a reality. This note is focused on the technical aspects of accessing tax data rather than the conceptual questions studied, but a sample list of tax policy questions and data required to study them is provided in Table 1. Staff in development partner organizations working on tax will have come across many of these questions. Graduate students or junior researchers should replace "policy question" by "research question", which would be derived from an understanding of the existing literature, remaining knowledge gaps and important questions that policy makers in low and middle-income countries face.

Below, I describe how to build a connection with a TA, make a successful project pitch, walk the tightropes of institutional politics, and prepare a data request.

Building a Connection

This section proposes some practical steps for building a connection with a TA and developing a joint project based on tax data. This is primarily intended for junior researchers. [2] Staff in development partner organizations will know which steps they can skip or have already completed.

- Find a context/country that you know well or have some connection to, ideally one that is not yet over-researched, to avoid overlap and potential conflicts of interest with other research teams. Or, work as Research Assistant for a more senior researcher to gain experience working with tax data, and potentially identify a spin-off topic for your own research.
- Identify a contact person or local champion in the TA. Ideally, someone more senior who has a relationship of trust with the TA would introduce you. Hopefully your first contact person will soon loop some of her colleagues into the dialogue.
- Spend time to understand
    a) the country's tax system, its particularities, and policy challenges [read World Bank and IMF reports, the country's tax laws, reports by the TA and Ministry of Finance, press releases and discussions on government websites, and above all, talk to people];
    b) the data structure (here, I mean to try to understand the tax return forms and other data formats –

[2] Also see Glennerster (2014, 2015) on how to create research partnerships with practitioners.

postpone direct discussions about the data for a bit);

c) then link a+b to the broader policy/research questions to study, derive a more specific formulation of the question and hypothesis, and define a methodological approach (which can be quasi-experimental, i.e. exploiting historical data and policy variation, a randomized field experiment, structural model estimation, or a combination of those).

- After many "learning meetings" and some informal discussions about the project, pitch your topic and methodological approach to your government counterpart. This can be in person or via videoconference. Then continue the dialogue by integrating feedback, adjusting your project to fit realities on the ground and policy needs, and, if needed, re-pitch your project (ideally at increasingly higher levels of the administration) until you have agreement on the project. Then start discussing the data request and associated logistics formally (you might have touched on this topic previously in an informal way).
- To make your project a reality, you either need high-level political buy-in (e.g. from the Director of the TA or a Deputy Minister of Finance – they can then request their technical staff to collaborate), or agreement from both the tax intelligence/IT department (which handles the data) and the technical staff who are experts of the topic the project focuses on (unless the latter are hierarchically superior to the data guardians and can request the data). The data vs policy split often corresponds to a TA vs Ministry of Finance split. Generally, it is ideal to have contacts in various parts of the TA and the Ministry of Finance, to ask background questions on legislation, administrative practice, get a second opinion on a topic etc.

In addition to the above, staff in development partner organizations could leverage synergies with technical assistance work, budget support lending (e.g. lending can support an MoU for data access, or a new data confidentiality law that stipulates ways for accessing data for research purposes), or investment projects (e.g. to support the establishment of data warehouse or datalab, or otherwise improve the TA's data infrastructure).

Ingredients for a Successful Project Pitch

Obviously, the key ingredient is a policy-relevant project which allows the government to improve on or learn something they would otherwise not be able to achieve, and which is in line with the TA's mandate and actual policy challenges. Needless to say, the project should also strike the balance between being innovative yet realistically feasible.

In addition, it helps to weave into your pitch (e.g. slides, write-up) some of the following:

- Evidence that other countries or other government institutions in the same country provide access to their data for analytical purposes.
- Examples of other projects using tax data that are policy-relevant and have a positive policy impact. This can be evidence from work by other people (maybe a mix of well-known work by senior researchers and work by people more similar to your background). Ideally choose topics/methods similar to the ones you are pitching.
    - The more specific you can be about the policy impact of the example project(s), the better (e.g. sustained changes in the TA practices informed by the project, or sustained improvement in tax revenue or distributional fairness).
- If applicable, evidence of your own track record in working with TAs/with tax administrative data in other contexts, and of your policy impact.

In terms of capacity building, local ownership and quality of the project (and potentially also ease of accessing data), it can be a good idea to identify not just a local champion but an actual co-author (Juliana Londoño-Vélez and Pierre Bachas have fared very well with this).

Walking the Tightropes of Institutional Politics

This a tricky terrain, and the politics of each country and TA are different, but here are a few thoughts (again, more targeted to graduate student than development bureaucrats, who would have heard such advice before):

- Be mindful of "your" country's development status and geopolitics when providing examples of successful collaborations in your pitch. Most governments like to be compared with "aspirational" peers, i.e. slightly more developed countries.
- Be mindful of internal politics in government agencies. External partners can have convening power (and the required innocence and independence) to bring competing government departments to the same table. Yet they might also inadvertently exacerbate internal challenges, e.g. by designing a field experiment that requires the collaboration of rival departments.
- Be mindful of the career concerns of your government counterparts. Depending on their position, career history and aspirations, some staff in the TA will have stronger incentives to collaborate with external partners and generate innovative findings than other staff. You can try to help ensure that the internal champions in your project get the visibility they deserve for their work, but also be careful about implicating them if part of the project turns out to be controversial.

Formulating a Data Request

Once the government counterparts have agreed to the project, a formal data request letter can be prepared. It would likely contain the following elements:

- Reiterate the (verbally agreed upon) purpose of the data use (e.g. policy question to be studied) and discuss the benefits of the study.
- Detail the data needed:
    - Type of dataset (e.g. annual corporate income tax declarations; if applicable, quarterly income tax declarations; all payment record related to the corporate income tax);
    - Sample (e.g. all corporations and unincorporated firms (i.e. self-employed individuals) in all tax offices);
    - Period covered (e.g. 2010-2018);
    - Identifiers: unique taxpayer ID (de-identified), tax year, declaration submission date, declaration number;
    - Variables: list the line items/boxes on the tax return that are needed (if the tax return isn't too long, it's often easiest to request all variables, which makes the request easier to deal with for the person extracting the data, prevents issues due to errors in variable selection, and limits the need for follow-up requests to add variables);
    - Any additional variables that need to be merged into the data (e.g. sector codes for all firms from the tax register).
- Specify the mode of access (see section above) if agreed upon or propose further discussions about the logistics.
- Whether or not you want to request an explicit ex-ante permission to be able to publish whatever results you find is a sensitive and context-specific question, but it is important to convey (at least implicitly) that you are not intending to prepare a top-secret report but rather a public good.

Once the data has been accessed, it will likely yield interesting results. Until then, it is important to maintain a regular exchange with the TA during the data analysis, communicate intermediate results, seek feedback and consult on the final dissemination strategy and policy discussion surrounding the results. After all, improving policy design – either directly or indirectly, by improving our knowledge of tax systems – should be a key objective of the analysis.

3) Publishing Findings Based on Tax Data

Whether or not governments should be consulted before publication of the results is a question up for debate. For staff in development partner organizations, it will usually be necessary to discuss research findings and dissemination strategies with the government before publishing. Academic researchers, however, might prefer to merely present research findings, as they worry that a consultation (or worse, a legal agreement) which gives government a veto right over the publication can harm the independence and objectivity of the study.

I have not encountered or heard of a situation in which a study was withdrawn, substantially altered or misrepresented due to government intervention. On the contrary, a strong ownership of the study's findings improves the chances that the study attracts the attention of policy makers – in the country in question as well as in peer countries and practitioner fora – and leads to positive change. Besides, requesting data from a government without promising consultations on the publication of results risks thwarting from the onset some studies that would be of high public interest yet potentially controversial.

As for publication in academic journals, work with tax data present the additional challenge that the data is highly confidential and cannot be published, which is contrary to editorial policy in most academic journals. Journals are generally happy to waive the data publication requirement for tax data, but this waiver must be requested at the time of submission to the journal. Researchers must also provide all replication codes and an explanation of how access to the data could be requested by other researchers desiring to replicate the findings in a published paper. There are ongoing discussions about the non-replicability problem for studies using confidential administrative data, so rules might become stricter in the future.

About the author(s):
Anne Brockmeyer, Senior Economist, World Bank's Macroeconomics, Trade, and Investment Global Practice
abrockmeyer@worldbank.org

Table 1: Typical Tax Policy Questions and Data Required to Answer Them

	Policy question	Required data	Other requirements to answer the question
1	Effect of a change in the rate/base of tax X (on reported tax base/payment); optimal rate/base for tax X	Taxpayer declarations for tax X, for some period around the tax policy change	Some variation in the applicability of the policy change (e.g. rate change applied only to certain types of taxpayers) Depending on whether optimal refers to revenue-maximizing or welfare-maximizing, answering this question may require estimates of other parameters, e.g. liquidity constraints
2	Effect of tax enforcement intervention/policy on reported tax base/payment	Administrative data that measures the aspects of compliance most likely affected by enforcement, directly and indirectly – e.g. through spillovers (can be tax declarations, registration forms, third-party reports)	Information on the timing and targeting of tax enforcement (e.g. list of taxpayers audited and audit date); information on how targeting was decided (e.g. random targeting, targeting based on some cutoff rule such as specific risk level)
3	Effect of tax policy or tax administrative practice on real outcomes (e.g. firm growth)	Data under 1. or 2. above + ideally survey data which provides a measure of real outcomes less susceptible to misreporting	Randomized or natural experiment which varies tax policy or tax administrative practice
4	Distributional properties of tax X (income or wealth taxes)	Taxpayer declarations for tax X, or household survey data containing information on tax X	
5	Distributional properties of tax X (consumption tax)	Consumption survey data (cf. Bachas et al 2019)	