Understanding Drivers of Property Tax Delinquency

Laura de Castro Zoratto

SUMMARY

The Brazilian city of Manaus suffers from rates of property tax delinquency so high that the authorities collect only 50 percent of property tax due. To uncover the drivers of delinquency and opportunities for reform, this project analyzed rich, diverse sources of data from Manaus. The team took the novel approach of combining household-level property tax records, administrative records of property values, census tracts data and geospatial data on each taxpayer. Using LASSO regression, researchers identified the most impactful drivers of delinquency probabilities in terms of socioeconomic factors and tax compliance history. They then analyzed geospatial data in relation to these drivers, to investigate inequities generated by the tax system, and how they affect compliance.

Tax liability in Manaus is based on observable proxies for home and land value, and varies across 65 tax sectors, with these borders and sector-specific tax rates set in 1983. This means taxpayers with divergent incomes or property value are often charged the same tax because they live in the same tax sector, while similar taxpayers who live in the same street can face very different tax rates because they arbitrarily fall into different tax sectors. To infer the effect of differential tax
rates in neighboring tax sectors on delinquencies, the team used Regression Discontinuity Design. The analysis showed that taxpayers’ physical proximity to a lower-tax sector was a significant determinant of delinquency rates, suggesting that perceived inequity drives non-compliance. This question can now be investigated further and, with training, members of the city government will be able to repeat the analysis to assess the impact of policies designed to address inequity in property tax rates.

**CHALLENGE**

A clear understanding of the drivers of property tax delinquency is crucial for municipal policymakers to design and implement effective tax policies. However, in many developing countries, the size of the sample space of potential variables, along with data sources that seem too disparate and inadequate to capture drivers, and a lack of technical expertise, have been a stumbling block for local governments’ efforts. For these reasons, big data in the form of historical tax compliance, census and geospatial data have seen limited use among municipal governments for informing taxation decisions.

In Brazil’s Amazonas state, the city of Manaus was able to collect only half of the property tax due each year. When a World Bank DPL required a reduction in tax delinquency as a prior action, achieving this proved more difficult than the authorities had expected. A particular feature of property tax liability in Manaus is that it is based on observable proxies for home and land value and depends on a scaling factor that varies across the city’s 65 tax sectors, with most borders set in 1983. As a result, taxpayers who are very different in terms of factors such as individual income or property value are often charged the same tax because they live within the same tax sector, while similar taxpayers who live across the street from one another can face very different tax rates because they arbitrarily fall into different tax sectors.

In response, the city government provided the project team with a series of datasets comprising de-identified taxpayer-level records on historical tax assessments, payments, attributes and appraisals of recorded property values, geospatial factors and census data. These data sources collectively contributed about 4,500 independent variables for about 700,000 observations. The high dimensionality of the data was compounded by the presence of many nominal features, which necessitated the use of binary variables, further increasing the dimensionality. To add to the complexity, missing values in many rows of the disparate datasets became a significant challenge, comprising about 20 percent of the total when the datasets were merged into a single database for analysis.

The team sought to clean the available data so it could be combined with geospatial data to investigate the relationship between inequities in the tax system and non-compliance. This would help the city government understand drivers of delinquency so it could respond with appropriate reforms.

**INNOVATION**

The first phase of the project involved the novel step of combining disparate datasets into a single database for analysis. These datasets comprised hundreds of thousands of rows of de-identified individual household-level tax
assessments, payments, property-value information and census data. Geospatial data on the approximate distance between taxpayers (at the “neighborhood block” level) and the nearest tax sector borders were then merged into the dataset.

The administrative data on property tax was combined with other administrative records on property sales and census data, producing a rich dataset, but one challenging to work with as it comprised thousands of potential explanatory variables. The database was composed of multiple individual tables, which were merged using one or more features (such as household ID) as the primary and composite keys. When a table with a smaller number of rows is merged into a larger one, the unmatched rows are coded as missing for the features in the larger table. For the large number of missing data scattered throughout the resulting database, the team used the technique of multivariate imputation by chained equations (MICE) for imputation when the missing values comprised less than 5 percent for a given variable – the threshold most commonly used in practice. When a variable contained missing values comprising more than the threshold, it was excluded from the model building.

To reduce the dimensionality of the data to a suitable size for analysis, the team used Least Absolute Shrinkage and Selection Operator (LASSO) regression to select from around 4,500 variables a subset of 10 to 20 most important ones, each of which independently contributed explanatory power. This enabled researchers to identify the most impactful drivers of the aggregate delinquency probabilities in terms of socioeconomic factors and tax compliance history. To avoid overfitting (whereby the analysis would correspond too closely to a particular set of data, failing to fit additional data or predict future observations reliably), the team took precautionary steps. These included five-fold cross validation to train and select the optimal model. In this technique, the data was randomly divided into five samples, with four used to test the fit of the model against the remaining sample, which was used for validation of the projected fit.

To infer the effect of differential tax rates in neighboring tax sectors on delinquencies, the team applied Regression Discontinuity Design to geospatial data, to calculate taxpayers’ physical proximity to their neighboring tax sector. In the first iteration of the analysis, the distance was measured at block level, since granularity at the property level was not available at the time. In the next iteration, the team was able to obtain the household-level geospatial data, which it will incorporate into the analysis. This imagery data was then combined with the historical tax compliance and census data for each taxpayer.

**Results**

In measuring delinquency, the team used measures that referred both to the percentage of tax liabilities not paid, and the percentage of taxpayers who have not paid. For both measures of delinquency, the LASSO models showed past delinquencies and certain property and household characteristics to be important. Of particular interest was taxpayers’ physical proximity to a lower-tax municipality. The RDD analysis showed that at the boundaries of the tax sectors, tax liabilities change discontinuously by an average of 17 percent. This is entirely due to the scaling factor, since appraised house values do not change at the
border. However, tax delinquency increases by an average of 9.5 percent on the high-tax side of the border, with higher levels of delinquency related to larger increases in tax levels across boundaries.

Taxpayers’ physical proximity to a lower-tax sector is therefore a significant determinant of delinquency rates, revealing taxpayers’ perceived inequity as a potential causal effect of non-compliance. This is a question for further investigation. The methodology itself proved a useful tool which, with training, members of the city government will be able to use for ongoing analysis to assess the impact of policies designed to address inequity across property tax boundaries.

In the next iteration of the project, the team will incorporate a more accurate distance measure at the property level, using the more granular geospatial data now available and the data on household-level income. This more detailed analysis will be useful to describe the setting and allow further investigation of the impact of inequities on compliance behavior. The team then aims to design an experiment to tease out different mechanisms through which inequities affect delinquency rates.

Lessons Learned

By being creative and undaunted in combining different sources of data for analysis, the team gained otherwise elusive behavioral insights and opened new avenues for research.

• **Be bold in combining different data sources**
  Combining administrative tax data with census data and exploiting their geographic features enabled the team to shed new light on the determinants of tax non-compliance – despite the dauntingly large and disparate dataset created. This bold approach provided the Manaus government with new intelligence about the city’s taxpayers and suggested further avenues of research.

• **Use data analysis to gain unique insight into human behavior**
  In sensitive areas such as tax delinquency, which perpetrators may be reluctant to discuss in a traditional survey, big data analysis offers insights into people’s behavior and motivators that would otherwise be hard to obtain. This offers policymakers invaluable evidence on which to base reform.