**Informality, Consumption Taxes and Redistribution**

README FOR DATA-SETS AND ANALYIS RELATED TO MEXICO REFORM

**Data availability statement**

Some of the data sets used for this project are confidential and were accessed on-site at INEGI’s microdata lab. Specifically, confidential data from the micro data-lab was used to analyze the Mexican VAT reform (Section 6 of the paper). This README document details the data for that analysis.

INEGI is the National Institute of Statistics and Geography (INEGI) in Mexico. Access to the microdata lab may be obtained upon registering with INEGI. Access to the microdata lab is only granted for public policy and academic research purposes. Researchers interested in access to the data may contact [microdatos@inegi.org.mx](mailto:microdatos@inegi.org.mx) . More information on process to access can also be found here: <https://www.inegi.org.mx/app/microdatos/solicitud/>

The process time that it takes to register with INEGI may vary depending on whether the university that the researcher is affiliated with has a pre-existing agreement with INEGI. If a pre-existing agreement is in place, registration likely takes one to two months. In case of no pre-existing university or institute agreement, the interested user needs to directly sign an agreement with INEGI, and this may take several months to process.

Access to the microdata must be done on-site at INEGI in Mexico City. There is also an option for remote execution of scripts and do files. This service was implemented during the global health pandemic, and it is active at the time of this README (May 2023) though there is no official policy as to whether it will remain active on the future. With the remote service, the researcher sends scripts via email; staff at INEGI execute the scripts. There is no option of remote access to the data.

When the registration is done, but before accessing the data on-site or remotely, it is mandatory for the researcher to undertake an in-person training at the INEGI office in Mexico City. This is a group training with specific dates, depending on the availability of the personnel at INEGI. For more information on the dates of the next training please reach out to: [microdatos@inegi.org.mx](mailto:microdatos@inegi.org.mx)

All replication scripts, including detailed explanations of data construction, etc. are available at the following DOI

**Registers used**

Economic Census, 2019 wave, INEGI, accessed 2020. ([https://www.inegi.org.mx/programas/ce/2019](https://www.inegi.org.mx/programas/ce/2019/#Informacion_general))

Economic Census, 2014 wave, INEGI, accessed 2017.

(<https://www.inegi.org.mx/programas/ce/2014>)

Economic Census, 2009 wave, INEGI, accessed 2017.

(<https://www.inegi.org.mx/programas/ce/2009>)

Microdata on Prices for the Consumer Price Index (CPI), 2012-2015, INEGI, accessed 2017.

<https://www.inegi.org.mx/temas/inpc/>

**Description of registers used**

The Economic Census is the confidential Census of Mexico’s firms done by INEGI every five years. The data sets include information on the universe of firm’s sales, costs, economic activity, taxes, etc. This data includes all firms regardless of whether they are formally registered. The year of the wave corresponds to the year in which the results are presented. For example, the 2019 Economic Census was conducted in 2018.

The Microdata on Prices for the Consumer Price Index is a confidential microdata used by INEGI to construct Mexico’s Consumer Price Index (CPI). The data is obtained from a selected set of goods in different stores from selected cities and localities each week. The observations are gathered at the product per store per week level and can be linked with the Economic Census data by the business identifier.

The CPI data is subject to on-going modifications, as INEGI conducts revisions for data-quality purposes and to comply with international best practices. For more information on these revisions, please refer to the following methodology document (in Spanish): <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825104177>.

For information about the data-layout of the CPI and Census datasets, and changes to data-protocols at the micro data-lab over time, please contact [microdatos@inegi.org.mx](mailto:microdatos@inegi.org.mx). Our research project at the micro data-lab has specifically been handled by Natalia Volkow ([natalia.volkow@inegi.org.mx](mailto:natalia.volkow@inegi.org.mx)).

**Stata dofiles / packages installed and used**

outreg2.do – This package allows us to extract components of regression tables and insert them in .tex format

reghdfe.do – This package allows for more computationally efficient estimation of a large number of fixed effects.

**Outline of folder structure**

Do-files

* The do-file ‘Master\_Mexico’ allows the user to run all do-files required for the full set of results. In addition, this do-file summarizes how the original data are used to generate analysis data-sets as well as how the analysis data-sets are used to generate all Figures and Tables.
* The remaining do-files in the DOFILES folder are the individual do-files that perform some specific tasks in order to create analysis data-sets and produce the Figures and Tables
* In the following section there is a detailed list of the analysis data-sets.

Folders

* The following folders follow the structure of the directories announced at the beginning of all do-files: DOFILES; ORIGINALDATA; CLEANDATA; WASTEDATA; RESULTS.
* Moreover, the folder ‘AUXILIARY’ contains documents that serve as additional information for various parts of the data-creation and analysis. These documents are referred to when applicable in the different do-files
* At the end of the document is a table which maps the Figures and Tables found in the RESULTS folder to their do-files and analysis data-sets.

**List of analysis datasets**

**CPI\_CE2014\_merged.dta**

Population definition:

* Price data for all products sampled by INEGI, Mexico’s National Statistics Office, in the context of creating the national Consumer Price Index (CPI). Further information on sample is in AUXILIARY folder of this replication do-file.

Panel structure:

* Monthly panel data 2012-2015. Contains the following key variables:
  + Store-type (detailed store-type classification which we use to classify stores as ‘modern’ versus ‘traditional’)
  + Geographical location of store (used to define treatment and control groups)
  + Price at product level (our main outcome variable)

Construction:

* To construct this dataset, start with the original CPI data. Prices are reported weekly, create monthly price variables by averaging across weeks. Use product classification to define sample for analysis: taxed products and non-tradeable products. Information on both classifications is contained in documents in the AUXILIARY folder of this replication package.
* In CPI data, define treatment versus control on the basis of border versus non-border location of city where price data is sampled. Border area includes: all cities within 20km of either the Northern border or the Southern border; and, all cities located in the states of Baja California, Baja California Sur, and Quintana Roo (see Figure A7). Save cleaned version of CPI data.
* Merge CPI data with 2014 Census data using the unique identifier CLEE.
* Use store-type information from Census to create store-classification (modern versus traditional). Traditional stores contain the following store-types: public markets; other street vendors; specialized stores; convenience stores; grocery stores. Modern stores contain the following store-types: supermarkets, price clubs, department stores.

**Censos\_2009-2014.dta**

Population definition:

* Retail Census for the years 2009 and 2014, containing information on all retailers operating in the country at the time.

Panel structure:

* Panel at the yearly level, connecting the 2009 Census to the 2014 Census (no data in intermediate years). Contains the following key variables:
  + Store-type (detailed store-type classification which we use to classify stores as ‘modern’ versus ‘traditional’)
  + Geographical location of store (used to define treatment and control groups)
  + Indicator variable for whether a unique store (identified by its identifier CLEE) is present in a particular Census

Construction:

* To construct this dataset, start with the 2009 Census. Since variables are pre-harmonized across Censuses, directly append with the 20014 Census.
* Define treatment versus control on the basis of border versus non-border location of municipality where price data is sampled. Border area includes: all municipalities within 20km of either the Northern border or the Southern border; and, all municipalities located in the states of Baja California, Baja California Sur, and Quintana Roo (see Figure A7).
* Use store-type information from either Census to create store-classification (modern versus traditional). Traditional stores contain the following store-types: public markets; other street vendors; specialized stores; convenience stores; grocery stores. Modern stores contain the following store-types: supermarkets, price clubs, department stores.
* Define the panel structure of the data by declaring the store ID (CLEE) and the time-variable (year)
* When the appended 2009-2014 Census is declared as a panel, fill out cases where the store is present in one Census but not in the other. In turn, define the main outcome variable: a dummy which equals 1 if the store is present in a particular Census, and 0 otherwise.

**Censos\_2014-2019\_id\_uelm.dta**

Population definition:

* Retail Census for the years 2014 and 2019, containing information on all retailers operating in the country at the time.

Panel structure:

* Panel at the yearly level, connecting the 2014 Census to the 2019 Census (no data in intermediate years). Contains the following key variables:
  + Store-type (detailed store-type classification which we use to classify stores as ‘modern’ versus ‘traditional’)
  + Geographical location of store (used to define treatment and control groups)
  + Indicator variable for whether a unique store (identified by its identifier CLEE) is present in a particular Census

Construction:

* To construct this dataset, start with the 2014 Census. Since variables are pre-harmonized across Censuses, directly append with the 2019 Census.
* Define treatment versus control on the basis of border versus non-border location of municipality where price data is sampled. Border area includes: all municipalities within 20km of either the Northern border or the Southern border; and, all municipalities located in the states of Baja California, Baja California Sur, and Quintana Roo (see Figure A7).
* Use store-type information from either Census to create store-classification (modern versus traditional). Traditional stores contain the following store-types: public markets; other street vendors; specialized stores; convenience stores; grocery stores. Modern stores contain the following store-types: supermarkets, price clubs, department stores.
* Define the panel structure of the data by declaring the store ID (ID\_UELM) and the time-variable (year)
* When the appended 2014-2019 Census is declared as a panel, fill out cases where the store is present in one Census but not in the other. In turn, define the main outcome variable: a dummy which equals 1 if the store is present in a particular Census, and 0 otherwise.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **List of analysis datasets** | **Nr of obs** | **Unique identifier** | **Inputs** | **Years** |
| CPI\_CE2014\_merged.dta | *1,584,900* | CLEE, calendar month-year | CPI, 2014 Census | 2012-2015 |
| Censos\_2009-2014.dta | *7,126,866* | CLEE, year | 2009 Census, 2014 Census | 2009 and 2014 |
| Censos\_2014-2019\_id\_uelm.dta | *7,894,170* | ID\_UELM\*, year | 2014 Census, 2019 Census | 2014 and 2019 |

Table of analysis datasets

Footnote \*: ID\_UELM corresponds to the new ID variables that INEGI is transitioning towards, which replaces CLEE. We used ID\_UELM for the 2014-2019 Census analysis in Table A6 to conform with INEGI’s data-protocols.

|  |  |  |
| --- | --- | --- |
| **Do-files** | **Analysis data input** | **Results** |
| Fig6.do | CPI\_CE2014\_merged.dta | Figure6A.eps; Figure6B.eps; Figure 6C.eps; Figure6D.eps; Figure 6E.eps; Figure6F.eps |
| FigA8.do | CPI\_CE2014\_merged.dta | FigureA8A.eps; FigureA8B.eps |
| TableA6.do | Censos\_2014-2019\_id\_uelm.dta; Censos\_2009-2014.dta | TableA6\_PanelA.tex; TableA6\_PanelB.tex |
| FigA7.do | [no pre-loaded data input, instructions for creating map] | FigureA7.pdf |
| FigC1.do | CPI\_CE2014\_merged.dta | FigC1A.eps, FigC1B.eps |

Figures and Tables in RESULTS folder with corresponding do-files and analysis data-sets