

README for ‘Local Infrastructure and the Development of the Private Sector: Evidence from a Randomized Trial’ by Daniel Rogger[®] Leonardo Iacovone[®] Luis Fernando Sanchez-Bayardo[®] Craig McIntosh

June 2025

Data Availability Statement

This paper uses microdata from INEGI's (Mexico's Statistics Agency; see <https://en.www.inegi.org.mx/>) Economic Census held in 2009, 2014 and 2019; and the Population Census held in 2020. Further details about the data, the period covered, and variables used are provided in the paper.

These Economic Census data need to be accessed through INEGI's microdata lab. All do files must be run either at INEGI's premises or with the assistance of the microdata lab's staff. Once results are estimated, they are double-checked by INEGI to ensure no sensitive data is exposed or leaked.

To access the data an email to microdatos@inegi.org.mx must be sent. A data access application must be filled and included in the email (https://www.inegi.org.mx/contenidos/app/microdatos/laboratoriodatos/doc/Solicitud_Uso.pdf) along with CVs, copies/scans of work and official IDs of those involved in the project (usually there is a user of the data and a project supervisor). The application must include the source to analyse (i.e. Economic Census 2009, 2014 and 2019 in this case) as well as the variables to include in the analysis. The catalog of existing variables as well as their names (in the database) can be taken from each census microsite (in their respective questionnaires and tables). INEGI then will reach back to continue with the process.

The Population Census data we use is public and downloaded directly from https://www.inegi.org.mx/programas/ccpv/2020/#datos_abiertos for the 2020 census and https://www.inegi.org.mx/programas/ccpv/2010/#datos_abiertos for the 2010 census.

Working with data in INEGI's Microdata lab

Since 2014, INEGI (Mexico's Statistics Agency) has provided access to microdata from the censuses, surveys and other sources of data it collects and regularly publishes for its analysis by researchers from academia, international organizations and government agencies.

Once the microdata lab grants access to the data and the latter is available for analysis, the user can start working with it. The user comes to the lab to do the analysis in its premises. She uses one of the computers available to connect to the remote desktop that contains the requested data. Users cannot go into the lab with notebooks, phones, electronics nor any other tool that may facilitate copying/writing data without INEGI's approval. The provided remote desktop has no access to internet, nor to peripherals such as USB (external data must be sent to the microdata lab email to have it saved into the desktop, and has very limited software available (Excel, Word, Stata and other software such as R and ArcGIS by request) and functionalities (Windows Explorer with sole access to essential folders). The user has access to a folder where she can save any do files and auxiliary files, and another one to save results to request for release.

Once an analysis is finished, the user can save in a folder with the project's code and today's date the results she wants to be released (along with do files used to estimate them) and send an email to the microdata lab requesting for the results. Results may take from a couple of hours to some days to be released, depending on their complexity and size, and on the availability of staff of the area that generated the data (which are responsible for doublechecking of results and make sure confidentiality rules are respected).

The microdata lab is very zealous in maintaining confidentiality (of firms in this case) of observations. This usually means that results too granular (i.e., with very few observations [less than 10], very localized, disaggregated; and/or a combination of these three aspects) will be most likely not released. Their objective is to provide access to microdata for research purposes while at the same time avoiding individual firms or small sectors to be exposed to third parties.

This Replication Package

The do files included here are intended to construct the main database and then to estimate the results presented in the paper. A folder called "Auxiliary data files" contains auxiliary data needed for some exercises, and since INEGI requires ado files to be sent to the workstation, a set of custom commands used in the paper.

Specifically, we send the following do files to INEGI:

000_HABITAT_MASTER.do
Prep 00 Declare globals.do
Prep 01 Main database.do
Prep 02 Population census data.do
Table 1 Descriptives.do
Table 10 Saturation.do
Table 2 Main regressions.do
Table 3 Exit Entry regressions.do
Table 4 Surviving firms regressions.do
Table 5 Main heterogeneity regressions.do
Table 6 Mechanism regressions.do
Table 7 Formality regressions.do
Table 8 CensusPop regs.do
Table 9 Spillovers.do
Figure 1 Firm dynamics.do
Figure 2 Densities.do
Table A10 Spillovers MarketAccess.do
Table A11 Saturation baseline.do
Table A12 Polygon results.do
Table A13 Saturation Exit Entry.do
Table A14 Saturation main.do
Table A2 Balance Posttreatment.do
Table A3 Balance initial values.do
Table A4 Exit Entry regressions.do
Table A5 Exit heterogeneity.do
Table A6 Entrants.do

Table A7 Banking access by ownership BCKUP.do

Table A7 Banking access by ownership.do

Table A8 Spillovers Exit Entry.do

Table A9 Spillovers Exit Entry MarketAccess.do

The outputs from these do files are generated by INEGI and subsequently provided to the research team.

Project File Structure

The file structure of the project is the following:

Z:.

a +---Procesamiento

a a +---Trabajo

a a a +---Codigo2020

a a a Access to transport.dta

a a a BuffDatabaseD_100m.dta

a a a BuffDatabaseD_1km.dta

a a a BuffDatabaseD_250m.dta

a a a BuffDatabaseD_500m.dta

a a a CensosPob_polygonW99.dta

a a a IntersectControl2013.csv

a a a IntersectControl2014.csv

a a a IntersectControl2015.csv

a a a IntersectControl2016.csv

a a a IntersectControl2017.csv

a a a IntersectTreated2013.csv

a a a IntersectTreated2014.csv
a a a IntersectTreated2015.csv
a a a IntersectTreated2016.csv
a a a IntersectTreated2017.csv
a a a Localidades in Habitat.dta
a a a Localidades in Habitat19.dta
a a a Manzanas habitat 2009(NEW).dta
a a a Manzanas habitat 2014(NEW).dta
a a a Manzanas habitat 2019(NEW).dta
a a a market_access_measures.dta
a a a PolygonDatabase.dta
a a a PolygonDummiesv3.dta
a a a PolygonSize_sqm.dta
a a a Prices 2008 1dig.dta
a a a Prices 2008 2dig.dta
a a a Prices 2008 3dig.dta
a a a Prices 2008 4dig.dta
a a a Prices 2008 5dig.dta
a a a Prices 2008 6dig.dta
a a a Prices 2018 1dig.dta
a a a Prices 2018 2dig.dta
a a a Prices 2018 3dig.dta
a a a Prices 2018 4dig.dta
a a a Prices 2018 5dig.dta
a a a Prices 2018 6dig.dta
a a a 000_HABITAT_MASTER.do

a a a Prep 00 Declare globals.do
a a a Prep 01 Main database.do
a a a Prep 02 Population census data.do
a a a SaturationData(2021-11-23)Luis.dta
a a a Table 1 Descriptives.do
a a a Table 10 Saturation.do
a a a Table 2 Main regressions.do
a a a Table 3 Exit Entry regressions.do
a a a Table 4 Surviving firms regressions.do
a a a Table 5 Main heterogeneity regressions.do
a a a Table 6 Mechanism regressions.do
a a a Table 7 Formality regressions.do
a a a Table 8 CensusPop regs.do
a a a Table 9 Spillovers.do
a a a Figure 1 Firm dynamics.do
a a a Figure 2 Densities.do
a a a Table A10 Spillovers MarketAccess.do
a a a Table A11 Saturation baseline.do
a a a Table A12 Polygon results.do
a a a Table A13 Saturation Exit Entry.do
a a a Table A14 Saturation main.do
a a a Table A2 Balance Posttreatment.do
a a a Table A3 Balance initial values.do
a a a Table A4 Exit Entry regressions.do
a a a Table A5 Exit heterogeneity.do
a a a Table A6 Entrants.do

a a a Table A7 Banking access by ownership BCKUP.do
a a a Table A7 Banking access by ownership.do
a a a Table A8 Spillovers Exit Entry.do
a a a Table A9 Spillovers Exit Entry MarketAccess.do
a a a +---custom commands
a a a iebaltab.ado
a a a iebaltab.sthlp
a a a iebaltab_regcoeff (ORIG).ado
a a a iebaltab_regcoeff1.ado
a a a iebaltab_regcoeff2.ado
a a a ieboilstart.ado
a a a ieboilstart.sthlp
a a a ieddtab.ado
a a a ieddtab.sthlp
a a a iedorep.ado
a a a iedorep.sthlp
a a a iedropone.ado
a a a iedropone.sthlp
a a a iefolder.ado
a a a iefolder.sthlp
a a a iegitaddmd.ado
a a a iegitaddmd.sthlp
a a a iegraph.ado
a a a iegraph.sthlp
a a a iekdensity.ado
a a a iekdensity.sthlp

```
a a a   iematch.ado
a a a   iematch.sthlp
a a a   iesave.ado
a a a   iesave.sthlp
a a a   ietoolkit.ado
a a a   ietoolkit.sthlp
a a +---Insumos
a   a +---Censos Economicos
a     Insumo2009.dta
a     Insumo2014.dta
a     Insumo2019.dta
a +---results
    T1 desc stats.csv
    T10 Saturation Regs.xls
    T2 Main regs.xls
    T3 ExitEntry regs.xls
    T4 Survivors regs.xls
    T5 Main heterogeneity regs.xls
    T6 Access finance and tech regs.xls
    T7 Formality Regs.xls
    T8 Residents Regs.xls
    T9 Spillover Regs.xls
    TA10 Spillovers MktAccess.xls
    TA11 Saturation baseline.xls
    TA12 Polygon regs.xls
    TA13 Saturation ExitEntry.xls
```


TA14 Saturation main.xls
TA2 Balance Post.xls
TA3 Balance Tables All firms.xls
TA3 Balance Tables Commerce and Services.xls
TA3 Balance Tables Manufacturing.xls
TA4 ExitEntry regs.xls
TA5 Exit heterogeneity.xls
TA6 Entrants.xls
TA7 Bank access by ownwership.xls
TA8 Spillovers ExitEntry.xls
TA9 Spillovers ExitEntry MktAccess.xls

Additional Output

In addition to the do files provided in this folder, there are additional output in the paper created by images and maps from external sources.

Figure A1

Maps using geospatial data from INEGI, and original Habitat data polygon data.

Figure A2

Google street maps.

Figure A3

<https://maps.app.goo.gl/CUtWRZ9WoJraPXPd7>

<https://maps.app.goo.gl/ammmgJqh9m9UvXrHA>

Figure A4-A7

Maps using geospatial data from INEGI, and original Habitat data polygon data.

Table A1

Sourced from McIntosh, Craig, Tito Alegría, Gerardo Ordóñez, and René Zenteno. 2018. “The neighborhood impacts of local infrastructure investment: Evidence from urban Mexico.” *American Economic Journal: Applied Economics*, 10(3): 263–86

Notes on figures

- Figure 2 is estimated in Stata and manually created in Excel.

About Maps (Figures A1, A4, A5, A6, A7) and geospatial data

Maps are created using INEGI's geospatial system, called Marco Geostadístico Nacional. Such system contains up to block-level geospatial data of cities and towns, which is crucial to locate Habitat polygons. Given the ever-changing nature of population (and thus land occupancy) data, INEGI updates geospatial data when a significant new surveying event occurs (Censuses and large Surveys). This meant that several versions of geospatial data had to be harmonized to a preferred one in order to have consistent maps and create consistent geospatial exercises.

Namely, 4 versions of geospatial were harmonized into a single version. Each version corresponds to a different event covered by the analysis:

- 1) Geospatial data of 2005 (Cartografía Geoestadística Urbana 2005), the one originally used by the Habitat project to select the randomized blocks covered by the program.
- 2) Geospatial data of 2009 (Cartografía Estadística Urbana, Cierre de los Censos Económicos 2009), covering the 2009 Economic Census.
- 3) Geospatial data of 2014 geospatial data (Cartografía Estadística Urbana, cierre de los Censos Económicos 2014, DENUE 01/2015), covering the 2014 Economic Census.
- 4) Geospatial data of 2019 (Marco Geoestadístico, septiembre 2019), covering the 2019 Economic Census.

Geospatial data of 2005, 2009 and 2019 were harmonized to the 2014 version at block-level detail. The strategy was to find as many identical blocks as possible between the version that had to be adjusted with the 2014 version. In the vast majority of cases, blocks exactly overlapped. Whenever a block did not match exactly, as long as a significant part of its area overlapped with a 2014 block (50% and above its area), it was considered to be part of that block. In some cases, manual case-by-case adjustments had to be made.

Once block data was harmonized, Habitat polygons were drawn, and buffer areas at 100m, 250m, 500m and 1km were estimated. Blocks located totally or partly within these buffer areas were selected to perform spillover exercises.

With all these geospatial layers, it was possible to create the maps included in the paper.

All exercises were done in either QGIS or ArcGIS, with some intermediate steps done using StatTransfer and Stata.