Data and Code Appendix
Assessing the Incidence and Efficiency of a Prominent Place Based Policy

By MATIAS BUSSO, JESSE GREGORY, AND PATRICK KLINE

This document describes the data sources, build scripts, and estimation routines for "Assessing the Incidence and Efficiency of a Prominent Place Based Policy."

I. Data Sources

We utilized three groups of data sources. First, confidential data from population and economic censuses. Second, publicly available data collected from a variety of sources. And third, data that we obtained from HUD via a FOIA request. These last twos group of datasets are in the archive PublicData.zip

A. Confidential Data

The empirical analysis in this paper uses proprietary individual data from the census of U.S. population and establishment level data from the U.S. manufacturing data both housed at the Center for Economic Studies (CES), Bureau of the Census. The data for this project were accessed via the Census Research Data Centers at the University of Michigan, UC Berkeley and Suitland. We rely on two primary sources of confidential data:

1) Individual and household level data from the 1980, 1990, and 2000 long-form of the Decennial Censuses of Population and Housing, including the Journey to Work component of the Decennial Census. See United States Census Bureau (1980, 1990a, 2000a).

2) Establishment level data from the Standard Statistical Establishment List (SSEL) and the Longitudinal Business Database (LBD) for 1987-2002. We coded each establishment to a 1990 census tract using an algorithm described in Appendix B based on the raw street addresses provided on the SSEL. See United States Census Bureau (1987-2000a, 1987-2000b)

   We use these data to construct a panel of Census tract level outcomes and covariates.

B. Publicly Available Data

1) Geographic identifiers on the 1980 and 2000 files use codes pertaining to the census geographic boundaries of their vintage. We map block of residence and block

of work identifiers in 2000 to 1990 Census tracts using the Census Block Relationship Files (CBRF). See United States Census Bureau (2000b). To map 1980 geographic identifiers to 1990 tracts we use the Census Tract Relationship Files (CTRF). See United States Census Bureau (1990b). These dataset are in the directories /PublicData/CBRF/ and /PublicData/CTRF

2) City level covariates for 1980 and 1990 were obtained from the County/City Databook (CCD) from the 1988 and 1994 files. See United States Census Bureau (1988, 1994). We downloaded the data from http://www2.lib.virginia.edu/ccdb/ on December 16, 2005. This data yields values of city level variables such as crime rate, percentage of workers in the manufacturing sector and percentage of workers working in the government. In cases where zones span multiple cities we assign all tracts the characteristics of the largest city in the metropolitan area. These datasets are in the directories /PublicData/CityData/ and /PublicData/Crime&Educ/ .

3) We also use metropolitan housing price data (all transactions, not seasonally adjusted) from the Office of Federal Housing Enterprise Oversight (OFHEO) to control for changes in metropolitan housing market conditions in the early 1990s. See Federal Housing Finance Agency (2012). This dataset is located in /PublicData/HPI/ and was downloaded on January 30, 2012.

4) The Census 2000 "Record Type 1" and "Record Type 2" TIGER/Lines Files (available at http://www.census.gov/geo/www/tiger/tiger2k/tgr2000.html) are a required input to our geocoding routine.

C. HUD Data

1) In order to construct a suitable control group for the EZs, we obtained 73 of the 78 first round EZ applications submitted to HUD by nominating jurisdictions via a Freedom of Information Act (FOIA) request. Data from the applications including the tract composition of rejected zones were entered manually into Stata. We then merged this information with publicly available data on the tract composition of EZs and future zones to create a composite set of treatment and control indicators for use in our analysis. In order to minimize errors and to have flexibility when dealing with multiple geographies, these data are organized at the census block level and have both 1990 and 2000 geographic correspondences. This dataset is in /PublicData/AdminData/.

II. Scripts

Scripts are organized into seven directories based upon the tasks that they perform. First, we provide a general description of the goal of the programs included in each directory, then we describe the purpose of each program. Finally, each program is properly documented.

- **/1.administrative/:** The programs in this subdirectory prepare administrative data files to be merged to data from the decennial Census and the LBD.

- **/2.extract_from_decennial/:** The programs in this subdirectory extract records from the 1980, 1990, and 2000 decennial Census of Population and Housing (Census) files, perform some basic recodes, and standardize geographic variables from the three vintages.

- **/3.extract_from_lbd/:** The programs in this subdirectory extract records from the 1987-2004 Longitudinal Business Database (LBD) and Standard Statistical Establishment List (SSEL) files, create "geocodes" from the establishments' raw street addresses, and create inverse probability weights to account for the fact that some establishments were not successfully geocoded.

- **/4.process_lbd/** The programs in this subdirectory create several measures from the raw LBD and SSEL records and collapse the microdata to create a panel of tract-level business outcomes.

- **/5.buildWork/** The programs in this subdirectory collapse the Census microdata at the tract-by-year level, merge on administrative data, and merge on the collapsed LBD and SSEL data.

- **/6.library/** This subdirectory contains subroutines (.ado files) that are called by the data processing and estimation programs.

- **/7.estimation/** The programs in this subdirectory perform estimation.

A. /1.administrative/

- **0.master_admin.do:** This program calls the other programs in this subdirectory in sequence.

- **1.admin4merge.do:** This program loads a file (blocklevel_admdata.dta) that contains a record of which, if any, of the HUD local subsidy programs each 1990 Census block and each 2000 Census block received and saves only the records for the 1990-vintage geography. The resulting file is later merged to the collapsed tract-level file that we create from the Census and the LBD.

- **2.cities_and_counties.do:** This program loads data from the County/City Databook and reshapes the information so that it may be merged to our tract-level panel.

- 3.**hpiextract2.do:** This program loads FHFA Housing Price Index data and reshapes the information so that it may be merged to our tract-level panel.

B. /2.extract_from_decennial/

- **0.master.bat:** This program calls the other programs in this subdirectory in sequence.

- **1a.Create_Decennial_Extract1980.sas:** This program extracts and merges variables from the person, household, and geography files of the 1980 decennial Census.

- **1b.Create_Decennial_Extract1990.sas:** This program extracts and merges variables from the person, household, and geography files of the 1990 decennial Census.

- **1c.Create_Decennial_Extract2000.sas:** This program extracts and merges variables from the person, household, and geography files of the 2000 decennial Census.

- **1d.Area1990.sas:** This program extracts the land area of each 1990 Census block.

- **2a.Recode_Decennial_Extract1980.do:** This program recodes the raw 1980 decennial Census extract and converts 1980 Census tracts to 1990 Census tracts.

- **2b.Recode_Decennial_Extract1990.do:** This program recodes the raw 1990 decennial Census extract.

- **2c.Recode_Decennial_Extract2000.do:** This program recodes the raw 2000 decennial Census extract and converts 2000 Census blocks to 1990 Census blocks.

- **2d.Create_Lists.do:** This program creates a list of counties to be looped over in later programs and saves the deciles of the wage distribution in sample counties in 1980, 1990, and 2000.

- **3.Create_POW_Missingness_Weights.do:** This program creates a set of inverse probability weights to account for the partial missingness of "places-of-work" records.

- **4.Add_POW_Missingness_Weights_to_Statefiles.do:** This program merges the place-of-work-missingness inverse probability weights to the decennial Census microdata.

- **5.Extract_Block_XY.sas:** This program extracts that latitude, longitude, and land area of each 1990 Census block.

- **6.Create_Tract1990_Area_and_Centroid.do:** This program computes the centroid (latitude and longitude) and total land area of each 1990 Census tract.

C. /3.extract_from_lbd/

- **1987_1992.sas:** This program calls each of the programs in this subdirectory in sequence.

/3.extract_from_lbd/01-Merge SSL-LBD Extracts/

- **1987_1992.sas:** This program extracts and merges records from the 1987-1992 SSEL and LBD files.

- **1993_1997.sas:** This program extracts and merges records from the 1993-1997 SSEL and LBD files.

- **1998_2001.sas:** This program extracts and merges records from the 1998-2001 SSEL and LBD files.

- **2002_2004.sas:** This program extracts and merges records from the 2002-2004 SSEL and LBD files.

- **99-Copy_SSLLBDs_to_Stata.sas:** This program copies the 1987-2004 SSEL-LBD SAS extracts to Stata.

/3.extract_from_lbd/02-geocoding/ /3.extract_from_lbd/02-geocoding/01-Extract Addresses for Geocoding/

- **Get_Addresses_From_Full_Data_Sets.sas:** This program keeps the street address information from the SSEL-LBD panel for each establishment in each year.

/3.extract_from_lbd/02-geocoding/02-Batch Geocoding/

- **Geocode.sas:** This program computes geocodes from the raw SSEL street addresses using the procedure described in Appendix II of Busso, Gregory, and Kline (2012).

/3.extract_from_lbd/02-geocoding/03-Longitudinal Imputation/

- **01-Create_Long_Imp_Dataset.sas:** This programs appends the results of the geocoding algorithm (both the successfully geocoded records and the non-matched records) and copies the result to a Stata dataset.

- **02-Longitudinal_Imputation_block.do:** This program attempts to impute a geocode for each non-matched establishment-year using any successfully-geocoded observations for the same establishment in other years.

- **03-Consolidate_Block_Data.sas:** This program creates a simplified Stata dataset containing all of the geocodes.

/3.extract_from_lbd/03-Create_Missingness_IP_Weights/

- **01-Create_Multivariate_County-Year_Missingness_IP_Weights.do:** This program creates a set of inverse probability weights to account for the fact that some establishments were not successfully geocoded.

D. /4.process_lbd/

- **master_lbd.do:** This program calls the other programs in this subdirectory in sequence.
- **00.flag_SSEL_Extract.sas:** This program extracts several variables from the LBD and SSEL that were not included on the initial extract.
- **01.compile_lbd.do:** This program cleans and labels the LBD extract.
- **02.flags_lbd.do:** This program merges the geocodes and flags to the basic SSEL-LBD extract.
- **03.description_all.do:** This program creates a codebook for each year of the SSEL-LBD panel.
- **04.LBD.do:** This program appends all years of the SSEL-LBD panel and keeps the variables that are required for our analysis.
- **05.LBD_vars.do:** This program creates measures that require the panel structure of the SSEL-LBD.
- **06.Create_LBD_full.do:** This program merges the administrative file to the SSEL-LBD panel file. It also creates new variables that are based on each establishment's EZ "treatment" status.
- **07.LBD_collapse.do:** This program collapses the SSEL-LBD data at the tract level in years 1987, 1992, 1997, and 2000 and performs some recodes.
- **08.LBD_mergeDecennial.do:** This program prepares the tract-level SSEL-LBD data to be merged with the decennial Census.
- **R1.LBD_simple_annual_collapse.do:** This program collapses a smaller set of SSEL-LBD variables at the tract level at an annual frequency.

E. /5.buildWork/

- **library.ado:** This program defines local macro variables that reference the subdirectories of the /data/ directory and defines two subroutines.
- **0.master_tract.do:** This program runs the other programs in this subdirectory to create a tract-level panel of Census-derived variables.
- **2.build_POR.do:** This program performs some recodes and collapses the data. The Census year and the unit of geography (either the Census *tract of residence* or Census *block of residence*) are passed to the program as arguments.
- **3.build_POW.do:** This program performs some recodes and collapses the data. The Census year and the unit of geography (either the Census *tract of work* or Census *block of work*) are passed to the program as arguments.

- **4.build_MigrationLogits.do:** This program estimates tract-level outmigration rates. It is used to construct one of the covariates in our estimation routine.
- **5.build_cells.do:** This program performs regression adjustments to place-of-residence-based variables.
- **6.build_lnwage.do:** This program performs regression adjustments to place-of-work-based variables.
- **7.compute_dist_to_cbd.do:** This program computes the distance from each tract to the central business district.
- **8.compile_work_tract.do:** This program merges the tract-level SSEL-LBD data, the city-level data, and the tract-level administrative data with the collapsed decennial Census data.
- **10.compute_lnvalue_city_means.do:** This program computes average log-housing values by city-year and by county-year.
- **code_industries.do:** This program recodes 3-digit industry classifications to broad 1-digit industry classifications.
- **code_rent80_intervals.do:** This program records the range (low endpoint and high endpoint) of the category responses to the 1980 Census rent question.
- **code_rent_intervals.do:** This program records the range (low endpoint and high endpoint) of the category responses to the 1990 and 2000 Census rent questions.
- **code_value80_intervals.do:** This program records the range (low endpoint and high endpoint) of the category responses to the 1980 Census housing value question.
- **code_value_intervals.do:** This program records the range (low endpoint and high endpoint) of the category responses to the 1990 and 2000 Census housing value questions.

F. /6.library/

- **balance.ado:** This subroutine creates a "balance" table comparing variable means in EZs, controls, and reweighed controls.
- **bo2.ado:** This subroutine computes Blinder-Oaxaca (aka Parametric Reweighting) treatment-effect estimates and reports analytical standard errors.
- **deg2rad.ado:** This subroutine converts degrees to radians.
- **earthdistance.ado:** This subroutine computes the distance between two longitude-latitude pairs.
- **formatable.ado:** This subroutine formats estimation output.

- **formatvarnames.ado:** This subroutine formats variable names for estimation output.
- **logit2.ado:** This subroutine estimates a logit, the propensity score and it also allows for variable trimming points.
- **movestuff.ado:** This subroutine is used to move estimation log files from the various specifications to the appropriate "/results/" folder.
- **overlap.ado:** This subroutine compares the percentiles of the distribution of the propensity score in treatment and control tracts. We used this as a specification check reported in earlier versions of the paper.
- **psmatch2.ado:** This is Leuven's and Sianesi's subroutine for performing propensity-score matching.
- **pstest.ado:** This is an auxilliary routine referenced by psmatch2.ado.
- **reweightvar.ado:** This subroutine accepts a variable name and a weight name as inputs and creates a new variable equal to the indicated variable multiplied by a standardized version of the indicated weight. An unweighted sum of the newly-defined variable returns the weighted mean of the indicated variable. This routine is useful for creating weighted means of multiple outcomes using a single "collapse" statement when not all outcomes are weighted by the same weighting variable.
- **spatialma_loo_norm.ado:** This subroutine computes a spatial moving average of each variable in the variable list for each tract included in the sample. For each tract j, the spatial moving average is computed excluding the value from tract j itself. The average is computed using a truncated normal kernel with a standard deviation of half a mile. The truncation point for the spatial moving average (in miles) is entered as an additional argument. In the paper we truncated the kernel at 1 mile.
- **sphdist.ado:** This subroutine computes the distance between two longitude-latitude pairs.
- **wild.ado:** This subroutine computes OLS treatment effect estimates and reports inference based on a clustered wild bootstrap-t procedure.
- **wild_bo.ado:** This subroutine computes Blinder-Oaxaca treatment effect estimates and reports inference based on a clustered wild bootstrap-t procedure.
- **wild_bo_ci.ado:** This subroutine computes Blinder-Oaxaca treatment effect estimates and clustered wild-bootstrapped confidence intervals.
- **wild_ci.ado:** This subroutine computes OLS treatment effect estimates and clustered wild-bootstrapped confidence intervals.

G. /7.estimation/

- **0.master.do:** This file uses the programs in this folder to compute the paper's main difference-in-differences estimates and to perform a series of robustness checks.
- **1.setup.do:** This program loads the subroutines (.ado files) that are used during estimation and defines global macros containing lists of outcome variables and lists of covariates.
- **2.dataprep.do:** This program prepares the data for estimation.
- **2a.workfile_winlose.do:** This program imputes local wages in tract-years in which no local workers were observed.
- **2b.workfile_placebo.do:** This program defines the placebo sample and defines "treatment" variables based on tracts' placebo status.
- **2c.workfile_percentile.do:** This program computes within-city percentiles of key outcome variables.
- **2d.workfile_indices.do:** This program computes six index variables that we use to summarize the balance of pre-treatment trends between EZs and control areas.
- **3.pscore.do:** This program loads the data, selects the estimation sample, and creates a balance table.
- **4.balance.do:** This program creates figures plotting 1980-1990-2000 outcome trends in EZs, controls, and reweighted controls.
- **5.estimation.do:** This program is used to conduct estimation. It accepts arguments specifying the estimator, the covariate list, and the inference method.
- **6.tables.do:** This program organizes estimation results into tables and adds significance *'s.
- **7.samplesizes.do:** This program counts the number of observations and the number of clusters included in our various regression specifications.
- **8.annual_lbd_plots.do:** This program plots SSEL-LBD-based outcomes by year in EZs, control areas, and reweighted control areas.
- **9.descriptive.do:** This program computes zone-level aggregate measures to be used in the paper's welfare analysis.
- **10.getCIs.do:** This program computes wild-bootstrapped confidence intervals for the treatment effect estimates used in the welfare analysis.
- **11.secondmoments.do:** This program tabulates the variance of census tract characteristics in EZs, control areas, and reweighted control areas.

- **12.results_to_excel.do:** This program arranges the estimation output to match the shape of our table shells.

III. Running Scripts

The set of scripts were written so that typically there is a master script that calls all the other codes within each directory. To reproduce the published results, organize the files in the appropriate directories and run the scripts in the following order:

```
stata-mp -b do  /1.administrative/0.master_admin.do

bash  /2.extract_from_decennial/0.master.bat

bash  /3.extract_from_lbd/0.master.bat

sas  /4.process_lbd/00.flag_SSEL_Extract.sas

stata-mp -b do  /4.process_lbd/master_lbd.do

stata-mp -b do  /5.buildWork/0.master_tract.do

stata-mp -b do  /7.estimation/0.master.do
```

Wait for each job to finish before submitting the next one.

IV. References

1) United States Census Bureau. 1980. “Decennial Census of Population and Housing.” United States Department of Commerce.

2) United States Census Bureau. 1988. “County and City Data Book: 1988.” United States Department of Commerce. Accessed online at http://www2.lib.virginia.edu/ccdb/ on December 16, 2005.

3) United States Census Bureau. 1987-2000a. “Longitudinal Business Database.” United States Department of Commerce.

4) United States Census Bureau. 1987-2000b. “Standard Statistical Establishment List.” United States Department of Commerce.

5) United States Census Bureau. 1990a. “Decennial Census of Population and Housing.” United States Department of Commerce.

6) United States Census Bureau. 1990b. “Census Tract Relationship File.” United States Department of Commerce.

7) United States Census Bureau. 1994. "County and City Data Book: 1994." United States Department of Commerce. Accessed online at http://www2.lib.virginia.edu/ccdb/ on December 16, 2005.

8) United States Census Bureau. 2000a. "Decennial Census of Population and Housing." United States Department of Commerce.

9) United States Census Bureau. 2000b. "Census Block Relationship File." United States Department of Commerce.

10) Federal Housing Finance Agency. 2012. "Metropolitan Statistical Areas and Divisions through 2012Q3 (Not Seasonally Adjusted)." accessed online on January 30, 2012 at http://www.fhfa.gov/webfiles/24667/3q12hpi_cbsa.txt