Description of replication files for:

Title: *Comparing Treatments across Labor Markets: An Assessment of Nonexperimental Multiple-Treatment Strategies*
Authors: Carlos A. Flores and Oscar A. Mitnik
Review of Economics and Statistics, December 2013, Vol. 95, No. 5: 1691–1707

This file contains information on the data and programs used to produce the results in the paper above. The information in this file is organized as follows:

1. Data
    1.1. NEWWS Data
    1.2. Local Economic Conditions Data
2. Programs
    2.1. Stata User-Generated Commands Needed
    2.2. Data Preparation Programs
    2.3. Programs Used to Obtain the Results in the Paper
    2.4. Estimation Code Programs
    2.5. Instructions on Where the Results in the Paper can be Found in the Output Generated by the Programs

All the files described below are located in the zip file **Flores_Mitnik_2013.zip**. Please address any questions to Carlos A. Flores (carlos.arturo.flores@gmail.com) or Oscar A. Mitnik (oscar@mitnik.net).

Note that because all confidence intervals in the tables are based on 1,000 bootstrap replications, some of these programs take a long time to run.[1]

1. DATA

1.1 NEWWS Data

The main data employed in the paper comes from the public use version of the National Evaluation of Welfare-to-Work Strategies (NEWWS) data. Despite the data being public, users are not allowed to distribute it. For instructions on gaining access to the public use version of the NEWWS data see: http://aspe.hhs.gov/hsp/newws/data-info.htm - info.

[1] Many of the programs used to generate the results in the paper are very computationally intensive, especially those used to generate some of the results in Table 5 (see paper for details). Some of the models in those programs (e.g., in Table5_9_results.do; see section 2.3 in this file) may require powerful computers to converge. For example, while some of those models failed to converge when the programs were run in regular desktops, the programs ran well when we used 8-core computers (which were used to generate all the results in the paper).

From the raw data provided to us after our request, we employed the data in the file N5pi1809.txt, with path: NEWWS1/Public/5Year/Fullsamp/data/N5pi1809.txt, in "CD #1" (see: http://aspe.hhs.gov/hsp/newws/data/2yr-client/npbcover.htm). The NEWWS data were read in SAS using the following program, which we provide: **read.sas**. This program uses the SAS input statements and variable label statements available at http://aspe.hhs.gov/hsp/newws/data-info.htm - info.

The variable labels used in the programs are given in the file **n5pilabl.txt**, which we provide (and is available at http://aspe.hhs.gov/hsp/newws/data-info.htm - overview). The file appears as one of the files in the downloadable zipped archive file p5fullsamp.zip (see subsection 5-Year Full Impact File in that website). More information about the data we employed can be found in that downloadable zipped archive file. Once the data set was read in SAS, we converted it to Stata format.

1.2 Local Economic Conditions Data

Another part of the data employed in the paper relates to the local economic conditions (LEC) of the sites analyzed. We employed three measures of LEC for the different cohorts within each site at the metropolitan statistical area (MSA): employment-to-population ratio, average real earnings, and unemployment rate. The information on employment (series number CA34, line code 20), population (CA05, 20) and private nonfarm earnings (CA05, 90) for the years 1989 to 1999 comes from the Bureau of Economic Analysis (www.bea.gov), while the information—for the same years—on unemployment rates (Local Area Unemployment Statistics) and consumer price index comes from the Bureau of Labor Statistics (www.bls.gov).

We provide the data on the three LEC variables we employed in the paper in the Stata file: **lec_data_msa_wide.dta**, which contains the following variables:

alphsite:	JOBS site (alphabetically) (1-7, for the 7 sites in the NEWWS)
radatp:	Random Assignment Date Phone 2-YYMMDD (1991-1994)
unemp_rate0:	Unemployment rate on the year of Random Assignment (*RA)*
unemp_rate1:	Unemployment rate one year after RA
unemp_rate2:	Unemployment rate two years after RA
unemp_rate_1:	Unemployment rate one year before RA
unemp_rate_2:	Unemployment rate two years before RA
tot_emp_pop0:	Total employment to population ratio on the year of RA
tot_emp_pop1:	Total employment to population ratio one year after RA
tot_emp_pop2:	Total employment to population ratio two years after RA
tot_emp_pop_1:	Total employment to population ratio one year before RA
tot_emp_pop_2:	Total employment to population ratio two years before RA
avg_tot_rern0:	Average total real earnings on the year of RA
avg_tot_rern1:	Average total real earnings one year after RA
avg_tot_rern2:	Average total real earnings two years after RA
avg_tot_rern_1:	Average total real earnings one year before RA
avg_tot_rern_2:	Average total real earnings two years before RA

2. PROGRAMS

The programs below were created using Stata 11.2 and thus should run in that version of Stata (or higher). In most programs there is the chance of specifying local macros with the paths of the desired directories where results will be saved, and from where data is read.

2.1 Stata User-Generated Commands Needed

The following Stata user-generated commands need to be installed before running any of the estimation programs in Section 2.4 below:

- *center*: install with this command:
```
ssc install center, replace
```

- *estout*: install with this command:
```
ssc install estout, replace
```

- *kdens*: install with this command:
```
ssc install kdens, replace
```

- *locpoly*: install with this command:
```
net install st0053_3.pkg, from(http://www.stata-
journal.com/software/sj6-4/) replace
```

- *mat2txt*: install with this command:
```
ssc install mat2txt, replace
```

- *moremata*: install with this command:
```
ssc install moremata, replace
```

- *svmat2*: install with this command:
```
net install dm79.pkg, from(http://www.stata.com/stb/stb56/)
replace
```

- *svmatf*: install with this command:
```
ssc install svmatf, replace
```

For simplicity, we provide a do-filed called **stata_addons_install.do** that runs all the installation commands specified above.

2.2 Data Preparation Programs

- **prepare_data.do**
  This program prepares the data employed in our analysis. For example, it constructs the specific sample used in our analysis (e.g., only females), it creates some of the variables used in our analysis, it merges the LEC variables with the NEWWS data, and defines growth rates of the LEC variables. Two datasets are created by this program: **analysis.dta**, which is the main dataset used to obtain all results in the paper, and the dataset **benchmark_analysis.dta**, which is identical to the analysis dataset, except that the treatment variable (site identification) has been assigned randomly, to create the "placebo experiment" (benchmark) results in Tables 2, 3 and 4 (see paper for details).

2.3 Programs Used to Obtain the Results in the Paper

- **main_results.do**
  This is the main program that generates the results presented in Tables 1 to 4 and Figure 1 in the paper. In addition, this program can be used to generate many of the results presented in the Internet Appendix of the paper. In general, this program only defines options that allow obtaining any type of results (including those in Table 5 and in the Appendix—see below). These options include (among others):
  - Which variables are used in each step of the program.
  - The number of sites considered (local "sites"), e.g., 4 or 5.
  - Whether the results are shown for the outcome in levels, differences, or both (local "did").
  - The amount of overlap imposed, including no overlap at all (local "suff_ovlp" defined by a value "q"), i.e., value "q" in paper (see paper for details, e.g., equation (7)).
  - Model used to estimate the GPS and whether the GPS is re-estimated after imposing overlap (local "suff_mthd"). For example, use "mnl" for multinomial logit, and "mnl_r" for re-estimating with MNL after imposing overlap.
  - LEC adjustment (local "lec_adjustment")
  - Rules for producing bootstrap results, figures, descriptive statistics, balancing analysis, etc.

  The options used to obtain the main results in the paper (Tables 1 to 4) are: multinomial logit model for the GPS, q=0.0025, outcome in levels and differences, GPS is not re-estimated after imposing overlap.

- **main_results_benchmark.do**
  This program is identical to main_results.do, but uses the benchmark_analysis.dta dataset and the randomly assigned site indicator to obtain the results (from the "raw mean, no overlap" estimator) for the bottom of Tables 2, 3 and 4.

- Programs to replicate each of the "lines" or "models" (e.g., type of GPS estimation) of Table 5 (lines 2 to 17, since line 1 just replicates results in Table 4):
  - Table5_2_results.do
  - Table5_3_results.do
  - Table5_4_results.do
  - Table5_5_results.do
  - Table5_6_results.do
  - Table5_7_results.do
  - Table5_8_results.do
  - Table5_9_results.do
  - Table5_10_results.do
  - Table5_11_results.do
  - Table5_12_results.do
  - Table5_13_results.do
  - Table5_14_results.do
  - Table5_15_results.do
  - Table5_16_results.do
  - Table5_17_results.do

  The "Table5_XX_results.do" programs are all similar to the "main_results.do" program. They just differ in which options or variables have been used for GPS estimation.

- **lec_dot_graphs.do**
  Program employed to obtain Figure 2 in the paper.

2.4 Estimation Code Programs

The following programs are written to run in Stata 11.2 or above.

- **main_estimation.do**
  This program is called from within the "main_results.do" and "Table5_XX_results.do" programs. This program calls the "estimation.ado" command, generates bootstrapped results for the treatment effects, generates tests of equality of means, and creates figures (not presented in the paper) with the estimated treatment effects (see below).

- **main_estimation_mata_inc.do**
  Mata code to calculate the *RMSD*, *MAD* and *Maximum Distance* measures (see paper for details); it is called by the "main_estimation.do" program.

- **b_bsi.ado**
  Command called by the "main_estimation.do" program to process bootstrapping results.

- **estimation.ado**
  This program gives the code for the various estimators of the treatment effects considered in the paper, and performs several summary statistics about the estimated GPS. This program calls the Mata functions created by the "gps_mata.do" and "balancing.do" programs. It also creates several alternative figures (see below) to better understand overlap and matching quality.

- **gps_mata.do**
  This defines the gps() Mata function that calculates the different GPS versions. This program has been compiled in Stata 11.2; "estimation.ado" uses the compiled version "gps.mo".

- **gmm_ipt_eval.ado**
  This Mata function is called by the gps() Mata function when estimating the GPS with IPT moment conditions.

- **balancing.do**
  This is a Mata function that checks the balancing of covariates after adjusting for the GPS (see paper for details). It has been compiled in Stata 11.2; the "estimation.ado" program uses the compiled version "balancing.mo".

2.5 Instructions on Where the Results in the Paper can be Found in the Output Generated by the Programs

For all Tables, the treatment values represent sites:
- Atlanta = 1
- Detroit = 3
- Grand Rapids = 4
- Portland = 6
- Riverside = 7

For Tables 2 to 5, the estimators 1 to 10 follow the order presented in the paper, and the outcomes 1 to 4 are as follows:
- levels = outcome 1
- levels LEC-adjusted = outcome 2
- differences = outcome 3
- differences LEC-adjusted = outcome 4

The results for the Tables and Figures in the paper, as well as those in the Internet Appendix, can be found in the following output files:

- **Table 1**
  A series of files in the results folder for the "main_results.do" program (default: results_5s_025p_mnl)
  - Part A raw means (and SDs): file "descriptive_stats.txt"

- Part A p-values joint equality of means Wald test: file "balancing_lev_joint_pval.csv"
- Part A RMSD: file "balancing_lev_joint_rmsds.csv"
- Part B joint equality of means across all sites: bottom of file "balancing_lev_joint_pval.csv"
- Part B difference of means each site versus other sites pooled: bottom of file "balancing_lev_t_bl_10.out" (adj_1 to adj_5 columns)
- Part C (overlap): bottom of file "descriptive_stats.txt"

- **Tables 2 to 4**
  Files in the results folder for the "main_results.do" program (default: results_5s_025p_mnl) for the 10 evaluated estimators, and files with the same names in the results folder for the "main_results_benchmark.do" program (default: results_5s_025p_mnl_benchmark) for the placebo experiment (benchmark) results.
  Note: For the benchmark results only estimator #1 (raw mean – no overlap) is used.
  - Table 2: Outcomes 1 to 4: files "table_5s_out_1.csv" to "table_5s_out_4.csv"
  - Table 3: Outcomes 1 to 4: files "estim_teffects_out_1.csv" to "estim_teffects_out_4.csv"
  - Table 4: Outcomes 1 to 4: files "table_exc_RIV_from_5s_out_1.csv" to "table_exc_RIV_from_5s_out_4.csv"

- **Table 5**
  Files located in the folders created by the do files (default: results_SUFF_Table5_LN, where "SUFF" changes for each GPS model, and "LN" represents the particular line in Table 5.
  Note: In all cases, only estimator 10 (IPW with covariates) results are presented in each "line" of Table 5.
  - Outcomes 1 to 4: files "table_exc_RIV_from_5s_out_1.csv" to "table_exc_RIV_from_5s_out_4.csv"

- **Figure 1**
  Different versions of Figure 1 are created in each result folder, in the sub-folder "figures". The version published in the paper is found in the main results folder (default: results_5s_025p_mnl), in the files "GPS_ovlp_quality.gph" (Stata) and "GPS_ovlp_quality.ps" (Postcript). Alternative versions are kernel-density based (GPS_kernel_density), LLR density-based (GPS_kernel_density_llr_dens) or LLR count-based (GPS_kernel_density_llr_count), and box-plots across sites (boxplot).
  Note: The folder contains also the representation of the treatment effects with figures using percentile-based confidence interval (prefix "p_") or normal-based confidence intervals (prefix "n_") after bootstrapping, for each of the estimators and the outcomes.

- **Figure 2**
  Figure 2 is created by the program "lec_dot_graphs.do". The files are figure_2.gph (Stata) and figure_2.ps (Postcript), and they are created in the same folder where the program is run, unless a folder for the files is specified in the program (local "figsdir").

- **Online Appendix Results**
  They follow the same logic of the above tables; the only change is the name of the folder where the results are stored, based on the particular options selected in the "main_results.do" Stata program.