Introduction to usgeogr

```
library(usgeogr)
```

Economists often utilize differences in policies across geographic regions in their identification strategies.[1]

The `usgeogr` package helps streamline such spatial identification strategies by providing several useful datasets and functions for working with commonly analyzed U.S. geographic regions.

This document introduces you to the `usgeogr` package. For more information, please visit the Github page.

Datasets

Baseline datasets

The below are general purpose datasets for U.S. geographic regions.

Census regions and divisions

`census_df` provides a list of U.S. states and their assigned census regions and divisions.

States

`state_df` provides a list of U.S. states and postal codes.

`border_coord_df` provides coordinates for mile long segments of state borders in the continental U.S.

Counties

`county_df` provides a list of counties in the continental U.S.

`adjacent_county_df` provides a list of all adjacent counties in the continental U.S.

ZIP codes

`zip_df` provides a list of all ZIP codes in the continental United States. If the ZIP can be mapped to a ZCTA, then the dataset provides information on its population (as of 2010), the coordinates of its centroid, and the distribution of land, water, and housing usage.

Census tracts

Coming soon!

Datasets for cross-border identification strategies

Cross-border identification strategies exploit differences in policies along state borders. In such strategies, the sample is constrained to units that reside within a “narrow region” of a state border. The identification assumption is that shocks correlated with the policy affect units on both sides of the border symmetrically. Therefore, comparisons of units on one side of the border versus the other isolates the effect of the policy.

When implementing a cross-border identification strategy, the analyst must decide what constitutes a “narrow region” of a state border. The most common choices are border counties (regardless of the unit of analysis) and distance strips.

Border county designs

In a border county design, the analyst constrains the sample to all units that reside within border counties. Panel estimators then exploit differences in policies across adjacent border counties over time.

To implement this identification strategy in practice, each border county needs to be matched with at least one adjacent cross-border neighbor. One issue, however, is that border counties do not line-up perfectly with one another; almost every border county in the U.S. is adjacent to at least two cross-border neighbors.

There are several methods for addressing this issue of non-uniqueness. The most commonly used methods vary by the unit of analysis of the study (e.g., county-level versus individual-level).

County-level studies

If the unit of analysis is at the county level, then the most common method for addressing non-uniqueness is to construct a full dataset of adjacent cross-border county pairs (including replicates of some counties).[2]

`cbcp_df` assigns border counties to cross-border county pairs following the methods in Dube, Lester, and Reich (ReSTAT, 2010). If a border county has $p > 1$ cross-border neighbors, then it will have $p > 1$ entries in the dataset. There are 1,184 unique border counties, 1,308 cross-border county pairs, and 2,616 rows in the dataset.

Within county-level studies

If the unit of analysis is more granular than the county level (e.g., the individual level), then the replicate-based method of Dube, Lester, and Reich (ReSTAT, 2010) cannot be feasibly implemented. This is because it will massively increase the sample size (e.g., for individual-level studies).[3]

The most commonly used method for addressing this issue is to assign each border county to a distinct “cross-border cluster”. There are numerous ways to perform this assignment. The `usgeogr` provides several choices for assigning border counties to clusters; these assignments are stored in the `cbcounty_df`. Only border counties are contained in `cbcounty_df`.

Each border county can only belong to one cluster under each assignment mechanism. The choices for cross-border cluster assignments are:

1. State border segments. This method assigns each border county to clusters defined by state border pair segments (e.g., AL-FL, NV-UT, etc.). If a border county is adjacent to more than one state border segment, then the border county is assigned to the closest border segment to its population center.[4]
2. Border strip segments. This method assigns each border county to clusters defined by 50 mile-long state border strips.
3. Couplets. This method assigns each border county to a cluster using the “couplet” and “relaxed couplet” algorithms.[5] The couplet algorithm sequentially assigns counties to clusters of size two (i.e., “couples”) with an objective of maximizing the total number of matched couplets. The relaxed couplet algorithm allows for more than two counties per cluster.

4. Centers of mass. This method assigns each border county to a cluster using the “max-method” and “relaxed max-method” algorithms.[6]. The max-method algorithm assigns counties to clusters based on their centrality and the number of counties they border.

Note that some counties may be unassigned to a cluster using the couplet or max-method algorithms. These should be removed prior to assigning any fixed effects or factors (to avoid null assignment). For analyses at the within-county level, the county cluster identifiers should simply be appended to the units of analysis. Analysts can also hand-pair counties if they are working with a small subset of states.

ZIP code-level studies

`cbzip_df` assigns ZIP codes to unique cross-border clusters based on the assignment of their cross-border counties. It also assigns ZIP codes to unique cross-border strips, discussed in the below subsection. This is just a special case of the within-county level assignments above.

Border strip designs

In a border strip design, the analyst constrains the sample to all units that reside within a certain distance of a state border (e.g., less than 50 miles). The analyst then assigns units to border strips of a chosen length (e.g., 20 mile strips along the border). Panel estimators then exploit differences in policies across units assigned to the same border strip but residing in different states.

ZIP codes as a proxy for location

It is often the case that the exact geographic location of the unit being studied (e.g. an individual) is unknown. To resolve this issue, ZCTA centroids are commonly used to proxy for a unit’s location.

`cbzip_df` provides the distance from each ZCTA centroid to the nearest state border.[7] It also assigns each ZCTA to its nearest 20 mile long border strip, along with adding identifiers for whether the ZIP code resides in a border county and the county’s cross-border cluster assignment.

Functions

Distance functions

`county_dist` returns the distance between county population centers. `zip_dist` returns the distance between ZCTA centroids. Both functions are vectorized.

`county_to_state_border` returns the distance between a county population center and the nearest state border. `zip_to_state_border` returns the distance between a ZCTA centroid and the nearest state border. Both functions are vectorized. Both functions provide an option for returning the nearest border identifier instead of the distance.

---

1. For example, Gopalan, Hamilton, Kalda, and Sovich (2019) utilize differences in the minimum wage across state borders to estimate the employment effects.↩
2. An adjacent cross-border county pair is defined as a pair of adjacent border counties that reside in different states. A single border county can belong to multiple adjacent cross-border county pairs. The definition can be traced to Dube, Lester, and Reich (ReSTAT, 2010).↩

3. This is because some counties have several cross-border adjacent neighbors. For example, counties may be geographically staggered so that each county borders approximately 2 other cross-state counties (creating overlap). Only if each cross-broder county had a single cross-border neighbor could we have a unique and natural mapping of pairs for the groupings. However, there is not a single such unique mapping in the United States.↩
4. A small set of counties will be mapped to a state border segment for which only one county or state is eventually assigned. Hence, variation from these border counties will be discarded in panel studies that utilize spatial variation across time.↩
5. The couplet algorithm proceeds as follows. Start with all counties in a donor pool. Select the first county. Pair the first county with one of its adjacent cross-border neighbors. This forms a cross-border cluster; assign no other counties to this cross-border cluster. Remove both assigned counties from the donor pool. Select another county and repeat until either no counties remain in the donor pool or all counties have been selected. If there are no possible cross-border assignments for a county in the donor pool, leave it unassigned but keep it in the donor pool. For each iteration, start by selecting the county and cross-border neighbor who are adjacent to the fewest counties (unconditionally). The algorithm leaves about 20 percent of counties without an assigned cluster; each cluster has exactly two members. For the relaxed couplet algorithm, assign the remaining counties to the cluster of their alphabetically first cross-border neighbor. These clusters will have more than two constituent counties.↩
6. The max-method algorithm proceeds as follows. For each county, calculate how many counties it is adjacent to other counties that lie across state borders. On the first pass of the algorithm, assign each county an identifer that corresponds to either itself or its adjacent county with the most cross-border neighbors, whichever has more cross-border neighbors (i.e., the center of gravity). If this identifier has at least two counties assigned to it, then let this county's FIPS code define a max-method cluster. If this identifier does not have at least two counties assigned to it, then mark the max-method cluster as unassigned. On the second pass, the relaxed max-method algorithm assigns unassigned counties to to max-method cluster of one of their cross-border neighbors. If this cross-border neighbor is also not matched, then a new identifier is constructed from their pairing based on the numerical order of their FIPS codes. Finally, we check again for at least two members in each cluster. Mark counties with only one member in their cluster as unassigned. Roughly 25% of counties are unassigned to a cluster using the max-method algorithm, and 4.3% are unassigned using relaxed max-method algorithm.↩
7. For ZIP codes that cannot be matched to a ZCTA, these distance measures and assignments are left null.↩