Vera INSTITUTE OF JUSTICE

March 2026

Incarceration and Inequality Data Explorer

“Similar Counties” Methodology

Overview

The Incarceration and Inequality Project (IIP) Data Explorer, created by the Vera Institute of Justice (Vera), allows users to examine trends in, and relationships between, measures of incarceration and a range of economic metrics—both at the county-level and for the entire country. The Explorer includes multiple views, allowing users to compare metrics for different demographic groups and levels of geography, providing easy-to-use information to stakeholders and advocates.

The Explorer is built on the foundation of Vera's IIP Public Dataset, incorporating calculated measures and the results of a multi-phase analysis, to provide a set of metrics that local advocates and other stakeholders can use to better understand their own counties. Vera will update the Explorer as further data releases become available.

This document contains a detailed description of the variables and methods underlying the "Similar Counties" sections of the Data Explorer. For a brief overview of this information, and details of the Explorer's data sources, please refer to the Technical Documentation Summary.

"Similar Counties"

Vera conducted a multi-phase machine learning analysis to create groups of counties with similar characteristics and trends across incarceration and economic indicators. Vera repeated a similar set of analyses four times, focusing each on one of four different domains: inequality, incarceration, economic opportunity, and economic stability. The result of each of these analyses was a cluster assignment (i.e., a number identifying which group the county belonged to). For each county, Vera combined urbanicity and cluster assignments (see below) to create a group of counties that were similar across the analyzed domains.

Before conducting any analysis, Vera data scientists filtered the dataset to the years with the most complete data across all variables at the county level, which were 2009 through 2019. The U.S. Census Bureau did not collect two variables, health insurance coverage and public assistance utilization, in the American Community Survey (ACS) until 2012 and 2010, respectively. Vera backward filled this data and linearly interpolated all other missing data where any single years of data were missing.

Vera took the following steps to generate each set of cluster assignments (aside from differences noted in each section):

1. Vera calculated time-series features (median and slope) across each variable in the domain to create a feature matrix.
    a. Median values capture the baseline level for a county, and slope values capture magnitude and direction of change over time. Together, these features allow the machine learning analysis to identify similarities across either or both baseline rates and changes over time.
2. Vera data scientists conducted k-means clustering using the time-series feature matrices.
    a. Vera Examined the silhouette widths and Uniform Manifold Approximation Projections (UMAP) to determine the appropriate number of clusters (k) for the next phase.
3. Vera data scientists applied Z-transformations to the time-series features to normalize and reduce skew.
4. Vera conducted Dynamic Time Warping (DTW) clustering on z-transformed time series features.

Vera used DTW clustering to allow for variations in the temporal alignment of trends; counties that experienced similar trends—but during different periods of time—can be considered similar, allowing users to discover other counties that may have been through similar experiences of incarceration and inequality in the past as well as concurrently.

Inequality clusters

Vera assigned counties inequality clusters within three urbanicity groups; rural, small/midsize metros, and suburban/urban counties. Other inputs to the model included inequality ratios for jail population rate, health insurance coverage rate, home ownership rate, median income, and unemployment rate.

Vera calculated the inequality ratios by dividing the relevant rate for the Black, Indigenous, and People of Color (BIPOC) population by the rate for the white population (or the white/not Latinx population, where possible to extract). For median income inequality ratios, the numerator is the weighted median income for the BIPOC population divided by the median income for the white (not Latinx) population. Vera weighted the BIPOC median income by the population size of each individual BIPOC group to more closely represent median income amongst those groups, as raw income data is not available for more precise calculations.

Because an analysis of racial inequality requires valid data for both BIPOC and white populations in each county, Vera took additional steps to handle missing or near-zero value data. To ensure the analysis did not rely significantly on interpolated or imputed data, Vera dropped any county that was missing more than two years of data total, or more than one consecutive year of data across any of the variables, and linearly interpolated others.

To handle extreme outliers and issues with dividing by zero, which occurred when calculating jail population rate and unemployment rate inequality ratios, Vera applied a small pseudocount to both the numerators and denominators. The pseudocounts were selected to balance smoothing of outliers without obscuring meaningful disparities. Vera determined these by examining the scale of each variable and conducting sensitivity testing. For the jail population rate inequality ratio, Vera selected $\varepsilon = 5$, and selected $\varepsilon = 0.01$ for the unemployment rate inequality ratio. Vera inverted these two inequality ratios, as in the section on indices above, for consistency in interpretation and alignment of directionality of results.

Inequality clusters did not utilize DTW clustering and instead used the k-means clustering results to assign counties to groups within urbanicities. DTW within urbanicity groups was not possible due to the number of observations available. The analysis resulted in two distinct groups of counties within each urbanicity group; the first contained more racially homogenous (typically majority white) populations and experienced slow to no population growth over the time period examined, and the second had more racially diverse populations and experienced population growth over the time period examined.

Counties that were removed from this analysis due to missingness were assigned to one of these two clusters using an additional round of k-means clustering that aligned time-series features of total population, racial demographics, and inequality indicators of the unassigned counties with those of the existing cluster groups.

Incarceration clusters

Vera conducted each of the four steps outlined above using total incarceration rates across all counties, regardless of urbanicity. This resulted in two distinct incarceration cluster groups; one with decreasing incarceration and one with increasing incarceration.

Economic opportunity clusters

The variables that Vera input into the four-step analysis for economic opportunity were the rate of undergraduate and above education, labor force participation rate, median income, poverty rate, and unemployment rate. This analysis produced two cluster groups that displayed differing baselines and trends across the time period. The first had worse economic opportunity outcomes that improved and stabilized over the time period studied, and the second had better outcomes that improved and then declined over the period.

Economic stability clusters

The variables Vera used for the analysis for economic stability included the count of businesses per 100,000 residents, health insurance coverage rate, home ownership rate, median rent as a percentage of household income, and public assistance and food stamps utilization rate. Again, two distinct cluster groups emerged; one with gradually increasing economic stability measures and one with sharply declining economic stability measures over the period studied.

Combined cluster groups

Vera constructed the final groups that appear in the “similar counties” section of the Data Explorer by combining the cluster assignments described above. Following this protocol, Vera grouped each county with all other counties that were assigned to the same set of clusters across all four domains. This produced 16 county groupings ranging in size from four to 316 counties.

Public Dataset

The foundational dataset that Vera used to create the Data Explorer and associated analyses is publicly available on GitHub, with additional details regarding raw variables and data sources.

A detailed list of the data points presented in the Data Explorer, including data sources and formulas for calculated variables, is also available in the Data Dictionary.

Contact Information

For more information about the data or methodology described in this brief, please contact the IIP team at Vera at iip@vera.org or V Thompkins by mail at 34 35th Street, Suite 4-2A Brooklyn, NY 11232.

The Vera Institute of Justice is powered by hundreds of advocates, researchers, and policy experts working to transform the criminal legal and immigration systems until they’re fair for all. Founded in 1961 to advocate for alternatives to money bail in New York City, Vera is now a national organization that partners with impacted communities and government leaders for change. We develop just, antiracist solutions so that money doesn’t determine freedom; fewer people are in jails, prisons, and immigration detention; and everyone is treated with dignity. Vera’s headquarters is in Brooklyn, New York, with offices in Washington, DC, New Orleans, and Los Angeles. For more information, visit vera.org.