Variable Definitions

Commuting-Zone Data Portal for *Image(s)*

Hans-Joachim Voth David Yanagizawa-Drott

March 2026

This document provides precise definitions of the variables available in the commuting-zone data portal. All measures are constructed from style vectors extracted from U.S. high school yearbook portraits, 1930–2010. Full details are in the paper: Voth and Yanagizawa-Drott, "Image(s)," available at https://www.jvoth.com/images_july_24.pdf.

Each portrait is represented by a sparse binary style vector v_i of 25 style attributes (hair length, tie, collar type, jewelry, glasses, etc.), classified using convolutional neural networks trained on Google Vertex AI.

1 Horizontal Norm Deviation (Individualism)

Horizontal norm deviation measures how far a student's style differs from that of contemporaneous same-gender peers within the same high school. For individual i in school s, gender g, year y, we compute the average cosine similarity against all other individuals in the same cohort:

$$S_i = \frac{\sum_{i}^{N}\left(\frac{1}{|C|}\sum_{j\in C}\frac{v_i \times v_j}{\|v_i\|\,\|v_j\|}\right)}{N}, \tag{1}$$

where v_i is the style vector for individual i; C is the set of style vectors for all other individuals of the same gender in the same school and year; $|C|$ is the cardinality of C (excluding individual i); and N is the total size of the cohort.

Higher values of S_i indicate greater similarity to peers (more conformity); lower values indicate more individualism. The variable `normdev_horizontal` in the dataset reports the commuting-zone–year average of $1 - S_i$, so that *higher* values indicate greater deviation from local peer norms.

Gender-specific variants: `normdev_horizontal_m` (men) and `normdev_horizontal_f` (women).

2 Vertical Norm Deviation (Persistence)

Vertical norm deviation measures how different a cohort's style is from that of students in the same high school twenty years earlier. It is defined analogously to horizontal norm deviation, but the comparison set C consists of all individuals of the same gender from the same school observed 20 years prior:

$$S_i^{\text{vert}} = \frac{1}{|C^{-20}|}\sum_{j\in C^{-20}}\frac{v_i \times v_j}{\|v_i\|\,\|v_j\|},$$

where C^{-20} is the set of same-gender style vectors from the same school, 20 years before year y.

Higher values indicate greater intergenerational similarity (persistence). The variable `normdev_vertical` reports the commuting-zone–year average of $1 - S_i^{\text{vert}}$, so that *higher* values indicate greater intergenerational stylistic change.

Gender-specific variants: `normdev_vertical_m` (men) and `normdev_vertical_f` (women).

3 Style Novelty

Style novelty captures the prevalence of style combinations that were observed among fewer than 1% of all students up to and including a given year. It measures the arrival of genuinely new stylistic combinations.

First, we compute the cumulative count of each style s for gender g from the start of the sample (t_0) to year t:

$$C_{s,g,t}(t_0,\, t) = \sum_{y=t_0}^{t} T_{s,g,y}, \tag{2}$$

where $T_{s,g,y}$ is the number of students adopting style s, of gender g, in year y.

A style is classified as *novel* if its cumulative count falls in the bottom 1% of the distribution of all cumulative style counts:

$$I_{s,g,t} = \begin{cases} 1 & \text{if } C_{s,g,y}(t_0,\, t) \leq Q_1\big(\{C_{s',g'}(t_0,\, t) \mid s' \in \mathcal{S},\ g' \in \{m, f\}\}\big), \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathcal{S}$ is the set of all styles ever observed, and Q_1 denotes the cut-off value for the top 1% of the (expanding) cumulative style distribution, ranked from lowest to highest.

The variable `stylenov` reports the commuting-zone–year share of students whose style is classified as novel.

Gender-specific variants: `stylenov_m` (men) and `stylenov_f` (women).

4 Norm Deviation Raw Index

The raw index (`normdev_raw_index`) combines horizontal and vertical norm deviation into a single summary measure at the commuting-zone–year level.

Gender-specific variants: `normdev_raw_index_m` (men) and `normdev_raw_index_f` (women).

5 Aggregation to Commuting Zones

All individual-level measures are aggregated to the commuting-zone–year level by taking the mean across all individuals within each commuting zone and year. Commuting zones follow the classification in Autor and Dorn (2013).

Citation. Voth, Hans-Joachim, and David Yanagizawa-Drott. "Image(s)." July 2024. https://www.jvoth.com/images_july_24.pdf