Co-Residence

An analysis of household configuration vs wealth

Project Description

Which household configurations make people happiest?

- Target Variable: **Gini Index** selected as a proxy for happiness
- Descriptive Analysis

Major Themes:

- Methodological Goal: Avoid line-deletion as if forecasting
- Sparsity: 88% after merging core dataset with target variable
- High Dimensionality: 150+ features
- Collinearity: Compositional data, low-variance features
- Time Leakage: Walk-forward fitting of scalers and transformers

The CoResidence National Database*

157 countries, 65 years, 809 observations, 178 features, 146 core features

Core Features:

Household Configuration

encoded with H

- HH: Household Headship
- HS: Size and Age Comp
- HT: Household Typology
- HR: Relationship to Head

Demographic Features:

Development, Population

encoded with P or D

- P: Population
- D: Human Development Index

Other

- T: Time, C: Country/Continent

***Citation**: Esteve, A., Galeano, J., Turu, A., García-Román, J., Becca, F., Fang, H., Pohl, M. L. C., & Trias Prat, R. (2023). Zenodo. https://doi.org/10.5281/zenodo.8142652

Target Selection

Original target used: **HDI** (**Human Development Index)**

HDI consists of **D**-coded features in the dataset.

Dropping D-features, MI (Mutual Info Regression) yielded these top ten →

Gini Index replaced HDI to retain constituent features.

The analysis uses Gini Index.

MI Score II	Code	Description
0.7981	HS20	Average number of children in the household (aged < 18)
0.7766	HS22	Average number of 0-9 individuals in the household
0.7567	HH39	Average number of children in male-headed households
0.7416	HS23	Average number of 10-19 individuals in the household
0.7276	HS18	Average number of 0-4 children in the household
0.7208	HS15	Proportion of households with at least one person 0-4 years old
0.7097	HH37	Average household size of male-headed households
0.6733	HH04	Proportion of 2-persons households of male-headed households
0.6660	HS17	Average household size
0.6638	HS02	Proportion of 2-persons households

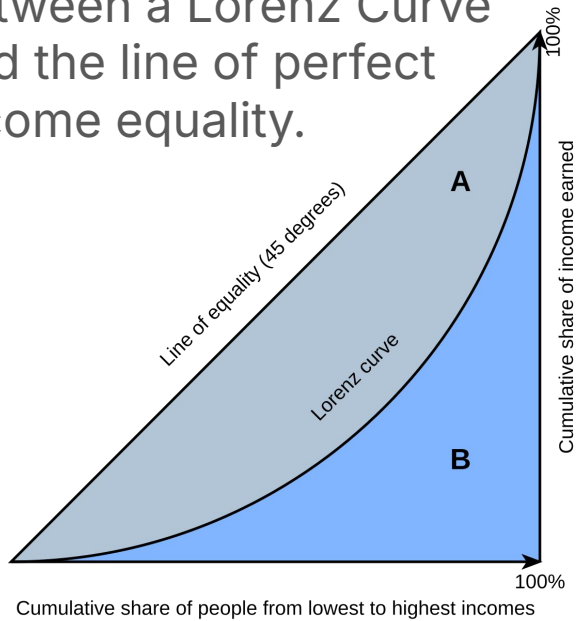
Mutual Information is the expected log departure of the joint distribution from the product of marginals.

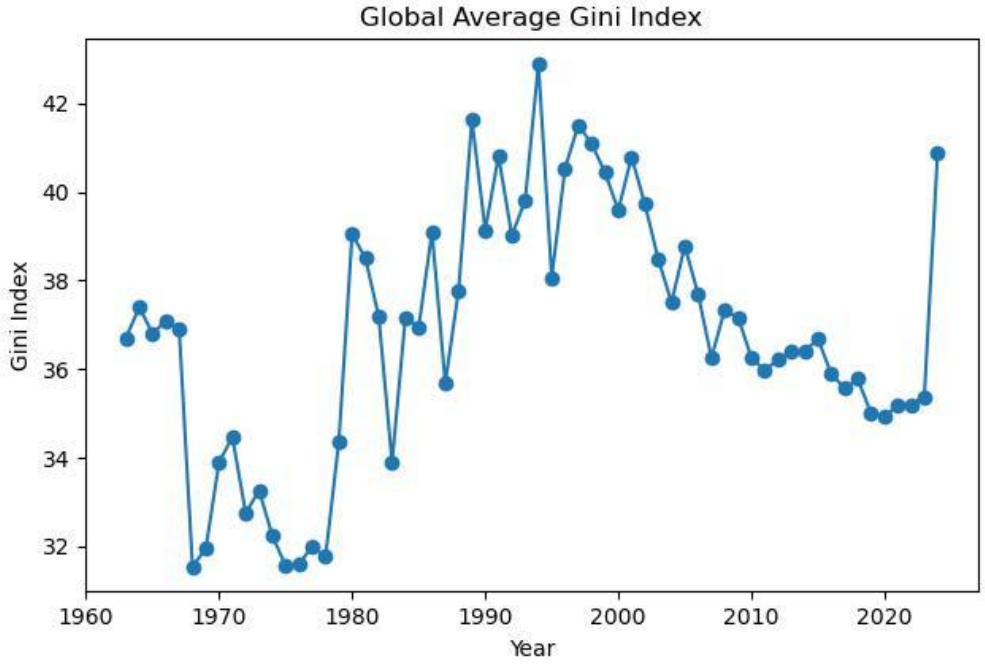
Target Variable

Gini Index

Gini Index - What Is It?

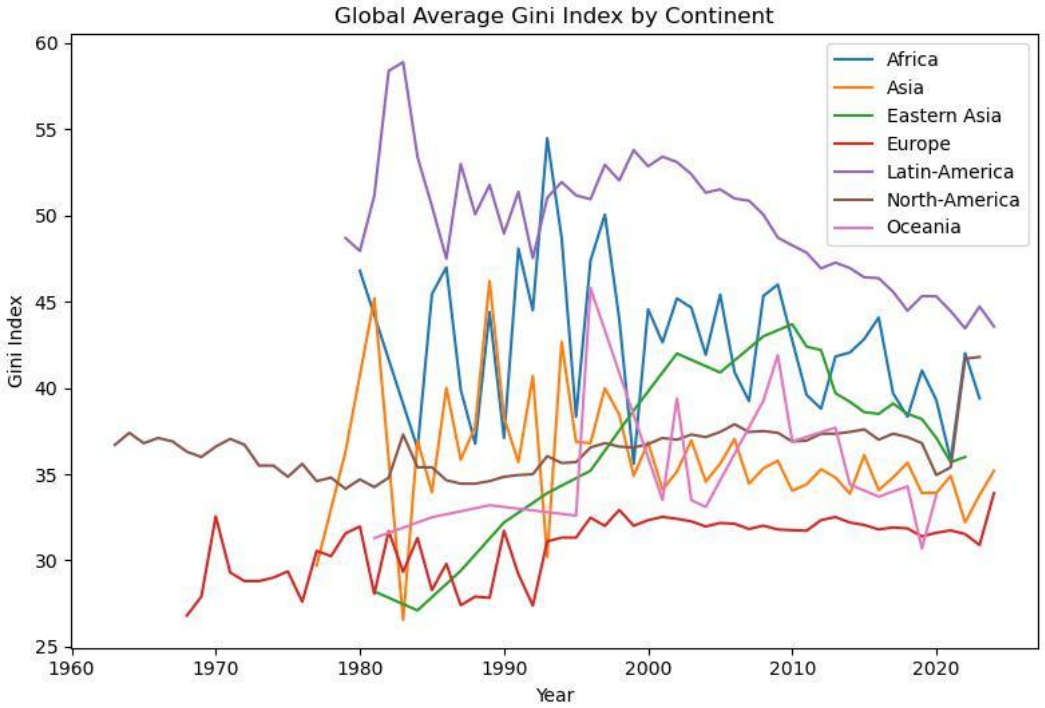
Gini Index is the area between a Lorenz Curve and the line of perfect income equality.





Data Source: World Bank - World Development Indicators - Last Updated 2025-12-19

Gini Index by Continent



Gini Index: Complications

Gini Index replaced HDI, to retain informative features which constitute the HDI.

Outer Merge conducted on Year, Country to keep all Gini values.

Extreme sparsity resulted, as Gini years not aligned with CoResidence years.

Many countries were deleted, due to insufficient data (details on next slide).

Missing Gini values imputed (no tscv) as attempted historical reproduction.

Models are evaluated exclusively on ground truth values, mitigating leakage.

Two target imputation approaches are tested (unsupervised): SVD and KNN

Country Deletion Criteria

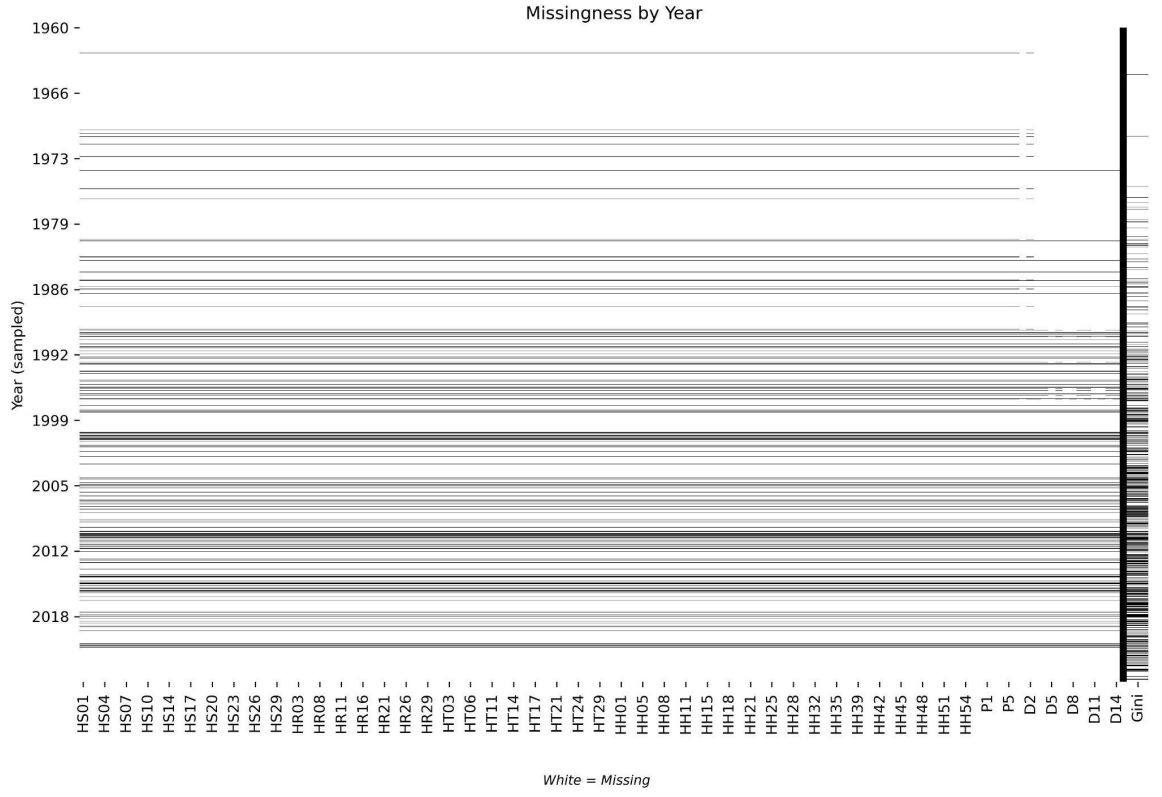
Post Gini-merge, **32 countries are deleted** due to insufficient data.
→ 125 out of 157 original countries retained

Country Retention Requirements:

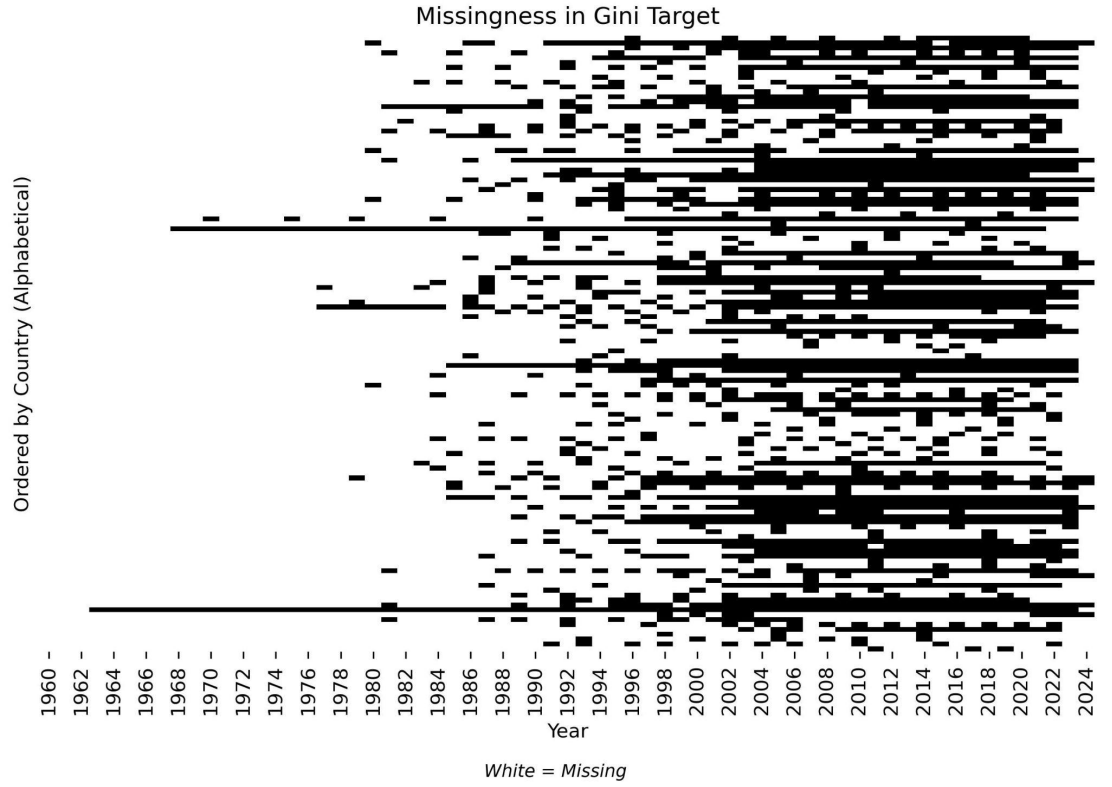
1. At least two ground-truth Gini Index values
2. Ground-truth Gini values least 8 years apart
3. At least two ground-truth CoResidence observations

Why? Trends require before-and-after, and time for change to unfold.

Overall Sparsity Post-Merge (post Country Deletion)



Historical Continuity: Filling Gaps in the Target



KNN Imputation

KNN Imputer fills missing values with the mean Gini of the k-nearest years. It is essentially a more sophisticated backfill strategy, which assumes that *nearby years predict missing years*.

Procedure:

1. Per country, randomly hide ~50% of values, and treat the rest as “training” points.
2. Try several KNN settings (k in {2,3,5,9,12}) using Year (min–max scaled) as the only predictor, fit KNN on the unhidden years, predict the hidden ones, to reveal the k with the lowest MAE.
3. Impute the full Gini series per country, using optimal k per country in KNNImputer.

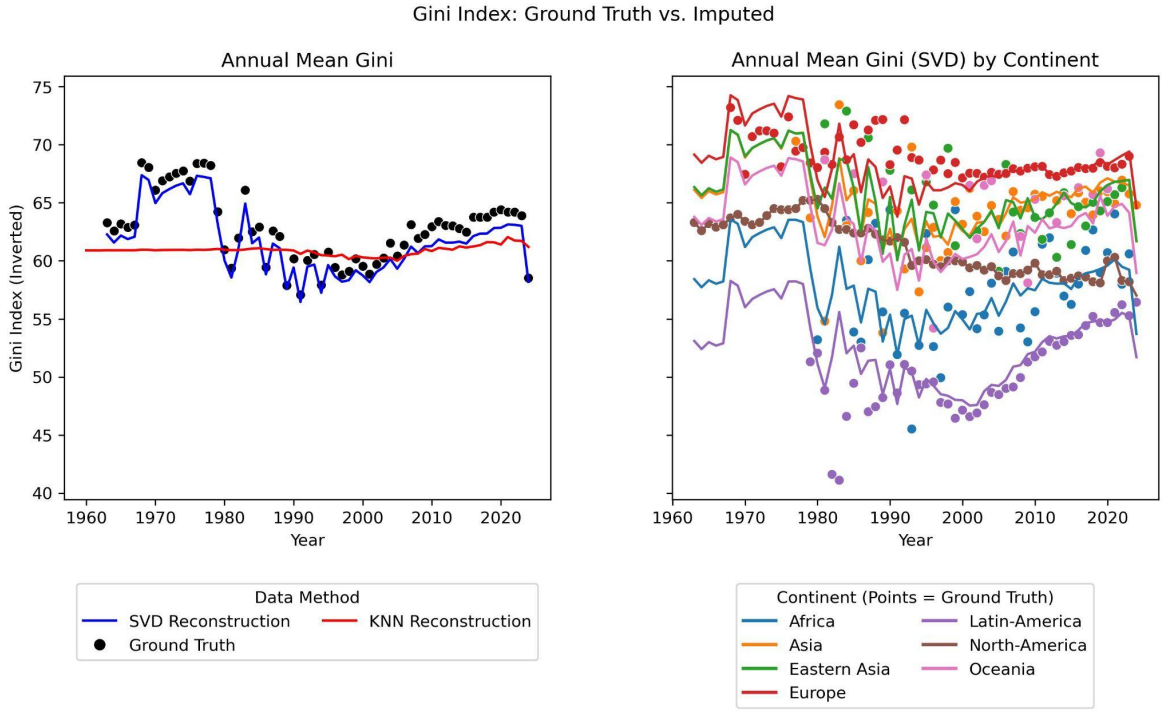
SVD Imputation (TWFE residuals-based initialization)

Missing target values are imputed as two-way fixed-effects predictions augmented by low-rank country–year interaction structure learned from observed *residuals*. This assumes *missing values have typical residuals at baseline*.

Procedure:

1. Pivot the target variable into a Country × Year matrix.
2. Estimate the TWFE baseline (global mean + country and year fixed effects) and create the residual matrix by subtracting this baseline.
3. Mean-fill missing entries in the residual matrix using available residual averages, with a zero fallback in residual space (fallback is equivalent to TWFE-only prediction when no residual structure is identifiable).
4. Apply TruncatedSVD to the residual matrix, reconstruct a low-rank approx capturing latent interaction structure and smoothing idiosyncratic variation.
5. Add the TWFE baseline back to the reconstructed residuals, and harvest reconstructed values only where ground truth values are missing (hybrid).

Visual Comparison: KNN vs. SVD Target Imputation

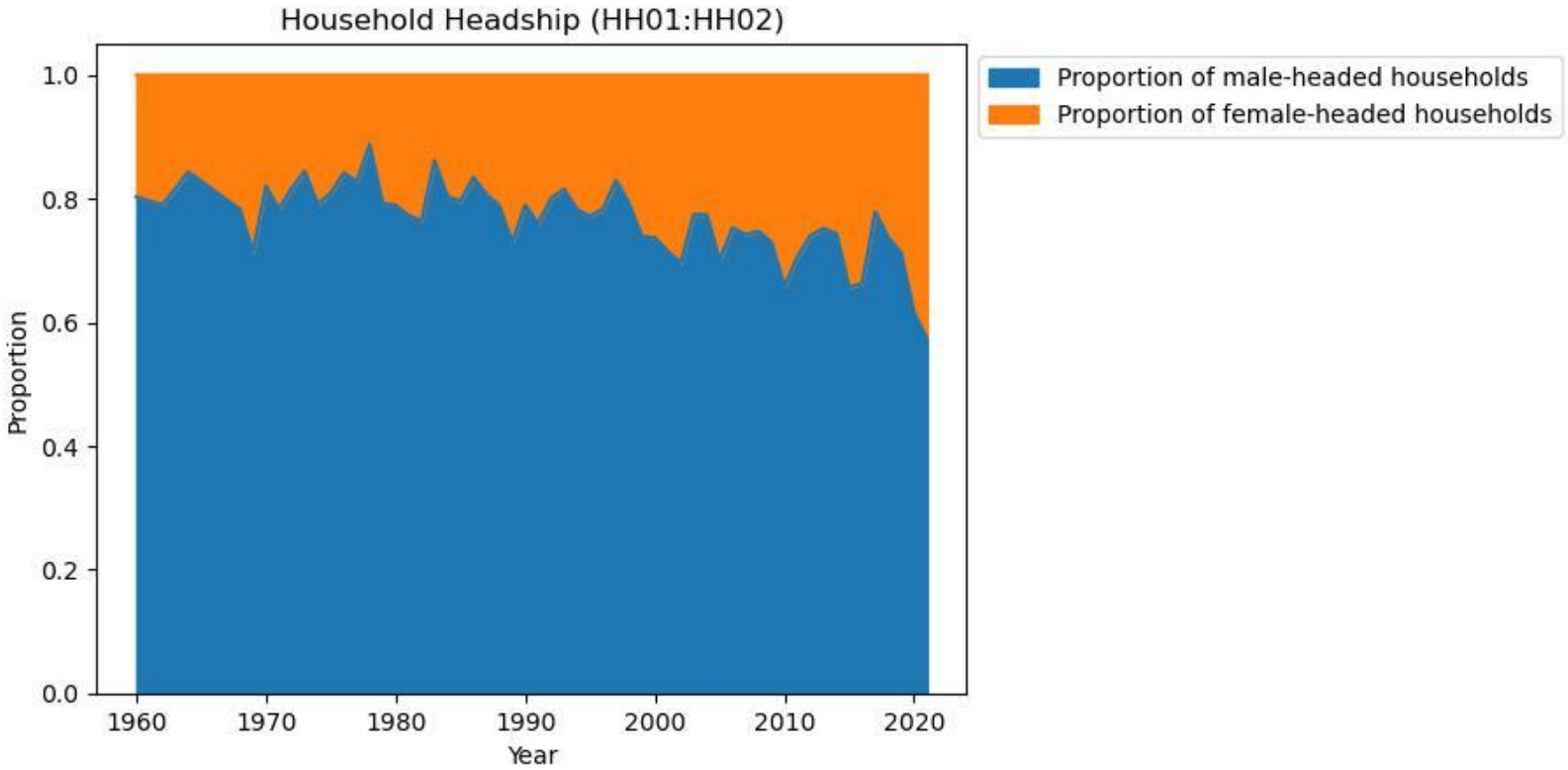


*Gini target in **inverted** such that 100 indicates total income equality, and 0 indicates total income inequality

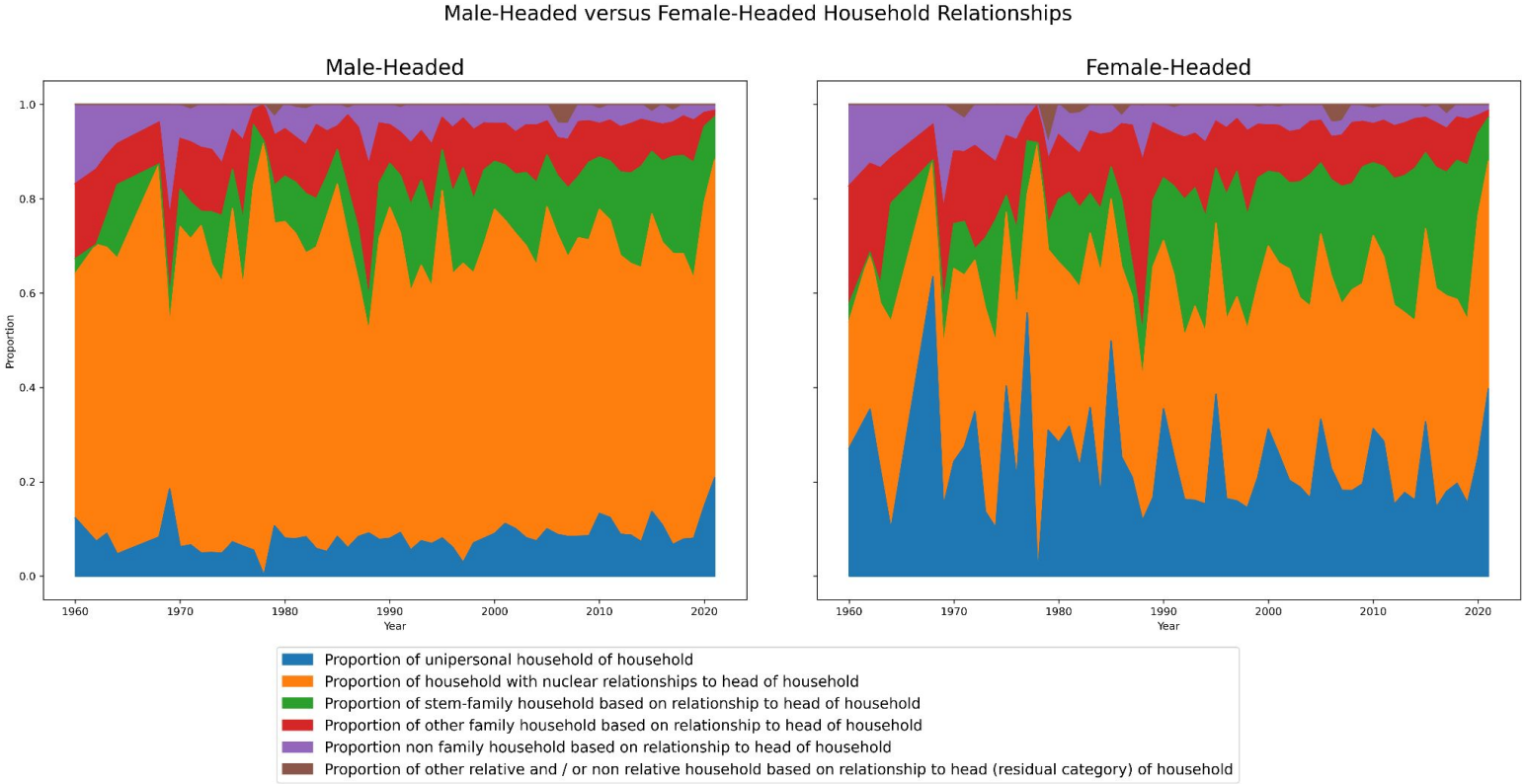
Exploratory Visuals

Notable Features

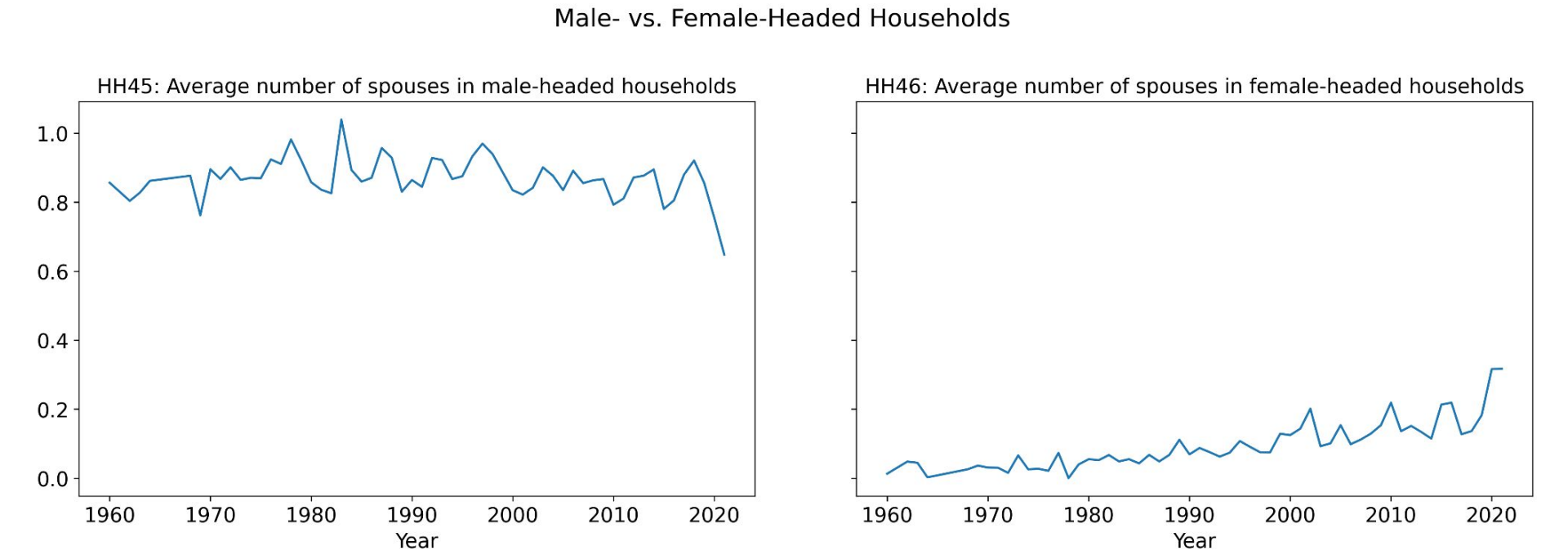
Increase in Female-Headed Households



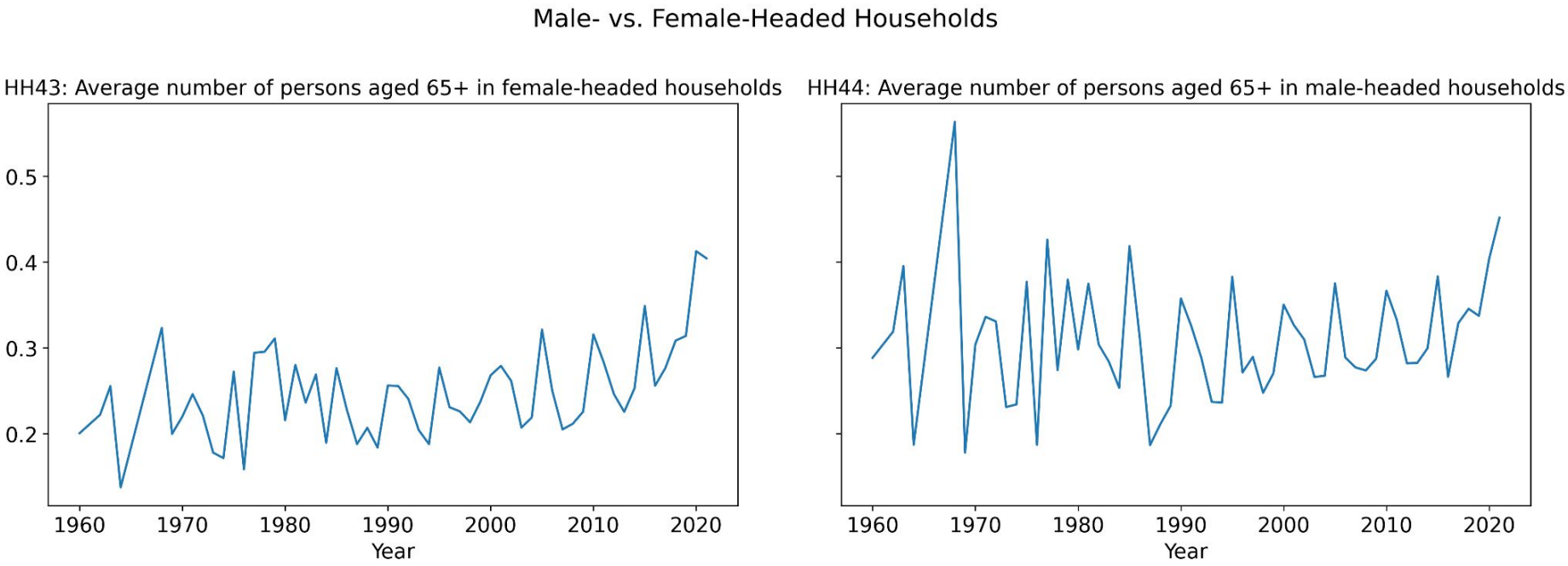
Increase in Unipersonal HHs; Nuclear Stays Constant



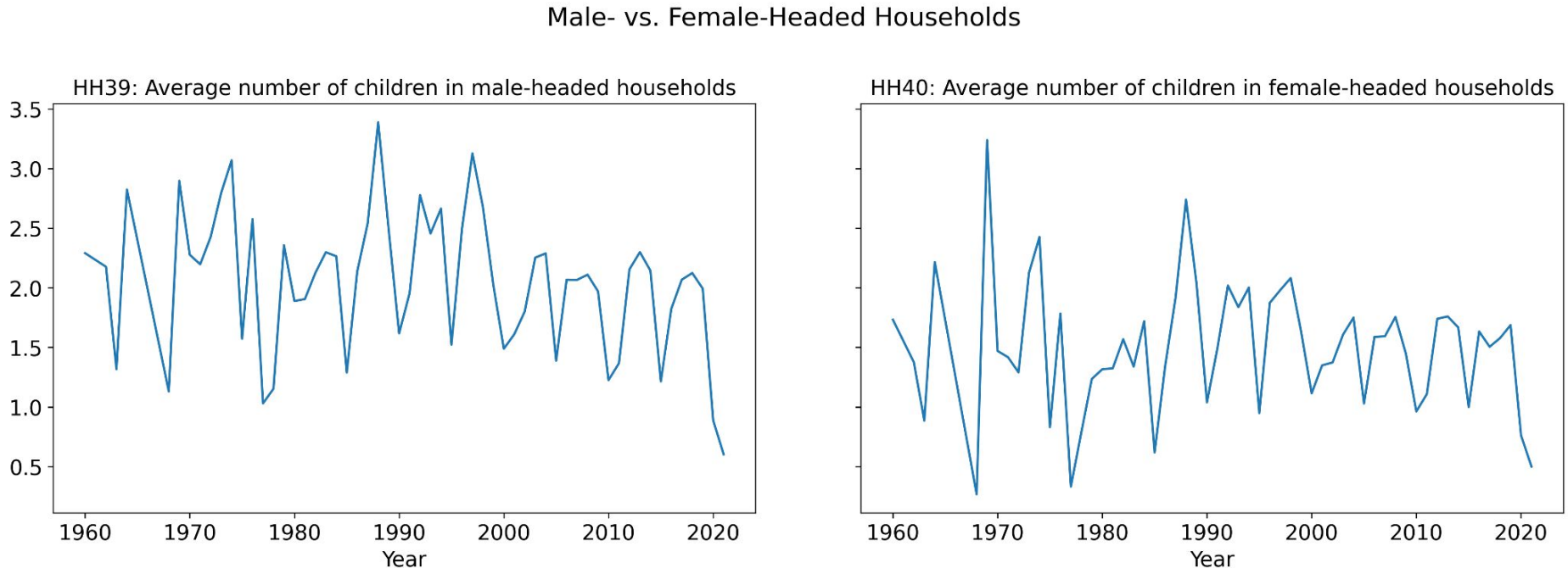
Male-Headed Fewer Spouses, Female- More Spouses



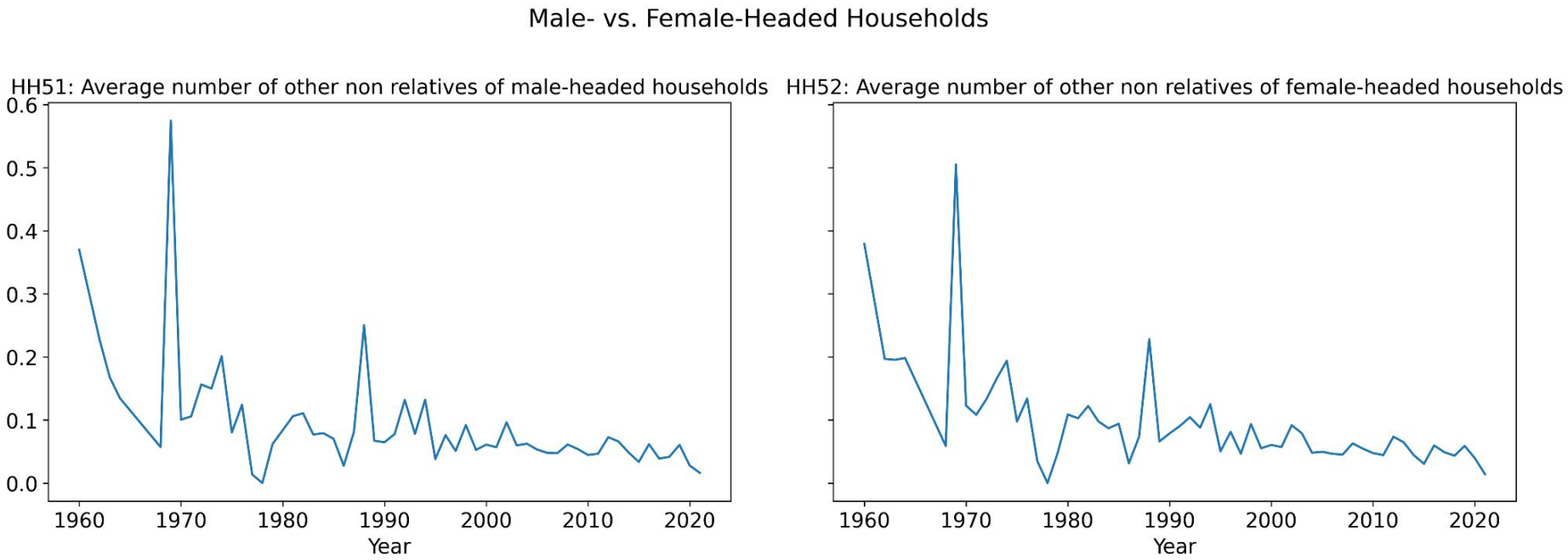
More Old People Everywhere Lately



Fewer Children Everywhere Lately



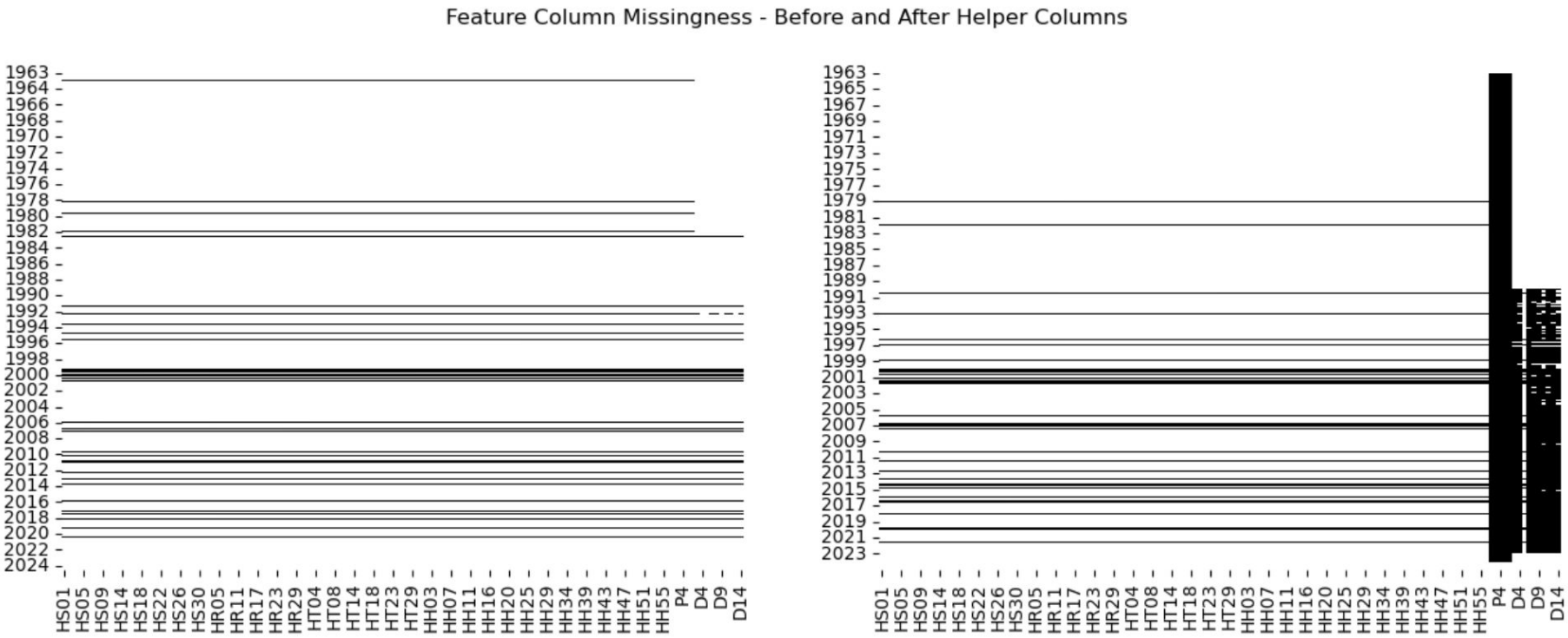
Non-Relatives Slight Decrease Lately



Feature Imputation

Three Main Strategies

Helper Columns Imported from United Nations Data



Some lines in early years appear to vanish but this is a result of plotting visualization idiosyncrasies.

Baseline Model vs. Imputation + Models

Baseline Model - **No Imputation** (XGBoost-only, which handles nulls)

1. Iterative Imputer (Bayesian Ridge) → All Models
2. Iterative Imputer fed to TruncatedSVD → All Models
3. TWFE initialization fed to TruncatedSVD → All Models

Models Used:

Linear: OLS, Lasso, Ridge

Non-Linear (or agnostic): Support Vector, Random Forest, KNN, XGB

Pros and Cons of Iterative Imputer Imputation

Pros

- Bayesian Ridge estimator assumes Gaussian priors, no need to initialize
- Predicts only missing cells. Does not transform ground truth values.
- Supports posterior sampling to propagate uncertainty in predictions

Cons

- Cannot handle rank deficiency. # Missing Values << # Helper Columns
- Too much sparsity → Convergence failures → Plaid approach needed
- Computationally slow and expensive, especially with sample_posterior
- Did not stabilize matrix under extreme sparsity

Iterative Imputer - “Plaid” Approach

Procedure:

1. Sort values by Year, Country to prepare for time-series split
2. Split into train set (earlier years), test set (later years)*
3. Fit scaler to train set only, apply globally
4. “Plaid” Approach: Create horizontal groupings (column groupings) to pair with helper columns, i.e. averages groups, proportion groups
5. Impute missing values blockwise (Bayesian Ridge), sample posterior**
6. Unscale imputed values

* Ideally this would be done per country, but too much sparsity in earlier years precluded this possibility.
**Blockwise imputation was necessary due to rank insufficiency, hence “Plaid” approach

Pros and Cons of SVD Latent Imputation

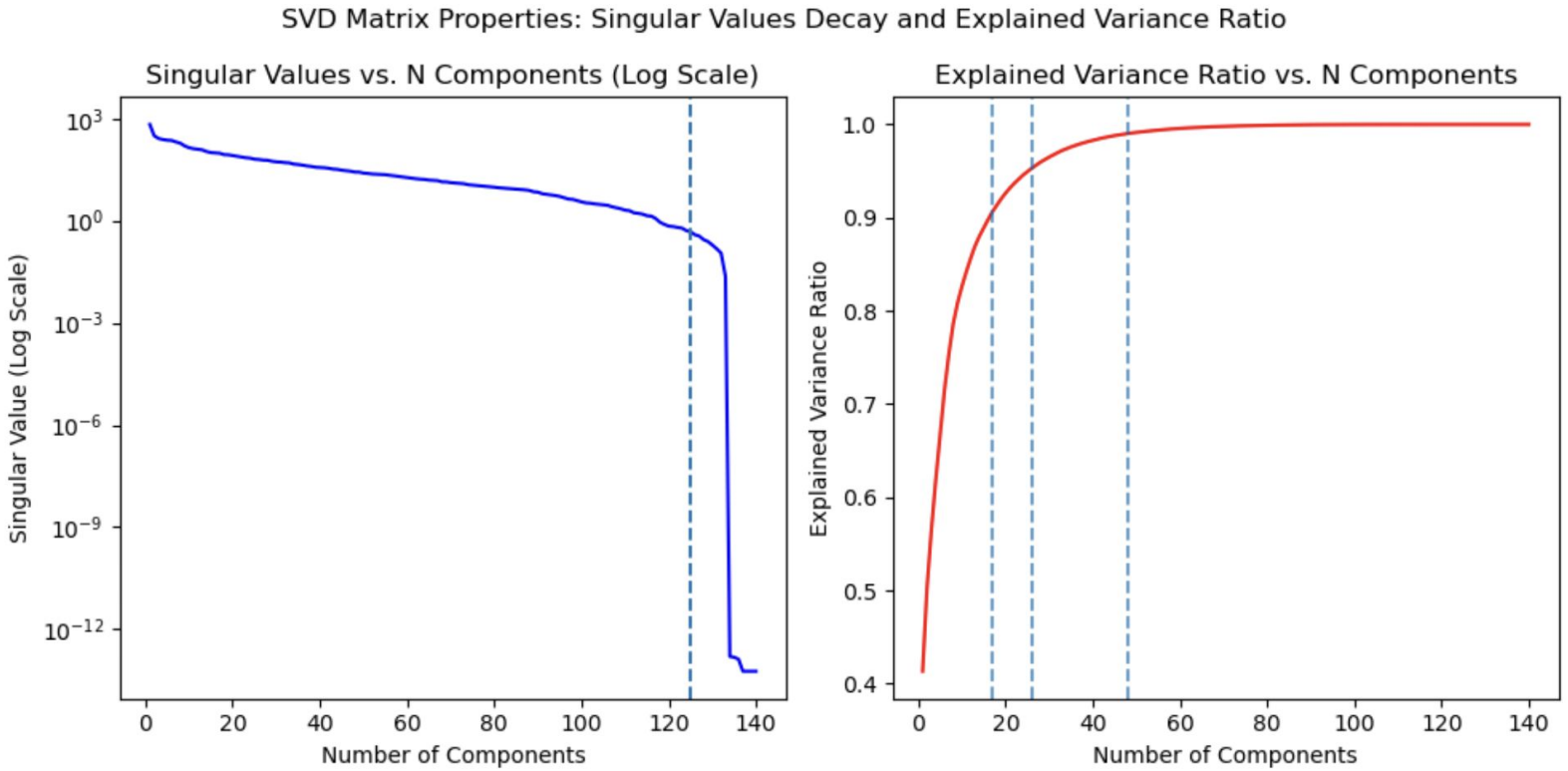
Pros

- Orthogonal Bases: Stable matrices, stable predictions, stable coefficients
- Iterative Imputer assumes conditional independence given predictors — whereas TWFE+SVD explicitly models global panel structure
- Dimensionality Reduction: Signal distilled and grafted to initialized values
- Latent Info: Singular Values represent directions of greatest energy

Cons

- Prior assumptions required to initialize missing values (TWFE baseline, etc.)
- Interpretability: Models ingest latent components, not native features
- Back-projection of beta vectors, allocation of permuted importances required
- No way to translate tree-model importances back to native features
- Permuted Importances lack sign, lack geometry

Truncated SVD - n_components



Implicit Latent Feature Imputation via **Truncated SVD**

Procedure:

1. Initialize missing values in feature matrix X with TWFE mean-fill*
2. Sort the filled X by panel ordering (Country, Year)
3. Create Cumulative Explained Variance plot to determine ideal k
   (run Truncated SVD with full range of possible k values)
4. Apply Standard Scaler via walk-forward strategy, fitting on train set only
5. Fit TruncatedSVD via walk-forward strategy, fitting on train set only
   (using optimal k values for defined thresholds of explained variance)
6. Feed latent representation to models, score on ground truth test set**

* Ideally initialization strategy is walk-forward but my data was too sparse in early years.
** X is not reconstructed but can be for forecasting purposes.

Model Comparison

Walk with Me: My Modeling Loop (TSCV)

Preprocessing - Time Series Split (n=3):

- Standard Scaler fitted on train sets only

Feature Engineering:

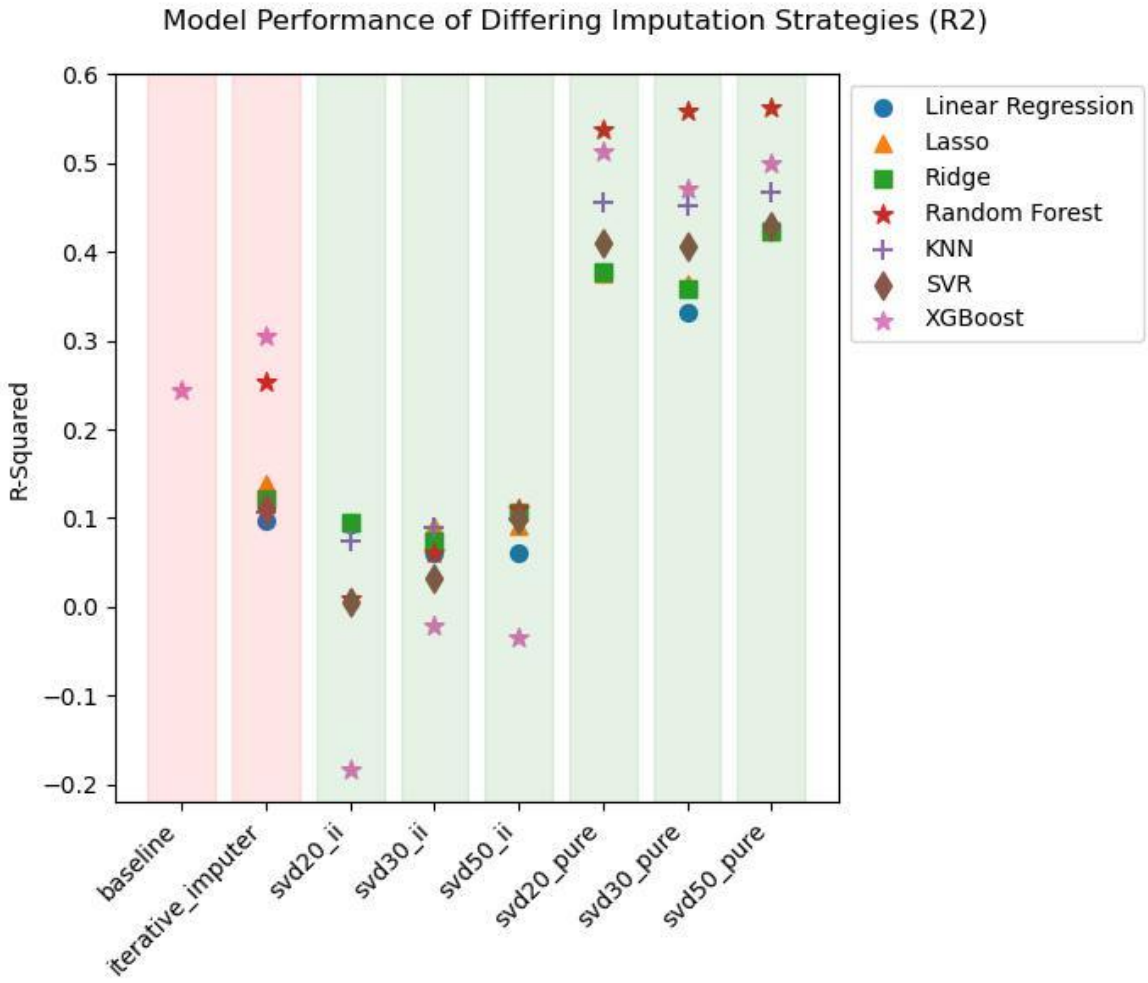
- Truncated SVD is fitted on train sets only

Modelling and Scoring:

- Hyperparameters tuned using MSE, models scored using R-Squared
- Models are tested on ground-truth rows only

Feature Importances

- Permutation Importances in loop (for cross-model and tree evaluation)
- Model-specific importances also retained for handling or validation


Model Performance of Differing Imputation Strategies (R2)
R-Squared
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
baseline
iterative_imputer
svd20_ii
svd30_ii
svd50_ii
svd20_pure
svd30_pure
svd50_pure
Linear Regression
Lasso
Ridge
Random Forest
KNN
SVR
XGBoost

Imputation x Model — Comparing Performance

Matrix instability reflects how sensitive a dataset is to small perturbations, typically arising from extreme sparsity and collinearity. For linear and other matrix-based models, unstable design matrices yield fragile predictions. While tree-based models do not explicitly invert matrices, they are still affected through degraded feature geometry, noisy splits.

Baseline: XGBoost-only due to nulls. Poor performance. Unstable matrix.

Iterative Imputer: Better but poor performance. Unstable matrix.

SVD (with Iterative Imputer initiation): Better but still poor performance.

SVD (with TWFE initiation): Remarkable improvement in performance.

Feature Importances

Which household configurations make people happiest?

Handling Interpretability Limitations

Best performing models fed **SVD latent components**, not native features!

Feature Importance Methods:

1) **Linear Model Beta** vectors back-projected to native feature space
2) **Permutation Importances** conducted on all models, for viable cross-comparison. PI allocated using row-normalized squared loadings
3) **Tree FIs** retained for validation but interpretability was lost

I place most faith in betas, because they are signed, and because back-projection assumptions were stronger than allocation assumptions.

Back-Projected Betas

As regression coefficients are signed linear weights, *we use vector space geometry to perform a basis transformation to feature space*.

1) The regression is trained on X expressed in the vector space basis ($Z=XV=U\Sigma$)
2) Linear weights in beta vector ($\hat{y}=Z\beta$) exist in the vector basis.
3) Columns of V are the orthonormal vectors defining the vector basis of feature space; rows of V define the feature basis.
4) Multiplying β by columns of V expresses the same linear functional in the original feature basis.
5) Each feature coefficient is a weighted combination of component coefficients, with weights given by the entries of V.
6) Per native feature, sum across the columns of V after weighting them by β to derive feature space coefficients.

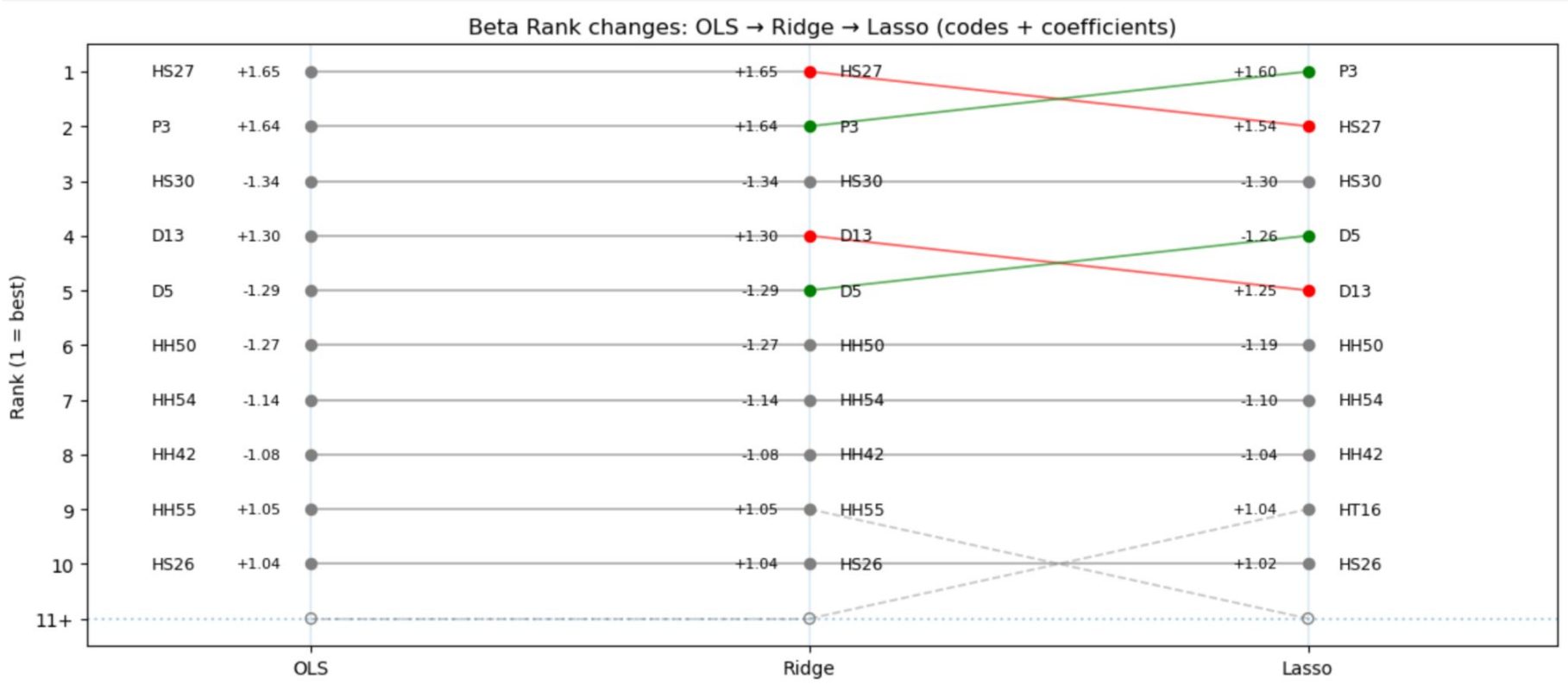
Permutation Importance Allocated Loadings Attribution

As *predictive importance is a variance-like magnitude*, we attribute PI to native features using variance geometry as the coordinate system.

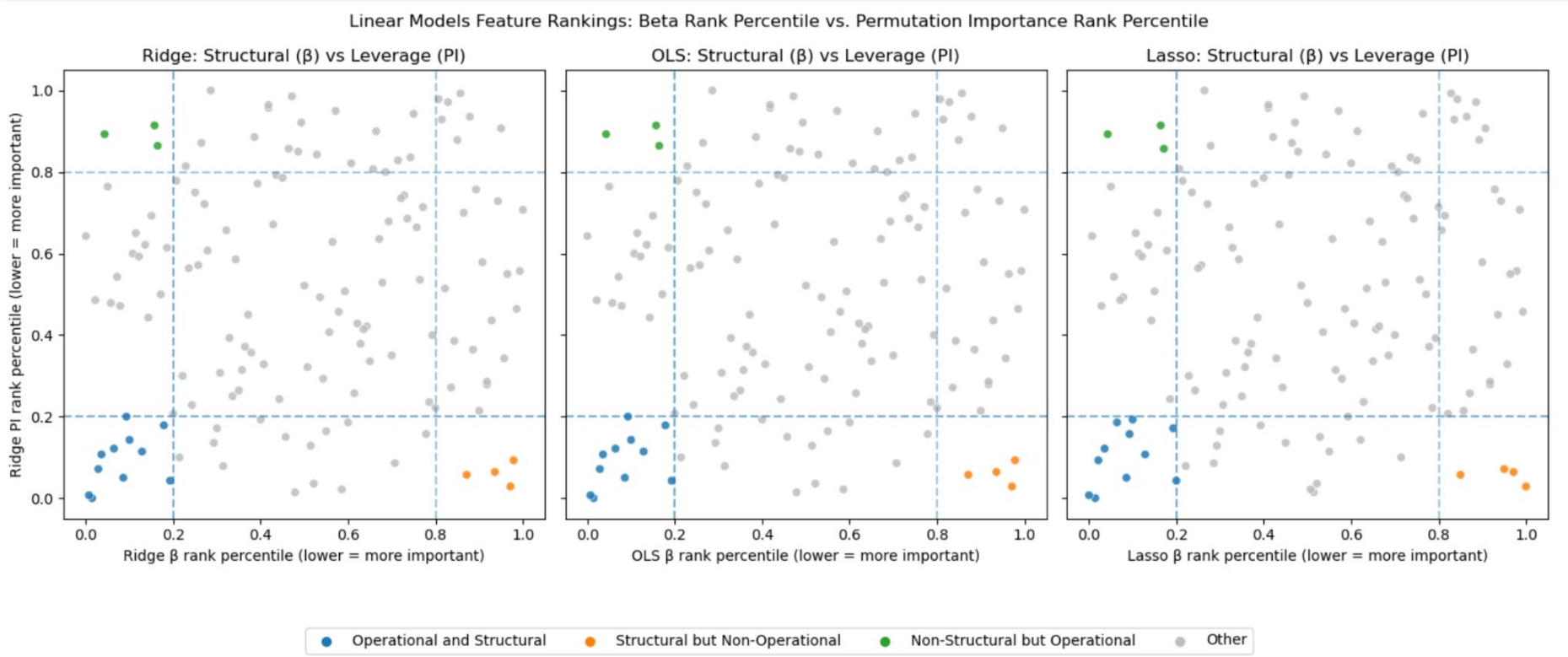
1) Singular vectors (latent components, i.e. columns of V) are orthonormal, thus variance decomposes additively across latent components.
2) Permutation importance assigns predictive weight PI to each orthonormal component.
3) Within each component, squared loadings partition that component's variance across native features.
4) Feature importance is obtained by distributing each component's permutation importance according to squared loadings:

$$\mathrm{FI}_i = \sum_j \mathrm{PI}_j \, v_{ij}^2$$

Back-Projected Beta Rankings: Ridge vs OLS vs Lasso



Quadrant Analysis: Beta Rankings vs. PI Rankings



Quadrant Analysis: Beta vs. PI

Structural but Non-Operational (Top β rank, Bottom PI rank)	Operational and Structural (Top β rank, Top PI rank)	Non-Structural but Operational (Bottom β rank, Top PI rank)
D12 – Expected years of schooling male ▲	P3 – Fertility rate ▲	HH54 – Average number of males adults in female-headed households ▼
HT29 – Average size of other family households based on relationship to head ▲	HS30 – Average number of 80+ individuals in the household ▼	HT06 – Proportion of households with non nuclear family relationships to head ▼
HS16 – Proportion of households with at least one person 65+ years old ▼	D5 – Gross Domestic income ▼	HH09 – Proportion of 7-persons households of male-headed households ▲
HS23 – Average number of 10-19 individuals in the household ▼	HH50 – Average number of other relatives of female-headed households ▼	
	HS26 – Average number of 40-49 individuals in the household ▲	
	HS12 – Proportion of 2-3 persons households ▲	
	HH38 – Average household size of female-headed households ▼	
	HH49 – Average number of other relatives of male-headed households ▲	
	HS14 – Proportion of 6+ persons households ▼	
	HH40 – Average number of children in female-headed households ▼	
	HH47 – Average number of childs in male-headed households ▲	

Ordered by OLS/Ridge beta percentiles (Top (low) → Bottom (high); ▲ = Positive Coef, ▼ = Negative Coef

Top Features Quadrant Analysis

Gender Double Standards: Coefficients often flip along gender lines for male-headed vs female-headed hhs

- Number of children in male-headed hh is positive, in female-headed hh is negative (HH47 vs HH40)
- Large hh bad if female-headed, good if male-headed (HH38 vs HH09)

Surprising Demographic Reversals

- Fertility Rate has a positive coefficient (historically viewed as negative to have too-high fertility rate)
- Gross National Income has a negative coefficient (historically viewed as positive development metric)

Aging Is Bad

- Proportion of households with at least one person 65+ years old, strongly negative (HS16)
- Average number of 80+ individuals in the household strongly negative (HS30)

Family Better than Randos

- Number of “other relatives” of female-headed household strongly negative (HH50)
- Average size of “other family” households based on relationship to head is positive (HT29)

Medium Better than Large

- Proportion of 2-3 person households is positive, 6+ person households is negative (HS12 vs HS14)

Quadrant Analysis Takeaway: *Replication Is Good*

Number One Feature in Quadrant Analysis: Fertility Rate

Fertility Rate is both operational (high PI ranking)

and structural (high Beta ranking).

Other “helper columns” are not dominant in top rankings,

suggesting this is a real signal, not due to imputation bias.

Household Headship: Male vs Female

Sign Agreement (hue) x Coefficient Spread (saturation)

OLS/Ridge beta (back-projected coefficients)

Male	Female	
HH03 - -0.369	HH14 - +0.442	- proportion of 1-person households of households
HH04 - +0.437	HH15 - +0.578	- proportion of 2-persons households of households
HH05 - +0.393	HH16 - +0.397	- proportion of 3-persons households of households
HH06 - +0.437	HH17 - +0.245	- proportion of 4-persons households of households
HH07 - +0.176	HH18 - -0.151	- proportion of 5-persons households of households
HH08 - -0.152	HH19 - -0.526	- proportion of 6-persons households of households
HH09 - +0.653	HH20 - -0.175	- proportion of 7-persons households of households
HH10 - +0.341	HH21 - -0.105	- proportion of 8-persons households of households
HH11 - +0.346	HH22 - -0.610	- proportion of 9-persons households of households
HH12 - +0.306	HH23 - -0.232	- proportion of 10-persons households of households
HH25 - -0.333	HH31 - +0.178	- proportion of unipersonal households of households
HH26 - +0.302	HH32 - +0.284	- proportion of households with nuclear relationships to head of households
HH27 - +0.156	HH33 - -0.123	- proportion of stem-family households based on relationship to head of households
HH29 - -0.180	HH35 - -0.163	- proportion non family households based on relationship to head of households
HH37 - +0.586	HH38 - -0.866	- average household size of households
HH39 - +0.455	HH40 - -0.630	- average number of children in households
HH41 - +0.764	HH42 - -1.081	- average number of adults in households
HH43 - -0.198	HH44 - -0.222	- average number of persons aged 65+ in households
HH45 - -0.551	HH46 - +0.600	- average number of spouses in households
HH47 - +0.618	HH48 - -0.333	- average number of childs in households
HH49 - +0.850	HH50 - -1.266	- average number of other relatives of households
HH51 - -0.109	HH52 - -0.624	- average number of other non relatives of households
HH53 - +0.356	HH54 - -1.139	- average number of males adults in households

Agreement (green) vs Disagreement (red)
Intensity = |βF − βM|

Double Standards Summary

Living Alone - Good for women; Bad for men

Large Households - Bad for women; Good for men

Avg Number of Spouses - Good for women; Bad for men

Adults in HH - Bad for women; Good for men

Nuclear Family - Good for everyone

Number of Children - Bad for women; Good for men

Non-Family - Bad for everyone

Aging - Bad for everyone

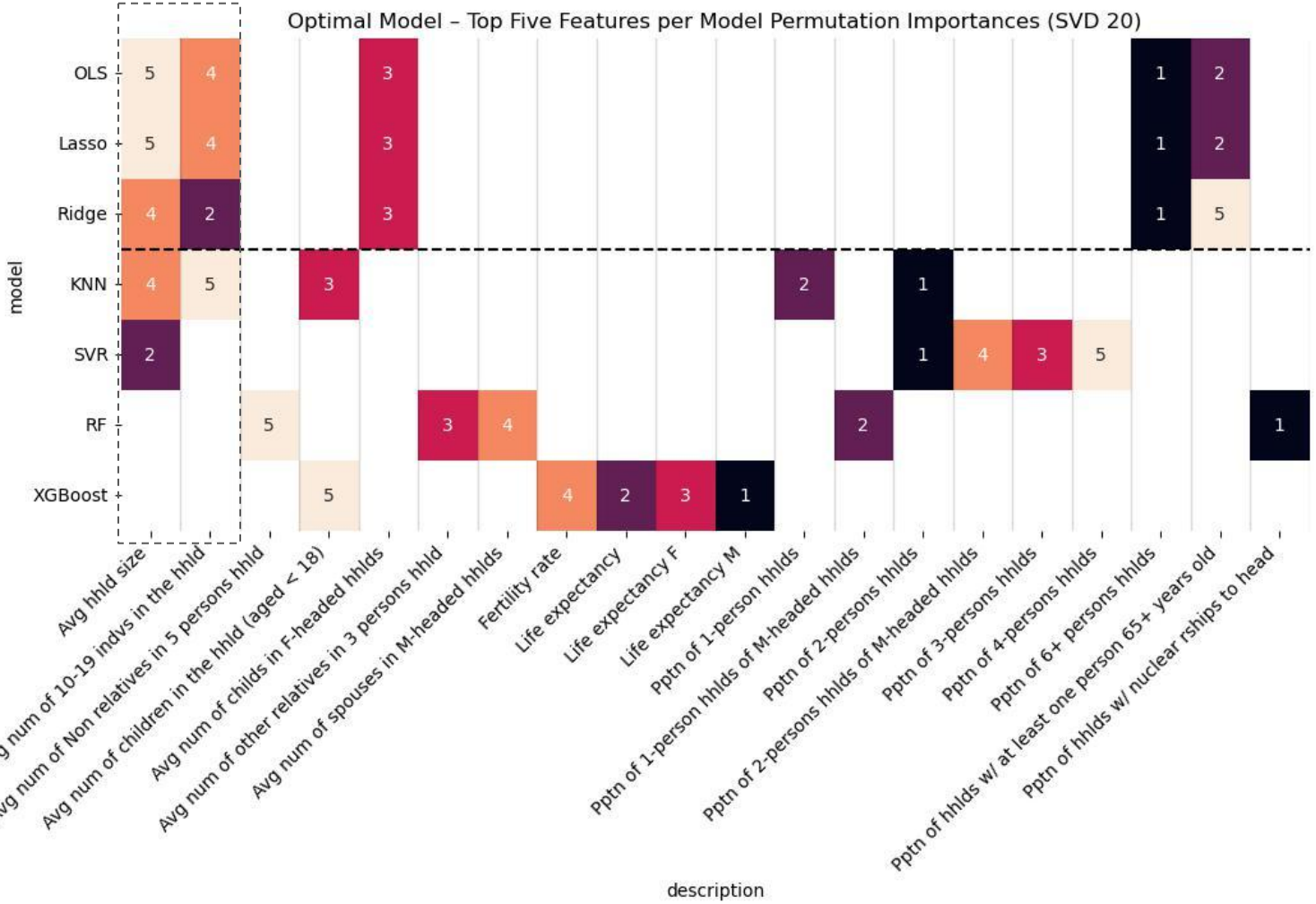
Decompression Patterns - Linear vs Non-Linear PI

Top Five Features change, depending on level of compression used.

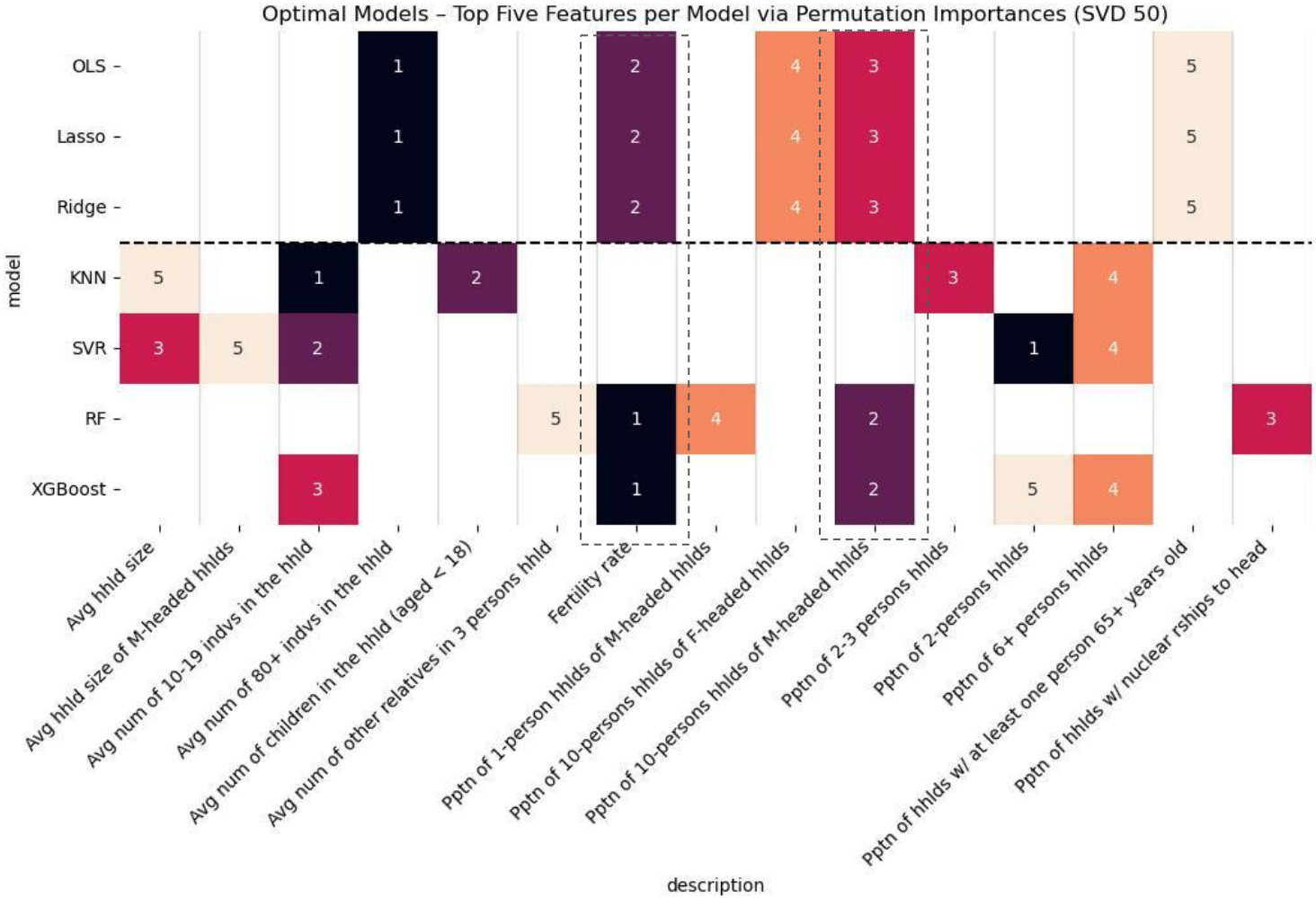
SVD with n_components=20 (suboptimal) vs. 50 (optimal)

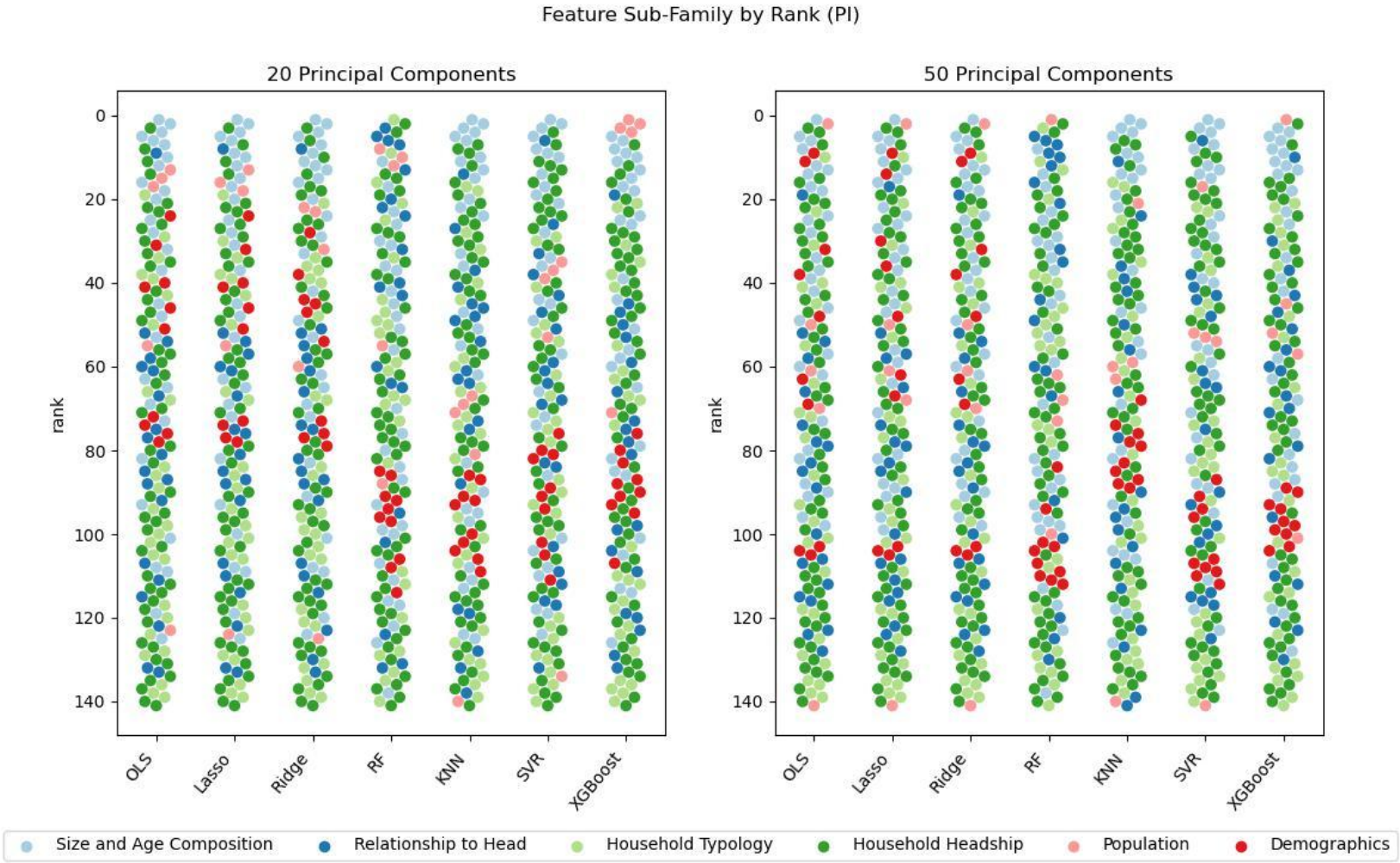
- **Non-Linear** models pick up Fertility Rate over Average Household Size under optimal decompression. With too much compression, these demographic features are subsumed under other global patterns.
- Linear vs. Non-Linear models highlight almost completely disjoint sets

Linear and Non-Linear Mostly Disjoint



Decompression Brings Fertility Rate to the Fore




Feature Sub-Family by Rank (PI)
20 Principal Components
50 Principal Components
rank
0
20
40
60
80
100
120
140
OLS
Lasso
Ridge
RF
KNN
SVR
XGBoost
Size and Age Composition
Relationship to Head
Household Typology
Household Headship
Population
Demographics

Overall Conclusions

Surprising

- Compression unexpectedly subsumes fertility under size and age
- Fertility Rate increase is good, Gross National Income increase bad!
- Linear and non-linear signal almost totally disjoint at the top

Unsurprising

- Most gendered features have opposing signs
- Medium households better than enormous households
- Family arrangements better than non-family

Remaining Questions

- Should I really be including every single feature, or prioritize?
- Did the SVD transformer bias towards linear structure? Is that bad?
- SVD has a random_state arg for multiple solutions. Does this affect FI?
- Did the twfe mean-fill initialization of SVD introduce bias?
- Did scaling the Central Log Ratio-transformed columns dilute the geometry of those columns?
- Can we trust any forecasts produced through the reconstructed X, when features are heavily engineered?
- Should I disaggregate data by female and male-headed households, since these seem to flip coefficients?

Next Steps

1) Improve initialization quality for SVD
    a) Play around with SoftImpute
    b) Bring in more helper columns for imputation
2) Manually reduce dimensionality with thematic groupings:
    a) Old People, Young People, Middle-Aged People
    b) Nuclear Family vs. Other, Family vs. Non-Family
    c) Huge Hh vs. Medium Hh vs. Tiny Hh
3) Reconstruct X in native feature space (using SVD bases) and visualize diff or forecast future based on salient features
4) Create weighted-Gini target to account for problem of equally-sliced-small-pie (where happiness proxy falls apart)

Links

GitHub Repo:
https://github.com/LittleBiggler/CoResidence

Slide Deck (read only):
https://docs.google.com/presentation/d/1nK782rSTxYEe0AHu3H8IU2pXppTRSnpFgtZipvn3quo/edit?usp=sharing