Detecting Community-Level Inflow Migration Patterns in Chicago Using Census Tract-level Inflow Data

Zixuan Zhou

MACS 30100

Motivation

- **Main Question**: Can we detect community-level migration patterns in Chicago using census data, and can socio-economic variables predict these patterns?
    - **Why it matters**: where demographic shifts impact economic and social dynamics
- **Existing Literature**
    i. **Greenlee (2019)** examines longitudinal household-level data to connect origin/destination mobility flows with income and neighborhood sociodemographic change trajectories

Motivation - Continued

- **Existing Literature**

  ii. **DeLuca (2018)** explores neighborhood change and residential mobility in Chicago by analyzing how neighborhood changes influence household-level decisions to move and how these dynamics shape residential patterns over time

  iii. **Yoon (2023)** provides a model to detect socio-economic and racial changes in Chicago, using heterogeneous **graph networks** to analyze community evolution, including racial displacement, poverty, and unemployment trends in Chicago neighborhoods

- **Motivation**

  - Computational methods are less employed in this field, but they can generate meaningful results for large-scale geographic-socioeconomic data
  - Detecting tract-level in-migration patterns and using socio-economic factors to predict such factors

Design

- **Two-Step Modeling Approach**:
    i. **Clustering:** using only variables **related** to inflow migration from 2018
    Goal: Identify distinct community migration patterns
    ii. **Classification:** train on variables **unrelated** to inflow migration from 2018; predict pattern using data from 2023
- **Goal:** Predict the identified migration patterns using other neighborhood features, and compare changes across time

Design - Continued

- **Why Clustering as a first step?**
  - Our data does not have well-defined labels, and there are too many variables related to inflow migration
  - We want **fewer dimensions for better interpretability**; we also hope to detect natural patterns - understanding the interplay between variables that characterize different migration patterns
  - Unsupervised methods (clustering and PCA) naturally discover patterns in migration without pre-defined labels

Data

Data Collection

Census data is often messy and requires additional steps for data collection.

- **Data Source**:
  - **American Community Survey (ACS):** 2018 & 2023, accessed via Census API, with 2018 data for comparison
  - **Transportation:** information on whether a region has railways. Accessed via Chicago Transit Authority (CTA)
  - **Geographic Shapes**: Chicago census tract level .shp files. Accessed via the Census Bureau
  - **Complimentary census data:** census data from the Urban Displacement Project

Data Collection - Continued

- **Selected Variables for Processing:**
  - Tract identifiers (FIPS code, year, county ID, state ID etc.)
  - Socioeconomic indicators: median home value, median rent, population, median income, education levels, **inflow migration by origin and income levels** etc.
  - **Changes** in housing price, income levels, and low-income households since last survey period
  - Transit data
- Example for example variables from ACS (see below)

```
# Variables to fetch from ACS 2023
variables_2023 = ['B01003_001E', # total population
                  'B02001_002E', # white population
                  'B11001_001E', # total households
                  'B19013_001E', # median household income in 2023 inflation adjusted dollars
                  'B25077_001E', # median home value in 2023
                  'B25064_001E', # median gross rent in 2023
                  'B25003_002E', # owner occupied housing units
                  'B25003_003E', # renter occupied housing units
                  'B15003_022E', # total number of people with bachelor's degree over age 25
                  'B15003_023E', # total number of people with master's degree over age 25
```

Data Collection - Continued

- **Data Size**:
  - Raw feature space:
    - ACS: ~ 3.5 million on a yearly basis, ~3600 variables for both years (See the ACS codebook 2023 for reference)
  - After Variable Selection:
    - ACS: 3815 samples (tract-year pairs from both 2018 and 2023) with 84 variables
    - Transit data: 444 samples, converted to a binary indicator if a tract has direct access to rails
    - Geographic shape files are not preprocessed

- **Data Examples**:

pop	white	hh	medhinc	mhval	mrent	ohu	rhu	total_bd
3726.0	1475.0	2190.0	57404.958678	245123.966942	1034.710744	689.0	1501.0	775.0
7588.0	1883.0	3038.0	41023.966942	197603.305785	1101.652893	823.0	2215.0	684.0
2609.0	987.0	1130.0	45552.892562	191652.892562	1067.768595	270.0	860.0	656.0
6311.0	3558.0	3185.0	54438.842975	252066.115702	1038.842975	859.0	2326.0	1585.0
4282.0	3349.0	2058.0	40509.917355	189586.776860	1031.404959	513.0	1545.0	999.0
3519.0	2298.0	2379.0	39876.033058	162479.338843	971.900826	341.0	2038.0	1157.0
3329.0	2171.0	1637.0	50716.528926	233223.140496	1083.471074	469.0	1168.0	770.0
2844.0	1806.0	1477.0	18119.008264	388842.975207	927.272727	177.0	1300.0	350.0
6708.0	3804.0	3066.0	59438.016529	210330.578512	1054.545455	1094.0	1972.0	1843.0
3573.0	1431.0	1652.0	61328.925620	285454.545455	1105.785124	820.0	832.0	1047.0

Preprocessing and Feature Engineering

- **Preprocessing**:
  - Cleaning: Removed negative values and duplicates; imputed for missing values; removed outliers
  - Adjusted all 2023 price levels to 2018 price levels
  - Performed log transformation on price-related variables to address skewness
- **Feature Engineering**
  - Computed for proportion/ratio variables
  - Computed for variables that represent changes since last survey period, such as percentage change in rental prices
- **Final output:** 36 variables on inflow migration, and 46 others (socioeconomic indicators, tract or year identifiers)

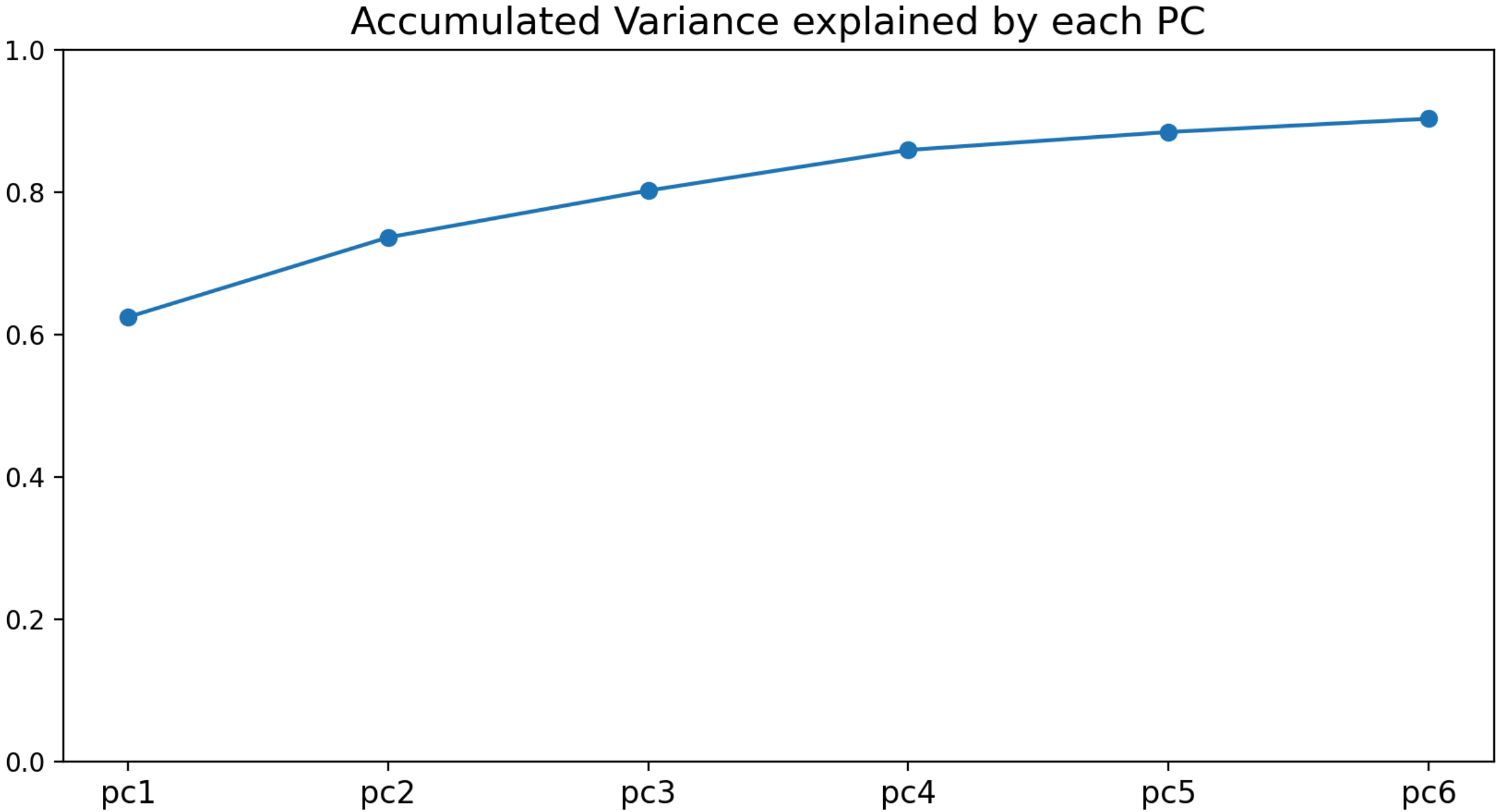
- More data examples

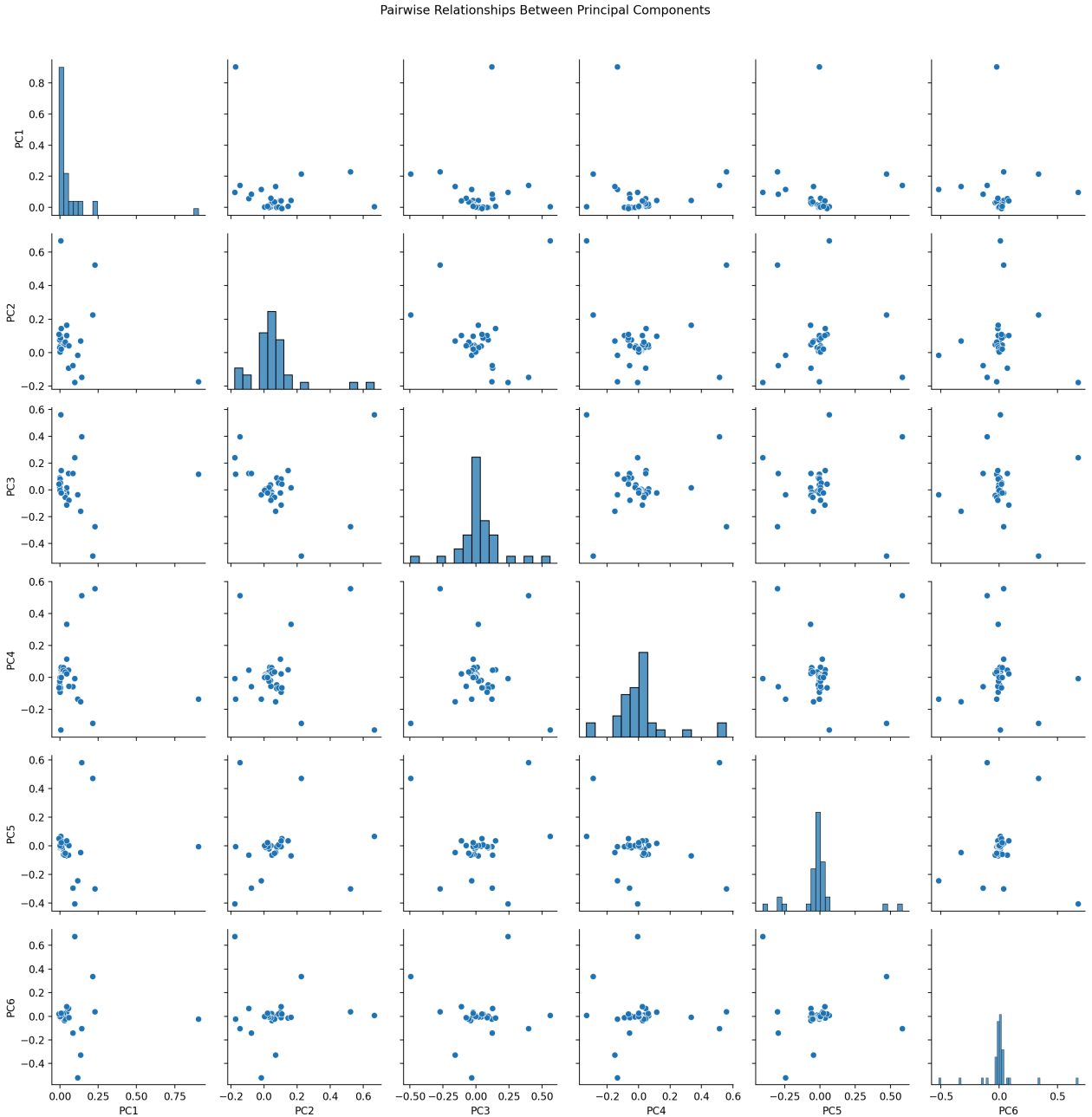
pop	hh	state	county	tract	FIPS	year	rail	total_li	pctch_real_mhval_sl	...	prop_hhinc_50000	prop_hhinc_60000	prop_hhinc_75000
2344.0	866.0	17	197	882900	17197882900	2018	0.0	259.0	0.010385	...	0.056582	0.050808	0.117783
3759.0	1183.0	17	197	883000	17197883000	2018	0.0	438.0	0.127786	...	0.060017	0.138631	0.090448
3726.0	1099.0	17	197	880111	17197880111	2018	0.0	104.0	0.090727	...	0.058235	0.035487	0.123749
5692.0	2429.0	17	197	883602	17197883602	2018	0.0	310.0	-0.173610	...	0.036641	0.114039	0.224784
2756.0	1439.0	17	197	883803	17197883803	2018	0.0	652.0	-0.005649	...	0.061154	0.079917	0.123697
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3943.0	1056.0	17	31	301701	17031301701	2018	0.0	559.0	0.097179	...	0.012311	0.091856	0.081439
4387.0	1089.0	17	31	301702	17031301702	2018	0.0	470.0	-0.066508	...	0.078972	0.089991	0.067034
3625.0	1052.0	17	31	301801	17031301801	2018	1.0	537.0	0.077217	...	0.000000	0.074144	0.074144
3509.0	866.0	17	31	301802	17031301802	2018	1.0	406.0	-0.082327	...	0.038106	0.101617	0.056582
5011.0	1348.0	17	31	301803	17031301803	2018	0.0	709.0	-0.010799	...	0.065282	0.085312	0.058605

Model Training

PCA

- **Goal:**
  - better understand the aspects our features represent
  - Reduce dimension for clustering to avoid the curse of dimensionality
- **PCA Parameters**:
  - PCA with explained variance threshold - 90%
  - Reduced dataset dimensionality from 36 features to 6 components
- Graph showing successful separation - no apparently linearity between the PCs


Accumulated Variance explained by each PC
1.0
0.8
0.6
0.4
0.2
0.0
pc1
pc2
pc3
pc4
pc5
pc6


Pairwise Relationships Between Principal Components
PC1
PC2
PC3
PC4
PC5
PC6

	PC1	PC2	PC3	PC4	PC5	PC6
prop_mov_wc_w_income	0.903923	-0.172464	0.118740	-0.135051	-0.005370	-0.023184
prop_mov_oc_w_income	0.005229	0.667485	0.561838	-0.330161	0.066116	0.007863
prop_mov_os_w_income	0.228162	0.523279	-0.273327	0.557253	-0.301847	0.037936
prop_mov_fa_w_income	0.045128	0.098268	-0.022165	0.114979	0.016221	0.036817
prop_mov_wc_9000	0.142291	-0.146214	0.397895	0.514351	0.581726	-0.103258
prop_mov_oc_9000	0.008033	0.145461	0.146365	0.047583	0.035899	-0.013989
prop_mov_os_9000	0.044949	0.164152	0.015443	0.333178	-0.068845	-0.008672
prop_mov_fa_9000	0.009087	0.034389	0.008472	0.063473	0.002233	0.005268
prop_mov_wc_15000	0.056557	-0.092046	0.123358	0.045354	-0.065644	0.067392
prop_mov_oc_15000	0.001477	0.039618	0.038224	-0.018105	-0.004419	-0.005018
prop_mov_os_15000	0.016691	0.030196	-0.008136	0.034404	-0.020767	0.009885
prop_mov_fa_15000	0.005760	0.004102	-0.001041	0.009536	-0.002315	0.001498
prop_mov_wc_25000	0.095748	-0.176465	0.240927	-0.006253	-0.406844	0.676304
prop_mov_oc_25000	-0.000370	0.077336	0.091228	-0.047078	-0.013214	0.001502
prop_mov_os_25000	0.021903	0.045608	-0.016590	0.061275	-0.059448	0.023696
prop_mov_fa_25000	0.006124	0.007660	0.006151	0.017769	0.002420	-0.001192
prop_mov_wc_35000	0.085236	-0.077249	0.122688	-0.058289	-0.296050	-0.140246
prop_mov_oc_35000	-0.000582	0.075983	0.089687	-0.067980	-0.005405	0.001481
prop_mov_os_35000	0.023597	0.044863	-0.031791	0.044010	-0.062502	-0.018693
prop_mov_fa_35000	0.005052	0.009051	-0.001809	0.011256	-0.011012	0.001898
prop_mov_wc_50000	0.115776	-0.015720	-0.035311	-0.135886	-0.243930	-0.520469
prop_mov_oc_50000	0.000736	0.102204	0.082115	-0.091103	-0.006417	-0.010799
prop_mov_os_50000	0.030818	0.047226	-0.041394	0.025137	-0.065565	-0.034403

PCA - Continued

- **Interpreting the PCs**
  - **PC1: LOCAL MIGRATION**
    - Extremely high loading on overall within-county migration; moderate loading on within-county high-income migration
  - **PC2: EXTERNAL MIGRATION**
    - Strongly driven by out-of-state and out-of-county migration and out-of-state migration
  - **PC3: REMOTE LOW-INCOME MIGRATION VS. LOCAL HIGH-INCOME MIGRATION**
    - Positive loadings on low-income migration ($9K, 0.40)
    - Strong negative loading on $76K+ within-county migration (-0.49)

PCA - Continued

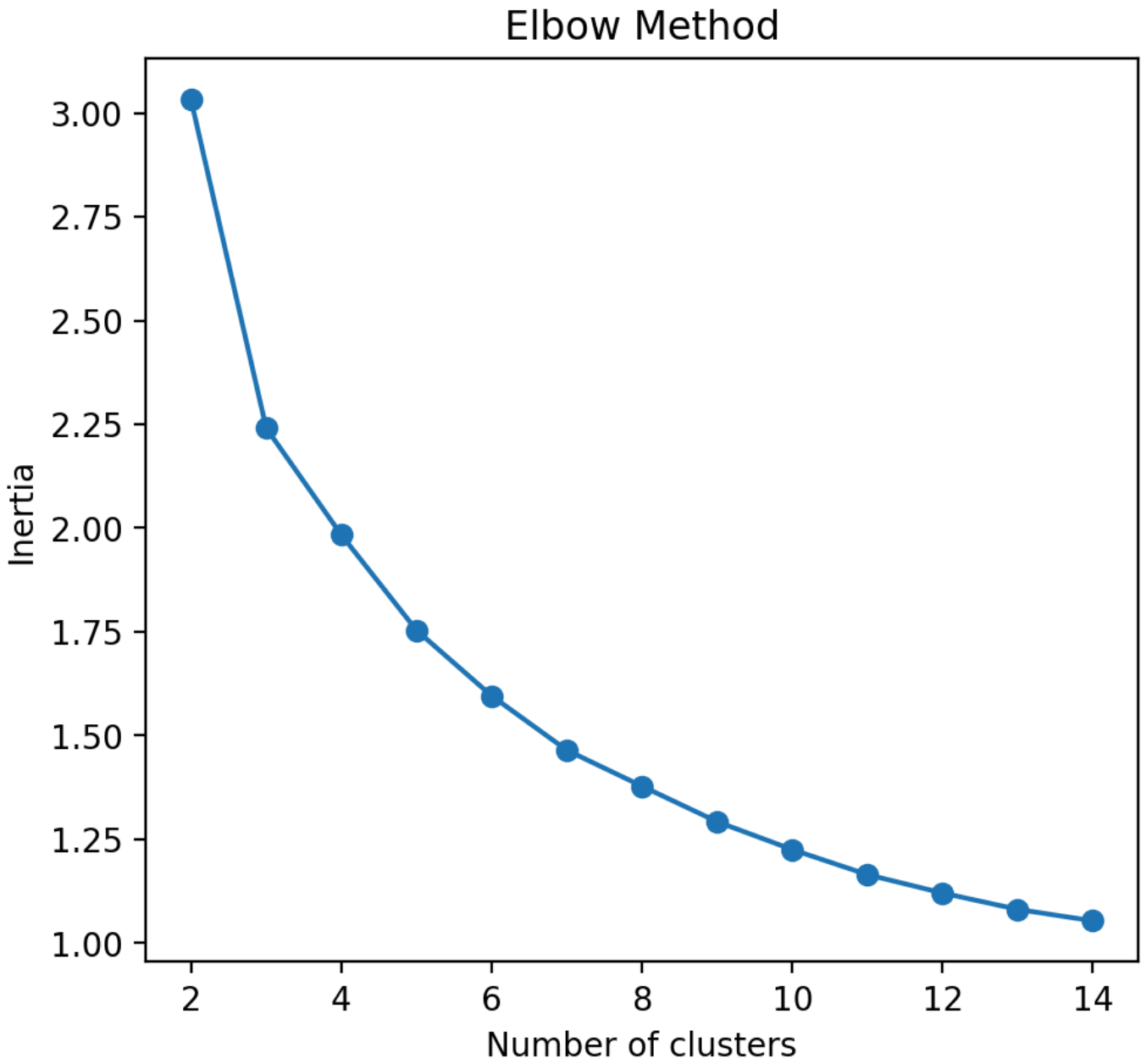
- **Interpreting the PCs**
  - **PC4: LOW-INCOME MIGRATION DESTINATION**
    - Strong positive loadings on $9K within-county (0.51) and out-of-state migration (0.56); Negative loadings on high-income migration categories
    - Areas with significant low-income migration from both local and distant sources
  - **PC5: INCOME EXTREMES**
    - High positive loading on $9K within-county (0.58) and $76K+ within-county (0.47)
    - Strong negative loadings on middle-income migration
    - Captures areas with both very low and very high income in-migration
  - **PC6: LOCAL LOW-INCOME MIGRATION VS. LOCAL MODERATE-INCOME MIGRATION**
    - Very high positive loading on $25K within-county (0.68); strong negative loading on $50K within-county (-0.52)

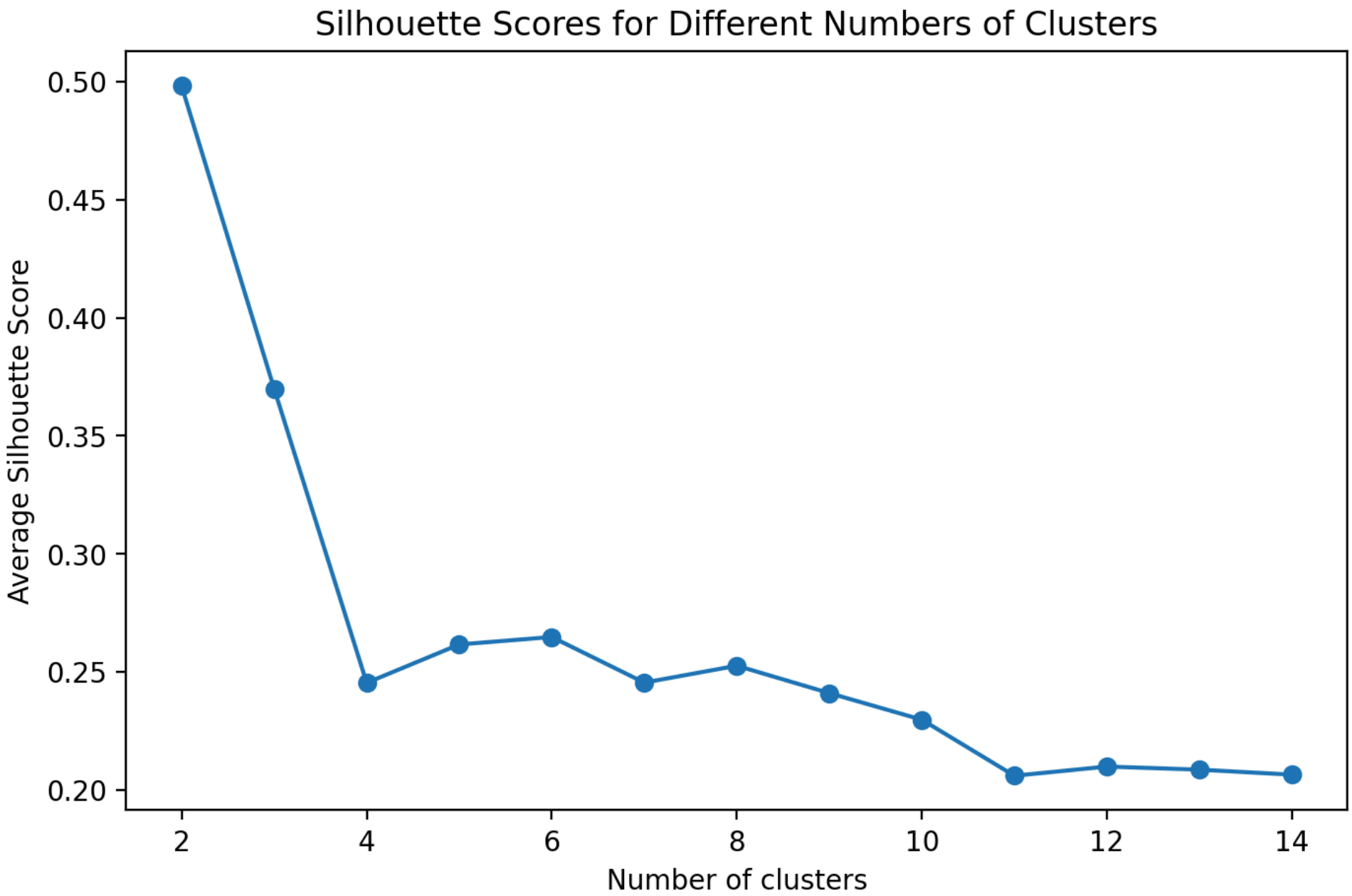
PCA - Continued

- **Downsides:** although reducing dimension to 6 achieves good balance between dimension complexity and interpretability, some of the PCs appear to measuring overlapping attributes, and are hard to interpret (for instance, PC3 is essentially a ratio)

Clustering

- **Selecting optimum number of clusters**
  - Experimented two models, K-means and KMedoids using the reduced-dimension data
  - Cluster numbers are determined using the Elbow Method and evaluated using Silhouette Score
  - Difference random seeds produce similar results
  - Optimum K:
    - K-means performs better in terms of silhouette scores
    - No sharp elbow in graph, but after K=5, an additional cluster produce less additional value
    - To balance interpretability and cluster quality, we may choose between K = 3 or K = 5; descriptive statistics on the mean of each PCs within cluster shows that it's easier to interpret across-cluster difference when K = 5

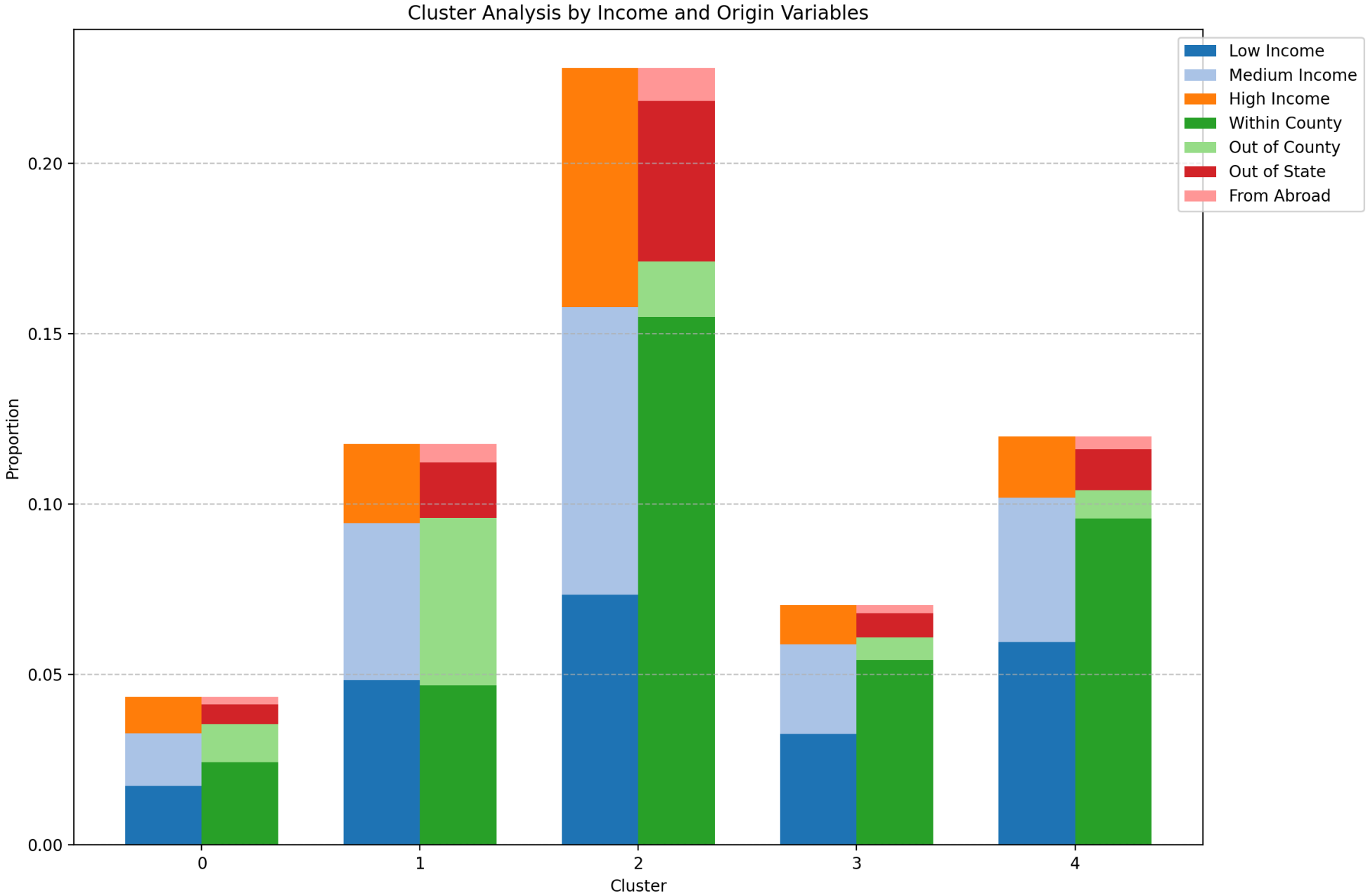

Elbow Method
Inertia
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
2
4
6
8
10
12
14
Number of clusters

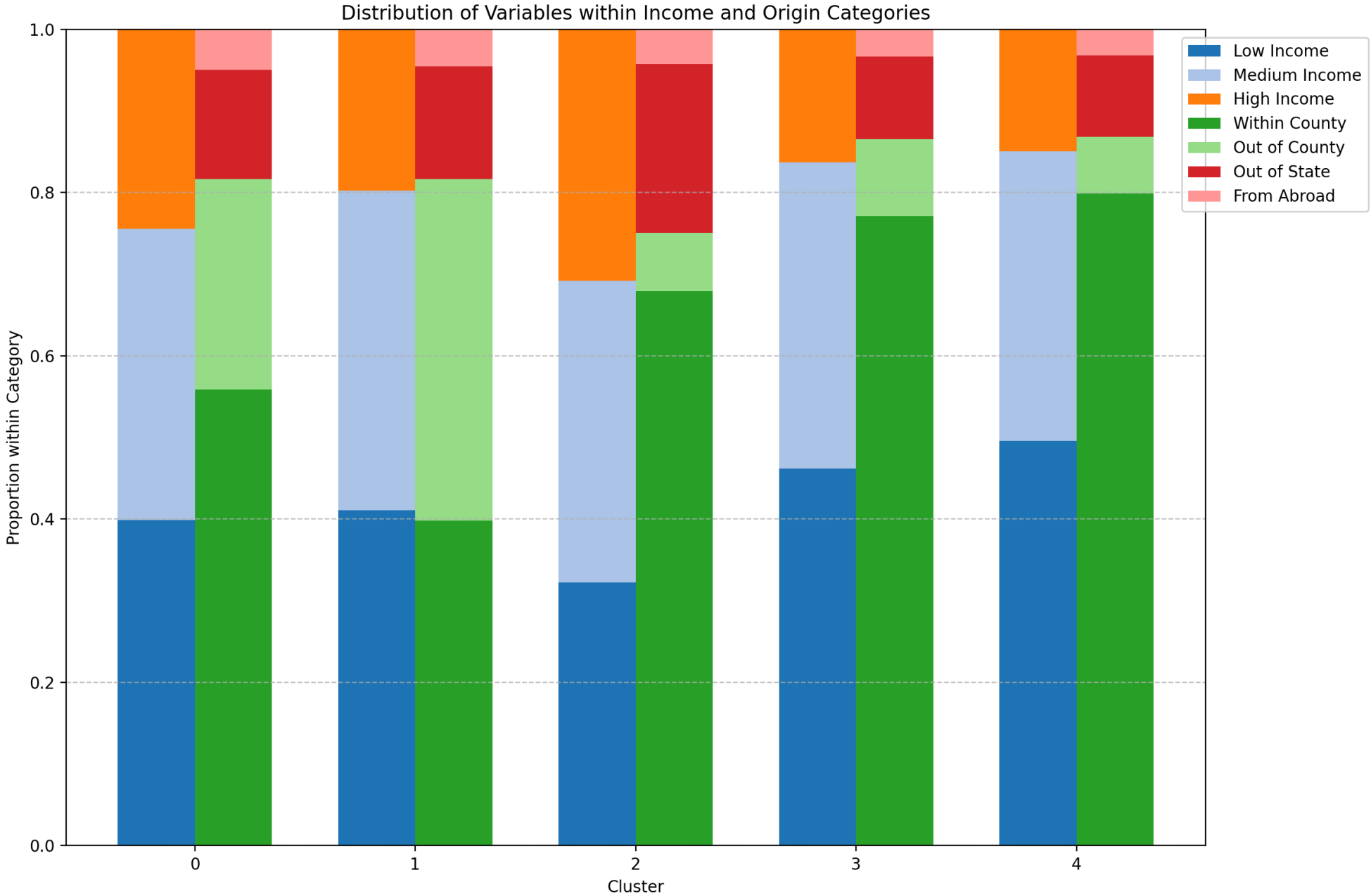

Silhouette Scores for Different Numbers of Clusters
Average Silhouette Score
0.50
0.45
0.40
0.35
0.30
0.25
0.20
2
4
6
8
10
12
14
Number of clusters

Clustering - Continued

- **Understanding cluster representations**
  - The descriptives of the PCs already gave us some information
  - Further check: aggregate the variables before PCA transformation by cluster - variables indicating place of origin and income levels

	PC1	PC2	PC3	PC4	PC5	PC6
cluster						
0	-0.0394	0.0026	-0.0051	0.0006	0.0007	0.0001
1	-0.0132	0.0337	0.0199	-0.0089	-0.0004	0.0002
2	0.1082	0.0143	-0.0114	0.0009	0.0021	0.0004
3	-0.0085	-0.0073	-0.0015	0.0008	-0.0006	-0.0006
4	0.0351	-0.0128	0.0077	0.0009	-0.0010	0.0005


Cluster Analysis by Income and Origin Variables
Low Income
Medium Income
High Income
Within County
Out of County
Out of State
From Abroad
0.20
0.15
0.10
0.05
0.00
Proportion
0
1
2
3
4
Cluster


Distribution of Variables within Income and Origin Categories
Proportion within Category
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Cluster
Low Income
Medium Income
High Income
Within County
Out of County
Out of State
From Abroad

Clustering - Continued

- **Tentative labeling:**
  - 0 - minimal inflow, attracts extreme population
  - 1 - moderate inflow, external working-class magnet
  - 2 - high inflow, internal high-income magnet
  - 3 - low inflow, attracts internal working-class
  - 4 - moderate inflow, internal working-class magnet

Sideline Task: Machine Learning for Predicting 2023 Inflow Types

- **Features used:** variables in the prepared data that's unrelated to inflow migration, besides region identifiers
- **Labels:** clusters generated through the previous step
- Normalized numeric values besides categorical variables ("rail" indicating whether the tract has direct access to rail)
- Used SVMSMOTE oversampling to handle **class imbalance**
- Models explored:
  - RF; fine-tuned using Random Search
  - XGBoost; fine-tuned using Random Search
  - Ensemble models: Stacking and Voting classifier

Model Performance Comparison:

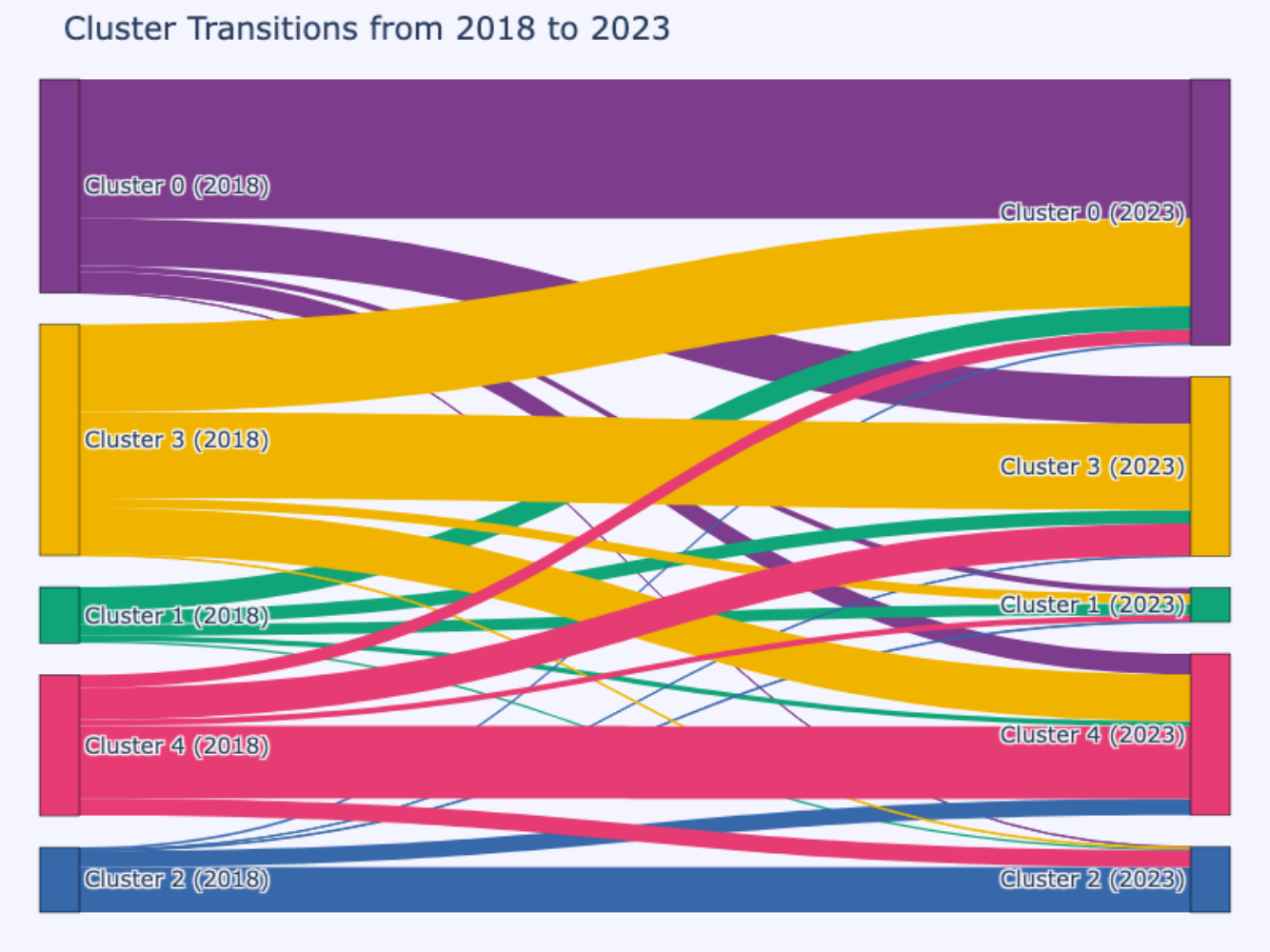
Model	Accuracy	Balanced Accuracy	F1 Score	Recall
Random Forest	0.5689	0.5674	0.5663	0.5689
Random Forest RS	0.5480	0.5500	0.5495	0.5480
XGBoost	0.5218	0.5226	0.5213	0.5218
XGBoost RS	0.5236	0.5290	0.5225	0.5236
Stacking	0.5497	0.5353	0.5465	0.5497
Voting	0.5462	0.5397	0.5444	0.5462

	Feature	Importance
33	prop_bd	0.069257
14	rhu_prop	0.064639
13	ohu_prop	0.056298
34	prop_grad	0.051941
15	pro_old_build	0.036682
10	log_iinc	0.034083
12	white_prop	0.031216
11	price_rent_ratio	0.030828
8	log_mhval	0.029860
31	prop_hhinc_200000	0.029638

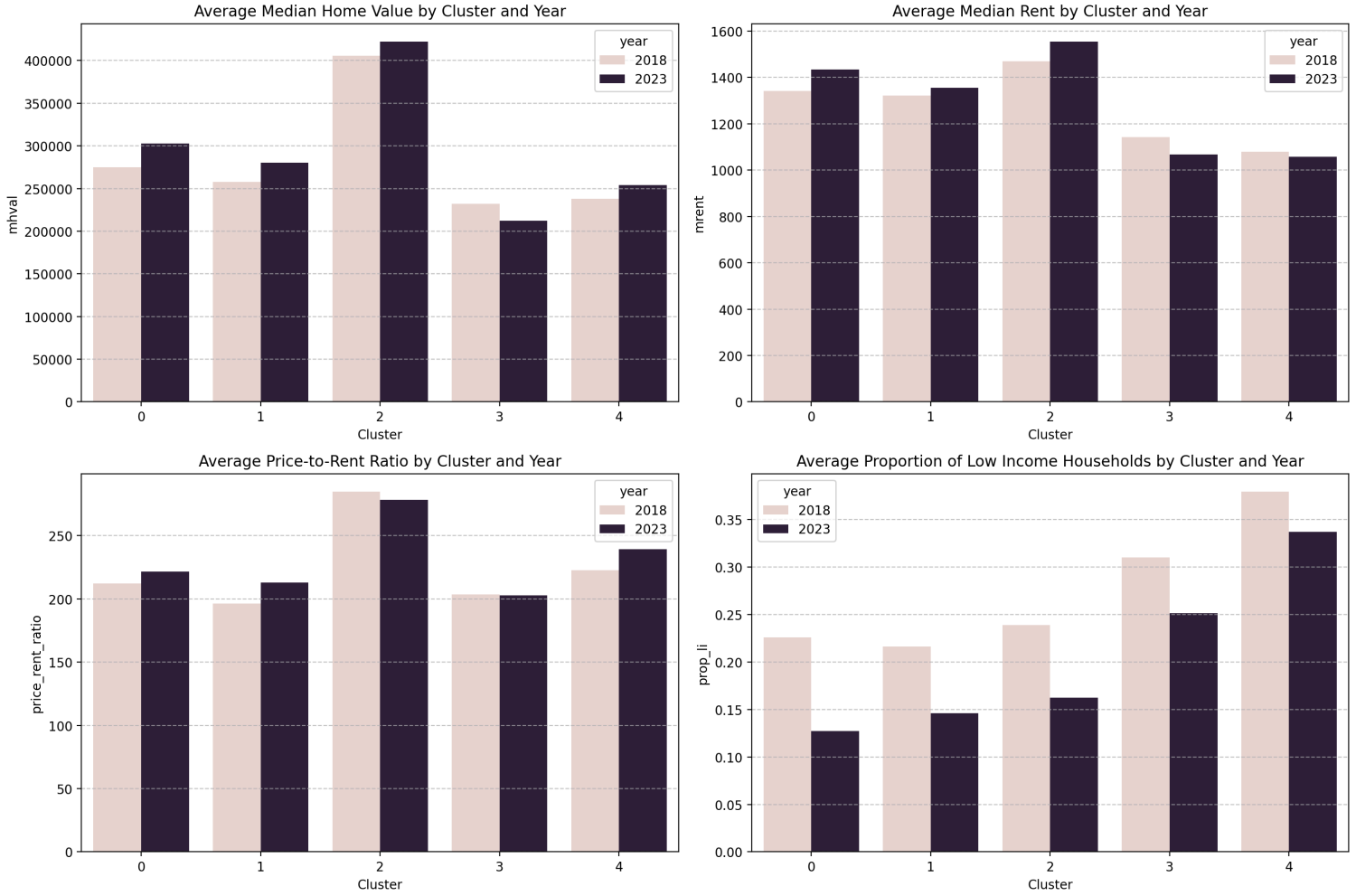
- Top 10 most predictive features: the most predictive features are ‘prop_bd’ (the proportion of individuals that have a bachelor degree) and ‘rhu_prop’ (proportion of renter occupied units)

Analysis

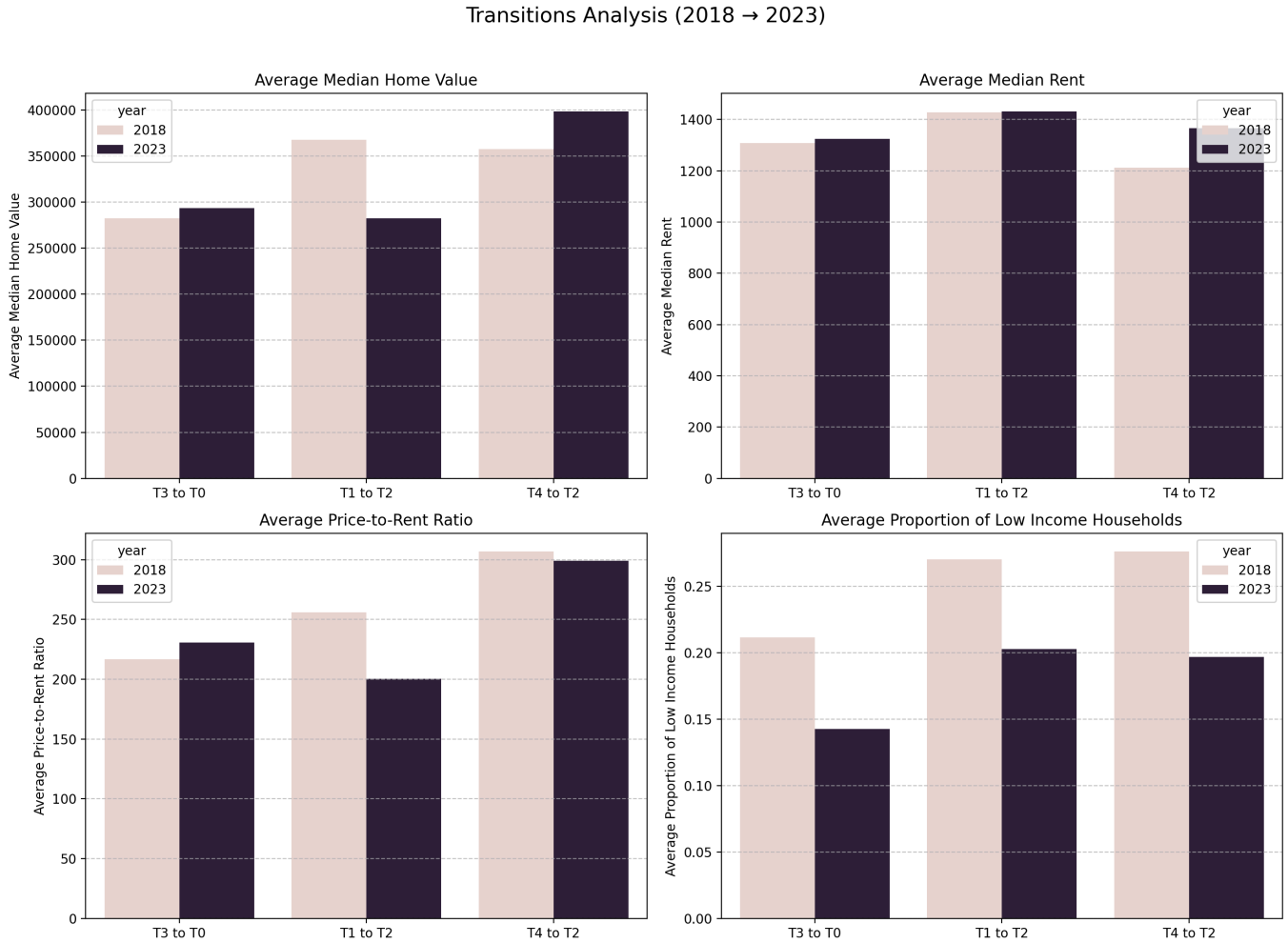
- Changes in cluster type

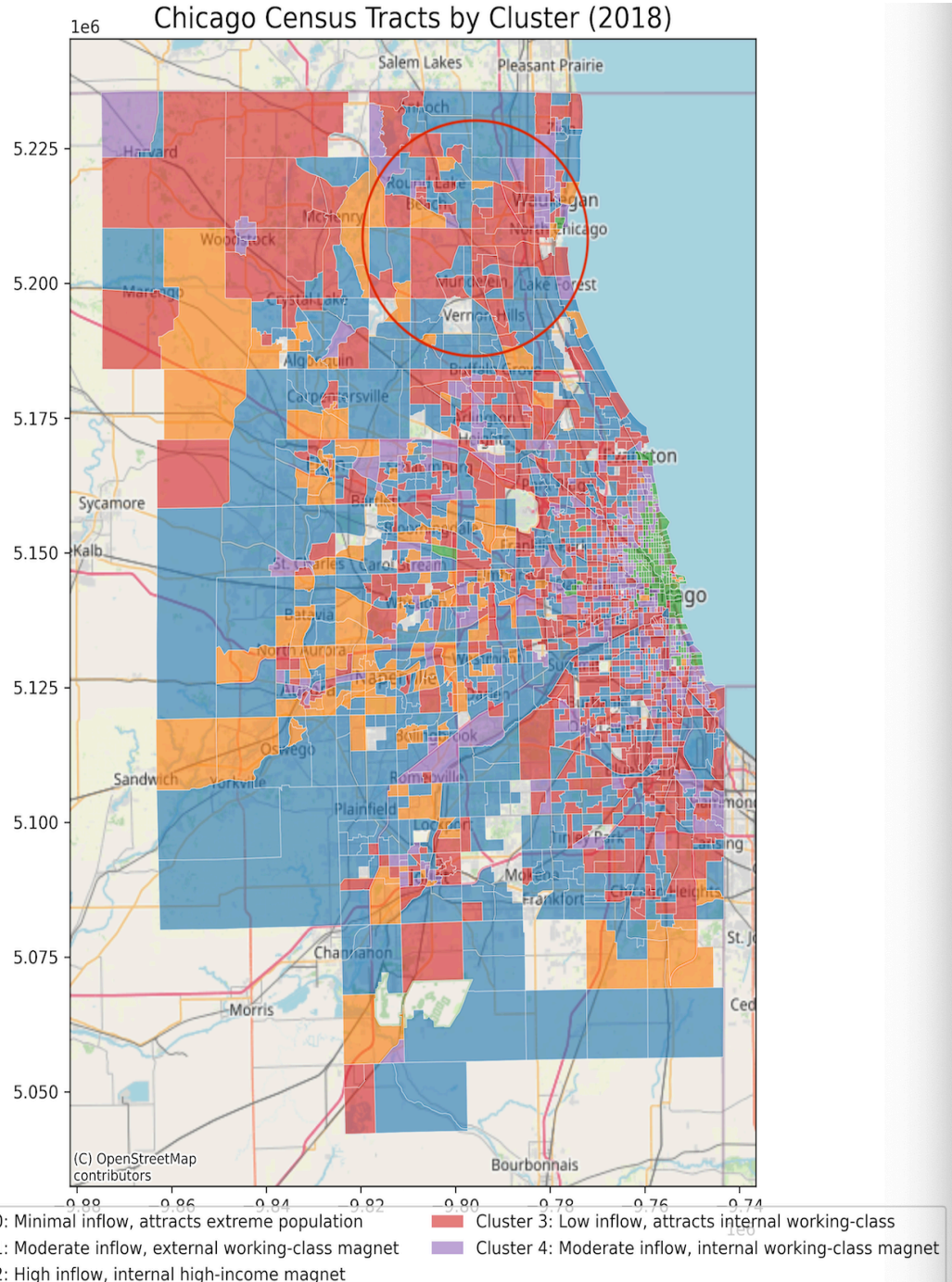


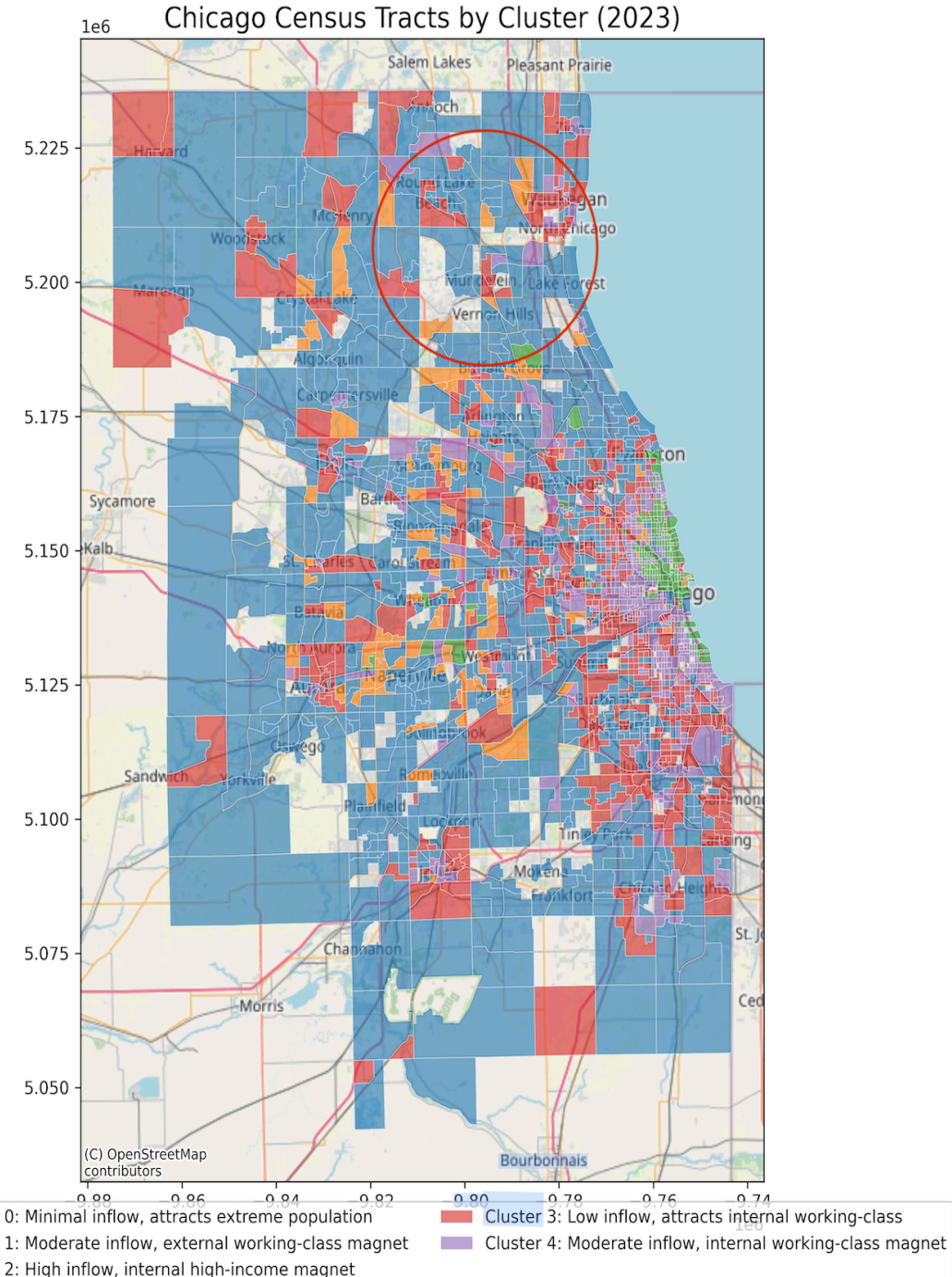
- Change in key features, 2018 - 2023



- Changes in key features for tracts that experienced certain transitions
- **Recall the labeling:** 0 - minimal inflow, attracts extreme population; 1 - moderate inflow, external working-class magnet; 2 - high inflow, internal high-income magnet; 3 - low inflow, attracts internal working-class; 4 - moderate inflow, internal working-class magnet




Chicago Census Tracts by Cluster (2018)
1e6
5.225
5.200
5.175
5.150
5.125
5.100
5.075
5.050
Salem Lakes
Pleasant Prairie
Harvard
Woodstock
Lake Forest
Sycamore
Kalb
Sandwich
Plainfield
Morris
Bourbonnais
(C) OpenStreetMap contributors
Cluster 0: Minimal inflow, attracts extreme population
Cluster 1: Moderate inflow, external working-class magnet
Cluster 2: High inflow, internal high-income magnet
Cluster 3: Low inflow, attracts internal working-class
Cluster 4: Moderate inflow, internal working-class magnet


Chicago Census Tracts by Cluster (2023)
1e6
5.225
5.200
5.175
5.150
5.125
5.100
5.075
5.050
Salem Lakes
Pleasant Prairie
Harvard
Woodstock
Lake Forest
Sycamore
Kalb
Sandwich
Plainfield
Morris
Bourbonnais
(C) OpenStreetMap contributors
Cluster 0: Minimal inflow, attracts extreme population
Cluster 1: Moderate inflow, external working-class magnet
Cluster 2: High inflow, internal high-income magnet
Cluster 3: Low inflow, attracts internal working-class
Cluster 4: Moderate inflow, internal working-class magnet

References

- Greenlee, Andrew J. 2018. “Assessing the Intersection of Neighborhood Change and Residential Mobility Pathways for the Chicago Metropolitan Area (2006–2015).” Housing Policy Debate 29 (1): 186–212. doi:10.1080/10511482.2018.1476898.
- DeLuca, Stefanie. 2018. “Residential Mobility and Neighborhood Change in Chicago.” Housing Policy Debate 29 (1): 213–16. doi:10.1080/10511482.2018.1524447.
- Kim, Namwoo, Hyeyeong Lee, and Yoonjin Yoon. 2023. “A Heterogeneous Attention Network Model for Longitudinal Analysis of Socioeconomic and Racial Inequalities in Urban Regions: Evidence from Chicago, IL.” In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, 1–4. Hamburg Germany: ACM. doi.org/10.1145/3589132.3625650.