Intergenerational Mobility In Education: A Segmentation Attempt On a Global Scale

Josh Vo
Denison University

Abstract

Intergenerational mobility (IGM) is a critical measure of societal equality and economic opportunity. This study conducted a segmentation analysis with the aim of categorizing 153 countries into distinct levels of educational IGM. We used two sets of variables (**Rank-based** and **Transition**) for our analyses because of their ability to capture data variance and independent qualities. Through k-means classification, we identified four clusters for each variable set that showed face-value validation through visualization. However, further mixed-method validation suggested limited discriminatory power of the clusters, hence no further support for our classification models. These results might indicate potential non-monotonic relationships between IGM and other indicators. The results also suggested a more broadened scope for IGM in future research as well as the need for a specific metric measuring educational IGM

References and Acknowledgement

Lloyd-Ellis, H. (2000). Public Education, Occupational Choice, and the Growth-Inequality Relationship. *International Economic Review*, *41*(1), 171–201.

Maoz, Y. D., & Moav, O. (1999). Intergenerational Mobility and the Process of Development. *The Economic Journal*, *109*(458), 677–697. https://doi.org/10.1111/1468-0297.00468

Narayan, A., Van Der Weide, R., Cojocaru, A., Lakner, C., Redaelli, S., Mahler, D. G., Ramasubbaiah, R. G. N., & Thewissen, S. (2018). *Fair Progress?: Economic Mobility across Generations around the World*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1210-1

Torche, F. (2015). Analyses of Intergenerational Mobility: An Interdisciplinary Review. *The ANNALS of the American Academy of Political and Social Science*, *657*(1), 37–62. https://doi.org/10.1177/0002716214547476

Vo, J. (2024). Intergenerational Mobility Project. *Github*. https://github.com/triuyenvo/Intergenerational_Mobility_Project

Van Der Weide, R., Lakner, C., Mahler, D. G., Narayan, A., & Ramasubbaiah, R. (2021). Intergenerational Mobility Around the World. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3981372

This project is part of my senior Data Analytics capstone research before graduation. I want to extend my gratitude and appreciation to Dr. Sarah Supp for her mentorship and extensive support through the research. I would also like to credit the Data Analytics Department at Denison University for the wonderful academic opportunities I have received in the past 4 year. Lastly, big thanks to Holly Duong for her insightful peer review on my paper.

Introduction

IGM refers to the level of association between adult children's socioeconomic standings compared to their parents. A higher association means less mobility–parents' backgrounds play a bigger role in shaping their children's future socioeconomic status. Educational IGM is a topic of international interest as it has been shown to be both a good indicator and driver of other critical domains such as economic efficiency, social equity, or development of human ability (Lloyd-Ellis, 2000; Maoz & Moav, 1999; Van Der Weide et al., 2021).

Previous research on educational IGM has neither (1) produced a robust index to measure IGM nor (2) categorized countries using this indicator in much depth. The widely-used categorization based on country type (i.e., developing vs. developed) cannot capture the complexity of IGM, potentially causing inaccurate generalization.

The current study was built off of The World Bank's original longitudinal IGM project (Van Der Weide et al., 2021). We conducted a segmentation analysis in hope to classify 153 countries into different IGM groups/levels. We specifically utilized two sets of variables (i.e., **Rank-based** and **Transition**) for our analyses because:

(1) - They captured much variance of our data

(2) - They are independent of other factors, only measuring vertical movement up and down educational standardized levels/ranks that are internationally recognized.

Methods

Data and Measures

We used two public datasets collected by The World Bank; both were already synthesized by the collectors and thereby required minimal cleaning effort. The first dataset was the Global Database On Intergenerational Mobility (GDIM; The World Bank, 2023; Van Der Weide et al., 2021) which included summary survey data from 153 countries collected between 1991 and 2017, representing 97% of the world population. The second dataset was the Global Gini Index Data (GGID; The World Bank, 2024) which contained annual Gini indices of 266 countries from 1963 to 2022.

Variables/Measures of interests	Description
CAT, YOS, 1-COR, 1-BETA	Four educational IGM metrics previously derived in the original report by The World Bank ⇒ used for model validation later
Gini Index	Economic metric for inequality
Rank-based set (12)	Movement after one generation. Selection criterion: parents below 50th percentile
Transition set (25 variables)	Movement after one generation on the ISCED scale. No selection criterion for parents

Statistical Analyses

We used a mixed-method approach to run our classification and validation. The following analyses were conducted on both **Rank-based** and **Transition** sets:

1. K-Means Clustering
   Chose the number of clusters through elbow graphs showing within-cluster sum of squares (WCSS).
2. Model Validation
   2.1. Dimension reduction - Margin check
   Run Principle Component Analysis (PCA) to select 2-3 principle components that capture most of the variable set's variance. Then graph clusters against PCs to visually check for clear margins.
   2.2. Between-group difference assessment
   Run Analyses of Variance (ANOVA) and Tukey's post hoc tests to compare the values of 5 descriptive metrics between each cluster. These 5 metrics included the Gini index and pre-derived education IGM measures (i.e., CAT, YOS, COR, BETA).
3. Segmentation assessment
   - Analyse validation results to describe each cluster's unique qualities.
   - Compare the classification output of both **Rank-based** and **Transition** sets to see if they show convergent results (i.e., using a *Cluster Matrix*)

Results

Clusters Produced by K-Means Clustering

Figure 1
K Selection for ***Rank-based Variable Set*** *– Elbow on WCSS Graph*

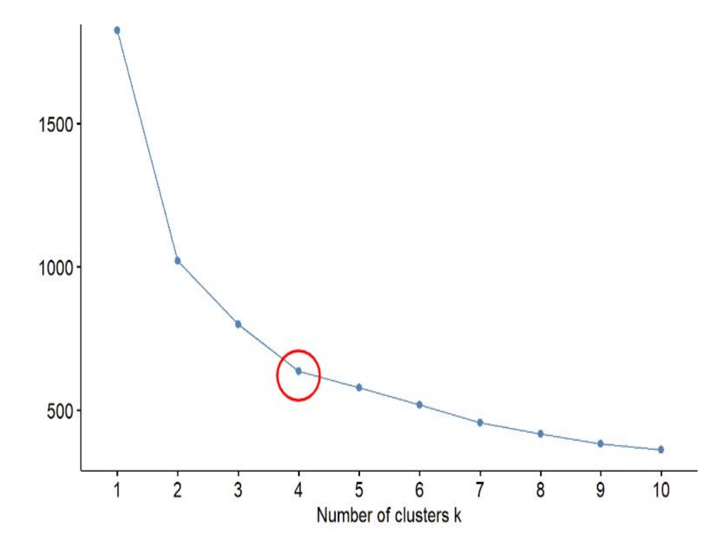
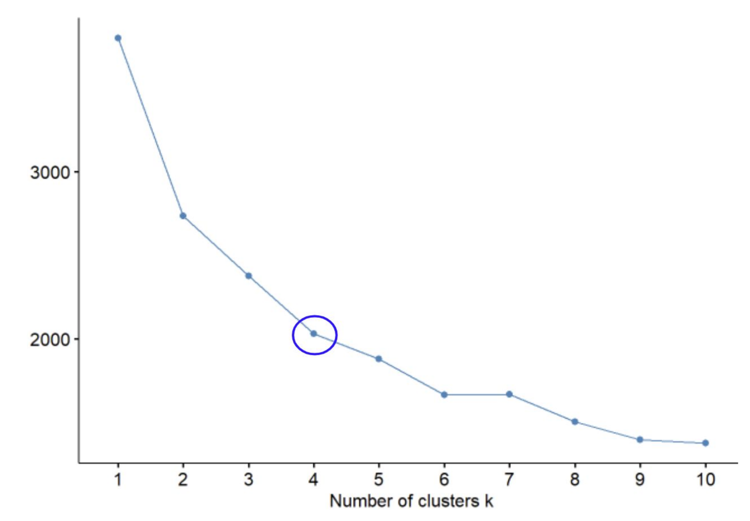


Figure 2
K Selection for ***Transition Variable Set*** *– Elbow Method on WCSS Graph*



⇒ **Figures 1 & 2** indicate that choosing 4 clusters would be optimal for K-means classification.

Validation

Figure 3
Clusters Based on ***Rank-based Variable Set*** *Plotted Against The First Three Principal Components*

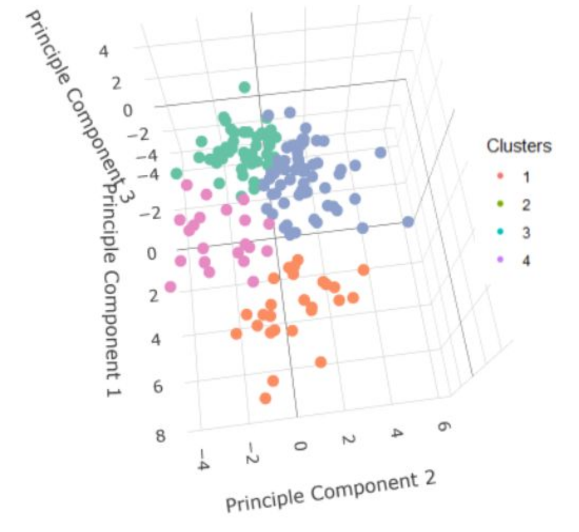


Figure 4
Clusters Based on ***Transition Variable Set*** *Plotted Against The First Three Principal Components*

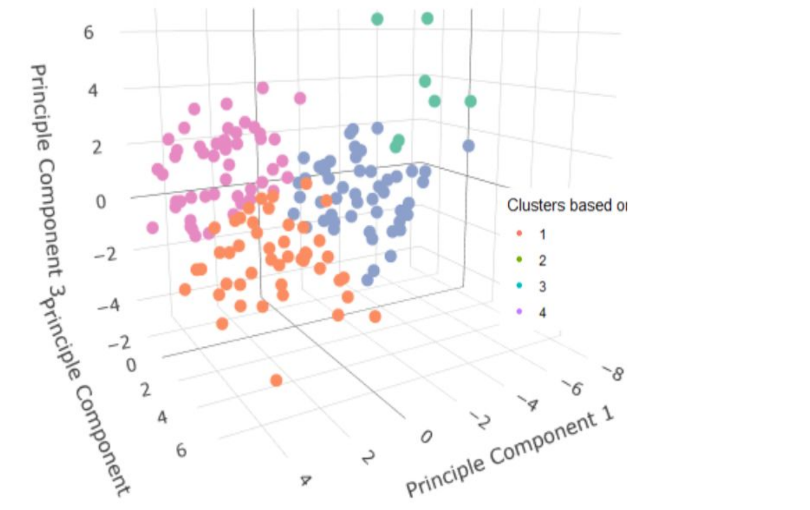


Table 1
Rank-based *Clusters' Mean & Standard Deviation – 5 Descriptive Metrics*

Metrics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CAT*	0.598 (0.139)^{a}	0.309 (0.174)^{b}	0.595 (0.165)^{a}	0.462 (0.138)^{c}
YOS*	0.700 (0.119)^{a}	0.432 (0.197)^{b}	0.654 (0.143)^{ac}	0.591 (0.135)^{c}
1 - COR*	0.486 (0.051)^{a}	0.643 (0.120)^{b}	0.646 (0.067)^{b}	0.518 (0.085)^{a}
1 - BETA*	0.495 (0.125)^{b}	0.387 (0.200)^{a}	0.658 (0.097)^{c}	0.398 (0.173)^{a}
Gini*	41.05 (8.54)^{a}	38.36 (5.79)^{a}	34.83 (6.29)^{b}	39.75 (7.60)^{a}

Table 1
Transition *Clusters' Mean & Standard Deviation – 5 Descriptive Metrics*

Metrics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CAT*	0.516 (0.150)^{a}	0.678 (0.115)^{b}	0.332 (0.139)^{c}	0.589 (0.589)^{a}
YOS*	0.633 (0.143)^{ab}	0.747 (0.112)^{a}	0.469 (0.165)^{c}	0.653 (0.104)^{b}
1 - COR	0.575 (0.148)^{a}	0.574 (0.107)^{a}	0.602 (0.111)^{a}	0.574 (0.095)^{a}
1 - BETA*	0.652 (0.120)^{a}	0.541 (0.182)^{a}	0.426 (0.170)^{b}	0.606 (0.138)^{a}
Gini*	45.74 (10.12)^{a}	36.50 (7.16)^{b}	40.97 (6.00)^{a}	34.86 (7.07)^{b}

Note. *Means with different superscripts are significantly different at $p \le .05$. Metrics with statistically significant ANOVA results ($p \le .05$) are signified with an asterisk (*).*

- Fairly separated clusters with some spatial overlapping (**Figures 3 & 4)** ⇒ Initial face validation for both K-means clustering models.
- Although 9 out of 10 ANOVAs returned significant results (**Tables 1 & 2**), Tukey's post hoc showed no consistent patterns in between-cluster differences ⇒ No further validation for both models.

Segmentation Assessment

Table 3
Classification Matrix Comparing Clusters Derived from ***Rank-based*** *versus* ***Transition*** *Variable Sets*

Clusters from two variable sets	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster 1	4	15	3	20	42
Cluster 2	0	3	23	0	26
Cluster 3	3	20	12	28	63
Cluster 4	0	8	12	2	22
Total	7	46	50	50	

➡

Between-set Ratio
42:26:63:22
vs.
7:46:50:50

Note. *Each row shows how a* ***Transition*** *cluster's observations are distributed in other* ***Rank-based*** *clusters. Conversely, each column shows how a* ***Rank-based*** *cluster's observations are distributed in other* ***Transition*** *clusters. Note that the order and label (i.e., 1,2,3,4) of the clusters do not matter.*

- Disproportionate between-set observation ratio.
- No prominent between-set overlapping (e.g., one **Rank-based** cluster made up most of one **Transition** cluster).
  ⇒ Classification outputs from **Rank-based** and **Transition** sets were not convergent.

Conclusion and Discussion

- The segmentation did not produced meaningful categorization.
- There could be non-monotonic relationships between IGM and other indicators.
- Our validation might have overly relied on Gini index and The World Bank's pre-derived measures.
- Future segmentation attempt could broaden the current educational scope to include other domains of IGM, considering intersectional topics.
- Future research should try to derive a specific metric for educational IGM.