From Some to None? Fertility Expectation Dynamics of Permanently Childless Women

Alison Gemmill[1]

Appendix

The selection of the clustering algorithm (Ward's or *k*-medoid) and number of clusters was jointly determined by evaluating several measures of partition quality provided in the Weighted Cluster package in R (for more details about the measures presented here, see Studer (2013)).

I use the following measures of partition quality summarized in the table below (adapted from Studer (2013)).

Name	Abbreviation	Range	Min/Max	Interpretation
Average Silhouette Width (weighted)	ASWw	[-1; 1]	Max	Coherence of assignments. High coherence indicates high between-group distances and strong within-group homogeneity.
Hubert's Gamma	HG	[-1; 1]	Max	Measure of the capacity of the clustering to reproduce the distances (order of magnitude).
Point Biserial Correlation	PBC	[-1; 1]	Max	Measure of the capacity of the clustering to reproduce the distances.
Hubert's C	HC	[0; 1]	**Min**	Gap between the partition obtained and the best partition theoretically possible with this number of groups and these distances.
Psuedo R^2	R2	[0; 1]	Max	Share of the discrepancy explained by the clustering solution (only to compare partitions with identical number of groups).
Psuedo R^2 (squared)	R2sq	[0; 1]	Max	As previous, but using squared distances.

I graph standardized values of each of these measures across 15 possible cluster solutions to assess the performance of the clustering algorithm and to aid in the selection of the number of clusters (Appendix Figure 1). As implied by the "Min/Max" column in the table, higher values for all metrics except HC suggest better performance.

As seen in Appendix Figure 1, the *k*-medoid clustering algorithm outperforms Ward's for most cluster solutions across all quality measures. Exceptions come from the 3-cluster solution using Ward's, which may be due to assumptions of hierarchical clustering underlying the algorithm. Notably, however, both measures of R2 and R2sq are higher for the *k*-medoid algorithm across all possible cluster solutions.

The selection of the number of clusters based on the *k*-medoid algorithm is aided by identifying local maxima for the ASWw, HG, and PBC measures, and local minima for the HC metric. Both ASWw and PBC show local maximum peaks at 2, 5, and 7 clusters. For both the HG and HC, however, no local maximum or minimum peaks emerge; instead, values for these measures suggest better performance as cluster number increases.

Using these quality measures as a guide, I choose a 5-cluster *k*-medoid solution to provide sufficient variation across groups without introducing excessive complexity with inadequately sized groups.

Appendix Figure 1. Quality measures of clustering results.