A Clustering Approach to Bottom-up Theories of Subjective Well-being: Is Life Satisfaction Differentially Predicted by Levels of Domain Satisfaction?

Wai Yu Amanda, Ng | Department of Statistical Science | University of Toronto (St George Campus)
Sofia Panasiuk | Department of Psychology | University of Toronto (St George Campus)

Introduction

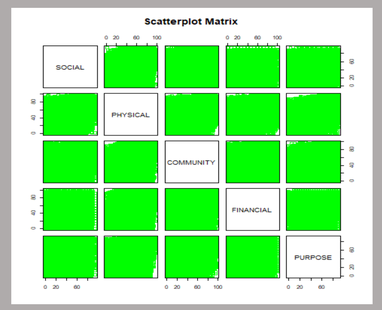
- Different individuals, at various stages of their lives, place more importance on one life domain than another.
- The study can be used to gain more understanding of how changes in well-being vary by hidden clusters in the population.
- Here, we will discuss various clustering method used and comment on the barriers we face.

The dataset

- Gallup-Sharecare Well-Being Index, 2014-2017
- Sample size: 529,237 (after data cleaning)
- Demographics: age, health, marital status, race, education
- Life domains: purpose, community, physical, financial and social well-being
- Subjective well-being measures: life satisfaction, positive affect and negative affect

Research questions

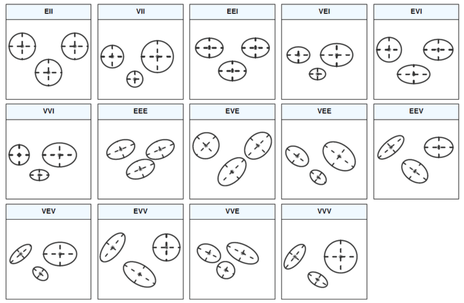
1. Are individuals readily clustered into groups according to life domains?
2. How many clusters exist and what are their qualities?
3. Do clusters differ from one another on demographical variables?
4. What is the overall subjective well-being of individuals in each of these clusters? Do clusters differ from one another on Subjective well-being?



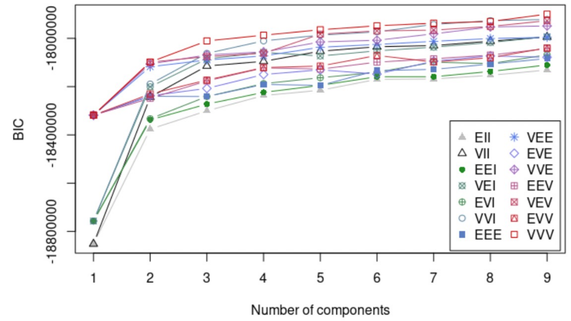
Methods

1. Model-based clustering using Finite Gaussian Mixture Model

- This is a relatively accurate clustering method with high statistical power and does not require us to specify the number of clusters prior to analysis.
- Different models are built by varying the distribution (spherical, diagonal, ellipsoidal), volume (equal, varying), shape (equal, varying) and orientation.
- **Expectation–maximization algorithm** is an iterative method used to optimize the parameters of the statistics model.
- **Bayesian information criterion (BIC)** scores a model based on its log-likelihood and complexity and reflects if overfitting occurs in the model. Generally, the higher the BIC score a model has, the better the model is. It also gives us the desired optimal number of clusters.



The best model will be determined by comparing different models' BIC values



*BIC value trend tends to increase as number of components increases

Problems encountered and thoughts:

- BIC gave us **"finitely many clusters"** because it is a strict criterion which results in high accuracy in the parameters and overlapping between clusters.
- However, it is not reasonable to **over-cluster data points** in our research project given limited time.
- Model-based clustering is a soft clustering method which clusters are allowed to overlap with each other. So, a datapoint (participant) is allowed to fall into multiple clusters. Hence, we may not be able to clearly identify which group does a participant is classified. Datapoint-cluster tracking is an important element of this research.

2. Dimensionality reduction

By conducting pairwise comparison analysis, we found out that all life domains correlates with each other, while PURPOSE is the most significant one.
We then used the caret package in R which can remove redundant dimensions because domains are highly correlated with each other.

	PURPOSE	COMMUNITY	PHYSICAL	FINANCIAL	SOCIAL
PURPOSE	1.00	0.54	0.52	0.49	0.61
COMMUNITY	0.54	1.00	0.36	0.42	0.45
PHYSICAL	0.52	0.36	1.00	0.42	0.44
FINANCIAL	0.49	0.42	0.42	1.00	0.44
SOCIAL	0.61	0.45	0.44	0.44	1.00

Pairwise table of correlation coefficients

3. K-means clustering

- This method requires us to pre-specify the number of clusters (which is doable) if we use the elbow method of determining the optimal cluster number.
- The optimal cluster number (elbow point) is 2.

Problems encountered and thoughts:

- The algorithm divided the data into 2 groups even if there are no obvious clusters.
- Eg. In the pairwise plots of each well-being domain dimension, there are **no any cluster popping out visually** (see plot on the left).
- This does not mean there are no clusters. But other clustering methods may be better suited for us than k-means hard clustering.

A Clustering Approach to Bottom-up Theories of Subjective Well-being: Is Life Satisfaction Differentially Predicted by Levels of Domain Satisfaction?

Wai Yu Amanda, Ng | Department of Statistical Science | University of Toronto (St George Campus)
Sofia Panasiuk | Department of Psychology | University of Toronto (St George Campus)

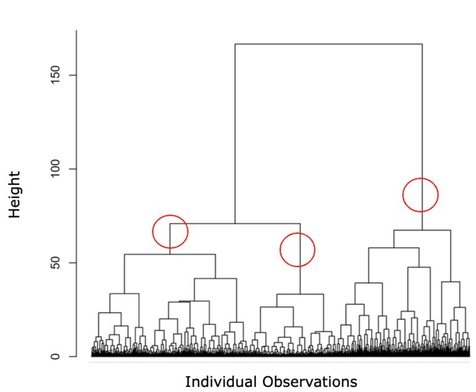
Methods (con't)

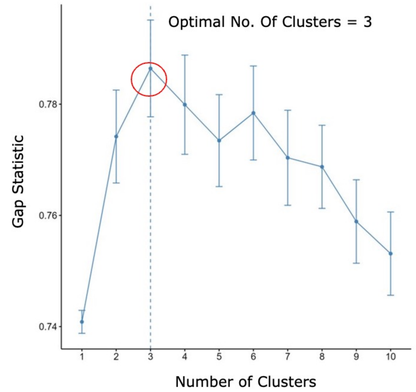
4. Agglomerative Hierarchical Clustering

- The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.
- To higher the gap statistics , the better the number of clusters it is.

Problems encountered and solutions:

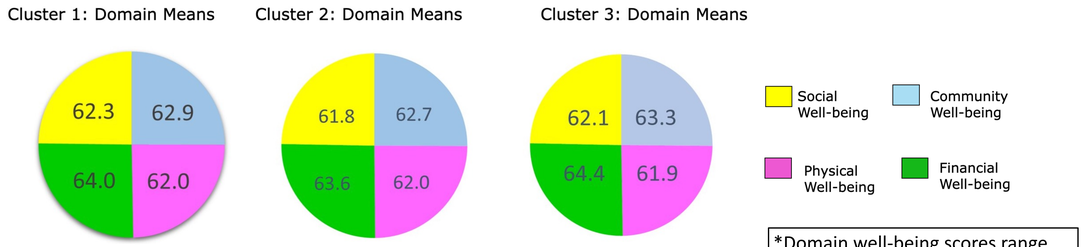
- It is very computationally expensive and slow. So, we sampled datapoints from the original dataset and send it to the Supercomputer to run the codes.
- We first separated the original dataset into "training dataset" and "testing dataset".
- In "training dataset", we randomly selected 10000 datapoints to run the algorithm.





- The algorithm gave us 3 as the optimal number of clusters.
- We confirmed it by using "for-loop" to randomly reselect 10000 datapoints from the "training dataset" and run the algorithm on the subsets.
- We also tested the result with the "testing dataset".

Clusters characteristics analysis



*Domain well-being scores range from 0 to a maximum score of 100

For continuous variables:

- All life domains (community, physical, financial and social well-being)
- Age
- All Subjective well-being measures (life satisfaction, positive affect and negative affect)

We ran T-test to confirm that we cannot reject the null hypothesis that there is **no significant differences** in the variables among the clusters identified.

For categorical variables:

Health	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Excellent	20.0	20.9	18.2
Good	31.3	31.0	32.6
Marginal	13.6	14.0	14.0
Neutral	30.2	29.2	30.6
Poor	4.8	4.9	4.5
Not Available	0.1	0.0	0.2

Race	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Asian	2.4	2.2	2.3
Black	9.5	9.6	8.3
Hispanic	9.1	8.4	9.6
Other	1.5	1.4	1.6
White	74.0	75.1	75.3
Not Available	3.5	3.2	3.0

Marital Status	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Divorced	10.7	10.4	10.7
Married	51.7	51.3	51.6
Separated	1.7	2.3	1.6
Single	19.2	19.7	19.0
Widowed	10.1	11.0	10.4
Not available	6.6	5.4	6.7

Education	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Less than high school	6.0	5.8	6.0
High School	23.5	23.4	23.0
Some college	24.3	25.5	24.9
College graduate	21.6	21.9	20.9
Post-grad	18.6	18.3	19.3
Technical School	4.0	3.7	4.2
Not available	2.0	1.5	1.6

Similarly, there is **no significant differences** in the variables among the clusters identified

Conclusion

- We successfully identified 3 sub-populations using the life domain variables.
- However, the clusters do not differ significantly from one another on life domains, demographical variables and subjective well-being.
- Meaning that we failed to identify the clusters unique characteristics.
- The problem of identifying sub-populations with higher/lower well-being is harder than we initially thought.

Discussion

- Null effect in cluster unique characteristics may due to:
  1. Unsuitable clustering method
  2. No cluster exist
  3. The variables we tested may not describe the clusters well. Perhaps, there are other undiscovered variables that can be used to define our clusters.
- People **don't neatly fit** into well-being domain clusters.
- Well-being exists on a continuous scale.