Cluster Analysis (CA) and Latent Class Analysis (LCA) in the identification of latent adverse childhood groups.

Juan Rivillas, School of Public Health, Department. Biostatistics and Epidemiology.
MRC Centre for Environment and Health. Email: j.rivillas-garcia20@imperial.ac.uk

Last updated: 10th Sep 2022.

Supervisors: Paolo Vineis, MD, MPH, FFPH, Faculty of Medicine, School of Public Health. Chair in Environmental Epidemiology. Email: p.vineis@imperial.ac.uk
Emilie Courtin, Ph.D. Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine LSHTM. Email: emilie.courtin@lshtm.ac.uk

Objective: To identify of the best fit model for the classification of high-risk groups.

In this document, I explain how cluster analysis and latent class analysis were used to classify participants into latent groups based on their adverse childhood experiences (ACE). In cluster analysis, the optimal numbers of clusters ranged from 2-6 across four applied approaches. While, in the laten class analysis 2-class were consistent in the models fit criteria as fit diagnostic criteria through the three datasets. The PolCA R package facilitated the identification of Latent Class Profiles of Adversity Childhood Experiences and tests different criteria to validate methods in the SABE datasets.

This document is structured as follows:

1) Models of the effects of adversity

In the ACEs literature, the dominant model of the effects of developmental adversity on later health is the cumulative risk model (the "ACE Score"). Multiple Individual Risks (MIR) is the most common alternative characterization for the ACE Score (Lanoue et al., 2020). Previous work contrasted cumulative risk and MIR models of the relationship between ACEs and adult health outcomes and identified limitations as predictors in statistical models. The cumulative risk model is the ACE Score, which indicates the number of exposures. The cumulative risk model answers the question, *"what is the impact of increasing numbers of events*?". Although, it is statistically robust approach, may not be necessarily the best characterization of the impacts of childhood adversity on adult health for all outcomes because it obscures the relative contributions of individual adversity event types. In contrast, MIR models include the presence or absence of multiple separate ACEs as independent predictors in a single regression model. This model answers the question, *"what is the impact of the occurrence of each specific event (given the presence/absence of the other events"?* Both models classify individuals based on ACEs into scores or categories rather than deriving the number of items and exposure coding strategy to choosing the optimal classification of people into adversity groups.

Additionally, differences by health outcomes have been identified in the literature. For instance, the multiple individual risks model is better than the categorical ACE Score in the depression outcome. While the cumulative risk and MIR models were of comparable fit but yielded different and complementary inferences for the consequences of obesity and cardiovascular diseases (Lanoue et al., 2020). Although, researchers agreed that Researchers working in the ACEs framework have conceptualized how ACEs might be related to outcomes by applying models like cluster analysis (Pamulaparty et al., 2016) or latent class analysis (Weller et al., 2020) to classify people rather than collapse them into ACEs into scores or sums of exposures.

Cluster analysis and Latent Class Analysis (LCA)

LCA and cluster analysis are similar. The two statistical procedures are considered "person-oriented analyse" (Weller et al., 2020)(Collins & Lanza, 2010), which use patterns of scores to identify individuals who can be grouped. In contrast, variable-centered approaches look for relationships among variables. A series of solutions are generated in both cluster analysis and LCA. Researchers use statistical and theoretical criteria to decide which solution is best. Cluster analysis and LCA make different assumptions about the data and use other statistical procedures despite these similarities. In cluster analysis, the assumption is that the cases with the most similar scores across the analysis variables belong in the same cluster ('Finite Mixture Modeling', 2006) (Nylund-Gibson & Choi, 2018). LCA assumes that latent classes exist and explain patterns of observed scores across cases. In cluster analysis, variable means are used to define the "nearness" of cases; therefore, analysis variables should be continuous. Because the analysis variables are categorical in LCA, cross-tabulations are used as the input information (Collins & Lanza, 2010). Case membership in clusters is determined in cluster analysis. In LCA, probabilities of class membership are obtained, not clear-cut assignments. However, the two statistical procedures can generate categorical classification variables for use in other analyses, and k-means clustering, and Latent Class Profiles are considered alternatives for identifying the optimal number of clusters and classes in childhood adversity groups. However, even though these two modelling choices can result in different conclusions, there is only limited evidence that directly contrasts them (Nylund-Gibson & Choi, 2018; Pamulaparty et al., 2016; Weller et al., 2020). A detailed description of the methods, advantages, and limitations of the two approaches is given in Table 1.

Table 1. Definition and estimation of K-mean clustering and Latent Class Analysis (LCA).

Item	K-means clustering	Latent Class Analysis (LCA)
Description	-The K-means algorithm is the most used and simplest method among all partitional clustering algorithms. -It is an iterative method which minimizes the sum of the squares for a given number of clusters. -Means clustering can be used to classify observations into k groups, based on their similarity. Each group is represented by the mean value of points in the group, known as the cluster centroid.	-LCA is also known as finite mixture modelling (ref), which are a popular method used to detect latent (or unobserved) heterogeneity in samples or for approximating general distribution functions. LCA is (Hagenaars & McCutcheon, 2002). -It is a special case of person-centered mixture 3odelling that identifies latent subpopulations within a sample based on patterns of responses to observed variables (B. O. Muthén & Muthén, 2000).
Assumption or questions	- The cases with the most similar scores across the analysis variables belong in the same cluster (Norusis, 1990). - This model answers the question: how many groups (clusters) share common characteristics or may be classified based on their similarity?	-The membership in unobserved classes can cause or explain patterns of scores across survey questions, assessment indicators, or scales (B. O. Muthén & Muthén, 2000; Wolke et al., 2013).
Methods to choosing the number of groups	K-means, Fuzzy C-means, Hierarchical agglomerative clustering (cluster dendograms), Hierarchical agglomerative clustering using" "average"" linkage, and Principal Component Analysis (PCA).	Cross-validation, domain-usefulness., entropy, extent of association with other data, information criteria, statistical tests, no small classes, and replicability,
Procedures	There are two approaches: Compute k-means for a range of k values and compute K-means algorithm several times with different initial cluster centers. K-Means reaches a state in which no points are shifting from one cluster to another. For measuring the quality of the clustering, the measure Sum of the squared error (SSE) is defined as $\mathrm{SSE}=\sum_{i=1}^{k}\sum_{x\in c_i} dist(c_i, x)^2$ Where dist is standard Euclidean distance between two objects in Euclidean space. The centroid (mean) of the ith cluster that minimizes the SSE is defined as $\sum_{i=1}^{k}\sum_{x\in c_i} dist(c_i, x)^2$	The model for traditional latent class analysis is then typically written as $\pi_{abc}^{ABC}=\sum_{x}\pi_{x}^{X}\pi_{a}^{A\mid X}\pi_{b}^{B\mid X}\pi_{c}^{C\mid X},$ where X is the latent class variable, π X x the size of class x and, for example, π A|X a is the probability that variable A takes on the value a in the latent class x. Equation 3 describes the probability of seeing any combination of values a, b, and c as depending solely on the differences in latent class sizes (π X|x) combined with how different these classes are in terms of the observed variables. Within the classes, the variables are unrelated "conditionally independent"), which is reflected in the product π A|X a π B|X b π C|X c .
Validation for selecting class model	Average Silhouette Method Gap Statistic Method	LL = log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; ALCPP = average latent class posterior probability (Predicted class memberships (by modal posterior probability)).
Advantages	-Highly scalable of the huge sum of data sets with (n * k * r) where r is the number of rounds, where n represent number of data items, k represents numbers of clusters.	
Limitations	-All the clustering techniques show ambiguity in clustering noisy data and outliers. -The Hierarchical clustering shows good results for small data sets and Fuzzy C means for the voluminous amount of data. K means technique has faster performance but finding the appropriate k value in the dataset is a challenge. -While applying cluster analysis we are contemplating that the groups exist. But this speculation may be false. The outcome of clustering should never be generalized.	-Presence of heterogeneity of the outcome probabilities within the true classes, which violates the assumption of conditional independence, and will require many classes to model the association in the data resulting in difficulties in interpretation LCA assigns individuals to classes based on their probability of being in classes given the pattern of scores they have on indicator variables (B. O. Muthén & Muthén, 2000). -Proper class assignment is not guaranteed. Also, because class assignment is based on probabilities, the exact number or percentage of sample members within each class cannot be determined. -Furthermore, researchers usually assign names to the identified classes and, because of the complexity of the classes, may advertently engage in ""naming fallacy"" wherein the name of the class does not accurately reflect the class membership.
Overview recommended	An overview on these models with many examples for applications is given in the publications of Harrell (2001), Dormann (2012), and Lavanya (2016). Harrell (2001), Hartigan, JA, and MA Wong. 1979	An overview on these models with many examples for applications is given in the recent monographs McLachlan and Peel (2000) and Frühwirth-Schnatter (2006), Nylund-Gibson & Choi, 2018).

2) Statistical analysis

Statistical package and year
Statistical analysis was performed using R Statistical Software (R Studio version 2021.09.1). Models were fit via methods for cluster analysis using the *cluster* (Martin Maechler & Peter Rouseeuw, 2022.) and *factoextra* packages (Alboukadel Kassambara & Fabia Mundt, 2022) commonly used for clustering algorithms and visualization of the results of Multivariate Data Analyses. To conduct Latent Class Analysis (LCA) were used *poLCA* package (Linzer & Lewis, 2011). Recommendation: load data as a .txt file to facilitate poLCA to process data efficiently. Multiple imputations by chained equations (MICE) package were used for the missing imputation data (van Buuren, 2018).

The code for this exploratory analysis is available at the following link: https://github.com/juanrivillas/Early-life-inequalities-and-biological-aging-Colombia-/blob/bd88edcec20809b6d06584ef2d368cec70f3b9e7/Scripts/clustering%20&%20LCA.Rmd

Estimation methods
There are six variables in the data: neglected food, household violence, migration, emotional abuse, early-life infection, and poor health reported at 15 years old. Then, the purpose is to classify participants based on the foregoing criteria and to identify profiles of adversities in childhood. The procedures between-category models and model comparisons are described as follows.

Clustering analysis
Two different clustering algorithms were used to compute k-means clustering:

- Compute k-means for a range of k values by varying k between 2 and 10, comparing the clustering results, and to choose the best k values.
- Compute the K-means algorithm several times with different initial cluster centres and select the lowest total within-cluster Sum of the square as the final clustering solution.

The completed steps to compute the K-means algorithm are described as follows:
1. Normalization and calculating distance matrix.
2. Select k points as initial centroids and repeat.
3. From K clusters by assigning each point to its closest centroid.
4. Recompute the centroid of each cluster until centroids do not change. K-Means reaches a state in which no points shift from one cluster to another, e.g., repeating until only 1% of the points change clusters. For measuring the quality of the clustering, I estimated the Sum of the squared error (SSE) or scatter.

The application of four different approaches allowed to compare k-mean clustering:

- K-mean clustering (the number of classes is fixed in advance): Elbow method, Average silhouette method, Gap statistic method, and Consensus-based algorithm.
- Hierarchical clustering (an unknown number of classes and helps to determine this optimal number): Hierarchical agglomerative clustering and Hierarchical agglomerative clustering using "average" linkage (This produces cluster dendrograms).

For this reason, *k*-means is considered as a supervised technique, while hierarchical clustering is considered as an unsupervised technique because the estimation of the number of clusters is part of the algorithm (Antoine Soetewey, 2020).

The R function *kmeans* (cluster package) was used to compute k-means, and the function *fviz_cluster* (*factoextra* package) to visualize the results (The description of this approach is in Table 1). The *kmeans* function has a nstart option that attempts multiple initial configurations and reports on the best output. For example, I added nstart = 25 to generate 25 initial structures. In the literature, this approach is often recommended by researchers. Iapplied a Screen Plot to illustrate the variabilities in clusters and cluster mapping to illustrate data points by the first two principal components that explained most of the variance.

In cluster analysis, the following four methods were used as clustering validation and criteria to find the optimal number of clusters for a *k*-means:

- The Elbow method looks at the total within-cluster sum of square (WSS) as a function of the number of clusters. The location of a knee in the plot is usually considered an indicator of the appropriate number of clusters. This bend suggests that additional clusters beyond the third have little value because it means that adding another cluster does not improve the partition. The higher the percentage, the better the score (and thus the quality) because *BSS* is large and *WSS* is small.

- The Silhouette method measures clustering quality and determines how well each point lies within its cluster. Average silhouette method (function *fviznbclust)* provides the silhouette coefficient (silhouette width) of observations for different values of k and measures the quality of a clustering (values more significant than 0.5 are desirable and means that the observation is well clustered). A high average silhouette width indicates how well each observation lies within its cluster.
- Gap statistic method (function *clusGap*) provides the gap statistic and standard error for an output. This approach can be utilized in any clustering method. The gap statistic compares the total intracluster variation for different values of k with their expected values under the null reference distribution of the data.
- Consensus-based algorithm (function *n clusters*) function allows to run many methods and take the number of clusters that is the most agreed upon (i.e., find the consensus).

Latent Class Analysis (LCA)

The completed steps of computing the algorithm are described as follows using previous work of mixture models: latent profile and latent class analysis by DL Oberski (Oberski, 2016):

1. Define a LCA model with the original database.
2. Fit a model for a specified number of classes. Select nclass as the initial model with two classes and then add classes until identifying the model with the best fit (Repeat).
3. Report results for fit statistical criteria in a table (*log-likelihood* =LL, Akaike'ss information criterion (AIC), and Bayesian information criterion (BIC)
4. Report validation and diagnostic criteria (most minor class count and size, entropy of fitted latent class models, and Predicted class memberships (ALCPP)).
5. We set fit statistical criteria and diagnostic criteria to look for more evidence that the chosen solution was the correct one.

Fit statistical criteria

- Log-likelihood (LL) based measure of "unexplained information" in a model, where smaller values are preferred. Akaike's information criterion (AIC) for evaluating how well a model fits the data it was generated from.
- Bayesian information criterion (BIC) is a method for scoring and selecting a mode and it is considered the most reliable fit statistic in LCA.

Validation and diagnostic criteria

- Smallest class count (n) and smallest class size (%).
- Entropy measures a random variable "information" or" "uncertainty"" of everyone to be assigned to the class. The entropy of a fitted latent class model is described as a characteristic, not a model selection criterion. But it is essential for consideration.
- average latent class posterior probability (Predicted class memberships by posterior modal probability (ALCPP)).

A full review of fit criteria can be found in other publications (Weller et al., 2020).

Within-category models and model comparisons

The model comparisons in this analysis are between k-means clustering and Latent Class Analysis (LCA) model. To make fair comparisons between those models, I first arrived at the best fitting model within each category and applied validation fit criteria to each model.

Data: cluster analysis and LCA, I applied to three different datasets: SABE original (n=23,894), biomarkers subsample with imputation values (n=4,092) and case complete dataset (excluding NA n=2812).

3) Results

This section is structured as follows:

1. Identification clusters of adversities in childhood
2. Identification Latent Class Profiles of adversity childhood experiences

Identification clusters of adversities in childhood

K-means and cross-validation of clustering quality are illustrated below.

K-means clustering

The methods used as clustering validation and criteria for selecting the clusters using subsample of biomarkers are presented in Table 2. However, this approach was applied to the SABE original and 4,092 biomarker datasets after imputation.

Table 2. The optimal number of clusters for a *k*-means is based on four cross-validation methods (n=4,092).

Number of clusters/ methods	Within cluster sum of squares by cluster	Silhouette coefficient (silhouette width)	Gap Statistic Method	Consensus-based algorithm
Cluster 2	17.5%	**0.63**		
Cluster 3	33.6%	**0.63**		**Optimal number (50% consensus)**
Cluster 4	48.7%	0.59		
Cluster 5	60.5%	0.58		
Cluster 6	**71.9 %**	0.55	**Optimal number**	
Cluster 7	63.6 %	0.55		
Cluster 8	66.1 %	0.56		
Cluster 9	66.6 %	0.55		
Cluster 10	69.4 %	0.55		

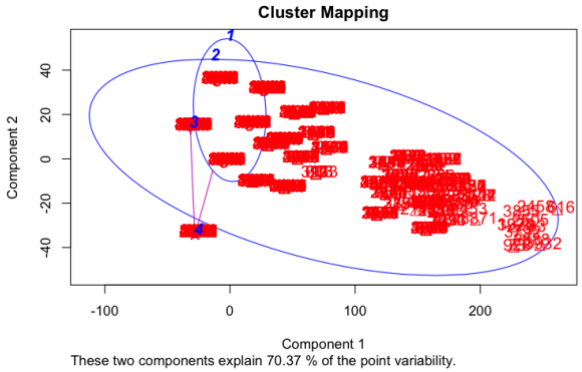
Interpretation of the outcome

- Here, the four approaches suggest a different number of clusters. The results indicate that 2, 3, or 6 are the optimal numbers of clusters; as we can see, these methods do not necessarily lead to the same result, but number of clusters were consistent between three datasets.
- The screen plot shows how to increase the number of clusters within-group sum of squares came down. So, in this ideal data number of clusters should be 3.
- Elbow method suggests 3 clusters (71.9%) in the three datasets.
- The Silhouette method suggests 2 or 3 well-separated clusters (obtained the highest clustering silhouette plot width: 0.63). Silhouette plot, which shows that our clustering using two, three, four, or six groups is good because there's no negative silhouette width, and most of the values are more significant than 0.5. Clustering silhouette plot width using four groups: 0.59 and six groups 0.55. Using SABE original dataset this ranging between 0.52 (2 clusters) to 0.64 (six clusters).
- The gap statistic comparisons of the total intracluster variation suggest that 6 is the optimal number of clusters.
- Based on a consensus-based algorithm, most methods suggest retaining 3 clusters, followed by a 2-clusters solution is the optimal choice of the cluster to retain. The selection of 3 clusters is supported by 14 (50.00%) methods out of 28 (Silhouette, Gap_Maechler2012, Gap_Dudoit2002, Hartigan, Scott, Marriot, trcovw, Tracew, DB, Duda, Pseudot2, Beale, Ball, Sdindex).

Figure 1 shows the number of clusters under the supervised techniques using different datasets and quality of clustering using Elbow, Silhouette, Gap Statistic methods and consensus algorithm.

Principal Component Analysis (PCA) is a helpful technique for exploratory data analysis that allows to visualize better the variation present in a dataset with many variables. Figure 2 shows PCA representing the variables in a two dimensions plane by each optimal number of clusters.

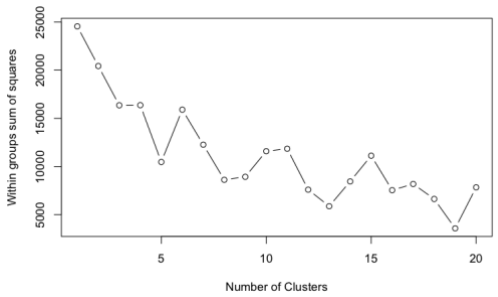
Figure 3. Cluster silhouette plot (Average silhouette width: 0.5)



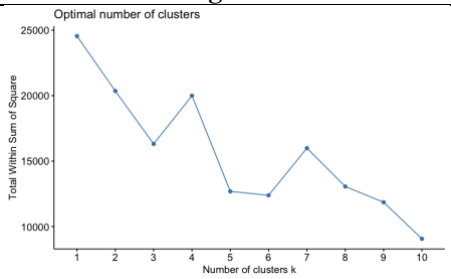
I also performed cluster plots based on the optimal choices of cluster to retain and *Hierarchical clustering analysis (HCLA) – Dendrograms* in the cross-validation methods to explore unknown number of classes and helps to determine this optimal number and revealed two and three clusters. This is a consistent finding across four cross-validation

methods in K-means clustering analysis: Silhouette coefficient (2 clusters) and Elbow method and Consensus-based algorithm (three clusters). The obtained dendrogram of the cluster structure for both methods across three datasets are not included in this document.

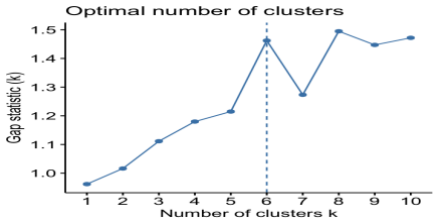
Figure 2. number of clusters under the supervised techniques: Screen plot and quality of clustering using Elbow, Silhouette, Gap Statistic methods and consensus algorithm.



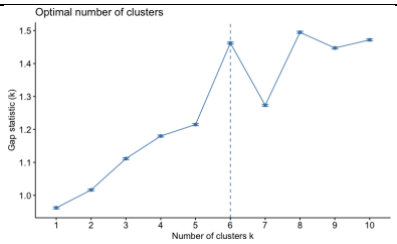
Screen plot of the number of clusters within groups



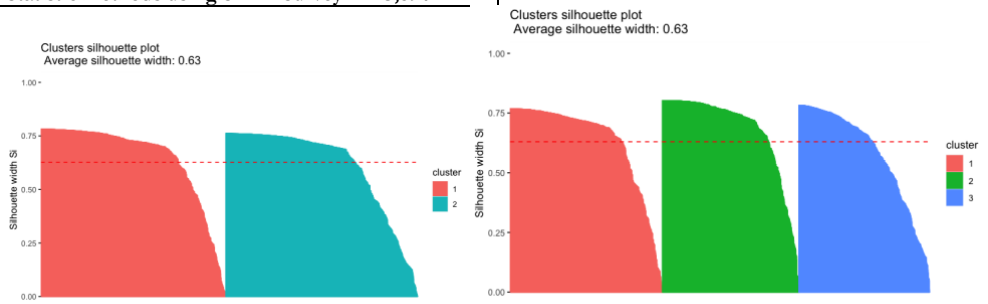
Elbow method using SABE Survey n=23,894



Gap Statistic methods using SABE Survey n=23,894

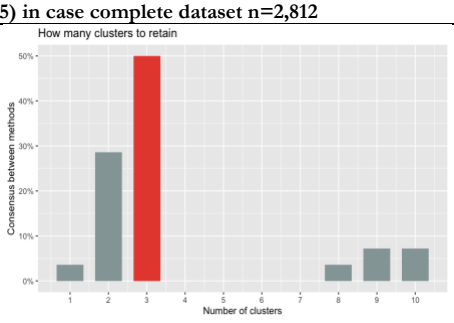


Gap Statistic methods using case complete dataset n=2,812



Cluster silhouette plot (Average silhouette width: 0.5) in case complete dataset n=2,812

SABE Survey n=23,894



Case complete dataset n=2,812

Identification Latent Class Profiles of Adversity Childhood Experiences

Criteria used for selecting a class model

Table 3 presents LCA results for different class models. The lowest LL, AIC, and BIC are shown in boldface, which indicates the model met the fit criteria. This includes the following descriptive fit indices: log-likelihood (LL); Akaike information criterion (AIC); Bayesian information criterion (BIC); Smallest class count (n) and size (%), entropy, and average latent class posterior probability (Predicted class memberships by posterior modal probability (ALCPP)). I run same approach in the three datasets (original sample, imputed biomarkers dataset and case complete dataset) with same variables.

The BIC was used to select the best laten model following evidence in the literature. However, I looked for more evidence that the chosen solution was the correct one. Here, we can also compare the values of the Akaike information criterion (AIC), which was the highest in the two-class model. Although entropy is not used to select a final model as it is more than a characteristic, it is essential to note the three-class model had the highest entropy (.65) but not an adequate entropy as suggested (above the cut-off of .80).

Table 3. Evaluating Latent Class Solutions

SABE original sample (n=23,984)

	Model fit criteria					Diagnostic criteria			
Models	LL	AIC	BIC	Chi-square goodness of fit	Likelihood ratio/deviance statistic	Largest class size (%)	Smallest class size (%)	Entropy	Predicted class memberships (ALCPP)
2 classes	**-6838.24**	13702.48	**13779.72**	72.64585	71.62059	0.7593	0.2407	0.512635	**0.8457**
3 Class	-6825.156	**13690.31**	13809.14	49.35696	45.45113	0.5182	0.2156	0.3612357	0.7905
4 Class	-6821.377	13696.75	13857.18	36.41236	37.8943	0.5413	0.1337	0.4312546	0.707
5 Class	-6818.145	13704.29	13906.31	31.43054	29.84655	0.683	0.0548	0.6267687	0.7255
6 Class	-6816.791	13715.58	13959.19	31.46284	28.72117	0.4602	0.0368	0.5291593	0.6419
7 Class	-6815.38	13726.76	14011.96	22.84352	25.89941	0.365	0.0358	0.453039	0.6366
8 Class	-6815.542	13741.08	14067.88	25.35561	26.2249	0.255	0.0204	0.400905	0.5011
9 Class	-6812.812	13749.62	14118.01	17.11802	20.76419	0.1772	0.0211	0.5068353	0.3738
10 Class	-6812.756	13763.51	14173.49	18.61296	20.65259	0.1476	0.0146	0.4419291	0.4001

Imputed dataset of biomarkers (n=4,092)

	Model fit criteria			Diagnostic criteria			
Models	LL	AIC	BIC	Smallest class count (n)	Smallest class size (%)	Entropy	Predicted class memberships (ALCPP)
2 Class	**-9679.634**	**19385.27**	**19467.39**	1,023	0.2548	0.4321435	**0.8724**
3 Class	-9657.111	19354.22	19480.56	531	0.1266	**0.6536626**	0.4888
4 Class	-9643.247	19340.49	19511.05	368	0.0936	0.3997616	0.5694
5 Class	-9642.421	19352.84	19567.61	450	0.1099	0.3459344	0.4651
6 Class	-9637.952	19357.9	19616.89	122	0.0348	0.4713346	0.5887

LCA models using case complete dataset (n=2,812)

Model	log-likelihood	resid. df	BIC	aBIC	cAIC	likelihood-ratio	Entropy
Model 1	-6992.206	57	14032.06	14013.00	14038.06	379.55185	-
Model 2	-6838.240	50	**13779.72**	**13738.42**	13792.72	71.62059	0.388
Model 3	-6824.507	43	13807.85	13744.30	13827.85	44.15455	0.493
Model 4	-6821.222	36	13856.87	13771.08	13883.87	37.58328	0.39
Model 5	-6817.721	29	13905.46	13797.43	13939.46	30.58211	0.469
Model 6	-6815.666	22	13956.94	13826.67	13997.94	26.47166	0.46

Interpretation of the outcome:

- The BIC suggests two well-separated class models. This result was consistent in the three datasets. In this case, 2-class model seems to be the best option for our data.
- In the LCA models, Model 2 shows the lower BIC in the case complete dataset.
- The model became unstable with the 6-class and 7-class model in the case complete dataset and imputed biomarkers dataset, respectively. In the SABE original dataset, 10-class model has negative degrees of freedom, which means that is not an acceptable model.

With most of these approaches suggesting 2 as the number of optimal classes, a final analysis was performed to extract the results using 6-class. Figure 4 present different models and then compare them.

Figure 4. LCA models

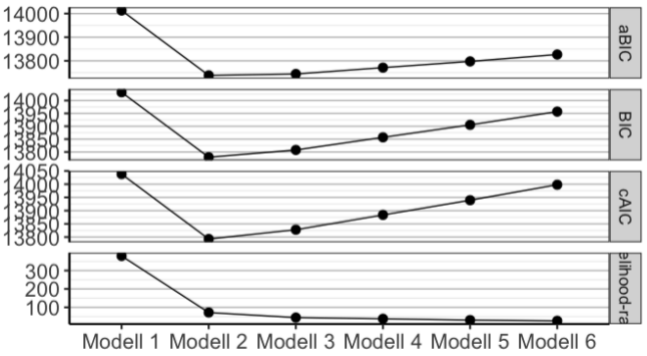
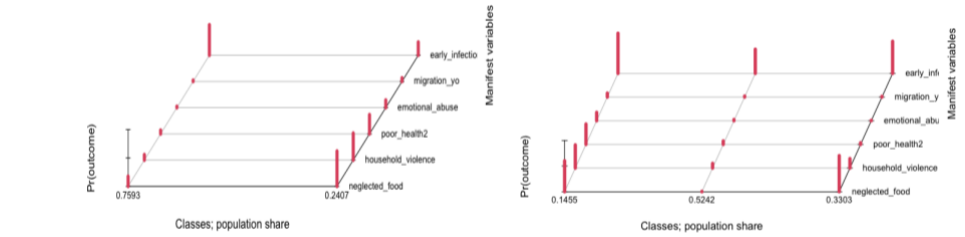
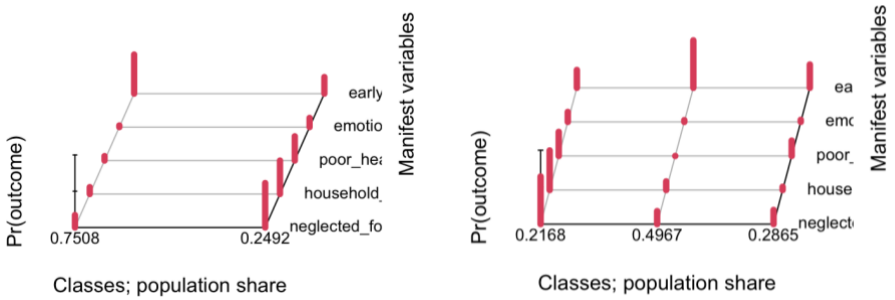


Figure 5 illustrates a profile plot of poLCA for the two-class and three-class models showing the estimated distribution of the six observed variables within each model. Each group of red bars represents the conditional probabilities, by latent class, of being rated positively by each of the six childhood adversities (labelled on the right side). Taller bars correspond to conditional probabilities closer to 1 of a positive rating. The two estimated latent classes correspond to a pair of types consistently rated as one positive (25%) and negative (74%) class representing 100% of the sample.

Figure 5: Estimation of the basic latent class model using the SABE data (poLCA function).



PolCA using complete SABE original sample



By observing the graph, we can see those participants in the group 2 (76%), have less adversities in the childhood (lower risk group). In contrasts, participants in the group 1 (24% of the subsample) reported greater adversities in childhood (high-risk group).

Common characteristics among groups:

- Low emotional abuse and migration
- Early-life infection

Class 1 characteristics (Low-risk childhood adversity)

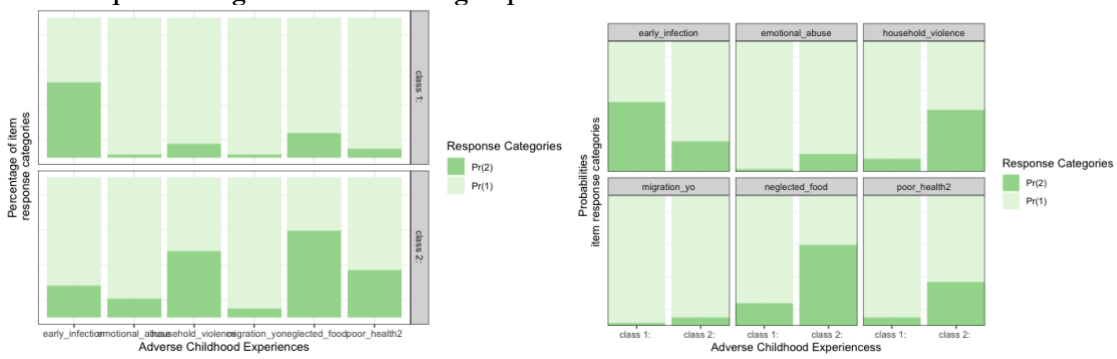
- Low emotional abuse and migration. High probability food security, good health, and not has witnessed domestic violence between parents. Conversely, high probability has had a serious infection.
- Within this group, early-life infection, household violence, and poor health contributed with higher adversity in childhood.

Class 2 characteristics (High-risk childhood adversity)

- Poor food environment, violent intimate home environment, and illness for 30 days (poor health).
- Low emotional abuse, migration, and early-life infection.

One-third of women and one- quarter of men were assigned to group 1. In both women and men, the probability of belonging to group 2 was the greatest (41% of women and 48% of men).

Figure 6. Response categories across latent groups.



Most older people (74%) were in the Promotive Factors class (Lower adversity class). This class had the lowest probability of food insecurity, household violence, poor health, emotional abuse, migration, and early-life infection tend to be positive emotional abuse, migration, and poor health. Conversely, a small percentage of the sample (26%) were in the Limited Access to Promotive Factors class (higher adversity group). This class had a higher probability of adverse childhood experiences.

The two-latent class model corresponds to three classes based on selected ACEs. Most older people (53%) were in the Promotive Factors class (lower adversity group, except for a higher probability of early-life infection), and the second class (33%) were in the intermediate with higher chances of food insecurity and early-life infection history. On the other hand, a small percentage (14%) of the sample was in the Limited Access to Promotive Factors class. This class had a higher probability of food insecurity, household violence, poor health, emotional abuse, migration, early-life infection tends to be positive, emotional abuse, migration, and poor health. The early-life infection had similar profiles across the two- and three-class models.

Extracting Results from cluster and LCA models

Table 4. Comparison methods and identification of best fist model

Methods/Dataset	SABE original sample n=23,892	Imputed biomarkers subsample n=4,092	Dataset excluding NA n=28,12
Cluster analysis	**Silhouette method= 2-3 clusters** Elbow method=3 clusters Gap method= 6 clusters Consensus methods=3 clusters	**Silhouette method= 2-3 clusters** Elbow method=3 clusters Gap method=6 clusters Consensus methods=3 clusters	**Silhouette method= 2-3 clusters** Elbow method=3 clusters Gap method=6 clusters Consensus methods=3 clusters
LCA	**2 classes**	**2 classes** High-risk group (n=1,023) Low-risk group (3,069)	**2 classes** High-risk group (n=434) Low-risk group (2,318)

Table 5 presents age and sex characteristic of the selected 2-class: low and high-risk groups. These two groups are included in the *dfbioage* as new variable in the descriptive analysis and biological age analysis.

Table 5. Selection of two latent groups in the datasets

	Low-risk childhood adversity **(N=2378)**	High-risk childhood adversity **(N=434)**	Overall **(N=2812)**
Age Mean (SD)	68.9 (6.97)	68.0 (6.59)	68.8 (6.92)
Age Median [Min, Max]	68.0 [60.0, 101]	66.5 [60.0, 92.0]	67.0 [60.0, 101]
female	1297 (54.5%)	236 (54.4%)	1533 (54.5%)
male	1081 (45.5%)	198 (45.6%)	1279 (45.5%)

Conclusions

In cluster analysis, the optimal numbers of clusters ranged from 2-6 across four applied approaches. The LCA revealed that 2-class as in the model fit criteria as fit diagnostic criteria in the three datasets. A lower value for BIC between models and datasets was consistent and allowed to pick the best one out. So based on the comparison methods and BIC value, we would choose model 2 in the LCA with two latent groups for our SABE aging data. The PolCA R package facilitates the identification of Latent Class Profiles of Adversity Childhood Experiences and tests different criteria to validate methods. Furthermore, I run models on the three types of SABE data as a sensitivity analysis.

References

Alboukadel Kassambara, & Fabia Mundt. (2022). *CRAN - Package factoextra. factoextra: Extract and Visualize the Results of Multivariate Data Analys*. https://cran.r-project.org/web/packages/factoextra/index.html

Antoine Soetewey. (2020). *The complete guide to clustering analysis: k-means and hierarchical clustering by hand and in R - Stats and R*. https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/

Collins, L. M., & Lanza, S. T. (2010). Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*, 1–295. https://doi.org/10.1002/9780470567333

Finite Mixture Modeling. (2006). *Finite Mixture and Markov Switching Models*, 1–23. https://doi.org/10.1007/978-0-387-35768-3_1

Lanoue, M. D., George, B. J., Helitzer, D. L., & Keith, S. W. (2020). Contrasting cumulative risk and multiple individual risk models of the relationship between Adverse Childhood Experiences (ACEs) and adult health outcomes. *BMC Medical Research Methodology*, *20*(1), 1–10. https://doi.org/10.1186/s12874-020-01120-w

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, *42*(10), 1–29. https://doi.org/10.18637/JSS.V042.I10

Martin Maechler, & Peter Rouseeuw. (n.d.). *cluster: 'Finding Groups in Data': Cluster Analysis Extended Rousseeuw et aCRAN - Package cluster*. Retrieved 24 June 2022, from https://cran.r-project.org/web/packages/cluster/index.html

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440–461. https://doi.org/10.1037/TPS0000176

Oberski, D. (2016). *Mixture Models: Latent Profile and Latent Class Analysis*. 275–287. https://doi.org/10.1007/978-3-319-26633-6_12

Pamulaparty, L., Rao, C., & Rao, M. (2016). Cluster Analysis of Medical Research Data using R. *Global Journal of Computer Science and Technology*, *16*(1).

van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. *Flexible Imputation of Missing Data, Second Edition*. https://doi.org/10.1201/9780429492259/FLEXIBLE-IMPUTATION-MISSING-DATA-STEF-VAN-BUUREN

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *46*(4), 287–311. https://doi.org/10.1177/0095798420930932