METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Clustering Longitudinal Data Using R: A Monte Carlo Study

Peter Verboon[1], Ron Pat-El[1]

[1] *Faculty of Psychology, Open University, Heerlen, The Netherlands.*

Methodology, 2022, Vol. 18(2), 144–163, https://doi.org/10.5964/meth.7143

Received: 2021-07-14 • **Accepted:** 2022-04-20 • **Published (VoR):** 2022-06-30

Handling Editor: Jost Reinecke, Bielefeld University, Bielefeld, Germany

Corresponding Author: Peter Verboon, Faculty of Psychology, Open University, P.O. Box 2960, 6401 DL Heerlen, the Netherlands. E-mail: peter.verboon@ou.nl

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

The analysis of change within subjects over time is an ever more important research topic. Besides modelling the individual trajectories, a related aim is to identify clusters of subjects within these trajectories. Various methods for analyzing these longitudinal trajectories have been proposed. In this paper we investigate the performance of three different methods under various conditions in a Monte Carlo study. The first method is based on the non-parametric k-means algorithm. The second is a latent class mixture model, and the third a method based on the analysis of change indices. All methods are available in R. Results show that the k-means method performs consistently well in recovering the known clustering structure. The mixture model method performs reasonably well, but the change indices method has problems with smaller data sets.

Keywords

longitudinal, clustering, Monte Carlo, change, trajectories, k-means, latent class, mixture model, R

The analysis of change in individuals and the development in time of groups of individuals is important in many research fields. With the emergence, or rather increased popularity of intensive longitudinal designs such as the Experience Sampling Method (ESM) (Bolger & Laurenceau, 2013; Hektner et al., 2007), statistical methods to analyze change over time have also become popular. Longitudinal data analysis concerns the analysis of change over time and despite differences between methods such as time

series or repeated measures, generally all such methods hold that the analysis of change over time is the analysis of the trajectory of growth.

Longitudinal data may be derived from experimental or observational studies. In longitudinal experiments differential between-subjects effects over time are usually the primary focus, and the level or the shape of the trajectory is often secondary. In observational studies the level or shape of the trajectory is often central to the analyses and comparisons between realistic groups (e.g. men versus women, educational level) of secondary importance. Another type of question is about differences between groups that were not defined beforehand, but are derived from the data (Muthén & Muthén, 2000). Such groups are called latent classes or clusters. A latent cluster consists of a homogeneous group of individuals and the grouping in longitudinal analyses can be based on shared levels and shapes of trajectories.

To analyze change a construct must be measured repeatedly, which will yield a set of trajectories, one for each subject. Other than in more general multivariate data, the repeated measures in longitudinal data have a dependence among measurements due to the ordering in time and therefore traditional regression techniques cannot be applied, since they assume independent observations. Another aspect in methods popular in the social sciences such as ESM is that this type of longitudinal data also differs from time series data because instead of a few random processes uniformly sampled over time, the more general longitudinal data consists of a large number of independent trajectories that are potentially irregularly sampled over time. Although it is possible to aggregate the individual change over time to a mean change over time, it is also possible to analyze the differential trajectories of growth between individuals in order to, for example, identify subgroups for which an intervention is successful.

With such longitudinal data the research questions are about within-subjects change such as: what is the individual and general level of change in the construct of interest, and what pattern or shape does this change have? But questions can also concern the between-subjects differences in change, such as: do some (groups of) individuals have a different level or pattern of change. The patterns of change can be captured in a functional form (parametric), usually a linear or quadratic pattern, but non-parametric patterns that describe the trajectories are also possible. When questions about the level and patterns of change are answered, the next level of interest lies in predicting these change levels or patterns by covariates. Vermunt (2010) provides an overview of various approaches and proposes a three-step approach that combines the clustering of growth trajectories via model-based methods and applying the predicted cluster-memberships as a dependent variable in a multinomial logistic analysis.

One way to analyze growth trajectories is to cluster them into partitions that reflect different trajectories of growth within a population (Muthén & Muthén, 2000). Several techniques have been developed for the clustering of longitudinal data, each with their strengths and weaknesses (Magidson & Vermunt, 2002), but so far it hasn't become clear

which technique should be preferred in a given context (Den Teuling et al., 2021; Martin & von Oertzen, 2015; Twisk & Hoekstra, 2012). The focus of such longitudinal clustering techniques is very similar to that of regular clustering techniques. If the research aim is to *find* groups or clusters in the longitudinal data, the aim of using a clustering method is to find between-person differences in within-person changes over time. There are many ways to obtain these clusters. Modeling the heterogeneity in growth trajectories is often more interesting than fitting a single model for the whole sample. Models that derive clusters or classes from the data are called latent class models.

Given the increased focus on intensive longitudinal design there is to be expected that researchers used to standard statistical software packages shift their attention to more flexible platforms for such data analysis such as the statistical programming language R (R-Core-Team, 2020). R has become very popular in various scientific fields (Lai et al., 2019) and due to being an open source platform can be adapted quickly to innovations in statistical methods. A global community of contributors constantly add to or improve statistical packages which are distributed through CRAN (Comprehensive R Archive Network, https://cran.r-project.org/). As such several packages have been developed and distributed that provide a library of functions specific to longitudinal cluster analysis. For applied researchers facing a longitudinal clustering situation, information on the quality and possibilities of different packages for their research problem may be very helpful.

In this article we will focus on three popular methods for longitudinal cluster analysis that are available in R (Version 4.1.0), reflecting different methods for clustering longitudinal data. These methods are: (1) the *k*-means method (*kml*); (2) the trajectory-method (*traj*); and (3) the latent-class mixed-model method (*lcmm*). These methods will be described below.

Longitudinal Clustering Methods

Generally statistical models that yield as outcomes an overall fit line and a distribution of fitted lines in longitudinal data are called latent growth models (LGM) in the context of Structural Equation Modeling (SEM). A fundamental concept in SEM is the modelling of factors that are *latent* which means they are not directly observed but are inferred through a mathematical model from observed variables. Often this relates to latent variables as a kind of (factor analytically) weighted average of several measured responses. In LGM however, the latent variables often refer to the intercept (ic) and the slope (sl) of the fitted linear growth pattern, see e.g. Berlin et al. (2014). But other growth patterns can also be modelled.

Another variant of this type of model is the Growth Mixture Model (GMM), where the term mixture refers to it being a mixture of (latent) growth models. GMM is a framework within the multilevel modeling (MLM) literature, which approaches growth

through the separation of variance into fixed and random effects. In longitudinal data a fixed effect assumes that the model intercept is time-invariant, and a random effect allows for testing whether the intercept is likely not time-invariant, i.e. suggests growth. Despite the different approaches to longitudinal data between LGM and GMM, from a comparison of the basic equations underlying both models, it can be seen that these two types of models are in their basic form essentially the same, see e.g. Hox et al. (2018) and Singer and Willett (2003), where the latent variables in SEM are analogous to the random effects in GMM.

The latent class growth model (LCGM) and GMM are closely related, where LCGM is a special type of GMM (e.g. Berlin et al., 2014). For each individual a probability is computed that indicates to which cluster (or class) an individual belongs. These clusters are not directly observed and are therefore considered latent in the LCGM, inferred through patterns in the data. With these latent clusters the variance and covariance estimates for the growth factors within each cluster are assumed to be fixed at zero. Because of this explicitly modelled assumption, all individual growth trajectories within a cluster are homogeneous. This assumption can be relaxed, however (Jung & Wickrama, 2008). The model that allows for variation within clusters is known under various names, among which GMM. An excellent overview of the differences between the model-based methods is given by van der Nest et al. (2020).

K-Means for Longitudinal Data: The kml Method

The first method to explore is the *k*-means for longitudinal data (*kml*) method from the *kml* R-package (Genolini et al., 2016, 2015) that is based on the *k*-means algorithm adapted for longitudinal trajectories. Like 'classic' *k*-means the *k*-means for longitudinal data is a partitional clustering method in which *k* clusters are specified, and an algorithm tries to partition the data in such a way that it minimizes within-cluster variance. At setup a user specifies the number of clusters to be identified in the data, which are the *k* in *k*-means. *K*-means is an algorithm which alternates between two steps. The initial step in the algorithm is to randomly assign each observation to a cluster in a given (fixed) number of clusters. The algorithm then optimizes the clustering solution by alternating between the two steps. In the first step, the centers of each cluster are computed. The second step consists in assigning each observation to its "nearest cluster." The alternation of the two steps is repeated until no further changes occur in the clusters or until the maximum number of iterations is reached (the default is 200). However, since *k*-means algorithms may converge to a local minimum, several starting values (the default is 20) are automatically applied in the *kml* method in order to increase the probability of obtaining a globally optimal solution. In order to determine the distance between an observation and a cluster center, the *k*-means algorithm uses a distance metric, such as the Euclidean distance or the Manhattan distance. There is no fundamental difference between classic *k*-means and longitudinal *k*-means, because no restrictions (e.g. linear

growth) on the trajectories are imposed. In longitudinal *k*-means the time points serve as the variables in classic *k*-means. The overall distance between two subjects takes the distances at all time points equally into account. However, some of the imputation methods available in this package are based on the longitudinal character of the data.

The *kml* R-package (Version 2.4.1) offers eight quality criteria, which help to select the most plausible number of clusters. These criteria are computed such that higher scores refer to relatively better solutions. It is up to the researcher to determine the number of clusters, based on both these criteria and the substantive knowledge of the longitudinal process. The *kml* method also offers eleven different imputation methods, which are helpful since substantial dropout in social sciences research is a common phenomenon. Furthermore, *kml* provides various types of distances (such as the Euclidean, which is the default, or the Manhattan) as criteria for assigning subjects to clusters, but distance functions can also be defined by the user. The choice of the initial configuration of the clusters may be important because a "good" start speeds up the computation and may lead to a better solution. There are various strategies available to set the initialization, see Genolini et al. (2015). In this study we have used the *kml* function with all the default options.

The traj Method

The second non-parametric method, implemented in the R-package *traj* (Version 1.2) (Leffondré et al., 2004; Sylvestre et al., 2006; Sylvestre & Vatnik, 2014), is a stepwise method that consists of three steps. In the first step 24 change indices are computed from the data, which form a new matrix of *N* subjects by 24 indices. These indices are statistical measures that according to Leffondré et al. (2004, p. 1050) "assess different aspects of the longitudinal pattern of change in an individual, that can discriminate between stable–unstable, increasing–decreasing, linear–nonlinear, and monotonic–nonmonotonic patterns, patterns, as well as identifying patients who tend to have abrupt changes." The indices are divided into four classes: elementary measures of change, measures of nonlinearity and inconsistency of change, measures sensitive to nonmonotonicity and to abrupt short-term fluctuations, and measures contrasting early versus later change. If two indices in this matrix correlate highly (> .95) with each other one of them is removed. The remainder of this matrix with change indices is analyzed by a principal component analysis to obtain a subset of the indices that describe the most important features of the trajectories. This subset is subsequently used in the third step to cluster the trajectories, using *k*-means clustering. So, the third step is a classic *k*-means, however not on the data themselves but on derivatives of the data as explained in step one (change indices) and two (PCA). Also, in this *kml* step various distances used to obtain the cluster solution can be chosen (the default is the Euclidean), and also the maximum number of iterations, the number of starting configurations (the default is 50), and the initialization of the starting configuration. Either a fixed number of clusters can be

specified or the program suggests the optimal number of clusters based on up to 30 different quality criteria (Leffondré et al., 2004). In this study all default options were taken for the second and third step. The first step has no choices to be made by the user.

The lcmm Method

The third method is the *lcmm* method (Proust-Lima et al., 2017, 2020). This is a model-based method that estimates the longitudinal growth patterns and the latent clusters using linear mixed model theory (Jung & Wickrama, 2008; van der Nest et al., 2020). This method comes under various names, such as growth mixture model (GMM), and belongs to a broader class of longitudinal growth models, for a recent overview, see van der Nest et al. (2020). The package used here is able to analyze a wide range of models, but for the purpose of this study we have focused on latent class mixture models or GMM for Gaussian longitudinal outcomes. The latent class mixture model (*lcmm*) assumes that the population is heterogeneous and composed of a fixed number of latent clusters of subjects characterized by their trajectories. The trajectories are modelled by a cluster-specific linear mixed model. The estimation in this method is based on maximum likelihood and goodness-of-fit measures for the estimated models are available. Models with different numbers of clusters (latent clusters) can thus be compared by inspecting the goodness-of-fit indices. This model is highly flexible since the fixed effect and the distribution of the random effects can be specified for each cluster. Predictors can be added to the model to find latent clusters and to find the parameters to optimally predict the trajectory of the dependent variable. The predictors in both steps may be different ones. Like the other methods, the result may depend on the starting values provided to the algorithm.

In this study the function 'hlme' from the *lcmm* package (Version 1.9.5) was used, where the variable time, indicating the time points, is used as a predictor for the fixed and the random part across the subjects. Since some clusters are define by quadratic growth, also the quadratic term of time has been added to fixed and random part of the model. No covariates were specified to predict latent cluster membership. The linear link function (default) is used here for Gaussian outcomes. No a-priori starting values for the clusters were given. The maximum number of iterations used was 100, the default value. For all other parameters the default values were taken, see Proust-Lima et al. (2020).

This Study: Research Question

We have given a very brief overview of three packages for longitudinal cluster analysis in R. There is little knowledge about how these three methods perform relative to each other and which method might be preferred by researchers interested in longitudinal clustering of their data. As such the aim of this study is to compare the quality of clustering solutions between these three methods in R, in order to identify the strengths

and weaknesses of each method and help practitioners in making an informed choice among these methods.

In order to compare the methods our research question is: do the three methods, namely *kml*, *traj* and *lcmm*, differ in their ability of recovering clustering solutions depending on (a) the amount of (longitudinal) time points, (b) the number of participants, (c) the number of clusters, and (d) the homogeneity of the clusters, i.e. within-cluster variance (error).

This study is not the first one that compares different longitudinal clustering methods. For instance, Twisk and Hoekstra (2012) compared five methods, including *k*-means and model-based models. The methods were compared on two real data sets (one of them was manipulated) and also the optimal number of clusters was a point of interest. They found that latent class growth analysis (LCGA) performed the best when classifying linear growth trajectories, but neither *k*-means nor one of the latent class methods were good at classifying trajectories that were both linear and quadratic. Similarly, Feldman et al. (2009) compared latent class methods, but did not compare the performance of these methods to other types of cluster analyses. The study by Martin and von Oertzen (2015) did compare *k*-means with model-based approaches in a Monte Carlo (MC) simulation. They expected that model-based methods would underperform relative to other methods in small sample sizes (N < 500). They not only evaluated the performance of models with the optimal number of clusters, but also how the models compared in choosing the number of clusters. Overall, they concluded that mixture-models outperformed other methods, including *k*-means. But only linear trends were evaluated and as such it is unclear how these methods would compare in classifying more complex growth patterns. A recent study by Den Teuling et al. (2021) addresses some of the limitations in these earlier studies. They compared longitudinal *k*-means to various mixture models, including a combination of *k*-means and a mixed-effect model. Specifically, they expanded on previous studies by simulating group trajectories that smoothly and slowly change over time instead of linear growth. In their simulations growth mixture modeling (GMM) and the two-step clustering approach (combined *k*-means and GMM) significantly outperformed the other methods across all scenarios, both in terms of group assignment and estimation of the group trajectories. Our study contributes to the existing literature by comparing longitudinal *k*-means and model-based clustering to a method that has not received much attention, namely the *traj*-method. In addition, we also evaluate the performance of these methods when classifying not only linear or slightly curved growth trajectories, but quadratic trajectories as well.

Method

Design of the Monte Carlo Study

In order to test the different longitudinal clustering methods data were simulated using a Monte Carlo method. The data generating procedure was to first define the number of clusters (M) that characterize the type of longitudinal change to be modelled into the data. Either three or six clusters (M = 3, 6) with distinct longitudinal patterns (trajectories) were specified. The three trajectories can be described as (1) stable low; (2) linear growth; (3) quadratic: first decline, then increase. For the generation of six clusters the following three trajectories were added to the previous three: (4) stable high; (5) linear decline; (6) quadratic: first increase, then decline.

More specifically, these are the trajectories of the six clusters

$$(\mathrm{C}_i, i = 1, .., 6)$$

- stable low:

$$C_1 = 0t + 2$$

- linear growth:

$$C_2 = 0.5t$$

- quadratic decline and increase:

$$C_3 = (t - [t_{max}/2])^2$$

- stable high:

$$C_4 = 0t + 4$$

- linear decline:

$$C_5 = -0.5t + 5.5$$

- quadratic increase and decline:

$$C_6 = 5 - 5\frac{C_3}{max(C_3)}$$

Where t means the fictitious time point starting at 1 for the first measurement and t_{max} stands for the maximum number of time points.

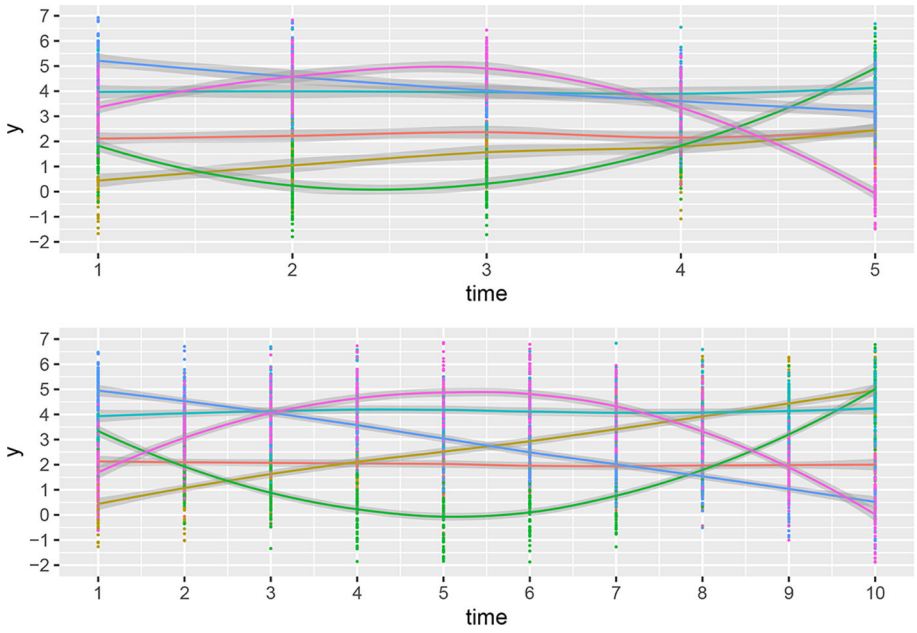
The Monte Carlo simulation subsequently varied the data sets on a fixed number of repeated measurements or time points ($T = 5, 10$), and the number of simulated subjects ($N = 25, 50$). So the N refers to the number of subjects at each time point in each cluster, yielding a total of data points in each data set of $T \times M \times N$. For all time points and subjects error (E) was added to the data, taken from a normal distribution with either $M = 0$ and $SD = 0.5$ ($E = \sim N(0, 0.5)$) for the low measurement error condition, or $E = \sim N(0, 1)$ for the high measurement error condition to simulate differences in cluster homogeneity.

The total number of cells in the simulated design were 2 (clusters) x 2 (time points) x 2 (subjects) x 2 (levels of error) = 16. For each method a new series of Monte Carlo simulated datasets were constructed, bringing the total number of cells in the design when also accounting for the clustering methods to 48. The number of replicated data sets per cell was $N = 500$. The default of 20 random starting values was used for the *kml* method for each analysis. The R-package *MonteCarlo* (Leschinski, 2019) (Version 1.0.6) was used for this study, which simplifies Monte Carlo simulation studies by automatically setting up loops to run over parameter grids. R code for generating the data sets and running the Monte Carlo simulations is provided as Supplementary Materials.

In Figure 1 two simulated data sets are shown, both with six clusters (trajectories), $N = 50$, and $E = N(0, 1)$.

Figure 1

Examples of Simulated Data With T = 5 (Top Panel) and T = 10 (Bottom Panel)



Dependent Variables

The Adjusted Rand Index

The Rand index (Rand, 1971) has been proposed as an objective measure for correspondence between two cluster partitions. This index was later adjusted for chance by Hubert and Arabie (1985). The adjusted Rand index (*ARI*) is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{n_{i.}}{2} + \sum_{j}\binom{n_{.j}}{2}\right] - \sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\binom{n}{2}}$$

Here, n is the total number of subjects to be clustered, n_{ij} refers to the number of subjects in the i^{th} and j^{th} cluster of respectively the true data and the computed result. The "dot notation" refers to column and row sums. The index has an upper bound of 1, which implies perfect correspondence between the two partitions, and 0 when the index equals its expected value when the clusters are independent. If the solution is worse than the expected values than the index can become negative.

The Calinski-Harabasz Index

The Calinski-Harabasz index (*CH*) (Calinski & Harabasz, 1974) is a measure for internal cluster validity. The index can be used when the true clustering is unknown and the validation of the cluster solution is made by using quantities inherent to the data. The *CH* (also known as variance ratio criterion) measures of how similar an object is to its own cluster compared to other clusters. Here similarity within a cluster is based on the distances from the data points in a cluster to its cluster centroid and between cluster separation is based on the distance of the cluster centroids from the global centroid.

For a given number of clusters, K, the *CH* is defined as:

$$CH = \frac{\sum_{k=1}^{K} n_k (c_k - c)^2}{K-1} \Big/ \frac{\sum_{k=1}^{K}\sum_{k=1}^{n_k} (y_i - c_k)^2}{N-K}$$

where, n_k and c_k are the number of data points and the cluster centroid of the k^{th} cluster, respectively, c is the overall centroid, and N is the total number of data points.

Higher values of *CH* indicate better cluster solutions, which implies that the clusters are well separated from each other. The lower bound of this statistic is 0, but there it has no upper bound, which depends on the data. In order to better compare the *CH* statistic across the three methods it was standardized, denoted as *CHs*, to fall in the interval $[0, 1]$ using a logistic transformation:

$$CHs = 2\left(\frac{1}{1 + e^{((-1/\lambda)^{*}chi)}} - 0.5\right)$$

Here, the factor λ is defined by

$$\lambda = \frac{\sigma_y}{(N - K)}$$

where σ_y is the standard deviation of the dependent variable (across all subjects and time points).

The Variability of the ARI Index

A fourth measure to compare the methods on is the variability of the *ARI* index, which is the standard deviation of the *ARI* across the replications. This measure may provide an indication of the robustness of the method. For methods with similar means of the *ARI* index, the one that shows less variability is probably preferred.

Results

The results of the Monte Carlo study for the four indices are presented in the tables below.

The Adjusted Rand Index

First, the recovery of the true clustering using the *ARI* is given. The results with respect to the *ARI* for the three cluster methods are given in Table 1.

The *kml* method has the highest *ARI* value in all conditions. However, in most conditions the *lcmm* method yields only slightly lower or similar results. Only in one condition (Error level = 0.5, Clusters = 3, Time points = 5, Subjects = 25) there appears a substantial difference between these methods, in favor of the *kml* method. With 10 time points *kml* almost perfectly recovers the target clustering. The *traj* method has lower *ARI* values than the other two methods in all conditions. Only with 6 clusters, 10 time points and little error *traj* yields good results.

With respect to the number of cluster, in particular the *traj* method benefits from more clusters. For the other two methods the results are mixed: in some conditions there is hardly any effect of the number of clusters, in some conditions the *ARI* is even somewhat smaller with six clusters compared to three. The number of subjects has little effect on the *traj* method. For the other methods 50 subjects yields similar or larger *ARI* values than 25 subjects.

Table 1

Mean Adjusted Rand Index for the Three Methods

		Error = 0.5		Error = 1.0	
Condition	**Method**	**Cluster = 3**	**Cluster = 6**	**Cluster = 3**	**Cluster = 6**
T = 5, S = 25	kml	0.929	0.893	0.521	0.575
	traj	0.483	0.540	0.220	0.272
	lcmm	0.574	0.852	0.416	0.572
T = 5, S = 50	kml	0.939	0.908	0.568	0.612
	traj	0.490	0.547	0.237	0.267
	lcmm	0.901	0.886	0.473	0.613
T = 10, S = 25	kml	1.000	1.000	0.960	0.958
	traj	0.678	0.891	0.228	0.444
	lcmm	0.962	0.916	0.946	0.897
T = 10, S = 50	kml	1.000	1.000	0.970	0.971
	traj	0.684	0.907	0.243	0.439
	lcmm	0.978	0.986	0.970	0.946

Note. Based on 500 replications.

The number of time points has a consistent effect on the *ARI*. With 10 time points the results are always better than with 5 time points. Likewise, the larger error level decreases the *ARI* values in all conditions.

The Calinski-Harabasz Index

Next, the *CH* and standardized *CHs* indices are given. The results for the Calinski-Harabasz index for the three methods are given in Table 2.

The results of the *CH* cannot be compared between different numbers of clusters, time points and subjects, because its value depends on the data. Within each condition the three methods can be compared and these results are consistent with the *ARI* results. The *kml* method performs best, closely followed by the *lcmm* method in most conditions, whereas the *traj* method yields lower values. As expected, with increasing random error this index clearly decreases in all conditions.

The results for the standardized Calinski-Harabasz Index for the three methods are given in Table 3.

Table 2

Mean Calinski-Harabasz Index for the Three Methods

		Error = 0.5		Error = 1.0	
Condition	**Method**	**Cluster = 3**	**Cluster = 6**	**Cluster = 3**	**Cluster = 6**
T = 5, S = 25	kml	101.540	272.537	32.774	79.731
	traj	49.804	63.464	13.176	30.711
	lcmm	69.031	249.348	26.306	77.025
T = 5, S = 50	kml	203.608	546.643	61.941	155.336
	traj	97.800	126.975	26.648	59.197
	lcmm	192.796	519.823	52.909	150.584
T = 10, S = 25	kml	152.579	252.779	39.510	64.078
	traj	76.515	171.461	9.131	28.938
	lcmm	144.360	202.191	38.413	59.935
T = 10, S = 50	kml	304.657	503.114	77.968	127.321
	traj	158.189	356.767	17.899	56.819
	lcmm	294.111	487.246	77.279	123.306

Note. Based on 500 replications.

Table 3

Standardized Mean Calinski-Harabasz Index for the Three Methods

		Error = 0.5		Error = 1.0	
Condition	**Method**	**Cluster = 3**	**Cluster = 6**	**Cluster = 3**	**Cluster = 6**
T = 5, S = 25	kml	0.321	0.595	0.087	0.175
	traj	0.161	0.159	0.035	0.068
	lcmm	0.220	0.550	0.070	0.169
T = 5, S = 50	kml	0.322	0.598	0.083	0.171
	traj	0.159	0.159	0.036	0.065
	lcmm	0.305	0.573	0.071	0.165
T = 10, S = 25	kml	0.215	0.319	0.048	0.073
	traj	0.109	0.220	0.011	0.033
	lcmm	0.204	0.256	0.047	0.068
T = 10, S = 50	kml	0.215	0.318	0.048	0.072
	traj	0.112	0.228	0.011	0.032
	lcmm	0.207	0.308	0.047	0.070

Note. Based on 500 replications.

The results for the *CHs* obviously yields the same pattern as the *CH* when comparing the three methods. For this index the number of clusters has an effect in the sense that

the six clusters conditions yield better results than three clusters. The number of subjects seems to have no effect on this index. The standardized *CH* index is almost always smaller for 10 time points compared to 5 time points.

The Standard Deviations of the ARI Across Replications

The results for the standard deviation of the ARI values across replications for the three methods are given in Table 4.

Table 4

Variability of the ARI Index for the Three Methods

		Error = 0.5		Error = 1.0	
Condition	**Method**	**Cluster = 3**	**Cluster = 6**	**Cluster = 3**	**Cluster = 6**
T = 5, S = 25	kml	0.058	0.072	0.117	0.057
	traj	0.170	0.071	0.126	0.090
	lcmm	0.380	0.108	0.191	0.086
T = 5, S = 50	kml	0.039	0.060	0.086	0.049
	traj	0.142	0.065	0.087	0.090
	lcmm	0.133	0.080	0.178	0.057
T = 10, S = 25	kml	0.002	0.000	0.045	0.049
	traj	0.327	0.079	0.167	0.071
	lcmm	0.135	0.135	0.103	0.128
T = 10, S = 50	kml	0.000	0.000	0.027	0.026
	traj	0.354	0.058	0.152	0.049
	lcmm	0.094	0.066	0.041	0.077

Note. Based on 500 replications.

The variability across the replications is smallest for the *klm* method in all conditions. With ten time points and little error the results are almost perfectly stable for this method. The other two methods show mixed results, which depend on the condition. The variability of the *traj* method is in almost all conditions larger with 3 clusters than with 6 clusters. Furthermore, there does not seem to be a clear pattern.

Discussion

This study compared three different methods for longitudinal cluster analysis and focused on three corresponding R-packages that are available on the R-repository CRAN. Studies have compared model-based longitudinal clustering methods (Hsu et al., 2018; Sijbrandij et al., 2019), but to our knowledge a comparison between model-based and

non-model-based methods has not been done. In our analysis of simulated datasets we found that longitudinal *k*-means consistently either outperforms the other methods or performs at least as well, regards of number of time points, clusters, measurement error or participants, in terms of clustering accuracy (*ARI*), clustering separation (*CH*), and stability (SD_{ARI}). Unsurprisingly, in all methods higher measurement error and less time points yielded worse clustering solutions than low error and more time points. Number of participants however did not seem to impact the quality of clustering separation. The model-based longitudinal clustering method *lcmm* resembled the performance of the longitudinal *k*-means method in many instances, especially with 10 points. However, with 5 time-points and 3 clusters *kml* decidedly outperforms *lcmm* in terms of clustering accuracy. In all conditions the *traj* method performs worse than *lcmm* or *kml*. The accuracy-gap between *traj* and both *kml* and *lcmm* closes somewhat with six clusters compared to three. The *traj* method demonstrates good results only with a high number of clusters and time points, together with low measurement error.

Our findings imply that longitudinal *k*-means is a surprisingly strong method that can complement latent class mixed modeling methods such as *lcmm*. Longitudinal *k*-means is a relatively easy, and computationally less complex approach, which consistently performs well and sometimes even better than *lcmm*. Researchers might favor more theory driven approaches to growth modeling for their finer control over the clustering of growth patterns, but the more data-driven approach in longitudinal *k*-means seems to offer researchers a very good starting point at the least, and might be a better choice when the amount of expected growth patterns and the number of repeated measurements is low. Even though *traj* performs well with a high number of growth patterns and repeated measurements, our findings would imply that *kml* and *lcmm* are better methods in general, even though it remains unclear if there are specific conditions in which *traj* might excel.

In our Monte Carlo approach, the number of clusters to be found was set equal to the number of clusters that were generated. Our study can be regarded as a comparison between methods in a best-case scenario: when the number of clusters in the analyses match the number of clusters in the population. The limitation of our approach is that it is unclear how well the methods and packages perform when used to explore an unknown number of clusters. It would be worthwhile to evaluate the performance of these methods and packages when used to recover the number of clusters. This has already been addressed in other studies (Den Teuling et al., 2021; Twisk & Hoekstra, 2012), and appears to be a challenging issue. Moreover, not only statistical considerations play a role in deciding on the right number of clusters, but also knowledge of the substantive research field (Ram & Grimm, 2009).

In similar vein the sample sizes in our Monte Carlo simulation were small ($N = 75$ to $N = 300$). We chose this sample size to reflect common scenarios in which people are followed for 5 or 10 measurement moments. In longitudinal studies *t* is often used as a

trade-off for *N*, such that observations are regularly regarded as sufficient with regards to the required power of the study. We have explored a large sample size for our MC conditions, which confirmed the results, in the sense that *kml* and *lcmm* profit from the larger samples and *traj* does not. Unfortunately, determining the sample size needed to adequately power each method for any given clustering-characteristic in a population is not straightforward. A study by Martin and von Oertzen (2015) provides an indication that even in small sample sizes such as in our study our clustering techniques might perform adequately. But Martin and von Oertzen (2015) consider $N = 150$ a small sample size which is still considerably higher than the minimum sample sizes in our study. Considering that latent variable models are generally considered to be techniques which require large sample sizes the question is whether sample size impacts all compared methods in our study equally, or whether latent techniques, such as *lcmm* might particularly underperform. This appears to be the case in the present study in the condition with the smallest number of data points and little error. We also explored a larger sample size, in which the results were in line with the findings from the smaller sample sizes.

The Calinsky-Harabasz index as used in this study is not optimal for comparing methods. The CH-index's strength lies in the process of finding the optimal number of clusters. As such the CH-index is more useful for within-dataset comparisons rather than between method-comparisons across different data sets. We still opted to include the CH-index because researchers applying clustering methods to their data are using CH-index, and we wanted to explore how the CH-index performs, even when applied to the best-case scenario with the true number of clusters. To allow for between method comparisons we standardized the CH-index between 0 and 1. In future studies it would be worthwhile to explore the performance of the different methods when using the CH-index in choosing the number of clusters. In this study the CH-index results in conclusions which are congruent with the other measures of fit, adding support to our finding that both *kml* and *lcmm* perform well and outperform *traj*.

One of our study's strengths is the use of simulated data, which carries the advantage that the underlying clustering structure is known. It is unclear, however, whether real data and the impact of problems in real data, such as uni- and multivariate outliers and noncentrality of residuals would significantly alter this study's conclusions. By varying the measurement error in the simulated data this study has attempted to reflect the noise in growth patterns in real data. However, in empirical longitudinal data noise levels may sometimes be larger or other sources of bias may influence the results, such as selective drop-out.

The present study did not concern the question of measurement invariance (De Roover, 2021) and its impact on the measurement of growth. However, further research is needed to address these issues. Our study did also not address how the three methods performed when model assumptions were violated, such as assumptions of normality or homoscedasticity, or when clusters are of unequal sizes and degree of measurement

error. As such, this study can be seen as a first step under optimal conditions. The next step could be to explore the robustness of these methods and whether our findings persist under divergent conditions.

The present study was limited to the freely available and open source R-packages, but besides the R-environment there is more software that can used for longitudinal clustering. For instance, MPLUS (Muthén & Muthén, 2017), which is a highly specialized program, by which models like *lcmm* can be tested. Another well-known program for analyzing longitudinal trajectories is the SAS procedure TRAJ (Jones et al., 2001), which is also model-based.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: Peter Verboon is a member of Methodology's Editorial Board, but played no editorial role for this particular article or intervened in any form in the peer review procedure.

Supplementary Materials

For this article, the R code used to construct the data sets and to run the Monte Carlo simulations is available via PsychArchives (for access see Index of Supplementary Materials below):

Index of Supplementary Materials

Verboon, P., & Pat-El, R. (2022). *Supplementary materials to "Clustering longitudinal data using R: A Monte Carlo study"* [Code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.7052

References

Berlin, K. S., Parra, G. R., & Williams, N. A. (2014). An introduction to latent variable mixture modeling (Part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology, 39*(2), 188–203. https://doi.org/10.1093/jpepsy/jst085

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research.* Guilford Press.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics: Theory and Methods, 3*(1), 1–27. https://doi.org/10.1080/03610927408827101

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling, 28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

Den Teuling, N. G. P., Pauws, S. C., & van den Heuvel, E. R. (2021). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics.*

Simulation and Computation. Advance online publication. https://doi.org/10.1080/03610918.2020.1861464

Feldman, B. J., Masyn, K. E., & Conger, R. D. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology, 45*(3), 652–676. https://doi.org/10.1037/a0014851

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software, 65*(4), 1–34. https://doi.org/10.18637/jss.v065.i04

Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., & Subtil, F. (2016). kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS One, 11*(6), Article e0150738. https://doi.org/10.1371/journal.pone.0150738

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. SAGE.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis techniques and applications* (3rd ed.). Routledge.

Hsu, H.-Y., Lin, J. J. H., & Skidmore, S. T. (2018). Analyzing individual growth with clustered longitudinal data: A comparison between model-based and design-based multilevel approaches. *Behavior Research Methods, 50*(2), 786–803. https://doi.org/10.3758/s13428-017-0905-7

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218. https://doi.org/10.1007/BF01908075

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29*(3), 374–393. https://doi.org/10.1177/0049124101029003005

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere, 10*(1), Article e02567. https://doi.org/10.1002/ecs2.2567

Leffondré, K., Abrahamowicz, M., Regeasse, A., Hawker, G. A., Badley, E. M., McCusker, J., & Belzile, E. (2004). Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *Journal of Clinical Epidemiology, 57*(10), 1049–1062. https://doi.org/10.1016/j.jclinepi.2004.02.012

Leschinski, C. H. (2019). *MonteCarlo: Automatic parallelized Monte Carlo simulations*. https://CRAN.R-project.org/package=MonteCarlo

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*, 37–44.

Martin, D. P., & von Oertzen, T. (2015). Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes. *Structural Equation Modeling, 22*(2), 264–275. https://doi.org/10.1080/10705511.2014.936340

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism, Clinical and Experimental Research, 24*(6), 882–891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x

Muthén, B., & Muthén, L. K. (2017). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). https://www.statmodel.com

Proust-Lima, C., Philipps, V., Diakite, A., & Liquet, B. (2020). lcmm: *Extended mixed models using latent classes and latent processes.* https://cran.r-project.org/package=lcmm

Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package *lcmm. Journal of Statistical Software, 78*(2), 1–56. https://doi.org/10.18637/jss.v078.i02

Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576. https://doi.org/10.1177/0165025409343765

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

R-Core-Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sijbrandij, J. J., Hoekstra, T., Almansa, J., Reijneveld, S. A., & Bültmann, U. (2019). Identification of developmental trajectory classes: Comparing three latent class methods using simulated and real data. *Advances in Life Course Research, 42*, Article e100288. https://doi.org/10.1016/j.alcr.2019.04.018

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sylvestre, M.-P., McCusker, J., Cole, M., Regeasse, A., Belzile, E., & Abrahamowicz, M. (2006). Classification of patterns of delirium severity scores over time in an elderly population. *International Psychogeriatrics, 18*(4), 667–680. https://doi.org/10.1017/S1041610206003334

Sylvestre, M.-P., & Vatnik, D. (2014). *Using* traj *package to identify clusters of longitudinal trajectories*. R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/traj/vignettes/trajVignette.pdf

Twisk, J., & Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: A comparison between methods. *Journal of Clinical Epidemiology, 65*(10), 1078–1087. https://doi.org/10.1016/j.jclinepi.2012.04.010

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research, 43*, Article e100323. https://doi.org/10.1016/j.alcr.2019.100323

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

Methodology
2022, Vol. 18(2), 144–163
https://doi.org/10.5964/meth.7143