The Global Journal of Sport and Exercise Science (GJSES)
Vol. 1, No. 2 (2025)
ISSN (Print): 3080-0307 ISSN (Online): 3107-037X
DOI (Crossref): https://doi.org/10.63750/nswme776

Longitudinal clustering in kinesiology 1: the bare essentials of the kml and kml3d packages in R

Keston Lindsay [1]

[1] Helen and Arthur Johnson Beth-El College of Nursing and Health Sciences, University of Colorado Springs, Colorado Springs, USA. klindsay@uccs.edu

Received on: 13 December 2025 | Accepted on: 23 December 2025 | Published on: 31 December 2025

© 2025 The Author(s). Open access under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

ABSTRACT

Longitudinal designs are common in the field of kinesiology, where there can often be high variability due to varying response patterns between participants. In studies where outcome changes are seen, there may be hidden groups of non-responders. Alternatively, in studies where no overall changes are seen, there may be hidden groups of responders. Longitudinal clustering is a method that is used to find unique trajectories in longitudinal designs. Cluster trajectories are a result of data quality, and clustering is not meant to be hypothesis driven. However, It may complement hypothesis-driven methods to better understand overall and nuanced change patterns. The purpose of this article is to demonstrate how to analyze a synthetic dataset using the *kml* and *kml3d* packages in R. It is meant to be relatively non-technical, easy to use, and interpret for inexperienced coders.

Keywords: repeated measures, longitudinal clustering, cluster analysis, multivariate repeated measures, R, longitudinal k-means

Corresponding Author:

(Keston Lindsay)

- Helen and Arthur Johnson Beth-El College of Nursing and Health Sciences, University of Colorado Colorado Springs, Colorado Springs, USA.
- Email: klindsay@uccs.edu

1. Introduction

Repeated measures designs are common in kinesiology, where investigators seek to determine how treatments affect participants over time. These designs range from simple univariate pretest/post-test designs (requiring paired t-tests [TT]), to multiple trials on singular (repeated measures ANOVA; RMA) or multiple outcomes (repeated measures MANOVA; RMM). They are

used in such areas as exercise physiology, strength & conditioning, rehabilitation, and others. Many of these studies report variability in their outcomes, mainly due to varying response patterns between participants. For example, Hubal, et al. (2005) found a wide range of responses to resistance training for strength and size of the biceps brachii in both men and women (n = 585). Ojeda-Aravena, et al. (2021) tested differences in performance between a Taekwondo group vs. a repeated sprint group, and found variability within groups for several outcomes. Ross et al. (2019) suggested that response variability to exercise is due to an underlying genetic component, and, along with others, recommend specialized statistical methods such as linear mixed modelling (Ross et al., 2019; Mota et al., 2023).

Studies that reject a null hypothesis could have hidden subgroups that are drivers of change, where other subgroups may change in other ways (e.g., less drastically), or not even change at all. In other words, while the average effect looks significant, there are participants or subgroups that respond differently from what is found on average. Conversely, variability could be extensive enough to result in the retention of a null hypothesis (i.e., the average effect is non-significant), though again, some may be responders. Clustering methods may help in this regard by capturing subgroups based upon different response patterns. The results could lead to more refined research designs, and individualized training protocols.

Other advantages of using longitudinal clustering include: possible improved interpretation of any time x group interactions in mixed designs; statistical support for anecdotal observations; trend observation for both small pilot studies and large cohorts; the discovery of complex and unexpected trends; support for targeted exercise protocols; and trend visualization. Therefore, these methods may complement more traditional statistical methods and models. Clustering methods are data-driven, descriptive, and exploratory. They should not be confused with hypothesis testing.

- ***Longitudinal clustering***

Several methods for longitudinal clustering exist (Den Teuling, et al., 2021). These include, but are not limited to, growth-based trajectory modeling, growth mixture modeling (Den Teuling, et al., 2021), latent class mixed modelling

(Proust-Lima et al., 2017), and longitudinal k-means (KML; Den Teuling, et al., 2021; Genolini, et al., 2015).

Generally, they have different assumptions, and there is no consensus to determine the "true" number of clusters in a dataset (Genolini, et al., 2015). Several indices exist for this purpose. Nonparametric indices include the Calinski & Harabasz criteria (Calinski & Harabasz, 1974; Genolini, et al., 2015)

While parametric ones include the Bayesian Information Criterion (Schwarz, 1978; Den Teuling, et al., 2021).

Additionally, different clustering algorithms have been shown to create different numbers of cluster trajectories (Mullin et al., 2021).

Ideally, no matter which algorithm is used, the clusters should make theoretical sense. Again, clustering methods are exploratory, and not hypothesis driven. Cluster membership and trajectory should be interpreted based upon theory and context.

- ***Longitudinal k-means***

In the scenario presented here, longitudinal k-means is used to better understand characteristics of responders and non-responders. Longitudinal k-means is not model based (Genolini, et al., 2015),

And is considered a non-parametric method. A major disadvantage is that there are no tests for goodness of fit. However, given that its use in this context is to explore the result of hypothesis driven methods, this is a small obstacle. In tandem with the criteria used to decide the best number of clusters, the nature of the clusters themselves must make theoretical sense. However, it is flexible (Genolini, et al., 2015)

And this package is easy to use. This package also does not support inclusion of covariates, as found in similar methods (Proust-Lima et al., 2017). These could restrict their use in studies where contextual variables (age, training history, etc.)

May be important for trajectory interpretation. The method is best used for studies where timepoints are equally spaced, in contrast to other methods that allow variable spacing between timepoints (Zhou et al., 2023).

Longitudinal k-means clustering has been used in studies of body mass index in children (Pryor et al., 2011), as well as sleep and obesity (Touchette et al., 2008). ***Statistical Programs and Packages***

R (R Core Team, 2025) is an open source and free statistical programming language. It is comprised of the base R package, and an abundance of add-on packages that perform different tasks. The Statistical Package for the Social Sciences (IBM SPSS; Armonk, NY) has a convenient and easy-to-use graphical user interface (GUI). It is popular as both a research and teaching tool, due to its relative ease of use (Jain & Sengar, 2024).

However, it is proprietary, and needs to be repurchased on a regular basis. In contrast, R is free and open-source. R also has a community that contributes add-on packages and tools for virtually any purpose and discipline through the Comprehensive R Archive Network (CRAN). However, it requires coding, and thus may intimidate some users (Mota et al., 2023).

On his website, Choueiry (2025) describes a random sample of over 76,000 research papers uploaded to PubMed central between 2016 and 2021. Of the >40,000 papers that mentioned the use of a statistical package, IBM SPSS was found to be the most popular program (~40 % of papers), followed by R (~20 %). IBM SPSS was reported to have the largest decline in use over a six-year trend, while R along with Prism (GraphPad Software, Boston, MA) had the largest rates of adoption. IBM SPSS was also found to be more popular with beginners, while R was found to be in more high impact journals. Depending on the researcher's needs, programs may also be used in tandem. R and IBM SPSS were also found to be the second most popular combination of programs used together.

- ***A brief introduction to R***

As of this writing, R may be downloaded here: https://www.r-project.org/. While not needed, R can be used via an interface called an integrated development environment (IDE). A free and popular IDE for R is R Studio (Posit, 2025). As the time of this writing, R Studio may be downloaded here: https://posit.co/download/rstudio-desktop/. This is not meant to operate R in the same "point and click" fashion, as found in IBM SPSS. However, it provides several benefits, such as a simplified interface for the import of data, add-on package import and management, code management, and visualization. Once R Studio is installed, datasets may be easily imported in any format using the "Import Dataset" dropdown menu. For this method, users should check to ensure all data are designated as "numeric" when importing.

- ***R Packages: kml and kml3d***

The packages kml and kml3d (Genolini, et al., 2015) are add-on packages designed to implement k-means on single (kml) and joint (kml3d) longitudinal variables. The k-means version used in the package is the Hartigan-Wong version (Hartigan & Wong, 1979). Genolini et al. (2015) explain that the kml and kml3d packages offer seven methods for initial configurations of the algorithm, and they offer detailed descriptions of each one. In brief, they are: randomK, randomALL, maxDist, kmeans+, kmeans++, kmeans-, and kmeans-. Depending on the nature of the data, the cluster solution (i.e., number of clusters) could be affected by these initial configurations.

Similarly to RMM/RMA in IBM SPSS, they accepts data in the wide format. This means that rows contain participant numbers, and columns are time points. This is in contrast to long format, where levels of time are contained within one column. Hence, for five timepoints in long format, each participant would have 5 rows of data, one for each point in time. They can implement several imputation methods for missing data. These methods include "trajectory mean", "trajectory median", in addition to seven others (Genolini et al., 2015).

They contain various criteria for optimum number of clusters, define starting conditions on k-means, and can visualize trajectories. A detailed description of these functions are described elsewhere (Genolini, et al., 2015).

Cluster membership can be appended to the dataset, which can then be used for follow-up analyses. All follow-up analyses are exploratory only. The code is beginner friendly for dedicated IBM SPSS users who are exploring R.

The objective of this article is to demonstrate the use of KML using kml/kml3d as an exploratory analysis in tandem with RMA to explore hidden trajectories in the data. It is meant to supplement the use of RMM and RMA to further describe high data variability, subgroups that drive change, and potential non-responders.

2. Method

- ***Premise and data***

A synthetic dataset was created using ChatGPT (OpenAI, 2025) to simulate an applied sports sciences question: how does a training program affect lower body strength (3-repetition max [3RM] in kg) and peak power in female volleyball players over five equal intervals over the course of an off-season. The dataset did not have all of the qualities of the prompt. For example, correlations between strength and power were strong instead of moderate, suggesting collinearity. However, it is of sufficient quality for this demonstration. The

dataset includes 36 participants with five timepoints, designed for exploration of hidden cohorts using clustering. The dataset ("DII_VB+SC_3clusters_WC_RM") is shared using this link: https://osf.io/mrw5v/files/osfstorage and may be used for the purpose of teaching and practice. The dataset and the script used to generate it are also provided as supplemental material for this article.

- ***Initial Repeated Measures ANOVA***

The repeated measures analysis was completed in IBM SPSS Version 30 (Armonk, NY). Generally, everyone improved as seen below. For strength, all timepoints differed significantly except timepoints 2 and 3. For power, timepoint 1 differed from timepoints 3, 4 & 5; time point 2 differed from time points 4 and 5; time point 3 differed from time points 1 and 5; time point 4 differed from time points 1 and 2; and time point 5 differed from time points 1-3. The descriptive statistics for lower body strength (3RM squat in kilograms) and power (measured in watts) are in Table 1. MANOVA and follow up ANOVA reports are in Table 2. Profile plots for strength and power are shown in figures 1 and 2 respectively.

Table 1 Descriptive statistics

Timepoint	Power (mean ± s. d.)	Strength (mean ± s. d.)
1	2798.3 ± 198.7	99.8 ± 9.0
2	2866.4 ± 124.6	103.2 ± 7.1
3	2910.2 ± 131.5	106.0 ± 7.3
4	2968.6 ± 147.7	108.8 ± 7.4
5	3010.3 ± 209.9	111.9 ± 10.5

Table 2 (M)ANOVA analyses

Statistical Method	ANOVA report
RM MANOVA	$\Lambda = .59$, $F(8, 278) = 10.4$, $p < .001$, $\eta^2_p = .23$
RM ANOVA (Strength)	$F(1.6, 56.3) = 23.6$, $p < .001$, $\eta^2_p = .40$*
RM ANOVA (Power)	$F(1.6, 56.6) = 14.8$, $p < .001$, $\eta^2_p = .30$*

*Sphericity violated, Greenhouse-Geisser adjustment applied

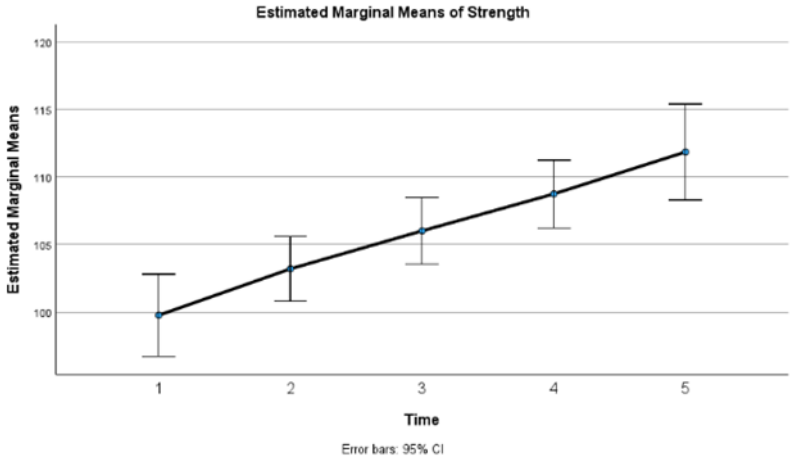


Figure 1. IBM SPSS generated profile plot for strength

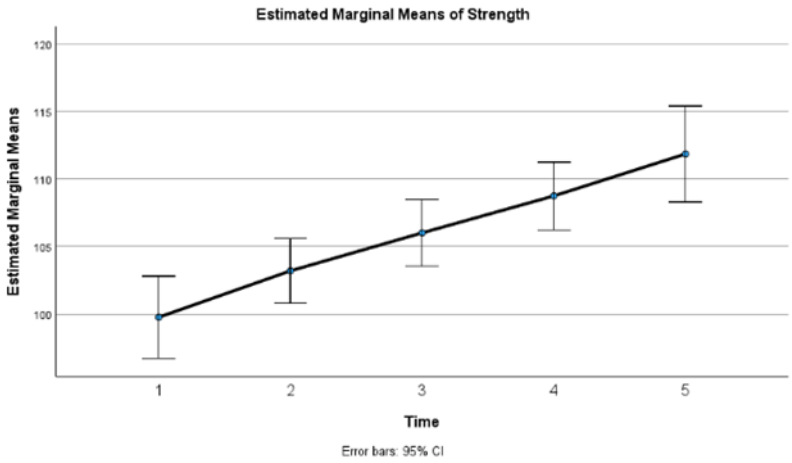


Figure 2. IBM SPSS generated profile plot for power

- ***Longitudinal k-means***

Longitudinal k-means (KML) was completed using R version 4.5.1 (2025) using the kml and kml3d packages (Genolini, et al., 2015). The RStudio desktop version 2025.05.1 (Posit, 2025) was used as the IDE for R. Once the dataset was imported into R Studio, it was named "sc3c". The multivariate data (i.e., joint trajectories) were analyzed first. Since the joint trajectory represented combined strength and power, it will henceforth be referred to as "performance". This was followed by the univariate data analysis for both strength and power. The list of commands is shown in Table 3 below.

Table 3. Summary of commands for kml3d, kml and data export

Command	Purpose
Performance commands	
sc3c<-as.data.frame(sc3c)	Converts an object into a data frame

perf_matrix<-cld3d(sc3c, timeInData = list(Strength = 2:6, Power = 7:11))	"ClusterLongData3d" command; builds cluster object for joint trajectory. Numbers designate column numbers for respective variables.
kml3d(perf_matrix	Performs clustering.
choice(perf_matrix)	Chooses best cluster solution for performance
perf_mem<-getClusters(perf_matrix, 3)	Assigns cluster membership based upon performance
sc3c$performance_clusters<-perf_mem<-getClusters(perf_matrix, 3)	Appends performance cluster membership as a new variable to dataset for further analysis.
Strength commands	
smatrix<-cld(sc3c, timeInData = 2:6)	"ClusterLongData" command; builds cluster object for strength only. Numbers designate column numbers for strength.
kml(smatrix, 1:5)	Runs longitudinal k-means between 1 and 5 clusters for strength.
plotAllCriterion(smatrix)	Optional command to compare optimal number of strength clusters using different criteria.
choice(smatrix)	Command to choose the best number of trajectories for strength.
strength_mem<-getClusters(smatrix, 3)	Assigns strength cluster membership.
sc3c$strength_cluster<-strength_mem	Appends strength cluster membership to data.
Power commands	
pmatrix<-cld(sc3c, timeInData = 7:11)	"ClusterLongData" command; builds cluster object for strength only. Numbers designate column numbers for power.
kml(pmatrix, 1:5)	Runs longitudinal k-means between 1 and 5 clusters for power.
plotAllCriterion(pmatrix)	Optional command to compare optimal number of power clusters using different criteria.
choice(pmatrix)	Software chooses the best number of single trajectories for power.
power_mem<-getClusters(smatrix, 3)	Assigns power cluster membership.
sc3c$power_cluster<-power_mem	Appends power cluster membership to data.
Export	

library(openxlsx)	Loads openxlsx package to export data to Excel spreadsheet
write.xlsx(sc3c, "sc3c.xlsx.", rownames = FALSE)	Exports Microsoft Excel spreadsheet to the folder you are using.
library(haven)	Loads the haven package to export to an IBM SPSS spreadsheet
write_sav(sc3c, "sc3c.sav")	Saves spreadsheet as stated above.

3. Workflow

- ***Multivariate analysis***

Once the cld3d, kml3d, and choice commands are run, the following image should appear for performance.

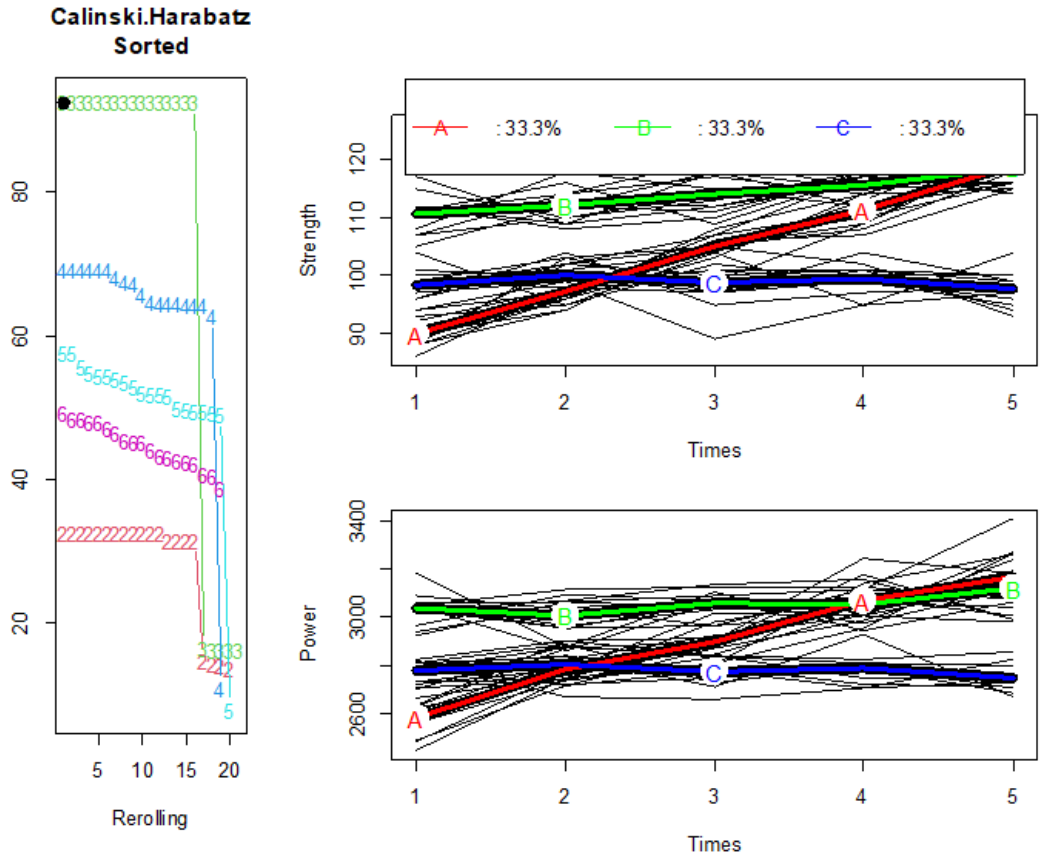


Figure 3. Three-cluster solution for performance using Calinski-Harabasz Criterion

The rectangular display on the right side shows the Calinski-Harabasz criterion. The cluster solution with the highest score should be interpreted, which is the 3-cluster solution. The getClusters command may then be used to assign cluster membership, and append to the dataset as described in Table 3 above.

The code can be modified as below: > kml3d(perf_matrix, nbClusters = 1:6, nbRedrawing = 50, "maxDist")

The nbClusters command specifies command can specify the number of clusters kml or kml3d should look for. The default is 2:6 clusters (comparison between 2 and 6 clusters). Depending on the data, it may be prudent to specify 1:6 in order to ensure that 1 cluster is not the best solution (i.e., to ensure that the group is not homogenous). The nbredrawing command designates the number of times k-means must be run with different starting conditions. The default kml command is equivalent to:

> kml3d(perf_matrix, nbClusters = 2:6, nbRedrawing = 20, "nearlyALL")

Where “nearlyAll” stands for “kmeans-“ followed by the alternation of “kmeans--” and “randomK” (Genolini, et al., 2015). Generally speaking, if “data” is the cluster object of your choice, the command:

> kml3d(data, nbClusters = 1:6)

or,

> kml3d(data, 1:6)

- ***Univariate analysis.***

For the univariate analyses (strength and power), the following images should be seen when the cld, kml, and choice commands are run.

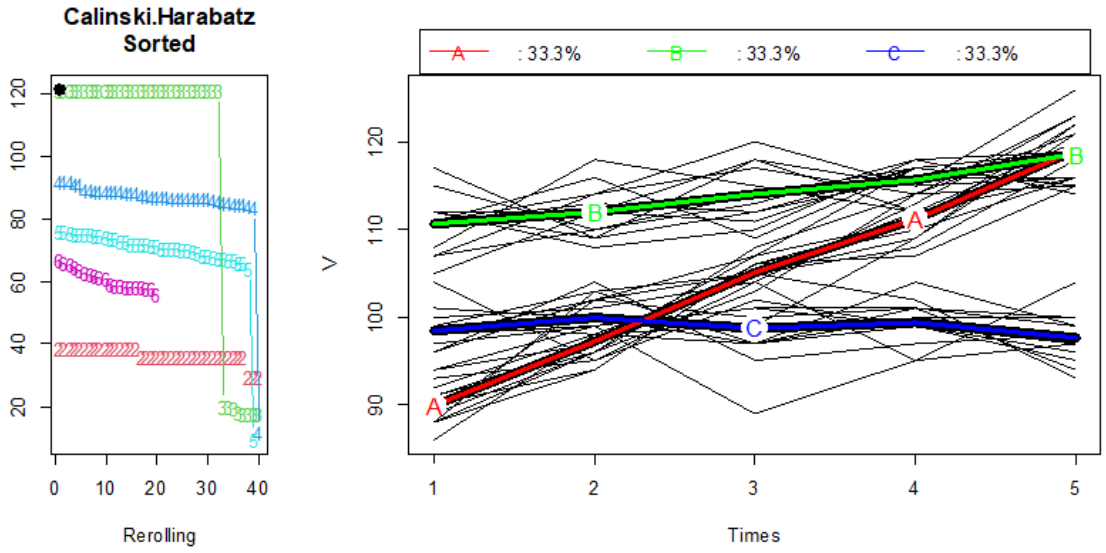


Figure 4. Three-cluster solution for strength using Calinski-Harabasz Criterion

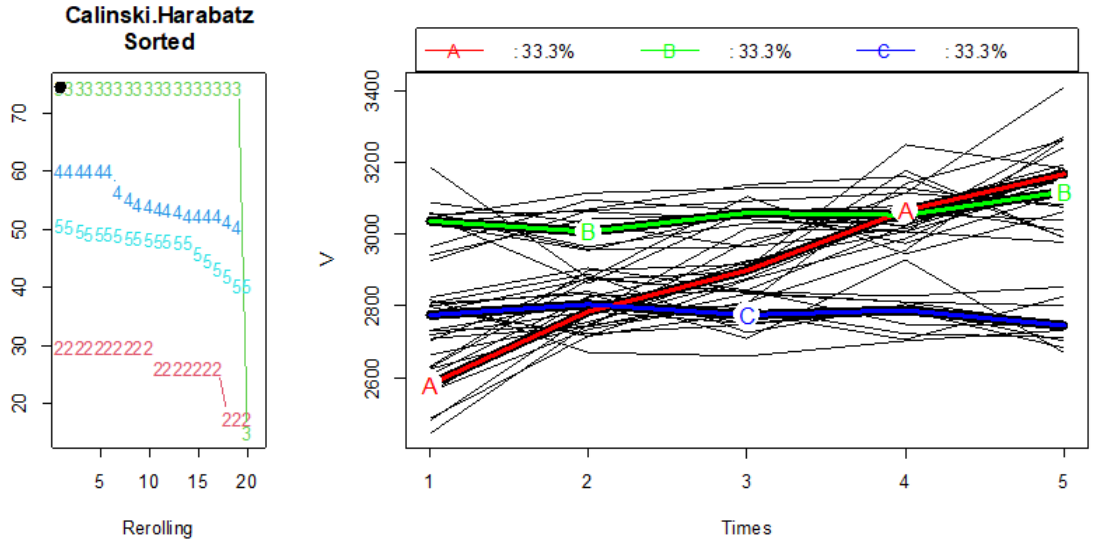


Figure 5. Three-cluster solution for power using Calinski-Harabasz Criterion

Note the similarity of both strength and power to performance. In practice, the number of clusters for all variables could differ. Such situations may necessitate individual inspection of cases to understand why. Similarly to performance, the get Clusters command may be used to acquire and append cluster membership.

Optionally, the Get All Criterion command produces the following figures for strength and power respectively. Although not unanimous, both figures show that three criteria prefer a three-cluster solution for both strength and power. This explains the three-cluster solution given by the choice command.

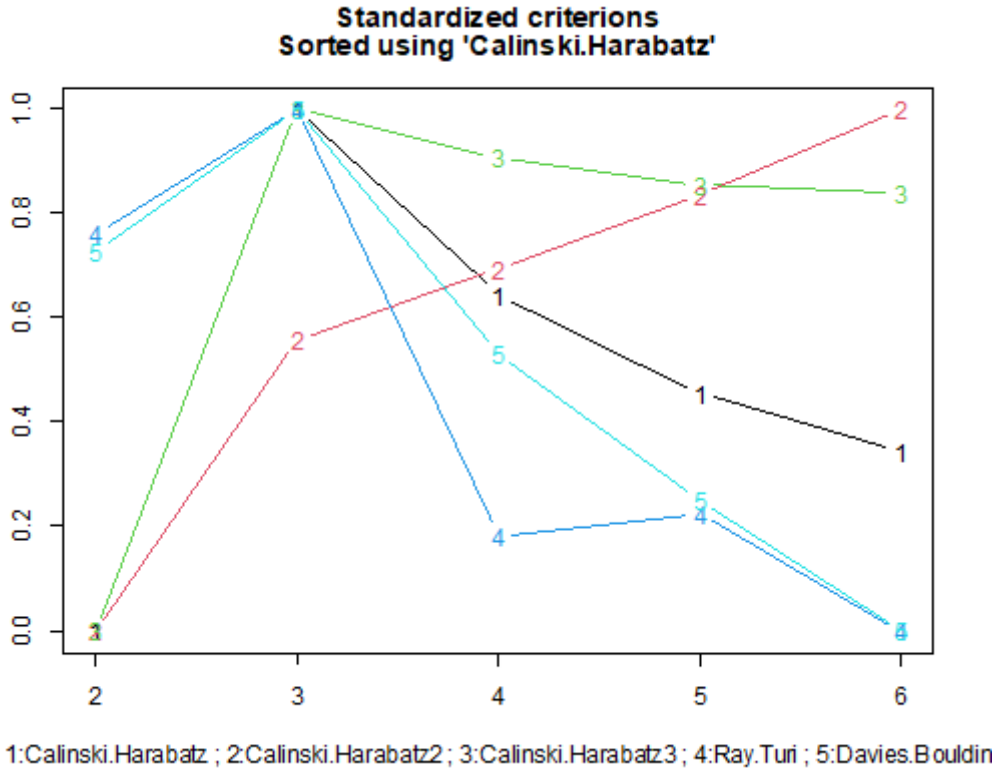


Figure 6. All cluster criteria for strength

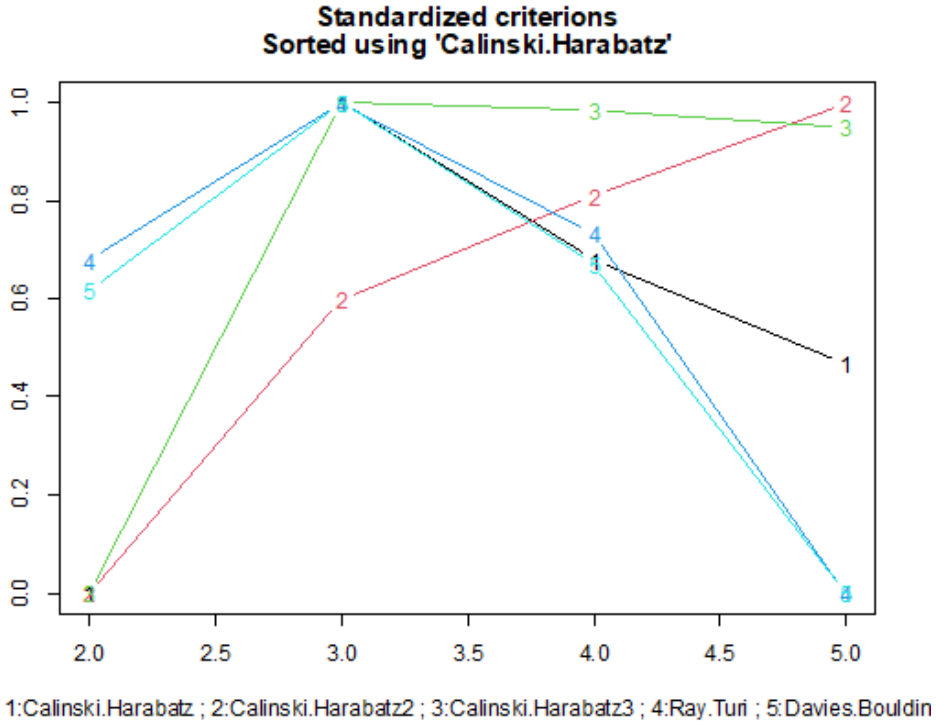


Figure 7. All cluster criteria for power

Once clusters are obtained, they may be statistically evaluated to explore differences between the clusters (Genolini, et al., 2015).

It should again be emphasized that cluster differences are a quality of the data only, and should not be generalized. While trends may be observed, cluster quality may not necessarily be consistent across different samples.

4. RESULTS Sample

- ***Analysis and Results sections***

This section is dedicated to providing a template for comprehensive and transparent reporting of analysis and findings.

- ***Analysis***

Descriptive statistics were performed using IBM SPSS version 30 (Armonk, NY). Repeated measures (RM) MANOVA for the effect of time on performance and follow up RM ANOVA analyses for both strength and power were also performed using IBM SPSS.

Longitudinal k-means was performed in R (R Core Team, 2025) using the kml (univariate strength and power), and the kml3d packages (multivariate performance). RStudio version 2025.05.1 (Posit, 2025) was used as the IDE for R. Exploratory comparisons between clusters were also completed using RM MANOVA and RM ANOVA in IBM SPSS, as described below.

Factorial (3 x 5) MANOVA was used to evaluate the effect of cluster membership and time on strength and power. Follow up mixed 2 x 5 mixed ANOVA with Sidak post hoc tests were then evaluated for each outcome.

5. Results

The null hypothesis for the first repeated measures MANOVA was rejected ($\Lambda = .59$, $F(8, 278) = 10.4$, $p < .001$, $\eta 2p = .23$). Similarly, the null hypothesis for the first strength [$F(1.6, 56.3) = 23.6$, $p < .001$, $\eta 2p = .40$)] and power [$F(1.6, 56.6) = 14.8$, $p < .001$, $\eta 2p = .30$] analyses were rejected. As sphericity was violated, Greenhouse-Geisser adjustments were applied to both the strength and power analyses.

Table 4 ANOVA analyses of clustered data

Statistical Method	Report
3 x 5 mixed MANOVA	$\Lambda = .02$, $F(16, 52) = 18.8$, $p = <.01$, $\eta^2_p = .85$
3 x 5 ANOVA (Strength)	$F(8, 132) = 37.4$, $p < .001$, $\eta^2_p = .69$
3 x 5 ANOVA (Power)	$F(8, 132) = 48.8$, $p < .001$, $\eta^2_p = .75$

- ***Cluster Interpretation.***

Longitudinal k-means revealed the highest Calinski-Harabasz criterion for three clusters each for strength, power, and performance. Participants 1-12 belonged to Cluster A, participants 13-24 to Cluster B, and participants 25-36 to Cluster C. As the null hypothesis for the higher order multivariate interaction (i.e., 3 x 5 MANOVA) was rejected, it was interpreted. Cluster A showed the most improvement, driving the most change in the entire cohort (see Table 5). Cluster B had the highest initial strength and power, and also trended upwards. At first glance, Cluster C appeared to be a subgroup of non-responders.

- ***Interaction of cluster and time on strength***

For strength, cluster B was higher than all other clusters at all time points except for cluster A at time point 5. At timepoint 1, C was higher than A. At timepoint 2, there were no differences between A and C. At timepoints 3, 4, and 5, A was higher than C. Cluster A's improvement was statistically significant across all times. Cluster B's improvement was less pronounced, but time point 5 was higher than timepoints 1, 2, and 3. Timepoint 4 was higher than timepoint 1. Cluster C was a group of non-responders. No other differences were observed.

- ***Interaction of cluster and time on power and performance***

For power, Cluster B was higher than the other clusters, except Cluster A at timepoints 4 and 5. Cluster A was lower than Cluster C at time point 1, and higher than Cluster C for timepoints 3, 4, and 5. Cluster A's improvement was statistically significant until timepoint 4, which didn't differ from timepoint 5. Clusters B and C were non responders. Hence Cluster A is a group of high responders, Cluster B is a group of higher performing marginal responders, and Cluster C is a group of non-responders.

- ***Variance explained***

The effect sizes on Tables 2 and 4 show that clustering was able to explain 62 % more variation in performance than RMM. Similarly, clustering was able to explain 29 % more variation in strength, and 45% more variation in power

than plain RMA. Descriptive statistics for all clusters at all timepoints are shown in Table 5.

Table 5. Cluster descriptive statistics

Variable/Timepoint	Cluster A	Cluster B	Cluster C
Strength/ Time 1	89.9 ± 2.4	110.7 ± 3.4	98.4 ± 2.6
Strength/ Time 2	97.2 ± 2.8	112.1 ± 3.1	100.0 ± 2.4
Strength/ Time 3	105.1 ± 2.4	114.1 ± 3.5	98.7 ± 4.4
Strength/ Time 4	111.3 ± 2.9	115.7 ± 2.1	99.4 ± 2.7
Strength/ Time 5	119.1 ± 3.1	118.7 ± 3.2	97.7 ± 3.0
Power/ Time 1	2582.3 ± 78.8	3035.6 ± 69.3	2776.9 ± 42.4
Power/ Time 2	2785.4 ± 58,1	3059.2 ± 54.7	2806.0 ± 78.1
Power/ Time 3	2897.7 ± 62.4	3052.0 ± 66.1	2773.7 ± 58.8
Power/ Time 4	3066.7 ± 84.4	3052.0 ± 66.1	2787.1 ± 62.0
Power/ Time 5	3166.4 ± 81.9	3118.3 ± 122.8	2746.4 ± 55.9

6. Discussion

The objective of this article was to demonstrate the use of longitudinal clustering as an exploratory method in tandem with repeated measures analyses using the kml/kml3d packages in R. Discovery and evaluation of hidden subgroups may be used to explain high variation, and may be used to create personalized exercise protocols, and further study non-responders.

Clustering should be used as an exploratory tool, and not a replacement for hypothesis testing. Depending on the data, different methods or initial configurations could produce different clustering solutions. Clustering may be used when researchers wish to see beyond the average effects of an intervention across time to see if meaningful subgroups exist. In a strength training study, everyone may improve, on average. However, clustering could reveal one group that steadily increased, while showing groups that plateau, slight, or non-responders. These findings may lend insight into generating new hypotheses, guiding follow-up studies, and modifying training programs.

- ***Limitations***

The kml3d package will only display two outcomes at a time, which may be a limiting factor for visualization in multivariate studies with more than two outcomes (Genolini et al., 2015). However, this may be a minor issue, as kml is capable of visualizing any singular outcomes that the researcher wishes. This

tutorial uses a synthetic dataset generated for demonstrative purposes only. Its generalizability is limited given the complex nature of real data. Researchers should exercise caution when extrapolating findings to experimental and applied settings and are strongly encouraged to test these methods using empirical data from actual studies.

6. Conclusion

The kml and kml3d packages are relatively easy to use, and the algorithm has been employed in previous studies in the health sciences. They may be a useful tool for students, junior researchers, and stakeholders who are more comfortable with GUI-driven packages like IBM SPSS. They may also be effective teaching tools for statistical programming courses, and to bridge the gap between using traditional ANOVA methods, and exploratory, data-driven methods. Stakeholders are encouraged to use kml/kml3d to gain more insight into their longitudinal data.

- ***Contributions***

The sole author bears the responsibility for all categories of this item.

- ***Acknowledgements None.***

Funding information None.

- ***Data and Supplementary Material Accessibility***

The dataset is available here:
https://osf.io/mrw5v/files/osfstorage

ID	Strength_T1	Strength_T2	Strength_T3	Strength_T4	Strength_T5	Power_T1	Power_T2	Power_T3	Power_T4	Power_T5
1	91	97	107	117	119	2561.7	2735.1	2811.5	2954.3	3265.0
2	89	102	107	111	122	2708.5	2744.2	2980.3	3078.9	3148.4
3	89	96	106	107	115	2628.9	2873.0	2897.1	3175.2	2990.4
4	88	94	106	110	116	2665.8	2757.0	2876.1	3057.3	3041.0

5	94	97	105	108	118	2582.4	2778.6	3018.2	3008.5	3135.3
6	90	94	106	111	119	2559.9	2823.2	2926.3	3007.6	3241.1
7	88	103	105	109	122	2607.8	2827.5	2843.8	3023.8	3168.6
8	86	98	99	109	121	2482.9	2773.7	2920.9	3050.4	3181.2
9	92	98	105	112	116	2486.8	2716.3	2872.6	2985.8	3187.1
10	88	96	108	114	115	2632.3	2900.9	2914.0	3070.6	3194.0
11	91	96	103	114	123	2446.5	2747.9	2904.8	3247.1	3184.6
12	93	95	104	113	123	2624.1	2747.2	2806.5	3141.4	3260.2
13	112	109	118	112	120	3050.1	2956.4	2964.3	3113.6	3082.1
14	117	109	112	116	116	3057.1	3062.9	3044.2	3007.3	2978.8
15	105	112	111	117	115	2964.3	3093.5	3067.1	2975.3	3113.9
16	115	110	113	118	114	3030.8	2954.3	3062.3	3079.7	3008.6
17	111	116	109	117	119	3028.6	3069.9	3136.6	3159.3	2989.8
18	112	108	110	118	119	2925.0	3066.2	3091.1	3116.2	3408.2
19	111	113	112	117	119	3045.7	3115.8	3126.3	3127.1	3074.8
20	108	118	115	112	120	3060.7	2963.2	3031.1	3036.2	3106.5
21	107	114	117	114	121	3185.2	2875.6	3104.9	2946.0	3062.2
22	111	114	120	115	116	3087.1	3030.1	2963.8	3017.8	3154.4
23	107	110	114	117	119	2941.6	3042.3	3053.6	3022.9	3271.5
24	112	112	118	115	126	3050.7	2863.0	3064.9	3022.1	3168.2
25	98	99	101	101	94	2733.7	2742.7	2834.8	2811.3	2748.3

26	99	98	97	104	99	2809.4	2889.7	2727.7	2806.3	2733.8
27	96	102	105	102	93	2782.6	2875.4	2841.0	2827.6	2854.4
28	99	103	97	100	100	2801.7	2842.1	2750.2	2788.4	2739.6
29	97	99	89	95	97	2807.8	2835.1	2709.5	2929.9	2669.5
30	96	104	95	97	98	2702.9	2880.1	2838.3	2812.4	2800.3
31	104	95	102	99	95	2799.0	2715.7	2781.1	2708.3	2828.0
32	101	100	97	99	97	2788.2	2721.5	2749.3	2795.5	2704.9
33	100	101	104	95	104	2734.2	2807.0	2794.6	2721.9	2712.3
34	94	99	101	99	96	2818.6	2671.7	2662.4	2705.0	2732.9
35	99	98	97	101	99	2824.9	2905.5	2843.6	2749.7	2748.5
36	98	102	100	101	100	2719.8	2786.0	2751.9	2788.3	2683.8

Instructions: Copy data into Excel spreadsheet to practice.

References

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

Choueiry, G. (2025). *Statistical software popularity in 40,582 research papers*. Quantifying Health. https://quantifyinghealth.com/statistical-software-popularity-in-research/

Den Teuling, N. G. P., Pauws, S. C., & van den Heuvel, E. R. (2021). *Clustering of longitudinal data: A tutorial on a variety of approaches* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2111.05469

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

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm.

Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108. https://doi.org/10.2307/2346830

Hubal, M. J., Gordish-Dressman, H., Thompson, P. D., Price, T. B., Hoffman, E. P., Angelopoulos, T. J., Gordon, P. M., Moyna, N. M., Pescatello, L. S., Visich, P. S., Zoeller, R. F., Seip, R. L., & Clarkson, P. M. (2005). Variability in muscle size and strength gain after unilateral resistance training. *Medicine and Science in Sports and Exercise, 37*(6), 964–972.

Jain, P., & Sengar, S. (2024). Unraveling the role of IBM SPSS: A comprehensive examination of usage patterns, perceived benefits, and challenges in research practice. *Educational Administration: Theory and Practice, 30*(5). https://doi.org/10.53555/kuey.v30i5.4609

Mota, J. A., Rajeb, M., Man, K., & Wohlgemuth, K. J. (2023). Linear mixed-effect modeling in kinesiology research: A tutorial in R. *Biomedical Journal of Scientific & Technical Research, 49*(3), 40613–40623. https://biomedres.us/fulltexts/BJSTR.MS.ID.007794.php

Mullin, S., Zola, J., Lee, R., Hu, J., MacKenzie, B., Brickman, A., Anaya, G., Sinha, S., Li, A., & Elkin, P. L. (2021). Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. *Journal of Biomedical Informatics, 122*, Article 103889. https://doi.org/10.1016/j.jbi.2021.103889

Ojeda-Aravena, A., Herrera-Valenzuela, T., Valdés-Badilla, P., Cancino-López, J., Zapata-Bastias, J., & García-García, J. M. (2021). Inter-individual variability of a high-intensity interval training with specific techniques vs. repeated sprints program in sport-related fitness of Taekwondo athletes. *Frontiers in Physiology, 12*, Article 766153. https://doi.org/10.3389/fphys.2021.766153

OpenAI. (2025). *ChatGPT* [Large language model]. https://chat.openai.com/chat

Posit. (2025). *RStudio Desktop* [Computer software]. https://posit.co/download/rstudio-desktop/

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

Pryor, L. E., Tremblay, R. E., Boivin, M., Touchette, E., Dubois, L., Genolini, C., Liu, X., Falissard, B., & Côté, S. M. (2011). Developmental trajectories of body mass index in early childhood and their risk factors: An 8-year longitudinal study. *Archives of Pediatrics & Adolescent Medicine, 165*(10), 906–912. https://doi.org/10.1001/archpediatrics.2011.153

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Ross, R., Goodpaster, B. H., Koch, L. G., Sarzynski, M. A., Kohrt, W. M., Johannsen, N. M., Skinner, J. S., Castro, A., Irving, B. A., Noland, R. C., Sparks, L. M., Spielmann, G., Day, A. G., Pitsch, W., Hopkins, W. G., & Bouchard, C. (2019). Precision exercise medicine: Understanding exercise response variability. *British Journal of Sports Medicine, 53*(18), 1141–1153. https://doi.org/10.1136/bjsports-2018-100328

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Touchette, E., Petit, D., Tremblay, R. E., Boivin, M., Falissard, B., Genolini, C., & Montplaisir, J. Y. (2008). Associations between sleep duration patterns and overweight/obesity at age 6. *Sleep, 31*(11), 1507–1514. https://doi.org/10.1093/sleep/31.11.1507

Zhou, J., Zhang, Y., & Tu, W. (2023). clusterMLD: An efficient hierarchical clustering method for multivariate longitudinal data. *Journal of Computational and Graphical Statistics, 32*(3), 1131–1144. https://doi.org/10.1080/10618600.2022.2149540