ARTICLE TEMPLATE

Contextualising the research problem: Improving cluster analysis insights into student learning

Emma Howard[a], Arthur White[a], and Jason Wyse[a]

[a]Trinity College Dublin, Dublin 2, Ireland

ARTICLE HISTORY
Compiled November 17, 2025

ABSTRACT

Cluster analysis is widely used in educational settings to gain insights into student learning. To justify their choice of clustering approach, authors often draw on methods used in earlier studies that they perceive to be similar. This can sometimes come at the expense of selecting a clustering method better suited to addressing their own study's goals. We argue that the selection of an appropriate clustering method should be closely connected to the context of the study. We demonstrate our argument through the use of the Open University Learning Analytics Dataset (OULAD), a well-established open access educational data set. Through a review of studies citing the OULAD, we identify seven possible motivations for clustering educational data and then focus on two of these: early identification of at-risk students and identification of similar groups of learners. We discuss the educational context behind the two motivations selected, describe which variable subsets from the OULAD might best align with the specific research questions that they motivate, and illustrate how a preferred clustering method may be highly influenced and driven by specifics of the research question. For example, the desire for an expression of uncertainty in cluster membership allocation. Fully reproducible R code is provided.

KEYWORDS
Cluster analysis; educational data mining; learning analytics; student profiles; online learning

1. Introduction

Cluster analysis is an unsupervised learning technique used to partition a set of data points or observations into homogeneous clusters or groups. Observations in the same cluster are more similar to each other than to observations in other clusters (James *et al.* 2021, Everitt *et al.* 2011). Variations of this definition include where observations may belong to multiple clusters, having competing degrees of membership to each. Traditionally, in educational research, cluster analysis has been deployed to group students with similar characteristics together, and these groups are often then related to academic achievement. For example, one debate in education focuses on whether students benefit equally from watching videos or attending lectures. Inglis *et al.* (2011) clustered students according to their use of module[1] resources (Virtual

CONTACT Emma Howard. Email: emhoward@tcd.ie

[1]Module is equivalent to course in American educational terminology

Learning Environment (VLE), lecture attendance and mathematics support centre usage) and found that the cluster with high lecture attendance achieved the highest module results on average. A recent review of clustering methods in educational data science (Le Quy *et al.* 2023) found that cluster analysis has been used for analysing students' behaviour and performance, predicting grades, for recommendations of learning resources or modules, supporting collaboration and teamwork, and analysing students' wellbeing. Educational cluster analysis is not limited to student data or quantitative studies. For example, Owen *et al.* (2015) used cluster analysis as part of a sequential mixed methods study to investigate UK teachers' patterns of use and attitudes towards the value of social media. A two-step clustering approach was implemented whereby hierarchical cluster analysis, followed by k-means clustering, identified a five cluster solution. Clusters were described as representing social media enthusiasts, social media engagers, social media sceptics, social media impartials and conscious luddites. Subsequently, at least one participant from each of the clusters were interviewed, which allowed for a deeper exploration of the nuances of teachers' perspectives on social media. Prevett *et al.* (2021) integrated a two-step cluster analysis approach within a qualitative inquiry framework to provide a deeper analysis of student interview data related to the values and approaches to learning mathematics.

In Le Quy et al.'s review of 133 publications, the most featured clustering methods were k-means (59%), hierarchical clustering (23%), fuzzy c-means (13%) and distinctive methods (7%). Of the nineteen other clustering methods featured in the publications in Le Quy *et al.* (2023), fourteen appeared in just one publication. Earlier reviews of clustering methods in education reinforce this. For example, Clatworthy *et al.* (2005) found that 76% and 54% of 59 studies reviewed used hierarchical and k-means clustering respectively. Clatworthy *et al.* (2005) reviewed the studies to identify whether they reported on the five cluster information items as defined by Aldenderfer and Blashfield (1984); these are the computer program used, the similarity measure used (where relevant), the cluster method used, how the number of clusters was determined, and evidence for validity of the clusters. They note that it was unclear whether specific methods were chosen because of their appropriateness, because previous studies had used them, or because they are the default methods within SPSS (the most popular software system being used in their identified studies). For example, k-means, while popular, is not specifically designed to handle mixed type data and an extension of k-means, k-prototype, may be more appropriate for mixed type data sets. Apart from factors relating to the nature of the utilised data, the choice of a cluster method may be influenced by contextual reasons. For example, in intervention studies, it may be necessary for financial reasons to identify smaller groups of students who demonstrate extreme behaviours. Alternatively, in studies aiming for practical applications based on the cluster solution, the area specialist may guide the choice of variables based on domain knowledge.

Cluster analysis is an exploratory method with no single optimal solution. Researchers often choose which clustering methods to implement based on the nature of their data or based on previous studies perceived as similar to their own. To identify a preferred solution, studies, for example Xu *et al.* (2013) and Nimy and Mosia (2023), have contrasted cluster solutions using Davies-Bouldin index, Calinski-Harabasz index, external data, cluster evolution, and silhouette width. Where clustering is most successfully applied, there is significant contextualisation of the research problem and aims, as argued by Hennig (2015). This includes that the data that are used reflect on, and have inferential relevance to, the research question. We argue that clustering methods can only be successfully applied in an educational data science setting if this

contextualisation has been explicitly acknowledged and discussed, and that, as well as using suitable metrics, it should directly inform how researchers choose between competing cluster solutions.

The central contributions of this paper are to: i) highlight the practical impact that the research question context should have on the different decisions needed when conducting cluster analysis in an educational setting; and ii) emphasise how the impact of these decisions should be understood within an appropriate contextual understanding of the research question that the analysis is being used to address. Therefore, we examine how the research question and educational motivation specifically influences what data are utilised, what cluster method is selected for analysis, and how the resulting cluster output is interpreted and evaluated. To allow readers to engage with the process themselves, the Open University Learning Analytics dataset (OULAD) has been chosen for this study, with fully reproducible analyses provided through the R language (R core team 2025). The OULAD (Kuzilek *et al.* 2017) has been used in at least 47 cluster analysis studies (see Section 2.2). We draw on these studies to identify potential research questions and educational motivations for performing cluster analysis. We show two valid cluster solutions for each motivation, and discuss under what circumstances one may be preferred over the other. To demonstrate the importance of contextualisation, our study addresses the following questions:

(1) What are the key motivations for applying cluster analysis to the OULAD?
(2) What subset of OULAD variables and cluster method(s) are appropriate to address research questions associated with a motivation?

Section 2 discusses the OULAD in detail, the educational motivations identified and the clustering methods used in this study. In Section 3, the study takes two identified educational motivations for clustering and for each motivation discusses the educational context, the OULAD data that aligns to the motivation and appropriate cluster methods. In addition, cluster analysis is performed with R code provided (in the paper and in supplementary files). Finally, in Section 4, the study conclusions are presented.

2. Materials & Methods

2.1. *Overview of the Open University Learning Analytics Dataset*

The OULAD contains information across 22 presentations of 7 Open University modules, and from 32,593 students (28,785 unique students) taking these modules in the 2013 and 2014 academic years. Data on each student includes assessment details, demographic details, registration details, and the logs of interactions with The Open University's Virtual Learning Environment (VLE) Moodle represented by daily summaries of student clicks. Modules are generally nine months long with a series of assessments throughout and an end of module examination. Modules can start in February (Summer offering denoted as 'B') or October (Winter offering denoted as 'J'). The OULAD has been anonymised (see Kuzilek *et al.* (2017) for further details) and is freely available as a collection of files (`https://analyse.kmi.open.ac.uk/open-dataset`). For statistical analysis, intensive preprocessing is required (Mihaescu and Popescu 2021). For ease of use, and reproducibility, this study will make use of the ouladFormat R package (Howard 2024). The ouladFormat package contains functions to load and format the OULAD for data analysis, allowing for reproducibility of data sets and

outputs.

As students' resource usage pattern in a module is strongly linked with the nature and timing of the assessment (Rust 2002), this study will be based on data from one Open University module presentation only. The social sciences module represented as 'BBB' with presentation '2013J' has been selected. Table 1 shows the due week and weight of the continuous assessment (CA) components of the module. When repeat students are removed, there are 1,453 students in the data set 'BBB-2013J' who have assessment, VLE, demographics and registration data. Repeat students are removed as they may display different resource usage patterns. Of the 1,453 students, 11% received distinctions, 55% passed, 23% failed and 11% withdrew. Students were predominantly female (89%) and 91% of students registered as having no disability. Students were divided into the following age categories: 0-35 (64%), 35-55 (36%) and ≥55 (0%).

Assessment Number	1	2	3	4	5	6	7	8	9	10	11
Week	3	7	8	14	14	19	19	24	24	30	30
Weight	5	18	1	1	18	1	18	1	18	1	18

Table 1. Due week and weight of the CA for module BBB-2013J

2.2. *Motivations for using Cluster Analysis on the OULAD*

2.2.1. Identification of Motivations

In order to identify educational motivations being addressed by cluster analysis, an investigative strategy was employed whereby the OULAD reference (Kuzilek *et al.* 2017) was searched on Google Scholar, and the 413 items (as of 19th of June 2024) citing the OULAD paper were taken as the pool of potential studies to draw from. The search string ('cluster*' OR 'unsupervised learning' OR 'kmeans' OR 'k-means') was applied to the pool of studies to reduce it further. The remaining 210 studies (when duplicates, non-English and inaccessible studies were removed) were reviewed by the first author to identify those which conducted cluster analysis on the OULAD. The research questions or/and motivation for the study were then extracted from the remaining 47 studies, and the studies reviewed. Upon reflecting on the goals and aims of the identified studies, the five most commonly explored motivations for using cluster analysis on the OULAD were:

(1) Early identification of at-risk students
(2) Identification of similar groups of learners
(3) Investigation of the link between student profiles and academic performance
(4) Evaluation of clustering methods
(5) Cluster analysis as part of an analysis pipeline

2.2.2. Cluster Analysis as the Primary Analysis Method

The studies categorised under the first three motivations are those where cluster analysis is the primary analysis method employed, and they have an educational-motivated objective. Motivation 1 involves identifying students at-risk of failing or dropping out early in the teaching term. This is in order to provide them with an intervention or support. This motivation is distinct from the other motivations as the data available for analysis are time limited up to when an intervention is to occur (see Section 3.1.1 for

further detail). Similar to motivation 1, studies categorised under motivation 3 could be considered as related to the identification of at-risk students. However, motivation 1 studies may be considered as proactive whereas motivation 3 studies are reflective. Motivation 3 requires students to have completed the module or course of interest so that student profiles may be linked to final academic achievement. For example, Kuzilek *et al.* (2019) clustered weekly VLE activity and found a connection between clusters representing resource usage levels and students' academic achievement. In contrast, motivation 2 is a broad classification (see Section 3.2.1 for examples). For a study to be considered under this category, cluster analysis must be the main analysis method used, the data featured have no time restrictions, and the objective is not to identify at-risk students by relating academic achievement to clusters.

In studies categorised under motivation 4, while cluster analysis is the primary analysis method employed, the focus is on comparing clustering methods using statistical evaluation. For example, Yu (2022) contrasts and statistically evaluates the cluster solutions provided by k-means, k-medoids and hierarchical clustering. Cluster labels output from each cluster method were used as an explanatory variable in a logistic regression where the response variable was a binary outcome indicating whether the student passed or failed. The preferred clustering method, hierarchical clustering, is chosen owing to its solution having the highest average silhouette width score, and having the highest pseudo R^2 and the lowest Bayesian Information Criterion (BIC) for its logistic regression.

2.2.3. Cluster Analysis as Part of an Analysis Pipeline

For motivation 5, cluster analysis represents one step in the study's analysis pipeline. Some studies, for example Sha *et al.* (2023) and Waheed *et al.* (2023), aim to overcome class imbalance in the OULAD by oversampling using k-means SMOTE with consideration of pass/fail or gender as protected variables. Subsequently, the authors perform predictive modelling on the new balanced data set. In other predictive modelling studies, such as Al-Zawqari *et al.* (2022), clustering is used as part of the variable selection process. Cluster analysis has been implemented as part of a recommendation system. For example, Bagunaid *et al.* (2022) calculate a student score variable based on each student's engagement and CA results. The student score was then clustered, dividing the cohort into three groups – excellent, average and poor. Next, the recommendation system focuses on the students in the average and poor learner clusters. Using the student score variable and an engagement variable as inputs into a reinforcement learning algorithm, Bagunaid *et al.* (2022) creates a series of action rules. These action rules are the basis for recommendations for students.

2.2.4. Summary of the Motivations

As these motivations are drawn from studies on a single data set, the OULAD, the range is limited. However, they are in line with those provided by Le Quy *et al.* (2023). In order to investigate how the research question informs the choice of data and the clustering method, we will focus only on motivations where cluster analysis is the primary method employed (motivations 1-4). Of these, in Section 3, we focus on two motivations to allow for depth of discussion. However, the approach taken is applicable to clustering analysis studies more broadly which are motivated by a research question. With motivation 1 and 3 being similar, albeit proactive versus reactive, we focus on motivation 1. As motivation 4 is defined by the focus on metrics for identifying the

preferred clustering method, and motivation 2 focuses on addressing an educational research question, we will also focus on motivation 2. Additionally, motivations 1 and 2 are complementary as motivation 1 addresses a specific, targeted question, while motivation 2 is more exploratory and in some sense reflective.

2.3. *Cluster Analysis*

We now outline the clustering methods that are to be used in Section 3. k-means and k-medoids are partition-based or algorithmic clustering methods. These are non-parametric clustering methods that do not assume an underlying parametric generative model for data. Gaussian finite mixture models and latent class analysis are model-based clustering methods which are parametric statistical approaches assuming a specific generative model. The generative model captures clustering and grouping by assuming each data point arises from one of a collection of sub-populations, each sub-population being described by its own parametric distribution. Points arise from sub-populations according to population proportions. The proportions are directly proportional in magnitude to the cluster/group membership size.

- *k-means:* is an iterative algorithm which partitions data into k clusters/groups (James *et al.* 2021, Maharaj *et al.* 2019, Everitt *et al.* 2011, Murphy *et al.* 2024). Each cluster is represented by its centroid, the mean of the data points within that cluster. k-means, also called c-means, is suitable for numerical data. The k-means algorithm is considered to be an understandable and computationally efficient algorithm. However, it tries to create spherical clusters of equal size and can have issues for high-dimensional data. k-means can be run in R using the *kmeans*() function from the *stats* package (R core team and contributors worldwide 2024).
- *Gaussian finite mixture models:* assumes that the data is generated by a finite mixture of Gaussian distributions with different parameters (Scrucca *et al.* 2016), and is generally used to cluster continuous data. Gaussian mixture models can be fitted in R using the *Mclust*() function from the popular *mclust* package (Fraley *et al.* 2024). Each model in mclust is denoted by a three-letter code, the volume-shape-orientation representation, which is explained as follows by Gormley *et al.* (2023, p. 577): 'The first letter denotes whether the volume is constrained to be equal (E) or varies (V) across clusters; the second letter denotes whether the shape is constrained to be equal (E) or varies (V) across clusters or if the clusters are spherical (I); and the final letter, which refers to the clusters' orientation, is subject to a similar interpretation'. The method is richer than k-means at the expense of making explicit parametric (Gaussian) assumptions about the data. Gaussian mixtures can provide a soft rather than hard clustering solution (see below bullet points) as a by-product of being formulated in a fully probabilistic way. It also allows for probabilistically grounded model determination techniques for k, such as BIC (Bruce *et al.* 2020), to be used.
- *k-medoids:* is also known as partitioning around medoids (PAM) or c-medoids (Maharaj *et al.* 2019, Everitt *et al.* 2011, Murphy *et al.* 2024). Its algorithm is similar to k-means, however, instead of clusters being represented by centroids, k-medoids selects actual data points (medoids) as exemplars of each cluster. It is often used when data are categorical in nature. k-medoids is less sensitive to noise and outliers, compared to k-means, but less computationally efficient. k-medoids can be run in R using the *pam*() function from the *cluster* package

(Maechler *et al.* 2024).

- *Latent class analysis (LCA):* is also known as latent structure analysis and is a model-based clustering method for categorical data. It assumes that the data are generated by latent, unobserved classes (sub-populations) and each class has its own probability distribution. Variables are assumed to be conditionally independent given the class membership information (Bouveyron *et al.* 2019). An extension to LCA, latent class regression, allows for the inclusion of covariates in the model to predict the latent classes. LCA can be run in R using the *poLCA*() function from the *poLCA* package (Linzer and Lewis 2024). LCA shares many of the same favourable properties of finite Gaussian mixtures as it is also a parametric clustering technique.

Computationally, all of the methods we have described have similarities in terms of how they are fit to data. Each iteration of the k-means clustering algorithm consists of two key steps. First, data points are assigned to the cluster to which their distance from the current estimate of the cluster centroid is minimal. Second, cluster centroids are recalculated using the updated cluster labels. These steps are repeated until no cluster labels change between iterations. The k-medoids algorithm proceeds in a similar fashion, except that instead of estimating cluster centroids at each iteration, cluster medoids are estimated using the updated cluster labels.

Gaussian mixture models and LCA models can both be estimated using Expectation-Maximisation (EM) algorithms. These algorithms are also similar in spirit to the k-means algorithm, in that each iteration consists of two key steps. In the E-step, the expected cluster membership of each observation is estimated, using current cluster parameter estimates. This is similar to the cluster allocation step in the k-means algorithm, except that the expected cluster membership is a soft cluster allocation, and hence assigns each observation a weight (formally, a posterior probability) of belonging to each cluster. Cluster parameters are then re-estimated during the M-step, using the data weights obtained during the E-step. The algorithm proceeds in this manner until it has converged. This occurs when differences in parameter estimates between successful iterations of the algorithm are deemed to be sufficiently small, or alternatively, when the log-likelihood increases a sufficiently small amount from one iteration to the next.

The k-means algorithm is a useful and efficient method for exploratory cluster analysis of continuous data. Its lack of a formal framework can however make more rigorous analysis challenging, and of the four methods we describe, it is the most restrictive in terms of the (implicitly) assumed cluster behaviour the algorithm seeks to identify. For example, cases where clusters exhibit different levels of variation, or correlation between variables, can often affect clustering performance. The k-medoids algorithm is a robust alternative to k-means, but also lacks an explicitly formal framework.

Gaussian mixtures and LCA models are both specified within a model-based clustering framework, and therefore share some similarities. Both types of cluster model can be specified explicitly as statistical models, and are more suitable when a rigorous, formal cluster analysis is required. The key difference between the models is that Gaussian mixtures are specified for continuous multivariate data, and LCA models are specified for categorical multivariate data.

Section 3 does not present how the number of clusters, k, is identified for different cluster analysis implementations. Code for k selection and related information is provided in the supplementary material. Selection of k depends on the cluster method used. The number of clusters can be selected using scree/elbow plots (Bruce *et al.*

2020), average silhouette width (Maharaj *et al.* 2019), expert opinion, or the BIC (Bruce *et al.* 2020) in the case of model-based clustering. Section 3 makes reference to some further clustering terminology which we outline briefly here:

- *Adjusted Rand Index (ARI):* is a measure of agreement between two different cluster solutions (Everitt *et al.* 2011, Hubert and Arabie 1985). The ARI is usually measured on a scale of 0 to 1 where 1 corresponds to perfectly matching cluster solutions. However, as the ARI takes into account agreements owing to chance between the two solutions in its calculation, it is possible that a computed ARI can be negative.
- *Hard cluster solution:* observations are classified to a single cluster. For example, there are two clusters in the solution and student 145 is assigned to Cluster 2. k-means and k-medoids produce hard clustering solutions.
- *Soft cluster solution:* observations have a probability of being assigned to each cluster. For example, there are two clusters in the solution and student 145 has probability of 0.2 of belonging to Cluster 1 and probability of 0.8 of belonging to Cluster 2. Model-based clustering methods produce soft clustering solutions, which can be transformed into hard clustering solutions simply by assigning each observation to the cluster for which its probability assignment is highest.

	Data type			Clustering solution	
Clustering Method	**Continuous**	**Categorical**	**Mixed**	**Soft clustering**	**Hard clustering**
k-means	✓	✗	✗	✗	✓
k-medoids	✓	✓	✓	✗	✓
k-prototype	✓	✓	✓	✗	✓
Hierarchical	✓	✓	✓	✗	✓
mclust	✓	✗	✗	✓	✓
LCA	✗	✓	✗	✓	✓
clustMD	✓	✓	✓	✓	✓

Table 2. Considerations for different clustering methods

3. Results & Discussion

In this section, we address motivations 1 and 2 which were outlined in Section 2.2. For each motivation we discuss the educational context, identify variables from the OULAD that might be used to address the motivation, and perform cluster analysis to address key research questions. The rationale behind the choice of clustering methods is discussed. Code implementing all methods is shown with the supplementary material including additional code for selection of k and reproduction of the figures included in the paper.

3.1. *Motivation: Early identification of at-risk students*

3.1.1. Educational Context

Learning analytics is defined as the 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs' (Siemens and Long 2011, p. 34). One core framework to the discipline of learning analytics is the academic or

learning analytics cycle (Campbell and Oblinger 2007). This consists of five steps: Capture, Report, Predict, Act, and Refine. An example of this in practice is an early warning system whereby students at risk of dropping out or failing a module are identified (Predict step) and an intervention provided to support them (Act step). This 'Act' step, is the real time support of students. For an intervention to be effective, the targeted students must be identified early on in the teaching term. Arguably, an optimal time for a supportive intervention is approximately half-way through the teaching term (Howard *et al.* 2018). This early identification of at-risk students limits the data that can be included in any analysis stage and thus in a Predict stage. While most studies identify at-risk students using predictive modelling (Jin *et al.* 2024, Alhakbani and Alnassar 2022), a number of studies use cluster analysis (Alshabandar *et al.* 2018, Palani 2020) to identify typical patterns of behaviour which might be used to predict at-risk students. Cluster analysis offers a key advantage over alternative predictive modelling approaches in such scenarios due to being an unsupervised method; it does not require a full set of labelled data (variables with an attached label or outcome) to train a cluster model. Unsupervised methods can identify frequently observed patterns of behaviour, which might then be mapped to categorisations such as at-risk based on expert insight.

Alshabandar *et al.* (2018) examined how to identify at-risk students in the Open University module 'BBB' for the 2013 cohort. Alshabandar *et al.* (2018, p. 3) split the VLE data into seven-time slices, where each time slice mapping is oriented around the final date of the assignment submission, and compared a Gaussian finite mixture model to supervised learning models linear regression and k-nearest neighbours for the identification of at-risk students over each of six time intervals. Gaussian finite mixture models performed comparatively to the supervised methods on measures of sensitivity, specificity, and other accuracy measures. Palani (2020) used cluster analysis (Gaussian finite mixture model, hierarchical and k-prototypes clustering methods) on the OULAD to identify students with low engagement levels at week 30 (out of nine months).

Borderline cases are an important consideration when identifying at-risk students at an early stage. While some students will exhibit clear patterns of poor or inconsistent engagement, indicating an urgent need for additional support, educators may also be keen to provide similar supports to students with weaker engagement levels whose difficulties are not so obviously clear cut. In such a setting, soft clustering solutions, which permit assessment of cluster uncertainty are particularly attractive. For example, suppose that a student had been assigned cluster probabilities of 49% and 51% between a cluster containing majority at-risk students and one which did not. In this case, the probability that this student is at-risk is clearly non-negligible, and it is likely that they would benefit from an early intervention. However, under a hard cluster assignment, this student could not be distinguished from other students in the same cluster who are less at-risk.

3.1.2. Cluster analysis case 1: Early identification of at-risk students

The data set used for this case consists of VLE data available up to week 14 inclusive of the teaching term and the average CA to week 14 inclusive. Demographic variables could also be included, but for this example we use VLE and average CA only as our variables. Students who withdrew prior to the end of week 14 are removed, as interest is in identifying students who are still at-risk. Students repeating the module are removed since these students are already at-risk (i.e., they have already failed

or withdrawn from the module in a previous presentation). The data set consists of numeric variables only for 1,382 students. The cluster methods employed are k-means and Gaussian finite mixture model-based clustering, which are both suitable for continuous numeric data. The VLE featured variables have a standard deviation in the range of 35-84, and the average CA has a standard deviation of 21. As the k-means cluster analysis method is sensitive to the scale of the featured variables, the variables have each been standardised (in other words, the data for each variable are centred by subtracting the variable mean and then rescaled by dividing by the variable standard deviation). This is commonly performed for cluster analysis. The following code is used to load and run the cluster analysis as described:

```
# Required packages for code
require(ouladFormat) # package for data set
require(tidyverse) # package for data manipulation etc.
require(mclust) # package for model-based clustering

# Load data for motivation 1 and standardises average_CA_score
mot1 <- combined_dataset(module = "BBB", presentation = "2013J",
  repeat_students = "remove", withdrawn_students = "remove",
  assessment = TRUE, na.rm = FALSE, VLE = "weekly",
  VLE_clicks = "standardise1", week_begin = 1,
  week_end = 14)$dataset_combined %>%
  dplyr::select(!(id_student:`14996`))
mot1$average_CA_score <- scale(mot1$average_CA_score)

# Implements k-means clustering and shows the cluster information
k_cluster <- kmeans(mot1, centers = 5, nstart = 10)
k_cluster

# Implements model-based clustering and shows the cluster information
m_cluster <- Mclust(mot1, G = 1:6)
summary(m_cluster)

# Calculate the ARI between the two solutions
adjustedRandIndex(k_cluster$cluster, m_cluster$classification)
```

Cluster results for both methods are visualised in Figure 1. The left-hand plots in this figure show students' standardised scores for each featured VLE variable, for each cluster. Scores are smoothed using a loess smoother in both cases. CA due dates are indicated on these plots, using vertical red lines. The right-hand side plots show boxplots of each student's average CA mark, grouped by cluster. Cluster results for k-means and Gaussian mixtures are shown in the top and bottom row of plots, respectively. The plots in these figures therefore show us the level of engagement exhibited by students within each cluster, as well as their subsequent academic performance.

From an educational perspective, of interest in Figure 1 are the flat patterns being exhibited across the VLE variables for most clusters. Studies, for example Rust (2002), have shown the strong relationship between the timing of students' resource usage and the due dates of CA. Therefore, it would be expected that there is an increase in resource usage prior to the due dates of the CAs (indicated by the vertical red lines in the left-hand plots in Figure 1) and a decrease in resource usage following the assessments. While it is expected that k-means and model-based clustering would produce different clustering solutions, the ARI score is 0.15, signalling limited agreement between the cluster solutions.

While standardising each variable individually is beneficial for implementing cluster analysis, this approach has not considered the repeated measures nature of the data

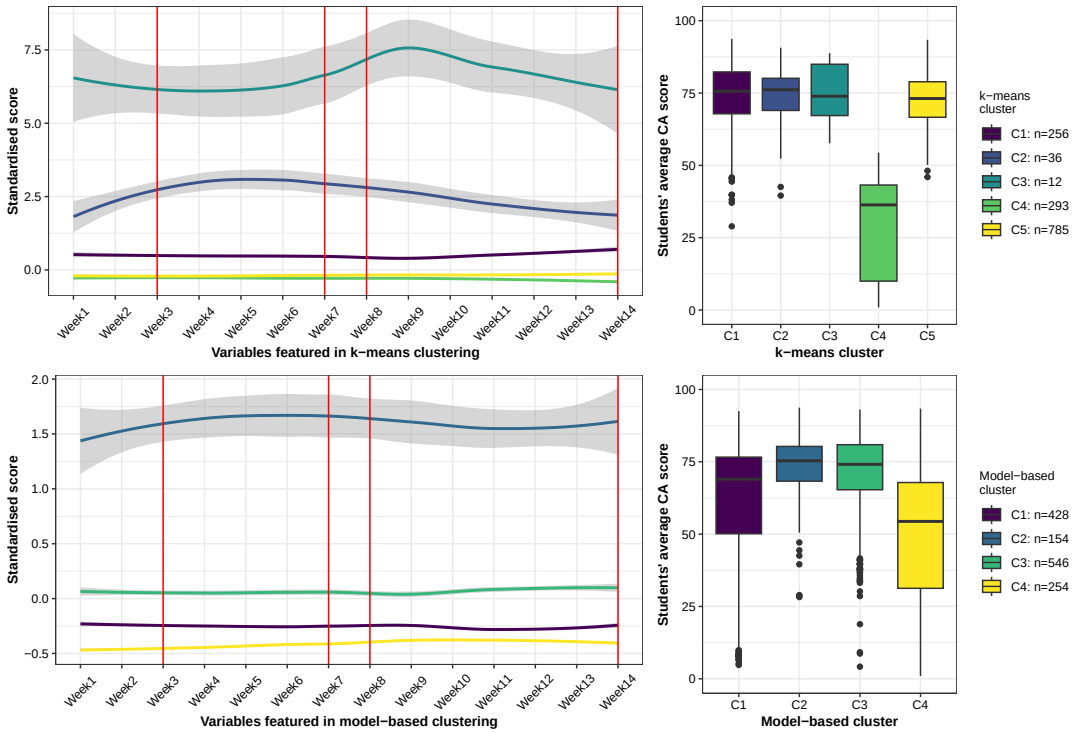

Figure 1. On the left-hand side, the plot shows students' standardised scores for each featured VLE variable and is smoothed for each cluster using a loess smoother. The red lines indicate weeks that contain CA due dates (see Table 1). On the right-hand side, a boxplot of students average CA by cluster grouping is shown. The top row represents the k-means cluster solution and the bottom row represents the model-based cluster solution.

set, and instead has removed much of the underlying structure of the data. In this data set, and likely in all data sets featuring VLE variables on a time basis, there is a strong positive correlation between the variables i.e., students who have high engagement in one time period are likely to have high engagement in the next time period. Additionally, as each variable is standardised individually, equal importance is assigned to the different variables. Subsequently, rather than standardising each variable individually, a global standardisation (or standardising across the data set) is applied to the VLE variables. This approach acknowledges the sequential relationship between the VLE variables and will go further in maintaining the integrity of any underlying signal. Note that with this approach, the average CA score variable is still standardised individually. To implement this change in standardisation, the argument VLE_clicks of the combined_dataset() function is changed to 'standardise2' with all other code remaining the same.

Cluster results for both methods using this alternative standardisation are visualised in Figure 2; this figure is organised in the same way as Figure 1. In line with expectations for the VLE data, the left hand plots in this figure show steady VLE usage for the first few module weeks (weeks 1-7), before a drop in VLE usage (weeks 7-12), followed by increased VLE use prior to the week 14 CA. While all clusters have a similar overall trajectory, we observe clear differences in engagement level between clusters, for each clustering method.

Students in the clusters with low VLE usage and low CA scores are considered at-risk; C4 for the k-means cluster solution, and C4 and C5 for the model-based cluster solution. Interestingly, for both cluster solutions, four clusters have approximately

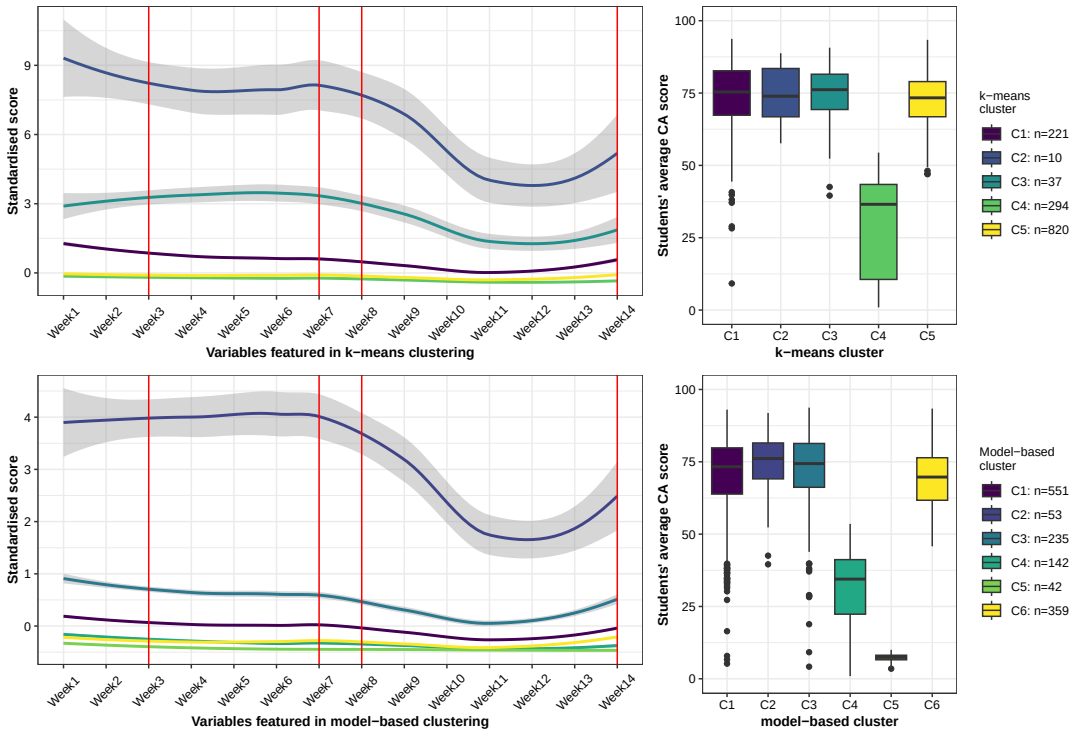

Figure 2. On the left-hand side, the plot shows students' standardised scores (based on global standardisation) for each featured VLE variable and is smoothed for each cluster using a loess smoother. The red lines indicate weeks that contain CA due dates (see Table 1). On the right-hand side, a boxplot of students average CA by cluster grouping is shown. The top row represents the k-means cluster solution and the bottom row represents the model-based cluster solution.

the same average CA of approximately 75% i.e., the students in these clusters have mastered the module material. For each cluster solution, the difference in the four clusters is the levels of VLE usage. This difference in VLE usage likely relates to student aptitude. The Model of School Learning, by (Carroll 1963, p. 729), proposes that all students can learn a task and aptitude is 'the amount of time needed to learn the task under optimal instructional conditions'. Carroll, in this model, relates the degree of learning by an individual to the factors of aptitude, ability to understand instruction, perseverance, opportunity to learn and the quality of instruction. Bloom (1971) builds on this, postulating that potentially over 90% of students can master what is being taught to them and proposes the theoretical framework of mastery learning.

While VLE usage is being used as a proxy for student engagement, the 'yellow cluster' from both solutions in Figure 2 might indicate an issue with this approach. The students in the yellow cluster for both clustering solutions have very low VLE engagement but high average CA. The Open University students receive module materials (including books) a few weeks prior to starting their module. Rather than accessing the VLE regularly, these students could potentially be relying on the physical materials that they received; for which no usage data are recorded. This emphasises the need to include both VLE data and CA data as featured variables in the cluster analysis when identifying at-risk students.

The ARI value between the cluster solutions has increased to 0.32; this now indicates low to moderate agreement. The k-means solution has clusters that vary substantially

in size (range is $10 - 820$), and has grouped the outliers together in a small cluster (C2: $n = 10$). Model-based clustering produced more evenly sized clusters (range is $42 - 551$) with smaller variable variation within the clusters. The model selected was the VEI (diagonal or ellipsoidal, varying volume, equal shape). In comparison, the k-means algorithm creates spherical clusters of equal volume and shape (equivalent to a EII model).

Returning to the motivation of the identification of at-risk students, once at-risk students have been identified (the Predict step), ideally a targeted intervention would be implemented for a small at-risk cohort (the Act step). The size of the cohort that could be accommodated would depend on the nature of the intervention/support planned, the ethics around fairness, the resources available etc. For k-means clustering, the cluster with the lowest engagement and CA average has a total of 294 students or 21% of the presentation cohort. Comparatively, the lowest group for model-based clustering has 42 students or 3% of the presentation (with the two lowest clusters representing 13% of the cohort). While both methods have identified at-risk students, the k-means solution may not be suitable for an Act step as it has identified a larger cohort of at-risk students. In addition, model-based clustering, unlike k-means, provides a soft-clustering solution. This may be useful, for example, if resources are limited and only those students who strongly belong to an identified at-risk cluster(s) are targeted for interventions. Therefore, we believe that generally the model-based clustering solution would be more appropriate for addressing the motivation.

It should be noted that there are extensions to k-means that will provide a soft solution, for example, fuzzy k-means. As demonstrated here, analysing data with a time element is a more complex clustering problem as there is likely to be strong correlation between variables. Casalino *et al.* (2019) addresses this for the OULAD by using DISSFCM (Dynamic Incremental Semi-Supervised Fuzzy C-Means) which is specifically designed for data stream classification and updates cluster classification as more data becomes available. Alternatively, StreamKM++ (Ackermann *et al.* 2012) could be used.

To validate the results, student demographics, and in this case the final result of students, can be cross-tabulated with the hard cluster solutions. The at-risk k-means cluster, Cluster 4, consists of the following breakdown: Distinction (1%), Pass (17%), Fail (62%), and Withdrawn (20%). Similarly, the at-risk model-based cluster, Cluster 5, breakdown consists of: Distinction (0%), Pass (0%), Fail (88%), and Withdrawn (12%). Both cluster solutions have identified the at-risk students. In practice, for an early warning system, the identification of at-risk students would occur early in the teaching term when no final results are known. Hence, cluster solutions would need to be validated by cross-tabulation with relevant demographic variables to check if the clusters are identifying differences in the cohort. These variables should not be included in the utilised data for the cluster analysis as it would bias the validation. For example, gender may be used if there are known historical differences in the module performance (see Table 3).

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
Female	0.89	0.96	0.93	0.85	0.86	0.88
Male	0.11	0.04	0.07	0.15	0.14	0.12

Table 3. Proportions of female and male students across the model-based clusters.

3.2. *Motivation: Identification of similar groups of learners*

3.2.1. Educational Context

The 'identification of similar groups of learners' provides a broad overview of the type of study included in this category. For example, in preparation for their study on investigating self-regulated learning on their main VLE data set, Eric (2023) implemented hierarchical clustering on the OULAD. Memon *et al.* (2020), using k-means and mini-batch k-means clustering on the OULAD, examined 'What are the crucial factors of students' behavioural data (activities performed) from VLEs that can identify procrastination behavior?'. The variables featured relate to number of previous attempts, date assignments submitted, studied credits, and overall score. In comparison, Treuillier and Boyer (2021, 2022) asked 'How to characterise learning dataset in terms of representativity of learning profiles'. They defined numeric learning indicators (performance, reactivity, engagement, regularity and curiosity) and used k-means clustering. The research question or aim influences the nature of the data to be clustered upon and therefore the clustering method (see Section 4 for application of contextualisation for these three studies).

Narrowing the scope of this motivation, the aim of this case will be 'to identify periods of withdrawal and disengagement by students'. To address this question, weekly binary indicators for activity will form our variables of interest. This allows us to explore periods of activity versus no activity (see Section 3.2.2 for more details). From an educational perspective, the limited information provided by binary data is a concern. Watt and Goos (2017, p. 135) state that 'Engagement is typically conceptualised as multidimensional, including behavioural, affective, and cognitive facets ... behavioural engagement refers to the extent to which students participate, including actual or intended enrolments, and degree of effort applied.' View counts, which provide more information than binary engagement variables, would be considered a very limited or granular proxy of behavioural engagement albeit one that is used regularly in research studies. Indeed Bergdahl *et al.* (2024, p. 1) have highlighted that learning analytics 'research overwhelmingly approaches engagement using observable behavioural engagement measures, such as clicks and task duration, with very few studies exploring multiple dimensions of engagement'. A potential concern in using binary data, a more restrictive form of count data, is the loss of information. However, Heuer and Breiter (2018) investigated the predictive power of binary daily variables as compared to total number of clicks on the given day in combination with demographic variables, and found them comparable. Although, this is likely dependent on the nature of the module/course.

While binary resource usage data would not accurately capture student engagement, it may be suitable depending on the research question, and can provide straightforward results which are easily visualised. Brooks *et al.* (2014) investigated students' usage of lecture capture videos. They used k-means clustering analysis on 13 binary variables; each variable representing a teaching week. A 1 indicated that the student had watched at least five minutes of the lecture capture video that week, and 0 otherwise. They identified five patterns of resource usage by students: High Activity, Just-In-Time, Early, Deferred, and Minimal. Those in the High Activity Cluster achieved the highest module marks, followed by the those in the Just-In-time Cluster. Alternatively, Carroll and White (2017) used LCA on binary variables related to students' use of module resources (lectures, tutorials, online and printed materials) across 12 teaching weeks and found four clusters named: Good Intentions; Conscientious Attenders, Late Online

Adopters; Conscientious Attenders, Early Online Adopters; and, Poorly Engaged.

3.2.2. Cluster analysis case 2: Identification of periods of withdrawal and disengagement

The data set used for this case consists of VLE weekly variables (for the full term of 39 weeks). Data consists of binary variables only. A 1 for a student in a given week indicates that a student accessed the VLE content in that week. Clustering is by k-medoids and LCA, both suitable for binary data. In k-means, the cluster centre represents the average of the cluster points. In comparison, with k-medoids, the cluster centre is represented by a medoid, a data point within the cluster that best represents that cluster. In the context of this data set, the medoid for each cluster would be a student who best represents a binary engagement pattern with the VLE module material. We run LCA using the poLCA() function of the poLCA package (Linzer and Lewis 2024). For poLCA, the formula and data arguments need to be in a particular format (e.g., rather than 0's and 1's being used for binary data, 1's and 2's must be used). The following code is used to load the specified data, run the cluster analysis as described and show summary clustering results:

```
# Required packages for code
require(ouladFormat) # package for data set
require(tidyverse) # packages for data manipulation etc.
require(cluster) # package for k-medoids clustering
require(poLCA) # package for latent class analysis (LCA)
require(mclust) # adjustedRandIndex() function for ARI calculation

# Load data for motivation 2
mot2 <- combined_dataset(module = "BBB", presentation = "2013J",
  repeat_students = "remove", withdrawn_students = "keep",
  VLE = "weekly", VLE_clicks = "binary",
  week_begin=1, week_end=39)$dataset_combined %>%
  dplyr::select(!(id_student:code_presentation))

# Implements k-medoids clustering and shows the cluster medoids
k_medoids <- pam(mot2, k = 6, nstart = 10)
k_medoids$medoids
k_medoids$clusinfo

# Format data for LCA
mot2LCA <- as.data.frame(mot2) + 1
model <- cbind(Week1, Week2, Week3, Week4, Week5, Week6, Week7,
  Week8, Week9, Week10, Week11, Week12, Week13, Week14,
  Week15, Week16, Week17, Week18, Week19, Week20, Week21,
  Week22, Week23,Week24, Week25, Week26, Week27, Week28, Week32,
  Week33, Week34, Week35, Week36, Week37, Week38, Week39) ~ 1

# Implements LCA once formatted and calculates ARI
LCA_clust <- poLCA(formula = model, data = mot2LCA, nclass = 7,
graphs = TRUE,  nrep = 10, maxiter = 3000)
adjustedRandIndex(k_medoids$clustering, LCA_clust$predclass)
```

For k-medoids clustering, a k=6 solution was chosen. Figure 3 shows the resource usage patterns of students in each cluster where a coloured block indicates that students accessed the VLE for that week. It was common to all clusters that students did not engage with the module in week 12. As the module began in October, this would likely align to the Christmas holidays. Figure 3 shows that CA clearly drives

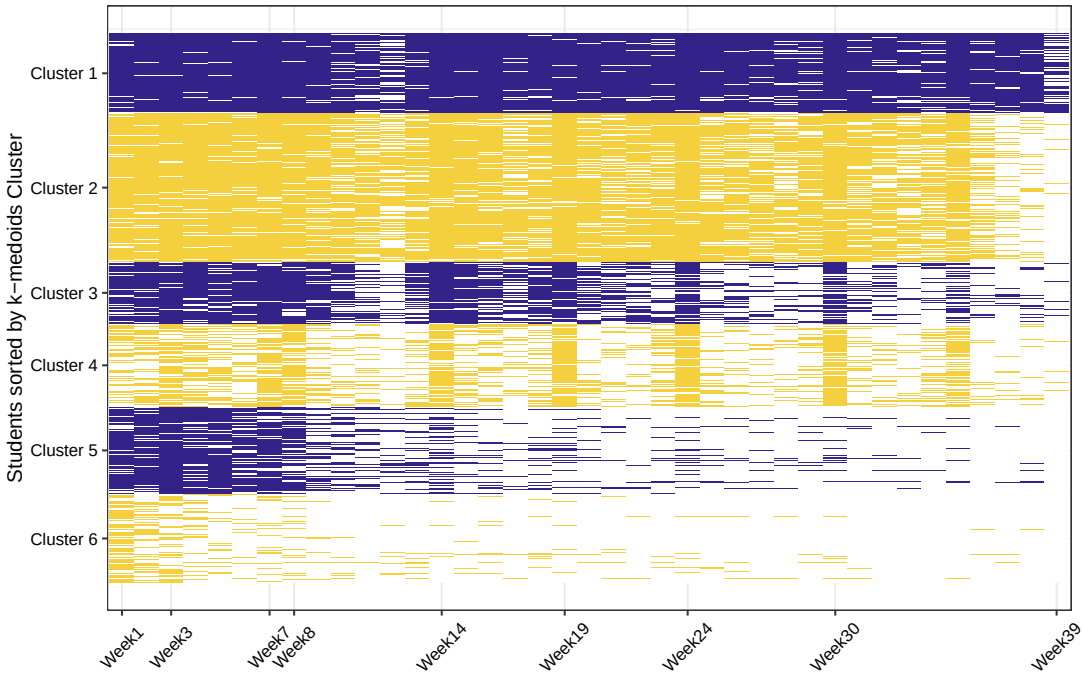

Figure 3. Each student is represented by a row, and each chronological week by a column. A filled in block indicates a student accessed the VLE content for that week. Weeks indicated on the x-axis are either the first week, the last week of VLE data or a week containing a CA due date.

engagement. There are three overarching patterns being displayed by the clusters. Students in Cluster 1 and 2 access module resources regularly throughout the semester; with engagement in the last few weeks differentiating which cluster the student belongs to. Students in Clusters 3 and 4 could be considered as just-in-time students; mainly accessing resources around the time that a CA is due. Clusters 5 and 6, consisting of approximately 32% of students, display patterns by students who withdrew or stopped engaging with the module (at different time points for the two clusters). As students tend to withdraw or stop engaging by week 12, the early identification of at-risk students (see Section 3.1) might be better placed to occur following the week 8 CA.

For LCA, a k=7 solution was chosen. Figure 4 displays the same three overarching patterns as seen in Figure 3. However, there are a few differences at the finer level. First, there are three clusters displaying withdrawal patterns (Clusters 4, 6, and 7) in approximately weeks 8, 14, 19, and 24. These account for 29% of the student cohort (5%, 8% and 16% respectively). For modules starting in the Autumn term (The Open University 2013), which includes BBB-2013J, students were liable for 0% of the fees up to the end of week 2, 25% of the fees after week 2 until the 31st of December (approximately week 12), 50% of the fees from the 1st of January until the 31st of March (approximately week 24/25), before becoming liable for the full amount. The timings of these withdrawal patterns are likely influenced by a combination of withdrawing after a CA or before another fee installment is due. This relationship could be investigated further by a qualitative follow-up study with withdrawn students.

In contrast to the k-medoid cluster solution, where there is residual engagement in later weeks, the LCA disengagement clusters have little or no engagement in later weeks. For example, for Cluster 6 in Figure 3 there are coloured blocks beyond weeks

7/8, but in Figure 4, Cluster 7 has very few coloured blocks beyond weeks 7/8.

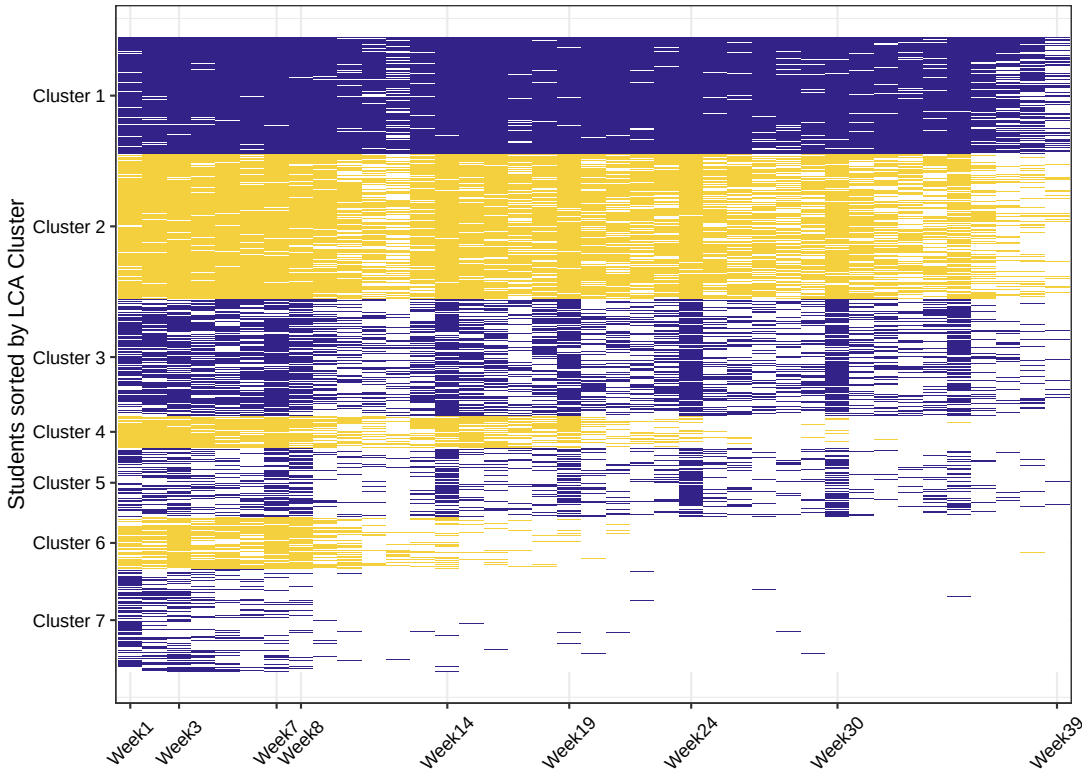

Figure 4. Each student is represented by a row, and each chronological week by a column. A filled in block indicates a student accessed the VLE content for that week. Weeks indicated on the x-axis are either the first week, the last week of VLE data or a week containing a CA due date.

The objective of this case was to identify periods of withdrawal and disengagement by students. The ARI between the cluster solutions is 0.42 indicating moderate but not strong agreement between the clusters. Both cluster methods have captured structures where students disengage, whether it is disengaging between CA due date, sometimes defined as procrastination, or disengaging entirely. The information provided by clustering could be used to estimate loss of fees from withdrawals, estimate staff costs, identify optimal intervention timings, or optimise the timings of module activities to keep students engaged. In such cases, k-medoids clustering might be preferred as the clustering algorithm is easier to understand and implement. Explainability can be particularly important if other stakeholders (e.g., university administrators or management) are involved in the study. LCA might be preferred in this case, as it provides a soft clustering solution and can be extended to 'latent class regression' or 'latent class models with covariates', where the probability of cluster membership is predicted by one or more covariates (e.g., demographic variables). An example of this is available in the supplementary materials. Overall, while both cluster solutions are valid, we would choose the LCA cluster solution as it is more well-defined with less residual engagement, thereby giving better estimates of the percentage of students withdrawing or disengaging throughout the module.

4. Discussion and Conclusion

This study demonstrates how clustering methods beyond the popular k-means and hierarchical clustering, can be applied to data in an educational settings. While, Table 2 summarises key properties of a number of clustering methods that we have mentioned, other motivations and data will require other clustering methods. For example, mini batch k-means (Wahyuningrum *et al.* 2021) or mixture models with variable selection (Fop *et al.* 2017) might be used for high dimensional data.

While the choice of clustering method(s) is influenced by the type of data utilised, this study aims to show how the educational context of the research study should also inform the cluster method chosen. In addition, the context should influence the choice of data utilised, including the variables chosen, the need for subsetting the observations, and potential transformations on the featured variables (for example global standardisation or dichotomisation). However, this is not always the case in (cluster analysis) studies. In some instances, the choice of a clustering method for a study is based on previous studies which are perceived to be of a similar nature or data-driven factors. For example, which cluster solution achieves the highest average silhouette width score. While these factors should be considered in the choice of cluster method, similar to Hennig (2015), we argue that contextualisation is also an important factor. In this paper, to raise awareness, we show two examples of contextualisation influencing data choices and cluster analysis decisions.

Drawing on studies citing the OULAD, this study identifies five common motivations for clustering of educational data (see Section 2.2) and focuses on two motivations (the early identification of at-risk students and the identification of similar groups of learners). For each motivation, the study demonstrates the importance of contextualisation by discussing the educational context behind the motivation, the variables from the OULAD that align to the motivation, and appropriate cluster methods for the motivation and data. The OULAD is an ideal data set to demonstrate this, as it is an established educational data set for cluster analysis research, allows for consideration of data to be utilised, and is open access. The two case studies presented are fully reproducible; with each data analysis decision documented and the full code available in the supplementary material.

For each motivation, two clustering methods were implemented and compared. Qualitatively, both solutions appeared similar in nature in each case. However, quantitatively, despite the two clustering methods using the same data, there was low to moderate agreement between the clustering solutions found for both motivations. Both clustering solutions are valid clustering solutions for their motivation, and the preferred choice should be influenced by contextual factors, for example, the need for a soft clustering solution when considering at-risk students, or a preference for identifying clusters of students with sparse but persistent engagement versus students whose engagement ends much more abruptly. Table 2 summarises properties of the clustering methods under discussion, and should serve as a rough guide for where to start when thinking about applying clustering. First determining the type of data being clustered, then following this, the type of clustering solution desired is a promising start. However, the final choice of cluster solution should be influenced by each study's research questions and objectives.

This study highlights the role of contextualisation in cluster analysis of educational data and encourages reflection on how data analysis decisions should be informed by the analyst's research objectives. To illustrate how this context has the potential to influence cluster analysis in an educational setting, we revisit the studies classified

in Section 3.2.1 under motivation 2, identification of similar groups of learners (Eric 2023, Memon *et al.* 2020, Treuillier and Boyer 2021, 2022). In each case, we consider whether a further cluster analysis could be beneficial, and if so, highlight what aspects of the additional cluster analysis are most relevant to the original study objective.

Eric (2023) investigated students' interactions with online resources as a proxy for self-regulation. It appears that students were clustered based on their total counts of VLE accesses, which links to self-regulation involving active participation. However, as noted by Eric (2023), theoretical models show that self-regulation involves various phases (e.g., forethought, planning, control and reflection) which can be linked to different learning behaviours and VLE activities. Considering this, additional variables that better contextualise the study aim, such as counts of different VLE activities, number of days CA was submitted in advance of deadlines, and number of days of activity in the teaching term, et cetera, could also be collected. In this case, the primary question of interest is how the research objective should inform what variables are used for the cluster analysis, and the choice of cluster method is secondary.

Memon *et al.* (2020) applied k-means clustering to the OULAD to identify students who procrastinate. They then visually compared the cluster labels estimated using this method to those assigned to students using expert opinion. Memon *et al.* (2020) clustered students multiple times using different sets of clustering variables: assessment score and studied credits; assessment submission dates and studied credits; assessment submission dates and obtained assessment scores; and, previous attempts and assessment scores. Memon *et al.* (2020, p.148) found that the cluster solutions resulted "visualizations of linearly separable data", that were not suitable to address the complex topic of procrastination; they hence preferred the expert labels to those obtained using data driven approaches. However, k-means cluster solutions will always lead to linearly separable solutions with strict boundary conditions. In contrast, alternative cluster methods such as k-medoids and mclust would permit more flexible cluster boundaries, potentially allowing for a more complex data-driven representation of procrastination. In addition, the four variables could be clustered upon together with the number of previous attempts treated as an ordered factor variable. Subsequently, mixed-type data cluster methods such as k-prototype and clustMD could be investigated.

Treuillier and Boyer (2021, 2022) both divided the DDD-2013B data into four sub-datasets (withdrawn, fail, pass, distinction). They removed outliers and then, using k-means, clustered each sub-dataset to identify learning personas of each subset. Considering the objective of the study, distinctive, well-defined personas are of interest, rather than a large number of profiles with a high degree of overlap. The optimal number of clusters k, identified using average silhouette width and Davies-Bouldin indices, is quite large for some sub-datasets (e.g., the pass data set has an optimal k of 10 for 451 students). It is possible that for smaller values of k, a cluster solution with a marginally weaker score with respect to the cluster selection indices could lead to a simpler and more clearly interpretable cluster solution which would be more preferable in practice. In this setting, a soft-label clustering method that helps users to assess the degree of overlap between clusters would be advantageous. Entropy-based methods, which are calculated using the output of a a soft-label cluster analysis, can be used to identify suitable choices of k that fit the data well while minimising cluster overlap (Biernacki *et al.* 2002), or to merge highly overlapping clusters (Baudry *et al.* 2010).

We hope this study raises awareness of the role of contextualisation and encourages reflection on data analysis decisions.

Disclosure Statement

The authors report there are no competing interests to declare.

Acknowledgements

We would like to thank Dr Anthony Brown for his input on the study.

Data Availability Statement

The Open University Learning Analytics dataset (OULAD) (Kuzilek *et al.* 2017) that support the findings of this study are freely available at https://analyse.kmi.open.ac.uk/open-dataset. The R code that support the findings of this study are available in the supplementary material.

References

Ackermann, M.R., *et al.*, 2012. Streamkm++: A clustering algorithm for data streams. *Journal of experimental algorithmics*, 17, 1–30.

Al-Zawqari, A., Peumans, D., and Vandersteen, G., 2022. A flexible feature selection approach for predicting students' academic performance in online courses. *Computers and education: Artificial intelligence*, 3, 100103.

Aldenderfer, M.S. and Blashfield, R.K., 1984. *Cluster analysis*. Newbury Park, CA: Sage Publishing.

Alhakbani, H.A. and Alnassar, F.M., 2022. Open learning analytics: A systematic review of benchmark studies using Open University Learning Analytics Dataset (OULAD). *In*: *Proceedings of the 2022 7th International Conference on Machine Learning Technologies*, ICMLT '22, New York, NY, USA. Association for Computing Machinery, 81–86.

Alshabandar, R., *et al.*, 2018. The application of gaussian mixture models for the identification of at-risk learners in Massive Open Online Courses. *In*: *2018 IEEE Congress on Evolutionary Computation (CEC)*. 1–8.

Bagunaid, W., Chilamkurti, N., and Veeraraghavan, P., 2022. AISAR: Artificial intelligence-based student assessment and recommendation system for e-learning in big data. *Sustainability*, 14 (17).

Baudry, J.P., *et al.*, 2010. Combining mixture components for clustering. *Journal of computational and graphical statistics*, 19 (2), 332–353.

Bergdahl, N., *et al.*, 2024. Unpacking student engagement in higher education learning analytics: A systematic review.

Biernacki, C., Celeux, G., and Govaert, G., 2002. Assessing a mixture model for clustering with the integrated completed likelihood. *Ieee transactions on pattern analysis and machine intelligence*, 22 (7), 719–725.

Bloom, B.S., 1971. *Handbook on formative and summative evaluation of student learning*. McGraw-Hill, Ch. Learning for Mastery.

Bouveyron, C., *et al.*, 2019. *Model-based clustering and classification for data science: With applications in r*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Brooks, C., *et al.*, 2014. Modelling and quantifying the behaviours of students in lecture capture environments. *Computers & education*, 75, 282–292.

Bruce, P., Bruce, A., and Gedeck, P., 2020. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

Campbell, J. and Oblinger, D., 2007. Academic Analytics. EDUCAUSE Quarterly.

Carroll, J., 1963. A model of school learning. *Teachers college record*, 64 (8), 1–9.
Carroll, P. and White, A., 2017. Identifying patterns of learner behaviour: What business statistics students do with learning resources. *Informs transactions on education*, 18 (1), 1–13.
Casalino, G., Castellano, G., and Mencar, C., 2019. Evolving fuzzy clustering for data analysis in Virtual Learning Environments. *In*: *6th ACM Celebration of Women in Computing: womENcourage 2019* , Rome, Italy.
Clatworthy, J., *et al.*, 2005. The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology*, 10, 329–358.
Eric, A.N., 2023. *An educational data mining model for promoting self-regulated learning on learning management systems*. Kenyatta University. Available from: `https://ir-library.ku.ac.ke/items/10f5e1fb-2e45-44f0-8aa5-289d6022f205/full`.
Everitt, B.S., *et al.*, 2011. *Cluster analysis*. Wiley.
Fop, M., Smart, K.M., and Murphy, T.B., 2017. Variable selection for latent class analysis with application to low back pain diagnosis. *The annals of applied statistics*, 11 (4), 2080–2110.
Fraley, C., *et al.*, 2024. mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation. `https://cran.r-project.org/web/packages/mclust/index.html`.
Gormley, I.C., Murphy, T.B., and Raftery, A.E., 2023. Model-based clustering. *Annual review of statistics and its application*, 10, 573–595.
Hennig, C., 2015. What are the true clusters? *Pattern recognition letters*, 64, 53–62.
Heuer, H. and Breiter, A., 2018. Student success prediction and the trade-off between big data and data minimization. *In*: *DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik*. Bonn: Gesellschaft für Informatik e.V., 219–230.
Howard, E., 2024. ouladFormat: Loads and formats the Open University Learning Analytics dataset for data analysis.
Howard, E., Meehan, M., and Parnell, A., 2018. Contrasting prediction methods for early warning systems at undergraduate level. *The internet and higher education*, 37, 66–75.
Hubert, L. and Arabie, P., 1985. Comparing partitions. *Journal of classification*, 2, 193–218.
Inglis, M., *et al.*, 2011. Individual differences in students' use of optional learning resources. *Journal of computer assisted learning*, 27, 490–502.
James, G., *et al.*, 2021. *An introduction to statistical learning: with applications in r*. Springer Texts in Statistics. Springer New York, NY. Available from: `https://books.google.ie/books?id=5dQ6EAAAQBAJ`.
Jin, L., *et al.*, 2024. Predictive modelling with the Open University Learning Analytics dataset (OULAD): A systematic literature review. *In*: A.M. Olney, I.A. Chounta, Z. Liu, O.C. Santos and I.I. Bittencourt, eds. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Cham. Springer Nature Switzerland, 477–484.
Kuzilek, J., Hlosta, M., and Zdrahal, Z., 2017. Open University Learning Analytics dataset. *Scientific data volume 4*, 1–8.
Kuzilek, J., *et al.*, 2019. Analysing student VLE behaviour intensity and performance. *In*: M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou and J. Schneider, eds. *Transforming Learning with Meaningful Technologies*, Cham. Springer International Publishing, 587–590.
Le Quy, T., Friege, G., and Ntoutsi, E., 2023. *A review of clustering models in educational data science toward fairness-aware learning*. Singapore: Springer Nature Singapore, 43–94.
Linzer, D. and Lewis, J., 2024. poLCA: Polytomous variable latent class analysis. `https://cran.r-project.org/web/packages/poLCA/`.
Maechler, M., *et al.*, 2024. cluster: "Finding groups in data": Cluster analysis extended Rousseeuw et al. `https://cran.r-project.org/web/packages/cluster/index.html`.
Maharaj, E., D'Urso, P., and Caiado, J., 2019. *Time series clustering and classification*. Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press.
Memon, M.M., *et al.*, 2020. Analysis of student procrastinatory behavior in virtual learning

environments using machine learning. *Journal of hunan university natural sciences*, 47 (10), 139–153.

Mihaescu, M.C. and Popescu, P.S., 2021. Review on publicly available datasets for educational data mining. *Wires data mining and knowledge discovery*, 11 (3), e1403.

Murphy, K., López-Pernas, S., and Saqr, M., 2024. *Dissimilarity-based cluster analysis of educational data: A comparative tutorial using r*. Cham: Springer Nature Switzerland, 231–283. Available from: https://doi.org/10.1007/978-3-031-54464-4_8.

Nimy, E. and Mosia, M., 2023. Web-based clustering application for determining and understanding student engagement levels in virtual learning environments. *E-journal of humanities, arts and social sciences*, 4(12), 4–19.

Owen, N., Fox, A., and Bird, T., 2015. The development of a small-scale survey instrument of uk teachers to study professional use (and non-use) of and attitudes to social media. *International journal of research & method in education*, 39 (2), 170–193. Available from: https://doi.org/10.1080/1743727X.2015.1041491.

Palani, K., 2020. Identifying at-risk students in virtual learning environment using clustering techniques. https://norma.ncirl.ie/4411/1/kamaleshpalani.pdf.

Prevett, P.S., *et al.*, 2021. Integrating thematic analysis with cluster analysis of unstructured interview datasets: an evaluative case study of an inquiry into values and approaches to learning mathematics. *International journal of research & method in education*, 44 (3), 273–286.

R core team, 2025. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

R core team and contributors worldwide, 2024. stats-package: The R stats package. https://rdrr.io/r/stats/stats-package.html.

Rust, C., 2002. The impact of assessment on student learning. *Active learning in higher education*, 3(2), 145–158.

Scrucca, L., *et al.*, 2016. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The r journal*, 8, 289–317.

Sha, L., Gašević, D., and Guanliang, C., 2023. Lessons from debiasing data for fair and accurate predictive modeling in education. *Expert systems with applications*, 228, 120323.

Siemens, G. and Long, P., 2011. Penetrating the fog: Analytics in learning and education. https://eric.ed.gov/?id=EJ950794.

The Open University, 2013. Fee Rules 2013/2014. https://help.open.ac.uk/documents/policies/fee-rules/files/70/fee-rules-2013.pdf.

Treuillier, C. and Boyer, A., 2021. Identification of class-representative learner personas. *In*: *LA4SLE 2021 - Learning Analytics for Smart Learning Environments*. Bolzano, Italy, 38–45. Available from: https://hal.science/hal-03549915/document.

Treuillier, C. and Boyer, A., 2022. A new way to characterize learning datasets. *In*: *In Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022)*. vol. 2, 35–44. Available from: https://www.scitepress.org/Papers/2022/109825/109825.pdf.

Waheed, H., *et al.*, 2023. Early prediction of learners at risk in self-paced education: A neural network approach. *Expert systems with applications*, 213, 118868.

Wahyuningrum, T., *et al.*, 2021. Improving clustering method performance using k-means, mini batch k-means, BIRCH and spectral. *In*: *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 206–210.

Watt, H.M.G. and Goos, M., 2017. Theoretical foundations of engagement in mathematics. *Mathematics education research journal*, 133–142.

Xu, B., *et al.*, 2013. Clustering educational digital library usage data: A comparison of latent class analysis and k-means algorithms. *Journal of educational data mining*, 5 (2), 38–68.

Yu, L., 2022. *Application and comparison of clustering methods to educational process data*. No. 29215215. University of Washington ProQuest Dissertations & Theses. Available from: https://www.proquest.com/openview/2c5eb499384924217536b4f8a702dd0d/1?pq-origsite=gscholar&cbl=18750&diss=y.