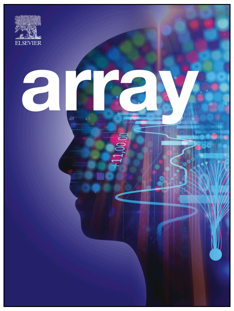
An Integrated Clustering Method for Pedagogical Performance

SAID, Raed A and MWITONDI, Kassim S. <http://orcid.org/0000-0003-1134-547X>

Published version

SAID, Raed A and MWITONDI, Kassim S. (2021). An Integrated Clustering Method for Pedagogical Performance. Array, p. 100064.



An Integrated Clustering Method for Pedagogical Performance

Raed A. Said, Kassim S. Mwitondi

PII: S2590-0056(21)00012-6

DOI: https://doi.org/10.1016/j.array.2021.100064

Reference: ARRAY 100064

To appear in: *ARRAY*

Received Date: 22 February 2021

Accepted Date: 12 April 2021

An Integrated Clustering Method for Pedagogical Performance

Raed A. Said[1] and Kassim S. Mwitondi[2]

[1]Canadian University Dubai
[2]Sheffield Hallam University, College of Business, Technology & Engineering
[2]k.mwitondi@shu.ac.uk

Abstract

We present an interdisciplinary approach to data clustering, based on an algorithm originally developed for the Big Data Modelling of Sustainable Development Goals (BDMSDG). Its application context combines mechanics of machine learning techniques with underlying domain knowledge–unifying the narratives of data scientists and educationists in searching for potentially useful information in historical data. From an initial structure masking, results from multiple samples of identified set of two to five clusters, reveal a consistent number of three clear clusters. We present and discuss the results from a technical and soft perspectives to stimulate interdisciplinarity and support decision making. We explain how the findings of this paper present not only continuity of on–going clustering optimisation, but also an intriguing starting point for interdisciplinary discussions aimed at enhancement of students performance.

Key Words: *Association Rules, Big Data, CHEDS, Data Mining, Data Science, Internship, Interdisciplinarity, Pedagogy, Propagated Clustering, SILPA, SMA Algorithm, Sustainable Development Goals, Unsupervised Modelling*

1 Introduction

The United Nations Sustainable Development Goals SDG [1] identifies good quality education as the foundation to creating sustainable development, improved quality of life, innovation and creativity. Investment in the sector of education and pedagogical innovations are well–documented, especially in the developed world. However, despite all the evidence on its impact on our livelihood, we are still witnessing huge gaps and variations in attainment and performance across the world. This paper presents an interdisciplinary approach to Big Data Modelling, based on two algorithms designed for machine learning techniques. A key motivation of this paper is to expand pathways for educationists and researchers in attaining unified efforts to uncover and analyse such factors in interdisciplinary contexts. It seeks to address the foregoing challenges by tracking undiscernible and potentially useful information hidden in multiple data attributes. Unlike in Miguis et al. [2], Brooks et al. [3] and Hua Leong and Marshall [4], where the focus was on the segmentation of the dynamics of static groups, this paper takes a Big Data modelling approach to tracking potential triggers of performance among University students (3639 observations on 19 variables) over an 11–year period (2005–2016). This work follows national guidelines of the Commission for Academic Accreditation (CAA) within the Ministry of Education (MoE) in the United Arab Emirates (UAE) which is authorized to license educational institutions, accredit programs and grant degrees and other academic awards across the country.

The Standards that guide the foregoing processes and the criteria that institutions must meet are specified in the Standards for Institutional Licensure and Program Accreditation [5]. It is clearly stipulated in SILPA [5] that institutions offering programs in professional fields such as medicine and other health-related disciplines, education, engineering and others must have to provide opportunities for learning through workplace experience, such as internships or practicums. Internships provide a structured practical learning experience where students are academically supervised and undergo a rigorous process to complement their theoretical learning. At the university degree level, internships are usually required as a part of the majors curriculum and as such they provide students with the opportunity to implement what they have learned theoretically while being supervised to insure they are on the right track. Research shows that through internships, students add more value to their knowledge by getting exposed to real life experimental learning experiences and opportunities. The paper is organised as follows. Section 1 presents the background, motivation, research aim, objectives and a brief review of relevant literature. Section 2 details the methods–data description and modelling techniques, followed by implementation, analyses, results and general discussions in Section 3. Finally, concluding remarks are drawn in Section 4, highlighting potential new research directions.

1.1 Motivation

Attaining good quality education is the ideal dream of all learners, institutions and nations across the world Attwell and Pumilia [6], Meusburger [7]. The United Nations identifies good quality education as the foundation to creating sustainable development, naturally leading to improved quality of life, innovation and creativity. In the modern era where we generate more data than we can process, the issue becomes both a challenge and an opportunity. In a typically academic environment where thousands of multilateral demographic students study multiple modules at different levels, the underlying and resulting data attributes are highly correlated sources of Big Data [8, 9]. Just what type of data, how much of it and how fast are questions that researchers have to deal with routinely. A key motivation of this paper is to expand pathways for educationists and researchers in attaining unified efforts to uncover and analyse such factors in interdisciplinary contexts. It is expected that this work will contribute to the work of the Center for Higher Education Data and Statistics (CHEDS) that collects vital educational data for the MoE. CHEDS [10] makes evidence-based decisions, influencing higher education policies and planning at both institutional and national levels. This helps the educational sector to enhance their strengths and ranking in the increasingly competitive world of higher education. Reports and analyses will help in advancing students learning experiences and curriculum designs.

1.2 Research Aim and Objectives

The aim of this paper is to highlight robust pathways for applying machine learning techniques in real–life applications in an interdisciplinary context [11]. It seeks to address the problem around **optimising naturally arising patterns in large datasets–**applying a clustering technique within an integrated generic algorithm in detecting and modelling potentially relevant educational performance data attributes. Its objectives, listed below, are two–fold. Objectives 1 through 3 focus on the technical aspects of the work, while 4 and 5 are on the underlying domain knowledge.

1 To capture multiple data attributes on students' performance across disciplines and carry out data cleaning, data wrangling and initial exploratory analyses for the purpose of gaining insights into the data.

2 To explore initial data for indications of inherent patterns based on selected key attributes–specialisation, level of study, gender and their potential impact on performance.

3 To assess the performance of a novel algorithm based on the mechanics of a standard clustering algorithm.

4 To highlight pathways for educationists, data scientists and other researchers to follow in engaging policy makers, development stakeholders and the general public in putting generated data to use.

5 To share findings with colleagues across disciplines and contribute towards unification scientific research.

1.3 Preliminary Studies

Attwell and Pumilia [6] emphasised the need for forging pedagogical competences in analysing and sharing results across disciplines. They particularly reiterated the use of open–source material in higher education, mainly for providing scholars and learners with easy access to data, information and knowledge. Data–driven investigations into aspects of teaching, learning and assessment have attracted interests of many researchers and professionals, not least educationists and data analysts for many years. This paper looks at the two as homing in to a common interdisciplinary problem and solution. While the former seek to enhance the learning process, the latter focus mainly on the tools, techniques that are deployed for learning enhancements. On face value, the two may be seen as representing soft and technical skills respectively, but together they form an interdisciplinary fabric upon which the learning process can thrive. In recent years, interdisciplinarity has been widely promoted as a learning methodology. For example, Aikat et al. [12] see an interdisciplinary gap in graduate education, as it "...remains largely focused on individual achievement within a single scientific domain." They argue that lacking interdisciplinary pedagogy deprives students of data-oriented approaches that could help them ”...translate scientific data into new solutions to today’s critical challenges.” Thus, they propose a data-centered pedagogy for graduate education that unifies the efforts of the educationist and the data scientist. This paper has been strongly influenced by the foregoing narratives [6, 12], which despite a ten–year gap between them, they didn’t exhibit a strong data–driven evidence. In searching for potentially useful information in the students data attributes, we shall be adopting their narrative.

2 Methods

We present the study methodology as a collection of projects, relating to cause-effect relationship between knowledge & development in a spatio–temporal context. The methodology, described below, focuses on gaining insights into the learning fabric of the sampled students, using identifiable attributes as drivers, to learn the concept via unsupervised. Its original ideas are in [8, 9], where it has been applied to map and deliver knowledge about societal SDG clusters.

2.1 Data Sources

A total of 3639 observations of individual students on 19 variables were obtained from a University data repository, in the United Arab Emirates, spanning across the period 2005 through 2016 inclusive. The final data attributes, summarised in Table 1, were the result of a laborious data preparation and cleaning process involving 4366 observations.

CODE	VARIABLE	TYPE	DESCRIPTION
IST	**Institution**	Character	The University where students are registered for their studies
GDR	**Gender**	Binary	Student gender
NTA	**Nationality**	Character	Home country of the student
CPS	**Campus**	Character	University campus where the student studies
TYP	**Type**	Character	Either started and continue or transferred from elsewhere
LVL	**Level**	Character	Level of study as in diploma, first degree or postgraduate
SPC	**Specialization**	Character	The broad specialization associated with student’s major
MJR	**Major**	Character	Student’s specific field of study
PCD	**ProgramCredits**	Numeric	Total number of credits on transcript counting to graduation
RCP	**RegCreditsPrev**	Numeric	Credits registered beginning of the previous Spring term
PVC	**PrevCreditsComplete**	Numeric	Credits completed successfully in the previous Spring term
RGC	**RegCredits**	Numeric	Credits registered for in the current academic period
CMC	**CumulativeCredits**	Numeric	Cumulative Credits over semesters
CGP	**CumulativeGPA**	Numeric	Cumulative GPA from the beginning to latest enrolment
QES	**QualifyingExitScore**	Percentage	Score from qualifying award- i.e, high school students GPA
INT	**InternSector**	Character	Industry, sector of the organization providing internship
BSG	**BeforeSemGPA**	Numeric	Recorded GPA before internship
ISG	**InSemGPA**	Numeric	Recorded in-semester GPA
ASG	**AfterSemGPA**	Numeric	Recorded GPA after internship

Table 1: Selected students’ data attributes

2.2 Implementation Strategy

Implementation strategy is driven by model optimisation achieved by harmonising data variability through Sampling-Measuring-Assessing (SMA) **Algorithm 1**[8, 9, 13]. The algorithm can be adapted for both unsupervised and supervised modelling scenarios and, in a typical unsupervised learning, where the goal is to cluster data objects according to some measures of homogeneity (heterogeneity), the focus is on parameter estimation and likelihood. Implementation starts with Exploratory Data Analysis (EDA), presenting the data in Table 1 in Fisher's correlation form as follows

$$\eta^2 = \frac{\sum_i^n \left(\hat{y} - \bar{y}\right)^2}{\sum_i^n \left(y_i - \bar{y}\right)^2} = 1 - \frac{SSE}{SST} = 1 - \frac{||e||^2}{\left(y'y - \eta\bar{y}\right)} = 1 - \frac{\sum_i^n e_i^2}{\sum_i^n \left(\hat{y} - \bar{y}\right)^2} \tag{1}$$

Equation 1 holds in a multiple regression scenario, where the deviations between the fitted values and the mean are replaced by the deviations due to the linear relationship Kim and Timm [14]. We can use cluster analysis [15] and [16] to group students according to this type of similarity measures. That is, given data $\mathbf{X} = [x_{i,j}]$ and, assuming k distinct clusters, i.e., $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$, each with a specified centroid, for each of the vectors $j = 1, 2, \ldots 10$, we can obtain the distance from $\mathbf{v}_j \in \mathbf{X}$ to the nearest centroid from each of the remaining points in set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\}$ as

$$\mathcal{D}_j\left(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\right) = \min_{1 \le l \le k} d\left(\mathbf{x}_l,\ \mathbf{v}_j\right) \tag{2}$$

where $\mathbf{x}_k \in \mathbf{X}$ and $d\left(.\right)$ is an adopted measure of distance and the clustering objective would then be to minimise the sum of the distances from each of the data points in $\mathbf{X}$ to the nearest centroid. That is, optimal partitioning of $\mathcal{C}$ requires identifying k vectors $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_k^* \in \mathbb{R}^n$ that solve the continuous optimisation function in Equation 3.

$$\min_{\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\} \in \mathbb{R}^n} f\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\right) = \sum_{j=1}^{p} \mathcal{D}_j\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\right) \tag{3}$$

Minimisation will depend on the initial values in $\mathcal{C}$ and hence if we let $z_{i=1,2,\ldots,n}$ be an indicator variable denoting group membership with unknown values, the search for the optimal solution can be through iterative smoothing of the random vector $x|(z = k)$, for which we can compute $\bar{\mu} = \mathbf{E}(x)$ and $\delta = \{\mu_k - \bar{\mu}|y = k \in \mathbf{c}_z\}$. Given labelled data, EDA outputs provide insights into the overall behaviour of the data particularly how the attributes relate to the target variable. Typically, SMA then learns the model in Equation 4, where D is the underlying distribution.

$$F\left(\phi\right) = \underbrace{(P)}_{x,y \sim D}\ \left[\phi\left(x\right) \neq y\right] \tag{4}$$

Algorithm 1 SMA-Sample, Measure, Assess

```
procedure SMA
    Set X = $[x_{i,j}]$ : Accessible Data Source
    Learn $F(\phi) = \underbrace{(P)}_{x,y\sim D} [\phi(x) \neq y]$ based on a chosen learning model
    Set the number of iterations to a large number $K$
    Initialise: $\Theta_{tr} := \Theta_{tr}(.)$ : Training Parameters
    Initialise: $\Theta_{ts} := \Theta_{ts}(.)$ : Testing Parameters
    Initialise: $\Pi_{cp} := \Pi_{cp}(.)$ : Comparative Parameters
    Initialise: $s$ as a percentage of $[x_{\nu,\tau}]$ , say 1%
    $s_{tr}$ : Training Sample $[x_{\nu,\tau}] \leftarrow [x_{i,j}]$ extracted from X = $[x_{i,j}]$
    $s_{ts}$ : Test Sample $[x_{\nu,\tau}] \leftarrow [x_{l \neq i,j}]$ extracted from X = $[x_{i,j}]$
    for $i := 1 \to K$ do: Set K large and iterate in search of optimal values
        while $s \le 50\%$ of $[x_{\nu,\tau}]$ do Vary sample sizes to up to the nearest integer 50% of $X$
            Sampling for Training: $s_{tr} \leftarrow X$
            Sampling for Testing: $s_{ts} \leftarrow X$
            Fit Training and Testing Models $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ with current parameters
            Update Training Parameters: $\Theta_{tr}(.) \leftarrow \Theta_{tr}$
            Update Testing Parameters: $\Theta_{ts}(.) \leftarrow \Theta_{ts}$
            Compare: $\Phi(.)_{tr}$ with $\Phi(.)_{ts}$ : Plotting or otherwise
            Update Comparative Parameters: $\Pi(.)_{cp} \leftarrow \Phi(.)_{tr,ts}$
            Assess: $P\left(\Psi_{D,POP} \ge \Psi_{B,POP}\right) = 1 \iff \mathbb{E}\left[\Psi_{D,POP} - \Psi_{B,POP}\right] = \mathbb{E}\left[\Delta\right] \ge 0$
        end while
    end for
    Output the Best Models $\hat{\mathcal{L}}_{tr,ts}$ based on $\mathbb{E}\left[\Delta\right] \ge 0$
end procedure
```

The SMA algorithm also caters for association rules, which can be used to investigate associations among the data attributes in Table 1 and data clustering, for investigating variations among the variables and the naturally emerging natural structures. The estimates can be obtained in various ways, one of the most common method is the Metropolis-Hastings algorithm, based on the original ideas of Markov Chain Monte Carlo (MCMC) simulation techniques [17], that allow for sampling from probability distributions as long as the density function can be evaluated.

2.3 Sequence of Analyses

Implementation goes through a sequence of logical steps. We deploy Exploratory Data Analysis (EDA) to provide initial insights into the general behaviour of the student data attributes. Ideally, EDA should guide through understanding interpretation of the analyses and results from data visualisation and other summaries. Based on the data insights from EDA, we adopt unsupervised modelling is implemented by deploying **Algorithm 1** based on Affinity Propagation Clustering (APC) algorithm as originally described in [18] and illustrated in [19]. Its original ideas are to merge data clusters until satisfactory levels of similarity (or dissimilarity) are achieved. This type of cluster merging is only possible if the dataset has inherent clusters not less than the initial number stipulated by the algorithm, hence the rationale for EDA. Further, it should be possible to repeatedly extract samples from the data that could then be merged into a cluster. Frey and Dueck [18] describe the merged clusters as exemplars that maximize the levels of average similarity. By repeated sampling and validation, we shall gain a better understanding of the influential factors in the formation of clusters. In the next exposition, we describe the mechanics of propagated clustering as deployed via **Algorithm 1**[8, 9]. If we let

$$\mathbf{X} = [x_{i,j}]\,, \text{where } \; i = 1, 2, ..., n \; \text{ and } \; j = 1, 2, \ldots, p \tag{5}$$

be the source dataset, with assumed k distinct clusters, we can extract repeatedly extract samples based on indicator variables $z_i = 1, 2, \ldots, n_z$ and $s_i = 1, 2, \ldots, n_s$, such $n_z + n_s \ll n$, as the initial potential joint examplar [**exemp**(z, s)] as the sample that maximizes the average similarity to all samples in the joint cluster $C[z \cup s]$, that is:

$$\textbf{exemp}(z, s) = \underset{i \in C[z \cup s]}{\operatorname{argmax}} \frac{\sum_{j \in C[z \cup s]} \mathcal{D}_{i,j}}{n_z + n_s} \tag{6}$$

where $\mathcal{D}_{i,j}$ is the similarity matrix with the indices corresponding to the i^{th} and j^{th} items in the two samples. The choice of the measure of similarity is application–dependent and user–defined. Then the merging objective is computed as

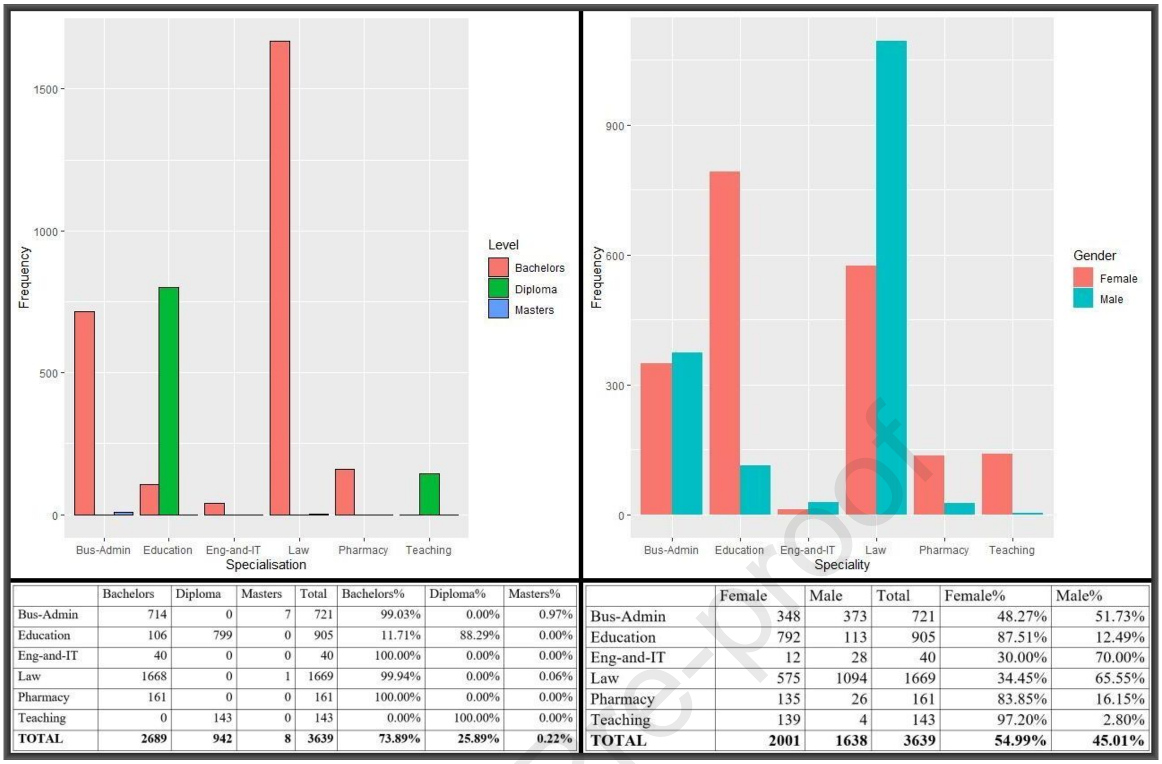
$$\textbf{obj}(z, s) = \frac{1}{2}\left[\frac{\sum_{\rho \in z} \mathcal{D}_{\textbf{exemp}(z,s)\rho}}{n_z} + \frac{\sum_{\nu \in s} \mathcal{D}_{\textbf{exemp}(z,s)\nu}}{n_s}\right] = \frac{n_s \sum_{\rho \in z} \mathcal{D}_{\textbf{exemp}(z,s)\rho} + n_s \sum_{\nu \in s} \mathcal{D}_{\textbf{exemp}(z,s)\nu}}{2 n_z n_s} \tag{7}$$

3 Implementation, Analyses and Results

Implementation goes through a sequence of logical steps. Insights gained from Exploratory Data Analysis (EDA) guide the applications of **Algorithm 1** based on Affinity Propagation Clustering algorithm as originally described in [18] and illustrated in [19]. EDA plays a crucial role in defining the research problem and objectives. We adopt it here as an initial step in grouping students according to some measures of similarity.

3.1 Graphical Data Visualisation

The two panels in Figure 1 provide basic insights into existing frequency structures in the data based on three key attributes–specialisation, level of study and gender. The most popular courses are law, education and business administration at bachelors and diploma levels. Females have a significant representation in the three most popular courses. They dominate in education, have a fare share in business administration and they make over 34% of law enrolment.



	Bachelors	Diploma	Masters	Total	Bachelors%	Diploma%	Masters%
Bus-Admin	714	0	7	721	99.03%	0.00%	0.97%
Education	106	799	0	905	11.71%	88.29%	0.00%
Eng-and-IT	40	0	0	40	100.00%	0.00%	0.00%
Law	1668	0	1	1669	99.94%	0.00%	0.06%
Pharmacy	161	0	0	161	100.00%	0.00%	0.00%
Teaching	0	143	0	143	0.00%	100.00%	0.00%
TOTAL	**2689**	**942**	**8**	**3639**	**73.89%**	**25.89%**	**0.22%**

	Female	Male	Total	Female%	Male%
Bus-Admin	348	373	721	48.27%	51.73%
Education	792	113	905	87.51%	12.49%
Eng-and-IT	12	28	40	30.00%	70.00%
Law	575	1094	1669	34.45%	65.55%
Pharmacy	135	26	161	83.85%	16.15%
Teaching	139	4	143	97.20%	2.80%
TOTAL	**2001**	**1638**	**3639**	**54.99%**	**45.01%**

Figure 1: Underlying distributions in the original students data

Alongside the key performance metrics, we shall use the baseline statistics above as the focal points of our analyses. The six panels in Figure 2 provide the underlying distributional patterns of the Grade Point Average (GPA) metric and they generally provide a rough idea about the number of clusters, hence highlighting the path towards unsupervised modelling. Our implementation strategy is driven by the structures in the two Figures 1 and 2.

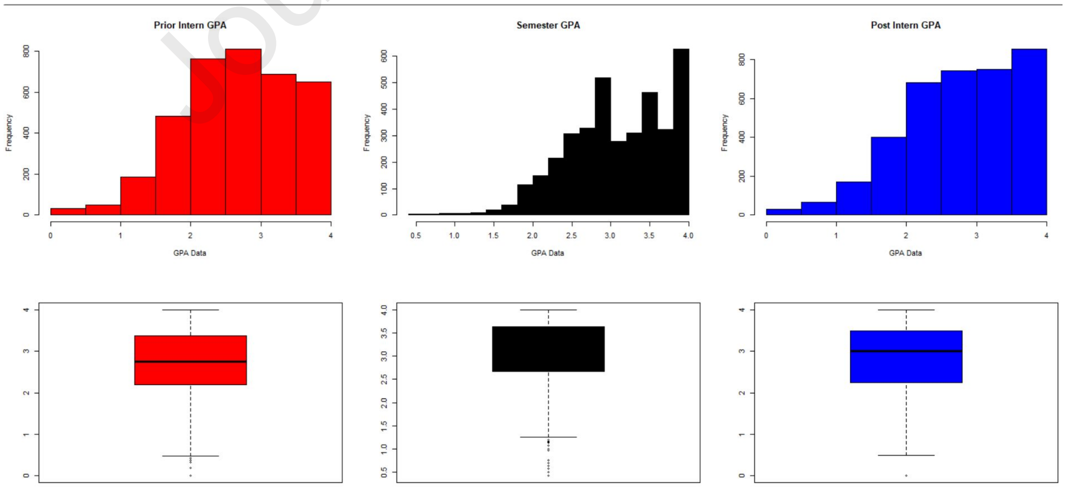


Figure 2: Histogram and boxplots for the GPA records before, in–semester and after internship

The six panels in Figure 2 exhibit the overal GPA distributions between prior and post–intern semesters, appearing to be fairly similar. As our interest is in detecting naturally arising structures in, we can examine the distributions from different bandwidths. Figure 3 shows that only at very low bandwidths we can detect underlying structured in each of the GPA category–more pronounced in the before semester than in the other two.

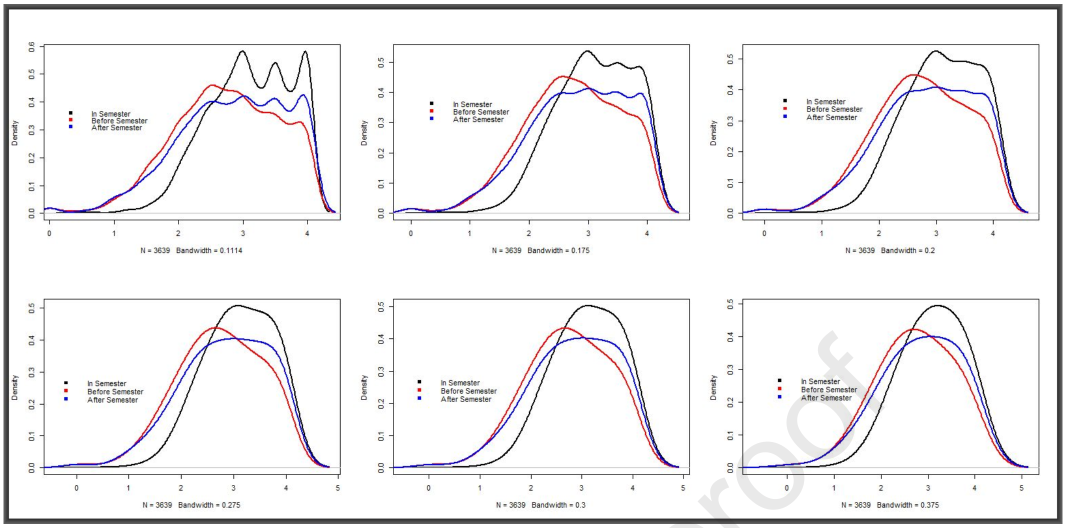


Figure 3: Densities for the three GPA classes plotted at different bandwidths

The average GPAs before, in–semester and after semester are 2.74, 3.11 and 2.84 respectively, suggesting either spurious clusters or masking in the top left panel in Figure 3. In the next exposition, we carry out further explorations by looking at the densities of the individual dominating categories–Law, Education and Business Administration.

3.2 Unsupervised Modelling

The Affinity Propagation Clustering algorithm generated heavily overlapping clusters for the GPA data. Figure 4 show patterns for two, three, four and five clusters, clock–wise from top left respectively. They both indicate a separation not based on the average GPA. Hence, we take a closer look at the data to establish the basis of the clusters' formation.

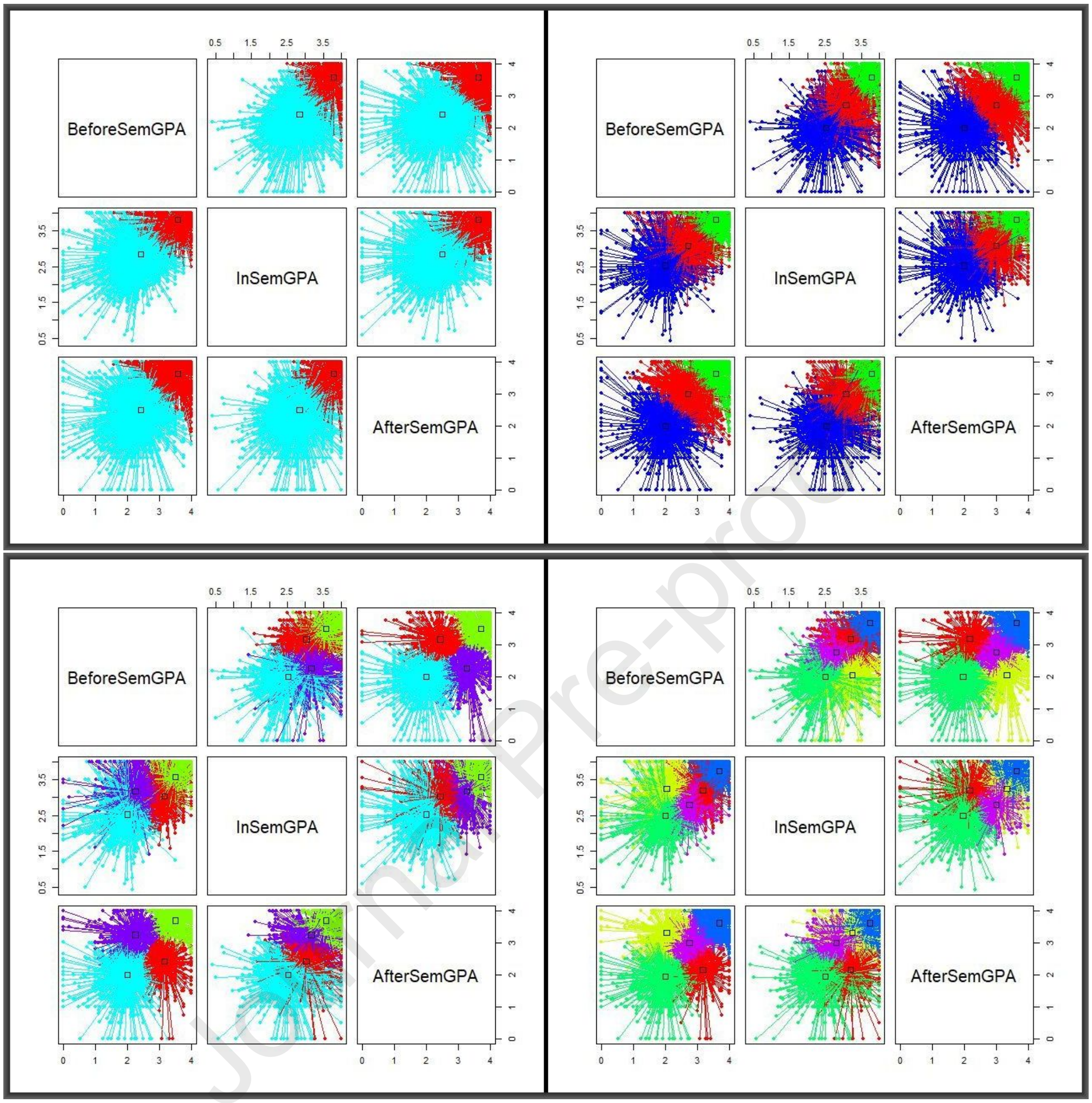


Figure 4: From top left clock–wise the four panels exhibit two, three, four and five clusters respectively

Table 2a shows the proportions of of cases, based on selected attributes, in each of the detected clusters. The rows in the first column, coded as **C12, C22** for cluster 1 and 2 in the two cluster group, to **C15** through **C55** for cluster 1 through 5 for the five cluster group. The remaining columns represent data from the attributes **Specialisation** and **Gender**. Table 2b shows the average GPA levels in each of the selected categories. These statistics are potentially useful in the sense that the choice of a course, specialisation and performance are conditional on various factors including the quality of teaching and delivery, course organisation and general management as well as assessment and feedback students receive[20]. Such statistics could help the CHEDS [10] in the UAE in making evidence-based decisions to guide and influence higher education policies and planning at all levels.

	Bus. Adm.	Educ.	Law	Female	Male
C12	0.216	0.233	0.458	0.541	0.458
C22	0.187	0.257	0.458	0.554	0.445
C13	0.214	0.245	0.451	0.535	0.464
C23	0.208	0.237	0.463	0.551	0.448
C33	0.168	0.262	0.463	0.566	0.433
C14	0.211	0.254	0.452	0.560	0.439
C24	0.215	0.233	0.459	0.548	0.451
C34	0.166	0.272	0.454	0.564	0.435
C44	0.201	0.234	0.468	0.522	0.477
C15	0.206	0.252	0.477	0.545	0.454
C25	0.200	0.210	0.505	0.486	0.513
C35	0.167	0.268	0.460	0.556	0.443
C45	0.213	0.232	0.461	0.552	0.447
C55	0.207	0.268	0.411	0.580	0.419

(a)

Bus. Adm.	Educ.	Law	Female	Male
3.553	3.564	3.551	3.561	3.550
2.521	2.493	2.505	2.500	2.508
2.930	2.889	2.915	2.905	2.921
3.662	3.655	3.640	3.650	3.652
2.139	2.187	2.188	2.192	2.173
2.869	2.835	2.858	2.849	2.861
3.627	3.639	3.621	3.629	3.627
2.085	2.161	2.134	2.151	2.126
2.877	2.860	2.863	2.849	2.878
2.859	2.813	2.821	2.812	2.843
2.903	2.867	2.877	2.847	2.901
2.044	2.117	2.100	2.108	2.095
3.657	3.669	3.643	3.652	3.656
2.856	2.830	2.858	2.843	2.855

(b)

Table 2: Sampled enrolment proportions and GPA averages in selected categories in Table 2a & 2b respectively

The two panels in Figure 5 correspond to values in Table 2a and 2b respectively. The horizontal axis on the left hand side panel corresponds to the three specialisation categories and gender in the order given in the two tables and the vertical axis represents the category percentage. The horizontal axis on the right hand side panel displays the 14 clusters as shown in the first column of Table 2a, while the vertical axis shows the average GPAs. By visual inspection through the line, cluster overlapping is evident–those on the same horizontal line have similar scores.

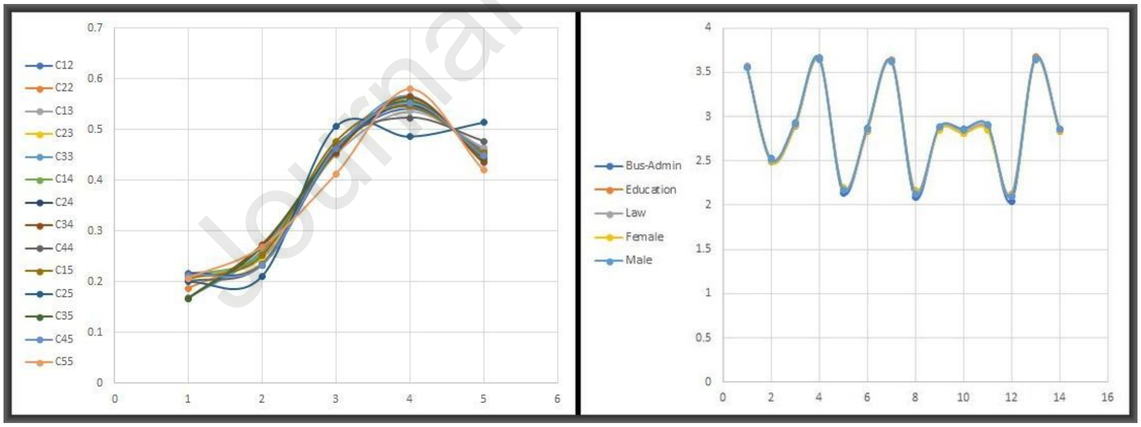


Figure 5: Enrolment proportions on the LHS panel and the GPA performance on the RHS

As noted earlier, the clusters in Figure 4 heavily overlap. Thus, to determine the optimal number of clusters in the sampled data, we refer back to the densities in Figure 3 and the enrolment proportions and GPA averages in Fugure 5. Repeated sampling through **Algorithm 1** yields in the consistent GPA average performance densities in Figure 6.

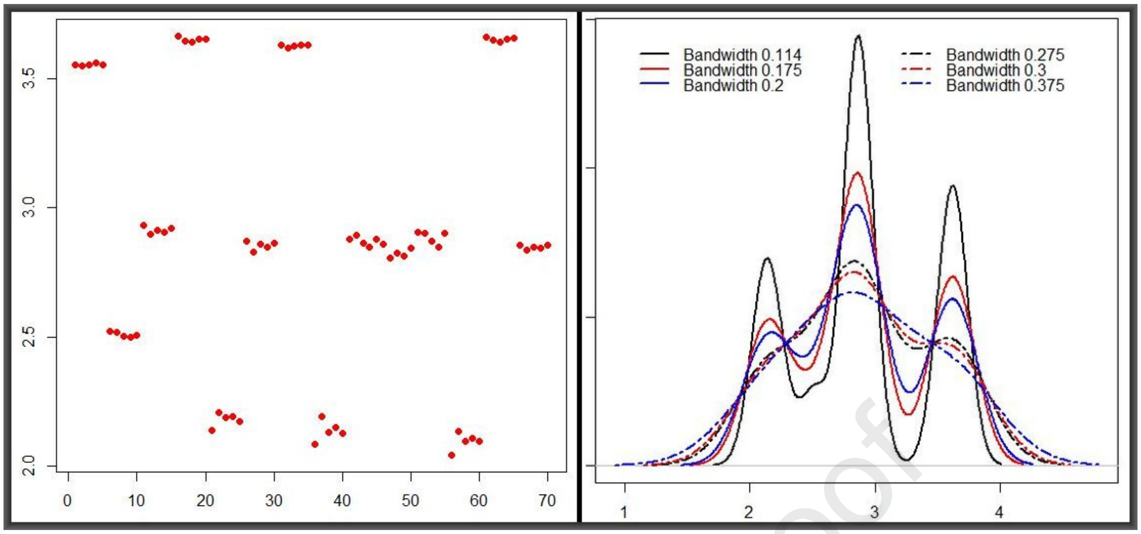


Figure 6: An optimal 3–cluster structure for the average GPA over multiple runs at different bandwidths

The patterns in Figure 6 are the best representations of the underlying structure in the sampled data. They were obtained based on multiple runs of sampling through the data inside the clusters in Figure 4. Both panels present a clear conclusion that in terms of GPA performance based on the sampled data, we can isolate three distinct clusters, centred around GPAs of 3.5, 3.0 and between 2.5 and 2.0. It is important to note that three clusters are dependent on both level and gender, which the two panels do not distinguish. While a table detailing the dominance in each category may be useful, it is imperative to interpret such data in conjunction with other relevant attributes, such as the left hand side panel of Figure 5. The data for each of the 14 clusters is available for potential future examinations.

4 Concluding Remarks and General Discussions

The paper sought to address a two–fold problem. On the one hand, it focused on the technical aspects of Big Data Modelling, for which it deployed the affinity clustering algorithm [18, 19] based on the mechanics of the SMA algorithm [8, 9]. On the other hand, it focused on the soft, interdisciplinary aspects of BDM–i.e., applying machine learning techniques to real–life applications in an interdisciplinary context. Objectives 1 through 3 were met in sub–sections 3.1 and 3.2. It is imperative to note that more analyses could have been carried out based on the settings in this paper. However, the scope for this application was confined to 3 of the original 19 attributes–i.e., Specialisation, Level and Gender, so as to accommodate the technical aspect of the set objectives under limited interdisciplinary interpretations. The findings presented in this paper are therefore intended to fulfil objectives 4 and 5–i.e., they should open new discussions and highlight novel paths for interdisciplinary research involving data scientists and educationists.

Even within this limited application, our findings show that there are great potentials in incorporating interdisciplinary approaches in university curricula, bringing together domain sciences on the one side and data science on the other. Further tests and evaluations of the SMA algorithm can conducted using a wide range of unsupervised and supervised techniques, with any combination of the 19 data attributes. **Algorithm 1** is also capable of handling association rules–originally developed for analysing shopping transactions–see Agrawal et al. [21]. In this particular application, association rules can play a unifying role between unsupervised and supervised modelling in that they can capture underlying rules of association among the students' data attributes. We expect the technical and soft aspects of the paper to increasingly attract attention to collaborative, interdisciplinary research activities in various sectors.

Finally, and as emphasised by Aikat et al. [12], our paper showed, via real data, that uncovering attainment and performance triggers cannot be confined to silos of domain knowledge, neither to algorithms developed by data scientists. A unified understanding can only be achieved through cross–institutional collaborative research, sharing data and findings. The outcomes of this work should provide useful inputs to the Center for Higher Education Data and Statistics

(CHEDS) of the United Arab Emirates in forging interdisciplinarity for educational performance enhancement.

Abbreviations

APC	Affinity Propagation Clustering
BDM	Big Data Modelling
BDMSDG	Big Data Modelling of Sustainable Development Goals
CAA	Commission for Academic Accreditation
CHEDS	Center for Higher Education Data and Statistics
EDA	Exploratory Data Analysis
GPA	Grade Point Average
MCMC	Markov Chain Monte Carlo
MoE	Ministry of Education
PCA	Principal Component Analysis
SILPA	Standards for Institutional Licensure and Program Accreditation
SMA	Sample-Measure-Assess
UAE	United Arab Emirates

Declarations

This section provides declarations relating to data availability, competing interests, funding, author contribution and acknowledgements. These declarations comprehensively cover the positions of each of the two authors of the paper.

Data Availability

As noted in Section 3.2, the data attributes used in this study were obtained via a semi–automated random selection and cleaning process by the authors. They were reformatted to fit in with the adopted modelling strategy–hence, the data is only available from the authors, who have retained both the raw and modified copies, should they be requested.

Competing Interests

Both authors declare that there are no competing interests in publishing this paper, be they financial or non–financial.

Funding

This work has not been supported by any grant, but rather it is an outcome of ordinary Research and Scholarly Activities (RSA) allocation to each of the two authors by their respective institutions.

Author Contribution

As a result of previous joint work, both authors contributed equally to this work–with KSM carrying out most of the data cleaning and automated selection and RAS providing the raw data and many of the insights into designing the analyses layout based on his experiences with the education system in the UAE.

Acknowledgements

This paper is part of on–going initiatives towards Big Data Modelling of Sustainable Development Goals (BDMSDG) and the application of the DSF both authors have been involved in over the last three years. We would like to thank many individuals and institutions who have discussed these initiatives with us at different stages of development. We particularly acknowledge the role of the Data Intensive Research Initiative of South Africa (DIRISA), through the South African Council for Scientific and Industrial Research (CSIR) (https://www.csir.co.za/), who have invited us a couple of times to Pretoria to present our findings. We also acknowledge the presentation opportunity we have had with the Joint Support-Center for Data Science Research (DS), through the Japanese Polar Environment Data Science Center (PEDSC) (http://pedsc.rois.ac.jp/en/), the United Nations World Data Forum (UNWDF) (https://unstats.un.org/unsd/undataforum/index.html) and the Sussex Sustainability Research Programme (SDG interactions) of the University of Sussex (https://www.sussex.ac.uk/ssrp/research/sdg-interactions). Most importantly, we are grateful to CHEDS in the MoE of the UAE for providing the raw data for this work.

References

[1] SDG, Sustainable Development Goals. 2015; `https://www.un.org/sustainabledevelopment/sustainable-development-goals/`.

[2] Miguis, V. L.; Freitas, A.; Garcia, P. J. V.; Silva, A. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems* **2018**, *115*, 36 – 51.

[3] Brooks, C.; Erickson, G.; Greer, J.; Gutwin, C. Modelling and quantifying the behaviours of students in lecture capture environments. *Computers & Education* **2014**, *75*, 282 – 292.

[4] Hua Leong, F.; Marshall, L. Modeling engagement of programming students using unsupervised machine learning technique. *GSTF Journal on Computing (JoC)* **2018**, *6*.

[5] SILPA, Standards For Institutional Licensure And Program Accreditation. *Ministry of Education, UAE* **2019**,

[6] Attwell, G.; Pumilia, P. The New Pedagogy of Open Content: Bringing Together Production, Knowledge, Development, and Learning. *Data Science Journal* **2007**, *6*, S211S219.

[7] Meusburger, P. In *Knowledge and the Economy*; Meusburger, P., Glückler, J., el Meskioui, M., Eds.; Springer Netherlands: Dordrecht, 2013; pp 15–42.

[8] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. *International Workshop on Data Science-Present and Future of Open Data and Open Science, 12-15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan* **2018**,

[9] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. An Interdisciplinary Data-Driven Framework for Development Science. *DIRISA National Research Data Workshop, CSIR ICC, 19-21 June 2018, Pretoria, RSA* **2018**,

[10] CHEDS, Center For Higher Education Data And Statistics. *Ministry of Education, UAE* **2018**,

[11] Parsons, M. A.; Øystein Godøy,; LeDrew, E.; de Bruin, T. F.; Danis, B.; Tomlinson, S.; Carlson, D. A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science* **2011**, *37*, 555–569.

[12] Aikat, J.; Carsey, T. M.; Fecho, K.; Jeffay, K.; Krishnamurthy, A.; Mucha, P. J.; Rajasekar, A.; Ahalt, S. C. Scientific Training in the Era of Big Data: A New Pedagogy for Graduate Education. *Big Data* **2017**, *5*.

[13] Mwitondi, K. S.; Said, R. A.; Zargari, S. A. A robust domain partitioning intrusion detection method. *Journal of Information Security and Applications* **2019**, *48*, 102360.

[14] Kim, K.; Timm, N. *Univariate and Multivariate General Linear Models*; New York: Chapman and Hall/CRC, 2006.

[15] Chapmann, J. *Machine Learning Algorithms*; CreateSpace Independent Publishing Platform, 2017.

[16] Kogan, J. *Introduction to Clustering Large and High-Dimensional Data*; Cambridge University Press, 2007.

[17] Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.

[18] Frey, B. J.; Dueck, D. Clustering by Passing Messages Between Data Points. **2007**, *315*, 972–976.

[19] Bodenhofer, U.; Kothmeier, A.; Hochreiter, S. APCluster: An R Package for Affinity Propagation Clustering. *Bioinformatics* **2011**, *27*, 2463–2464.

[20] Burgess, A.; Senior, C.; Moores, E. A 10-year case study on the changing determinants of university student satisfaction in the UK. *PLOS ONE* **2018**, *13*, 1–15.

[21] Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.* **1993**, *22*, 207–216.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: