[Array 11 (2021) 100064](https://doi.org/10.1016/j.array.2021.100064)

Contents lists available at [ScienceDirect](http://www.sciencedirect.com/science/journal/25900056)

Array

journal homepage: [www.elsevier.com/journals/array/2590-0056/open-access-journal](http://www.elsevier.com/journals/array/2590-0056/open-access-journal)

An integrated clustering method for pedagogical performance

Raed A. Said [a](#_bookmark0), Kassim S. Mwitondi [b](#_bookmark1),[\*](#_bookmark3)

a *Canadian University Dubai, United Arab Emirates*

b *Sheffield Hallam University, College of Business, Technology* & *Engineering, UK*

A R T I C L E I N F O

*Keywords:* Association rules Big data

CHEDS

Data mining Data science Internship

Interdisciplinarity Pedagogy

Propagated clustering SILPA

SMA Algorithm

Sustainable development goals Unsupervised modelling

A B S T R A C T

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 pedagogical 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.

1. Introduction

The United Nations Sustainable Development Goals SDG [[1](#_bookmark21)] iden- tifies good quality education as the foundation for creating sustainable development, improved quality of life, innovation and creativity. In- vestment 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 an algorithm designed for machine learning tech- niques. The main 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](#_bookmark22)], Brooks et al. [[3](#_bookmark23)] and Hua Leong and Marshall [[4](#_bookmark24)], 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 vari- ables) 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](#_bookmark25)]. It is clearly stipulated in SILPA

[[5](#_bookmark25)] 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 major's curriculum and as such they

\* Corresponding author.

*E-mail address:* [k.mwitondi@shu.ac.uk](mailto:k.mwitondi@shu.ac.uk) (K.S. Mwitondi).

<https://doi.org/10.1016/j.array.2021.100064> Received 22 February 2021; Accepted 12 April 2021

Available online 22 April 2021

2590-0056/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

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 fol- lows. Section [1](#_bookmark2) presents the background, motivation, research aim, ob- jectives and a brief review of relevant literature. Section [2](#_bookmark4) details the methods–data description and modelling techniques, followed by implementation, analyses, results and general discussions in Section [3](#_bookmark11). Finally, concluding remarks are drawn in Section [4](#_bookmark20), highlighting poten- tial new research directions.

* 1. *Motivation*

Attaining good quality education is the ideal dream of all learners, institutions and nations across the world Pumilia [[6](#_bookmark26)], Meusburger [[7](#_bookmark27)]. The United Nations identifies good quality education as the foundation for 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](#_bookmark28),[9](#_bookmark29)]. Just what type of data, how much of it and how fast it changes 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 ex- pected 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](#_bookmark30)] 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. Re- ports and analyses will help in advancing students learning experiences and curriculum designs.

* 1. *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 interdisci- plinary context [[11](#_bookmark31)]. 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 me- chanics 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. *Preliminary studies*

Attwell and Pumilia [[6](#_bookmark26)] 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 educa- tion, 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 re- searchers 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](#_bookmark32)], see an interdisci- plinary 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](#_bookmark26),[12](#_bookmark32)], 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 attri- butes, we shall be adopting their narrative.

1. 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 derive from predictive modelling derive from [[8](#_bookmark28), [9](#_bookmark29)], where it has been applied to map and deliver knowledge about so- cietal SDG clusters.

* 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](#_bookmark5), were the result of a laborious data preparation and cleaning process involving 4366 observations.

* 1. *Implementation strategy*

Implementation strategy is driven by model optimisation achieved by harmonising data variability through Sampling-Measuring-Assessing (SMA) [Algorithm 1](#_bookmark7) [[8](#_bookmark28),[9](#_bookmark29),[13](#_bookmark33)]. The algorithm can be adapted for both unsupervised and supervised modelling scenarios and, in a typical un- supervised learning, where the goal is to cluster data objects according to some measures of homogeneity (heterogeneity), the focus is on param- eter estimation and likelihood. Implementation starts with Exploratory Data Analysis (EDA), presenting the data in [Table 1](#_bookmark5) in Fisher's correlation

Table 1

Selected students’ data attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| 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

distribution.

*F*(*ϕ*)= (*P*) [*ϕ*(*x*) /= *y*] (4)

|*x*,{*y*~z*D*}

Algorithm 1

SMA-Sample, Measure, Assess

1: Procedure SMA

2: Set X = [*xi*,*j* ] : Accessible Data Source

3: Learn *F*(*ϕ*) = (*P*) [*ϕ*(*x*) /= *y*] based on a chosen learning model

|{z}

*x*,*y*~*D*

4: Set the number of iterations to a large number *K*

5: Initialise : Θ*tr* := Θ*tr* (.) : Training Parameters 6: Initialise : Θ*ts* := Θ*ts* (.) : Testing Parameters

|  |  |  |  |
| --- | --- | --- | --- |
| LVL | Level | Character | Level of study as in diploma, first |
| SPC | Specialisation | Character | degree or postgraduate  The broad specialisation |
| MJR | Major | Character | associated with student's major  Student's specific field of study |
| PCD | ProgramCredits | Numeric | Total number of credits on |
| RCP | RegCreditsPrev | Numeric | transcript counting to graduation  Credits registered beginning of |
| PVC | PrevCreditsComplete | Numeric | the previous Spring term  Credits completed successfully in |
| RGC | RegCredits | Numeric | the previous Spring term  Credits registered for in the |
| CMC | CumulativeCredits | Numeric | current academic period  Cumulative Credits over |

7: Initialise : Π*cp* := Π*cp* (.) : Comparative Parameters

8: Initialise : *s* as a percentage of [*xν*,*τ* ], say 1%

9: *str* : Training Sample [*xν*,*τ* ] ← [*xi*,*j* ] extracted from X = [*xi*,*j* ]

10: *sts* : Test Sample [*xν*,*τ* ] ← [*xl*/=*i*,*j*] extracted from X = [*xi*,*j*]

11: for *i* := 1 → *K* do: Set K large and iterate in search of optimal values

12: while *s* ≤ 50% of [*xν*,*τ* ] do Vary sample sizes to up to the nearest integer 50% of *X*

13: Sampling for Training : *str* ← *X*

14: Sampling for Testing : *sts* ← *X*

s *L tr*,*ts*∝Φ(.)*tr*,*ts* with current parameters

c

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | semesters | 15: | Fit Training and Testing Model |  |
| CGP | CumulativeGPA | Numeric | Cumulative GPA from the | 16: | Update Training Parameters : | Θ*tr* (.) ← Θ |
|  |  |  | beginning to latest enrolment | 17: | Update Testing Parameters : | Θ*ts* (.) ← Θ*ts* |

*tr*

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

18: Compare : Φ(.)*tr* with Φ(.)*ts* : Plotting or otherwise 19: Update Comparative Parameters : Π(.)*cp* ← Φ(.)*tr*,*ts*

20: Assess : *P*(Ψ*D*,*POP* ≥ Ψ*B*,*POP* ) = 1 ⇔ E[Ψ*D*,*POP* — Ψ*B*,*POP* ] = E[Δ] ≥ 0

21: end while

22: end for

ASG AfterSemGPA Numeric Recorded GPA after internship

23: Output the Best Models

24: end procedure

*L*c*tr*,*ts* based on E[Δ] ≥ 0

form as follows

P (b — )

*n y y* 2

*η*2 = *i* = 1 — (*y* — *y*)

P*n* 2

*i*

*i*

*SSE SST*

= 1 —

||*e*||2 (*y*' *y* — *ηy*)

*ne*2

P

= 1 — *i i*

P*n* 2

(b*y* — *y*)

*i*

(1)

The SMA algorithm also caters for association rules, which can be used to investigate associations among the data attributes in [Table 1](#_bookmark5) 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-

Equation [(1)](#_bookmark8) holds in a multiple regression scenario, where the de- viations between the fitted values and the mean are replaced by the deviations due to the linear relationship [[14](#_bookmark34)]. We can use cluster analysis

[[15](#_bookmark35)] and [[16](#_bookmark36)] to group students according to this type of similarity measures. That is, given data X = [*xi*,*j*] and, assuming *k* distinct clusters, i.e., *C* = {*c*1, *c*2, …, *ck*}, each with a specified centroid, for each of the vectors *j* = 1, 2, …10, we can obtain the distance from v*j* ∈ X to the nearest centroid from each of the remaining points in set {x1, x2, …x*k*} as

*D j*(x1, x2, …x*k*)= min *d* x*l*, v*j* (2)

1≤*l*≤*k*

where x*k* ∈ X and *d*(.) is an adopted measure of distance and the clus- tering objective would then be to minimise the sum of the distances from each of the data points in X to the nearest centroid. That is, optimal

partitioning of *C* requires identifying *k* vectors x\*, x\*, …, x\* ∈ R*n* that solve the continuous optimisation function in Equation [(3)](#_bookmark9).

1 2 *k*

{x ,x min }∈R*n f* (x1, x2, …, x*k*)= X *D j*(x1, x2, …, x*k*) (3)

*p*

Hastings algorithm, based on the original ideas of Markov Chain Monte

Carlo (MCMC) simulation techniques [[17](#_bookmark37)], that allow for sampling from probability distributions as long as the density function can be evaluated.

* 1. *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](#_bookmark7) based on Affinity Propagation Clustering (APC) algorithm as originally described in [[18](#_bookmark38)] and illustrated in [[19](#_bookmark39)]. Its original ideas are to merge data clusters until satisfactory levels of sim- ilarity (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

1 2 ,…,x*k*

*j*=1

that could then be merged into a cluster Frey and Dueck [[18](#_bookmark38)]. describe

Minimisation will depend on the initial values in *C* and hence if we let *zi*=1,2,…,*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 *μ* = E(*x*) and *δ* = {*μk* —*μ*|*y* = *k* ∈ 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)](#_bookmark6), where *D* is the underlying

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](#_bookmark7) [[8](#_bookmark28),[9](#_bookmark29)]. If we let

X = *xi*,*j* , where *i* = 1, 2, …, *n* and *j* = 1, 2, …, *p* (5) be the source dataset, with assumed *k* distinct clusters, we can extract

repeatedly extract samples based on indicator variables *zi* = 1, 2, …, *nz* and *si* = 1, 2, …, *ns*, such *nz* + *ns* ≪ *n*, as the initial potential joint exam- plar [exemp(*z*, *s*)] as the sample that maximizes the average similarity to all samples in the joint cluster *C*[*z* ∪*s*], that is:

an initial step in grouping students according to some measures of similarity.

* 1. *Graphical data visualisation*

exemp(*z*, *s*)= argmaxP*j*∈*C*[*z*∪*s*] *D i*,*j*

(6)

The two panels in [Fig. 1](#_bookmark12) provide basic insights into existing frequency

*i*∈*C*[*z*∪*s*]

*nz* + *ns*

structures in the data based on three key attributes–specialisation, level of study and gender. The most popular courses are law, education and

where *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 simi- larity is application–dependent and user–defined. Then the merging objective is computed as

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.

Alongside the key performance metrics, we shall use the baseline

obj(*z*, *s*)= 1 P*ρ*∈*z D* exemp(*z*,*s*)*ρ* + P*ν*∈*s D* exemp(*z*,*s*)*ν* = *ns* P*ρ*∈*z D* exemp(*z*,*s*)*ρ* + *ns* P*ν*∈*s D* exemp(*z*,*s*)*ν*

(7)

2 *nz ns* 2*nzns*

1. 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](#_bookmark7) based on Affinity Propagation Clustering algorithm as originally described in [[18](#_bookmark38)] and illustrated in [[19](#_bookmark39)]. EDA plays a crucial role in defining the research problem and objectives. We adopt it here as

statistics above as the focal points of our analyses. The six panels in [Fig. 2](#_bookmark13) 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 [Figs. 1](#_bookmark12) [and 2](#_bookmark12).

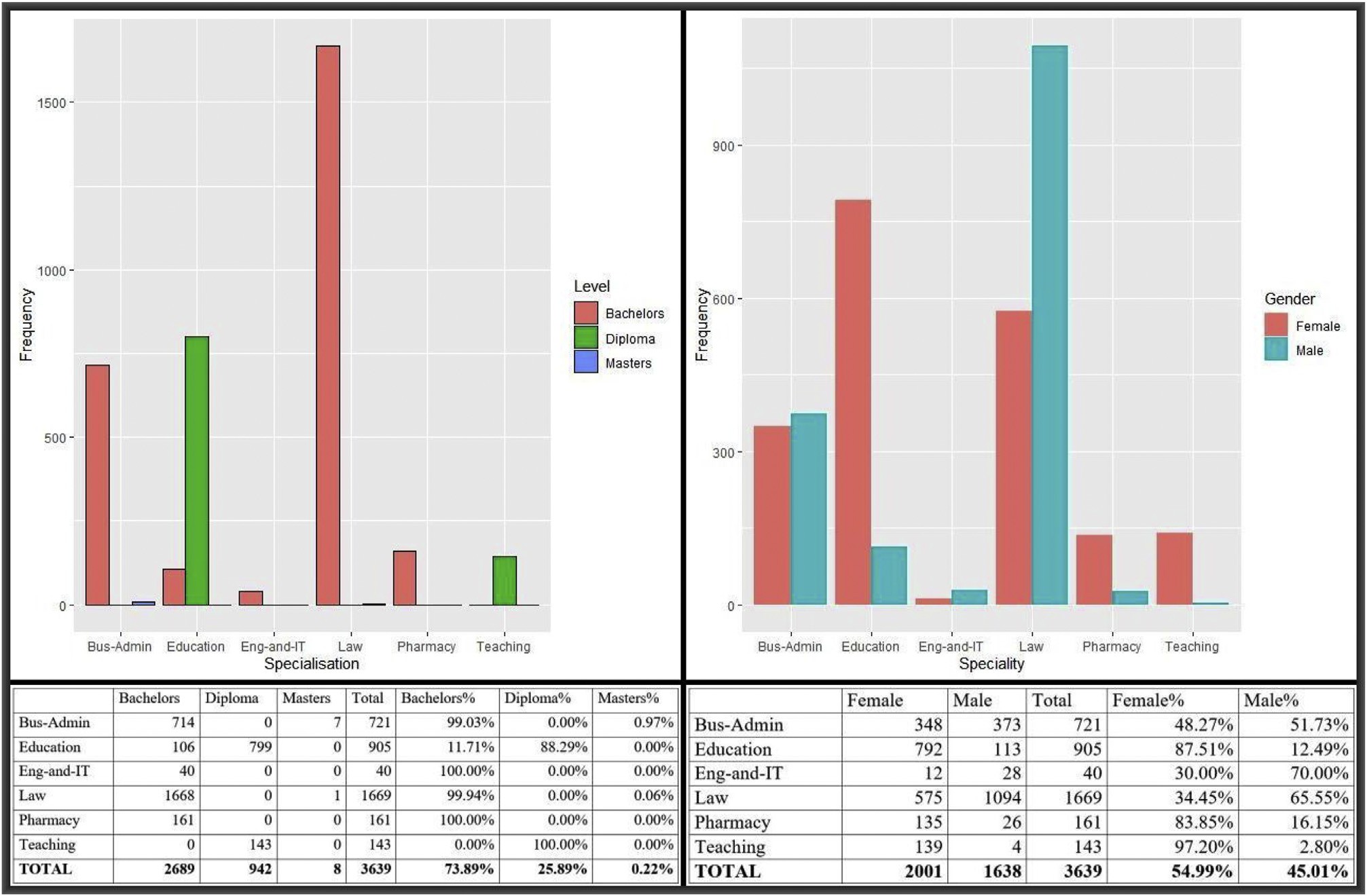


Fig. 1. Underlying distributions in the original students data.

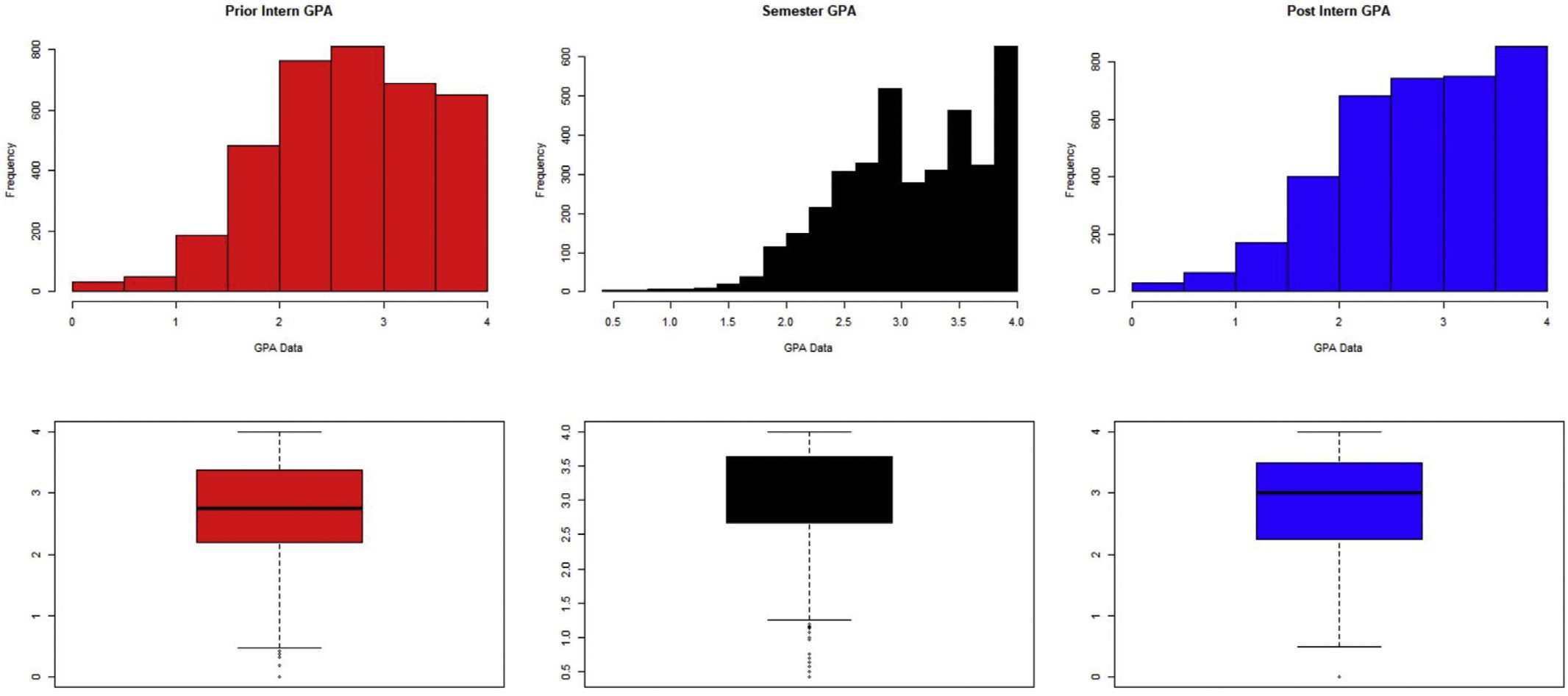


Fig. 2. Histogram and boxplots for the GPA records before, in–semester and after internship.

The six panels in [Fig. 2](#_bookmark13) exhibit the overall 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. [Fig. 3](#_bookmark14) 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.

The average GPAs before, in–semester and after semester are 2.74,

3.11 and 2.84 respectively, suggesting either spurious clusters or mask- ing in the top left panel in [Fig. 3](#_bookmark14). In the next exposition, we carry out further explorations by looking at the densities of the individual domi- nating categories–Law, Education and Business Administration.

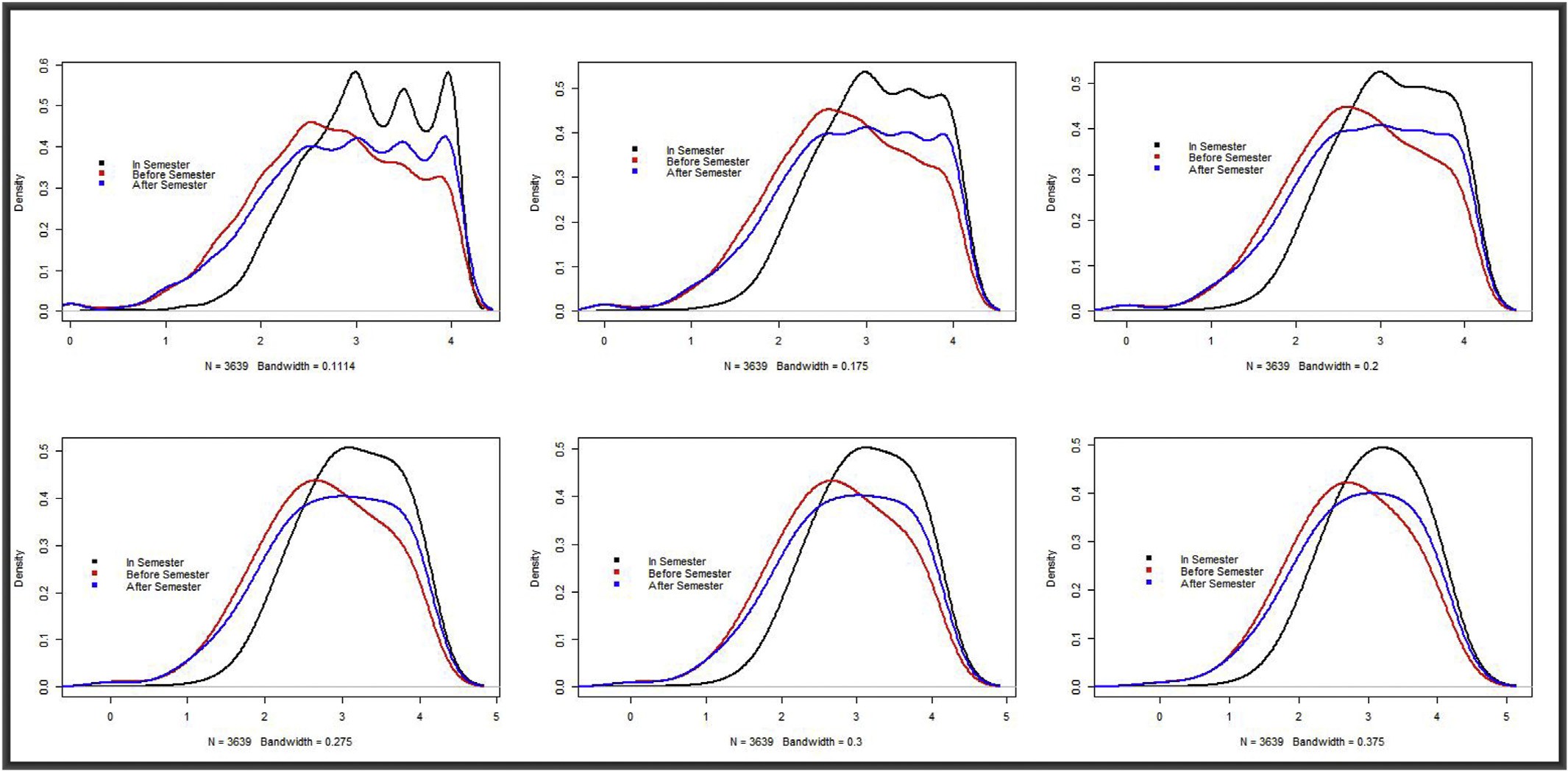


Fig. 3. Densities for the three GPA classes plotted at different bandwidths.

* 1. *Unsupervised modelling*

The Affinity Propagation Clustering algorithm generated heavily overlapping clusters for the GPA data. [Fig. 4](#_bookmark16) 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.

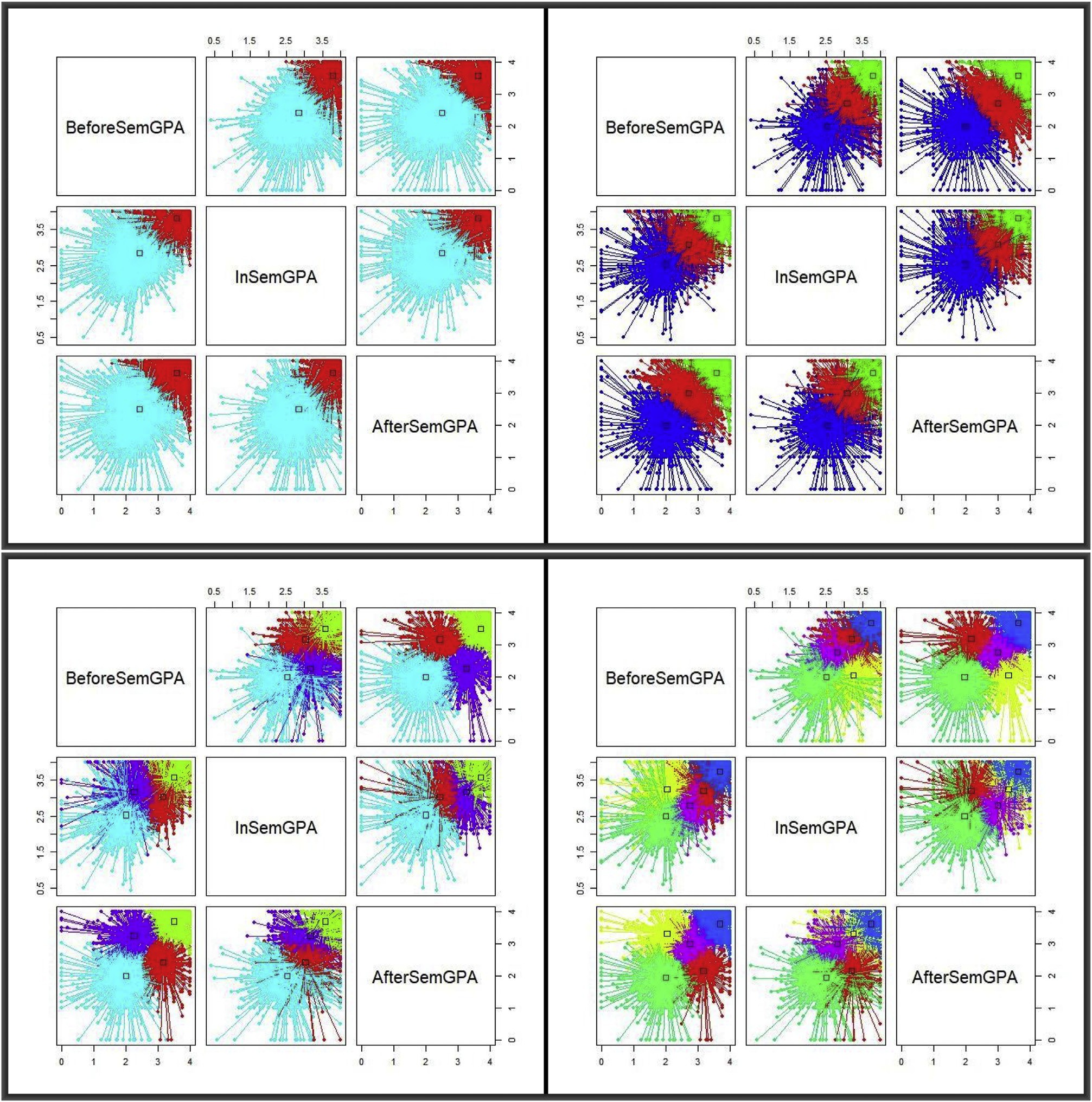


Fig. 4. From top left clock–wise the four panels exhibit two, three, four and five clusters respectively.

[Table 2](#_bookmark17)a shows the proportions 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 2](#_bookmark17)b 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](#_bookmark40)]. Such statistics could help the [[10](#_bookmark30)] in the UAE in making evidence-based decisions to guide and influence higher education pol- icies and planning at all levels.

Table 2

Sampled enrolment proportions and GPA averages in selected categories in [Table 2](#_bookmark17)a and 2b respectively.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | a |  | | | | | |
|  | 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 |  |
|  | b |  |  |  |  |  |  |
|  |  | Bus. Adm. | Educ. | Law | Female | Male |  |
|  | C12 | 3.553 | 3.564 | 3.551 | 3.561 | 3.550 |  |
|  | C22 | 2.521 | 2.493 | 2.505 | 2.500 | 2.508 |  |
|  | C13 | 2.930 | 2.889 | 2.915 | 2.905 | 2.921 |  |
|  | C23 | 3.662 | 3.655 | 3.640 | 3.650 | 3.652 |  |
|  | C33 | 2.139 | 2.187 | 2.188 | 2.192 | 2.173 |  |
|  | C14 | 2.869 | 2.835 | 2.858 | 2.849 | 2.861 |  |
|  | C24 | 3.627 | 3.639 | 3.621 | 3.629 | 3.627 |  |
|  | C34 | 2.085 | 2.161 | 2.134 | 2.151 | 2.126 |  |
|  | C44 | 2.877 | 2.860 | 2.863 | 2.849 | 2.878 |  |
|  | C15 | 2.859 | 2.813 | 2.821 | 2.812 | 2.843 |  |
|  | C25 | 2.903 | 2.867 | 2.877 | 2.847 | 2.901 |  |
|  | C35 | 2.044 | 2.117 | 2.100 | 2.108 | 2.095 |  |
|  | C45 | 3.657 | 3.669 | 3.643 | 3.652 | 3.656 |  |
|  | C55 | 2.856 | 2.830 | 2.858 | 2.843 | 2.855 |  |

The two panels in [Fig. 5](#_bookmark18) correspond to values in [Table 2](#_bookmark17)a and [2b](#_bookmark17) 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 2](#_bookmark17)a, 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.

present a clear conclusion that in terms of GPA performance based on the sampled data, we can isolate three distinct clusters, centered 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

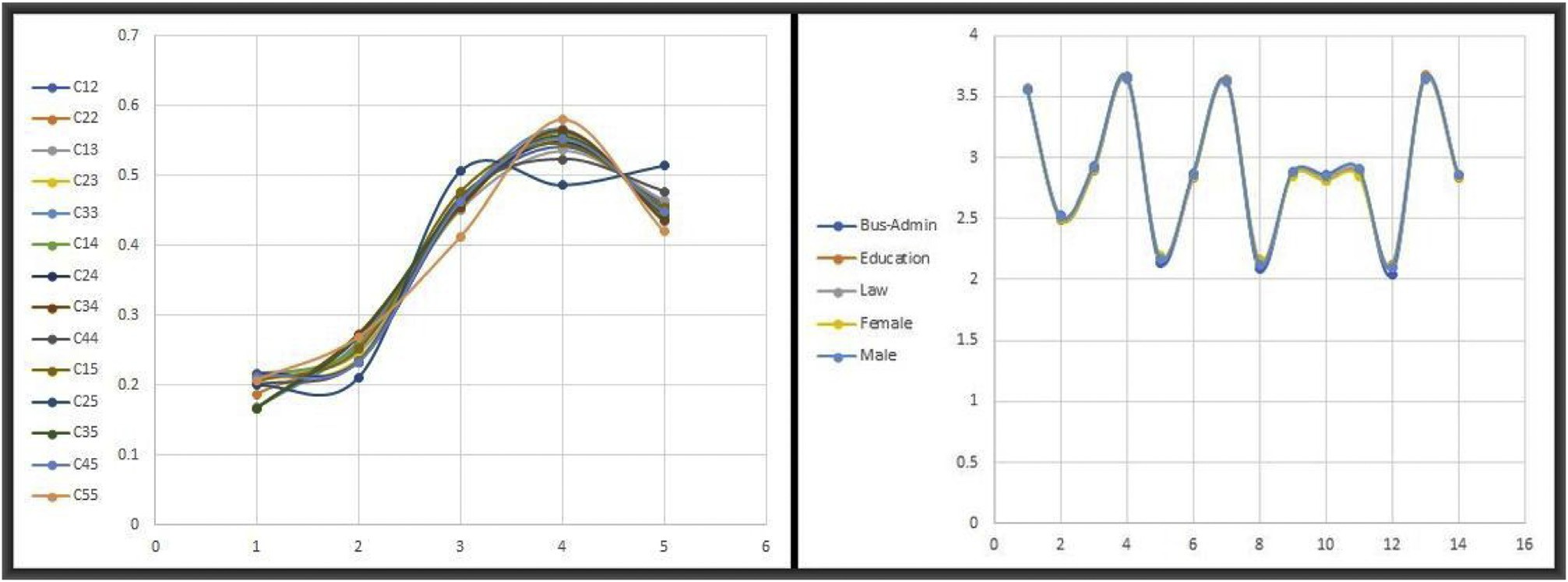


Fig. 5. Enrolment proportions on the LHS panel and the GPA performance on the RHS.

As noted earlier, the clusters in [Fig. 4](#_bookmark16) heavily overlap. Thus, to determine the optimal number of clusters in the sampled data, we refer back to the densities in [Fig. 3](#_bookmark14) and the enrolment proportions and GPA averages in [Fig. 5](#_bookmark18). Repeated sampling through [Algorithm 1](#_bookmark7) yields in the consistent GPA average performance densities in [Fig. 6](#_bookmark19).

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 [Fig. 5](#_bookmark18). The data for each of the 14 clusters is available for po- tential future examinations.

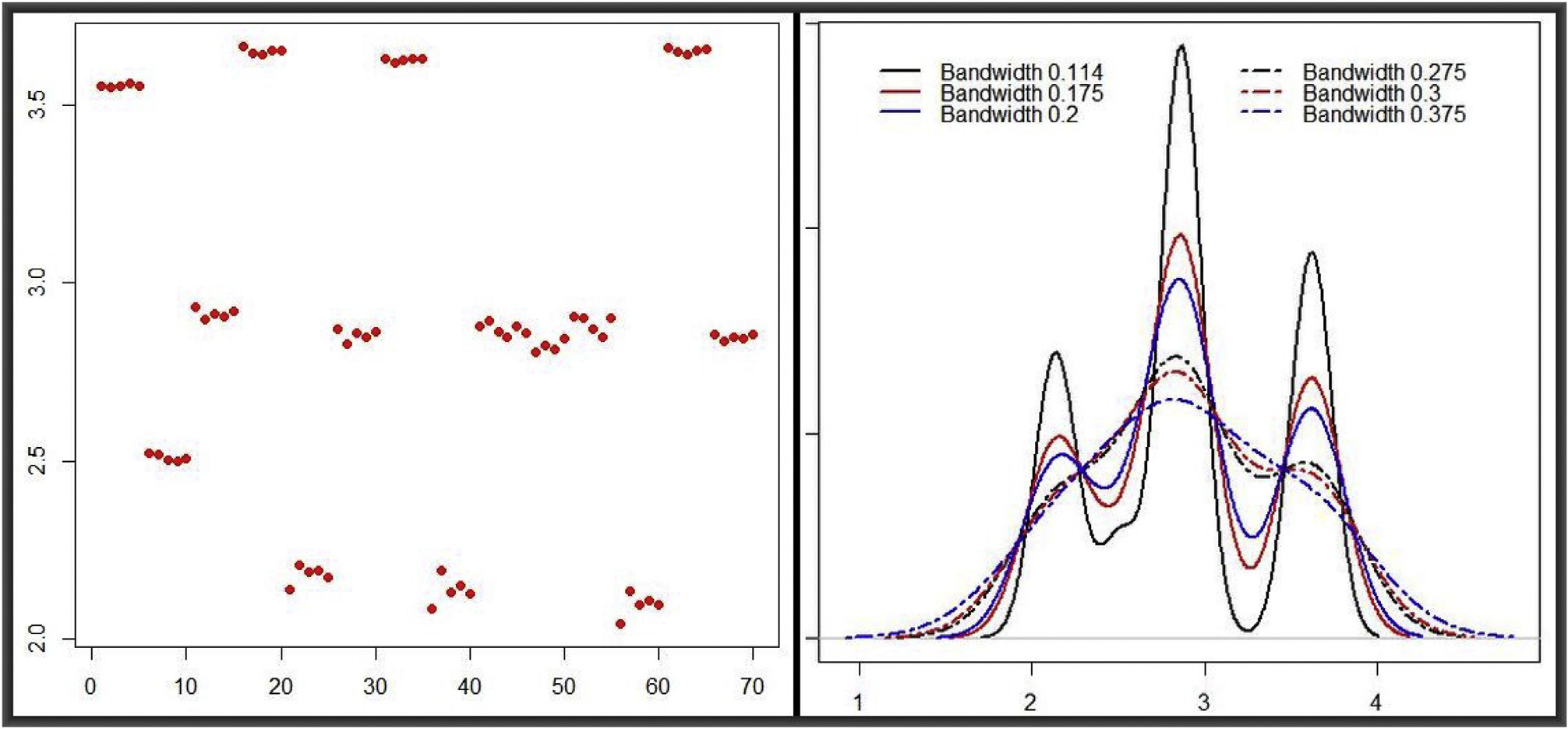


Fig. 6. An optimal 3–cluster structure for the average GPA over multiple runs at different bandwidths.

The patterns in [Fig. 6](#_bookmark19) 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 [Fig. 4](#_bookmark16). Both panels

1. 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](#_bookmark38),[19](#_bookmark39)] based on the me- chanics of the SMA algorithm [[8](#_bookmark28),[9](#_bookmark29)]. 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. Ob- jectives 1 through 3 were met in sub–sections [3.1 and 3.2](#_bookmark10). 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 interdisci- plinary research involving data scientists and educationists.

Even within this limited application, our findings show that there are great potentials in incorporating interdisciplinary approaches in uni- versity curricula, bringing together domain sciences on the one side and data science on the other. Further tests and evaluations of the SMA al- gorithm can conducted using a wide range of unsupervised and super- vised techniques, with any combination of the 19 data attributes. [Algorithm 1](#_bookmark7) is also capable of handling association rules–originally developed for analysing shopping transactions Agrawal et al. [[21](#_bookmark41)]. In this particular application, association rules can play a unifying role between unsupervised and supervised modelling in that they can capture under- lying rules of association among the students’ data attributes. We expect the technical and soft aspects of the paper to increasingly attract atten- tion to collaborative, interdisciplinary research activities in various sectors.

Finally, and as emphasised by Aikat et al. [[12](#_bookmark32)], 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 Emir- ates in forging interdisciplinarity for educational performance enhancement.

Declaration of competing interest

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.

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 dis- cussed these initiatives with us at different stages of development. We particularly acknowledge the role of the Data Intensive Research Initia- tive 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 Japa- nese Polar Environment Data Science Center (PEDSC) ([http://pedsc.rois.](http://pedsc.rois.ac.jp/en/) [ac.jp/en/](http://pedsc.rois.ac.jp/en/)), the United Nations World Data Forum (UNWDF) ([https://uns](https://unstats.un.org/unsd/undataforum/index.html) [tats.un.org/unsd/undataforum/index.html](https://unstats.un.org/unsd/undataforum/index.html)) and the Sussex Sustainabil- ity 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.

Nomenclature

a The first letter

w Some other letter

+x A special symbol

1. Some letter
2. The last letter

*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

Data availability

As noted in Section [3.2](#_bookmark15), 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.

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 auto- mated 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.

References

1. SDG. Sustainable Development Goals. 2015. [https://www.un.org/sustainabledeve](https://www.un.org/sustainabledevelopment/sustainable-development-goals/) [lopment/sustainable-development-goals/](https://www.un.org/sustainabledevelopment/sustainable-development-goals/).
2. [Miguis VL, Freitas A, Garcia PJV, Silva A. Early segmentation of students according](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref2) [to their academic performance: A predictive modelling approach. Decis Support](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref2) [Syst 2018;115:36](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref2)–[51](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref2).
3. [Brooks C, Erickson G, Greer J, Gutwin C. Modelling and quantifying the behaviours](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref3) [of students in lecture capture environments. Comput Educ 2014;75:282](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref3)–[92](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref3).
4. [Hua Leong F, Marshall L. Modeling engagement of programming students using](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref4) [unsupervised machine learning technique. GSTF J Comput 2018;6](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref4).
5. [SILPA. Standards For Institutional Licensure And Program Accreditation. UAE:](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref5) [Ministry of Education; 2019](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref5).
6. [Attwell G, Pumilia P. The New Pedagogy of Open Content: Bringing together](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref6) [production, knowledge, development, and learning. Data Sci J 2007;6:S211S219](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref6).
7. [Meusburger P. In: Meusburger P, Glückler J, el Meskioui M, editors. Knowledge and](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref7) [the economy. Dordrecht: Springer Netherlands; 2013. p. 15](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref7)–[42](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref7).
8. [Mwitondi K, Munyakazi I, Gatsheni B. Amenability of the united Nations](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref8) [sustainable development goals to Big data modelling. In: International workshop on](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref8) [data science-present and future of open data and open science, 12-15 nov 2018,](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref8) [joint support centre for data science research. Mishima, Shizuoka, Japan: Mishima](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref8) [Citizens Cultural Hall; 2018](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref8).
9. [Mwitondi K, Munyakazi I, Gatsheni B. An Interdisciplinary Data-Driven Framework](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref9) [for Development Science. Pretoria, RSA: DIRISA National Research Data Workshop;](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref9) [2018. CSIR ICC, 19-21 June 2018](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref9).
10. [CHEDS. Center For Higher Education Data And Statistics. UAE: Ministry of](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref10) [Education; 2018](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref10).
11. [Parsons MA, Godøy Øystein, LeDrew E, de Bruin TF, Danis B, Tomlinson S,](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref11) [Carlson D. A conceptual framework for managing very diverse data for complex,](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref11) [interdisciplinary science. J Inf Sci 2011;37:555](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref11)–[69](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref11).
12. [Aikat J, Carsey TM, Fecho K, Jeffay K, Krishnamurthy A, Mucha PJ, Rajasekar A,](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref12) [Ahalt SC. Scientific training in the era of Big data: A New Pedagogy for graduate](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref12) [education. Big Data 2017;5](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref12).
13. [Mwitondi KS, Said RA, Zargari SA. A robust domain partitioning intrusion detection](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref13) [method. Journal of Information Security and Applications 2019;48:102360](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref13).
14. [Kim K, Timm N. Univariate and Multivariate General Linear Models. New York:](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref14) [Chapman and Hall/CRC; 2006](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref14).
15. [Chapmann J. Machine Learning Algorithms. CreateSpace Independent Publishing](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref15) [Platform; 2017](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref15).
16. [Kogan J. Introduction to Clustering Large and High-Dimensional Data. Cambridge](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref16) [University Press; 2007](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref16).
17. [Hastings WK. Monte Carlo sampling methods using Markov Chains and Their](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref17) [Applications. Biometrika 1970;57:97](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref17)–[109](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref17).
18. [Frey BJ, Dueck D. Clustering by passing messages Between data Points 2007;315:](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref18) [972](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref18)–[6](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref18).
19. [Bodenhofer U, Kothmeier A, Hochreiter S APCluster. An R package for Affinity](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref19) [Propagation Clustering. Bioinformatics 2011;27:2463](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref19)–[4](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref19).
20. [Burgess A, Senior C, Moores E. A 10-year case study on the changing determinants](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref20) [of university student satisfaction in the UK. PloS One 2018;13:1](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref20)–[15](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref20).
21. [Agrawal R, Imieli](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref21)n´[ski T, Swami A. Mining Association rules Between sets of items in](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref21)

[large databases. SIGMOD Rec 1993;22:207](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref21)–[16](http://refhub.elsevier.com/S2590-0056(21)00012-6/sref21).