Recurrence Quantification Analysis explained and applied to Educational Science

B.M.L. Hogervorst

Applied Mathematics, Delft University of Technology
Mekelweg 4, 2628 CD Delft

Thesis Committee:

Prof. Dr. A.J. Cabo,	Supervisor
Drs. E. Papageourgiou,	Supervisor
Drs. I.A.M. Goddijn	

July 18, 2024

Preface

This thesis was written as a final project to obtain a Bachelor's degree in Applied Mathematics from the TU Delft under the supervision of Dr. A.J. Cabo and Drs. E. Papageorgiou Statistics section of Applied Mathematics at the TU Delft

Firstly, I would like to greatly thank Elli Papageorgiou for her guidance throughout this project. Often times I was lost in the field of educational psychology and Elli always guided me with amazing explanations and dedicated advice. Her enthusiasm was contagious and made me interested in pursuing research further.

Secondly, I want to thank Annoesjka Cabo for her support and valuable feedback. Her expertise and insights contributed significantly to the improvement of this work.

Lastly, I would like to thank Ingeborg Goddijn for taking a seat in my thesis committee.

Laymen Abstract

This thesis explores the relationship between how students stay engaged with their online homework exercises and their performance. Student Engagement (SE) is an umbrella term that covers students' involvement, focus, participation and persistence on a task. This research was done to gain information about this relationship and can be used to inform students and teachers about how they can improve their behavior to achieve better results. The method used to answer the research question is called Recurrence Quantification Analysis (RQA). RQA analyzes if the engagement shown by a student at different times is similar. SE was looked at in three different variables: how much time students spent studying, how many exercises they solved and whether they completed their work on time. In this study the behavior of 144 mechanical and civil engineering students following a Linear Algebra course at the TU Delft was analyzed. The results showed that high-performing students show more overall engagement than middle- and low-performing students. Furthermore, they do this using study sessions that have varying durations and also vary in the total exercises attempted. Therefore students are advised to engage with study materials as much as possible and consider doing this in different study sessions that differ in their length.

Abstract

Student Engagement (SE) is a critical factor when researching student performance. This thesis intends to research the correlation between SE and performance through the lens of Complex Dynamical System (CDS) theory. Recognizing that SE is influenced by multiple interdependent variables that cause non-linear and not fully predictable behavior, SE is analyzed through Recurrence Quantification Analysis (RQA). RQA uses the distance between time series data points to visualize and quantify the dynamic characteristics of a CDS, such as repetition, periodicity and predictability. In this thesis, Time Spent, Attempts Made and On-Time Rate are used as indicators of SE to examine if correlation exists between student performance and recurrence regarding SE. To analyze if there does exist a relationship between these aspects, a dataset of 144 civil and mechanical engineering students following a Linear Algebra course at TU Delft was used. Recurrence in SE was quantified using the following RQA variables: Recurrence Rate (RR), Determinism (DET), Average Diagonal Line Length ($Davg$), Trapping Time (TT),and Shannon's Entropy of diagonal line lengths ($ENTR$). However, eventually $ENTR$ was not included as it did not provide interpretable information regarding SE. The findings suggest that high-performing students show more engagement overall and less recurrence in some SE indicators, namely Time Spent per study session and Attempts Made per study session. These insights suggest that students should engage with online study materials as much as possible in study sessions that vary in length and exercises attempted to perform optimally. This thesis concludes with recommendations for further research analyzing the effects between recurrence in SE and student performance.

Contents

Chapter 1

Introduction

Student Engagement (SE) in the learning process is an influential factor when researching the performance of a student [13, 23]. We define engagement as “the intensity of productive involvement with an activity”, based on the paper of Ben-Eliyahu et al. (2018) [8]. This definition includes a student’s involvement, focus, participation, and persistence on a task. SE research has been vastly focused on online learning settings, especially in higher education, where the use of online learning is widespread. Therefore, examining SE in this specific context provides useful information to improve teaching and studying behavior.

Furthewebrmore, SE can be effectively understood using the theory of Complex Dynamical Systems (CDS). A CDS is defined as a system influenced by multiple interdependent variables over time that cause non-linear and not fully predictable behavior [20]. SE comprises various dimensions (affective, behavioral, cognitive) [8]. The relationship between SE dimensions is complex and heterogeneous. The relationship between these dimensions is thus not inherently linear and not fully predictable [8]. SE is a dynamic process that evolves over time, influenced by factors related to the student, the educational institution, and the broader environment [31]. Therefore, viewing SE as a complex dynamical system provides a comprehensive framework for analyzing how SE evolves over time.

However, research in educational psychology that treats phenomena of interest as a CDS, rarely examines these phenomena using CDS methodology or tends to translate CDS to linear cause-effect models that do not adequately describe the theory [18]. Hence, in this research an approach based on CDS-oriented methods is used, which does not require linearity assumptions and takes the order of events into account.

1.1 Clustering

Variable-centered approaches with regard to SE give valuable information as they capture how a population behaves on average [31]. These methods are used often in educational research to prove or falsify educational theories that describe relationships between different variables [28]. However, an “average” learning pattern often fails to adequately describe many students, as it overlooks the unobserved heterogeneity among them [17]. Therefore, in this thesis students will be clustered based on their performance, under the assumption that the full population might have different sub-populations that differ in their human behavior and experiences [28, 31]. More specifically, the clustering based on performance was chosen, since multiple studies have shown that SE is positively related to performance [29, 12, 23]. The main goal of this clustering is to analyze if groups of students, who perform differently, also show different recurrence patterns of SE indicators.

1.2 Recurrence Quantification Analysis

The main method applied is Recurrence Quantification Analysis (RQA). RQA is used to visualize and quantify characteristics of categorical or continuous time series data [24]. RQA uses the distance between time series data points to visualize the dynamic characteristics of a complex system, such as repetition, periodicity and predictability in a Recurrence Plot (RP). Furthermore, RQA quantifies these characteristics through multiple variables. Due to the fact that RQA is suited for analyzing non-linear systems and takes temporality into account, the method has become well-established in recent years and is described by Richardson et al. (2014) as "one of the most robust and generally applicable methods for assessing the dynamics of biological and human behavior" [27].

1.3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique, used to visualize higher dimensional data [19]. PCA has the goal to extract the main variance from a higher-dimensional dataset, to represent it as a 2-dimensional plot. This is done by calculating the Principal Components, which are linear combinations of the variables explaining the greatest variance in the data. This method is applied to the values obtained for the RQA variables for every cluster to visualize differences between them. Furthermore, the Principal Components give information about the structure of the variables. A similar value for variables in a Principal Component shows that these variables show correlation over the whole dataset. Therefore, PCA also provides information about the overall relationships between all the RQA variables.

1.4 Permutational analysis of variance

Permutational analysis of variance (PERMANOVA) is a multivariate method used to analyze the differences between the clusters [6]. PERMANOVA is used in this research since the distribution of the obtained variables does not follow a normal distribution and this assumption is not necessary to implement PERMANOVA. PERMANOVA compares the distance within clusters with the distance between clusters to determine if there is a statistically significant difference in the composition of the clusters. The test is applied to the data to obtain information about the differences between the clusters regarding the attained values for the RQA variables.

1.5 Thesis Outline

This thesis has two main goals: firstly, to explain the mathematical foundations behind the RP and the RQA, and secondly, to apply this methodology to analyze the relationship between SE and student performance. This relationship will be analyzed by first defining the three SE indicators used based on the theoretical definition of SE given by Ben-Eliyahu et al. (2018) [8] and prior literature on online SE, such as Saqr & López-Pernas (2021) [31]. Next, students will be clustered based on their study performance to form more homogeneous sub-populations. Then, time series data on these SE indicators will be analyzed using RQA. As stated previously, RQA is suited for analyzing the repetition and periodicity of systems, which is why this method will be used to analyze students' recurrence of SE indicators. The quantified recurrence of SE indicators of these clusters will then be compared by using PCA and PERMANOVA. Eventually, the main question answered is: "How is student performance related with recurrence in SE on online homework exercises?"

1.6 Hypothesis

To formulate a hypothesis it is important to consider that the field of educational science is heavily influenced by the definitions used. Not every research paper defines SE in the same way. Furthermore, the context in which SE is studied (type of students, study task, etc.) often varies from study to study. Therefore, the results from the papers introduced can be used to get an indication of expected results, but cannot directly be associated with this study.

Firstly, Franciškovič et al. (2024) [4] did not use RQA, but did find that: "Greater consistency in weekly session durations and the number of weekly sessions correlated with improved exam outcomes". Poquet et al. (2023) [25] did use RQA and found similar results. Through the measure of Entropy of behavioral states, which is explained in Chapter 2.1.5, they found that students who exhibited steep changes in their study behavior, performed worse on exams than other students.

Based on these two studies, our hypothesis is to find that high-performing students will show more recurrence in SE indicators. Practically, higher recurrence is shown in a higher value in the following RQA variables RR, DET, D_{avg}, TT and LAM which are introduced in Chapter 2.1.5.

Chapter 2

Methods

2.1 Recurrence Quantification Analysis

It is important to distinguish between two types of RQA, namely continuous and categorical RQA. In the categorical case, used in Li et al. (2022) [32] and Poquet et al. (2023) [25], the variables can only take a finite amount of values [1]. For example, in the case of researching study behavior, a time series could report the actions a student undertakes in studying, like "reading", "practicing", "planning", etc. In this thesis, continuous data will be studied, which is data that can take an infinite amount of values. An example of continuous data would be the amount of time spent studying in a week. In this thesis RQA was implemented using the "crqa" function from the R package "crqa" [2].

2.1.1 Recurrence Plot

A time series is a set of observations, $\mathbf{x}_t$, all recorded at a specific time t [10]. During this research, discrete time series will be used, which means that the times at which the observations are observed are a discrete set. An example of such a discrete set would be consecutive weeks. It is important to note that in this research only 1-dimensional time series will be used (every observation contains one data entry). However, all the methods explained can also be applied to higher dimensional data and for completeness, the definitions will be given for the higher dimensional case.

First, the formal definition of a Recurrence Plot (RP) as used in this research will be given. Suppose we have an m-dimensional time series of length N, where the observed data points are denoted by $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. An RP plots the time series against itself. So both the x- and y-axis are of length N. An RP then gives information about when time series entries are "similar". In the categorical case, time series entries are seen as "similar" if they describe the same action. In the continuous case, the two time series entries are seen as "similar" if their distance is smaller than a chosen margin, ϵ. So the formal definition is:

Definition 2.1.1. Let $\epsilon > 0$ be the defined margin. Then for an m-dimensional time series of length N with observed data points $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, the value given at x-coordinate i and y-coordinate j is defined as:

$$R_{i,j} = \Theta(\epsilon - ||\mathbf{x}_i - \mathbf{x}_j||)$$

Where $||\cdot||$ is a distance function of choice, discussed in Chapter 2.1.2, and Θ is the Heaviside step-function, which returns the value 1 whenever the input is positive and zero otherwise. In other words, the RP only gets assigned value 1 at coordinate (i, j) if the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is smaller than the margin, ϵ, and else gets value 0. Whenever $R_{i,j} = 1$, the point (i, j)

is called a Recurrence Point. ϵ is the maximum distance two data points can have, to still be considered “similar”. When translating this mathematical definition into the actual plot, every 1 is represented with a dot and every 0 gets no representation. A very simplistic example of an RP, created from a time series of length 4, is shown in Figure 2.1.

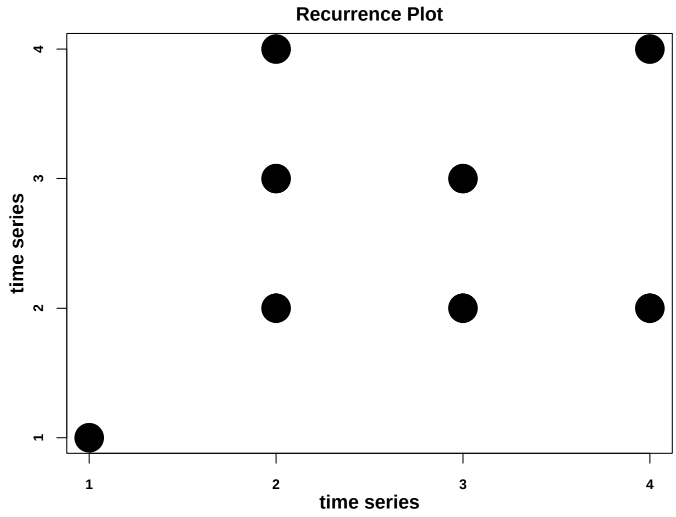


Figure 2.1: Representation of simplistic Recurrence Plot

Line Of Identity

It is important to note that in Figure 2.1 there are recurrence points all through the main diagonal. This is not an exception, but indeed this is always the case for a recurrence plot, which can be shown easily.

Lemma 1. *An RP always has recurrence points along the main diagonal.*

Proof. Consider a point on the main diagonal, so $i = j$, then the following is obtained:

$$||\mathbf{x}_i - \mathbf{x}_j|| = ||\mathbf{x}_i - \mathbf{x}_i|| = 0 < \epsilon$$

Showing that along the main diagonal the distance is always smaller than the margin and therefore the points on the main diagonal will always be recurrent. □

This diagonal line is called the Line of Identity (LOI). In Figure 2.2 the LOI is highlighted in red and the non-trivial recurrence points (showing that the second data point is similar to both the third and fourth data point) in blue.

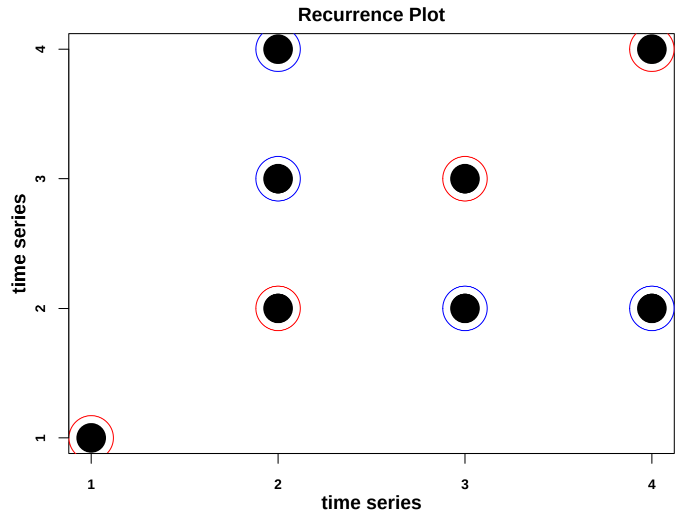


Figure 2.2: Highlighted representation of simplistic Recurrence Plot

Symmetry

Another element that stands out from Figure 2.1 is that the plot is symmetric along the main diagonal, this is indeed also always the case. The only way an RP is not symmetrical along the main diagonal is if a margin dependent on the x-coordinate is chosen, but this is not discussed in this thesis. In the case of a constant margin it is quite straightforward that the RP becomes symmetrical.

Lemma 2. *An RP with a constant margin, ϵ, is always symmetrical about its main diagonal.*

Proof. Suppose the time series has length N. To obtain symmetry in the RP, it is needed that, for all pairs in the set: $\{(i,j) \mid 1 \leq i \leq N, 1 \leq j \leq N\}$, $R_{i,j} = R_{j,i}$. This is indeed the case by definition of a distance function, discussed in Chapter 2.1.2.

$$R_{i,j} = \Theta(\epsilon - ||\mathbf{x}_i - \mathbf{x}_j||) = \Theta(\epsilon - ||\mathbf{x}_j - \mathbf{x}_i||) = R_{j,i}$$

□

2.1.2 Distance functions

So far the distance between two time series entries has been used, but not defined exactly. Again the definitions for distance will be given for the higher dimensional case, even though in this thesis only 1-dimensional time series will be used. First, the formal definition for a distance function will be given.

A distance function, denoted by $||\cdot||$, is a measurement of how far apart two data points are. If the data points are $\mathbf{x}$ and $\mathbf{y}$ this is denoted by $||\mathbf{x} - \mathbf{y}||$.

Definition 2.1.2. Let $D \subseteq \mathbb{R}^m$ be the space where the data points are defined. Then A well-defined distance function, $||\cdot|| : D \to [0, \infty)$, satisfies four properties:

$$\forall \mathbf{x} \in D : ||\mathbf{x} - \mathbf{x}|| = 0 \tag{i}$$

$$\forall \mathbf{x} \neq \mathbf{y} \in D : ||\mathbf{x} - \mathbf{y}|| > 0 \tag{ii}$$

$$\forall \mathbf{x}, \mathbf{y} \in D : ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{y} - \mathbf{x}|| \tag{iii}$$

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in D : ||\mathbf{x} - \mathbf{z}|| \leq ||\mathbf{x} - \mathbf{y}|| + ||\mathbf{y} - \mathbf{z}|| \tag{iv}$$

Every distance function that satisfies these conditions could in theory be used for RQA, but in practice a few functions are used, which will now be discussed.

Manhattan

The Manhatten distance function sums over the absolute differences for all data entries, so in m-dimensional space:

$$||\mathbf{x} - \mathbf{y}||= \sum_{i=1}^{m} |x_i - y_i|$$

Euclidian

The Euclidian distance calculates the smallest distance of “travelling” directly from one data point to another, which is defined as:

$$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

In the 1-dimensional case, it is obvious that Manhattan distance and Euclidian distance obtain the same value.

Minkowski

A more generalized version of the Euclidian distance is Minkowski's distance, where the value of $p > 0$ can be chosen:

$$||\mathbf{x} - \mathbf{y}||_p = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Minimum and Maximum

The definitions for the Minimum and Maximum functions are straightforward, but will be given for completeness.

$$||\mathbf{x} - \mathbf{y}||_{max} = max\{|x_i - y_i|\} \qquad ||\mathbf{x} - \mathbf{y}||_{min} = min\{|x_i - y_i|\}$$

2.1.3 Choice of distance function

The choice for a distance function in most cases is not that obvious. If it is the case that all of the data entries must be close to each other, the Maximum distance function is suitable. Similarly, if only one entry needs to be similar the Minimum function can be chosen. In all other cases the other distance functions give a comparable indication of how similar two data points are. In this thesis the computations were done with the Euclidian distance measure, which in the 1-dimensional case is equal to the Manhattan, Minimum and Maximum distance.

2.1.4 Remarks about distance in higher-dimensional space

In this thesis multiple RQA analyses will be done using 1-dimensional time series of different variables. There is also the possibility of combining these variables into a higher-dimensional time series. However, this gives rise to two problems. Firstly, if the distance has to be calculated in a higher-dimensional space, choices need to be made with regard to what the influence is of each variable. It does not make sense to use the exact same function for different variables, that have different distributions, and let each variable have the same impact. Therefore, the influence of each variable needs to be determined and it is not quite clear how this needs to be done. Secondly, if RQA results are produced it is not quite clear which variable has caused similarity between data points and therefore the results are hard to interpret. Therefore, the choice was made to perform multiple 1-dimensional analyses.

2.1.5 RQA variables

To analyze the occurrence of repetition, periodicity or predictability of any complex dynamical system multiple variables will be introduced which quantify the characteristics of an RP. The variables will be explained in great detail regarding the origin of every formula to give the reader intuition regarding the direct meaning of every variable. We follow the research done by Marwan and Webber (2015) regarding this subject [24].

Recurrence Rate

Firstly, it is of great importance what part of all data points is recurrent. This is described by the variable the Recurrence Rate (RR). As seen in Figure 2.1.1 there will always be recurrence points along the main diagonal. Therefore, only the points where $i \neq j$ are considered. The Recurrence Rate is calculated by dividing the total number of recurrence points by the total number of points. The total amount of recurrence points is found by summing over $R_{i,j}$ for all points where $i \neq j$, or in formula: $\sum_{i \neq j=1}^{N} R_{i,j}$, where N is the total amount of data points in the time-series. Every data point is compared with every other data point so a total of N^2 comparisons is made. N is subtracted from the total amount of data points since the LOI is ignored in this formula.

Definition 2.1.3. The Recurrence Rate (RR) of an RP is given by:

$$RR = \frac{1}{N^2 - N} \left(\sum_{i \neq j=1}^{N} R_{i,j} \right) \cdot 100\% \tag{2.1}$$

Diagonal Lines

Diagonal lines occurring in an RP give information about the deterministic structure of a system. The length of a diagonal line in an RP is not the Euclidean distance, but rather the amount of recurrence points contained in this line. In mathematical terms, a diagonal line of at least length l starting at point (i, j) has the following attributes: $R_{i,j} = R_{i+1,j+1} = \ldots = R_{i+l-1,j+l-1} = 1$. A diagonal line gives valuable information since it shows that for different times in the time series, the system has obeyed the same "rules". To put it more concretely, suppose that $R_{i,j} = 1$ and $R_{i+1,j+1} = 1$ (diagonal line of length 2). Then the data at time i is similar to the data at time j and the data at time $i + 1$ is similar to the data at time $j + 1$. This means that at different times the system has undergone a comparable transformation, or in other words has followed the same "rules". Therefore Diagonal Lines in a Recurrence Plot are an indication of a deterministic system and give information about if there are underlying mechanisms that could be quantified.

By taking the product over l recurrence points that follow each other diagonally a binary operator can be defined showing if there is a diagonal line of length l starting at point (i, j), in formula:

$$\prod_{k=0}^{l-1} R_{i+k,j+k} = \begin{cases} 1 & \text{if} \quad \text{diagonal line of length at least } l \\ 0 & \text{else} \end{cases} \tag{2.2}$$

If the extra conditions that $R_{i-1,j-1} \neq 1$ and $R_{i+l,j+l} \neq 1$ are applied, the number of diagonal lines of exactly length l is obtained. Summing over all, i, j, the following formula for the number of diagonal lines in the RP of exactly length l is obtained.

Definition 2.1.4. The amount of diagonal lines in an RP of length l is denoted by $H_d(l)$ and given by:

$$H_D(l) = \sum_{i=j=1}^{N} (1 - R_{i-1,j-1})(1 - R_{i+1,j+1}) \prod_{k=0}^{l-1} R_{i+k,j+k} \tag{2.3}$$

Using this variable an important quantification of diagonal lines within the RP is introduced, namely the average diagonal line length, which is denoted by D_{avg} respectively. For D_{avg} a minimum length for the diagonal lines needs to be defined and is denoted by d_{min}. Typically, $d_{min} = 2$, which is the value that also will be used in this thesis [24].

Definition 2.1.5. The average length of the diagonal lines in an RP is denoted by D_{avg} and is given by:

$$D_{avg} = \frac{\sum_{l=d_{min}}^{N} lH_D(l)}{\sum_{l=d_{min}}^{N} H_D(l)} \tag{2.4}$$

Percent Determinism

Next, the proportion of recurrence points that occur in a diagonal structure is quantified, which is denoted by Percent Determinism (DET). This gives information about how deterministic (rule-obeying) the system is [24]. Again a minimal length for the diagonal lines needs to be chosen, d_{min}, which is kept at 2. As discussed in Definition 2.1.3 the total amount of recurrence points is shown in the variable RR. Furthermore, the total amount of points contained in diagonal structures of length l is also known, namely $H_D(l)$, as explained in Definition 2.1.4.

Definition 2.1.6. The part of the recurrence points that occur in diagonal structures in an RP is denoted by DET and given by:

$$DET = \frac{\sum_{l=d_{min}}^{N} lH_D(l)}{\sum_{i=j=1}^{N} R_{i,j}} \cdot 100\% \tag{2.5}$$

Note that in this definition the LOI is included. Furthermore, if $d_{min} = 1$, all recurrence points are included in diagonal structures and thus $DET = 100\%$.

Vertical/Horizontal Lines

Next to diagonal lines in the Recurrence Plot, vertical lines also give away important information about the characteristics of a system. First note that due to the fact that a Recurrence Plot is symmetrical as explained in Lemma 2, the analysis of vertical and horizontal lines will give the same result. Next, the interpretation of vertical structures will be discussed. A vertical line of length l, starting in coordinate (i, j), occurs whenever $R_{i,j} = R_{i,j+1} = \ldots = R_{i,j+l-1} = 1$. So the data at time i is similar to all data points at times $j, j+1, \ldots, j+l-1$. This means that most likely those states do not differ much compared to each other and the time series is quite stationary from time j to time $j+l-1$.

Next, it will be discussed how vertical/horizontal lines in a Recurrence Plot can be quantified.

Definition 2.1.7. The amount of vertical/horizontal lines of length l in an RP is denoted by $H_V(l)$ and given by:

$$H_V(l) = \sum_{i,j=1}^{N} (1 - R_{i,j-1})(1 - R_{i,j+l}) \prod_{k=0}^{l-1} R_{i,j+k} \tag{2.6}$$

The second variable introduced regarding the vertical structures is the so-called Trapping Time (TT), which describes the average length of the vertical structures. The name Trapping Time is quite fitting since it describes how long a state is "trapped" (does not change a lot).

Definition 2.1.8. The Trapping Time (TT) shows the average length of the vertical/horizontal structures in an RP and is given by:

$$TT = \frac{\sum_{l=v_{\min}}^{N} lH_V(l)}{\sum_{l=v_{\min}}^{N} H_V(l)} \tag{2.7}$$

Definition 2.1.9. Laminarity (LAM) shows the percentage of recurrence points occurring in vertical structures and is given by:

$$LAM = \frac{\sum_{l=v_{\min}}^{N} lH_V(l)}{\sum_{i,j=1}^{N} R_{i,j}} \cdot 100\% \tag{2.8}$$

Shannon's entropy of the diagonal lines

Shannon's entropy of the diagonal lines ($ENTR$) [33] measures how uniformly the lengths of the diagonal lines are distributed. First the formal definition will be given and then this formula will be derived.

Definition 2.1.10. Shannon's Entropy of the diagonal lines ($ENTR$) shows how uniformly the lengths of the diagonal lines are distributed and is given by:

$$ENTR = \sum_{l=d_{min}}^{d_{max}} -p(l) \ln p(l) \quad \textit{where, } p(l) = \frac{H_D(l)}{\sum_{l=d_{min}}^{N} H_D(l)} \tag{2.9}$$

$p(l)$ denotes the chance of picking a diagonal line of length l from all of the diagonal lines, which is defined by the number of lines of length l divided by the total number of diagonal lines.

Now a measure of "surprise" is introduced. Intuitively "surprise" can be seen as a number quantifying how surprised one would be to find a certain income. Logically, it exhibits an inverse relationship with probability. If a certain outcome has a high probability one would not be "surprised" to find that certain outcome. The direct inverse of probability cannot be taken as "surprise", since then a probability of 1, will give a surprise of 1, while it would be certain to find that outcome and thus not surprising. Therefore the logarithm of the inverse of probability is taken. The "surprise" associated with finding a diagonal line of length l out of all the diagonal lines is denoted by $s(l)$.

$$s(l) = \ln\left(\frac{1}{p(l)}\right) \tag{2.10}$$

Following this, the final definition of Shannon's Entropy is given. Multiplying $s(l)$ for all $l \in \{d_{min}, \ldots, d_{max}\}$ by $p(l)$ gives the expectation for the total "surprise", which is the definition of Shannon's entropy. Thus it is a measure of how much "surprise" is expected of the distribution of the lengths of the diagonal lines or how (un)predictable this distribution is. The mathematical definition for Shannon's Entropy will be given and rewritten to the form commonly seen in the literature.

$$\begin{aligned} ENTR &= \sum_{l=d_{min}}^{d_{max}} p(l)s(l) && (2.11)\\ &= \sum_{l=d_{min}}^{d_{max}} p(l)\ln\left(\frac{1}{p(l))}\right) && (2.12)\\ &= \sum_{l=d_{min}}^{d_{max}} -p(l)\ln p(l) && (2.13) \end{aligned}$$

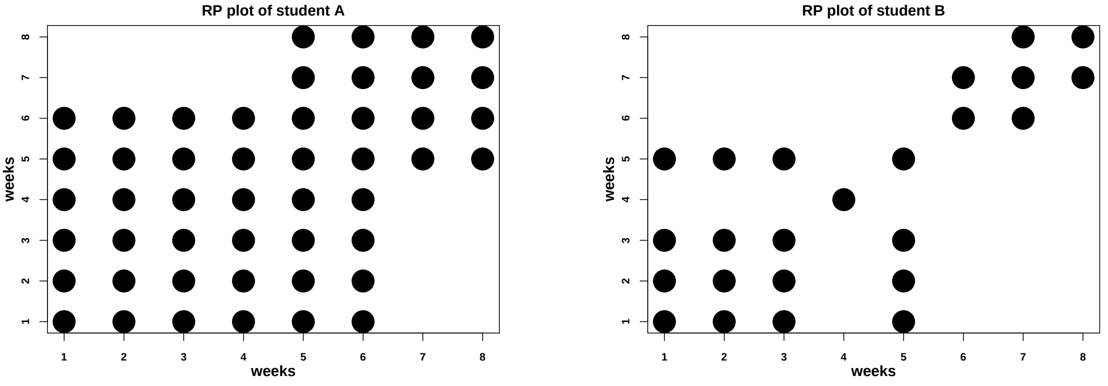
The value of Shannon's Entropy is maximized if and only if the distribution of the lengths of the diagonal lines is uniform. This Theorem is proven in Appendix A.2, using Jensen's inequality, which is stated and proven in Appendix A.1.

2.1.6 Application to example students

To give more intuition as to what exactly happens in the RP and how this translates into the RQA variables, concrete examples of four "typical" students will be given. Consider four students A, B, C and D for whom the number of attempted exercises was tracked for a period of 8 weeks. Student A could be considered as "steady buildup", student B as "late start", student C as "random behavior" and student D did the work in the beginning and the end of the term. Their total Attempts Made per week can be captured in time series, for example:

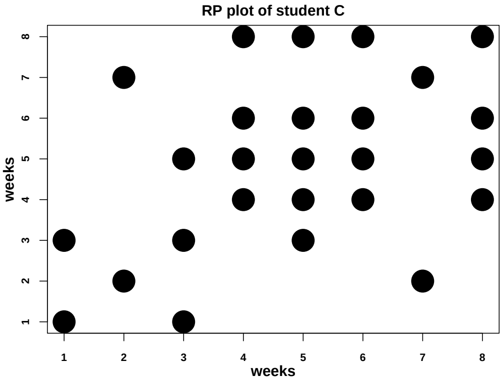
$$\begin{aligned} A &= [20, 20, 25, 25, 30, 30, 40, 40]\\ B &= [0, 0, 0, 20, 0, 60, 70, 80]\\ C &= [0, 50, 10, 30, 20, 30, 60, 30]\\ D &= [40, 40, 20, 0, 0, 45, 40, 45] \end{aligned}$$

Note that the total Attempts Made across the 8 weeks for all students are exactly the same (230). If a margin of 10 is chosen, the following RP's are generated, with the corresponding RQA variables given in Table 2.1.

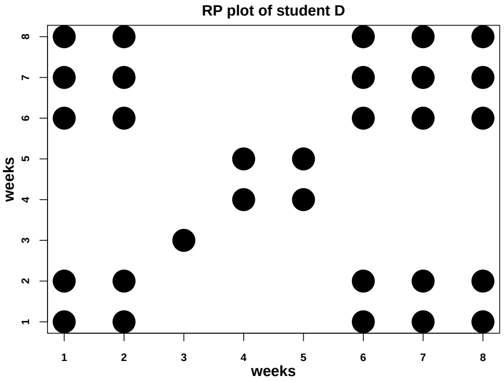


(a) RP of Attempts Made per week

(b) RP of Attempts Made per week



(a) RP of Attempts Made per week



(b) RP of Attempts Made per week

Student	RR	DET	D_{avg}	TT	$ENTR$	LAM
A	75,00	91,67	4,89	6,00	1,58	100,00
B	37,50	66,67	3,20	2,71	0,50	79,17
C	40,63	61,54	3,20	3,25	0,50	50,00
D	46,88	66,67	2,86	2,42	0,41	96,67

Table 2.1: values of RQA variables of Time Spent per study session

The values for these variables are quite alike for students B, C and D. This shows that even though the study behavior of these students was quite different, the recurrence they have shown in this indicator is quite similar. Only student A stands out with high values for the RQA variables, which is expected, since this student was categorized as “steady”. It is also noteworthy that the variable $ENTR$ for student A is significantly higher. This is caused by the fact that the RP of student A contains a lot of diagonal lines of different lengths. The other RP’s do not have longer diagonal lines and therefore the lengths of the diagonal lines are mostly 2. So the distribution of the lengths of the diagonal line is not uniform at all, causing a lower value for $ENTR$, as explained in Chapter 3.8.

2.2 Principal Component Analysis

After RQA is applied too many variables are available per student to directly visualize the obtained information. Therefore, Principal Component Analysis (PCA), a dimensionality reduction technique, is applied to the data. PCA transforms the original variables into two Principal Components, which capture the most significant patterns in the original variables. In other words, PCA is a useful method to visualize higher-dimensional data, while keeping the essential characteristics of the data.

PCA works by first standardizing the different variables, so that the results are not biased by variables that have values on a larger scale. Denote the number of data points by n and let $\mathbf{x}_i$ be a vector containing all the variables for data point i. Suppose there are d variables contained in every data point. Then standardizing is done as follows:

$$\mathbf{x}_{i_{standardized}} = \frac{\mathbf{x}_i - \mu(\mathbf{x})}{\sigma(\mathbf{x})} \tag{2.14}$$

Where $\mu(\mathbf{x})$ is the vector containing all the means of the d variables and $\sigma(\mathbf{x})$ contains the standard deviation for the d variables. Let $\mathbf{X}$ be the matrix which has $\mathbf{x}_{i_{standardized}}$ as its columns. Then the covariance matrix $\mathbf{C}$ is calculated to quantify the relationships between variables. This is done since if two variables are highly correlated with each other, they contain redundant information for the 2-dimensional plot.

Definition 2.2.1. The covariance matrix $\mathbf{C}$ shows how correlated all variables are with each other and is defined by:

$$\mathbf{C}_{i,j} = \frac{\mathrm{Cov}(i,j)}{n-1} = \frac{\sum_{k=1}^{n} \mathbf{X}_{i,k}\mathbf{X}_{k,j}}{n-1} \tag{2.15}$$

Where it was used that the mean of every variable is 0, since the data was normalized. The division by $n-1$ is because this is equal to the Degrees of Freedom of $\mathrm{Cov}(i,j)$. The degrees of Freedom are defined as the total number of independent pieces of information that influence the covariance. $n-1$ is obtained since the only independent values are the multiplications of one data point with all the others, i.e. $n-1$. This definition can be summarized in the form of matrix multiplication.

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \tag{2.16}$$

Next, the eigenvalues and eigenvectors of $\mathbf{C}$ are calculated. By the definition of eigenvectors, the direction of an eigenvector is invariant to the multiplication with matrix $\mathbf{C}$. The eigenvalues represent the magnitude of the variance in the direction of the corresponding eigenvector. Since $\mathrm{Cov}(i,j) = \mathrm{Cov}(j,i)$ and the covariance is always a real number, the covariance matrix is real and symmetric and therefore the eigenvalues are real [15]. Additionally, the number of eigenvalues is equal to the dimension of the covariance matrix (including multiplicities) [16]. So if the data points have d variables, there will be d eigenvalues that are greater or equal to zero and. Also, there will be an orthonormal basis consisting of d eigenvectors for $\mathbb{R}^d$, which are called Principal Components.

The main objective of the PCA plot is to explain the main variance in the data and thus the two Principal Components with the greatest eigenvalues are chosen as axes for this plot, as they explain the biggest part of the variance. These two Principal Components represent the linear combination of the d variables that explain the greatest variance in the data. Two things can be deduced from the Principal Components; Firstly, variables that have the largest values in the Principal Components explain the greatest part of the variance within that Principal Component. Secondly, variables that have the same sign show correlation in the data.

To plot the data with the newly found axes, the standardized data is multiplied with a matrix, whose columns correspond with the eigenvectors that had the greatest eigenvalues. If these eigenvectors are $\mathbf{v}_1, \mathbf{v}_2$, then $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2]$. The new data points are summarized in matrix $\mathbf{A}$, containing the coordinates of the data points as rows:

$$\mathbf{A} = \mathbf{X}^T\mathbf{V} \tag{2.17}$$

It is important to note that PCA assumes that the variables have a linear relationship with each other. This might not necessarily be the case and thus PCA is only used to indicate how the clusters differ from each other. Plots were created for every SE indicator to visualize the values of the RQA variables for every cluster. These plots can be seen in Figures 4.2, 4.5, 4.7, 4.9 and 4.10, where the differences of the clusters will be discussed per SE indicator. The ellipses shown in these figures represent 95% confidence intervals for the respective clusters. If a new data point of these clusters is added to the plot it has 95% chance to be within the confidence interval.

2.3 Permutational Multivariate Analysis of Variance

Permutational Multivariate Analysis of Variance (PERMANOVA) is a multivariate method used to compare means of different groups [6]. The data that will be obtained does not follow a normal distribution and therefore PERMANOVA was chosen to analyze if there exist differences between the clusters. Unlike many other statistical method, such as traditional Multivariate Analysis of Variance (MANOVA), PERMANOVA does not require any normality assumptions. PERMANOVA primarily functions by comparing the inter-cluster distances with the intra-cluster distances. The only assumption used in PERMANOVA is that the original observations are interchangeable under the assumption that the null hypothesis is true. In other words, this means the distribution of every subset of the data has a similar multivariate dispersion (variance) [6]. This assumption is made because a significant difference between the clusters can also be caused by differences in dispersion [6]. So when this assumption is satisfied the test only measures the "real" differences between the clusters. This assumption is satisfied for the obtained data and shown in 3.9. The PERMANOVA test was carried out by using the function "adonis2" from the package "vegan" [3].

A multivariate test was chosen since the main interest of this research is to analyze the recurrence of SE indicators as a whole. Performing tests per RQA variable would not give the most complete information about recurrence. The first step in PERMANOVA is defining a distance matrix.

Definition 2.3.1. The distance matrix indicates the dissimilarity between each pair of data points. Suppose there are n data points, $\mathbf{x}_i$. Let $||\cdot||$ be the chosen distance function. Then the distance matrix $\mathbf{M}$ is given by:

$$\mathbf{M} = \begin{pmatrix} 0 & ||\mathbf{x}_1 - \mathbf{x}_2|| & ||\mathbf{x}_1 - \mathbf{x}_3|| & \cdots & ||\mathbf{x}_1 - \mathbf{x}_n|| \\ ||\mathbf{x}_2 - \mathbf{x}_1|| & 0 & ||\mathbf{x}_2 - \mathbf{x}_3|| & \cdots & ||\mathbf{x}_2 - \mathbf{x}_n|| \\ ||\mathbf{x}_3 - \mathbf{x}_1|| & ||\mathbf{x}_3 - \mathbf{x}_2|| & 0 & \cdots & ||\mathbf{x}_3 - \mathbf{x}_n|| \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ||\mathbf{x}_n - \mathbf{x}_1|| & ||\mathbf{x}_n - \mathbf{x}_2|| & ||\mathbf{x}_n - \mathbf{x}_3|| & \cdots & 0 \end{pmatrix}$$

Subsequently, PERMANOVA calculates the sum of squared distances between the data points of a cluster. This is eventually compared with its centroid and compares it to squared distances between the different centroids.

Definition 2.3.2. A centroid of data points is defined as the average of these points. Let $\mathbf{G}_j$ be the centroid of cluster j and suppose cluster j has n data points which are d-dimensional, denoted as $\mathbf{x}_i$. Then $\mathbf{G}_j$ can be calculated as follows:

$$\mathbf{G}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \tag{2.18}$$

Definition 2.3.3. Suppose there are k clusters, then the total squared distances within the clusters, SS_{within}, is given by:

$$SS_{within} = \sum_{l=1}^{k} \sum_{\mathbf{x} \in \text{cluster } l} ||\mathbf{x} - \mathbf{G}_l||^2 \tag{2.19}$$

This calculation is actually done using the distance matrix, $\mathbf{M}$. This is possible due to a theorem stating that the sum of squared distances from individual points to their centroid is equal to the sum of squared interpoint distances divided by the number of points, a theorem first formulated by Christiaan Huygens in the 17th century [7]. This theorem is used as for many distance measures the calculation of a central data point (centroid) may be problematic [6]. For completeness, the proof of this theorem is given in Appendix A.3.

In practice, the distances are thus calculated without ever calculating the centroids. Using Huygens theorem the distance within all clusters is thus calculated as follows:

$$SS_{within} = \sum_{l=1}^{k} \left(\frac{1}{2n_l} \sum_{\mathbf{x} \in \text{cluster } l} \sum_{\mathbf{y} \in \text{cluster } l} ||\mathbf{x} - \mathbf{y}||^2 \right) \tag{2.20}$$

Now, still the calculations for the sum of squared distances between clusters cannot be done, since Huygens' theorem cannot be applied. Here the fact is used that the sum over the squared distances within clusters and the squared distances between clusters is equal to the total sum of the squared distances of the whole data set [7].

Definition 2.3.4. Denote the total sum of squared distances by SS_{total}, let the centroid of the whole dataset be denoted by $\mathbf{G}$ and suppose that there are m total data points, then SS_{total} is given by:

$$SS_{total} = \sum_{i=1}^{d} ||\mathbf{x}_i - \mathbf{G}||^2 \tag{2.21}$$

Again using Huygens' theorem SS_{within} can be rewritten to:

$$SS_{total} = \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} ||\mathbf{x}_i - \mathbf{x}_j||^2 \tag{2.22}$$

Now the the sum of the squared distances, denoted by $SS_{between}$, can easily be calculated as the difference between SS_{total} and SS_{within}. Next, MD_{within} and $MD_{between}$ are introduced. These are measures that are obtained by dividing SS_{within} and $SS_{between}$ by their Degrees of Freedom. This ensures that MD_{within} and $MD_{between}$ are unbiased estimators for the variation within and between clusters. Similarly to the covariance, if a cluster has n_j data points, the sum of squared distances of one cluster has $n_j - 1$ Degrees of Freedom. Therefore SS_{within} has $d - k$ degrees of freedom, this is obtained by summing over the clusters. The variance attributed between the clusters is equal to the total variance minus the variance explained within the clusters. Therefore the degrees of freedom of $SS_{between}$ is equal to the degrees of freedom of SS_{total} minus the degrees of freedom of SS_{within}. Thus if every cluster has n_l data points and the total amount of data points is d, the degrees of freedom of $SS_{between}$ is given by:

$$Df_{between} = (d-1) - \sum_{l=1}^{k} (n_l - 1) = d - 1 - d + k = k - 1 \tag{2.23}$$

Where k is the number of clusters. So MD_{within} and $MD_{between}$ are defined as follows:

Definition 2.3.5. The mean sum of squares within and between clusters are denoted by:

$$MD_{within} = \frac{SS_{within}}{d-k} \tag{2.24}$$

$$MD_{between} = \frac{SS_{between}}{k-1} \tag{2.25}$$

The F-statistic is defined to give information about the ratio between MD_{within} and $MD_{between}$ and thus gives information about the relative magnitude of the variance between groups compared to the variance within groups. Specifically, a higher F-statistic indicates that the variability between groups is significantly larger than the variability within groups, suggesting that the differences between the groups are statistically significant.

Definition 2.3.6. The F-statistic for the PERMANOVA test shows the ratio between the variance between clusters and the variance within clusters and is given by:

$$F = \frac{MD_{between}}{MD_{within}} \tag{2.26}$$

The PERMANOVA method tests the null hypothesis that there is no difference between the group means [6]. The alternative hypothesis is that there exist two clusters that have a significantly differing mean. The calculation of the p-value is done by randomly assigning all the data points to clusters for a given number of permutations, N (the default is 999). In this thesis, 10.000 permutations were used to calculate the p-values. For these random permutations of the clusters the same calculations are done and an F-statistic is calculated. The eventual p-value is given by the number of times the random permutations bring forth a larger F-statistic than the given distribution divided by the number of permutations.

Definition 2.3.7. The p-value for the PERMANOVA test gives indication as to with what certainty the conclusion can be drawn that the means of clusters are statiscally different and is defined by:

$$p = \frac{\sum_{i=1}^{N} \mathbb{1}\left(F_{permutation,i} \geq F\right)}{N} \tag{2.27}$$

Where $\mathbb{1}$ is the indicator function and F is the statistic observed by the real distribution of the clusters. This p-value is a measure of certainty that the means of the clusters of the real observation are not different from each other.

Chapter 3

Application to data from PRIME

The PRogramme of Innovation in Mathematics Education (PRIME) is part of the Interfaculty Teaching from the department of Applied Mathematics at TU Delft, responsible for redesigning mathematics courses for engineers [26]. PRIME has the following goals [11]: improve study results, enhance the connection between mathematics and engineering and stimulate the active participation and motivation of students. In this chapter, data gathered by PRIME will be analyzed to investigate the relation between SE and course performance. Firstly the SE indicators, Time Spent, Attempts Made and On-Time Rate are discussed. Secondly, the data and the program it was gathered from (Grasple) are introduced. Thirdly, the clustering of the students is addressed. Fourthly, the parameters and units of analysis are discussed. Finally, an interpretation of the RQA variables in the context of SE is given, followed by the results.

3.1 Study overview

Given the vast diversity of SE interpretations in the literature, it is important to define the SE indicators used rigorously. This will be done based on the theoretical definition of SE given by Ben-Eliyahu et al. (2018) [8] and the practical application on online SE, as used in Saqr & López-Pernas (2021) [31].

3.1.1 Time Spent

The definition used in this research for the variable Time Spent is the following: "The total amount of minutes a student actively spends on online homework exercises per unit of analysis". How the amount of active minutes was calculated is discussed in Chapter 3.5. This variable was also used in Saqr & López-Pernas (2021) [31] with the name "Total session duration". Time Spent is a well studied indicator of SE, but research suggests that it should be examine in combination with other indicators to better understand SE [13]. A student can for example have an online exercise open, but not be actively occupied with the task at hand. In the definition of Ben-Eliyahu et al. (2018) [8], Time Spent can be seen as a subpart of involvement.

3.1.2 Attempts Made

Attempts Made is defined as follows: "The total amount of attempts (correct or incorrect) a student makes on online exercises per unit of analysis". In Saqr & López-Pernas (2021) [31], this variable is also used, under the name Frequence of Action. Which indicated the frequencies of interactions of multiple activities, such as videos, forums, etc. In the dataset used for this research the information gathered was about when students answered exercises and thus there was no information about different types of actions. This definition of Attempts Made includes the first time a student attempts an exercise and also when a previously (in)correctly answered exercise is tried again. In light of the definition used in Ben-Aliyahu et al. (2018) [8], this definition encaptures involvement, but also persistence, as it includes “extra” attempts a student undertakes.

3.1.3 On-Time Rate

On-Time Rate is defined as: “The fraction of exercise sets that were attempted before the deadline, out of all the assigned exercise sets for each week”. Firstly, it is important to note that the On-Time Rate is only defined for a weekly defined time series. The reasons for this will be discussed in Chapter 3.6. Furthermore, it stands out that in this definition not the term “exercises”, but rather the term “exercise sets” is used. This is because in the data it cannot be seen, which exercise exactly is attempted by a student, but only which exercise set. In the definition of Attempts Made it is possible to keep the term “exercises”, since it does not matter which exact exercise is attempted.

Bourquet (2024)[9] revealed a significant negative relationship ($r = -0.6$) between the deviation from the recommended study path and final examination scores. The recommended study path was defined as the expected completion dates for exercises. The deviation was defined as “the sum of the day differences between the expected completion date and the actual first completion date of an activity by the student”. This shows that completing exercises close to the assigned date leads to a better understanding of the study material. Therefore, the choice was made to also analyze if recurrence in this SE indicator is correlated with performance.

3.2 Data

The initial dataset comprised 236 first-year computer science, civil engineering, and mechanical engineering students following a first-year Linear Algebra course. Since the course load for the computer science students was greater and the profile of these students is quite different from mechanical and civil engineering students, the choice was made to not include computer science students in this research. After having deleted the computer science students, 108 mechanical engineering students and 60 first-year civil engineering students were left. Both of these groups took the same Linear Algebra course at the TU Delft. They followed the course during the same 8-week period, had the same structure and had to make the same assignments. Therefore the decision was made not to differentiate between these students. The students who did not achieve any grade for the final exam were also excluded, leaving 159 students. 15 students took the exam, but had fewer than 3 study sessions, therefore there was almost no data to analyze. Furthermore, the RQA variables that would be obtained for these students would greatly skew the overall distribution. From these 144 students that were left 108 identified as male, 31 as female, 2 as another gender not specified, 1 as non-binary and 2 preferred not to say. 118 students were younger than 20 years old and 26 students were between the ages of 20 and 25. Course performance was measured by the final exam grade. The average grade of the course was 5.51 with a standard deviation of 1.78.

3.3 Clustering

K-means clustering was used to identify different groups of students based on their grades for the final exam. K-means clustering works as follows. If the optimal number of clusters is n, the algorithm picks n random starting points (centers). Every observation (student) is then associated with the center nearest to it. Of all the formed clusters the gravity centers are calculated. These are the new centers and again every observation is paired with the center nearest to it. These steps are iterated until convergence (no difference between steps).

Firstly, the optimal number of clusters needs to be determined. This is done by using the Elbow-method. This method calculates the Within-Cluster Sum of Squares (WCSS) for a certain range of number of clusters. WCSS is defined as the sum of the distance between all data points and the center of their assigned cluster. If G is defined as the set of all the grades and μ_i as the center associated with grade g_i mathematically WCSS can be described as follows:

$$\text{WCSS} = \sum_{g_i \in G} |g_i - \mu_i|^2$$

The Elbow-method plots the WCSS value calculated for certain numbers of clusters. In this plot the Elbow-point is located at the position on the x-axis where the graph starts to diminish at a lower rate, creating an angle (elbow) in the plot. The number of clusters associated with this location in the plot, is considered to be the optimal amount of clusters since adding extra clusters does not have that great of an impact in lowering the distance between data points and their assigned centers. The Elbow-plot showing 1 to 10 clusters can be seen in Figure 3.1

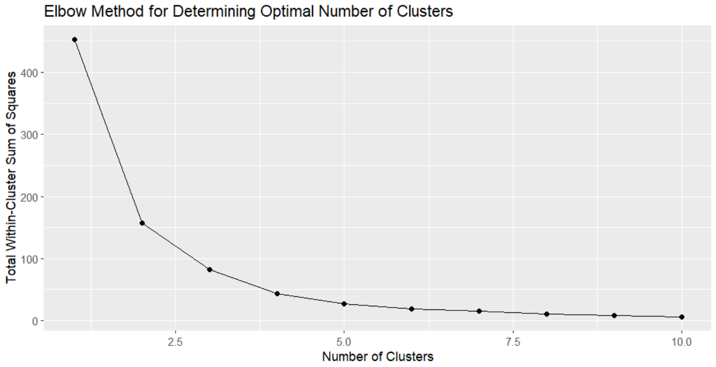


Figure 3.1: Elbow plot showing WCSS for clustering based on grades

Based on this plot 2,3 or 4 clusters all seem plausible. The choice was made to focus on 3 clusters. Choosing 3 clusters creates clusters of the form “low”-performance, “middle”-performance and “high”-performance. In this way differences between students just passing the course and students performing excellently, can also be researched, unlike the case when only having 2 clusters. If 4 clusters are chosen, the difference between the number of students in the largest cluster (54) and the smallest cluster (18) becomes quite significant.

The following distribution of students is obtained:

Cluster	number of students	center	sd
“low”-performing	39	3.26	1.00
“middle”-performing	62	5.52	0.59
“high”-performing	43	7.53	0.74

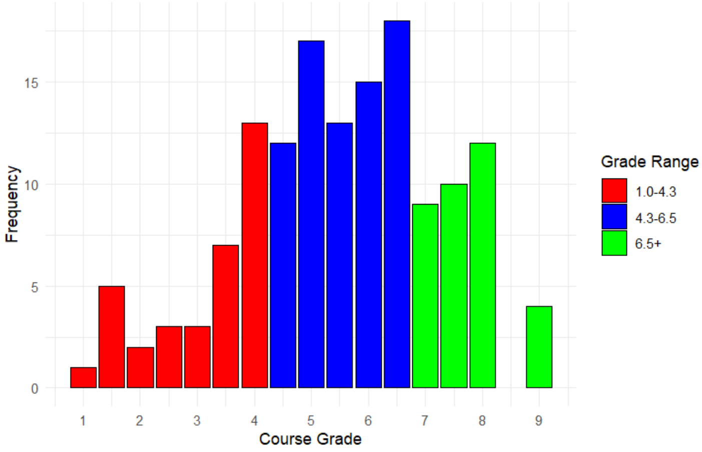
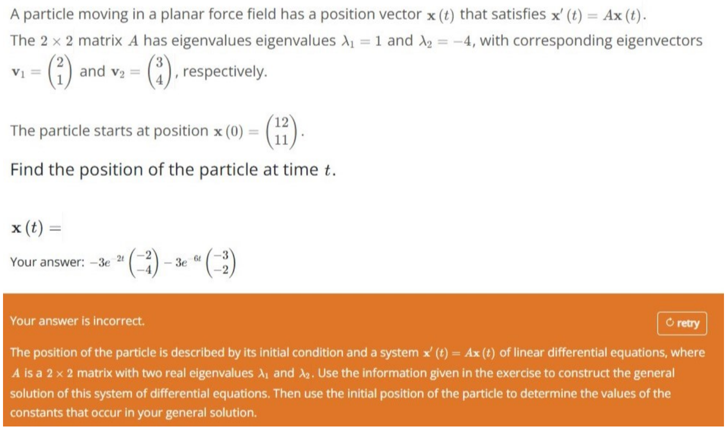


Figure 3.2: Distribution of clusters based on grades

3.4 Online Learning Platform

The students were asked to complete the assigned exercises in the online learning platform Grasple [14]. In Grasple the exercises were split up per lecture and subsequently into subtopics. Every subtopic had a different amount of exercises and a different minimum of correct exercises to complete the topic. Whenever an exercise is answered incorrectly, Grasple provides a hint for the exercise and students get an extra attempt. If the answer provided is wrong again, the correct answer is given by the program. In Figure 3.3 an example can be seen of how the program gives feedback. Students were given the possibility to make 1220 exercises over the 8 weeks of the course. 656 of these exercises were labeled as “optional”. Datasets on student online activity were exported from Grasple. This data was processed to develop the examined SE indicators.



A particle moving in a planar force field has a position vector $\mathbf{x}(t)$ that satisfies $\mathbf{x}'(t) = A\mathbf{x}(t)$.
The 2×2 matrix A has eigenvalues eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -4$, with corresponding eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$, respectively.

The particle starts at position $\mathbf{x}(0) = \begin{pmatrix} 12 \\ 11 \end{pmatrix}$.

Find the position of the particle at time t.

$\mathbf{x}(t) =$

Your answer: $-3e^{2t}\begin{pmatrix} -2 \\ -4 \end{pmatrix} - 3e^{6t}\begin{pmatrix} -3 \\ -2 \end{pmatrix}$

Your answer is incorrect.

The position of the particle is described by its initial condition and a system $\mathbf{x}'(t) = A\mathbf{x}(t)$ of linear differential equations, where A is a 2×2 matrix with two real eigenvalues λ_1 and λ_2. Use the information given in the exercise to construct the general solution of this system of differential equations. Then use the initial position of the particle to determine the values of the constants that occur in your general solution.

Figure 3.3: Example of how the Grasple program works

3.5 Data adaptation

One of the variables in the dataset obtained from Grasple was “exercises answered at”, showing when a student answered a certain exercise. This variable is used to determine the total amount of time spent per unit of time. This method is used in several other papers. Liu et al. (2015) [5] researched the optimal threshold to describe different study sessions of a student and found a time cap of 45 minutes of inactivity to describe one session. However, Francišković et al.

(2024) used a time cap of 3 minutes [4], Jovanovic et al. (2019) [21] used a time cap of 15 minutes and Jovanovic et al. (2021)[22] used a time cap of "the 85th percentile of the time gaps between two successive learning actions". Looking at the distribution of the differences between study actions for the data analyzed in this thesis, which is shown in Figure 3.4, it can be seen that the greatest part of the differences is below 5 minutes. Using the 85th percentile of this distribution (2.5 minutes) resulted in unreasonably many study sessions, as multiple students had more than 100 study sessions. Therefore the time cap of 15 minutes was chosen, as this resulted in a reasonable average of 30.4 study sessions per student. So a study session is defined as: A period of successive exercises attempted, where the time difference between these exercises was not larger than 15 minutes.

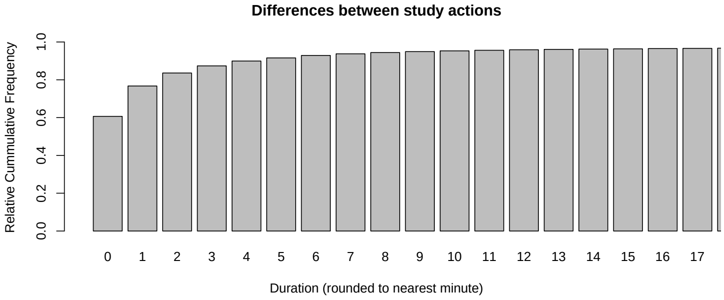


Figure 3.4: Distribution of time difference between study actions

Using this assumption the number of Attempts Made per study session per student and the Time Spent in each session was determined. To determine the Attempts Made and the Time Spent per day/week, the same assumption was used, only then the study sessions per day/week were added together. The one influence this approach has is that a study session can spread over 2 different days/weeks. This occurs whenever a certain study sessions starts before midnight and ends after. So it might occur that one study session influences 2 days/weeks, but this does not have a meaningful impact on the analysis.

3.5.1 Distribution of number of study sessions

Using the cap of 15 minutes the following amount of study sessions per student was found which are visualized in Figure 3.5. It stands out that there are no students in the high-performing cluster that have a relatively large amount of study sessions, there is however one student in this cluster that falls of the plot with a total of 186 study sessions.

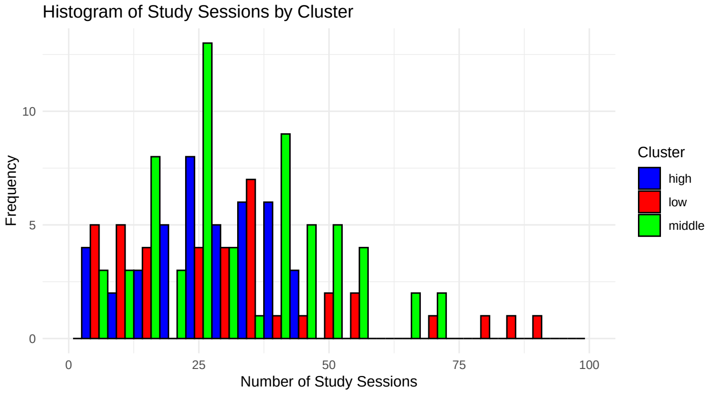


Figure 3.5: Histogram of number of study sessions per student

3.6 Unit of analysis

To perform an analysis of the recurrence of the SE indicator per student, a unit of time has to be defined. Initially, three units of analysis were seen as interesting, namely study sessions, days and weeks for the variables Time Spent and Attempts Made. For the On-Time Rate, only weeks was seen as a logical measure of time, since study sessions do not have deadlines and not every day had deadlines. Using the three units of time, RP's were produced to gain initial information about what the characteristics are of every unit of analysis for Time Spent and Attempts Made. For all of the different engagement factors plots were created, but to give the reader insight into the advantages and disadvantages of all options, the plots for Time Spent will be discussed. Using the initial value of 10% of the diameter of the data the following margins were found. The maximum time spent in one day was 495 minutes and the minimum was 0 minutes. Therefore a margin of 50 minutes will be used for this unit of analysis. For Time Spent per week a minimum of 0 minutes and a maximum of 1121 minutes was found and a margin of 110 minutes will be used. Regarding study sessions a minimum of 0 and a maximum of 190 minutes were obtained and a margin of 20 minutes will be used. As an example the three RP's created based on the behavior of student “S138” can be seen in Figures 3.6, 3.7, 3.8

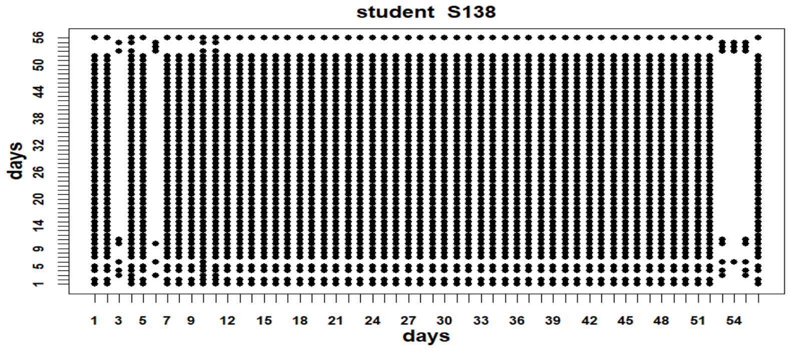


Figure 3.6: Recurrence Plot Student 138 with days as unit of analysis

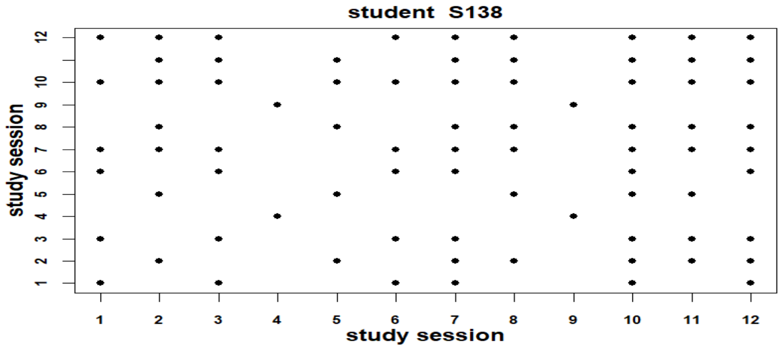


Figure 3.7: Recurrence Plot Student 138 with study sessions as a unit of analysis

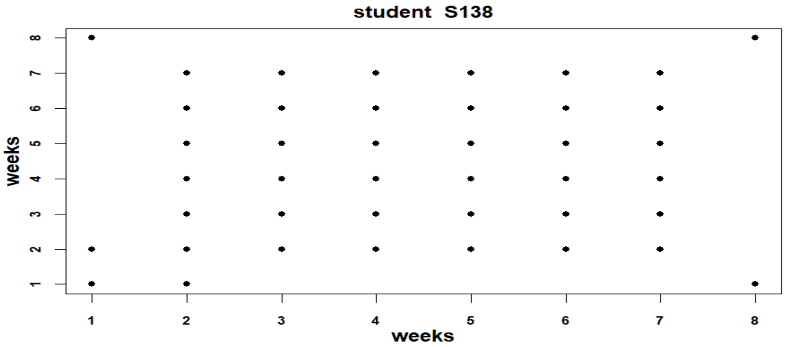
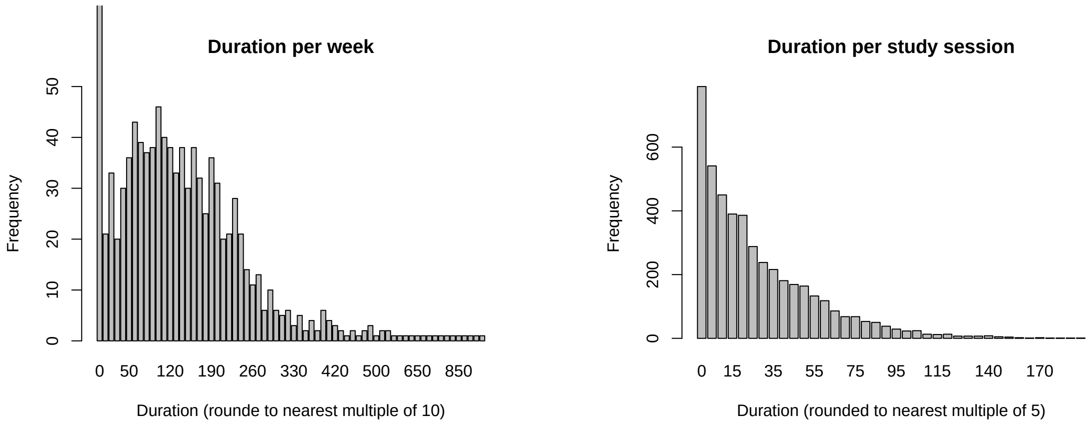


Figure 3.8: Recurrence Plot Student 138 with weeks as a unit of analysis

In the plot using days as the unit of analysis (Figure 3.6) a lot of recurrence points are shown, which in itself is not a problem. However, in this case these points are almost all occurring due to the fact that this student has a lot of days where 0 minutes were spent studying. To be exact from a total of 56 days, no activity was recorded for 48 days. This does not give us a lot of useful information to analyze but does have a great impact on the RQA variables. It is important to note that this student is not an exception, but rather follows the rule since of all 9016 days analyzed, on 6819 days no time was spent studying. Therefore the decision was made that using days as a unit of analysis would not be beneficial.

To determine if weeks and/or study sessions as a unit of analysis contain interesting aspects the histograms in Figure 3.9a and Figure 3.9b show how Time Spent is distributed for both Time Spent per study session and Time Spent per week.



(a) Occurrences of Time Spent per week

(b) Occurrences of Time Spent per session

In both plots there still are a lot of occurrences of 0 minutes spent. In fact, the number of occurrences of 0 minutes spent per week is so great that it falls off the plot. In reality, this happens in 330 of the total 1288 weeks. It could be argued that the same problem occurs as with the analysis of days. However, it is important to note that these data points also have value for our analysis. Firstly, it shows that a student was inactive in a week which is an important fact to take into account when researching study behavior. Secondly, especially for the main question posed in this thesis the focus lies on recurring behavior, which also includes not studying at all.

Looking at the histogram of the study sessions it seems as if the impact of the sessions lasting 0 minutes is even greater, however this histogram shows rounded values. The actual number of sessions with a duration of 0 minutes is 232 out of a total of 4588 study sessions. The less drastic impact of the study sessions with a duration of 0 minutes, is because a study session of 0 minutes only occurs when a student only attempts one single exercise, which does not happen often. Furthermore, it is important to note that the variance of the distribution of the study sessions is notably lower. The choice for analyzing the behavior of students with study sessions as a unit of analysis seems straightforward. Nevertheless, there is another issue with this unit of analysis. By the definition used for study sessions, expressed in Chapter 3.5, the number of study sessions is not set per student. This results in time series of varying sizes. This can have a great impact on the RQA variables. If for example, a student has only two study sessions, the Recurrence Rate can drop from 100% to 50%, if these two sessions are not recurrent. Therefore, students with less than three study sessions were not considered in the analysis. Both weeks and study sessions will be used as a unit of analysis to calculate the RQA variables. An analysis will then be made of the impact of the different units of time.

3.7 Value of margins

To produce the RP's per student and afterward calculate the RQA variables, choices need to be made regarding the margins used. For the margin, ϵ, as discussed in Chapter 2.1.1 the research already done using RQA suggests using a margin with a value smaller than 10% of the diameter of the data (the greatest difference between the observations) [24]. Here the maximum and minimum over all students are used, and thus no different margins per student are defined, else this would give unequal results comparing students. The parameter of 10% of the diameter was initially used to see if this produced interesting results. For all of the different variables and units of analysis, which will be discussed in Chapter 3.6, this diameter was taken as a starting point. If the Recurrence Rate with this initial margin was too high for specific variables, lower margins were chosen based on the distribution of the data. RR was used to indicate usable values since it directly quantifies what part of the data points is seen as recurrent, which is the direct goal of the value of the margin. In this thesis, an average RR in the neighborhood of 50-60% was aimed for. It should be noted that we are mostly interested in researching differences between the clusters and not in the exact numbers of recurrence.

As discussed in Chapter 3.6 initially a margin of 20 was used for the SE indicator Time Spent per study session. This margin was not altered since it produced an RR around 50%, as can be seen in Table 4.1. The initial margin for Time Spent per week caused a mean RR of 71,20%. After using trial and error with lower margins, a value of 70 was chosen, since it produced a mean RR of 56,38% which was more in the neighborhood of the value found using study sessions.

Similarly to how a margin was determined for Time Spent, the maximum and minimum number of Attempts Made per unit of analysis were used to get an indication of the margin of this variable. In Figure 3.11 the distribution can be seen for the Attempts Made per study session. Even more so than is the case with the distribution of Time Spent, this plot is heavily right-skewed. Therefore the maximum is disproportionately large in comparison with the rest of the data. For study sessions, a maximum of 370 attempts was found. Using the guidelines of 10% of the diameter of the data a margin of approximately 40 was obtained. Intuitively this margin would cause a very large RR, and this is indeed the case. Using the data of all three clusters an RR of 82,42% was found. This was seen as too high to gain valuable information and therefore a lower margin was chosen. For study sessions, a margin of 20 was chosen. This margin was chosen as it caused a reasonable RR (61,14%). Furthermore, the Recurrence Plots of the students showed varying structures.

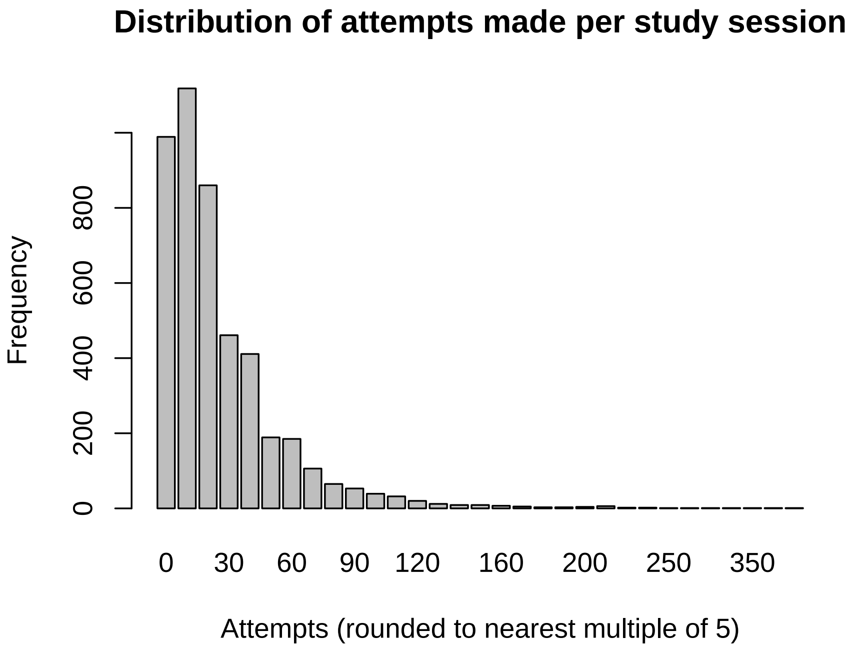


Figure 3.10: Occurrences of Attempts Made per session

In Figure 3.10 the distribution can be seen for the Attempts Made per week. Again this plot is heavily right-skewed. Therefore the maximum is disproportionately large in comparison with the rest of the data. A maximum of 2248 Attempts Made per week was found. Using the guidelines of 10% of the diameter of the data a margin was obtained of 220. Intuitively this margin would cause very large RR, and this is indeed the case. Using the data of all three clusters a RR of 91,31% was found. This was seen as too high to gain valuable information and therefore a lower margin was chosen. For Attempts Made per week, a margin of 70 was used. This margin was chosen as it caused a reasonable RR (60,94%) over all students.

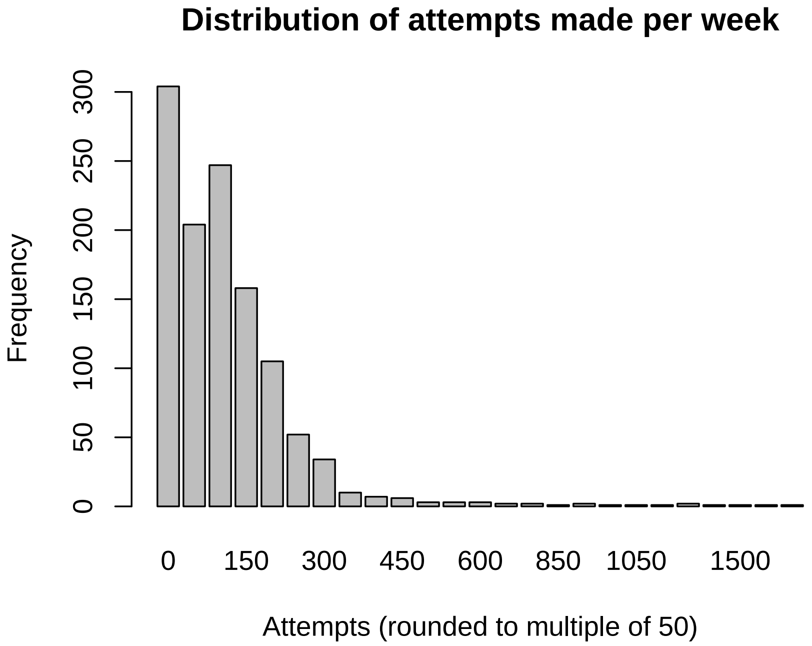


Figure 3.11: Occurrences of Attempts Made per week

For the On-Time Rate, a minimum of 0 and a maximum of 0.935 were found. The total distribution of the On-Time Rate per student can be seen in 3.12. Based on the distribution of the values and the produced RR, namely 46.97%, the decision was made to keep the margin at 10% of the diameter, which was chosen as 0.1.

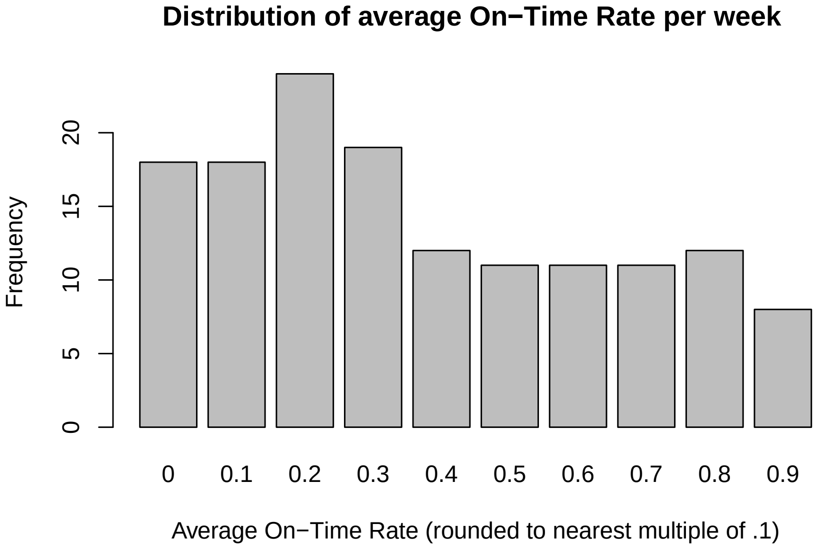


Figure 3.12: Occurrences of Mean On-Time Rate

3.8 Interpretation of RQA variables

Now that most of the methodology has been set out, first an indication of what the RQA variables, discussed in Chapter 2.1.5, mean in the context of this research. In Chapter 3.5 examples are given of how the time series data for different students look. In this section the interpretation of the RQA variables will be given for a "general" SE indicator. Later the five engagement indicators will be analyzed one by one. These indicators are: Time Spent per study sessions, Attempts Made per study session, Time Spent per week, Attempts Made per week and On-Time Rate per week.

Recurrence Rate

As discussed previously, the Recurrence Rate (RR) shows what percentage of the data is similar to each other. What conclusions can be drawn from this information? A high RR shows that a student has similar behavior for a large part of the time series: either the study sessions of a student or the weeks of the course. Note that this does not give any indication of the exact observed engagement value. It only gives information about similarity. When linking this to the main research question, a high RR indicates overall recurrence in the engagement of a student. For the SE indicator Time Spent a high RR means that the student usually spends the same amount of time when studying. The interpretation for Attempts Made is similar. A high RR for the On-Time Rate indicates that for a large part of the weeks a comparable share of the assigned exercises was completed on time.

Percent Determinism

The value for Percent Determinism (DET) shows the part of recurrence points that occur in diagonal structures (with a minimal length of 2). DET is highly dependent on RR. If there is almost no recurrence, then the value of DET can easily become relatively high. A high value for DET means that there are a lot of recurrence points that are contained in diagonal structures. This means that at two different times in the time series, a student showed comparable engagement and one time step later still showed comparable engagement. This can be interpreted as a student following a certain study "pattern". In the case of Time Spent this could mean that after a shorter study session, a student always has a longer study session. In general, a high DET indicates that the engagement of a student follows a certain pattern.

Average Diagonal Line Length

The definition of the Average Diagonal Line Length (D_{avg}) is straightforward. In the context of SE, it shows that when a student repeats a similar sequence of values for the engagement factor twice, how long this sequence is on average. It should be emphasized that this variable should always be considered in combination with other variables. If for example only one diagonal line is shown in the RP, this is automatically the value of D_{avg} .

Laminarity

The Laminarity (LAM) shows what percentage of the recurrence points are contained in vertical/horizontal lines. A vertical line shows that the value of an engagement factor for multiple consecutive time steps is similar to one other value. In the context of SE, it indicates that a student has had multiple similar study sessions/weeks after each other. It is therefore an indicator of successive recurrence in the study process.

Trapping Time

The definition of the Trapping Time (TT) is similar to D_{avg} only then defined for vertical lines. It thus shows if a student has a sequence of similar time series entries, how long on average such a sequence is. The same caution as with D_{avg} should be applied to this variable.

Shannon's Entropy of the diagonal lines

As said earlier Shannon's Entropy of the diagonal lines ($ENTR$) is a measure of the expected "surprise" in the distribution of the lengths of the diagonal lines. In other words, $ENTR$ has a low value if one length occurs many more times than others. Using Jensen's inequality, which is proven in Appendix A.1, a proof is given in Appendix A.2 considering the fact that $ENTR$ is maximized if and only if the distribution of the lengths of the diagonal lines is uniform. Thus a low value for $ENTR$ can be interpreted as a distribution different from uniform, i.e. one length occurs much more often than others. In the context of SE a low $ENTR$ can be seen as an indication of that a great part of the diagonal lines, of which the interpretation is given in 3.8, are of the same length. In practice it is hard to directly interpret this variable. Therefore, it does not give concrete information about the SE behavior of a student. Especially considering the main question posed in this thesis, $ETNR$ does not give helpful information. The decision was thus made to not include $ENTR$ in further analyses, as it would only influence the results, while not containing any useful information. In Poquet et al. (2023) [25] Entropy was indeed used as a measure of recurrence in SE. However, it is important to note in that case The Entropy of the behavioral states was used, instead of the Entropy of the diagonal lines, which is a completely different variable.

3.9 Normality of the data

Before it can be analyzed if the clusters are indeed different. First information needs to be gathered about the distribution of the RQA variables. Many tests that are used for interpreting the difference between clusters use the assumption of normality of the data, such as the t-test and Multivariate Analysis of Variance (MANOVA). To test the normality of the data for every combination of engagement indicator, cluster and RQA variable the Shapiro-Wilk test [30] for normality was applied. The Shapiro-Wilk test has the null hypothesis that the data is normally distributed and thus a low p-value indicates that the data is not normally distributed. For every combination of SE indicator, cluster and RQA variable the normality was tested and the distribution of the p-values can be seen in Figure 3.13
A substantial part of p-values are below the significance level of 0.05 and even below 0.01. From this fact, the conclusion can be drawn that most of the distributions violate the assumption of normality. Therefore, PERMANOVA was used to analyze if there were significant differences between the RQA variables of the different clusters.

3.10 Variance of the data

To use the PERMANOVA test to analyze the differences between the clusters, the assumption of interchangeability, discussed in Chapter 2.3, needs to be satisfied. Therefore the "PERMDISP" test is applied to see if the clusters vary in their variance. This test is applied using the "betadisper" function from the package "vegan" [3]. This test has the null hypothesis that there are no differences between the dispersion of the clusters. A low p-value therefore shows that there are no significant differences regarding this aspect. The obtained results are shown in 3.1 and show that there are no significant differences in the variance of the clusters. Therefore, the PERMANOVA test can be applied to the data.

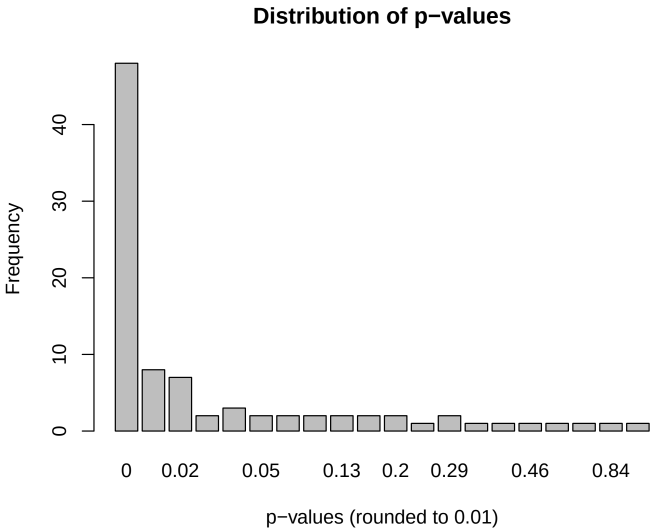


Figure 3.13: p-values for every combination of engagement indicator, RQA variable and cluster

Indicator	p-value
Time Spent study sessions	0.262
Attempts Made study sessions	0.194
Time Spent weeks	0.190
Attempts Made weeks	0.470
On-Time Rate weeks	0.951

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3.1: p-values obtained by “betadisper”

Chapter 4

Analysis of Results

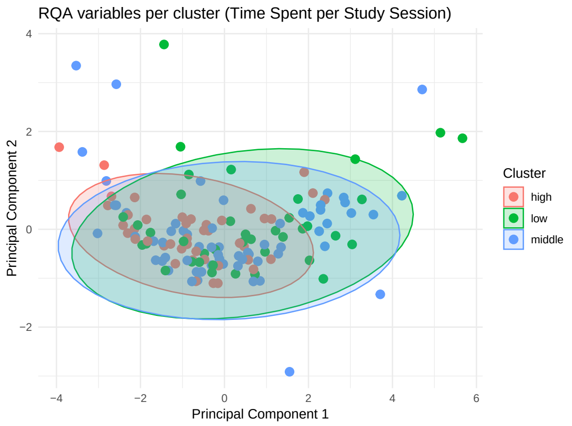
For the five different engagement indicators (Time Spent per study sessions, Time Spent per week, Attempts Made per study sessions, Attempts Made per week and On-Time Rate per week), five different RQA variables were calculated (RR, DET, D_{avg}, LAM and TT). $ENTR$ was not used, since it did not provide useful information about the recurrence in SE. In this chapter, the analysis of these results is discussed. The main intention is to determine if a significant difference between these variables for the different clusters exists. The partitioning of the clusters based on grades can be seen in Figure 3.2. First, the overall means and standard deviations for every RQA variable for every cluster will be discussed. Second, the plots and Principal Components obtained by PCA will be introduced and thirdly, the PERMANOVA results will be analyzed.

4.1 Time Spent per study session

The margin of 20 for Time Spent per study sessions produced the following values for the RQA variables. To give a more complete overview of how every cluster had spent time over the course, a variable showing the total amount of time spent is also included under the name $TotalTime$.

Variable	Grade	M	SD
RR	Lowest	55.11	14.30
	Medium	53.67	14.27
	Highest	48.12	10.39
DET	Lowest	74.35	10.72
	Medium	73.83	11.31
	Highest	68.27	9.54
D_{avg}	Lowest	3.43	0.62
	Medium	3.27	0.52
	Highest	3.12	0.34
TT	Lowest	3.95	1.23
	Medium	3.62	1.22
	Highest	3.17	0.85
LAM	Lowest	78.91	15.08
	Medium	76.40	19.60
	Highest	70.50	15.92
$TotalTime$	Lowest	826.88	681.06
	Medium	862.17	484.57
	Highest	947.63	735.98

Table 4.1: RQA Variables obtained for SE indicator Time Spent per study session



variable	PC1	PC2
λ	1.865	0.892
RR	0.463	-0.158
DET	0.482	-0.098
D_{avg}	0.374	0.777
TT	0.485	0.155
LAM	0.423	-0.581

Figure 4.2: PCA plot of Time Spent per study session and corresponding Principal Components

From the data based on study sessions as a measure of time, seen in Table 4.1 it can be deduced that the means of the clusters are inversely related to performance, i.e. the highest-performing cluster has the lowest means for all of the RQA variables. Furthermore, the total time spent is related with performance, which is shown by the variable $TotalTime$. It is shown that the highest-performing cluster has the highest amount of total time spent on average, then the middle-performing cluster and the lowest-performing students have the lowest amount of total time spent on average.

The clusters created by PCA seem to be quite different at first glance, based on the plot shown in Figure 4.2. The high-performance cluster is more toward the left side of the plot. PC1 shows similar positive values for all the RQA variables. This shows that the RQA variables are all correlated with each other and also that overall the high-performance cluster has lower values for the RQA variables. Along the y-axis, showing the values for PC2, no great differences between the clusters seem to exist.

The difference between the clusters is confirmed by the PERMANOVA test, showing a significant difference between the clusters, with a p-value of 0.012. The PERMANOVA test only shows if there exists a significant difference between the clusters, but not between which clusters. Therefore, the test was performed again in a pairwise manner. These results are summarized in Table 4.3

Compared clusters	F-statistic	p-value
Low & High	7.0644	0.002 ***
Low & Middle	0.773	0.445
Middle & High	3.998	0.025 **

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4.3: F-statistic and p-values produced by PERMANOVA

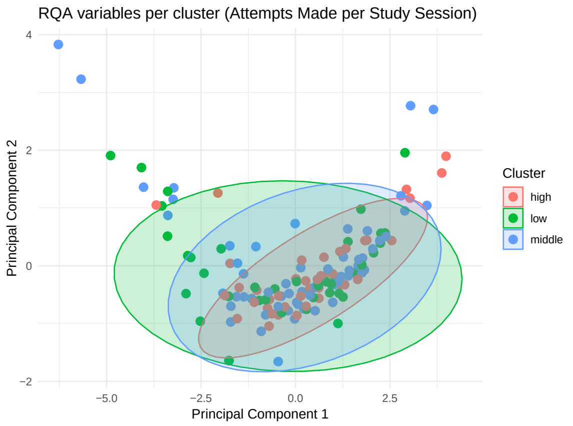
From the results in Table 4.3 it is concluded that the high-performing cluster differs from both the middle- and low-performing cluster. Based on PC1 and the plot, it can be seen that the high-performing cluster has lower values regarding the RQA variables. This is again confirmed by the summarizing statistics given in Table 4.1. Interpreting the RQA variables in the context of SE, the outcome of this analysis is that high-performing students have more variability in the duration of their study sessions, as they have lower values for the RQA variables.

4.1.1 Attempts Made per study sessions

The results obtained for study sessions using a margin of 20 can be found in Table 4.4. To give a more complete overview of the amount of attempts of every cluster a variable showing the total amount of Attempts Made is also included under the name $TotalAttempts$.

Variable	Grade	M	SD
RR	Lowest	65.31	20.09
	Medium	60.86	16.82
	Highest	57.75	14.15
DET	Lowest	81.65	13.61
	Medium	77.76	15.05
	Highest	76.49	13.72
D_{avg}	Lowest	4.13	1.45
	Medium	4.05	1.57
	Highest	3.69	0.79
TT	Lowest	5.54	3.03
	Medium	5.06	3.15
	Highest	4.27	1.86
LAM	Lowest	84.73	15.86
	Medium	82.47	15.44
	Highest	79.55	17.75
$TotalAttempts$	Lowest	737.23	642.90
	Medium	832.85	612.32
	Highest	928.00	1224.34

Table 4.4: RQA Variables obtained for SE indicator Attempts Made



variable	PC1	PC2
λ	1.950	0.903
RR	-0.469	-0.204
DET	-0.465	-0.317
D_{avg}	-0.415	0.632
TT	-0.463	0.432
LAM	-0.420	-0.521

Figure 4.5: PCA plot of Attempts Made per study session and corresponding Principal Components

Similar patterns are found for Attempts Made in comparison with Time Spent. Again an inverse relationship between performance and the RQA variables is shown using study sessions as a unit of analysis, as seen in Table 4.4. Furthermore, there exists a positive relation between the mean of the total Attempts Made by each cluster and the performance of students, i.e. the highest-performing cluster has the highest mean, then the middle-performing cluster and the low-performing cluster has the lowest mean.

In the PCA plot obtained for Attempts Made per study session (Figure 4.5), it is shown that the clusters have similar patterns. However, it is also clear that the cluster for high-performers has less variance than the other clusters in the Principal Components. This is also shown in Table 4.4, where the standard deviation for the higher cluster is significantly lower for most RQA variables. So the students of the high-performance cluster all show relatively similar behavior regarding the recurrence in the Attempts Made per study session. PC1 indicates that the primary variance in the data is explained by the overall values of the RQA variables, which are highly correlated, as evidenced by their similar negative values. The cluster of high-performing students seems to be slightly more towards the right side of the plot, indicating overall lower values for the RQA variables. This would indicate the the highest-performing students show less recurrence in the amount of Attempts Made per study session. Along the y-axis, showing the values for PC2, no great differences between the clusters seem to exist.

From the PERMANOVA results the conclusion can be drawn that the variance between the clusters is relatively greater than the variance within clusters, as the F-statistic has a value of 1.600. Together with the p-value of 0.175 given for the hypothesis that the clusters have no significant differences in the reported RQA variables, the conclusion can be drawn that the clusters show some difference in their recurrence behavior in the Attempts Made per study session, but this is not statistically significant. Since no significant difference was found, no pairwise comparisons between the clusters were carried out.

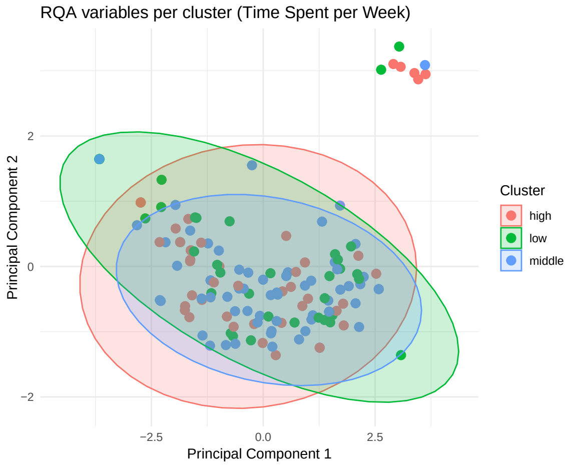
4.1.2 Time Spent per week

The summarizing statistics for the RQA variables for Time Spent per week using a margin of 70 can be found in Table 4.6. No clear pattern over all of the RQA variables is shown. For all of the variables, different clusters have the highest or lowest value.

The visual representation, shown in Figure 4.7, of the RQA values of the clusters obtained by analyzing the SE indicator of Time Spent per week, indicates that there are no great differences

Variable	Grade	M	SD
RR	Lowest	58.01	22.08
	Medium	55.59	17.29
	Highest	56.03	17.52
DET	Lowest	71.76	18.13
	Medium	77.78	15.28
	Highest	73.33	17.72
D_{avg}	Lowest	4.05	1.12
	Medium	3.69	0.86
	Highest	4.09	1.51
TT	Lowest	3.82	2.04
	Medium	3.44	1.45
	Highest	3.58	1.44
LAM	Lowest	75.86	21.56
	Medium	75.12	20.97
	Highest	77.03	22.44

Table 4.6: RQA Variables obtained for SE indicator Time Spent per week



variable	PC1	PC2
λ	1.771	1.027
RR	-0.518	0.124
DET	-0.474	-0.194
D_{avg}	0.172	0.911
TT	-0.498	0.340
LAM	-0.479	0.033

Figure 4.7: PCA plot of Time Spent per week and corresponding Principal Components

between the clusters. No cluster is clearly more on one side of the plot. PC1 reveals that almost all of the RQA variables are highly correlated with each other, with the exception of D_{avg}, which is slightly positive in contrast to the other variables. It is once more confirmed that D_{avg} behaves differently for this data with the high value shown for PC2. This suggests that the average diagonal line length is the cause of a lot of the variance in the data. In the context of SE; some students have the same pattern of Time Spent per week occurring repeatedly and others have no set rules in this pattern. Overall, it can be concluded that the clusters do not differ in the recurrence shown in the time they spent per week. It is noticeable that there were indeed differences between the clusters when study sessions were used as a unit of analysis. This might indicate that a week does not capture the complex dynamics of SE. Furthermore, the short length of the time series might also cause unusual values for the RQA variables. The differences between using study sessions or weeks are discussed in more detail in Chapter 4.2.

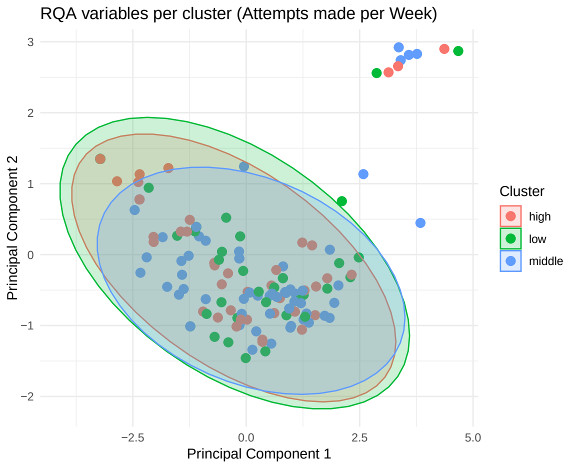
The results from the PERMANOVA test once again show that there are no great difference between the clusters regarding this SE indicator. A low F-statistic of 0.589 was found and a high p-value of 0.703. Since there are no significant differences between the clusters, no pairwise analyses were performed.

4.1.3 Attempts Made per week

The summarizing statistics for the RQA variables for Attempts Made per week using a margin of 70 can be found in Table 4.6. Using weeks as a unit of analysis, the medium-performing cluster most often has the lowest values for the RQA variables. Usually, the lowest-performing cluster has the highest values, but these do not differ much from the highest-performing cluster. A clear overall difference cannot be concluded from these values solely.

Variable	Grade	M	SD
RR	Lowest	63.22	21.18
	Medium	58.57	19.30
	Highest	62.29	19.90
DET	Lowest	73.10	16.94
	Medium	72.46	15.88
	Highest	75.59	16.45
D_{avg}	Lowest	4.04	1.18
	Medium	4.05	1.57
	Highest	4.04	1.27
TT	Lowest	4.15	2.08
	Medium	3.70	1.67
	Highest	3.94	1.83
LAM	Lowest	81.34	19.76
	Medium	76.74	21.22
	Highest	79.25	21.10

Table 4.8: RQA Variables obtained for SE indicator Attempts Made per week



variable	PC1	PC2
λ	1.793	1.025
RR	-0.520	0.108
DET	-0.475	0.057
D_{avg}	0.152	0.925
TT	-0.504	0.300
LAM	-0.477	-0.198

Figure 4.9: PCA plot of Attempts Made per week and corresponding Principal Components

For the plot obtained by PCA high resemblance can be seen between all the clusters. PC1 reveals that almost all of the RQA variables are highly correlated with each other, with the exception of D_{avg}, which is slightly positive in contrast to the other variables. It is once more confirmed that D_{avg} behaves differently for this data with the high value shown for PC2. This behavior was also seen when PCA was applied to Time Spent per week. This suggests that the average diagonal line length is the cause of a lot of the variance in the data. In the context of SE; some students have the same pattern of Attempts Made per week occurring repeatedly and others have no set rules in this pattern.

The PERMANOVA test reveals no statistically significant differences between the clusters, reporting an F-statistic of 0.552 and a p-value of 0.731. Therefore, again no further pairwise analyses were carried out.

An aspect that should be commented on regarding this SE indicator, is the great resemblance with all the values obtained for Time Spent per week. It is shown that the composition of the PC's is highly similar and also the values for the F-statistic and the p-value are alike. Naturally, the Time Spent and Attempts Made per week are correlated with each other. However, it is noticeable that there are greater differences between Time Spent and Attempts Made when the analysis is done using study sessions. This might indicate that weeks summarize the SE behavior too much, while study sessions do capture the complex dynamics. Based on the statistics from the PERMANOVA test, the conclusion is again drawn that the clusters do not show differences in behavior regarding the recurrence of the number of Attempts Made per week. As was the case with Time Spent, again the clusters show less variance when weeks is used as opposed to study sessions. Possible reasons that have caused this are identical to explanations discussed in the case of Time Spent in Chapter 4.1.2.

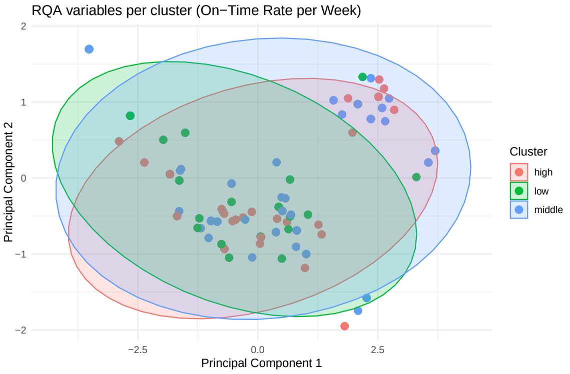
4.1.4 On-Time Rate per week

The values found for the RQA variables of the SE indicator On-Time Rate per week with a margin of 0.1 are reported in Table 4.5. To give more context regarding the behavior of every cluster, the average On-Time Rate, indicated by AVG_{ontime}, is also included.

For the RQA variables, the lowest-performing cluster usually has the highest mean. The only variable violating this principle is D_{avg}. So there exists a small indication that lower-performing students show more recurrence in the part of exercises they have completed on time each week.

Variable	grade	M	SD
RR	lowest	50.50	19.58
	medium	46.08	21.32
	highest	45.04	17.48
DET	lowest	72.51	18.22
	medium	70.96	16.15
	highest	71.85	17.23
D_{avg}	lowest	4.10	1.30
	medium	4.53	1.57
	highest	4.22	1.58
TT	lowest	3.46	1.72
	medium	3.13	1.73
	highest	2.95	1.36
LAM	lowest	74.42	25.60
	medium	67.39	29.95
	highest	67.55	29.52
Avg_{ontime}	lowest	0.33	0.29
	medium	0.37	0.27
	highest	0.41	0.26

Table 4.5: values of RQA variables of On-Time rate per week



variable	PC1	PC2
λ	1.862	0.790
RR	-0.467	0.336
DET	-0.414	-0.147
D_{avg}	0.387	0.831
TT	-0.489	0.417
LAM	-0.472	0.047

Figure 4.10: PCA plot of On-Time Rate per week and corresponding Principal Components

Additionally, a positive relationship exists between the means of the average On-Time rate per student and performance.

Based on the PCA plot, the behavior of the clusters regarding the RQA variables of the percentage of exercises per week that were completed on time, seems to differ slightly, as can be seen in Figure 4.10. Particularly, there are quite some students in the middle-performing cluster towards the right side of the plot. These are students with lower values for the RQA variables, as can be deduced from PC1. However, this is not confirmed in the summarizing statistics shown in Table 4.5. Again from PC2 the value for D_{avg} stands out, indicating the high variability in the average diagonal line length.

The values obtained from the PERMANOVA test indicate that there are no statistically significant differences between the clusters regarding On-Time Rate per week with an F-statistic of 0.795 and a p-value of 0.516.

4.2 Differences between study sessions and weeks

Multiple things stand out when comparing the analyses made for study sessions and weeks. Firstly, it seems counterintuitive that the differences between the clusters are that much greater when study sessions are used as a unit of analysis instead of weeks. This might be explained by the fact that weekly-based analysis summarizes the SE behavior in one larger amount of time. Therefore, the complex dynamics of SE are captured in one value per week, which does not accurately describe the behavior of a student. It can thus occur that students show completely different engagement every day, but that this is not captured in the corresponding week.

Secondly, in PC1 for all of the different PCA plots all the RQA variables show correlation with each other. The only variable that stands out is D_{avg} and only in the cases where weeks is used as a unit of analysis. This is against what is expected, since when for example the values RR and DET increase, there are more recurrence points and more of them occur in diagonal structures. Logically, this would cause the average lengths of the diagonal lines in the RP to increase. For weeks as a measure of time, this relationship is not found and it shows that perhaps weeks is not the best measure to use. The negative correlation is somewhat explained by the short length of the time series, as when the RP has such small dimensions there are not many possibilities for longer diagonal lines to occur.

Chapter 5

Conclusions and Future Research

5.1 Conclusions

This thesis explored the relationship between Student Engagement (SE) and student performance among 144 first-year mechanical and civil engineering students at TU Delft, utilizing Recurrence Quantification Analysis (RQA) to analyze recurrence in SE indicators. The primary objective was to determine if groups of differently performing students showed significantly different behavior in the recurrence of SE.

K-means clustering was used to identify 3 groups of students that had different results on their final exam. SE was examined through three indicators, Time Spent, Attempts Made and On-Time Rate. Both study sessions and weeks were used as a unit of analysis, since beforehand both were seen as possibly valuable. Recurrence Quantification Analysis (RQA) was employed to quantify and visualize the dynamic characteristics of these indicators through the use of five RQA variables: RR, DET, D_{avg}, TT and LAM. Principal Component Analysis (PCA) was used to visualize the values of the RQA variables for the different clusters. Permutational Multivariate Analysis of Variance (PERMANOVA) was used to quantify the differences between the clusters.

Summarizing statistics

Regarding the summarizing statistics, some patterns did indeed emerge. For the two SE indicators Time Spent per study session and Attempts made per study session, a negative correlation existed between performance and the mean values for the RQA variables. In other words, the high-performance cluster reported the lowest values for the quantified recurrence of these SE indicators. These results are in contrast with the findings of Franciškovič et al. (2024) [4] and Poquet et al. (2023) [25], discussed in the hypothesis (Chapter 1.6). For the other three SE indicators: Time Spent per week, Attempts Made per week and On-Time Rate per week, no clear pattern emerged from the means of the RQA variables. The total Time Spent, total Attempts Made and average On-Time rate were also reported for every cluster and these all exhibited a positive correlation with performance. Combining these results, the conclusion can be drawn that this study found that high-performing students engage more with the course material overall, but do this in study sessions that vary more in their duration and total exercises attempted.

PERMANOVA analysis

The overall differences between the means of the clusters were further evaluated in Chapter 4, using PERMANOVA. For all the SE indicators using weeks as a unit of analysis no significant differences were found between the clusters. Analyzing Attempts Made per study session resulted in an F-statistic of 1.600 and p-value of 0.175. Although indicating a greater difference for these RQA variables between the clusters than the RQA variables for the SE indicators based on weeks, this result was not statistically significant. The F-statistic (3.511) and p-value (0.012) found for Time Spent per study session, were statistically significant. After further analysis using only 2 clusters at a time, it was found that the means of the low- and high-performers were significantly different ($p = 0.002$) and the means of the middle- and high-performers as well ($p = 0.024$). Therefore, it was concluded that the recurrence shown in behavior concerning the Time Spent per study session for high-performers was significantly lower than that of middle- and low-performers. This finding is again in contrast with Franciŝković et al. (2024) [4]. In the context of study behavior, this would indicate that high-performers spend more time in total studying, but do this in study sessions that vary more in duration than middle- and low-performers.

Unit of analysis

Furthermore, the question was raised regarding which unit of analysis would be most effective in this research. weeks as a measure of time measured inactivity better and study sessions had less of an impact of measurements of 0 minutes. Also, using weeks as a measure of time always produced time series of length 8 per student, while the amount of study sessions per student varied. After having completed the analyses, most of the significant results were found using study sessions as a unit of analysis. There were two problems occurring with weeks as a measure of time. First of all, SE is hard to capture in an average per week, since the way a student is engaged with the material can change every second. Using weeks to measure SE causes a loss of information about the complex dynamics of SE. Furthermore, using weeks causes the time series to always have length 8. Even though this length is then constant across all students, it is quite short for applying RQA, because the RQA variables can be heavily influenced by small changes. This was especially evidenced by the negative correlation shown between D_{avg} and the other RQA variables when weeks were used as a measure of time. In summary, it is advised to use the measure of study sessions as it measures the complex dynamics of SE more closely. Especially, in the cases where not that many weeks can be analyzed, the use of study sessions is recommended.

Results in practice

In this research, much attention was spent on the definition of RQA variables and RQA-centered analysis. However, the most important goal is to draw conclusions about how students should actively engage with their study materials to achieve better results. Therefore, the theoretical conclusions drawn will be related to the behavior in SE. First of all, the study showed that higher-performing students made a larger amount of exercises, spent more time on the course and completed more exercises before the deadline. In addition to this, higher-performing students showed less recurrence in the Attempts Made and Time Spent per study session, however, this result was only statistically significant for Time Spent. An explanation for these observed results could be that higher-performing students adjust their approach according to the difficulty of the material. This would mean that they use more time and do more exercises for topics that are difficult to grasp or are important for the final exam, while lower-performing students are evenly engaged throughout all the different topics of the course. Whether this is indeed the case requires further investigation and is a topic for possible future research. discussed in Chapter 5.2.

5.2 Future Research

This research was conducted to add to the existing literature on the relationship between student performance and recurrence in Student Engagement. As the findings were mostly dissimilar to papers using a likewise approach, it would be prudent to conduct more similar experiments with other datasets to be able to draw a decisive conclusion about the relationship between recurrence shown in SE and student performance.

Explanation of the Results

The results of this research found that high-performing students overall engage more with the course material but do this in study sessions that vary more in their duration and total exercises attempted. It would be interesting to see when high-performers have longer and shorter study sessions. This could lead to more concrete advice regarding effective study behavior. If it is for example shown that on similar topics the high-performers have very few and short study sessions, this would be an indication of that this topic is not essential for achieving a high grade.

Variable-centered Analysis

In this thesis, a person-centered analysis was conducted to determine if there existed differences in the recurrence of SE between clusters of students based on their grades. This was done to examine if there existed similarities on a group scale between students performing similarly. However, to analyze the direct relationship between SE and performance variable-centered analysis could also be applied. Due to lack of time, this analysis was not carried out in this thesis, but a methodology was thought out. Similar to the research of Francišković et al. (2024) [4] Machine Learning algorithms could be applied to all of the RQA variables found for every student. After the performance of the models is evaluated, feature importance can be applied, to measure what RQA variables had the biggest influence on predicting student performance. Using this method, not the differences between groups of students would be analyzed, but rather the overall relationship between recurrence in SE and performance.

Higher-Dimensional RQA

For this research, the conscious choice was made to perform 1-dimensional RQA to make sure that any results would be interpretable. However, for this data it is also possible to use multi-dimensional RQA (MdRQA). The results of combining multiple SE indicators when investigating recurrence could give a more complete picture of SE. Despite this benefit, other important choices need to be considered in this case. For every SE indicator a distance metric needs to be selected, the impact of each indicator needs to be determined and the total margin needs to be chosen. Therefore, it would be interesting to be able to determine what influence every SE indicator has on engagement as a whole. Nevertheless, these concepts are quite abstract and thus an exact method would be hard to define.

Chapter 6

Discussion

In this thesis multiple assumptions have been made and therefore, the results cannot be directly interpreted without stating the shortcomings of this research.

6.1 Homogeneous research group

Firstly, the students analyzed do not represent the whole population. They were mostly male, attend higher education and a significant number were in the same age group. There was no information available about their socio-economic background, so no conclusions can be drawn regarding whether the group was representative for the population in this apect. In general, these results can thus not be extrapolated to the whole population. Furthermore, only students of the TU Delft were analyzed over one specific course. Thus the results can also not be directly extrapolated to the higher education system as a whole.

6.2 Definitions of Student Engagement (Indicators)

Secondly, it should be noted that SE is both a very broad and complex concept. Therefore, the variables used in this thesis to indicate a student's engagement do not encompass the full scope of SE. Specifically, the factors used in this thesis mostly provide information about the duration and occurrence of SE, and do not take the "effectiveness" of this engagement into account. Consequently, the results obtained in this thesis about SE should always be interpreted with caution. Research incorporating more factors that are linked to SE can give more decisive results regarding SE.

6.3 Value of margins

In many other studies where RQA was used to analyze the relationship between performance and recurrence of SE, categorical time series were used, such as Li et al. (2022) [32] and Poquet et al. (2023) [25]. These time series described the types of action a student undertook. When applying RQA to a categorical time series no margin needs to be defined as the same action is classified as recurrent and different actions are defined as non-recurrent. In the case of this thesis continuous time series were used to apply a different methodology than the existing literature. This introduces the need for a margin. Even though there exists literature on guidelines for the margin, such as Marwan and Weber (2015) [24], no exact "perfect" value can be defined. Therefore, the margins in this thesis were chosen based on the diameter of the data, the distribution of the data and the value of RR produced. This choice was made subjectively and thus many other margins could also have been used to assess the data. To gain more certainty about the correlations found, the results derived from different margins should be analyzed.

6.4 Definition of study session

To obtain values for the SE indicators for a study session of a student a time cap had to be chosen as a maximum time of inactivity to still be considered as one study session. As discussed in Chapter 3.2, many different time caps have been used in the existing literature. Therefore, the choice for this time cap was again subjective and based on the distribution of the differences between successive study actions. This choice majorly influences the amount of study sessions and the value obtained for the SE indicators. Therefore, also for this assumption, different values can be tested to gain more certainty about the results.

6.5 Other study activity

Lastly, only data from Grasple was obtained to analyze SE. However, students always have the option to undertake additional work beyond this online program. Consequently, the indicators of SE that have been used in this research can never fully gauge the engagement of a student over the whole course. To better capture the full behavior of a student a greater analysis needs to be made as to the other materials that a student accessed and in what way this was done.

Appendix A

A.1 Jensen's Inequality

Theorem 3 (Jensen's Inequality). *Let f be a concave function on $[a,b]$ and $y_1,\ldots,y_n \in [a,b]$, then*

$$f(y_1)+\ldots+f(y_n) \leq n\cdot f(\frac{y_1+\ldots+y_n}{n})$$

Proof. Proof by induction.
Base step $n=2$: By the definition of a concave function, for any $\lambda \in [0,1]$ the following inequality holds:

$$(1-\lambda)f(y_1)+\lambda f(y_2) \leq f((1-\lambda)y_1+\lambda y_2)$$

Substituting $\lambda = \frac{1}{2}$, gives the required inequality:

$$\frac{1}{2}f(y_1)+\frac{1}{2}f(y_2) \leq f(\frac{1}{2}y_1+\frac{1}{2}y_2) \tag{A.1}$$

$$f(y_1)+f(y_2) \leq 2f(\frac{1}{2}y_1+\frac{1}{2}y_2) \tag{A.2}$$

Induction step: Suppose that for $n=k$ the following inequality holds:

$$f(y_1)+\ldots f(y_k) \leq k\cdot f(\frac{y_1+\ldots+y_k}{k}) \tag{A.3}$$

We will show that the inequality then also holds for $n=k+1$, let $y_1,\ldots,y_k,y_{k+1}$, be the considered points. Then by the definition of a concave function, the following inequalities hold.

$$\frac{k}{k+1}f\left(\frac{y_1+y_2+\cdots+y_k}{k}\right)+\frac{1}{k+1}f(y_{k+1}) \leq f\left(\frac{k\cdot\frac{y_1+y_2+\cdots+y_k}{k}+y_{k+1}}{k+1}\right) \tag{A.4}$$

$$k\cdot f\left(\frac{y_1+y_2+\cdots+y_k}{k}\right)+f(y_{k+1}) \leq (k{+}1)f\left(\frac{k\cdot\frac{y_1+y_2+\cdots+y_k}{k}+y_{k+1}}{k+1}\right) \tag{A.5}$$

$$f(y_1)+\ldots+f(y_k)+f(y_{k+1}) \leq (k{+}1)f\left(\frac{y_1+\ldots+y_k+y_{k+1}}{k+1}\right) \tag{A.6}$$

Where A.6 holds by the Induction Hypothesis A.3, this concludes the induction step and the proof.

□

A.2 Proof uniform distribution gives maximum value for Shannon's Entropy

Theorem 4. *The maximum value of Shannon's Entropy of the diagonal lines is obtained if and only if the distribution of the lengths of the diagonal lines is uniform. That is, if there are n* *different lengths occurring in the RP, for all those lengths l, we have $p(l) = \frac{1}{n}$. The maximum value of the Entropy then is* $\ln(n)$.

Proof. We know Shannon's Entropy is given by $-\sum_{l=d_{min}}^{d_{max}} p(l) \ln p(l)$, with the extra condition that $\sum_{l=d_{min}}^{d_{max}} p(l) = 1$. We only consider the lengths l for which $p(l) > 0$, as other terms do not contribute to the sum. For $p(l) > 0$ the function for Shannon's Entropy is concave and therefore Jensen's inequality for concave functions can be applied in A.8. In this case $f(x) = -x\ln(x)$. This results in the following inequality:

$$\begin{aligned} ENTR &= \sum_{l=d_{min}}^{d_{max}} -p(l)\ln p(l) && \text{(A.7)}\\ &\leq n \cdot -\frac{\sum_{l=d_{min}}^{d_{max}} p(l)}{n} \cdot \ln\left(\frac{\sum_{l=d_{min}}^{d_{max}} p(l)}{n}\right) && \text{(A.8)}\\ &= n \cdot -\frac{1}{n} \cdot \ln\left(\frac{1}{n}\right) && \text{(A.9)}\\ &= -\ln\left(\frac{1}{n}\right) && \text{(A.10)}\\ &= \ln(n) && \text{(A.11)} \end{aligned}$$

In Step A.8 Jensen's inequality was applied. This shows that the maximum bound for Shannon's Entropy is $\ln(n)$, where n is the number of different lengths of diagonal lines occurring in the RP. Next, it will be shown that the maximum only occurs when the distribution of the lengths of the diagonal lines is uniform.

If the distribution is uniform, it is trivial that the maximum is obtained. Suppose that the distribution is not uniform, that is $\exists p(l_1), p(l_2) \ni p(l_1) \neq p(l_2)$ and for all lengths not equal to l_1 or l_2, $p(l_i) = \frac{1}{n}$, where $n \geq 2$. Suppose without loss of generality that $p(l_1) > p(l_2)$. Since one must be larger than the average and one must be greater, $p(l_1)$ and $p(l_2)$ can be written as follows for some $0 < \epsilon < \frac{1}{n}$:

$$p(l_1) = \frac{1}{n} + \epsilon$$
$$p(l_2) = \frac{1}{n} - \epsilon$$

Then,

$$\begin{aligned} ENTR &= -\left(\underbrace{\frac{1}{n}\ln\frac{1}{n} + \ldots + \frac{1}{n}\ln\frac{1}{n}}_{n-2\text{ times}} + (\frac{1}{n}+\epsilon)\ln\left(\frac{1}{n}+\epsilon\right) + (\frac{1}{n}-\epsilon)\ln\left(\frac{1}{n}-\epsilon\right)\right) && \text{(A.12)}\\ &= -\left(\frac{n-2}{n}\ln\frac{1}{n} + (\frac{1}{n}+\epsilon)\ln\left(\frac{1}{n}+\epsilon\right) + (\frac{1}{n}-\epsilon)\ln\left(\frac{1}{n}-\epsilon\right)\right) && \text{(A.13)} \end{aligned}$$

Next, the derivative is taken with respect to ϵ, to determine for what value of ϵ the maximum is obtained.

$$\begin{aligned} \frac{\mathrm{d}ENTR}{\mathrm{d}\epsilon} &= (\frac{1}{n}+\epsilon)\frac{1}{\frac{1}{n}+\epsilon} + \ln(\frac{1}{n}+\epsilon) - (\frac{1}{n}-\epsilon)\frac{1}{\frac{1}{n}-\epsilon} - \ln(\frac{1}{n}-\epsilon) && \text{(A.14)}\\ &= \ln(\frac{1}{n}+\epsilon) - \ln(\frac{1}{n}-\epsilon) && \text{(A.15)} \end{aligned}$$

Solving for this equation equal to zero, $\epsilon = 0$, is obtained, i.e. $p(l_1) = p(l_2)$, which is a contradiction. It can be shown that this is indeed the maximum, by calculating the second derivative with respect to ϵ.

$$\frac{\mathrm{d}^2 ENTR}{\mathrm{d}\epsilon^2} = \frac{1}{\frac{1}{n} + \epsilon} - \frac{1}{\frac{1}{n} - \epsilon} < 0$$

□

A.3 Proof of Huygen's Theorem

Theorem 5. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be data points with centroid $\mathbf{G}$, then for any distance measure $||\cdot||$, that is induced by an inner product space, $\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{G}||^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{x}_i - \mathbf{x}_j||^2$. In other words, the sum of squared distances from individual points to their centroid is equal to the sum of squared interpoint distances divided by the number of points.*

Proof. Firstly it is important to note that the sum of square interpoint distances divided by the number of points is equal to $\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{x}_i - \mathbf{x}_j||^2$, since every distance is counted twice. Then the following equality is obtained, by the characteristics of an inner product space:

$$
\begin{aligned}
\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{x}_i-\mathbf{x}_j||^2 &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{x}_i||^2+||\mathbf{x}_j||^2-2\langle\mathbf{x}_i,\mathbf{x}_j\rangle \\
&= \frac{1}{2n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{x}_i||^2+\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{x}_j||^2-2\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\right) \\
&= \frac{1}{2n}\left(n\sum_{i=1}^{n}||\mathbf{x}_i||^2+n\sum_{j=1}^{n}||\mathbf{x}_j||^2-2\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\right) \\
&= \frac{1}{2n}\left(2n\sum_{i=1}^{n}||\mathbf{x}_i||^2-2\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\right) \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_j\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+\frac{1}{n}\langle-\sum_{i=1}^{n}\mathbf{x}_i,\sum_{j=1}^{n}\mathbf{x}_j\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+\frac{1}{n}\langle\sum_{j=1}^{n}\mathbf{x}_j-2\sum_{i=1}^{n}\mathbf{x}_i,\sum_{j=1}^{n}\mathbf{x}_j\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+\frac{1}{n}\left(\langle\sum_{j=1}^{n}\mathbf{x}_j,\sum_{j=1}^{n}\mathbf{x}_j\rangle-2\langle\sum_{j=1}^{n}\mathbf{x}_j,\sum_{i=1}^{n}\mathbf{x}_i\rangle\right) \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+n\langle\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j,\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j\rangle-2\langle\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i,\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+n\langle\mathbf{G},\mathbf{G}\rangle-2\langle\sum_{i=1}^{n}\mathbf{x}_i,\mathbf{G}\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i||^2+\sum_{i=1}^{n}||\mathbf{G}||^2-2\sum_{i=1}^{n}\langle\mathbf{x}_i,\mathbf{G}\rangle \\
&= \sum_{i=1}^{n}||\mathbf{x}_i-\mathbf{G}||^2
\end{aligned}
$$

□

Appendix B

B.1 Code of Weeks

This code has both the calculations for the Time Spent and the Attempts made. Note that to change the variable analysed the only change that needs to be made is the time series variable in the 'crqa' function. The code for Study Sessions is not included separately as the main calculations are also used for weeks.

```
    library("nonlinearTseries")
library("ggplot2")
library("tseriesChaos")
library("crqa")
library("lubridate")
library("tidyverse")
library("factoextra")
library("MASS")
library("NbClust")
library("readxl")
library("statip")


#importing data
data_civil <- read_excel("path_to_df")
data_mechanical <- read_excel("path_to_df",
                              col_types = c("text", "text", "numeric",
                                            "numeric", "text", "numeric", "numeric",
                                            "text", "text", "numeric", "numeric",
                                            "numeric", "numeric", "numeric",
                                            "text", "numeric", "text", "numeric",
                                            "numeric", "numeric", "numeric",
                                            "numeric", "text"))


#combining datasets
total_df = rbind(data_civil, data_mechanical)
df = total_df


#deleting entries where no exercise was answered
df = df[!is.na(df$exercise_answered_at),]
```

```
#formatting dataset
colnames(df)[2] = "extra"
colnames(df)[1] = "student_id"

#reading grades
final_grades_age_gender = read_excel("path_to_df")

#deleting students that were not taken into account
final_grades_age_gender = final_grades_age_gender[
                          final_grades_age_gender$course_grade != "NV",]
final_grades_age_gender = final_grades_age_gender[
                          !is.na(final_grades_age_gender$course_grade),]
final_grades_age_gender = final_grades_age_gender[
                          final_grades_age_gender$course != "cse", ]

#making sure only students that have a grade
df <- df %>%
  semi_join(final_grades_age_gender, by = "student_id")

#determining students that do have a grade, but are not in grasple df
students_not_in_df <- final_grades_age_gender %>%
  anti_join(df, by = "student_id")

#we do not want to take these students into account so deleting from grades df
df_grades <- final_grades_age_gender %>%
  anti_join(students_not_in_df, by = "student_id")

#formatting times in correct way
df$exercise_answered_at  = ymd_hms(df$exercise_answered_at)


#variables where everybody has same begin and end date (i.e. whole period)
df = df[order(df$exercise_answered_at), ]
base_day = date(df$exercise_answered_at[1])
end_day = date(tail(df, n=1)$exercise_answered_at)

#ordering on students
df = df[order(df$student_id),]


#needed to loop through the students
#students = unique(df_grades_delete_low_sessions[["student_id"]])
students = unique(RQA_info_all_student$student_id)


#opening pdf file where plots per student will be saved
#pdf(paste('RPweeks_time_high_eps=90.pdf', sep = ''),width = 14)

#store RQA variables per student to cluster later
RQA_info_all_student <- data.frame(RR = numeric(),
                                   DET = numeric(),
                                   maxL = numeric(),
                                   L = numeric(),
                                   ENTR = numeric(),
                                   LAM = numeric(),
                                   TT = numeric(),
                                   student_id = character(),
                                   stringsAsFactors = FALSE)

weeks_all_students <- data.frame(time_spent = numeric(),
                                stringsAsFactors = FALSE)

attempts_all_students <- data.frame(attempts_made = numeric(),
                                    stringsAsFactors = FALSE)

for (student in students){

  #get data set per student
  df_per_student = df[df$student_id %in% c(student), ]

  #sorting actions based on time answer given
  df_per_student = df_per_student[order(df_per_student$exercise_answered_at), ]

  #starting variables per student
  base_time = df_per_student$exercise_answered_at[1]
  prev_time = df_per_student$exercise_answered_at[1]

  #creating the dates to loop through
  dates <- seq(as.Date(base_day), as.Date(end_day), by=1)

  #empty df to store all the times spent
  time_per_day <- data.frame(time_spent = numeric(),
                             stringsAsFactors = FALSE)

  time_per_week <- data.frame(time_spent = numeric(),
                              stringsAsFactors = FALSE)

  attempts_per_day <- data.frame(attempts_made = numeric(),
                             stringsAsFactors = FALSE)

  attempts_per_week <- data.frame(attempts_made = numeric(),
                             stringsAsFactors = FALSE)

  total_time_spent_per_student = 0
  total_attempts_per_student = 0
  #iterate through all times per student
  for (date in dates){

    #creating dataframe per date and ordering
    df_per_date = df_per_student[date(df_per_student$exercise_answered_at)
```

```
%in% c(as.Date(date)), ]
df_per_date = df_per_date[order(df_per_date$exercise_answered_at), ]

#starting times
base_time = df_per_date$exercise_answered_at[1]
prev_time = df_per_date$exercise_answered_at[1]

day_time = 0

#attempts per day
attempts = nrow(df_per_date)
attempts_per_day = attempts_per_day
%>% add_row(attempts_made = as.numeric(attempts))

#looping through times
for (times in df_per_date$exercise_answered_at){
  #calculate difference between two times (if > 45 session done)
  difference = difftime(times, prev_time)
  difference = as.numeric(difference, units = 'mins')


  if (difference > 15){

    #calculate time work in this study session
    time_worked = difftime(prev_time, base_time, units = 'mins')

    prev_time = times

    #add to time worked on this day
    day_time = day_time + time_worked
    base_time = times

  }
  else{
    #difference not great enough so still in same session and we update time
    prev_time = times
  }
}

#adding time of final session
time_worked = difftime(times, base_time, units = 'mins')
day_time = day_time + time_worked

#adding total amount of time of the day
if (length(day_time) == 0){
  time_per_day <- time_per_day %>% add_row(time_spent = 0)
}
else{
  time_per_day <- time_per_day %>% add_row(time_spent = as.numeric(day_time))
  total_time_spent_per_student = total_time_spent_per_student +
    as.numeric(time_worked)

  }
}

#combining each seven days to get the data per week
for (i in seq(1,8, by = 1)){
  time_per_week <- time_per_week %>% add_row(time_spent =
  sum(as.numeric(time_per_day$time_spent[
  as.numeric(7 * i -7): as.numeric(7 * i)])))
  weeks_all_students = weeks_all_students %>% add_row(time_spent =
  sum(as.numeric(time_per_day$time_spent[
  as.numeric(7 * i -7): as.numeric(7 * i)])))
  attempts_per_week = attempts_per_week %>% add_row(attempts_made =
  sum(as.numeric(attempts_per_day$attempts_made[
  as.numeric(7 * i -7): as.numeric(7 * i)])))
  attempts_all_students = attempts_all_students %>% add_row(attempts_made =
  sum(as.numeric(attempts_per_day$attempts_made[
  as.numeric(7 * i -7): as.numeric(7 * i)])))
}

rqa.analysis = crqa(ts1 = time_per_week,
ts2 = time_per_week,
method = 'crqa',
radius = 70,
embed = 1,
delay = 1)

#storing the information of RP per student
RQA_info_all_student <- RQA_info_all_student %>%
add_row(RR = rqa.analysis$RR,
        DET = rqa.analysis$DET,
        maxL = rqa.analysis$maxL,
        L = rqa.analysis$L,
        ENTR = rqa.analysis$ENTR,
        LAM = rqa.analysis$LAM,
        TT = rqa.analysis$TT,
        student_id = student)

#starting plot of the RP's

# find the size of the RP
xdim   = nrow(rqa.analysis$RP)
ydim   = ncol(rqa.analysis$RP)

# transform the RP into a square matrix
rqa.analysis$RP = matrix(as.numeric(rqa.analysis$RP), nrow = xdim, ncol = ydim)

# figure out where the recurrent points are
ind = which(rqa.analysis$RP == 1, arr.ind = T)
```

```
  # create all time series of all possible samples on axis
  tstamp = seq(1, xdim, 1)

  # create the shell of the plot
  par(mar = c(3.8, 3.8, 2,2), font.axis = 2, cex.axis = 1,
      font.lab = 2, cex.lab = 1.2)

  plot(tstamp, tstamp, type = "n",
       xlab = "", ylab = "", main=paste("student ", student),
       xaxt = "n", yaxt = "n",
       font = 2, cex.axis = 1)

  # add recurrent points to the plot
  points(ind[,1], ind[,2], cex = 0.9, pch = 16)

  # Add ticks on the x and y-axis
  axis(1, at = seq(1, xdim, length.out = length(tstamp)), cex.axis = 0.8)
  axis(2, at = seq(1, ydim, length.out = length(tstamp)), cex.axis = 0.8)

  # add x- and y-axis labels to the plot (so they’re not too far away)
  mtext("weeks", at = mean(tstamp), side = 1, line = 2, cex = 1.2, font = 2)
  mtext("weeks", at = mean(tstamp), side = 2, line = 2, cex = 1.2, font = 2)
}
graphics.off()
rownames(RQA_info_all_student) = students

#if students with NaN for TT, this means that they had no vertical lines at all,
#so therefore value 0
RQA_info_all_student <- RQA_info_all_student %>%
  mutate(TT = ifelse(is.nan(TT), 0, TT))

students_in_both = semi_join(df_grades, RQA_info_all_student, by = "student_id")
excel_RQA = left_join(students_in_both, RQA_info_all_student, by = "student_id")
#write_xlsx(excel_RQA, "RQAinfo_weeks_attempts_eps=70.xlsx")


RQA_info_all_student_low <- RQA_info_all_student %>%
  semi_join(cluster_low, by = "student_id")

RQA_info_all_student_middle <- RQA_info_all_student %>%
  semi_join(cluster_middle, by = "student_id")

RQA_info_all_student_high <- RQA_info_all_student %>%
  semi_join(cluster_high, by = "student_id")


#adding the name of the cluster each student belongs to
excel_RQA$cluster <-
ifelse(excel_RQA$student_id %in% cluster_low$student_id, "low",
ifelse(excel_RQA$student_id %in% cluster_middle$student_id, "middle",
ifelse(excel_RQA$student_id %in% cluster_high$student_id, "high", NA)))

####Applying Principal Component Analysis ####
preprocess_params <- preProcess(excel_RQA[, c('RR', 'L', 'DET',
'ENTR', 'LAM', 'TT')],
method = c("center", "scale"))
data_standardized <- predict(preprocess_params, excel_RQA)

# Apply PCA
pca_weeks_time <- prcomp(data_standardized[, c('RR', 'L', 'DET',
'ENTR', 'LAM', 'TT')])

# Create a data frame with the PCA results
pca_df_weeks_time <- data.frame(PC1 = pca_weeks_time$x[,1],
                                PC2 = pca_weeks_time$x[,2],
                                Cluster = excel_RQA$cluster)

# Plot the PCA results
ggplot(pca_df_weeks_time, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "RQA variables per cluster (Time Spent per Week)",
  x = "Principal Component 1", y = "Principal Component 2") +
  theme_minimal() +
  stat_ellipse(aes(fill = Cluster), alpha = 0.2, geom = "polygon")

####Applying PERMANOVA ####
variables = excel_RQA[, c('RR', 'L', 'DET', 'ENTR', 'LAM', 'TT')]
variables = scale(variables)

dist_matrix <- dist(variables, method = "euclidean")

adonis_result_weeks_time <- adonis2(dist_matrix ~ excel_RQA$cluster,
                                    data = excel_RQA,
                                     permutations = 1e4)
```

```
####distributions of all study_sessions (plots)####
#distribution attempts per week
attempts_all_students$rounded = round(attempts_all_students$attempts_made/50)*50
table_attempts = table(attempts_all_students$rounded)
barplot(table_attempts,
        main = 'Distribution of attempts made per week',
        xlab = 'Attempts (rounded to multiple of 50)',
        ylab = 'Frequency')


#round to nearest multiple of 10
weeks_all_students$time_spent_rounded = round(weeks_all_students$time_spent/10)*10
time_freq_table <- table(weeks_all_students$time_spent_rounded)

barplot(time_freq_table,
        main = 'Duration per week',
        xlab = 'Duration (rounde to nearest multiple of 10)',
        ylab = 'Frequency',
        ylim = c(0,50))
```

B.2 Code of On-Time Rate

```
library("nonlinearTseries")
library("ggplot2")
library("tseriesChaos")
library("crqa")
library("lubridate")
library("tidyverse")
library("factoextra")
library("MASS")
library("NbClust")
library("readxl")
library("statip")

#### clustering ####
#reading grades
final_grades_age_gender = read_excel("path_to_dataframe")

#deleting students that were not taken into account
final_grades_age_gender = final_grades_age_gender[
                          final_grades_age_gender$course_grade != "NV",]
final_grades_age_gender = final_grades_age_gender[
                          !is.na(final_grades_age_gender$course_grade),]
final_grades_age_gender = final_grades_age_gender[
                          final_grades_age_gender$course != "cse", ]

#making sure only students in df that have a grade
df <- df %>%
  semi_join(final_grades_age_gender, by = "student_id")
#determining students that do have a grade, but are not in grasple df
students_not_in_df <- final_grades_age_gender %>%
  anti_join(df, by = "student_id")

#we do not want to take these students into account so deleting from grades df
df_grades <- final_grades_age_gender %>%
  anti_join(students_not_in_df, by = "student_id")
```

```
#kmeans clustering
grades_cluster_3 = kmeans(df_grades$course_grade, centers = 3, nstart = 25)

#dataframes per cluster
cluster_middle = df_grades[grades_cluster_3$cluster == 3, ]
cluster_high = df_grades[grades_cluster_3$cluster == 1, ]
cluster_low = df_grades[grades_cluster_3$cluster == 2, ]


#### importing data ####
data_civil <- read_excel("path_to_dataframe")
data_mechanical <- read_excel("path_to_dataframe",
                              col_types = c("text", "text", "numeric",
                                            "numeric", "text", "numeric", "numeric",
                                            "text", "text", "numeric", "numeric",
                                            "numeric", "numeric", "numeric",
                                            "text", "numeric", "text", "numeric",
                                            "numeric", "numeric", "numeric",
                                            "numeric", "text"))


#combine civil en mechanical students
df = rbind(data_civil, data_mechanical)
df = df[!is.na(df$exercise_answered_at),]


#making sure right column has the name student_id
colnames(df)[2] = "extra"
colnames(df)[1] = "student_id"


#### new ####

#importing the deadlines of every exercise
df_deadlines <- read_excel("path_to_df")
df_deadlines$deadline <- as.Date(df_deadlines$deadline, format="%Y-%m-%d")
df$exercise_answered_at = as.POSIXct(df$exercise_answered_at,
                                     format = "%Y-%m-%d %H:%M:%S")

#the selection of these students is dependent on the RQA dataframe of another script
#as with studysessions script the students were determined with at least 3 sessions
students = unique(RQA_info_all_student[["student_id"]])
students = sort(students)

#counting the exercises that were assigned per week
exercises_per_week <- data.frame(week_deadline = 1:7) %>%
  left_join(df_deadlines %>% count(week_deadline), by = "week_deadline") %>%
  replace_na(list(n = 0))
```

```
df = df %>%
  left_join(df_deadlines, by = "subject_id")

#df to store all RQA info of students
RQA_info_all_student_comp <- data.frame(RR = numeric(),
                                      DET = numeric(),
                                      maxL = numeric(),
                                      L = numeric(),
                                      ENTR = numeric(),
                                      LAM = numeric(),
                                      TT = numeric(),
                                      student_id = character(),
                                      avg_comp = numeric(),
                                      stringsAsFactors = FALSE)


#pdf(paste('RP_high_comprate_eps=hundreth.pdf', sep = ''),width = 14)
for (student in students){

  #considering only data of one student
  df_per_student = df[df$student_id %in% c(student), ]


  #df with only first attempts at every subject
  df_student_first <- df_per_student %>%
    group_by(subject_id) %>%
    filter(exercise_answered_at == min(exercise_answered_at))

  #adding a column showing for if exercises were attempted on time
  df_student_first <- df_student_first %>%
    mutate(on_time = ifelse(exercise_answered_at < deadline, 1, 0))


  #counting how many exercises per week were on time
  df_weekly_counts <- df_student_first %>%
    group_by(week_deadline) %>%
    summarize(on_time_count = sum(on_time == 1))


  #adding 0 to the weeks where no exercises were attempted
  complete_weeks <- data.frame(week_deadline = 1:7)
  df_complete <- complete_weeks %>%
    left_join(df_weekly_counts, by = "week_deadline") %>%
    replace_na(list(on_time_count = 0))

  #dividing the attempted exercises by the total amount of exercises
  df_complete <- df_complete %>%
    left_join(exercises_per_week, by = "week_deadline") %>%
    mutate(compliance_rate = on_time_count / n)
```

```
#performing RQA analysis
rqa.analysis = crqa(ts1 = df_complete$compliance_rate,
                    ts2 = df_complete$compliance_rate,
                    method = 'crqa', radius = 0.1,
                    embed = 1,
                    delay = 1)

#storing the information of RP per student
RQA_info_all_student_comp =
RQA_info_all_student_comp %>% add_row(RR = rqa.analysis$RR,
                                      DET = rqa.analysis$DET,
                                      maxL = rqa.analysis$maxL,
                                      L = rqa.analysis$L,
                                      ENTR = rqa.analysis$ENTR,
                                      LAM = rqa.analysis$LAM,
                                      TT = rqa.analysis$TT,
                                      student_id = student,
                                      avg_comp = mean(df_complete$compliance_rate))

#starting plot of the RP's

# find the size of the RP
xdim   = nrow(rqa.analysis$RP)
ydim   = ncol(rqa.analysis$RP)

# transform the RP into a square matrix
rqa.analysis$RP = matrix(as.numeric(rqa.analysis$RP), nrow = xdim, ncol = ydim)

# figure out where the recurrent points are
ind = which(rqa.analysis$RP == 1, arr.ind = T)

# create all time series of all possible samples on axis
tstamp = seq(1, xdim, 1)

# create the shell of the plot
par(mar = c(3.8, 3.8, 2,2), font.axis = 2, cex.axis = 1,
    font.lab = 2, cex.lab = 1.2)

plot(tstamp, tstamp, type = "n",
     xlab = "", ylab = "", main=paste("student ", student),
     xaxt = "n", yaxt = "n",
     font = 2, cex.axis = 1)

# add recurrent points to the plot
points(ind[,1], ind[,2], cex = 0.9, pch = 1)

# Add ticks on the x and y-axis
axis(1, at = seq(1, xdim, length.out = length(tstamp)), cex.axis = 0.8)
axis(2, at = seq(1, ydim, length.out = length(tstamp)), cex.axis = 0.8)
```

```
  # add x- and y-axis labels to the plot (so they’re not too far away)
  mtext("weeks", at = mean(tstamp), side = 1, line = 2, cex = 1.2, font = 2)
  mtext("weeks", at = mean(tstamp), side = 2, line = 2, cex = 1.2, font = 2)
}
graphics.off()
rownames(RQA_info_all_student_comp) = students


RQA_info_all_student_comp <- RQA_info_all_student_comp %>%
  mutate(TT = ifelse(is.nan(TT), 0, TT))

#splitting the results into the three clusters
RQA_info_all_student_low <- RQA_info_all_student_comp %>%
  semi_join(cluster_low, by = "student_id")

RQA_info_all_student_middle <- RQA_info_all_student_comp %>%
  semi_join(cluster_middle, by = "student_id")

RQA_info_all_student_high <- RQA_info_all_student_comp %>%
  semi_join(cluster_high, by = "student_id")

students_in_both = semi_join(df_grades, RQA_info_all_student_comp, by = "student_id")
excel_RQA = left_join(students_in_both, RQA_info_all_student_comp, by = "student_id")
%write_xlsx(excel_RQA, "RQAinfo_Weeks_complance_eps=hundreth.xlsx")


#adding the name of the cluster each student belongs to
excel_RQA$cluster <- ifelse(excel_RQA$student_id %in% cluster_low$student_id, "low",
                ifelse(excel_RQA$student_id %in% cluster_middle$student_id, "middle",
                ifelse(excel_RQA$student_id %in% cluster_high$student_id, "high", NA)))


####Applying PCA####
preprocess_params <- preProcess(excel_RQA[, c(’RR’, ’L’, ’DET’,
                                ’ENTR’, ’LAM’, ’TT’)], method = c("center", "scale"))
data_standardized <- predict(preprocess_params, excel_RQA)

# Apply PCA
pca_weeks_ontime <- prcomp(data_standardized[, c(’RR’, ’L’, ’DET’,
                                              ’ENTR’, ’LAM’, ’TT’)])

# Create a data frame with the PCA results
pca_df_weeks_ontime <- data.frame(PC1 = pca_weeks_ontime$x[,1],
                                  PC2 = pca_weeks_ontime$x[,2],
                                  Cluster = excel_RQA$cluster)

# Plot the PCA results
ggplot(pca_df_weeks_ontime, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "RQA variables per cluster (On-Time Rate per Week)",
  x = "Principal Component 1", y = "Principal Component 2") +
  theme_minimal() +
  stat_ellipse(aes(fill = Cluster), alpha = 0.2, geom = "polygon")

####Applying PERMANOVA####
#using the adonis function to determine if the RQA variables for the cluster are different
variables = excel_RQA[, c(’RR’, ’L’, ’DET’, ’ENTR’, ’LAM’, ’TT’)]
variables = scale(variables)

dist_matrix <- dist(variables, method = "euclidean")

adonis_result_weeks_ontime <- adonis2(dist_matrix ˜ excel_RQA$cluster,
                                      data = excel_RQA, permutations = 1e4)
```

Bibliography

[1] Alan Agresti. “Categorical Data Analysis”. In: Wiley, 2013. Chap. Discrete-Continuous Variable Distinction. URL: https://mybiostats.wordpress.com/wp-content/uploads/2015/03/3rd-ed-alan_agresti_categorical_data_analysis.pdf.

[2] Coco I. Moreno et al. *crqa*. 2023. URL: https://cran.r-project.org/web/packages/crqa/index.html.

[3] Jari Oksanen et al. *Vegan*. 2023. URL: https://cran.r-project.org/web/packages/vegan/index.html.

[4] Teodor Sakal Franciskovic et al. “Predicting Students’ Final Exam Scores Based on Their Regularity of Engagement with Pre-Class Activities in a Flipped Classroom”. In: (2024). URL: %22https://www.scitepress.org/Papers/2024/126818/126818.pdf.

[5] Zhenhui Liu et al. “Modeling the learning behaviors of massive open online courses”. In: (2015), pp. 2883–2885. DOI: 10.1109/BigData.2015.7364110. URL: https://ieeexplore.ieee.org/document/7364110.

[6] Marti J. Anderson. “A new method for non-parametric multivariate analysis of variance”. In: *Austral Ecology* (2001). URL: https://ecoevol.ufg.br/adrimelo/div/Anderson-2001-AustEcol_non-parametric_manova.pdf.

[7] Jonathan D. Bakker. “Applied Multivariate Statistics in R”. In: Pressbooks, 2024. Chap. PERMANOVA. URL: https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/permanova/.

[8] Adar Ben-Eliyahu et al. “Investigating the Multidimensionality of Engagement: Affective, Behavioral, and Cognitive Engagement Across Science Activities and Contexts”. In: *Contemporary Educational Psychology* 53 (Jan. 2018). DOI: 10.1016/j.cedpsych.2018.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0361476X17300334.

[9] Marie-Luce Bourguet. “Demonstrating the impact of study regularity on academic success using learning analytics”. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 2024, pp. 736–741. DOI: 10.1145/3636555.3636845. URL: https://doi.org/10.1145/3636555.3636845.

[10] Peter J. Brockwell and Richard A.. Davis. “Time series: Theory and Methods”. In: Springer, 1987. Chap. Examples of Time Series. URL: https://link.springer.com/book/10.1007/978-1-4899-0004-3.

[11] Annoejska J. Cabo and Renate G. Klaassen. “Active learning in redesigning mathematics courses for engineering students”. In: 2018. URL: https://pure.tudelft.nl/ws/portalfiles/portal/47533309/20180630Final_CDIO2018_Paper.pdf.

[12] Armando P. Delfino. “Student Engagement and Academic Performance of Students of Partido State University”. In: *Asian Journal of University Education* 15 (2019). URL: https://www.researchgate.net/publication/338275769_STUDENT_ENGAGEMENT_AND_ACADEMIC_PERFORMANCE_OF_STUDENTS_OF_PARTIDO_STATE_UNIVERSITY.

[13] Barbara Flunger et al. “The Janus-faced nature of time spent on homework: Using latent profile analyses to predict academic achievement over a school year”. In: *Learning and Instruction* 39 (2015), pp. 97–106. DOI: https://doi.org/10.1016/j.learninstruc.2015.05.008. URL: https://www.sciencedirect.com/science/article/pii/S0959475215300086.

[14] Grasple. *Grasple.* URL: https://www.grasple.com/.

[15] Andre Hensbergen and Nikolaas Verhulst. “Open Linear Algebra Book”. In: TU Delft, 2024. Chap. 8.1.

[16] Andre Hensbergen and Nikolaas Verhulst. “Open Linear Algebra Book”. In: TU Delft, 2024. Chap. 6.2.

[17] Marian Hickendorff et al. “Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis”. In: *Learning and Individual Differences* 66 (2018), pp. 4–15. DOI: https://doi.org/10.1016/j.lindif.2017.11.001. URL: https://www.sciencedirect.com/science/article/pii/S1041608017301863.

[18] Jonathan C. Hilpert and Gwen C. Marchand. “Complex Systems Research in Educational Psychology: Aligning Theory and Method”. In: *Educational Psychology* (2018). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6701846/.

[19] Harold Hotelling. “Analysis of a complex of statistical variables into principal components”. In: *Journal of Educational Psychology* 24 (1933), pp. 417–441. URL: https://doi.org/10.1037/h0071325.

[20] Michael J. Jacobson. “Complexity Conceptual Perspectives for Research About Educational Complex Systems”. In: *The Journal of Experimental Education* 88.3 (2020), pp. 375–381. DOI: 10.1080/00220973.2019.1652138. URL: https://www.tandfonline.com/doi/full/10.1080/00220973.2019.1652138.

[21] Jelena Jovanovic et al. “Predictive power of regularity of pre-class activities in a flipped classroom”. In: *Computers Education* 134 (2019), pp. 156–168. DOI: https://doi.org/10.1016/j.compedu.2019.02.011. URL: https://www.sciencedirect.com/science/article/pii/S0360131519300405.

[22] Jelena Jovanović et al. “Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success”. In: *Computers Education* 172 (2021), p. 104251. DOI: https://doi.org/10.1016/j.compedu.2021.104251. URL: https://www.sciencedirect.com/science/article/pii/S0360131521001287.

[23] Cui Yunhuo Lei Hao and Zhou Wenye. “Relationships between student engagement and academic achievement: A meta-analysis”. In: *Social Behavior and Personality: an international journal* 46 (2018), pp. 517–528. URL: https://www.researchgate.net/publication/324183400_Relationships_between_student_engagement_and_academic_achievement_A_meta-analysis.

[24] Norbert Marwan and Charles L. Webber. “Mathematical and Computational Foundations of Recurrence Quantifications”. In: *Recurrence Quantification Analyasis: Theory and Best Practices.* Ed. by Charles L. Webber Jr. and Norbert Marwan. Springer International Publishing, 2015, pp. 3–43. DOI: 10.1007/978-3-319-07155-8_1. URL: https://doi.org/10.1007/978-3-319-07155-8_1.

[25] Oleksandra Poquet, Jelena Jovanovic, and Abelardo Pardo. “Student Profiles of Change in a University Course: A Complex Dynamical Systems Perspective”. In: *LAK23: 13th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 2023, pp. 197–207. DOI: 10.1145/3576050.3576077. URL: https://doi.org/10.1145/3576050.3576077.

[26] PRIME. *PRIME*. URL: https://www.tudelft.nl/en/eemcs/the-faculty/departments/applied-mathematics/education/prime/.

[27] Michael J. Richardson, Rick Dale, and Kerry L. Marsh. “Complex Dynamical Systems in Social and Personality Psychology: Theory, Modeling, and Analysis”. In: *Handbook of Research Methods in Social and Personality Psychology*. Ed. by Harry T. Reis and Charles M.Editors Judd. Cambridge University Press, 2014, pp. 273–274. URL: https://www.researchgate.net/publication/259892479_Complex_dynamical_systems_in_social_and_personality_psychology_Theory_modeling_and_analysis.

[28] Bart Rienties et al. “Unpacking the intertemporal impact of self-regulation in a blended mathematics environment”. In: *Computers in Human Behavior* 100 (2019), pp. 345–357. DOI: https://doi.org/10.1016/j.chb.2019.07.007. URL: https://www.sciencedirect.com/science/article/pii/S074756321930250X.

[29] George D. Kuh Robert M. Carini and Stephen P. Klein. “Student Engagement and Student Learning: Testing the Linkages”. In: *Research in Higher Education* 47 (2006), pp. 1–32. URL: https://link.springer.com/article/10.1007/s11162-005-8150-9.

[30] M. B. Wilk S.S. Shapiro. “An Analysis of Variance Test for Normality”. In: *Biometrika* 52 (1965), pp. 591–611. DOI: https://doi.org/10.2307/2333709. URL: https://www.jstor.org/stable/2333709.

[31] Mohammed Saqr and Sonsoles López-Pernas. “The longitudinal trajectories of online engagement over a full program”. In: *Computers Education* 175 (2021). DOI: https://doi.org/10.1016/j.compedu.2021.104325. URL: https://www.sciencedirect.com/science/article/pii/S0360131521002025.

[32] Xudong Huang Shan Li Juan Zheng and Charles Xie. “Self-regulated learning as a complex dynamical system: Examining students’ STEM learning in a simulation environment”. In: *Learning and Individual Differences* 95 (Mar. 2022). DOI: 10.1016/j.lindif.2022.102144. URL: https://www.sciencedirect.com/science/article/pii/S1041608022000310.

[33] Claude E. SHANNON. “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* (1948). URL: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.