ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Μελέτη των παραγόντων κινδύνου που σχετίζονται με την πανδημία COVID-19 με χρήση Τεχνικών Πολυμεταβλητής Ανάλυσης

Ιωάννα Π. Χαμπεσή

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Οκτώβριος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Μελέτη των παραγόντων κινδύνου που σχετίζονται με την πανδημία COVID-19 με χρήση Τεχνικών Πολυμεταβλητής Ανάλυσης

Ιωάννα Π. Χαμπεσή

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Οκτώβριος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Κούτρας Μάρκος, Καθηγητής (Επιβλέπων)
- Τήνιος Πλάτων, Καθηγητής
- Τριανταφύλλου Ιωάννης, Επικ. Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Study of Risk Factors Related to the COVID-19 Epidemic Using Multivariate Analysis Techniques

By

Ioanna P. Champesi

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
October 2024

Στους γονείς μου
Μαρία και Παναγιώτη και
στο φίλο μου Μανώλη.

Ευχαριστίες

Ολοκληρώνοντας την παρούσα Μεταπτυχιακή Διατριβή, δεν θα μπορούσα να μην αναφερθώ στα άτομα που συνετέλεσαν σε αυτή. Αρχικά, θα ήθελα να ευχαριστήσω θερμά τα μέλη της τριμελούς επιτροπής για το χρόνο που αφιέρωσαν στην αξιολόγηση της Διατριβής και ιδιαιτέρως τον επιβλέποντα Καθηγητή Μάρκο Κούτρα για την αδιάκοπη καθοδήγησή του στην εκπόνηση αυτής. Επιπλέον, θα ήθελα να ευχαριστήσω, τους συμφοιτητές και τους φίλους μου, και ιδιαίτερα τον Γιώργο Κουνιό, για την εμψύχωσή τους καθ' όλη την διάρκεια του Μεταπτυχιακού. Τέλος, ευχαριστώ από τα βάθη της καρδιάς μου την οικογένειά μου η οποία με στηρίζει αδιαλείπτως σε κάθε κίνησή μου.

Περίληψη

Η πανδημία COVID-19 έχει επιφέρει σημαντικές παγκόσμιες προκλήσεις, απαιτώντας λεπτομερή ανάλυση δεδομένων μεγάλου όγκου με στόχο τη λήψη αποφάσεων δημόσιας υγείας. Η παρούσα διπλωματική εργασία μελετά την εφαρμογή μεθόδων πολυμεταβλητής ανάλυσης για την εξερεύνηση μη κλινικών παραγόντων που συμβάλλουν στην επίδραση της πανδημίας COVID-19 στην κοινότητα της Μαδρίτης. Παρόλο που η υπάρχουσα βιβλιογραφία επικεντρώνεται κυρίως στα κλινικά αποτελέσματα, η εργασία έχει ως στόχο να διευρύνει την κατανόηση της πανδημίας μέσω της εξέτασης δημογραφικών, κοινωνικοοικονομικών και κλιματολογικών μεταβλητών. Η έρευνα χρησιμοποιεί την Ανάλυση Κύριων Συνιστωσών, την Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων και την k-means μέθοδο ανάλυσης κατά συστάδες για την επεξεργασία και ερμηνεία δεδομένων υψηλών διαστάσεων. Αυτές οι τεχνικές είναι κατάλληλες για την αναγνώριση σύνθετων προτύπων και σχέσεων που μπορεί να υπάρχουν στα δεδομένα, βοηθώντας στην ανίχνευση κρίσιμων παραγόντων που σχετίζονται με την πανδημία COVID-19. Βασιζόμενοι στη μελέτη των Pérez-Segura et al.[66], εφαρμόζουμε και παρουσιάζουμε πώς μπορούν οι προαναφερθείσες τεχνικές να εφαρμοστούν σε πραγματικά δεδομένα για την ανακάλυψη κρίσιμων παραγόντων κινδύνου που συνέβαλαν στην έξαρση της πανδημίας. Τα ευρήματα αναδεικνύουν την ανάγκη για προσαρμοσμένες στρατηγικές δημόσιας υγείας, λαμβάνοντας υπόψη τις διαφορετικές συνθήκες σε διάφορες περιοχές. Μέσα από αυτήν την ανάλυση, στοχεύουμε στην καλύτερη κατανόηση των δεδομένων, και κατ' επέκταση τη βαθύτερη κατανόηση της πολυδιάστατης επίδρασης της πανδημίας.

Abstract

The COVID-19 pandemic has introduced significant global challenges, requiring detailed analysis of vast datasets to guide public health decisions. This thesis studies the implementation of multivariate analysis methods to explore the non-clinical factors influencing the pandemic's impact in the community of Madrid. While much of the existing research has focused on clinical outcomes and medical interventions, the present study aims at broadening the understanding by examining demographic, socioeconomic, and climatological variables. The research employs Principal Component Analysis, Partial Least Squares Regression, and k-means clustering to process and interpret high-dimensional data. These techniques are suitable for uncovering complex patterns and relationships that may be present in the datasets, thereby shedding light on some of the key factors associated with COVID-19 transmission and severity. Building on the study by Pérez-Segura et al. [66], the thesis demonstrates how these techniques can be applied to real-world data to uncover critical risk factors. The findings emphasise the need for tailored public health strategies, considering the diverse conditions across different regions. Through this exploration, it aims to enhance the statistical power of data analysis, ensuring more accurate and comprehensive insights into the pandemic's multidimensional impact.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
BIC	Bayesian information criterion
CO	Carbon Monoxide
COVID-19	Coronavirus Disease of 2019
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DRTs	Dimensionality Reduction Techniques
GMM	Gaussian Mixture Models
ICA	Independent Component Analysis
LVs	Latent Variables
LLE	Locally Linear Embedding
MIC	Mean Item Complexity
NO2	Nitrogen Dioxide
NO	Nitrogen Monoxide
NIPALS	Non-linear Iterative Partial Least Squares
NMF	Non-negative Matrix Factorization
OPTICS	Ordering Points To Identify the Clustering Structure
PLSR	Partial Least Squares Regression
PM2.5	Particulate Matter
PCA	Principal Component Analysis
PCs	Principal Components
RSD	Respiratory System related Death
RMSE	Root Mean Square Error
SIMPLS	SIMplified of the Partial Least Squares
SVD	Singular Value Decomposition
SO2	Sulphur Dioxide
SS	Sum of Squared
t-SNE	t-Distributed Stochastic Neighbor Embedding
TAI	Total Accumulated Infections

Dis	Total Dispersion
UMAP	Uniform Manifold Approximation and Projection
WHO	World Health Organization

List of Abbreviations in Greek

AIDS	Σύνδρομο Επίκτητης Ανοσοανεπάρκειας
AIC	Κριτήριο Πληροφορίας Akaike
ANOVA	Ανάλυση Διασποράς
BIC	Κριτήριο Πληροφορίας Bayes
CO	Μονοξείδιο του Άνθρακα
COVID-19	Νόσος του Κορωνοϊού 2019
DBSCAN	Χωρική Ομαδοποίηση Εφαρμογών με Θόρυβο με Βάση Την Πυκνότητα
DRTs	Τεχνικές Μείωσης Διαστάσεων
GMM	Μοντέλα Μίξης Γκαουσιανών Κατανομών
ICA	Ανάλυση Ανεξάρτητων Συνιστωσών
LVs	Λανθάνουσες Μεταβλητές
LLE	Τοπικά γραμμική εμφύτευση
MIC	Μέση Πολυπλοκότητα Στοιχείου
NO2	Διοξείδιο του Αζώτου
NO	Μονοξείδιο του Αζώτου
NIPALS	Μη-γραμμική Επαναληπτική Μέθοδος Μερικών Ελαχίστων Τετραγώνων
NMF	Μη-αρνητική Παραγοντοποίηση Πινάκων
OPTICS	Διάταξη Στοιχείων για τον Εντοπισμό της Δομής Ομαδοποίησης
PLSR	Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων
PM2.5	Σωματιδιακή Ύλη
PCA	Ανάλυση Κύριων Συνιστωσών
PCs	Κύριες Συνιστώσες
RSD	Θάνατοι Σχετικοί με το Αναπνευστικό Σύστημα
RMSE	Ρίζα Μέσης Τετραγωνικής Απόκλισης
SIMPLS	Απλοποιημένη Ανάλυση Μερικών Ελαχίστων Τετραγώνων
SVD	Διάσπαση (Πίνακα) Σε Ιδιάζουσες Τιμές
SO2	Διοξείδιο του Θείου

SS	Άθροισμα Τετραγώνων
t-SNE	Στοχαστική Εμφύτευση Γειτνίασης με Βάση την Κατανομή t
TAI	Συνολικές Συσσωρευμένες Μολύνσεις
Dis	Συνολική Σκέδαση
UMAP	Προσέγγιση και Προβολή Ομοιόμορφης Πολλαπλότητας
WHO	Παγκόσμιος Οργανισμός Υγείας

CHAPTER 1

Introduction

The 2019 Coronavirus Disease (COVID-19) pandemic has presented unprecedented challenges worldwide, necessitating extensive analysis of large datasets to aid public health decision-making. As the virus spreads rapidly, understanding transmission dynamics, identifying at-risk populations, and predicting patient outcomes became crucial. Governments and healthcare providers had to quickly adapt, relying heavily on data to provide effective responses. This prompted the need for complex analytical algorithms capable of processing high-dimensional datasets.

The existing literature on COVID-19 has focused mostly on the clinical aspects of the pandemic, which is justifiable considering the urgent need to develop therapies, vaccines, and public health guidelines globally. Significant strides have been made in these areas, resulting in the development and distribution of vaccines and treatment protocols that have saved countless lives (Polack et al. [67], Baden et al. [25]). Additionally, numerous studies have focused on the clinical consequences of hospitalised COVID-19 patients (Zhou et al. [82], Evans et al. [36]). The impact of the pandemic was enormous globally but uneven across countries. According to the World Health Organization (WHO), some countries have managed to avoid any coronavirus cases since the pandemic started in early 2020. Many of these places are islands in the Pacific and Atlantic Oceans, likely benefiting from their isolation and being surrounded by the sea. Their success may also be due to strict travel policies. In contrast, countries like China, Italy, and Spain were heavily affected (Haug et al. [44]). This difference can be attributed to various clinical and non-clinical factors. Quarantines, non-pharmaceutical intervention strategies, climate, geographic coordinate system, and socioeconomic conditions are among the potential factors that influenced the impact of COVID-19 on different countries and communities. Despite

this, many aspects remain unclear, prompting ongoing scientific research to identify non-clinical risk factors that make societies vulnerable to these types of events. In the case of COVID-19 pandemic, this task is particularly difficult due to the short duration of the pandemic and the complex relationships among factors from different aspects, making the answer to this question multi-dimensional and requiring a collaborative approach. Thus, researchers have turned to the development of more individualised approaches. The factors that can be used to express aspects of the COVID-19 pandemic in some countries or even continents are quite diverse, thereof making it almost impossible to create a standardised (generalised) model. Under these conditions, it is clear that “one-size-fits-all” approaches in the aftermath of a pandemic cannot be effectively developed, and the need for tailored strategies is evident due to the heterogeneity of the factors across the world (Nicola et al. [62], Hale et al. [42]).

Multivariate analysis has emerged as a vital tool in this context, offering the capability to process and interpret complex datasets and identify associations, interactions, and patterns among multiple variables analysed simultaneously. The primary goal is to simplify high-dimensional data while retaining essential information, allowing for more accurate identification of the dataset's structure (Jolliffe and Cadima [50], Koutras [13]). This type of analysis enables researchers to uncover patterns and relationships that are not immediately apparent, providing a deeper understanding of the subject under investigation. For instance, multivariate analysis has helped to identify key risk factors associated with severe COVID-19 outcomes, enabling targeted interventions for vulnerable populations (Pérez-Segura et al. [66]). Their study demonstrates the application of multivariate techniques in COVID-19 research. By examining the complex interactions between various demographic, socioeconomic, and climatological variables, the study provides insights into the factors contributing to COVID-19 spread and severity in the community of Madrid. Such studies highlight the critical role of multivariate analysis in advancing our understanding of COVID-19 and its transmission within multi-dimensional spaces.

Among the most popular multivariate methods are Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR), and k-means clustering. While these

methods may not be the best for every type of data analysis, they are particularly suitable for the subject analysed in this thesis. These dimensionality reduction and clustering techniques provide insights into large and complex datasets. Their application is especially useful in the context of the COVID-19 pandemic, as evidenced by the analysis conducted by Pérez-Segura et al. [66]. This thesis builds on the techniques of Pérez-Segura et al. [66] to demonstrate how multivariate methods can be leveraged in complex analysis. Our goal is not to dispute the established procedures or conclusions but to enhance the statistical power of the analysis by applying alternative methods. This approach aims to address potential gaps in the dataset and uncover additional insights. Specifically, we set out to achieve three key aims: First, we will collect and preprocess the data according to their framework to ensure a solid analytical foundation. Second, we will implement the multivariate techniques used in their study [66]. Third, we aim to explore alternative dimensionality reduction approaches, with a particular focus on evaluating PLSR's effectiveness in improving model performance and interpretability. Beyond overall model performance, we will investigate the contribution of individual variables and their associations with COVID-19 by analysing patterns within the DRTs components and latent variables. Through these aims, we seek to extend the findings of Pérez-Segura et al. [66] and provide further insights into the applicability of PLSR for analysing complex datasets.

Chapter 2 explores dimensionality reduction techniques, focusing on PCA and PLSR. These methods simplify complex datasets and enhance interpretability. PCA is particularly useful in exploratory data analysis, identifying underlying structures within the data (Abdi and Williams [20], Pearson [65], Koutras [13]). PLSR combines features of PCA and multiple regression, finding latent variables that capture significant variation in the response variable, thus increasing predictive power.

Chapter 3 explores clustering techniques, primarily k-means clustering. This partitioning method classifies datasets into distinct clusters summarised by their centroids. k-means clustering effectively identifies homogeneous subgroups within a dataset, useful in fields like image analysis, bioinformatics, social sciences and pattern recognition (MacQueen [59]).

Chapter 4 applies the methodologies from Chapters 2 and 3 to analyse the COVID-19 pandemic, drawing insights from the study conducted by Pérez-Segura et al. [66]. This chapter demonstrates how these techniques can be employed to understand non-clinical risk factors associated with COVID-19 transmission and severity, aiming to enhance understanding of the disease's impact on specific communities.

Chapter 5 offers concluding remarks, summarising the findings and highlighting the significance of multivariate analysis techniques in advancing our understanding of complex phenomena like the COVID-19 pandemic. It discusses the implications of the research conducted and suggests potential avenues for future exploration in the field of multivariate analysis and public health.

CHAPTER 2

Dimensionality Reduction

2.1 Introduction

Understanding data analysis approaches can often prove perplexing, especially when it comes to exploring complex high-dimensional spaces. Within the latter, a new set of challenges emerges, commonly categorised under the term “curse of dimensionality”. Unlike lower-dimensional spaces, where information is more densely populated, high-dimensional ones experience sparsity, making it difficult to draw accurate conclusions. With the increase in dimensions, the amount of data required for conducting a reliable analysis grows exponentially, often exceeding practical limits. Alongside the challenges encountered in conventional techniques, the additional assumptions lurking in high-dimensional data further complicate analysis and interpretation, making it increasingly challenging to gain a deep understanding of the data. To address these complications, Dimensionality Reduction Techniques (DRTs) have been developed to project high-dimensional information into a lower-dimensional space while preserving a large percentage of the information contained in the original data. By decreasing dimensionality, DRTs allow more effective analysis and visualisation of complex datasets. This thesis explores the intricacies of dimensionality reduction, examining its importance, methods, and implications in data analysis (Van der Maaten and Hinton [74], Koutras [13]).

Facilitating the exploration and interpretation of large and complex datasets plays a crucial role in modern data analysis. DRTs can enhance the efficiency of machine learning algorithms by reducing computational complexity and improving predictive accuracy (Van der Maaten and Hinton [74]). In today's vastly information-driven world, the large amounts of data pose a significant difficulty for analysts to understand, making DRTs essential tools. While some aspects are inevitably missing, the

advantages of applying such techniques outweigh this concern (Ntotsis and Karagrigoriou [64]). Some of the key aspects that highlight the transformative power of DRT in simplifying data analysis, are:

- **Multicollinearity:** Multicollinearity is the phenomenon where independent variables are highly correlated, posing a significant challenge in regression analysis. Dimensionality reduction can solve this problem by transforming the correlated variables into a new set of variables that are not correlated with each other, making the model more stable and easier to interpret.

- **Noise Reduction:** By retaining only the most important features, researchers can focus on key information, potentially leading to improved model accuracy.

- **Computational Efficiency:** Fewer dimensions not only simplify the model but also enhance computational efficiency. For example, machine learning algorithms trained on lower-dimensional data are more efficient in terms of execution time and computational resources.

- **Visualisation**: Visualization is an essential part of data analysis. Reducing the data dimensionality to two or three dimensions, can help researchers gain insights into the underlying structure and relationships.

- **Overfitting:** In most cases, datasets contain numerous features; hence, reducing their number makes the model less prone to overfitting.

- **Non-linear to Linear Data**: Through DRTs, non-linear data can be transformed into linearly separable data, leading to a more easily analysable model.

2.2 Principal Component Analysis

Large datasets are accompanied by a plethora of problems, increasing the potential for inconsistencies in both analysis and interpretation. Principal Component Analysis (PCA) stands as a fundamental tool for reducing data complexity. Introduced by Pearson [65] and further developed by Hotelling ([47],[48]), PCA transforms correlated variables into a new set of

uncorrelated ones (Principal Components - PCs), which are ordered so as to capture the maximum variance of the original data, facilitating easier interpretation and analysis (Koutras [13]). By compressing the most significant variation into the initial PCs, PCA provides a condensed yet accurate representation of the original dataset. This approach enhances the interpretability of statistical tests and analysis results without sacrificing data fidelity. PCA finds applications in several scientific areas such as finance, biology, image processing and social sciences. In finance, it helps identify underlying factors that drive asset price movements; in biology, it aids in deciphering complex gene expression patterns; in image compression and reconstruction of visual data, it enables faster transmission and storage; in social sciences it helps understanding patterns among highly correlated variables (Varmuza and Filzmoser [75], Hoffmann and Bradley [46]).

To better understand how PCA functions, it is essential to explore how the principal components are derived. The process begins by identifying the first PC, which captures the largest amount of variance in the data. Below, we walk through the mathematical steps involved in deriving the first principal component.

2.2.1. Deriving the first PC

Let us consider a sample of n individuals, each with p observed characteristics (random variables) $X_1, X_2, \dots, X_P$. We represent the data in an $n \times p$ matrix X, where each element x_{ij} contains the value of the j-th characteristic for the i-th individual

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \tag{2.1}$$

Then, we define a new variable Y, which is a linear combination of the original variables $X_1, X_2, \dots, X_P$

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p,$$

where $\alpha_1, \alpha_2, \dots, \alpha_p \epsilon \mathcal{R}$ are coefficients that determine the linear combination. For each observation i, the corresponding value of the new variable Y is given by

$$y_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}, \qquad i = 1,2, \dots, n.$$

The goal of PCA is to determine the coefficients $\alpha_1, \alpha_2, \dots, \alpha_p$ so that the variance of Y is maximised. This is equivalent to maximising the total sum of squares of Y, which can be expressed mathematically as

$$Dis(Y) = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{2.2}$$

Starting from the sum of squared distances (SS) of the observations from the mean of Y, we have

$$SS_Y = \sum_{r=1}^{n} \sum_{i=r+1}^{n} (y_r - y_i)^2 = \frac{1}{2} \sum_{r=1}^{n} \sum_{i=1}^{n} (y_r - y_i)^2 =$$

$$\frac{1}{2} \sum_{r=1}^{n} \left[\sum_{i=1}^{n} (y_r - \bar{y})^2 - 2 \sum_{i=1}^{n} (y_r - \bar{y})(y_i - \bar{y}) + \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]. \tag{2.3}$$

The sum of cross terms ($\sum_{i=1}^{n} (y_r - \bar{y})(y_i - \bar{y})$) simplifies because of the mean property $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$, thus Eq. (2.3) takes the form

$$SS_Y = \frac{1}{2} \sum_{r=1}^{n} \left[\sum_{i=1}^{n} (y_r - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 \right] =$$

$$\frac{1}{2} \sum_{r=1}^{n} [n(y_r - \bar{y})^2 + Dis(Y)] = \frac{1}{2} \big(nDis(Y) + nDis(Y) \big) = nDis(Y),$$

thus, we have proved that

$$SS_Y = nDis(Y). \tag{2.4}$$

Therefore, the problem of maximising SS_Y of Eq. (2.4) is equivalent to finding the coefficients $\alpha_1, \alpha_2, \dots, \alpha_p$ for which the $Dis(Y)$ of Eq. (2.2) is maximised. Hence, it comes down to the conclusion that the first PC is the linear combination that maximise the $Var(Y)$, i.e.,

$$\max_{\alpha_1,\alpha_2,\dots,\alpha_p} Var(Y)$$

subject to the constraint

$$\sum_{j=1}^{p} a_j^2 = 1 \Leftrightarrow \|a\| = 1.$$

This maximisation is achieved by considering the eigenvector corresponding to the largest eigenvalue of the covariance matrix S, where $S = \frac{1}{n-1} X'X$. The eigenvector provides the coefficients $\alpha_1, \alpha_2, \dots, \alpha_p$ that define the first PC.

While the eigenvector mathematically defines the first PC, understanding its geometric interpretation provides a deeper insight into how these components relate to the data. By considering the geometric representation, we can better grasp how the data is transformed in the reduced dimensional space. The geometric interpretation can be used to express $Dis(Y)$ through the matrix $X = (x_{ij})$. If we denote by $\bar{x}_j$ the sample means for the values, we collected for each characteristic X

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,, \qquad j = 1,2,\dots,p\,, \tag{2.5}$$

then, $\bar{y}$ can be expressed as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}).$$

If we denote the scores of the PC for each individual by $f = y_i - \bar{y}$, we get

$$f = \alpha_1(x_{i1} - \bar{x}_1) + \alpha_2(x_{i2} - \bar{x}_2) + \cdots + \alpha_p\left(x_{ip} - \bar{x}_p\right),$$

which can be simplified as

$$f = \boldsymbol{z_i'a}, \qquad i = 1,2,\dots,n \tag{2.6}$$

where, $\boldsymbol{\alpha}$ is the vector of coefficients ($[\alpha_1, \alpha_2, \dots, \alpha_p]$) and $\boldsymbol{z_i}$ is the vector of centred data for the i-th observation ($[x_{i1} - \bar{x}_1, x_{i2} - \bar{x}_2, \dots, x_{ip} - \bar{x}_p]$).

The total dispersion of the set N along the vector $\boldsymbol{a}$ will be denoted as $Dis_\alpha(N)$, that is

$$Dis_\alpha(N) = Dis(Y) = \sum_{i=1}^{n} f_i^2 = \boldsymbol{f'f} = (Z\boldsymbol{a})'(Z\boldsymbol{a}) = \boldsymbol{a}'(Z'Z)\boldsymbol{a}. \tag{2.7}$$

Maximising the dispersion with respect to $\boldsymbol{\alpha}$, subject to the constraint $\boldsymbol{\alpha'\alpha} = 1$, yields the first PC –this is a well-known result from linear algebra. More specifically, if the matrix $Z'Z$ contains non-negative eigenvalues, denoted $\lambda_1, \lambda_2, \dots, \lambda_p$, and assuming that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, with corresponding unit eigenvectors $\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_p}$, we can establish the following

1. The vector $\boldsymbol{\alpha}$ that optimises the expression $\boldsymbol{a}'(Z'Z)\boldsymbol{a}$ is the eigenvector $\boldsymbol{u_1}$ associated with the highest eigenvalue λ_1.
2. The highest value of the quadratic form is equal to λ_1, thus

$$\max_{\|a\|=1} Dis_\alpha(N) = \max_{\|a\|=1} Dis_{u_1}(N) = \max_{\|a\|=1} \boldsymbol{a}'(Z'Z)\boldsymbol{a} = \lambda_1.$$

The variable Y, which maximises this expression, is referred to as the first principal component. This provides a concise explanation of how the first PC is derived by finding the direction (eigenvector) that maximises the variance (associated with the largest eigenvalue) in the dataset.

2.2.2 Deriving the rest PCs

Building on the framework provided for the first PC, the concept of principal components can be extended to identify additional components that capture the remaining variance in the data, while being orthogonal to the previous components. After determining the first, which corresponds to the direction along which the variance λ_1 is maximised, the procedure continues to identify additional PCs. These components are orthogonal to the previous ones and account for the remaining variance in the dataset. To formalise this process, the concepts of orthogonality and maximisation come into play, ensuring that each subsequent PC is computed in a way that preserves these important properties.

Each subsequent PC corresponds to an eigenvector of the matrix $Z'Z$. The j-th PC is associated with the unit eigenvector u_j that maximises the $\boldsymbol{a}'(Z'Z)\boldsymbol{a}$, under the condition that $\boldsymbol{a}$ is orthogonal to all previously found eigenvectors $\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_{j-1}}$. The j-th PC captures the variance λ_j, corresponding to the j-th largest eigenvalue of $Z'Z$

$$Dis_{u_j}(N) = \boldsymbol{u_j'}(Z'Z)\boldsymbol{u_j} = \lambda_j, \qquad j = 1,2, \dots, p. \tag{2.8}$$

The total dispersion $Dis(N)$ of the dataset can be decomposed as the sum of the dispersions along all PCs

$$Dis(N) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \boldsymbol{u_j'}(Z'Z)\boldsymbol{u_j}\,. \tag{2.9}$$

This means that each PC contributes λ_j to the total variance of the data.

The proportion of the total variance explained by the first j PCs is given by

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_j}{Dis(N)} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \qquad 1 \leq j \leq p. \tag{2.10}$$

This indicates the cumulative proportion of variance captured by the first j components.

In practice, when p is large, often only the first few PCs (those associated with the largest eigenvalues) are retained, as they capture the majority of the variance. The matrix $Z'Z$ can be approximated by summing only the terms corresponding to these leading eigenvalues

$$Z'Z \approx \lambda_1 \boldsymbol{u_1 u_1'} + \lambda_2 \boldsymbol{u_2 u_2'} + \cdots + \lambda_j \boldsymbol{u_j u_j'}, \qquad 1 \leq j \leq p. \tag{2.11}$$

If the eigenvalues $\lambda_{j+1} \geq \ldots \geq \lambda_p$ are relatively small, then the approximation in Eq. (2.11) provides a good representation of the original data in a lower-dimensional space.

Note that many functions in the programming language *R*, such as the function *principal*() from package "psych" that we utilise in Chapter 4, often represent the eigenvectors in scaled forms related to the principal components

$$\boldsymbol{v_j} = \sqrt{\lambda_j}\boldsymbol{u_j}, \qquad 1 \leq j \leq p.$$

The eigendecomposition, i.e., the process of finding eigenvectors and eigenvalues of the input matrix, is typically implemented and applied after scaling the data. This decomposition is generally recommended in high-dimensional cases to formulate the PCs due to its lower computational demands. An alternative to eigendecomposition is the Singular Value Decomposition (SVD), which decomposes the original data matrix into three matrices: one containing the left singular vectors, a diagonal matrix holding the singular values, and one containing the right singular vectors. While SVD offers numerical stability and versatility in handling non-square matrices, it requires the calculation of all singular values and vectors, potentially increasing computational complexity. Additionally, it may not offer as straightforward an interpretation of results as eigendecomposition.

Beyond SVD, other methods such as iterative approaches including Von Mises Iteration, also known as power iteration (Von Mises and Pollaczek-Geiringer [77], Roughgarden and Valiant [17]), Lanczos iteration (Lanczos [57]), and the Jacobi eigenvalue algorithm (Golub and Van der Vorst [40]), could also be used for formulating scores and loadings. These iterative methods, while potentially computationally efficient for large datasets, may require careful parameter tuning and do not guarantee convergence for all datasets.

In this study, we opt for eigendecomposition mainly due to its efficiency in handling high-dimensional data in terms of both time and memory consumption. It facilitates the identification of key patterns and structures within datasets by revealing the dominant directions of variation and serves as a foundational tool in several DRTs aiding in the exploration and interpretation of complex data relationships. All these factors make the method a reasonable choice for large-scale data analysis tasks, ensuring both accuracy and efficiency.

In general, the eigendecomposition can be performed either on the covariance or the correlation matrix. When the eigendecomposition is performed on the covariance matrix, it helps to identify the directions in which the data varies the most. On the other hand, when the eigendecomposition is performed on the correlation matrix, it helps in finding the directions in which the data vary the most, but this time it signifies how each feature is related to the others while considering their different scales or units of measurement. The correlation matrix is often preferred because it allows for easier comparison of values and facilitates clearer communication of information regarding any linear patterns present in the dataset.

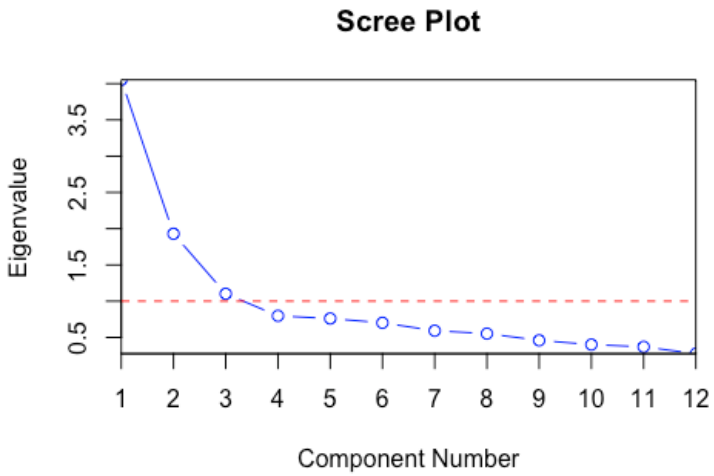
One of the most crucial tasks in PCA is determining the number of components to retain. This decision is critical as it directly influences the interpretability and performance of subsequent analysis (Jolliffe [49]). Various methods exist to ascertain an optimal number of components, each with its own rationale, advantages, and drawbacks. Among the most effective and commonly used methods are (Koutras [13])

1. **Kaiser's Criterion**: Kaiser's criterion suggests retaining components with eigenvalues greater than one, as these components explain more variability than a single original variable and are therefore considered worthy of retention.

While straightforward and easy to implement, Kaiser's criterion may not always be applicable, especially in datasets with complex structures or non-linear associations (Kaiser [52]).

2. **Catell's Scree Test** (Scree Plot): One popular approach involves plotting the eigenvalues against the number of components. The "elbow" point on the resulting plot (see e.g. Fig. 2.1), where the eigenvalues begin to level off, is typically considered as an indicator of the optimal number of components to retain. This approach allows for a visual assessment of the proportion of variance explained by each component. While simple and intuitive, the subjective determination of the "elbow" point can sometimes be challenging, leading to potential ambiguity in the selection process (Cattell [31]).

Fig 2.1: A sample scree plot produced in R. The Kaiser criterion is shown in red.
Source: Wikipedia [18]



3. **Cumulative Variance**: By selecting the number of components that enclose a high percentage of variance (usually at least 70% or 80%), researchers can effectively capture the structure of the original data (see Eq. (2.11)). This approach provides a clear criterion for component selection, ensuring that a significant amount of information is retained. However, it may result in

retaining more components than necessary, which can (potentially) lead to overfitting and increased computational complexity (Wehrens [78]).

Selecting the optimal number of components involves careful consideration of various factors, including the proportion of variance explained, model complexity, and predictive performance. While no single approach is considered superior, a combination of them, tailored to the specific characteristics of the dataset and the research objectives, can help ensure a robust and meaningful reduction of the dimensions. In most cases, it is advisable to utilise the initial 2 or 3 PCs, as they can be effectively visualised in 2- or 3-dimensional spaces.

As with many statistical methods, PCA works under certain assumptions. If these assumptions are violated, the validity and interpretability of the results may be compromised. The core assumptions of PCA are:

1. **Linearity:** The first assumption is that the relationships between independent variables are linear.

2. **Large Sample Size:** PCA performs better with a large sample size, which ensures that the covariance matrix is accurately estimated. Small sample sizes can potentially lead to unstable results.

3. **Independence of Principal Components:** The derived PCs are orthogonal, meaning they are statistically independent of each other, simplifies interpretation.

4. **Normality (optional):** PCA is generally insensitive to deviations from normality. However, when the data is normally distributed, the PCs more accurately reflect the underlying structure of the data, facilitating more meaningful interpretation. PCA can still be applied to non-normally distributed data, though the results may be less reliable.

Assessing and addressing these assumptions is essential for ensuring the validity, reliability, and interpretability of PCA results in practical data analysis scenarios (Laerd Statistics [14]). While PCA offers numerous benefits, it is crucial to recognise its limitations and potential drawbacks. In cases of large datasets or intricate multicollinearity structures, PCA may lead to model overestimation or underestimation, affecting the accuracy of analysis. Researchers must carefully evaluate dataset characteristics and explore alternative techniques, such as non-linear DRTs, to address these challenges effectively.

Overall, PCA represents a vital tool for data analysts, providing an efficient approach to dimensionality reduction. By simplifying data representation without compromising accuracy, PCA enables researchers to uncover hidden insights and make well-justified decisions across diverse applications.

2.3 Univariate Partial Least Squares Regression

Exploring beyond PCA, another DRT frequently used is the Partial Least Squares Regression (PLSR). While PCA uncovers patterns and reduces dimensionality by maximising variance in the predictor matrix X, PLSR focuses on finding components that explain the variation in the predictor matrix X while also maximising the relationship with the response variable Y. In other words, it aims to identify patterns in X that are most useful for predicting Y. Although both techniques share the same goal of simplifying complex data, each brings a unique perspective to the table.

As discussed in Section 2.2, PCA aims to capture the maximum variance in the data by creating PCs. These components are ordered by the amount of variance they explain, allowing for data visualisation and dimensionality reduction. However, PCA may not always be optimal for modelling tasks, particularly when the goal is to maximise the predictive power of the model. This is where PLSR can be useful, as it uses both predictor inputs and response outputs (Rosipal and Kramer [69]). To complement this, PLSR creates new predictors, often referred to as Latent Variables (LVs), which are linear combinations of the original predictor variables. These LVs are constructed so

that they are highly associated with the response variable, thus capturing predictive relationships between the two groups of variables. By doing so, it identifies a parsimonious representation of the predictor variables that most accurately corresponds to the response variable (Wold et al. [80], Wehrens [78]).

To better understand the decomposition of data in PLSR, it is useful to examine the approach outlined by Ng [15]. Let $X = [\boldsymbol{x_1} \dots \boldsymbol{x_m}]$ be an $n \times m$ mean-centred matrix and let $Y = [\boldsymbol{y_1} \dots \boldsymbol{y_p}]$ be a $n \times p$ mean-centred matrix. In the univariate case, where $Y = y_1$ is an $n \times 1$ vector, the aim is to decompose the matrices X and Y as follows

$$X = TP'$$

and

$$Y = UQ'. \tag{2.12}$$

Here, T $(n \times l)$ and U $(n \times l)$ are matrices of latent scores (l denotes the number of extracted LVs), representing the new variables that are linear combinations of the original predictors and response variables. The matrices P and Q are loadings that define these linear combinations. Note that the scores, even though they are denoted as in PCA, their formulas differ. In the case of PLSR, scores represent the new coordinates of the samples in the LV space. PLSR is (commonly) employed with the use of NIPALS (Non-linear Iterative Partial Least Squares) (Wold [79]) or SIMPLS (SIMplified of the Partial Least Squares) (De Jong [34]) algorithm to iteratively derive the LVs. These algorithms iteratively calculate the weights, scores, and loadings. The steps for the NIPALS algorithm are as follows.

- **Step I (Initialisation):**
  - Choose an initial vector $t = X_j$ for some column of j of X.
  - Set $u = Y_j$ for some column j of Y.

- **Step II (Iterative Updates):**
  - Repeat the following procedure until convergence[1], i.e., until t stop changing

$$p := \frac{X'u}{\|X'u\|},$$

$$t := Xp,$$

$$q := \frac{Y't}{\|Y't\|},$$

$$u := Yq.$$

  - These steps ensure that the LVs t and u capture the most covariance between X and Y.

- **Step III (Deflation)[2]:**
  After determining the first pair of LVs, deflate the matrices X and Y to remove the information captured by the current LVs

$$X := X - tp'$$

and

$$Y := Y - uq'.$$

[1] In the case when the Y matrix has only one variable, we can set q to 1 and the last two steps of the loop can be omitted.

[2] The term "deflation" in the context of PLSR refers to the process of removing the information that has been captured by the current LV from the original matrices X and Y. This ensures that subsequent LVs explain different aspects of the data, rather than repeating the information already captured.

- **Step IV (Repetition):**
  Repeat Steps II and III to extract further LVs until the desired number of them is obtained.

Through this iterative process, PLSR captures the essential patterns in the data by projecting it onto a lower-dimensional space. This is done while simultaneously maximising the covariance between the predictors and the response variable. Note that the iterative process in PLSR can also be viewed through the lens of eigenvalue problems. The vectors p and q derived during the NIPALS iterations are not directly eigenvectors, but rather weight vectors that aim to maximise the covariance between X and Y through iterative projections.

Once the LVs T and U are determined, a regression model can be built up to relate X and Y through the latent variables, namely

$$Y = XP\beta Q', \tag{2.13}$$

where β is the matrix of regression coefficients that links the latent scores T and U.

When it comes to the assumptions that must be met, PLSR shares some with PCA due to their conceptual similarities in dimensionality reduction. PLSR must meet the same assumptions as PCA (i.e., linearity, absence of multicollinearity, adequate sample size, interval or ratio data, and normality). Additionally, PLSR has unique assumptions (Wold et al. [80]) such as

1. Homoscedasticity: PLSR assumes homoscedasticity, meaning that the variance of the residuals is constant across all levels of the independent variables. Heteroscedasticity can lead to inefficient estimates and affect the reliability of the model.

2. Independence: PLSR assumes that the observations are independent of each other. Violation of this assumption, such as in the presence of autocorrelation in time-series data, can lead to biased estimates and incorrect inference.

These assumptions are critical for ensuring the validity and reliability of the PLSR model. While some deviations can be tolerated, significant violations may require alternative methods or pre-processing steps to address these issues.

2.4 Multivariate Partial Least Squares Regression

One advantageous aspect of PLSR is its flexibility to handle both univariate and multivariate responses. In the case of a univariate response, as discussed, PLSR operates similarly to traditional regression models, where a single response variable is predicted based on a set of predictor variables. Thus, the univariate approach is useful for predicting a single outcome of interest.

However, PLSR can also be extended to a multivariate approach when dealing with multiple response variables. This capability allows to simultaneously model the relationships between predictor variables and multiple response variables, capturing complex dependencies among them. This multivariate approach is especially useful when analysing datasets with correlated responses or when the aim is to determine how predictors influence multiple outcomes simultaneously. A multivariate approach is beneficial because it accounts for the correlations between response variables, providing more robust and comprehensive modelling results. Additionally, by jointly modelling multiple responses, PLSR can reveal shared patterns and underlying relationships that may not be apparent when analysing each response variable separately. Nevertheless, implementing PLSR in multivariate scenarios, can potentially introduce challenges, particularly in terms of computation and interpretation. The calculation becomes more complex as the model must estimate the relationships between predictor variables and multiple response variables simultaneously. Additionally, interpreting the coefficients and contributions of

predictor variables to each response variable becomes more nuanced in the multivariate setting.

Table 2.1 summarises the similarities and dissimilarities between PCA and PLSR (Wold et al. [80], Van der Maaten and Hinton [74], Wehrens [78], Rosipal and Kramer [69]).

TABLE 2.1

Similarities and dissimilarities between PCA and PLSR.

PCA VS PLSR	
SIMILARITIES	**DISSIMILARITIES**
Dimensionality Reduction: Both PCA and PLSR tries to reduce the dimensions of original dataset by creating new variables (called PCs and LVs respectively).	**Supervision**: The primary distinction between PCA and PLSR lies in their supervision status. PCA is an unsupervised technique, meaning it operates solely based on the structure of the predictor variables. In contrast, PLSR is a supervised technique that explicitly incorporates information from both predictor and response variables to construct new variables
Linear Transformations: Both methods apply linear transformations to derive the new variables. PCA derives principal components by finding linear combinations of the original variables; PLSR constructs components that are linear combinations of both the predictor variables and the response variable.	**Objective**: PCA aims to maximise the variance of the data, capturing as much information as possible in the new variables. On the other hand, PLSR aims to maximise the covariance between the predictor variables and the response variable, focusing on predictive modelling rather than data exploration.
Orthogonality: The new variables created by PCA and PLSR are orthogonal to each other.	**Interpretation**: While PCA provides insight into the underlying structure of the data by identifying patterns and relationships among variables, PLSR is more focused on prediction, making it a preferred choice when the goal is to build a predictive model.

2.5 Other Dimensionality Reduction Techniques

In addition to the techniques discussed in detail above, there are a variety of DRTs, each of which has unique advantages and limitations. Below, we briefly document these alternative techniques to inform the reader of their existence and provide a broader context; however, they will not be implemented in the application of this thesis.

- **Sparse PCA** is a variant of PCA that encourages sparsity in the principal components, resulting in a more interpretable representation of the data. It selects a subset of the original features while retaining most of the variance. Sparse PCA is useful for feature selection and dimensionality reduction in high-dimensional datasets. However, finding the optimal sparsity level can be challenging, and the resulting representation may not always be intuitive (Bertsimas et al. [28]).

- **Kernel PCA** is a nonlinear extension of PCA that maps data into a higher-dimensional space using a kernel function before performing PCA. This allows capturing nonlinear relationships in the data. Kernel PCA is flexible and can handle complex data structures, however, it requires tuning of kernel parameters and may be computationally demanding for large datasets (Schölkopf et. al. [70]).

- **Independent Component Analysis** (ICA) is a technique used to separate a multivariate signal into independent components. It assumes that the observed data are a linear combination of independent sources and aims to recover these sources. ICA is particularly useful for blind source separation and feature extraction. However, it relies on strong statistical assumptions and may struggle with mixed signals or non-Gaussian distributions (Comon [33]).

- **Autoencoder** is a neural network architecture used for unsupervised learning of efficient data representations. It consists of an encoder and a decoder, which

learn to compress and reconstruct the input data, respectively. Autoencoders are versatile and can capture complex patterns in the data. However, they may suffer from overfitting and require careful architecture design and training (Kramer [56]).

- **Uniform Manifold Approximation and Projection** (UMAP) is a DRT that emphasises in preserving both local and global structures in the data. It constructs a low-dimensional representation by optimizing a low-dimensional embedding to match the high-dimensional data nearest neighbours. UMAP is known for its potential to handle large datasets efficiently. However, like t-SNE, UMAP may require parameter tuning and can be sensitive to local density variations.

- **t-Distributed Stochastic Neighbor Embedding** (t-SNE) is widely used for visualizing high-dimensional data by mapping them to a lower-dimensional space while preserving the local structure. It is particularly effective for exploring clusters and patterns in the data. However, t-SNE can be computationally demanding and sensitive to hyperparameters, requiring careful tuning (Van der Maaten and Hinton [74]).

- **Locally Linear Embedding** (LLE) aims to preserve local relationships within the data by reconstructing each data point as a linear combination of its neighbours. This technique is useful for unfolding nonlinear manifolds and capturing the underlying structure of the data. LLE is robust to noise and outliers, but it may struggle with very high-dimensional data and requires careful selection of parameters.

- **Isomap** is a method for nonlinear dimensionality reduction that focuses on preserving the global geometry of the data. By constructing a graph representing the data's intrinsic geometric structure, Isomap embeds the data into a lower-dimensional space. One advantage of Isomap is its potential to capture nonlinear

relationships in the data. However, it can be sensitive to noise and outliers, which may affect the quality of the embedding (Tenenbaum et al. [73]).

- **Non-negative Matrix Factorization** (NMF) decomposes a non-negative data matrix into two lower-dimensional matrices, representing parts-based representations of the original data. It is commonly used for feature extraction and modelling tasks. NMF is interpretable and suitable for sparse and non-negative data. However, it requires careful initialization and may be trapped in local optima.

- **Random Projection** is a simple yet effective technique for dimensionality reduction. It projects high-dimensional data onto a lower-dimensional subspace using a random matrix. Despite its simplicity, random projection can preserve pairwise distances reasonably well and is computationally efficient. However, it may not capture complex nonlinear relationships in the data as effectively as other methods.

DRTs can be classified using several criteria, each offering unique insights into data analysis. Understanding the diverse classifications and methodologies of them, enables researchers to implement these techniques effectively in various tasks. Whether seeking to simplify data structures, uncover hidden patterns, or enhance predictive models, DRTs offer a versatile way for exploring the complexities of high-dimensional data.

One fundamental classification of DRTs revolves around feature extraction and feature selection methods. Feature selection techniques, shift through variables to identify and remove irrelevant or redundant ones, resulting in a streamlined dataset. On the other hand, feature extraction techniques like PCA and PLSR create new variables that condense and summarise information from the original dataset.

Additionally, DRTs can be categorised based on the presence of class labels in the data, distinguishing between supervised and unsupervised approaches. Supervised

techniques like PLS use class labels to extract discriminant information or LVs from the data. In contrast, unsupervised techniques such as PCA and Isomap operate without class labels, identifying patterns and relationships based solely on the data's intrinsic structure.

Furthermore, DRTs can be categorised based on the data structure, and more precisely by taking into account whether it occurs linear or non-linear relationship between the variables of the examined dataset. Linear techniques, like PCA excel in capturing linear relationships between variables. On the other hand, non-linear methods such as Isomap and Kernel PCA are optimal at uncovering complex, non-linear patterns within the data.

Considering the distinctions outlined above, PCA emerges as an unsupervised linear feature extraction technique, highlighting complex patterns hidden within data structures without the need for external guidance. Conversely, PLSR takes on the role of a supervised linear feature extraction technique, leveraging class labels to guide the extraction of information and enhance predictive modelling capabilities.

Closing this Section, we mention that PCA is essentially a data-driven tool aimed at uncovering patterns in terms of variance explained, while PLSR is more geared towards predictive modeling and capturing the relationships between variables. The choice between the two techniques depends on the specific objectives and needs of the analysis. PLSR is often described as a supervised analogy of PCA, but their connection extends beyond the simple dichotomy between supervised and unsupervised methods. Despite their similarities, the theoretical frameworks and mathematical algorithms used are quite different. PLSR typically utilises algorithms such as SIMPLS (De Jong [34]) or NIPALS (Wold [79]) to iteratively generate LVs. These algorithms are designed to incorporate both predictor and response variables and maximise the covariance between them, which helps in capturing the predictive relationships between these variables. In contrast, PCA does not rely on such specific algorithms designed for regression tasks. It focuses only on the structure of the predictor variables and aims to maximise the variance of the data.

CHAPTER 3

Cluster Analysis

3.1 Introduction

Cluster analysis, or shortly clustering, is a crucial technique in data analysis for finding groups of related data points. It is used to explore hidden patterns or structures within large collections of data. This allows researchers to understand corresponding groups of similar data points, providing insights into the underlying structure of the data. Clustering involves partitioning a dataset into subsets (also called clusters) such that items belonging to the same cluster are more similar to each other than to those in different clusters. As the number of features or dimensions in a dataset increases, clustering faces unique challenges. High-dimensional data tends to be sparse, making it harder to identify meaningful clusters. Unlike lower-dimensional spaces, where clusters are more apparent, high-dimensional spaces require a significantly larger amount of data to maintain statistical reliability. This sparsity complicates both the analysis and interpretation of data, making it challenging to gain accurate insights (Hastie et al. [43]). To overcome these issues, many clustering algorithms have been engineered to work efficiently with high-dimensional data. These algorithms are tailored to discover complex patterns that are not easily observed, enhancing the processing and understanding of the data. Clustering reduces complexity in the data and increases the efficiency of machine learning models and other analytic methods (Aggarwal [21]).

Clustering is a well-known concept that has been practiced for ages. From an early age, we automatically group objects and entities in our minds, such as distinguishing dogs from cats, basketball from football, or even identifying someone's biological gender. It is through the process of grouping and sub-grouping that learning happens, as we organise patterns observed in everyday activities (Jain et al. [51]). The majority of the approaches have been developed since the mid-1960s. However, the

computational intensity required for these techniques initially hindered their practical use. Over the last couple of decades, the rise of microcomputers has led to the creation of many statistical software packages for various machines and platforms, specifically for clustering data. By 2020, even complex datasets with many variables and large sample sizes can yield potentially useful results with relatively little effort (Hastie et al. [43]). Clustering has had a significant impact on many scientific fields. For example, the classification of chemical elements in Mendeleev's periodic table in the 1860s was a landmark application, crucial for understanding atomic structure. In astronomy, the classification of stars into dwarfs and giants using the Hertzsprung-Russell diagram of temperature versus luminosity significantly influenced the development of the theory of the universe's creation (Kaufman and Rousseeuw [54], Everitt et al. [37]).

Clustering methods are a major branch of modern data analysis and have applications across all scientific disciplines for exploring and interpreting large, complex datasets. Grouping data can reveal hidden patterns and relationships that researchers might not otherwise discover. This is why these techniques are highly effective in many fields such as healthcare, social and economic sciences, and ecology, as they inform researchers about the dynamics within their objects of study, enabling wiser decisions. Furthermore, they can enhance the performance of machine learning models by simplifying the data and reducing noise (Berkhin [27]).

There are various clustering techniques, each offering unique advantages and limitations. Understanding these aspects may help in choosing the most appropriate method for specific data analysis tasks. In Section 3.2, we briefly document several techniques to provide a broader context, but these techniques will not be implemented here. This thesis focuses on the k-means clustering algorithm (also known as the MacQueen algorithm) due to its relevance to the application of this research (see Chapter 4).

K-means partitions data into k clusters by assigning each data point to the cluster whose mean is closest. It is frequently selected for its simplicity and computational

efficiency, particularly in handling large datasets (Koutras [13], Hastie et al. [43]). Furthermore, the technique is computationally fast and typically converges. (terminates) after a few iterations, often reaching the final solution early in the process, with only minimal adjustments needed afterward. This efficiency makes it particularly suitable for big data applications, as it requires minimal memory and computational power. The algorithm's ability to quickly form clusters, often resulting in groups of approximately equal size, enhances its utility in large-scale analyses (Koutras [13]). Moreover, k-means is versatile and applicable across diverse domains. Its methodology is straightforward, as outlined by Bishop [29]. Below, we outline the steps for its implementation.

Algorithmic Procedure I

Step I (Initialisation): The first step is initialisation, where k initial centroids μ_k are chosen. These centroids can be selected randomly or by using techniques like k-means++ (Arthur and Vassilvitskii [24]) to improve convergence. In the case of random initialisation, each centroid μ_k is chosen from the data points x_i

$$\mu_k = x_i, \qquad \text{for } k = 1,2,\dots,K.$$

Step II (Assignment): In the assignment step, each data point x_i is assigned to the nearest centroid. The assignment is based on the Euclidean distance between the data point and each centroid. The data point x_i is assigned to the cluster whose centroid μ_k is closest

$$c_i = argmin_k \|x_i - \mu_k\|^2$$

Here, c_I is the index of the centroid closest to x_I and $\left|\left|x_i - \mu_k\right|\right|$ denotes the Euclidian distance.

Step III (Update): The update step involves recalculating the centroids. Each centroid μ_k is updated by taking the mean of all data points assigned to that cluster

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_i} x_i$$

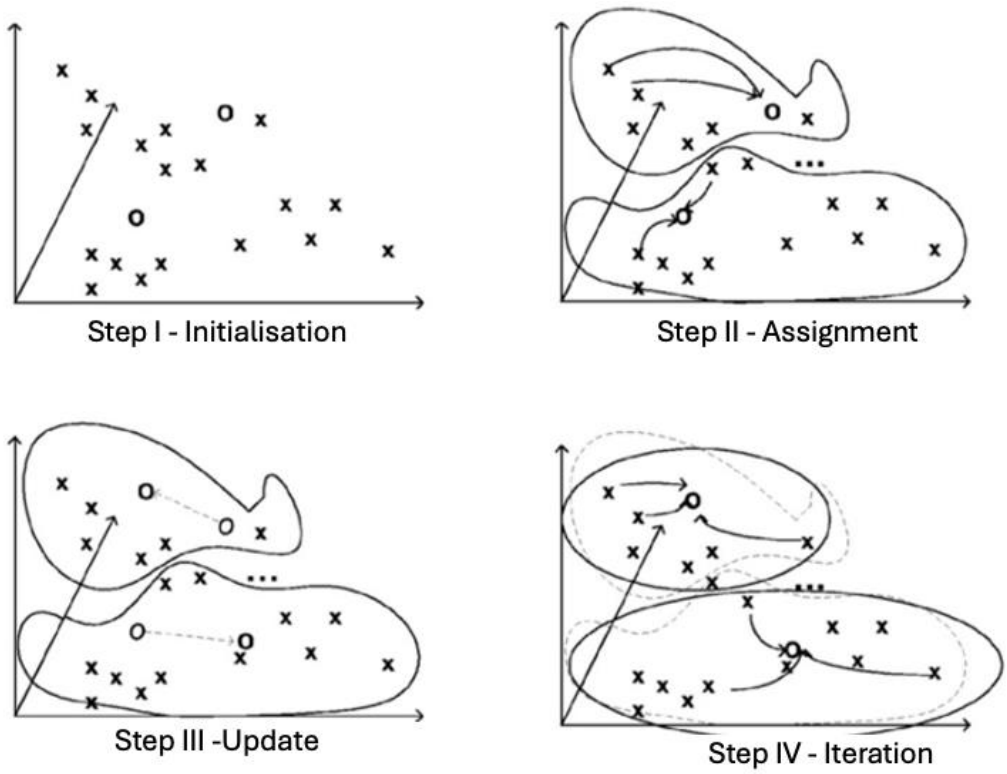
where C_k is the set of points assigned to the k-th cluster, and $|C_k|$ is the number of points in cluster k.

Step IV (Iteration): The assignment and update steps are repeated iteratively. Specifically, the assignment and update steps are alternated until the centroids no longer change significantly or a predefined number of iterations is reached (See Fig 3.1). Convergence is typically assessed by checking if the changes in centroids positions are below a certain cutoff ε, i.e.,

$$\left|\left|\mu_k^{t+1} - \mu_k^t\right|\right| < \varepsilon, \qquad for\ all\ k,$$

where μ_k^t is the centroid of cluster k at iteration t, and ε is a positive cutoff point.

Fig 3.1: Visual representation of the Algorithmic Procedure I for the implementation of k-means.
Retrieved from Koutras [13]

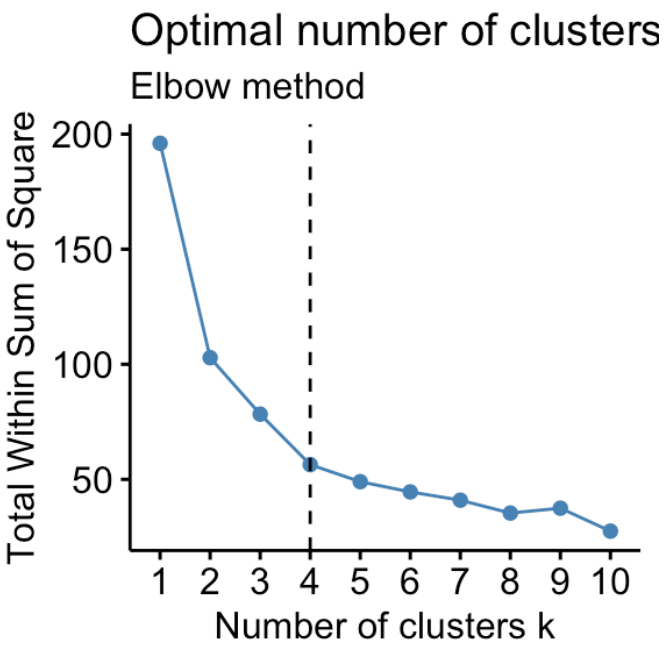


While k-means clustering offers numerous advantages, it also presents certain limitations. One of the major challenges is determining the optimal number of clusters. Deciding the best number of clusters is not straightforward, and common approaches such as the Scree Test, Silhouette Analysis, or Gap Statistics are often employed for this purpose (Kaufman and Rousseeuw [54]).

One commonly used technique is the Scree Test (also known as elbow method), which involves plotting the within-cluster sum of squares against the number of clusters and identifying the point where the rate of decrease levels off (known as the “elbow” in the curve). This point indicates the optimal *k* for testing. A visual representation of the technique can be seen in Fig. 3.2, which illustrates the results of the elbow method, where the total within-cluster sum of squares is plotted against the number of clusters.

The “elbow” of the curve, where the rate of decrease significantly slows, can be seen at $k = 4$, indicating that three clusters are optimal in this example.

Fig 3.2: Scree Test portraying the optimal number of clusters to be kept (4). Retrieved from Kassambara [12]



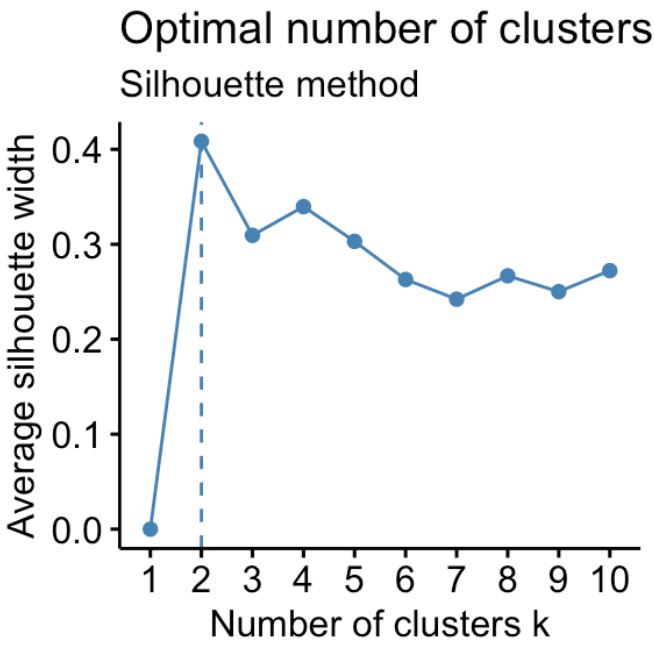
On the other hand, Silhouette Analysis measures how well each data point fits within its assigned cluster compared to other clusters. The silhouette score for each point is calculated by taking the difference between the average distance to points in its own cluster and the average distance to points in the nearest cluster, normalised by the maximum of the two. The score ranges from -1 to 1, where

- a score close to 1 indicates that the point is well-clustered and far from the neighbouring clusters,
- a score near 0 suggests that the point is on or near the boundary between clusters, and
- a negative score indicates that the point might be assigned to the wrong cluster.

Higher average silhouette scores across all data points suggest better-defined and more distinct clusters. This can be used as a metric for evaluating the quality of clustering results and selecting the optimal number of clusters. Fig. 3.3 illustrates an example of

a Silhouette Plot, showing how the silhouette score changes with different numbers of clusters. The optimal number of clusters is often where the silhouette score is highest, as seen in this example where k =2 clusters yield the highest average silhouette score.

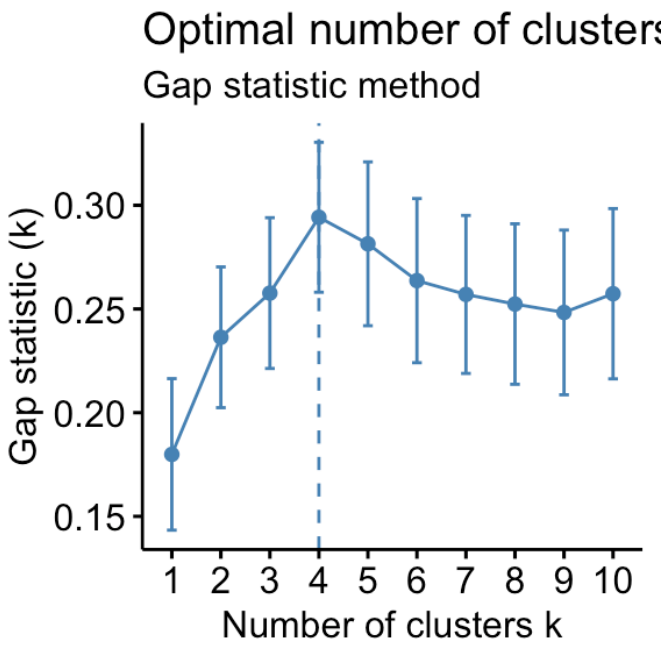
Fig 3.3: Sillouete Plot portraying the optimal number of clusters to be kept (2). Retrieved from Kassambara [12]



Another approach, Gap Statistics compare the total within-cluster variation - essentially the sum of squared distances between each data point and its cluster centroid- for different numbers of clusters with the expected values under a null reference distribution. The expected values refer to the total within-cluster variation that would be observed if the data were uniformly distributed without any inherent cluster structure. The null reference distribution represents this random uniform distribution of data points. The method identifies the number of clusters that maximises the Gap Statistic, which is the difference between the observed within-cluster variation and these expected values, thereby determining the optimal clustering structure (Hastie et al. [43]). Fig. 3.4 presents a Gap Statistics plot, where the gap statistic is plotted against the number of clusters. The number of clusters at which the gap statistic is maximised, (here $k = 4$) is considered the optimal choice. This point

represents the maximum difference between the observed clustering result and what would be expected if the data had no real clusters.

Fig 3.4: Gap Statistics Plot portraying the optimal number of clusters to be kept (4) Retrieved from Kassambara [12]



Additionally, the efficiency of the k-means algorithm can depend significantly on the choice of initial centroids. Poor initialisation can lead to suboptimal clustering results, a phenomenon often referred to as the “local minimum problem”. To mitigate this, methods like k-means++ (Arthur and Vassilvitskii [24]) are employed to improve the selection of initial centroids by spreading out the initial points, thus increasing the likelihood of converging to a better solution.

The technique also assumes that clusters are spherical and of similar size. This assumption may not always hold true for real-world datasets, a fact leading to suboptimal clustering when the actual data clusters vary in shape and size. Moreover, it is susceptible to outliers, which can distort the placement of centroids and affect the overall clustering outcome. Outliers, being distant from other data points, can disproportionately influence the centroid calculations, leading to skewed clusters.

Addressing these issues often requires preprocessing steps such as outlier detection and removal, or using more robust clustering methods that are less sensitive to outliers.

In the context of our application, k-means clustering will be used to extract and group district data. The goal is to cluster districts based on various parameters to uncover similarities and patterns among them. By clustering these districts, we aim to identify meaningful patterns and similarities across different attributes. This understanding is crucial for making decisions and optimising resource allocation within the city of Madrid. Understanding the application of this technique to specific datasets may reveal hidden patterns and structures that might otherwise go unnoticed. This study aims to illustrate practical applications of the technique in data analysis, demonstrating how analysts can effectively leverage it to enhance their analytical processes, drawing upon foundational research.

Summarising, k-means is a non-hierarchical clustering method. This method is iterative and based on the principle of the centre of the group, known as the centroid. The centroid represents the mean value of each variable for a single group. During clustering, each observation is placed in a group based on its proximity to the centroids of all groups. Essentially, for each observation, we calculate its distance from each centroid and assign it to the group with the minimum centroid distance. The main difference between these methods lies in when and how they update the centroids and reassign each observation. Euclidean distance is the standard and often ideal approach for such calculations, though other distance metrics can also be effective. The power of these methods lies in the continual iterative updating of centroids as observations are assigned to groups. This iterative process refines the groups over time until the clustering stabilises, resulting in no further changes. This approach efficiently groups data points into clusters, facilitating the identification of patterns and structures within large datasets. Non-hierarchical methods are widely used in various fields, from clinical research to climatology, due to their simplicity and effectiveness.

3.2 Alternative Clustering Techniques

In addition to the techniques discussed in detail above, there exists a variety of clustering techniques, each of which has unique advantages and limitations. Understanding these can help in choosing the most appropriate for specific data analysis tasks. Below, we briefly document several alternative techniques in order to inform the reader of their existence and provide a broader context; however, they will not be implemented in the application of this thesis.

- **Hierarchical Clustering:** Hierarchical clustering builds a hierarchy of clusters by either iteratively merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive). This method does not require the number of clusters to be specified and generates a dendrogram to visualise the cluster hierarchy. However, it may be computationally intensive for large datasets and is sensitive to noise or outliers (Nielsen [63]).

- **Agglomerative Clustering:** A specific type of hierarchical clustering, agglomerative clustering starts with each data point as its own cluster and merges the closest pairs of clusters iteratively. It is simple to understand and implement and provides a comprehensive cluster hierarchy. Nonetheless, it is computationally intensive for large datasets and sensitive to the choice of distance metric (Kononenko and Kukar [55]).

- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** DBSCAN groups points that are close to each other based on a distance metric and marks points in low-density regions as outliers. This method can find clusters of arbitrary shapes and is robust to noise. However, it requires careful tuning of its parameters and may struggle with clusters of varying densities (Ester et al. [35]).

- **Gaussian Mixture Models (GMM):** GMM assumes that data points are generated from a mixture of several Gaussian distributions with unknown parameters. It provides a probabilistic clustering approach that can capture more complex cluster shapes. However, GMM requires specifying the number of clusters and can be computationally expensive (Bishop [29]).

- **Spectral Clustering:** Spectral clustering uses the eigenvalues of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. It can handle non-convex clusters and is often more effective for data that is not well separated. However, it is computationally expensive and sensitive to the choice of similarity matrix (Von Luxburg [76]).

- **Ordering Points To Identify the Clustering Structure (OPTICS):** OPTICS is an extension of DBSCAN that creates an augmented ordering of the database, representing its density-based clustering structure. This method can identify clusters of varying densities. However, it is more complex and computationally intensive than DBSCAN (Ankerst et al. [23]).

- **Mean Shift:** Mean Shift iteratively shifts each data point towards the mode (highest density point) of its neighborhood until convergence. It does not require specifying the number of clusters and can find arbitrarily shaped clusters. However, it is computationally expensive and requires careful selection of the bandwidth parameter (Cheng [32]).

- **Affinity Propagation:** Affinity Propagation uses a message-passing algorithm between data points to identify exemplars that best represent the clusters. It does not require specifying the number of clusters and can identify clusters of varying shapes and sizes. However, it is computationally expensive and may produce many clusters if not tuned properly (Frey and Dueck [39]).

The choice of the appropriate method is determined by the properties of the dataset at hand and the research objectives of the clustering process. The existence of many clustering techniques assists data science researchers in addressing high-dimensional and complex data problems, enabling them to make the best use of their insights.

CHAPTER 4

Multivariate Analysis of Risk Factors of the COVID-19 Pandemic in the Community of Madrid, Spain

4.1 Introducing the Subject: An Overview

In late 2019, the first cases of coronavirus emerged in China's Wuhan Province. By January 2020, a new strain, SARS-CoV-2, was confirmed, sparking a global pandemic that impacted countries unevenly (WHO [19]). While some, like Spain and Italy, faced significant challenges, others were hardly affected. This discrepancy has led to increased scientific interest in understanding why certain societies are more susceptible to COVID-19. Researchers have explored various environmental factors, such as weather conditions and pollution levels, to determine their role in the virus's spread (Shakil et al. [71], Briz-Redón and Serrano-Aroca [30]). Despite numerous studies, reaching a scientific consensus has proven difficult due to several reasons. Differences in variable measurement, the selection of variables, the analytical methods applied, and spatiotemporal considerations are some of the challenges (Haug et al. [44]). This has prompted the need for more comprehensive research to overcome these issues.

To address these gaps, Pérez-Segura et al. [66] considered a multivariate approach. They considered a wide range of factors such as socioeconomic conditions and atmospheric pollution by focusing on Madrid’s community during the pandemic's initial wave. By employing advanced statistical methods, the study aimed on providing

deeper insights into the complex interplay between environmental factors and COVID-19 transmission. Thus, rather than seeking universal conclusions, they attempted to uncover localised vulnerabilities and identify practical strategies to enhance societal resilience. By understanding the unique challenges faced by different communities, their paper contributes to more effective disaster preparedness efforts in the case of similar events in the future.

4.2 The primary aim

The primary goal of the study [66] is to investigate how environmental factors affect COVID -19 infections in the Madrid community. To achieve this, the authors employed a method involving multiple variables, mainly focusing on climatological, pollution, and social aspects. Their approach can be characterised as a three-step algorithmic process exploiting multivariate techniques. Initially, potential risk factors were identified and selected based on existing literature. Then, a feature extraction DRT, namely PCA, was applied to reduce the number of variables while retaining most of the variability of the original variables. Finally, a k-means clustering analysis was conducted to categorise the territory based on these new risk factors, followed by statistical methods to characterise each cluster. Linear regression models were employed to understand the influence of each component on the total number of cases for the entire territory and for each cluster.

The data collection, pre-processing, cleaning, and analysis were conducted based on the procedures outlined in the Pérez-Segura et al. [66]. The statistical analysis was performed using *R version 4.4.1* along with several libraries, including "*sf*", "*tidyverse*", "*psych*", "*extrafont*", "*cluster*", "*sandwich*", "*lmtest*", "*reshape2*", "*factoextra*", "*agricolae*", "*ggspatial*", "*nortest*", and "*pls*". This thesis does not aim to challenge the established procedures and findings of the paper. Instead, it seeks to explore how the statistical power of the analysis can be enhanced through alternative DRTs.

A deeper dive into the data used shows that the total number of municipalities in the community of Madrid was considered, excluding the municipality of Madrid itself and

instead utilising its 21 districts for better comparability. Thus, a total of 199 districts and municipalities were included, with data sourced from publicly available databases. The analysis focused on the first-wave data from March to May 2020, when the first lockdown was imposed in Spain. Data for the target variable, i.e., the total accumulated infections (TAI) in the Madrid community, were obtained from the open dataset "COVID 19 -TIA" (Datos Abiertos Comunidad de Madrid [7]) by the municipalities and districts of Madrid. Absolute values were used instead of ratios per 100,000 population to account for population differences. To address the missing data (~21%) caused by confidentiality –where cases with fewer than 6 instances were replaced with NA– a constant value of 5 was used for imputation. This simple and consistent approach helps avoid distorting the data while allowing the analysis to proceed efficiently. Additionally, a Box-Cox logarithmic transformation was applied to further normalize the data and improve its suitability for analysis.

The independent variables used in the analysis are summarised in Table 4.1, with population size included as a control variable in the regression models due to its influence on the total number of cases. These variables were selected based on studies in the literature that support their impact on COVID-19 transmission. They are categorised into three dimensions (groups): Socioeconomic (Section 4.2.1), Pollution (Section 4.2.2), and Climatological (Section 4.2.3). Each dimension addresses different but interconnected factors related to COVID-19 infection and transmission, offering a comprehensive approach to understanding the pandemic.

4.2.1 The socioeconomic dimension

The socioeconomic aspect of the study focuses on four key factors: two demographic elements and two economic indicators. The proportion of individuals aged 65 and older and population density constitute the demographic elements. Research has highlighted the heightened vulnerability of specific age groups, particularly the elderly, to COVID-19 (Kang and Jung [53]). This underscores the importance of understanding population density, which plays a pivotal role in disease transmission. Studies, such as Carozzi's [6], have validated a positive correlation between higher population density and

increased transmission rates. However, findings from Sun et al. [72] suggest that population density did not significantly impact infection rates during COVID-19 initial wave, though it remains relevant due to subsequent lockdown measures implemented post-surge.

TABLE 4.1

Summary of independent variables.

Variable	Measure
	Socioeconomic Dimension
Density	Population density
Age	Percentage of the population over 65 years old
Income	Per capita income by municipality from Estimate of the Municipal Gross Domestic Product
Workers	Percentage of workers supported by social security
	Pollution Dimension
CO	Carbon monoxide, micrograms per cubic meter
NO	Nitrogen monoxide, micrograms per cubic meter
NO_2	Nitrogen dioxide, micrograms per cubic meter
SO_2	Sulphur dioxide, micrograms per cubic meter
Ozone	Ozone, micrograms per cubic meter
PM2.5	Particulate matter < PM2.5
RSD	Respiratory System related Death rate
	Climatological Dimension
Temperature	Temperature April average level
Humidity	Humidity April average level
	Control Variable
Population	Number of people

Data Sources: Pérez-Segura et al. [66], Datos Abiertos Comunidad de Madrid [7], Ayuntamiento de Madrid [2], Ayuntamiento de Madrid [3], Bankinter [5], Instituto de Estadística [11], Ayuntamiento de Madrid [4], Datos Abiertos Comunidad de Madrid [8], Portal de Datos Abiertos del Ayuntamiento de Madrid [16], Instituto National de Estadística [9], AEMET [1], Instituto de Estadística [10].

The role of economic status in pandemic transmission has been evident in previous outbreaks, such as the Spanish flu and Acquired Immunodeficiency Syndrome (AIDS) (Grantz et al. [41], Prual et al. [68]). The study by Hawkins et al. [45] revealed associations between geographic incidence and socioeconomic status. Additionally, insufficient economic resources can result in limited access to hygiene facilities and substandard living conditions, including overcrowded housing, which poses challenges to implementing social distancing measures during lockdowns. Pérez-Segura et al. [66] incorporates economic factors like per capita income and the proportion of workers relying on social security. Including the percentage of workers accounts for economic disparities within regions, as average income alone may not accurately portray the quality of the labour market.

Overcrowding poses a significant challenge to maintaining social distancing measures during lockdowns, increasing the risk of infection. Economic challenges, including unemployment or limited job opportunities, may drive individuals to engage in risky behaviours to secure income. Theoretical work by Ahmed et al. [22] suggests that impoverished populations may experience poorer health outcomes due to inadequate living conditions, heightening their vulnerability to COVID-19. Therefore, including these economic factors in this study provides a more in-depth understanding of the socioeconomic dynamics at play.

4.2.2 Pollution dimension

Numerous studies have identified a relationship between air pollution and COVID-19, impacting both infection rates and mortality (Fernández et al. [38], Marquès et al. [60]). Although the exact pathways through which pollution influences these outcomes remain unclear, the consistent findings across studies indicate a genuine correlation. Examining pollution involves two different variables, each assessing pollution in diverse ways across regions. One variable involves measuring pollutants like carbon monoxide (CO), nitrogen monoxide (NO), sulphur dioxide (SO2), nitrogen dioxide (NO2), ozone, and particulate matter (PM2.5) in the Madrid community. The other examines the proportion of deaths attributed to respiratory issues -Respiratory System related Death rate (RSD), providing an indirect indicator of chronic pollution levels in

the region. The latter is reinforced by the work of Zheng et al. [81], who indicated that prolonged exposure to pollutants correlates with increased COVID-19 case numbers.

4.2.3 Climatological dimension

Temperature and humidity play crucial roles in respiratory illnesses like influenza, but their influence on COVID-19 remains uncertain. While Bashir et al. [26] noted a positive correlation between temperature and transmission, Ma et al. [58] identified a negative association. Conversely, studies such as those by Mollalo et al. [61] found no link between these factors. In this context, a literature review by Briz-Redon and Serrano-Aroca [30] reports that 33 studies indicate a negative association, while only 6 show a positive one, and 7 studies report no significant association. Similarly, ambiguity surrounds the impact of humidity. According to the same review by Briz-Redon and Serrano-Aroca [30], most studies incorporating humidity demonstrate a negative correlation, with fewer indicating a positive correlation or no correlation at all (13, 3, and 6 studies, respectively).

4.3 An alternative statistical approach

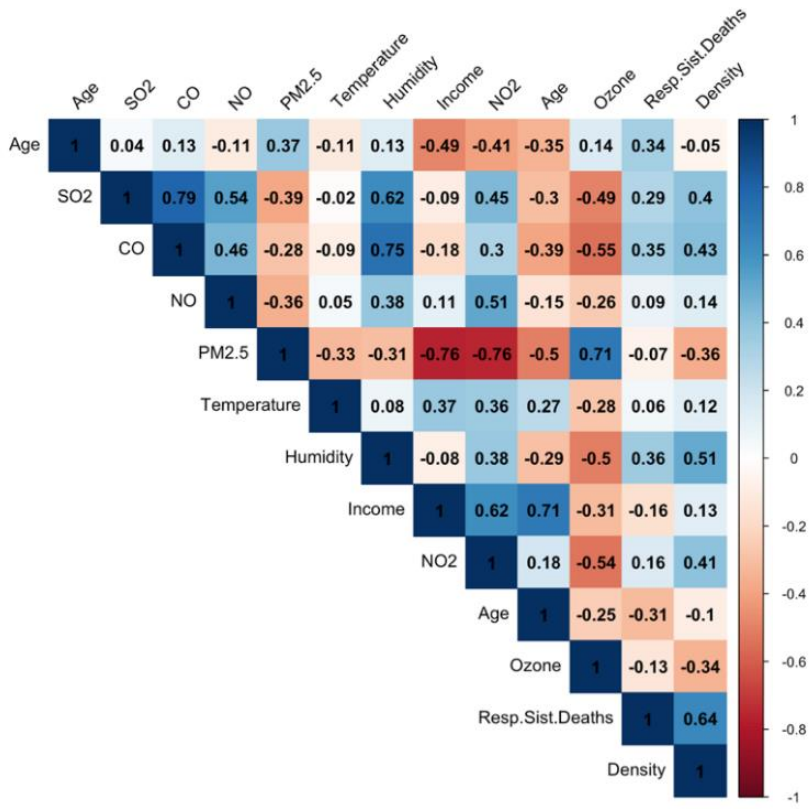
In the paper by Pérez-Segura et al. [66], the authors used PCA to analyse the data. In this thesis, we propose using PLSR as an alternative DRT. We will first replicate the analysis conducted with PCA and then perform our own analysis using PLSR. The final regression results from both methods will be compared to evaluate the differences. For consistency and comparability with the PCA approach, the latent variables derived from PLSR will be referred to as components within this application.

Our intention to investigate PCA and similar DRTs stems from an observation made during the data pre-processing phase. Significant correlations were noted in the correlation matrix of independent variables, even after scaling the data. As illustrated in Fig. 4.1, correlations often exceed 0.5 in absolute terms. Therefore, exploring DRTs

like PCA and PLSR could offer valuable insights of the data and (potentially) improve the robustness of the analysis.

FIGURE 4.1

Correlation matrix between independent variables.



Unlike PCA, which focuses on the associations within the independent matrix alone, PLSR simultaneously considers both the associations within the independent matrix (X) and those between the independent matrix and the target variable (Y). This characteristic makes PLSR particularly suitable for our analysis, aligning with our goal of predicting the target variable based on the independent variables. While it can be

seen as a supervised extension of PCA, it differs in its assumptions and methodology. By incorporating both independent and target variables, this approach potentially offers a more comprehensive understanding of the data relationships and may improve the predictive performance of our model. Thus, employing PLSR alongside PCA could enhance the robustness of our analysis and provide valuable insights for regression modelling.

It is important to stress that we maintain the clusters identified in the original paper across all DRTs to ensure comparability. The original study emphasised that the distinctions among clusters were influenced by longitudinal characteristics, so we have chosen to follow this approach to remain consistent with their findings.

In the final part of our study, we conduct regression analysis to evaluate the effectiveness of these approaches within the districts of Madrid. By adhering to the established clusters and analysing them in the context of Madrid's districts, we aim to offer a comprehensive comparison of the applied DRTs.

4.4 Results and discussion

4.4.1 Principal Component Analysis

Initially, it is worth mentioning that the assumptions necessary for applying PCA have been examined using the metrics outlined in Pérez-Segura et al. [66]. The same (or equivalent) metrics mentioned in Chapter 2.3 were used to evaluate the appropriateness of the data for PLSR. More precisely, Bartlett's test of sphericity supported the null hypothesis of variance independence, while the Kaiser-Meyer-Olkin measure revealed adequate sampling adequacy for the data in all cases. The results indicate that the data satisfy the requirements for both DRTs, confirming the validity and reliability of our analysis.

The Scree Test plot depicted in Fig. 4.2 suggests retaining 3 components. Additionally, in PCA, the cumulative variability accounts for approximately 71%, while in PLSR, it reaches about 85% when retaining the first 3 components.

Furthermore, additional metrics, such as Mean Item Complexity (MIC) and the hypothesis test for component sufficiency, yield identical results. Specifically, the MIC is calculated as 1.5, and the test for the sufficiency of three components gives a Root Mean Square Error (RMSE) of 0.07, with an empirical chi-square value of 168.

FIGURE 4.2

Scree plot for PCA models.

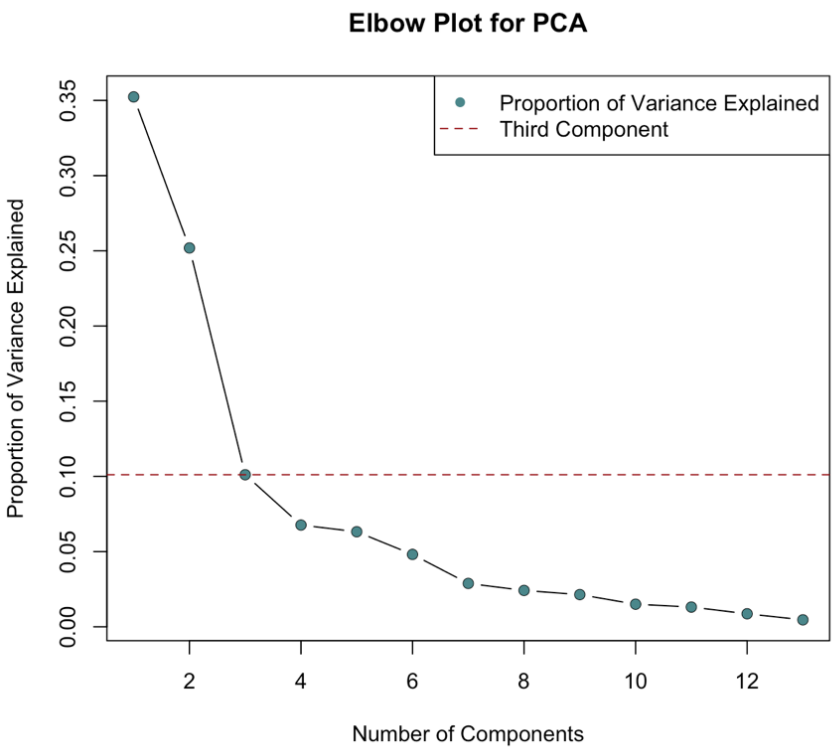


TABLE 4.2

PCA components loadings.

	Pollution and Density	Particulate Matter and Temperature	Socioeconomic
Age		-0.57	
SO2	**0.87**		
CO	**0.86**		
NO	**0.74**		
Ozone	-0.43	**-0.83**	
RSD		0.54	
Density	**0.73**		
PM2.5		**0.92**	
NO2	0.53	**0.68**	
Temperature		**0.76**	
Humidity	-0.56	-0.50	
Income			**0.87**
Workers			**0.76**

Note: Loads less than 0.4 have been suppressed. The most representative factor loadings are indicated in bold type.

Table 4.2 presents the output of PCA, as described by Pérez-Segura et al. [66]. Based on it, three components emerged from the initial 13 variables, accounting for 71% of the original variability. The components were interpreted based on factor loadings detailed in Table 4.2. Although these empirical dimensions may differ from the initially projected theoretical ones due to the lack of validated scales, they still effectively summarise the 13 risk factors. Notably, population density showed a strong association with a climatic factor, which is logical as pollution often correlates with human population density.

TABLE 4.3

PLSR components loadings.

	Air Quality and Density	Respiratory System Deaths	Environmental and Demographic Influences
Age	-0.13	-0.27	**0.50**
SO2	**0.35**	-0.13	0.42
CO	**0.32**		**0.57**
NO	0.27		
Ozone	**-0.37**	0.22	0.35
RSD	0.17	**0.83**	-0.33
Density	**0.32**		0.40
PM2.5	0.20		**-0.61**
NO2	**0.38**		-0.28
Temperature		-0.31	**-0.75**
Humidity	**-0.37**	0.12	
Income	0.14	-0.28	0.39
Workers	0.26	-0.23	0.15

Note: Loads zero or close to zero have been suppressed. The most representative factor loadings are indicated in bold type.

In Table 4.3, the components of PLSR are analysed to understand their relationships with the predictor variable. Component 1 can be labelled as “Air Quality and Density”, since pollutants such as SO2, CO, NO2, Ozone, and Humidity display the highest loadings, indicating strong associations with this component. Density in this component exhibits the same level of significance as the pollutant variables. Component 2, labelled as “Respiratory System Deaths”, is distinct as only one variable emerges as significantly more influential than others, with a loading more than twice as important compared to the others. Given the nature of this study, which aims to explore the impact of these factors during the pandemic and considering that RSD significantly increased during this period, component 2 potentially contains meaningful information. In component 3, designated as “Environmental and Demographic Influences”, temperature and PM2.5 exhibit strong negative loading values, indicating an inverse relationship, while Age and CO display positive loading values. These interpretations provide valuable insights into how the predictor variables contribute to the different components identified through the PLSR analysis.

When comparing Tables 4.2 and 4.3, several observations emerge:

- PLSR highlights Age and RSD as significant variables, unlike PCA. This difference arises from PLSR's consideration of associations within the X-matrix and between the X-matrix and Y-matrix. Age is a known factor in mortality rates, including in the case of COVID-19, while RSD are a major cause of COVID-19-related deaths.

- In contrast, PLSR does not emphasise the importance of income and workers as much as PCA approach do. This discrepancy may stem from the lack of strong evidence linking these variables to COVID-19 or other independent variables. This pandemic impacted countries with varying economic strengths. Even in PCA, income and workers show limited connections to other variables, particularly in the third component.

- Overall, while these differences exist, both DRTs generally involve variables that contribute to at least one component, validating their selection. It is important to note that these observations do not necessarily suggest PLSR is better than PCA but rather document differences in outcomes

The RMSE was calculated between the response and the three components generated by each approach to assess their predictive ability. PLSR outperformed PCA with an RMSE value of 0.87, in contrast to PCA’s value of 4.23.

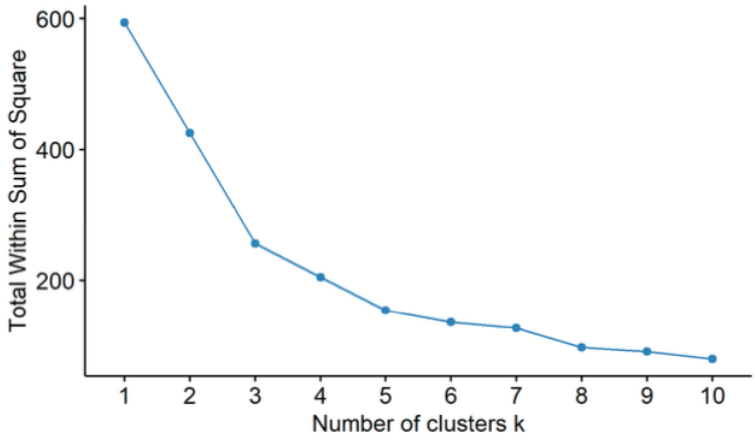
4.4.2 Cluster Analysis

For the purpose of facilitating comparative analysis and gaining a deeper understanding of the underlying mechanisms of each DRT, we employed the k-means clustering methodology as delineated in Pérez-Segura et al. [66]. In their work, they utilised the Scree Test of the within-cluster sum of squares for each number of clusters (see Fig. 4.3), resulting in the selection of three clusters.

FIGURE 4.3

Scree plot for the selection of clusters.

Source: Pérez-Segura et al. [66]



This enabled us to discern the diverse ecosystems within the community of Madrid based on the considered risk factors.

The rationale behind the selection and naming of clusters, as portrayed by Pérez-Segura et al. [66], serves as the basis for our cluster analysis methodology. The clustering described in the study aimed to identify distinct ecosystems within the community of Madrid based on various risk factors (Fig. 4.4). Upon replicating the procedure, it was observed that the clusters resulting from k-means have been slightly altered for better cluster interpretability. It was observed that the “Madrid-City” cluster, comprising the city's 21 districts, exhibited elevated levels of pollution, population density, and socioeconomic indicators compared to the other clusters (Fig. 4.5).

FIGURE 4.4

Cluster between Madrid's districts.

Source: Pérez-Segura et al. [66]

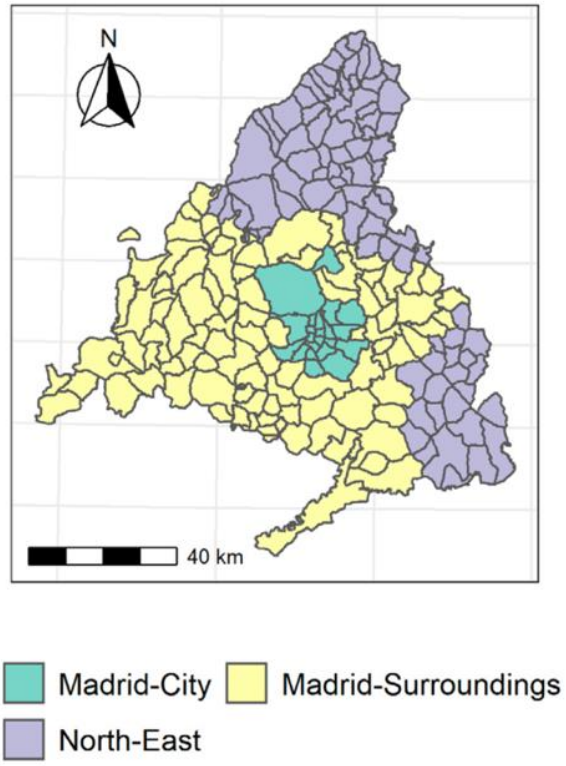
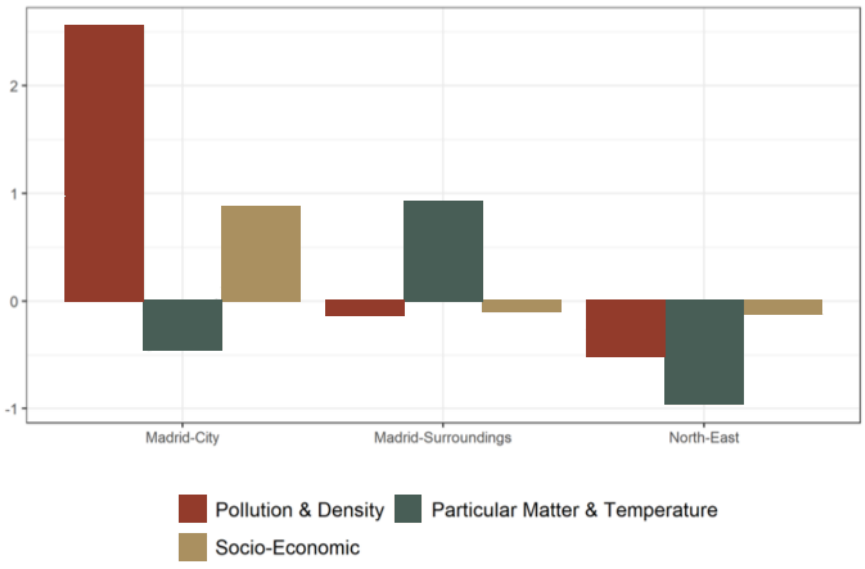


FIGURE 4.5

Components by clusters in Madrid's districts.

Source: Pérez-Segura et al. [66].



Conversely, concerning particulate matter and temperature factors, Madrid demonstrated a negative score like that of the “North-East” cluster. The “Madrid-Surroundings” cluster encompassed 96 municipalities, including densely populated areas, where particulate matter and temperature factors were notable. While pollution and population density were lower than in the “Madrid-City” cluster, they surpassed those of the “North-East” cluster. Additionally, this region exhibited the poorest socioeconomic conditions. Lastly, the “North-East” cluster consisted of 82 municipalities, predominantly rural areas with fewer than 5000 inhabitants, characterised by lower pollution levels and socioeconomic status.

Table 4.4 provides a detailed summary of each cluster's characteristics based on the average scores of the original variables. The Analysis of Variance (ANOVA) analysis revealed significant differences in means across all variables in at least one cluster. To get a better understanding, a post hoc examination (Scheffé´s test) was conducted to pinpoint the specific clusters with differences. The “Madrid-City” cluster emerged as the most polluted area in the region, followed by the “Madrid-Surroundings” cluster, while the “North-East” cluster showed lower pollution levels, something reasonable due to population density. Similarly, the socioeconomic aspect followed a similar pattern, with “Madrid-City” recording the highest scores. However, the measurements of particulate matter and temperature showed a different trend. In this case, the “North-East” cluster appeared to be the most humid, whereas the “Madrid-Surroundings” cluster experienced the highest temperatures. Interestingly, “Madrid-City” displayed the lowest scores for both variables.

TABLE 4.4

Comparison of the mean values of the original variables by cluster.

Source: Pérez-Segura et al. [66].

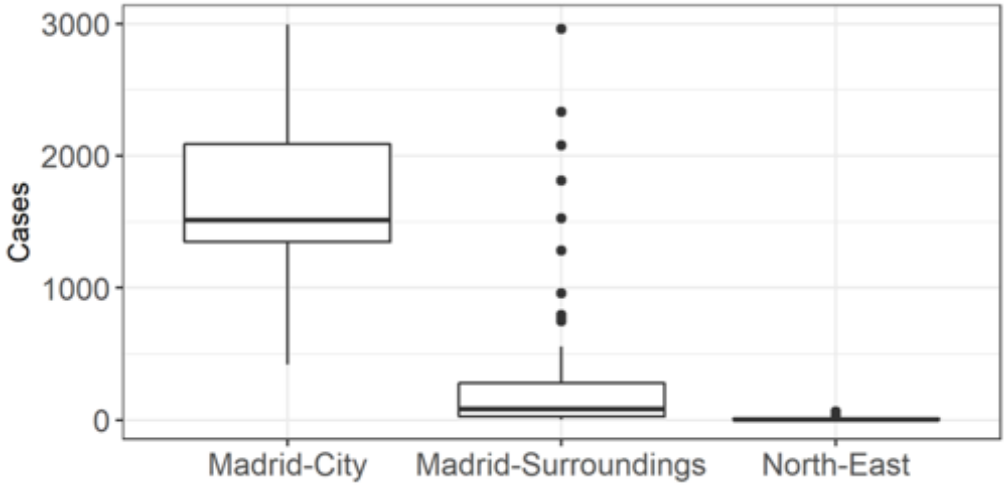
Variable	Cluster			ANOVA (p-Value)
	Madrid-City	Madrid-Surroundings	North-East	
Social Dimension				
Density	14.25	0.85	0.05	<0.01
SD	9.82	1.32	0.05	
Income	40,590.98	21,877.74	24,445.71	<0.01
SD	17,966.81	11,808.98	14,445.71	
% Workers	0.6	0.31	0.22	<0.01
SD	0.39	0.2	0.11	
Climate Dimension				
Humidity	65.96	71.16	75.98	<0.01
SD	0	3.56	0.54	
Temperature	9.69	12.83	10.89	<0.01
SD	0	0.92	1.6	
Pollution Dimension				
PM2.5	4.05	6.66	**3.45**	<0.01
SD	0.22	0.69	1.07	
SO_2	4.33	1.29	1	<0.01
SD	1.56	0.5	0	
CO	50.4	0.87	0.45	<0.01
SD	12.33	0.69	0.19	
NO	4.15	1.82	1	<0.01
SD	4.57	0.9	0.01	
NO_2	17.01	14.04	1.68	<0.01
SD	5.65	7.25	0.63	
Ozone	53.08	53.9	80.62	<0.01
SD	4.61	3.84	4.24	

Dark gray, light gray and white are used to indicate that there are statistical differences between means. The darkest shade refers to the highest mean value, the light gray to the mean value and the white to the lowest mean value. When there are only statistically significant differences between two groups, only dark gray and light gray are used to distinguish the groups. Differences in means were tested by Scheffé´s methods at a significance level of 0.05.

FIGURE 4.6

Boxplot of COVID-19 cases by cluster.

Source: Pérez-Segura et al. [66].



The diagram represents Q_2 (median value, line inside the box), the Q1 (lower boundary of the box), the Q3 (upper boundary of the box), the maximum and the minimum (vertical line) and the outliers (dots).

Fig. 4.6 shows how COVID-19 spread across each cluster. It is evident that the "Madrid-City" area had the highest number of infections. Additionally, the "Madrid-Surroundings" cluster had many outliers, mainly made up of densely populated cities like Leganes (2963 cases), Alcala de Henares (2331 cases), and Mostoles (2079 cases), among others. Statistical tests confirmed significant differences in average infection levels among the three areas. These results highlight how risk factors can outline areas with different infection rates, although they do not pinpoint the exact influence of each factor.

4.4.3 Regression Analysis

Tables 4.5 and 4.6 present the outcomes of the regression models conducted for the entire territory ("C. Madrid") and each cluster using PCA and PLSR.

TABLE 4.5

Regression models with PCA components.

	C. Madrid			North-East			Madrid-Surroundings			Madrid-City		
	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value
(Intercept)	3.17*** (0.08)	37.80	<0.01	1.17** (0.37)	3.13	<0.01	3.69*** (0.35)	10.52	<0.01	6.1*** (0.21)	28.80	<0.01
1st component	0.48*** (0.08)	5.98	<0.01	-0.62 -0.49	-1.27	0.20	0.77* (0.34)	2.25	0.02	0.00 (0.07)	0.14	0.89
2nd component	0.75*** (0.06)	12.36	<0.01	-0.10 (0.12)	-0.87	0.38	0.07 (0.40)	0.17	0.86	-0.13 (0.27)	-0.48	0.63
3rd component	0.16* (0.06)	2.31	0.02	-0.15 (0.12)	-1.22	0.22	0.25 (0.16)	1.54	0.12	0.02 (0.03)	0.75	0.45
Population	0.00*** (0.00)	11.66	<0.01	0.00*** (0.00)	17.06	<0.01	0.00*** (0.00)	7.88	<0.01	0.00*** (0.00)	7.58	<0.01
Adjusted R-squared		0.82			0.83			0.66			0.81	
p-values		<0.01			<0.01			<0.01			<0.01	
white test (p-values)		<0.01			0.03			0.05			0.34	
Model Additional Information												
AIC		504.51			48.65			268.77			-2.93	
BIC		524.27			63.08			284.16			3.33	
Loglik		-246.25			-18.32			-128.38			7.46	

*** p < 0, ** p < 0.01, * p < 0.05.

The results for PCA are provided in Table 4.5. Overall, the regression results suggest that the significance and importance of the components vary across regions. While "C. Madrid" and "Madrid-Surroundings" models highlight the relevance of specific components, the "North-East" and "Madrid-City" models indicate that these components may not be crucial predictors in those areas. Across all models, the population variable consistently shows high significance, underscoring its universal importance as a factor in the analysis.

More analytically, in the "C.Madrid" model, all components are significant and positively linked. The second component is the most influential, with a strong coefficient and highly significant p-value, suggesting it plays a crucial role in the model. The first component also contributes significantly, though to a lesser extent, while the third component, despite being the weakest among the three, still shows a statistically significant effect. The latter suggests that higher scores on the variables, including socioeconomic status, increase the risk of contracting COVID-19. This finding goes against existing theories on socioeconomic status and COVID-19 transmission. However, it is essential to note that we are analysing variables at a municipal level, thus intra-municipal variability is not considered. Therefore, our results suggest that municipalities with higher socioeconomic status (more workers

and higher incomes) might have more COVID-19 cases, although there could still be economic segregation within these municipalities affecting infection risk.

The partial models yield limited information due to the lack of statistical significance in the parameters. More precisely, the “C.Madrid” model, the “North-East” and “Madrid-City” models provide limited insights as none of the components reach statistical significance. The coefficients for the components are not only non-significant but also small in magnitude, indicating that these factors do not substantially contribute to explaining the variation in the dependent variable within this region. Despite this, the population variable, consistent with other models, shows high significance, reinforcing its importance across different regions. On the other hand, the “Madrid-Surroundings” model presents a somewhat mixed picture. The first component is statistically significant, suggesting it has a moderate impact on the dependent variable. However, the second and third components do not achieve significance, with the third component showing borderline relevance (for significance level 10%). This implies that while the first component may be an essential predictor in this region, the other components do not offer meaningful explanatory power. Nevertheless, as in other models, the population variable remains highly significant, confirming its strong influence.

This lack of significance in the partial models can potentially be attributed to two factors: (1) the small number of observations in each cluster and (2) the homogeneous behaviour of the target variable within each cluster.

TABLE 4.6

Regression models with PLSR components.

	C. Madrid			North-East			Madrid-Surroundings			Madrid-City		
	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value	Coef. (SD)	t	p-value
(Intercept)	0.39* (0.19)	1.93	0.04	1.78*** (0.5)	3.57	<0.01	1.82** (0.60)	3.00	<0.01	6.11*** (0.43)	14.15	<0.01
1st component	0.01 (0.14)	0.20	0.89	-0.06 (0.21)	-0.29	0.77	0.22 (0.29)	0.77	0.44	0.01 (0.11)	0.14	0.88
2nd component	0.41 (0.31)	1.34	0.17	0.36 (0.43)	0.84	0.40	-3.97* (1.51)	-2.62	0.01	-0.09 (0.48)	-0.19	0.85
3rd component	0.40 (0.34)	1.19	0.23	-0.36 (0.40)	-0.90	0.37	4.44** (1.37)	3.23	0.001	0.08 (0.44)	0.18	0.85
Population	0.00** (0.00)	2.51	<0.01	0.00*** (0.00)	13.83	<0.01	0.00 (0.00)	1.35	0.17	0.00*** (0.00)	8.35	<0.01
Adjusted R-squared		86%			83%			73%			81%	
p-values		<0.01			<0.01			<0.01			<0.01	
white test (p-values)		<0.01			0.05			<0.01			0.33	
Model Additional Information												
AIC		463.97			49.55			247.92			-2.02	
BIC		483.73			63.99			263.31			4.24	
Loglik		-225.98			-18.77			-117.96			7.01	

*** p < 0, ** p < 0.01, * p < 0.05.

The results from the PLSR model for the entire territory and each cluster are presented in Table 4.6. In summary, the regression results illustrate the varying influence of the components across different regions. While the population variable consistently show significance in most models, the components themselves do not always hold statistical significance, particularly in the North-East and Madrid-City models. The Madrid-Surroundings model stands out for the significance of the second and third components, although with differing effects. These findings suggest that the explanatory power of these components is region-specific, with population being the most consistent predictor across all models.

More analytically, in the “C. Madrid” model, the intercept is marginally significant with a p-value of 0.04. However, the three components do not show statistical significance, as indicated by their high p-values (0.89, 0.17, and 0.23, respectively). This suggests that these components do not contribute meaningfully to the model in this region. Interestingly, the population variable remains significant ($p < 0.01$), consistent with previous models, indicating that population continues to play a crucial role in explaining the dependent variable in “C. Madrid”.

In the “North-East” and “Madrid-City” models, the intercept is highly significant ($p < 0.01$), showing a strong baseline effect. However, like the “C.Madrid” model, the

components do not demonstrate statistical significance, with p-values well above the conventional threshold of 0.05. This lack of significance suggests that these components do not hold substantial explanatory power in these regions. Despite this, the population variable is extremely significant ($p < 0.01$), confirming its critical importance in this model as well.

The “Madrid-Surroundings” model presents a more complex picture. The intercept is significant ($p < 0.01$), indicating a strong baseline effect. The first component is not significant ($p = 0.44$), suggesting it does not substantially influence the dependent variable in this region. However, the second component is significant ($p = 0.01$), though it has a negative coefficient, indicating an inverse relationship with the dependent variable. The third component is also significant ($p = 0.001$), with a strong positive coefficient, highlighting its importance in the model. Unlike in other regions, the population variable is not significant ($p = 0.17$), which contrasts with its significance in the other models, suggesting that population may not be as crucial in this particular region.

While some clusters showed significant associations between components and cases, others displayed less pronounced effects. Population density almost consistently emerged as a significant factor across clusters, influencing the distribution of COVID-19 cases. However, non-significant intercepts and components in some clusters suggest a need for further investigation to fully understand transmission dynamics

4.4.3.1 C.Madrid: Regression Model Comparisons

The regression model using PCA components for “C.Madrid” demonstrated an adjusted R-squared value of approximately 0.82, indicating that the model explains about 82% of the total variation of COVID-19 outcomes. All components in this model exhibited significant p-values (<0.05), suggesting their importance in explaining the variance within the dataset. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were 504 and 524, respectively, indicating the model's goodness of fit and parsimony. The White test p-value is below the conventional significance level of 0.05, suggesting potential heteroscedasticity.

In contrast, the PLSR model demonstrated superior performance with an adjusted R-squared value of 0.86, indicating that the model explains about 86% of the total variation of COVID-19 outcomes. While the overall fit of the model was better, some components showed non-significant p-values (>0.05), suggesting a potential lack of statistical significance. The AIC (464) and BIC (484) values were lower compared to the PCA model, further indicating better model fitness and parsimony. Similar to the PCA model, the PLSR model exhibited potential heteroscedasticity based on the White test p-value.

Additionally, the PCA model indicate that the pollution and climate components had a greater influence than the socioeconomic component on the overall model. In contrast, the PLSR model indicates that RSD, as well as environmental and demographic characteristics, play the most important roles, which is in accordance with many research studies. Notably, PLSR is the only method that formulates a component whose identity is solely based on RSD.

4.4.3.2 Madrid-Surroundings: Regression Model Comparisons

The PCA models for the “Madrid- Surroundings” demonstrated an adjusted R-squared value of 0.66, indicating that the model explains about 66% of the total variation of COVID-19 outcomes. Only the first component is significant (p=0.02), suggesting pollution and density factors play crucial role in this region. The AIC and BIC values were 269 and 284, respectively, indicating the model's goodness of fit and parsimony. The p-value obtained from the White test is exactly 0.05, which aligns with the conventional significance threshold, suggesting that the assumption of homoscedasticity may be questionable.

In contrast, the PLSR model demonstrated superior performance with an adjusted R-squared value of 0.73, indicating that the model explains about 73% of the total variation of COVID-19 outcomes. While the overall fit of the model was better, the second and third components showed significant p-values (<0.05), suggesting RSD, Environmental and Demographic factors play a crucial role in this region. The AIC

(248) and BIC (263) values were lower compared to the PCA model, further indicating better model fitness and parsimony. The White test p-value is below the conventional significance level of 0.05, and lower than the one in the PCA models, suggesting potential heteroscedasticity.

4.4.3.3 North-East & Madrid-City: Regression Model Comparisons

The models applied to both the “North-East” and “Madrid-City” regions showed notable similarities, with only minor, non-significant differences. This suggests a high degree of consistency in performance across all methodologies.

CHAPTER 5

Concluding Remarks & Future Expansions

The global COVID-19 outbreak has prompted scientists to investigate the factors that increase susceptibility to the virus. While most of the studies have focused on environmental factors, there remains much to learn.

In this thesis, we began our investigation by employing the same PCA as described and implemented by Pérez-Segura et al. [66] to condense thirteen identified risk factors into three distinct components, which accounted for 71% of the initial variance. These new variables included aspects such as pollution and population density, particulate matter and temperature, and socioeconomic status. Subsequently, these variables were used to map the various ecosystems within the region according to these risk dimensions. A regression analysis was then conducted using the formulated components as explanatory variables to examine potential differences in COVID-19 infection rates across the identified clusters. Additionally, we employed an alternative approach by using PLSR as DRTs, instead of PCA. Doing so, we aimed to explore whether this alternative method could yield improved outcomes compared to the original approach. Each DRTs provided valuable insights into COVID-19 risk factors. While PCA model demonstrated statistical significance for most components, the PLSR model exhibited superior overall fitness and parsimony. For two clusters ("C. Madrid" and "Madrid-Surroundings"), PLSR yielded significantly higher adjusted R^2 and Log-likelihood values, along with lower AIC and BIC values, compared to PCA. For the remaining two clusters ("North-East" and "Madrid-City"), both DRTs performed similarly with no significant differences. However, the presence of non-significant p-values across all models suggests caution in interpreting certain risk factors.

The PLSR model can be considered as the preferred choice for analysing COVID-19 risk factors due to its superior model fit and parsimony. Despite encountering some non-significant components, the overall performance of the PLSR model surpasses that of PCA. Further investigation is warranted to validate the significance of individual risk factors identified by both approaches. As discussed in Chapter 3 and inside the work by Pérez-Segura et al. [66], the cluster models do not provide substantial information due to the non-significance of the parameters and the observed low sampling power. Thus, the lack of significance may be attributed to these issues. Overall, PLSR aligns with many previous studies, reinforcing the rationale behind the findings (Wehrens [78]).

A key finding of our study is that PLSR identified Respiratory System Deaths (RSD) as a significant factor in the formulation of its components. Specifically, the second component was heavily influenced by this variable, as shown in Table 4.3. In contrast, PCA model did not highlight RSD. This discrepancy, combined with the established strong association between this variable and COVID-19, underscores the importance of PLSR for modelling tasks related to this pandemic. This identification suggests PLSR's enhanced suitability for such tasks. Furthermore, this finding calls for further investigation into which dimensionality reduction and clustering techniques are most effective for modelling complex phenomena like pandemics.

Potential future work could involve implementing the DRTs mentioned in Section 2.5 and comparing the variables that emerge as key factors. For instance, Sparse PCA can yield even more informative outputs by introducing sparsity in the components. Furthermore, future work could include the utilisation of Gaussian Mixture Models (GMM) as an alternative to k-means clustering. This technique provides a more flexible approach by allowing clusters to have different shapes and sizes, accommodating data that does not conform to the rigid boundaries imposed by k-means. Implementing such models could lead to improved cluster identification, particularly for datasets with overlapping clusters or non-linear associations.

Comparing the results of GMM with k-means could offer deeper insights into the structure of the data and provide a more comprehensive understanding of key variables.

References

Online Sources

1. AEMET. (n.d.). OpenData. https://opendata.aemet.es/centrodedescargas/inicio (accessed on 8 June 2021).
2. Ayuntamiento de Madrid. (n.d.). Datos anuales. https://www.madrid.es/portales/munimadrid/es/Inicio/El–Ayuntamiento/Estadistica/Areas–de–informacion–estadistica/Mercado–de–trabajo/Afiliaciones–a–la–Seguridad–Social/Datosanuales/?vgnextfmt=default&vgnextoid=9c7d29bc75a80510VgnVCM1000000b205a0aRCRD&vgnextchannel=f26a62a006986210VgnVCM2000000c205a0aRCRD (accessed on 8 June 2021).
3. Ayuntamiento de Madrid. (n.d.). Distritos en cifras (Información de Distritos). https://www.madrid.es/portales/munimadrid/es/Inicio/El–Ayuntamiento/Estadistica/Distritos–en–cifras/Distritos–en–cifras–Informacion–de–Distritos/?vgnextfmt=default&vgnextoid=74b33ece5284c310VgnVCM1000000b205a0aRCRD&vgnextchannel=27002d05cb71b310VgnVCM1000000b205a0aRCRD (accessed on 8 June 2021).
4. Ayuntamiento de Madrid. (n.d.). Estimación del Producto Interior Bruto Municipal. http://www.madrid.org/iestadis/fijas/estructu/economicas/contabilidad/epibmb15tab.htm (accessed on 26 May 2021).
5. Bankinter. (n.d.). Los barrios de Madrid con mayor renta media. https://www.bankinter.com/blog/finanzaspersonales/barrios–madrid–mayor–renta (accessed on 8 June 2021).
6. Carozzi, F. (2020). Urban density and COVID–19, *Social Science Research Network*. https://papers.ssrn.com/abstract=3643204 (accessed on 3 February 2021).
7. Datos Abiertos Comunidad de Madrid. (n.d.). COVID–19–TIA por municipios y distritos de Madrid. https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos (accessed on 26 May 2021).
8. Datos Abiertos Comunidad de Madrid. (n.d.). Municipios de la Comunidad de Madrid. https://datos.comunidad.madrid/catalogo/dataset/municipio_comunidad_madrid (accessed on 8 June 2021).
9. Instituto Nacional de Estadística. (n.d.). Encuesta nacional de salud. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=resultados&idp=1254735573175 (accessed on 8 June 2021).
10. Instituto de Estadística. (n.d.). Padrón anual: Resultados definitivos. https://www.madrid.org/iestadis/fijas/estructu/demograficas/padron/estructupopc.htm (accessed on 8 June 2021).
11. Instituto de Estadística. (n.d.). Trabajadores afiliados a la Seguridad Social en alta que trabajan en la Comunidad de Madrid. https://www.madrid.org/iestadis/fijas/estructu/sociales/iss20.htm (accessed on 8 June 2021).
12. Kassambara, A. (n.d.). Determining the optimal number of clusters: 3 must–know methods, *Datanovia*. Retrieved June 8, 2024, from https://www.datanovia.com/en/lessons/determining–the–optimal–number–of–clusters–3–must–know–methods/.

13. Koutras, M. (2022). *Applied multivariate analysis: Lecture notes for the postgraduate course of the Department of Statistics and Insurance Science*, University of Piraeus.
14. Laerd Statistics. (2024). Principal component analysis (PCA) using SPSS statistics, *Laerd Statistics*. Retrieved from https://statistics.laerd.com/spss–tutorials/principal–components–analysis–pca–using–spss–statistics.php.
15. Ng, K. S. (2013). A simple explanation of partial least squares, *The Australian National University*.
16. Portal de Datos Abiertos del Ayuntamiento de Madrid. (n.d.). Calidad del aire. Datos diarios años 2001 a 2021. https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default (accessed on 8 June 2021).
17. Roughgarden, T., and Valiant, G. (2015). CS168: The modern algorithmic toolbox lecture #8: PCA and the power iteration method.
18. Wikipedia contributors. (n.d.). Scree plot, In *Wikipedia, The Free Encyclopedia*. Retrieved May 18, 2024, from https://en.wikipedia.org/wiki/Scree_plot.
19. World Health Organization (WHO). (2020). Novel coronavirus—China, https://www.who.int/csr/don/12–january–2020–novel–coronavirus–china/en/ (accessed on 26 May 2021).

Books and Scientific Papers

20. Abdi, H. and Williams, L. J. (2010). Principal component analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(4), 433–459.
21. Aggarwal, C. C. (2015). *Data mining: The textbook*, Springer.
22. Ahmed, F., Ahmed, N., Pissarides, C. and Stiglitz, J. (2020). Why inequality could spread COVID–19, *Lancet Public Health*, **5**, e240.
23. Ankerst, M., Breunig, M. M., Kriegel, H. P. and Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, ACM Press, 49–60.
24. Arthur, D. and Vassilvitskii, S. (2007). k–means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM–SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, 1027–1035.
25. Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., et al. (2021). Efficacy and safety of the mRNA–1273 SARS–CoV–2 vaccine, *New England Journal of Medicine*, **384**(5), 403–416.
26. Bashir, M. F., Ma, B., Komal, B., Bashir, M. A., Tan, D., and Bashir, M. (2020). Correlation between climate indicators and COVID–19 pandemic in New York, USA, *Science of The Total Environment*, **728**, 138835.
27. Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, Springer, 25–71.
28. Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2022). Solving large-scale sparse PCA to certifiable (near) optimality, *Journal of Machine Learning Research*, **23**, 13:1–13:35.
29. Bishop, C. M. (2006). *Pfattern recognition and machine learning*, Springer.
30. Briz–Redón, Á. and Serrano–Aroca, Á. (2020). The effect of climate on the spread of the COVID–19 pandemic: A review of findings and statistical and modelling techniques, *Progress in Physical Geography: Earth and Environment*, **44**, 591–604.

31. Cattell, R. B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research*, **4**(1), 245–276.
32. Cheng, Y. (1995). Mean shift, mode seeking, and clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(8), 790–799.
33. Comon, P. (1994). Independent component analysis: A new concept? *Signal Processing*, **36**(3), 287–314.
34. De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **18**(3), 251–263.
35. Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A density–based algorithm for discovering clusters in large spatial databases with noise, In E. Simoudis, J. Han, and U. M. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD–96)*, AAAI Press, 226–231.
36. Evans, R. A., McAuley, H., Harrison, E. M., Shikotra, A., Singapuri, A., Sereno, M., et al. (2021). Physical, cognitive, and mental health impacts of COVID–19 after hospitalisation (PHOSP–COVID): A UK multicentre, prospective cohort study, *The Lancet Respiratory Medicine*, **9**(12), 1275–1287.
37. Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster analysis*, John Wiley & Sons.
38. Fernández, D., Giné–Vázquez, I., Liu, I., Yucel, R., Ruscone, M. N., Morena, M., García, V. G., Haro, J. M., Pan, W., and Tyrovolas, S. (2021). Are environmental pollution and biodiversity levels associated with the spread and mortality of COVID–19? A four–month global analysis, *Environmental Pollution*, **271**, 116326.
39. Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points, *Science*, **315**(5814), 972–976.
40. Golub, G. H., and Van der Vorst, H. A. (2000). Eigenvalue computation in the 20th century, *Journal of Computational and Applied Mathematics*, **123**(1–2), 35–65.
41. Grantz, K. H., Rane, M. S., Salje, H., Glass, G. E., Schachterle, S. E., and Cummings, D. A. T. (2016). Disparities in influenza mortality and transmission related to sociodemographic factors within Chicago in the pandemic of 1918, *Proceedings of the National Academy of Sciences*, **113**, 13839–13844.
42. Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., et al. (2021). A global panel database of pandemic policies (Oxford COVID–19 Government Response Tracker), *Nature Human Behaviour*, **5**(4), 529–538.
43. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.), Springer.
44. Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars–Larrive, A., Loreto, V., et al. (2020). Ranking the effectiveness of worldwide COVID–19 government interventions, *Nature Human Behaviour*, **4**(12), 1303–1312.
45. Hawkins, R., Charles, E., and Mehaffey, J. (2020). Socio–economic status and COVID–19–related cases and fatalities, *Public Health*, **189**, 129–134.
46. Hoffmann, L. D., and Bradley, G. L. (2004). *Calculus for business, economics, and the social and life sciences* (8th ed.), 575–588.
47. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**, 417–441, 498–520.
48. Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**(3/4), 321–377.
49. Jolliffe, I. (2002). *Principal components analysis* (2nd ed.), Springer.

50. Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: A review and recent developments, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**(2065), 20150202.
51. Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review, *ACM Computing Surveys*, **31**(3), 264–323.
52. Kaiser, H. F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement*, **20**, 141–151.
53. Kang, S.–J., and Jung, S. I. (2020). Age–related morbidity and mortality among patients with COVID–19, *Infect Chemother*, **52**, 154–164.
54. Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons.
55. Kononenko, I. and Kukar, M. (2007). *Machine learning and data mining*, Elsevier.
56. Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks, *AIChE Journal*, **37**(2), 233–243.
57. Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *Journal of Research of the National Bureau of Standards*, **45**(4), 255–282.
58. Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., and Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID–19 in Wuhan, China, *Science of The Total Environment*, **724**, 138226.
59. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.
60. Marquès, M., Rovira, J., Nadal, M., and Domingo, J. L. (2021). Effects of air pollution on the potential transmission and mortality of COVID–19: A preliminary case–study in Tarragona Province (Catalonia, Spain), *Environmental Research*, **192**, 110315.
61. Mollalo, A., Vahedi, B., and Rivera, K. M. (2020). GIS–based spatial modeling of COVID–19 incidence rate in the continental United States, *Science of The Total Environment*, **728**, 138884.
62. Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al–Jabir, A., Iosifidis, C., et al. (2020). The socio–economic implications of the coronavirus pandemic (COVID–19): A review, *International Journal of Surgery*, **78**, 185–193.
63. Nielsen, F. (2016). Hierarchical clustering, In *Introduction to HPC with MPI for data science*, 195–211, Springer.
64. Ntotsis, K. and Karagrigoriou, A. (2021). *The impact of multicollinearity on big data multivariate analysis modeling*. In Y. Dimotikalis, A. Karagrigoriou, C. Parpoula, & C. H. Skiadas (Eds.), *Applied modeling techniques and data analysis 1: Computational data analysis methods and tools*, 187-194, John Wiley & Sons.
65. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572.
66. Pérez–Segura, V., Caro–Carretero, R., and Rua, A. (2021). Multivariate analysis of risk factors of the COVID–19 pandemic in the Community of Madrid, Spain, *International Journal of Environmental Research and Public Health*, **18**(17), 9227.
67. Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid–19 vaccine, *New England Journal of Medicine*, **383**(27), 2603–2615.

68. Prual, A., Chacko, S., and Koch–Weser, D. (1991). Sexual behaviour, AIDS and poverty in Sub–Saharan Africa, *International Journal of STD & AIDS*, **2**, 1–9.
69. Rosipal, R. and Kramer, N. (2005). Overview and recent advances in partial least squares, In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe–Taylor (Eds.), *Subspace, latent structure and feature selection*, 34–51, New York, NY: Springer.
70. Schölkopf, B., Smola, A., and Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**(5), 1299–1319.
71. Shakil, M. H., Munim, Z. H., Tasnia, M., and Sarowar, S. (2020). COVID–19 and the environment: A critical review and research agenda, *Science of The Total Environment*, **745**, 141022.
72. Sun, G.–Q., Wang, S.–F., Li, M.–T., Li, L., Zhang, J., Zhang, W., Jin, Z., and Feng, G.–L. (2020). Transmission dynamics of COVID–19 in Wuhan, China: Effects of lockdown and medical resources, *Nonlinear Dynamics*, **101**, 1981–1993.
73. Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**(5500), 2319–2323.
74. Van der Maaten, L. J. P., and Hinton, G. E. (2008). Visualizing data using t–SNE, *Journal of Machine Learning Research*, **9**, 2579–2605.
75. Varmuza, K. and Filzmoser, P. (2009). *Introduction to multivariate statistical analysis in chemometrics*, CRC Press.
76. Von Luxburg, U. (2007). A tutorial on spectral clustering, *Statistics and Computing*, **17**(4), 395–416.
77. Von Mises, R., and Pollaczek–Geiringer, H. (1929). Praktische Verfahren der Gleichungsauflösung, *Zeitschrift für Angewandte Mathematik und Mechanik*, **9**, 152–164.
78. Wehrens, H. (2011). *Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences*, Springer.
79. Wold, H. (1975). Path models with latent variables: The NIPALS approach, In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi (Eds.), *Quantitative sociology: International perspectives on mathematical and statistical model building*, 307–357, Academic Press.
80. Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS–regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130.
81. Zheng, P., Chen, Z., Liu, Y., Song, H., Wu, C.–H., Li, B., Kraemer, M. U. G., Tian, H., Yan, X., Zheng, Y., et al. (2021). Association between coronavirus disease 2019 (COVID–19) and long–term exposure to air pollution: Evidence from the first epidemic wave in China, *Environmental Pollution*, **276**, 116682.
82. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID–19 in Wuhan, China: A retrospective cohort study, *The Lancet*, **395**(10229), 1054–1062.