UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Corso di Laurea Magistrale
in Management

Tesi di Laurea

SPECTRAL CLUSTERING: ANALISI DELLA TEORIA E APPLICAZIONE ALLO STUDIO DELLA VISITOR EXPERIENCE ALLA PINACOTECA TOSIO MARTINENGO

Relatore: Chiar.ma Prof.ssa PAOLA ZUCCOLOTTO

Correlatore: Dott. MATTEO VENTURA

Laureanda:
ELENA ROMANO

Matricola n. 724530

Anno Accademico 2022/2023

Indice

Introduzione

Nel mondo in continua evoluzione del marketing e della gestione aziendale, la segmentazione del mercato è una strategia essenziale per comprendere e adattarsi alle complesse dinamiche dei mercati. La suddivisione delle unità statistiche, come i clienti o i consumatori, in gruppi omogenei chiamati cluster, rappresenta un pilastro per l'elaborazione di strategie mirate. I cluster sono gruppi di unità statistiche che condividono caratteristiche simili tra loro, ma che si distinguono significativamente dagli altri cluster in base a variabili rilevanti per l'analisi. Questa suddivisione è di cruciale importanza in quanto permette alle aziende di comprendere meglio le preferenze e i comportamenti dei propri clienti, prendere decisioni più informate e personalizzare le proprie strategie di marketing.

La *Cluster Analysis*, una tecnica statistica ampiamente utilizzata, gioca un ruolo fondamentale nella segmentazione del mercato e nell'analisi esplorativa dei dati. Questa tecnica, trova applicazione in una vasta gamma di discipline scientifiche, tra cui la statistica, l'informatica, la biologia, le scienze sociali, la psicologia e, ovviamente, il marketing. La sua capacità di identificare pattern nascosti e relazioni tra dati, semplificando così la comprensione di insiemi complessi di informazioni, è essenziale per condurre analisi esplorative approfondite.

Nel contesto di questa tesi, esploreremo in profondità una tecnica di clustering innovativa e particolarmente potente: lo *Spectral Clustering*. Questo algoritmo, basato sulla Teoria dei Grafi, sfrutta le proprietà degli autovalori e degli autovettori per suddividere in cluster dati non etichettati. La sua capacità di rivelare pattern complessi nei dati e di superare le limitazioni dei tradizionali metodi di clustering lo rende straordinariamente utile in un ampio spettro di applicazioni, tra cui, appunto, il campo del marketing.

L'obiettivo principale di questa tesi è duplice. Innanzitutto, intendiamo fornire una panoramica esaustiva dei fondamenti teorici relativi alla segmentazione del mercato, alla *Cluster Analysis* e allo *Spectral Clustering*. Questo fornirà una base solida per comprendere il contesto in cui si inserisce la nostra ricerca.

Infine, intendiamo dimostrare l'applicazione concreta di queste teorie attraverso un caso di studio. Utilizzando dati reali raccolti durante uno studio sulla *Visitor Experience* presso la Pinacoteca Tosio Martinengo di Brescia, mostreremo come lo *Spectral Clustering* possa rivelare pattern comportamentali tra i visitatori e come questi risultati possano essere utilizzati per ottimizzare le strategie di marketing del museo.

La nostra ricerca si articola in quattro capitoli chiave. Nel primo capitolo,

introduciamo i concetti fondamentali relativi alla segmentazione del mercato, alla *Cluster Analysis* e allo *Spectral Clustering*. Esploriamo il ruolo cruciale della segmentazione del mercato nelle strategie aziendali e presentiamo lo *Spectral Clustering* come un'innovativa tecnica di clustering.

Il secondo capitolo ci porta nel cuore del clustering, concentrandosi sul grafo di similarità. Spieghiamo come questo strumento permetta di rappresentare dati come un insieme di vertici interconnessi da archi, riflettendo la misura di similarità tra i dati stessi. Questo concetto è fondamentale per lo *Spectral Clustering*, che sfrutta il grafo di similarità per rivelare pattern complessi nei dati.

Nel terzo capitolo, approfondiamo il concetto di matrice Laplaciana del Grafo e la sua importanza nella riduzione della dimensionalità. Questa matrice gioca un ruolo cruciale nell'analisi degli autovettori e degli autovalori, fondamentali per il funzionamento dello *Spectral Clustering*. Esploriamo le diverse varianti della matrice Laplaciana e come influenzino la rappresentazione dei dati.

Il quarto e ultimo capitolo rappresenta l'applicazione concreta dello *Spectral Clustering* in un contesto di ricerca di mercato. Utilizzando dati reali raccolti in uno studio sulla *Visitor Experience* presso la Pinacoteca Tosio Martinengo di Brescia, dimostriamo come lo *Spectral Clustering* possa rivelare pattern comportamentali tra i visitatori e migliorare le strategie di marketing del museo. Questo capitolo è il culmine del nostro percorso, in cui teoria e pratica si incontrano per produrre risultati tangibili.

Attraverso questo lavoro di ricerca, miriamo a dimostrare l'efficacia dello *Spectral Clustering* nell'analisi dei dati complessi e nell'ottimizzazione delle strategie di marketing. Speriamo che questa tesi possa contribuire alla comprensione delle potenzialità dell'analisi dei cluster e dell'applicazione dello *Spectral Clustering* in diversi contesti.

Con questa introduzione generale, ci immergeremo ora nell'analisi dettagliata di ciascun capitolo, esplorando i concetti teorici e presentando i risultati dell'applicazione pratica.

1. Introduzione al clustering

Nel contesto del marketing e della gestione aziendale, la segmentazione del mercato è una strategia fondamentale che mira a suddividere le unità statistiche, come i clienti o i consumatori, in gruppi omogenei chiamati cluster. Il termine cluster significa letteralmente "grappolo", ma il suo utilizzo nella lingua inglese può variare da "gruppo" a "sciame". L'obiettivo è creare gruppi di unità statistiche simili tra loro all'interno di ciascun cluster, ma significativamente diversi dagli altri cluster, in base alle variabili rilevanti per l'analisi. Questa suddivisione è di cruciale importanza per comprendere e gestire efficacemente la complessità dei mercati. Attraverso l'analisi dei cluster, le aziende possono ottenere una visione dettagliata delle diverse caratteristiche e preferenze dei propri clienti, consentendo loro di adattare le proprie strategie di marketing e prendere decisioni più informate e mirate (Bassi F., 2022).

La segmentazione del mercato può essere effettuata utilizzando diverse tecniche statistiche, tra cui la *Cluster Analysis*. Queste tecniche sono ampiamente utilizzate in diverse discipline. Ad esempio, in psichiatria vengono impiegate per affinare le categorie diagnostiche esistenti, mentre in archeologia vengono utilizzate per studiare la relazione tra vari tipi di manufatti. Nel campo delle ricerche di mercato, i metodi di *Cluster Analysis* sono invece applicati per creare gruppi di consumatori con diversi modelli di acquisto.

A un livello più generale, l'utilizzo di uno schema di classificazione, come la *Cluster Analysis*, consente di organizzare in modo conveniente un vasto insieme di dati, facilitandone la comprensione e il recupero delle informazioni in modo efficiente. Se i dati possono essere raggruppati validamente in un numero ridotto di cluster, le etichette dei gruppi forniscono una descrizione sintetica dei modelli di somiglianze e differenze presenti nei dati. Ad esempio, nel settore delle ricerche di mercato, la suddivisione di un ampio numero di intervistati in base alle loro preferenze per determinati prodotti può aiutare ad identificare un "prodotto di nicchia" per un particolare tipo di consumatore. La necessità di sintetizzare grandi quantità di dati in modo efficace è diventata sempre più importante con l'avvento dei grandi database disponibili in diverse aree scientifiche, e l'analisi dei cluster, insieme ad altre tecniche di analisi multivariata, viene spesso denominata *data mining* (Zou, 2020).

Nel contesto di questa ricerca, forniremo una panoramica generale sulla *Cluster Analysis* e presenteremo le principali tecniche impiegate in questo campo. presteremo particolare attenzione allo *Spectral Clustering*, un'innovativa tecnica di clustering che sfrutta la Teoria dei Grafi e le proprietà degli autovalori e degli

autovettori per individuare pattern e strutture nascoste nei dati.

Sebbene l'approfondimento dello *Spectral Clustering* sia previsto nei capitoli successivi, in cui analizzeremo l'algoritmo in profondità e discuteremo i risultati ottenuti attraverso un esperimento condotto su dati reali, in questa sezione introdurremo brevemente il suo funzionamento per fornire una visione generale.

In aggiunta, discuteremo i limiti delle tecniche di segmentazione tradizionali e come lo *Spectral Clustering* possa superarli efficacemente.

1.1 Generalità sulla Cluster Anaysis

Per *Cluster Analysis* si intende un insieme di procedure e metodologie utilizzate per identificare gruppi all'interno di una popolazione[1] di dati. Le sue origini risalgono al 1939, quando Tryon R. C. ha pubblicato una monografia intitolata "*Cluster Analysis*" (Tryon, 1939), ma è stato solo a partire dagli anni '60 che questa metodologia è stata ampiamente proposta e studiata. Sokal e Sneath hanno svolto un ruolo fondamentale nella sua diffusione, pubblicando nel 1963 il libro "*Principles of numerical taxonomy*" (Sokal, 1963), che ha offerto una prima esposizione sistematica delle tecniche di clustering.

Negli anni successivi, la *Cluster Analysis* ha trovato applicazione in diversi campi di ricerca. Hartigan, nel 1975, ha fornito un importante sommario degli studi pubblicati che documentano i risultati ottenuti con la *Cluster Analysis* (Hartigan J. A., 1979).

In generale, ogni qualvolta si deve classificare una grande mole di informazioni in gruppi espressivi e trattabili, la *Cluster Analysis* rappresenta uno strumento prezioso. Queste tecniche sono state utilizzate per affrontare una vasta gamma di problemi, dalla biologia alla psicologia, dalla geografia all'economia. In particolare, nel contesto manageriale, la *Cluster Analysis* si è dimostrata uno strumento prezioso per la definizione di strategie commerciali (Smart, 2005).

Nel campo manageriale, comprendere le caratteristiche, i bisogni e i comportamenti degli acquirenti è essenziale per sviluppare una strategia di marketing efficace (Paul E. Green, 1967). La *Cluster Analysis* offre un approccio analitico che consente di identificare gruppi omogenei di clienti, consentendo alle aziende di adattare l'offerta di prodotti e servizi in base alle specifiche esigenze di ciascun gruppo (Brian S. Everitt, 2001). Ciò permette di ottimizzare le politiche di

[1] Insieme finito o infinito di unità statistiche oggetto d'indagine.

vendita e di creare una comunicazione mirata per massimizzare l'efficacia delle iniziative di marketing (Bassi F., 2022).

Ad esempio, supponiamo che un'azienda operi nel settore dell'elettronica di consumo. Utilizzando la *Cluster Analysis*, può suddividere la sua base di clienti in gruppi omogenei sulla base di fattori come le preferenze di prodotto, il livello di reddito, l'età e il comportamento di acquisto. Questa segmentazione del mercato consente all'azienda di sviluppare offerte di prodotti specifiche per ciascun gruppo, adattando il design, il prezzo e la promozione in base alle esigenze e alle preferenze di ogni segmento di clientela (Brian S. Everitt, 2001).

Il termine *Cluster Analysis* si riferisce quindi ad un insieme di procedure utilizzate per individuare sottoinsiemi di oggetti all'interno di un insieme di dati. Questi sottoinsiemi, chiamati *cluster*, sono mutuamente esclusivi e tendenzialmente omogenei al loro interno.

Le tecniche di *Cluster Analysis* creano i gruppi in modo tale che ogni osservazione sia molto simile a tutte le altre che appartengono allo stesso gruppo in funzione di alcuni criteri prestabiliti dal ricercatore. L'obiettivo è massimizzare la similarità intra-cluster, cioè la similarità tra gli oggetti all'interno di ciascun cluster, e allo stesso tempo massimizzare l'eterogeneità inter-cluster, cioè la differenza tra gli oggetti di cluster diversi.

Come rappresentato nella Figura 1.1, alla fine del procedimento di *Cluster Analysis*, ci si aspetta che i cluster finali esibiscano un'alta omogeneità interna, il che significa che gli oggetti all'interno di ciascun cluster sono molto simili tra loro. Allo stesso tempo, ci si aspetta un'alta eterogeneità esterna, che indica che gli oggetti appartenenti a cluster diversi sono significativamente diversi tra loro (Barbarito, 2011).

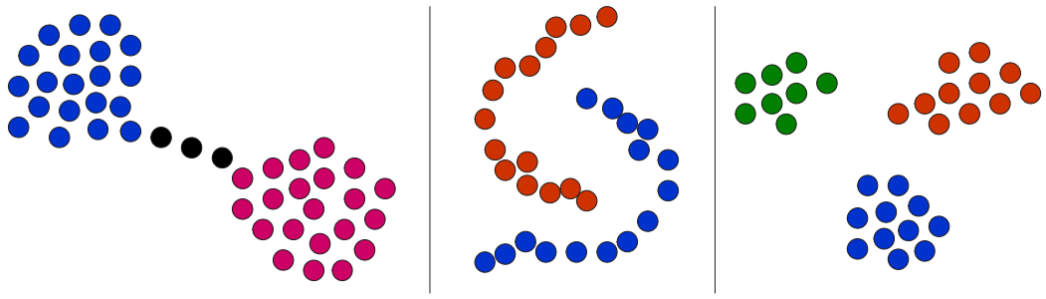


Figura 1.1 – Cluster dotati di coesione interna e separazione esterna.
Fonte: nostre elaborazioni.

In pratica, una buona *Cluster Analysis* dovrebbe consentire di identificare gruppi

distinti e rilevanti all'interno dei dati, fornendo una comprensione più approfondita della struttura nascosta o dei modelli presenti nei dati stessi.

Tramite una rappresentazione visuale le proprietà di coesione dei cluster possono essere rese evidenti, senza definirle in modo esplicito e rigoroso. Teniamo presente che non esiste una singola definizione che risulti sufficiente per ogni situazione: i tentativi di rendere i concetti di omogeneità e separazione matematicamente precisi in termini di espliciti indici numerici hanno condotto a numerosi e differenti criteri.

Punto di partenza di ogni applicazione di *Cluster Analysis* è la disponibilità di un collettivo statistico (anche campionario) di *n* elementi ciascuno rappresentato da *p* variabili qualitative e/o quantitative. I dati si possono organizzare in una matrice $\boldsymbol{X}$ di dimensione $n \times p$, dove ogni riga *i* rappresenta un soggetto e ogni colonna *j* rappresenta una variabile ($i = 1,2,\dots,n;\ j = 1,2,\dots,p$):

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

(1.1)

Tale matrice è spesso chiamata bimodale, in riferimento al fatto che a righe e colonne è associato un diverso significato. Le variabili in $\boldsymbol{X}$ possono essere continue, ordinali, categoriche o di natura mista e qualche cella può essere vuota. Vedremo che la natura mista delle variabili e le celle vuote complicano molto il processo di clustering dei dati.

Alcune tecniche di classificazione provvedono ad una preventiva trasformazione della matrice $\boldsymbol{X}$ in una matrice quadrata $n \times n$ unimodale di similarità, dissimilarità, o distanza (in generale si può parlare di prossimità). La *Cluster Analysis* è essenzialmente un processo di scoperta della struttura a gruppi dei dati e non deve essere confusa con la differenziazione o assegnazione, in cui i gruppi sono noti a priori e l'analisi deve costruire regole per inserire i singoli oggetti in uno dei gruppi noti.

Nelle Figure 1.2 e 1.3 è possibile osservare visivamente come la *Cluster Analysis* abbia l'obiettivo di suddividere i dati in *k* gruppi, mirando a ottenere un'elevata omogeneità interna e una marcata separazione esterna tra i cluster. Questo processo evidenzia la capacità della tecnica statistica di riconoscere la presenza di raggruppamenti nei dati nell'iperspazio in cui sono definiti (Barbarito, 2011).

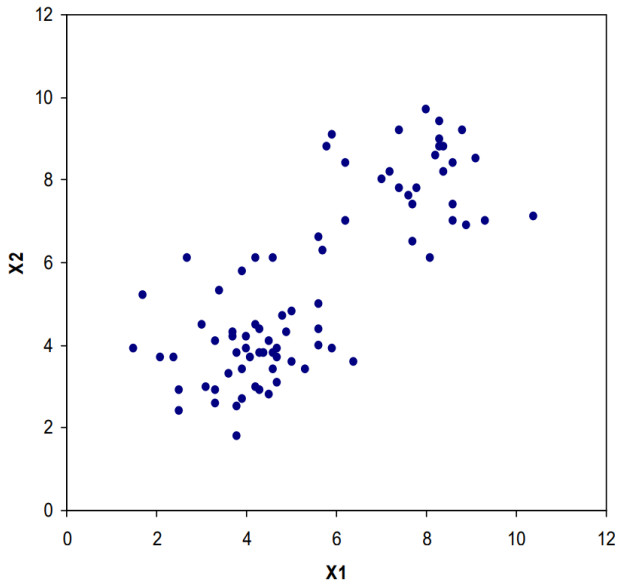


Figura 1.2 – Dati bidimensionali che non contengono alcuna classe.
Fonte: Statistica per il Marketing, Prof.ssa Paola Zuccolotto (A.A. 2022-2023).

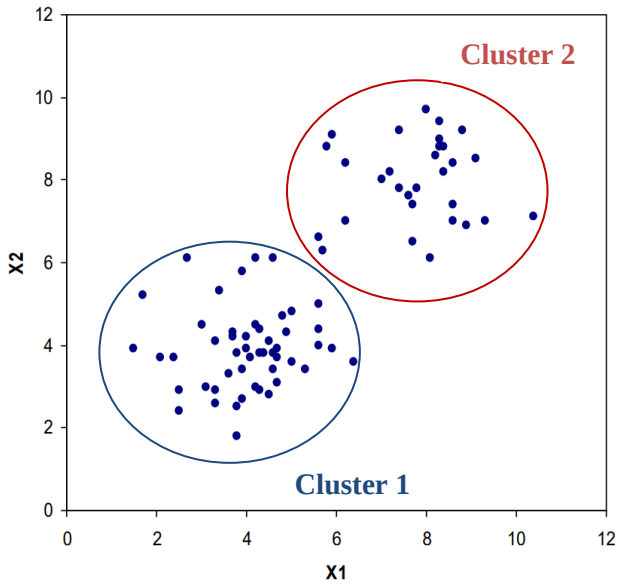


Figura 1.3 – Dati bidimensionali riorganizzati in cluster.
Fonte: Statistica per il Marketing, Prof.ssa Paola Zuccolotto (A.A. 2022-2023).

Per raggiungere questo obiettivo, la *Cluster Analysis* utilizza una misura di distanza appropriata per valutare la similarità o dissimilarità tra le unità statistiche. Le misure di distanza più comuni includono la distanza euclidea, la distanza di Manhattan o la correlazione (Bandyopadhyay & Saha, 2013). La scelta della misura di distanza dipende dal tipo di dati e dal contesto di studio.

La *Cluster Analysis* è un metodo esplorativo e non necessita di alcuna assunzione a priori, tuttavia, richiede una serie di decisioni da parte del ricercatore, prima, durante e dopo l'analisi, come riportato nella Tabella 1.4, che elenca tali decisioni.

Tabella 1.4 – Le decisioni nella Cluster Analysis.

Prima	• Scelta delle variabili; • Criteri di similarità-distanza;
Durante	• Tecniche di aggregazione; • Numero dei gruppi da ottenere;
Dopo	• Valutazione della qualità della soluzione; • Scelta fra le diverse possibili soluzioni alternative.

Ovviamente scelte diverse portano a risultati diversi, pertanto, questa componente di arbitrarietà è stata oggetto di notevoli critiche, ma, evidentemente, nelle scienze il fattore di soggettività accomuna tutti i procedimenti di analisi multivariata dei dati. È infatti tipico dei procedimenti di conoscenza scientifica un processo di riduzione e di semplificazione controllata delle informazioni disponibili, per favorire la comprensione dei fenomeni.

La definizione di clustering non è univoca e non esiste una procedura standard per identificare i diversi cluster. In generale, la *Cluster Analysis* può essere implementata percorrere alcuni passi principali:

- **Scelta delle unità di osservazione**: è necessario determinare le unità statistiche (soggetti) da includere nell'analisi. Questa selezione dipende dal contesto e dall'obiettivo dell'analisi.
- **Scelta delle variabili**: nel caso della *Cluster Analysis*, per suddividere *n* unità statistiche in *k* gruppi si possono utilizzare sia variabili numeriche che categoriali, o includere entrambi i tipi. È fondamentale selezionare attentamente le variabili da utilizzare, poiché i cluster saranno costruiti in base a queste variabili. Le unità statistiche all'interno di ciascun cluster saranno omogenee tra loro rispetto alle variabili selezionate.
- **Omogeneizzazione delle scale di misura**: se le variabili selezionate hanno scale di misura diverse, è necessario eseguire una normalizzazione o una standardizzazione per garantire che le differenze di scala non influiscano sull'analisi.
- **Scelta del tipo di distanza**: dopo aver scelto le variabili è necessario decidere quale misura di distanza utilizzare per valutare la similarità o la

dissimilarità tra le unità statistiche. La scelta della misura di distanza è importante per il successo dell'algoritmo di clustering. È importante considerare che la definizione di similitudine o dissimilitudine tra le unità statistiche dipende dal tipo di dati analizzati. Le misure di distanza possono dunque variare a seconda del tipo di dati considerati: per variabili quantitative vengono spesso utilizzate misure di distanza euclidee o basate sulla correlazione, mentre per variabili qualitative si ricorre a misure di distanza basate sulla dissimilarità o sulla similarità dei profili. Pertanto, è necessario adattare le tecniche di *Cluster Analysis* in base al tipo di dati trattati.

- **Scelta dell'algoritmo di classificazione**: una volta definita la misura di distanza, si possono applicare diverse tecniche o algoritmi per identificare i cluster. Esistono diverse metodologie, come il *k-means*, il metodo agglomerativo o il metodo divisivo, che utilizzano diverse strategie per creare i cluster.
- **Analisi e interpretazione dei cluster ottenuti**: una volta formati i cluster, è necessario analizzarli per comprendere come variano le variabili utilizzate per la segmentazione all'interno dei cluster stessi (etichettatura) e come si differenziano dalle variabili non utilizzate per la segmentazione (descrizione). Questa fase richiede una comprensione approfondita dei dati e delle caratteristiche dei cluster identificati (Bassi F., 2022).

1.2 Principali tecniche di Cluster Anaysis

Nella ricerca di strutture nascoste all'interno dei dati, il clustering si è dimostrato uno strumento fondamentale. L'obiettivo principale del clustering, come già anticipato, è quello di raggruppare gli oggetti simili tra loro in modo da creare sottoinsiemi omogenei e distinguibili. Nel corso degli anni, sono stati sviluppati diversi approcci per affrontare questa sfida.

Gli approcci principali al clustering possono essere distinti in due categorie fondamentali: gerarchici e non gerarchici (o partizionali). Ognuno di questi offre vantaggi e limitazioni specifiche, ma hanno tutti in comune l'importanza attribuita alla scelta della misura di distanza (Gülagiz, 2017). Infatti spesso il successo di un algoritmo di clustering consiste nello scegliere la misura adatta. La misura di distanza rappresenta un criterio essenziale per valutare la similarità o la dissimilarità tra due punti nel processo di raggruppamento.

Nella letteratura, esistono diverse misure di distanza comunemente utilizzate, come la distanza euclidea, la distanza di Manhattan, la distanza di Minkowski e

molte altre (Bandyopadhyay & Saha, 2013). La scelta della misura di distanza dipende strettamente dalla natura dei dati e dagli obiettivi del clustering.

Negli approcci gerarchici, l'obiettivo è costruire una struttura gerarchica di cluster che rappresenti una serie di partizioni "nidificate". Questo tipo di approccio offre una visione più completa della struttura dei dati, consentendo di esplorare diverse scale di raggruppamento. Tuttavia, gli approcci gerarchici possono essere dal punto di vista computazionale più intensivi e possono richiedere più risorse rispetto agli approcci non gerarchici.

Gli approcci non gerarchici, noti anche come partizionali, cercano invece di creare partizioni "piane" dei dati, senza una struttura gerarchica. Questi approcci sono spesso più efficienti dal punto di vista computazionale e possono essere adatti a dataset di grandi dimensioni. Tuttavia, possono essere meno flessibili nel rappresentare la complessità della struttura dei dati rispetto agli approcci gerarchici.

1.2.1 Cluster Analysis gerarchica

Le procedure agglomerative costituiscono la variante di clustering più usata: esse generano una serie di partizioni dei dati in k cluster, che vanno da $k = n$ (ogni cluster contiene un singolo individuo), fino a $k = 1$ (un unico cluster che contiene tutti gli individui).

La *Cluster Analysis* gerarchica è un metodo di clustering che organizza i dati in una struttura gerarchica di cluster, in cui i cluster possono essere visualizzati come un albero chiamato dendrogramma. Questa tecnica si basa sulla similarità o dissimilarità tra gli elementi del dataset per creare i cluster.

La *Cluster Analysis* gerarchica può essere eseguita utilizzando due approcci principali: agglomerativo e divisivo. Nel metodo agglomerativo, si inizia considerando ogni osservazione come un cluster separato e successivamente si uniscono iterativamente i cluster più simili tra loro fino a ottenere un singolo cluster contenente tutti i dati.

Inizialmente, viene calcolata una misura appropriata di distanza o similarità tra le coppie di osservazioni, come la distanza euclidea, la correlazione o altre misure specifiche al dominio dei dati (Bandyopadhyay & Saha, 2013).[2] Questa misura viene utilizzata per valutare quanto due osservazioni siano simili o diverse l'una dall'altra.

[2] https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e#:~:text=Average%2Dlinkage%20is%20where%20the,distance%20metrics%20in%20hierarchical%20clustering.

Successivamente, si procede unendo i cluster che sono più vicini tra loro in base alla misura di distanza o similarità scelta e al tipo di algoritmo utilizzato (ad esempio, metodo del legame singolo, completo, di Ward, …, come si vedrà con maggior dettaglio in seguito). Questo processo di unione viene ripetuto iterativamente, fino a quando tutti i dati sono raggruppati in un unico cluster (strategia *bottom-up*) (Lance G. N., 1966).

I metodi divisivi operano in modo contrario rispetto ai metodi agglomerativi: si parte invece con un unico grande cluster che contiene tutte le osservazioni e si procede a dividerlo in cluster più piccoli in modo iterativo. Inizialmente, viene calcolata la distanza o la similarità tra tutte le coppie di osservazioni nel cluster. Successivamente, si seleziona una coppia di osservazioni che presenta la massima dissimilarità e si suddivide il cluster iniziale in due sottocluster. Questo processo viene ripetuto iterativamente, dividendo i cluster in sottocluster più piccoli fino a quando ogni osservazione costituisce un cluster separato (strategia *top-down*) (Bridges Jr, 1966). Questi metodi possono essere computazionalmente impegnativi se si considerano tutte le possibili divisioni in due sottogruppi di un cluster di *k* oggetti, che sono $2^{k-1} - 1$ divisioni (Brian S. Everitt, 2001).

Tuttavia, per i dati che consistono di variabili binarie, esistono metodi di divisione relativamente semplici ed efficienti dal punto di vista computazionale, noti come metodi monotetici di divisione. Questi metodi dividono i cluster in base alla presenza o all'assenza delle variabili, in modo che ogni cluster contenga membri con attributi specifici, che sono tutti presenti o tutti assenti. Pertanto, i dati utilizzati con questi metodi devono essere organizzati in una matrice a due modalità, in cui ogni variabile è binaria (Brian S. Everitt, 2001).

Il termine "monotetico" si riferisce al fatto che viene utilizzata una sola variabile per la suddivisione in ogni fase del processo. Al contrario, i metodi politetici utilizzano tutte le variabili in ogni fase. Sebbene i metodi di divisione siano meno comuni rispetto ai metodi agglomerativi, presentano il vantaggio, evidenziato da Kaufman e Rousseeuw (1990), di rivelare fin dall'inizio la struttura principale dei dati, che è di interesse per la maggior parte degli utenti.

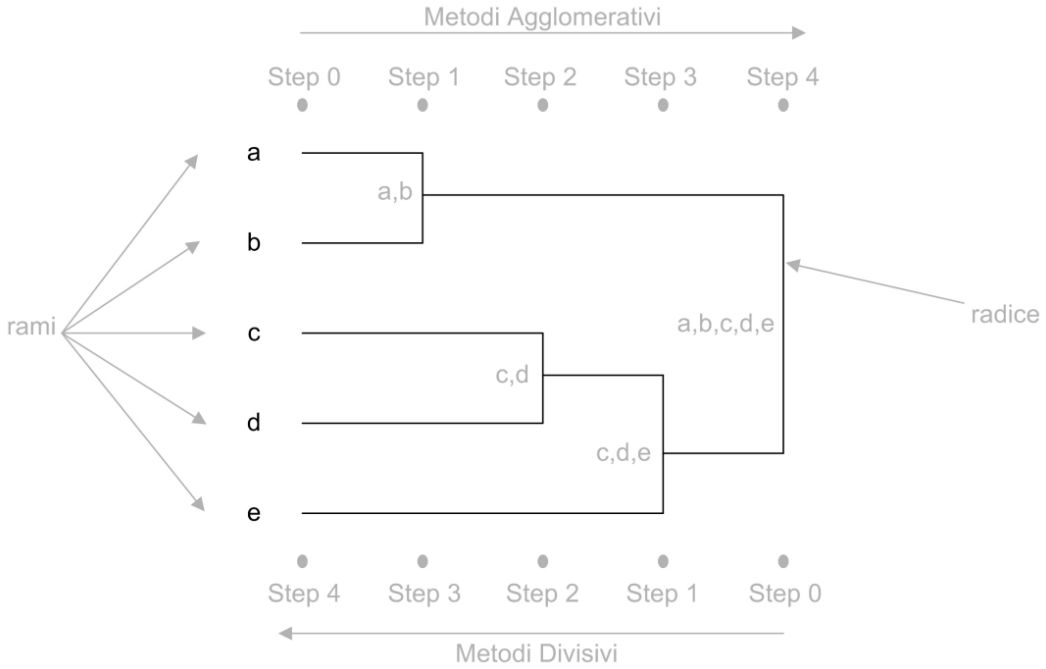


Figura 1.5 – Esempio di struttura gerarchica ad albero.
Fonte: nostre elaborazioni.

Lo scopo di entrambe le varianti è il medesimo: arrivare, tramite successive suddivisioni o sintesi, a una partizione ottimale, operando su una matrice di prossimità che rappresenta le relazioni tra gli elementi del dataset. La Figura 1.5 illustra il funzionamento delle due diverse varianti, agglomerativa e divisiva, dell'algoritmo di clustering gerarchico.

La caratteristica principale degli algoritmi gerarchici nella *Cluster Analysis* è che le divisioni o fusioni sono irrevocabili. Questo significa che una volta che due elementi sono stati assegnati a cluster separati, non possono essere nuovamente uniti nello stesso cluster, né viceversa. I metodi agglomerativi consentono solo fusioni, quelli divisivi solo partizioni e non sono ammessi algoritmi gerarchici di tipo misto; pertanto, se due individui sono stati assegnati a cluster diversi, non potranno in seguito essere nuovamente membri di uno stesso cluster (Lancia G. N., 1967). Il termine "gerarchico" deriva proprio da questa caratteristica dei metodi gerarchici, in cui le decisioni di divisione o fusione dei cluster seguono una struttura a cascata, in cui le modifiche apportate in fasi precedenti influenzano le decisioni nelle fasi successive.

Poiché nelle tecniche gerarchiche agglomerative, il processo di clustering prosegue fino a quando tutti i dati sono racchiusi in un unico cluster contenente tutti gli individui, mentre le tecniche divisive suddividono l'insieme dei dati fino

ad avere *k* gruppi, ognuno dei quali contenente un singolo individuo, l'investigatore deve possedere un criterio che determini la terminazione dell'algoritmo nel momento in cui si è raggiunto il numero ottimale di cluster.

La struttura nidificata che si ottiene con questo algoritmo, chiamata anche albero binario, poiché le fusioni o le divisioni avvengono a coppie, può essere rappresentata visivamente attraverso un diagramma bidimensionale chiamato dendrogramma. Il dendrogramma mostra le fusioni e le divisioni che si sono verificate ad ogni stadio dell'analisi, consentendo di valutare il livello dell'albero che fornisce la migliore partizione dei dati (Lancia G. N., 1967). Un esempio di dendrogramma è mostrato in Figura 1.6.

Attraverso la lettura del dendogramma, è possibile comprendere come si è svolto il processo di agglomerazione e identificare i cluster a diversi livelli di dettaglio. Ogni vertice nel dendrogramma rappresenta un gruppo o una fusione di gruppi, mentre i rami rappresentano la sequenza delle fusioni. La lunghezza dei rami può essere interpretata come una misura della distanza o dissimilarità tra i gruppi (Lance G. N., 1966).

La scelta del livello di partizione ottimale nel dendrogramma dipende dall'obiettivo dell'analisi. A seconda del contesto e delle domande di ricerca, potrebbe essere necessario identificare cluster di dimensioni specifiche, individuare gruppi distinti o trovare una suddivisione che massimizzi la similarità interna e minimizzi la similarità tra i gruppi. L'interpretazione del dendrogramma può aiutare nella selezione del punto di taglio appropriato per ottenere la partizione desiderata (Bridges Jr, 1966).

Le tecniche di classificazione gerarchiche sono ampiamente utilizzate in biologia, sociologia, biblioteconomia e in tutti quei campi in cui è implicita una struttura gerarchica nei dati.

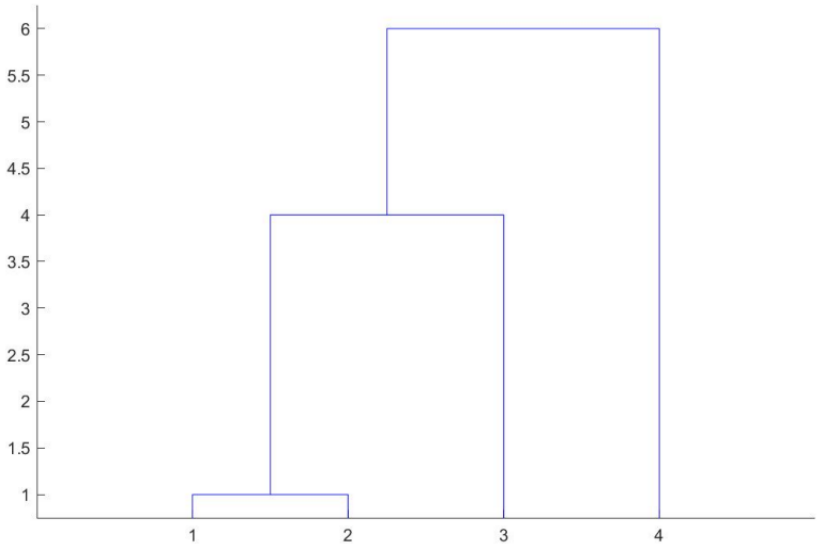


Figura 1.6 – Esempio di dendogramma con 4 elementi.
Fonte: https://it.mathworks.com/help/stats/linkage.html

Nella *Cluster Analysis*, l'implementazione più comune è quella di tipo agglomerativo, quindi per il resto della spiegazione ci concentreremo solo su questo approccio.

Nell'approccio agglomerativo della *Cluster Analysis*, uno dei passaggi fondamentali consiste nell'utilizzo di una matrice di distanza $\boldsymbol{D}$ di dimensione $n \times n$, dove n rappresenta il numero di individui nel dataset. Questa matrice di distanza viene calcolata in base alla scelta predefinita della misura di distanza, che può variare a seconda del tipo di dati e delle caratteristiche del problema. Alcune delle misure di distanza comunemente utilizzate includono la distanza euclidea, la distanza di Manhattan e la distanza di correlazione. La misura di distanza scelta dipende dalla natura dei dati e dalle specifiche esigenze dell'analisi, e ha un impatto diretto sulla configurazione dei cluster risultanti.

Ogni cella della matrice di distanza $\boldsymbol{D}_{n \times n}$ indica la distanza o la dissimilarità tra due oggetti o unità statistiche del dataset. Ad esempio, se si considerano n punti in uno spazio multidimensionale, la matrice di distanza conterrà n righe e n colonne, dove ogni elemento (i, j) rappresenta la distanza o la similarità tra le unità statistiche i e j.

Il processo di clustering gerarchico agglomerativo si sviluppa attraverso i seguenti passaggi:

- **Inizializzazione**: il processo inizia con n unità statistiche che devono

essere classificate. Ogni unità statistica rappresenta un gruppo iniziale. Inoltre, viene utilizzata una matrice delle distanze $\boldsymbol{D}$ di dimensione $n \times n$, che è simmetrica e rappresenta le distanze o similarità tra le unità statistiche.

- **Selezione**: identifica le due unità statistiche più vicine nella matrice $\boldsymbol{D}_{n\times n}$ rispetto alla misura di prossimità fissata inizialmente e le unisce per formare un gruppo cluster.
- **Aggiornamento**: il numero totale di cluster viene ridotto di uno (da *n* a $n-1$) unendo i due gruppi selezionati in un unico cluster. Successivamente, viene ricalcolata la matrice delle distanze. Le due righe (o colonne) relative ai due cluster precedentemente selezionati vengono sostituite con una singola riga di distanze che rappresenta il nuovo gruppo formato dalla fusione. Di conseguenza, la matrice delle distanze viene ridotta di dimensione, diventando una matrice di dimensione $(n-1)\times(n-1)$, in cui la riga e la colonna corrispondenti ai due cluster uniti sono state aggiornate.
- **Ripetizione**: ripete iterativamente i passi (2) e (3) per questo processo iterativo per $(n-1)$ volte.
- **Arresto**: la procedura si arresta quando tutti gli elementi vengono incorporati in un unico cluster.

Durante il processo di clustering gerarchico, oltre alla definizione della distanza, è necessario specificare anche il metodo di calcolo delle distanze tra i gruppi, noto come *linkage*. Ci sono diverse tipologie di *linkage* utilizzate nel clustering gerarchico, tra cui:

- **Metodo del legame singolo**: il primo metodo è il più semplice tra i metodi di clustering gerarchico ed è detto metodo del legame singolo (*single linkag*e) o del vicino più vicino (*nearest- neighbour technique*) (Florek K., 1951; Sneath, 1957; Johnson, 1967). L'assunto base di questa tecnica è identificare la distanza (o similarità) fra due gruppi con quella fra i loro membri più vicini (o più simili).

  Il grado di vicinanza fra due gruppi è stabilito prendendo in considerazione solo le informazioni relative a due oggetti più vicini, ignorando quelle che si riferiscono a tutti gli altri oggetti appartenenti ai gruppi. La distanza tra due gruppi *A* e *B* è definita come la distanza minore rilevata tra la coppia di individui (*i*, *j*) con $i \in A$, $j \in B$, in altri termini si considera il minimo delle $n_A \times n_B$ distanze tra ciascuna delle unità del gruppo *A* e ciascuna delle unità del gruppo *B*, dove n_A e n_B sono rispettivamente le numerosità dei gruppi *A* e *B*:

$$d_{AB} = \min_{i \in A, j \in B} d_{ij}.$$

(1.2)

La Figura 1.7 illustra chiaramente il concetto di misura della distanza nel contesto del metodo del legame singolo.

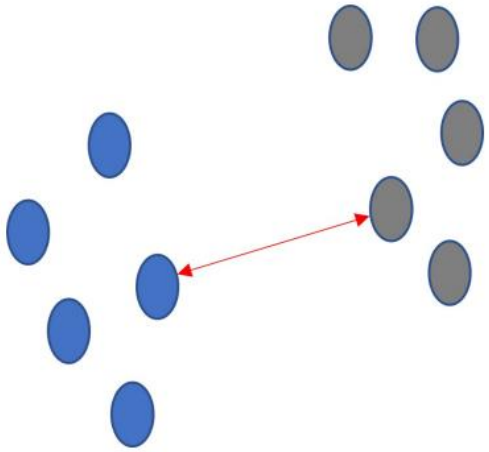


Figura 1.7 – Misura della distanza nel caso del legame semplice.
Fonte: nostre elaborazioni.

La tecnica del legame singolo gode di diverse importanti proprietà matematiche. Si può dimostrare infatti che la sequenza di partizioni che si ottengono è invariante rispetto a trasformazioni monotoniche delle variabili (Jardine N., 1971). Questa tecnica è quindi una delle poche che risultano insensibili a trasformazioni delle variabili che conservano l'ordine dei valori nella matrice di similarità.
Anche se gode di importanti proprietà matematiche la tecnica del legame singolo si rivela di regola di scarsa utilità. Essa ha infatti la tendenza a concatenare quasi tutti i casi in un unico grande gruppo; si mantengono separati solo piccoli gruppi o casi isolati.

- **Metodo del legame completo**: l'assunto base del metodo del legame completo (*complete linkage*), chiamato anche del vicino più lontano (*farthest neighbor technique*), è opposto a quello della tecnica del legame singolo: la distanza (similarità) fra due gruppi è identificata con quella fra i membri più lontani (o più simili). La distanza tra due gruppi A e B è definita come la distanza maggiore rilevata tra la coppia di individui (i, j) con $i \in A$, $j \in B$, in altri termini si considera il massimo delle $n_A \times n_B$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo:

$$d_{AB} = \max_{i \in A, j \in B} d_{ij}.$$

(1.3)

La tecnica manifesta la tendenza a identificare gruppi relativamente compatti, che risultano – nello spazio multidimensionale delle variabili – di forma ipersferica, composti da oggetti fortemente omogenei rispetto alle variabili impiegate (Ackerman M., 2016). La Figura 1.8 offre una chiara rappresentazione della misura della distanza nel contesto del metodo del legame completo.

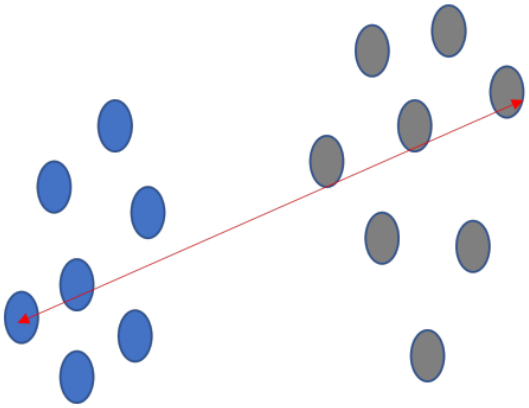


Figura 1.8 – Misura della distanza nel caso del legame completo.
Fonte: nostre elaborazioni.

- **Metodo del legame medio**: a differenza dei primi due criteri, con il metodo del legame medio (*average linkage*), per determinare la distanza fra due gruppi *A* e *B*, si prendono in considerazione tutte le distanze fra gli n_A oggetti membri del primo rispetto a tutti gli n_B oggetti membri del secondo.
  Con la tecnica del legame medio la distanza fra due gruppi si computa in base alla media aritmetica di tali distanze (Sokal R. R., 1958; McQuitty, 1964).

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}.$$

(1.4)

La Figura 1.9 offre una rappresentazione visiva della misura della distanza nel contesto del metodo del legame medio.

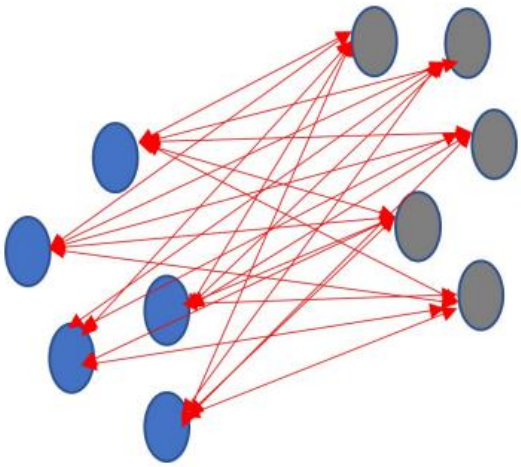


Figura 1.9 – Misura della distanza nel caso del legame medio.
Fonte: nostre elaborazioni.

- **Metodo del centroide**: la tecnica del centroide fa riferimento ad una rappresentazione spaziale degli oggetti da classificare, infatti, per ogni gruppo si definisce centroide il punto nello spazio multidimensionale che ha come coordinate la media aritmetica di tutti gli oggetti appartenenti al gruppo. La distanza fra i gruppi è in questo caso identificata dalla distanza fra i rispettivi centroidi (Jarman, 2020).
  La distanza tra due gruppi *A* e *B* di numerosità n_A e n_B è definita come la distanza tra i rispettivi centroidi (medie aritmetiche), $\bar{x}_A$ e $\bar{x}_B$.

$$d_{AB} = d(\bar{x}_A, \bar{x}_B).$$

(1.5)

  Dopo la fusione dei gruppi *A* e *B*, il centroide del nuovo gruppo formato *AB* è dato da:

$$\bar{x}_{AB} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n_A + n_B}.$$

(1.6)

  La Figura 1.10 presenta una visualizzazione grafica della procedura di calcolo della distanza nel metodo del centroide.

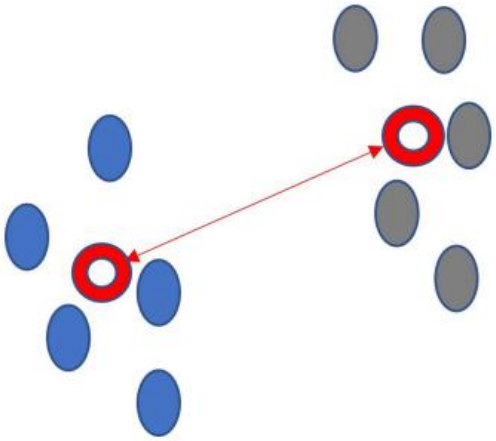


Figura 1.10 – Misura della distanza nel caso del metodo del centroide.
Fonte: nostre elaborazioni.

II metodo del centroide e il metodo del legame medio presentano delle interessanti analogie da considerare: il metodo del legame medio considera la media delle distanze tra le unità di ciascuno dei due gruppi, mentre il metodo del centroide calcola le medie di ciascun gruppo, e in seguito misura le distanze tra di esse.

- **Metodo di Ward**: nel metodo di Ward, l'obiettivo è minimizzare l'aumento della somma dei quadrati delle distanze quando due cluster vengono uniti (Ward, 1963). Questo significa che si cerca di minimizzare la varianza all'interno dei cluster combinati. L'algoritmo calcola la distanza tra i cluster utilizzando una misura di distanza come la distanza euclidea o la distanza di Manhattan. La fusione viene effettuata tra i due cluster che generano il minimo incremento della somma totale dei quadrati delle distanze intra-cluster:

$$E = \sum_{m=1}^{k} E_m,$$

(1.7)

dove:

$$E_m = \sum_{l=1}^{n_m} \sum_{h=1}^{p_h} (x_{ml,h} - \bar{x}_{m,h})^2,$$

(1.8)

in cui $\bar{x}_{m,h} = \left(\frac{1}{n_m}\right) \sum_{l=1}^{n_m} x_{ml,h}$ (la media del *m*-esimo cluster per la *h*-esima variabile), essendo $x_{ml,h}$ il punteggio sulla *h*-esima variabile ($h = 1, \dots, p$)

per l'l-esimo oggetto ($l = 1, \dots, n_m$) nel m-esimo cluster ($m = 1, \dots, k$). Questo aumento è proporzionale alla distanza euclidea al quadrato tra i centroidi dei cluster uniti (Brian S. Everitt, 2001).

La *Cluster Analysis* gerarchica offre numerosi vantaggi. Uno di questi è che non richiede di specificare a priori il numero di cluster desiderati. Grazie alla struttura gerarchica, è possibile esplorare diversi livelli di dettaglio e "tagliare" il dendrogramma al livello opportuno per ottenere il numero di cluster desiderato. Questa flessibilità consente di adattare l'analisi alle caratteristiche dei dati e agli obiettivi dell'analisi.

Inoltre, la *Cluster Analysis* gerarchica fornisce una rappresentazione visuale intuitiva della struttura dei dati attraverso il dendrogramma. Questo diagramma bidimensionale permette di identificare facilmente i cluster simili e di comprendere le relazioni tra di essi. Si tratta di una rappresentazione grafica che agevola l'interpretazione dei risultati e può aiutare a prendere decisioni informate (Bridges Jr, 1966).

Tuttavia, come in ogni metodo di clustering, ci sono alcune sfide da considerare. La *Cluster Analysis* gerarchica può essere più difficoltosa da gestire con dataset enormi, poiché richiede un maggiore sforzo computazionale e di memoria. Inoltre, può essere sensibile al rumore e ai valori anomali presenti nei dati, il che può influenzare la qualità dei cluster formati. Sensibilità che può essere attenuata utilizzando un legame di tipo singolo (Jarman, 2020).

Un ulteriore problema connesso al metodo gerarchico agglomerativo consiste nel fatto che i metodi del collegamento singolo, completo e medio falliscono nell'identificare cluster di forma non sferica. Tuttavia, il metodo del collegamento singolo presenta però una caratteristica utile: può essere utilizzato per individuare punti singolari, noti come *outliers*. Questi punti vengono considerati singoli elementi e possono essere scartati se sono sufficientemente distanti dal loro elemento confinante più vicino.

D'altro canto, l'utilizzo del legame semplice può portare al problema del *chaining* o concatenamento. Questo fenomeno si verifica quando cluster tra loro distanti ma connessi da una fila di punti intermedi a causa della presenza di rumore nei dati di input, vengono erroneamente uniti (Brian S. Everitt, 2001). Le Figure 1.11 e 1.12 illustrano la situazione esposta.

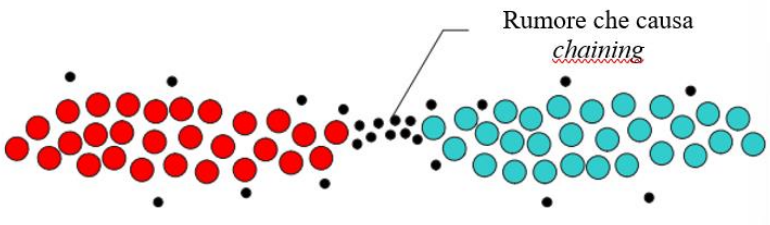


Figura 1.11 – Fenomeno del Chaining.
Fonte: nostre elaborazioni

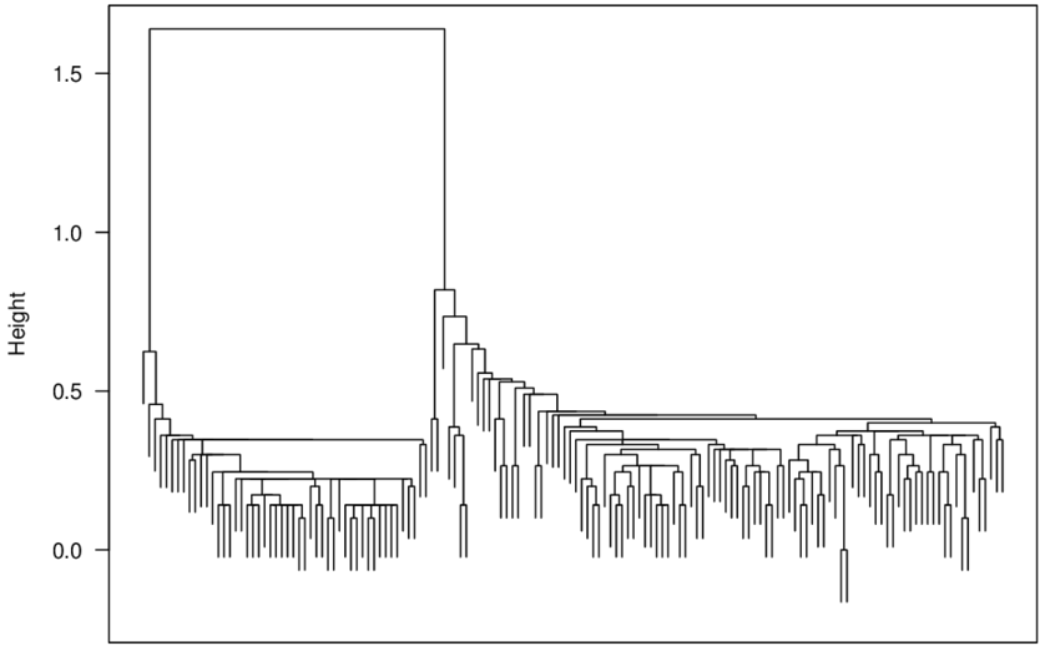


Figura 1.12 – Dendogramma affetto dal problema del Chaining.
Fonte: nostre elaborazioni

Un'altra sfida significativa riguarda la determinazione del numero ottimale di cluster, che rappresenta il punto cruciale in cui l'algoritmo deve concludere la sua esecuzione. Questa determinazione può essere complessa e richiede l'applicazione di metodi o criteri specifici per valutare e scegliere il numero appropriato di cluster.

I problemi menzionati e le possibili soluzioni saranno affrontati in modo approfondito nel paragrafo (1.4) dedicato ai "Limiti delle tradizionali tecniche di segmentazioni e come superarli con lo *Spectral Clustering*". In questo paragrafo, verranno esaminati dettagliatamente i limiti delle tecniche di clustering tradizionali, come il clustering gerarchico agglomerativo, e verrà presentato lo *Spectral Clustering* come un possibile approccio per superare tali limitazioni.

1.2.2 Cluster Analysis non gerarchica

Oltre alla *Cluster Analysis* gerarchica, esiste un'altra importante tecnica di clustering chiamata *Cluster Analysis* non gerarchica o partizionale.

Le tecniche di *Cluster Analysis* non gerarchica si concentrano principalmente sulla suddivisione delle unità statistiche in gruppi basandosi su una misura di similarità o dissimilarità. Prima di avviare il processo di clustering, vengono definite le caratteristiche o le variabili rilevanti per la creazione dei gruppi. Successivamente, si seleziona un algoritmo di clustering appropriato per assegnare gli oggetti a gruppi in base alla loro similarità.

A differenza dell'approccio gerarchico, la *Cluster Analysis* non gerarchica mira a suddividere direttamente gli oggetti in un insieme predefinito di cluster senza creare una struttura gerarchica ad albero. Invece di creare sottogruppi all'interno di gruppi più ampi, la *Cluster Analysis* non gerarchica fornisce una singola partizione delle *n* unità in *k* gruppi, con valore di *k* fissato a propri. L'algoritmo esplora tutte le possibili partizioni degli *n* soggetti in *k* gruppi e determina la partizione migliore in assoluto, quella che massimizza la coesione interna e massimizza la separazione esterna tra i gruppi (Likas A., 2003).

L'obiettivo finale dell'approccio non gerarchico è ottenere una suddivisione chiara e rilevante delle unità statistiche in gruppi omogenei, consentendo una migliore comprensione delle strutture e delle relazioni presenti nei dati. Questo tipo di clustering offre un approccio estremamente flessibile per l'organizzazione dei dati in cluster, consentendo di specificare il numero desiderato di gruppi e cerca attivamente di trovare la soluzione ottimale attraverso l'utilizzo di criteri di valutazione appropriati.

Uno degli algoritmi più comuni utilizzati nella *Cluster Analysis* non gerarchica è il *k-means* (Hastie T., 2001; MacQueen, 1967; Hartigan J. A., 1979), il quale risulta particolarmente utile quando si dispone di un numero elevato di unità statistiche.[3]

Il *k-means* è un algoritmo di clustering iterativo appartenente alla famiglia dei *centroid-based model*. Gli algoritmi di questa famiglia sono del tipo iterativo e i cluster sono rappresentati da un vettore centrale, che può non appartenere all'insieme di dati da analizzare.

L'algoritmo *k-means* richiede di specificare in anticipo il numero desiderato di

[3] Per il contesto in questione, si ritiene che un numero elevato di unità statistiche corrisponda a una quantità superiore a 100.

cluster (*k*) e opera attraverso un processo iterativo. Durante ogni iterazione, assegna le unità statistiche a un cluster e aggiorna i centroidi[4] dei cluster in base alla minimizzazione di una funzione di distanza,[5] come ad esempio la distanza euclidea tra i punti e il centro del cluster. Concretamente, l'algoritmo *k-mean*s cerca soluzioni localmente ottimali utilizzando come criterio di raggruppamento *F*, la somma della distanza al quadrato (L^2) tra ciascun elemento e il centroide del cluster più vicino. Questo criterio è talvolta indicato come criterio dell'errore quadratico. Pertanto, ne consegue che (Pena J.M., 1999):

$$F = \sum_{m=1}^{k} \sum_{l=1}^{n_m} \|x_{ml} - \bar{x}_m\|.$$

(1.9)

Dove, *k* è il numero di cluster, n_m il numero di oggetti del cluster *m*, x_{ml} l'l-esimo oggetto dell'*m*-esimo cluster e $\bar{x}_m$ è il centroide dell'*m*-esimo cluster che si definisce come (Pena J.M., 1999):

$$\bar{x}_m = \frac{1}{n_m} \sum_{l=1}^{n_m} x_{ml},$$

(1.20)

dove, $m = 1, \dots, k$. L'algoritmo converge quando i centroidi dei cluster non subiscono più modifiche o quando viene soddisfatto un criterio di arresto predefinito (Likas A., 2003).

L'algoritmo *k-means* può essere riassunto come segue:

1. **Definizione del numero di cluster**: si decide il numero *k* di cluster in cui verrà suddiviso il dataser.
2. **Inizializzazione dei centroidi**: si scelgono i *k* centroidi (diversi) in modo casuale nello spazio dei dati, e ci si assicura che siano abbastanza lontani tra di loro, questa condizione è utile per la convergenza.
3. **Calcolo della distanza**: viene calcolata la distanza di ogni oggetto del dataset rispetto ai centroidi. La distanza più utilizzata è quella euclidea, ma si possono utilizzare altre definizioni di distanza.
4. **Assegnazione degli oggetti ai cluster**: gli oggetti dei cluster vengono assegnati in base alla vicinanza ai centroidi.
5. **Aggiornamento dei centroidi**: si calcolano i nuovi centroidi considerando gli elementi assegnati ai cluster, come la media delle posizioni di tutti gli

[4] *k* punti definiti nello spazio *p*-dimensionale in cui sono definiti i punti corrispondenti alle unità statistiche da raggruppare.

[5] Quasi tutti i metodi di clustering partizionale si basano sull'idea di ottimizzare una funzione *F*, definita "criterio di clustering", che si spera traduca le nozioni intuitive sui cluster in una formula matematica ragionevole.

oggetti appartenenti al cluster.

6. **Iterazione e convergenza**: si itera a partire dal punto (3) fino a quando non ci sono più variazioni.

Il vantaggio principale di questa categoria risiede nella semplicità dell'implementazione. Dal punto di vista operativo, questo approccio è una scorciatoia lontana dall'idea di esplorare tutte le possibili partizioni degli *n* soggetti in *k* gruppi. Nonostante ciò, il risultato ottenuto con il *k-means*, essendo una tecnica iterativa, sarà particolarmente sensibile condizioni iniziali di partenza, ossia ai centroidi scelti all'inizio dell'iterazione (Pena J.M., 1999).

Nonostante l'ampia gamma di applicazioni in cui viene utilizzato, l'algoritmo *k-means* non è esente da inconvenienti. Dal punto di vista della qualità della soluzione, il raggiungimento dell'ottimo globale non viene garantito, ovvero non garantisce di trovare la partizione migliore che massimizza la coesione interna e la separazione esterna. Infatti come illustrato nelle Figure 1.13 e 1.14, la scelta dei centroidi iniziali può portare a una partizione dei dati differente.

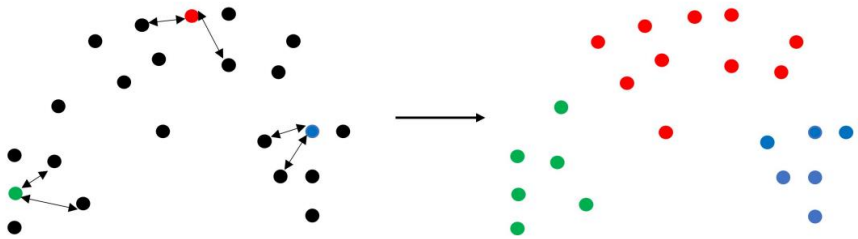


Figura 1.13 – Esempio di risultato ottenuto con il k-means nel caso di k=3. Fonte: nostre elaborazioni.

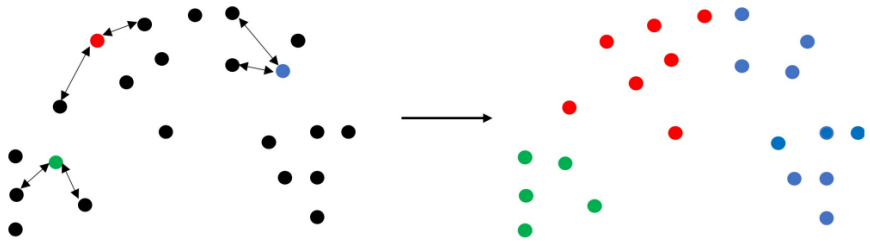


Figura 1.14 – Altro esempio di risultato ottenuto con il k-means nel caso di k=3. Fonte: nostre elaborazioni.

Nonostante ciò, l'algoritmo *k-means* offre la garanzia di convergere verso una delle migliori partizioni possibili, consentendo un notevole risparmio di tempo rispetto all'esplorazione completa delle combinazioni.

Come molti metodi di clustering, l'algoritmo *k-means* presuppone che il numero di cluster (*k*) nel database sia noto a priori, il che, ovviamente, non è necessariamente vero nelle applicazioni reali (Pena J.M., 1999). La scelta di specificare in anticipo il numero desiderato di cluster (*k*) può essere vantaggiosa in molte applicazioni in cui l'obiettivo è anche quello di scoprire il numero "corretto" di cluster. Tuttavia, questa operazione può risultare complessa e soggettiva, poiché spesso manca una conoscenza predefinita sul numero ottimale di cluster da utilizzare. Una scelta errata del valore di *k* può portare a una suddivisione inefficace o inaccurata dei dati.

Una buona pratica è eseguire l'algoritmo *k-means* per diversi valori di *k* e confrontare le variazioni tra i risultati impiegando criteri appropriati per selezionare il valore più adatto di *k* (Milligan G. W., 1985). Inoltre, la validità del clustering può essere valutata considerando la distanza tra i cluster e la varianza all'interno dei cluster. Un buon clustering è caratterizzato da una grande distanza tra i cluster e una bassa varianza all'interno di ciascun cluster.

Il problema delle condizioni iniziali di partenza non è esclusivo dell'algoritmo *k-means*, ma condiviso con molti algoritmi di clustering che funzionano come una strategia di *hill-climbing* il cui comportamento deterministico porta a un minimo locale dipendente dalla soluzione iniziale e dall'ordine delle istanze. Sebbene non vi sia alcuna garanzia di raggiungere un minimo globale, almeno la convergenza dell'algoritmo *k-means* è assicurata (Selim S. Z., 1984).

Milligan (1980) mostra la forte dipendenza dell'algoritmo *k-means* dalla clusterizzazione iniziale e suggerisce che è possibile ottenere buone strutture finali di cluster utilizzando il metodo gerarchico di Ward (1963) per fornire all'algoritmo *k-means* i cluster iniziali. Fisher (1996) propone di creare i cluster iniziali costruendo un clustering gerarchico iniziale basato sul lavoro di Fisher (1987). Higgs et al. (1997) e Snarey et al. (1997) suggeriscono di utilizzare un algoritmo *MaxMin* per selezionare un sottoinsieme del database originale come centroidi iniziali per stabilire i cluster iniziali. In un lavoro, Meila e Heckerman (1998) presentano alcuni risultati sperimentali di un'istanza dell'algoritmo EM che ricorda il *k-means* con tre diversi metodi di inizializzazione (uno dei quali è un metodo di clustering agglomerativo gerarchico).

La maggior parte dei metodi di inizializzazione che abbiamo menzionato sopra non sono solo metodi di inizializzazione. Sono essi stessi metodi di clustering e, se utilizzati con l'algoritmo *k-means*, danno luogo a un algoritmo di clustering ibrido. Pertanto, questi metodi di inizializzazione soffrono dello stesso problema dell'algoritmo *k-means* e devono essere dotati di un clustering iniziale (Cao, 2009).

1.3 Introduzione allo Spectral Clustering

Negli ultimi anni, lo *Spectral Clustering* si è affermato come una valida alternativa alle precedenti tecniche di Clustering. I motivi sono vari, innanzitutto la semplicità dell'implementazione, per la quale è sufficiente solamente una libreria di algebra lineare, e, in secondo luogo, risultati sorprendenti che superano molte difficoltà considerate insormontabili (von Luxburg, 2007).

Nonostante la sua apparente semplicità, lo *Spectral Clustering* si basa su una teoria complessa, la Teoria dei Grafi, una materia vasta e profonda che coinvolge diverse discipline scientifiche, e sulle proprietà degli autovalori e degli autovettori delle matrici associate ai dati.

Lo *Spectral Clustering* permette di raggruppare le unità statistiche in cluster coerenti. Risulta particolarmente utile quando i dati presentano una struttura complessa e non possono essere facilmente separati in modo lineare nello spazio delle caratteristiche.

Lo *Spectral Clustering* è un algoritmo che permette di affrontare la complessità di dataset ad alta dimensionalità sfruttando le informazioni contenute negli autovalori e autovettori di specifiche matrici costruite a partire dal grafo di similarità o direttamente dal dataset. Questo approccio consente di ridurre la dimensionalità dei dati, favorendo una successiva esecuzione efficiente del clustering.

La costruzione del grafo di similarità non è univoca, tuttavia è fondamentale che tutti i metodi rispettino la relazione locale tra i punti dati. Esistono diverse tipologie di grafi utilizzate nello *Spectral Clustering*, tra cui (von Luxburg, 2007):

- ***ε-neighborhood Graph***: questo tipo di grafo collega i punti che si trovano entro una certa distanza (ε) l'uno dall'altro.
- ***Fully Connected Graph***: in questo caso, il grafo collega tutti i punti del dataset, assegnando pesi alle connessioni basati sulla similarità.
- ***k-Nearest Neighbor Graph***: questo tipo di grafo collega ogni punto ai suoi *k* punti più vicini nel dataset.

Indipendentemente dalla tipologia di grafo utilizzata, è possibile rappresentare quest'ultimo mediante una matrice di adiacenza ($\boldsymbol{W}$), chiamata anche matrice di similarità, dove le righe e le colonne rappresentano i vertici del grafo, mentre i valori all'interno della matrice indicando le connessioni tra i vertici, esprimendo i rispettivi pesi. Nel caso del *ε-neighborhood graph*, il valore sarà semplicemente 1

se i vertici sono connessi entro una determinata distanza (ε). Se due vertici non sono collegati, il valore nella matrice di adiacenza sarà nullo.

Dal grafo è inoltre possibile derivare la matrice dei gradi (*Degree matrix*, $\boldsymbol{D}$), che è una matrice diagonale che fornisce informazioni sul numero di collegamenti di ciascun vertice.

Attraverso l'utilizzo di queste matrici, è possibile ricavare la matrice Laplaciana, sebbene la sua definizione non sia unica nella letteratura. Esiste infatti un intero campo di studio dedicato allo studio di queste matrici, noto come la Teoria Spettrale dei Grafi (Chung, 1997). Di seguito si riportano le varie tipologie di matrici Laplaciane (von Luxburg, 2007):

- Matrice Laplaciana non normalizzata ($\boldsymbol{L}$);
- Matrice Laplaciana normalizzata simmetrica ($\boldsymbol{L_{sym}}$);
- Matrice Laplaciana normalizzata non simmetrica ($\boldsymbol{L_{rw}}$).

Queste diverse formulazioni delle matrici Laplaciane offrono differenti approcci per la rappresentazione dei grafi e per la gestione delle informazioni di similarità (affinità) tra i vertici.

La scelta del tipo di grafo e della matrice Laplaciana da utilizzare per rappresentarlo non segue delle regole rigide, ma deve essere valutata caso per caso, in base alle caratteristiche del dataset. È importante selezionare la combinazione di grafo e matrice Laplaciana che si adatta meglio alle peculiarità e alla struttura dei dati per ottenere risultati ottimali nel clustering (von Luxburg, 2007).

Gli autovalori nulli della matrice Laplaciana forniscono informazioni sul numero di gruppi di elementi connessi nel grafo. Questo può essere utile nel caso in cui il numero di cluster desiderato non sia noto a priori. Gli autovalori non nulli, ma con valori vicini a zero, indicano la presenza di collegamenti molto "deboli" che potrebbero potenzialmente portare alla creazione di ulteriori cluster. Di solito, è buona norma cercare il primo "salto" significativo tra autovalori per avere un'idea del numero di cluster da utilizzare (von Luxburg, 2007). Gli autovettori associati ai *k* autovalori più piccoli andranno a definire lo spazio dimensionale nel quale i dati reali verranno mappati per l'esecuzione del clustering.

Successivamente, si procede alla decomposizione spettrale della matrice di similarità. Questo processo comporta il calcolo degli autovalori e degli autovettori della matrice. Gli autovalori rappresentano l'importanza relativa delle diverse modalità di raggruppamento degli oggetti, mentre gli autovettori definiscono le

combinazioni lineari dei dati che corrispondono a tali modalità (Chung, 1997).

Dopo aver ottenuto gli autovettori, si selezionano i primi *k* autovettori associati agli autovalori più grandi, dove *k* rappresenta il numero desiderato di cluster. Questi autovettori vengono utilizzati come nuove rappresentazioni dei dati.

Successivamente, si applica un algoritmo di clustering, come il *k-means* o un'altra tecnica di partizionamento, ai nuovi dati rappresentati dagli autovettori. L'algoritmo di clustering assegna gli oggetti ai cluster sulla base della loro vicinanza nello spazio delle caratteristiche definito dagli autovettori selezionati (Ng, Jordan, & Weiss, 2001).

1.3.1 Algoritmo di Spectral Clustering

Assumiamo che i nostri dati siano costituiti da *n* "osservazioni" $x_1, \dots, x_n$ che possono essere oggetti arbitrari. Misuriamo le loro somiglianze a coppie $x_1 = s(x_i, x_j)$ mediante una funzione di similarità simmetrica e non negativa e denotiamo la matrice di similarità corrispondente con $\boldsymbol{S} = (s_{ij})_{i,j=1,\dots,n}$.

Esistono diverse varianti dello *Spectral Clustering*, tra cui una versione non normalizzata e due versioni normalizzate (a seconda di quale delle matrici Laplaciane normalizzate del grafo viene utilizzata) (von Luxburg, 2007).

In questo contesto di ricerca sullo *Spectral Cluste*ring, scegliamo di adottare l'algoritmo di *Normalized Spectral Clustering* proposto da Ng et al. (2001), che utilizza la matrice Laplaciana normalizzata simmetrica $\boldsymbol{L_{sym}}$. L'algoritmo, fondato sui principi della Teoria dei Grafi e dell'analisi degli autovalori, rappresenta un approccio efficace per la suddivisione di dati non etichettati in cluster coerenti e rilevanti.

L'algoritmo di *Normalized Spectral Clustering*, proposto da Ng et al. (2002), definisce una serie di fasi che guidano l'implementazione del metodo. Queste fasi rappresentano un processo strutturato e ben definito per ottenere una rappresentazione spettrale dei dati e l'individuazione dei cluster desiderati.

Tale algoritmo presenta le seguenti fasi:

- **Input**: l'algoritmo richiede due input principali:
  - La matrice di similarità $\boldsymbol{S}$ di dimensione $n \times n$, che rappresenta la similarità tra i punti del dataset. La similarità può essere calcolata utilizzando diverse misure, come la distanza euclidea, a seconda del tipo di dati e del dominio del problema.
  - Il numero *k* di cluster desiderati per la suddivisione dei dati non etichettati,

ovvero il numero di gruppi in cui si desidera raggruppare i punti.

Nel Capitolo 2, approfondiremo in dettaglio il calcolo della matrice di similarità e le diverse misure utilizzate.

- **Costruzione del grafo di similarità**: per trasformare i dati in una rappresentazione grafica, è necessario costruire un grafo di similarità utilizzando una delle tecniche descritte nella sezione (2.5), come il *ε-neighborhood graph*, il *fully connected graph* o il *k-nearest neighbor graph*. Questo grafo cattura le relazioni di similarità tra i punti del dataset. Otteniamo una matrice di adiacenza pesata $\boldsymbol{W}$:

$$\boldsymbol{W} = (w_{ij})_{i,j=1,\dots,n},$$

(1.13)

dove $w_{ij} \geq 0$ denota la similarità (affinità) tra x_i e x_j.
La matrice di adiacenza simmetrica $n \times n$ non negativa $\boldsymbol{W}$ associata al grafo di similarità è la cosiddetta matrice di similarità (Favati & al., 2020). [6]

- **Calcolo della matrice Laplaciana normalizzata**: si calcola la matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$ a partire dalla matrice di adiacenza $\boldsymbol{W}$. la matrice Laplaciana normalizzata è una misura che cattura la struttura del grafo di similarità ed è definita come:

$$\boldsymbol{L_{sym}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}},$$ [7]

(1.42)

Dove $\boldsymbol{I}$ è la matrice identità di dimensione $n \times n$,[8] $\boldsymbol{W}$ è la matrice di adiacenza pesata del grafo di similarità e $\boldsymbol{D}$ è la matrice diagonale dei gradi dei vertici del grafo.

La matrice dei gradi $\boldsymbol{D}$[9] è definita come la matrice diagonale con i gradi

[6] La matrice di adiacenza $\boldsymbol{W}$ viene poi utilizzata come matrice di affinità nel processo di *Spectral Clustering*, dove gli algoritmi cercano di identificare gruppi di punti simili o correlati all'interno del grafo (Bhissy, Faleet, & Ashour, 2014).
[7] I lettori che hanno familiarità con la teoria spettrale dei grafi potrebbero essere abituati a vedere la matrice Laplaciana come $1 - \boldsymbol{L}$. Tuttavia, per semplificare la nostra discussione successiva e poiché ciò influisce solo sugli autovalori (da λ_i a $1 - \lambda_i$) e non sugli autovettori, preferiamo utilizzare $\boldsymbol{L}$ (Ng, Jordan, & Weiss, 2001).
[8] La matrice identità, indicata con $\boldsymbol{I}$, è una matrice quadrata in cui tutti gli elementi della diagonale principale sono pari a 1, mentre tutti gli altri elementi sono pari a 0. La matrice identità è importante perché rappresenta un elemento neutro rispetto alla moltiplicazione di matrici. La presenza della matrice identità nella definizione delle matrici Laplaciane normalizzate dei grafi serve a introdurre le opportune normalizzazioni per le matrici di adiacenza e dei gradi dei vertici, al fine di ottenere misure rilevanti per la struttura del grafo nella teoria spettrale.
[9] La matrice $\boldsymbol{D}$ è una matrice diagonale i cui elementi (i, i) sono la somma della riga i-esima di $\boldsymbol{W}$.

$d_1, \dots, d_n$ sulla diagonale, dove

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

(1.53)

Il calcolo della matrice Laplaciana normalizzata sarà approfondito nel Capitolo 3, in cui verranno presentati i dettagli matematici e le intuizioni dietro questa misura.

- **Calcolo degli autovalori e degli autovettori**: si calcolano i primi *k* autovalori[10] $\lambda_1, \dots, \lambda_k$ e i corrispondenti autovettori $u_1, \dots, u_k$ della matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$. Nel Capitolo 3, approfondiremo in dettaglio il calcolo degli autovalori e degli autovettori della matrice Laplaciana normalizzata.
- **Creazione della matrice dei punti**: si costruisce una matrice ***U*** di dimensione $n \times k$, in cui ogni colonna rappresenta un autovettore corrispondente a un autovalore. Questa matrice cattura le informazioni di raggruppamento contenute negli autovettori. Il calcolo della matrice dei punti ***U*** sarà approfondito nel Capitolo 3.
- **Normalizzazione delle righe**: si normalizzano le righe della matrice ***U*** per ottenere una nuova matrice ***T*** di dimensione $n \times k$, in cui ogni riga ha norma 1. Questo processo di normalizzazione viene eseguito come segue:
  per ogni elemento t_{ij} in ***T***, impostiamo:

$$t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{\frac{1}{2}}},$$

(1.64)

  per $i = 1, \dots, n$ e $j = 1, \dots, k$. Dove, u_{ij} è l'elemento corrispondente nella matrice ***U***. Nel Capitolo 3, approfondiremo in dettaglio il calcolo della matrice normalizzata ***T***.
- **Creazione dei vettori di rappresentazione**: per ogni riga *i* della matrice ***T***, si ottiene un vettore y_j di dimensione *k* che rappresenta il punto corrispondente. Si ottiene quindi un insieme di vettori $(y_j)_{i=1,\dots,n}$, i quali rappresentano i punti del dataset trasformati nello spazio dei *k* autovettori normalizzati.
- **Clustering dei punti**: si applica l'algoritmo *k-means* sui punti $(y_j)_{i=1,\dots,n}$ utilizzando *k* cluster. L'algoritmo *k-means*, come descritto nella sezione 1.2.2, mira a identificare *k* centroidi che minimizzano la somma dei quadrati delle distanze tra i punti e i centroidi dei rispettivi cluster. Come risultato, si

[10] Con "i primi *k* autovalori" ci riferiamo agli autovalori corrispondenti ai *k* autovalori più piccoli.

ottengono k cluster $C_1, \dots, C_k$.

- **Output**: fornisce come output i cluster $A_1, \dots, A_k$, dove ogni A_i contiene gli indici dei punti che appartengono al cluster C_i. In altre parole, $A_i = \{j \mid y_j \in C_i\}$ (von Luxburg, 2007).

Nell'algoritmo di *Spectral Clustering* la strategia principale consiste quindi nel cambiare la rappresentazione dei punti astratti x_i in punti $y_j \in \mathbb{R}^k$. È grazie alle proprietà delle rappresentazioni dei grafi mediante matrici Laplaciane che questo cambio di rappresentazione risulta utile. Questo cambio di rappresentazione potenzia le proprietà di raggruppamento dei dati, consentendo di individuare facilmente i cluster nella nuova rappresentazione. In particolare, l'algoritmo di clustering *k-means* semplice non ha difficoltà a individuare i cluster in questa nuova rappresentazione (von Luxburg, 2007).

Il metodo proposto da Ng et al. (2001) è stato ampiamente studiato e ha dimostrato la sua efficacia in diverse applicazioni di clustering. Nel contesto della nostra ricerca, impiegheremo questo algoritmo come fondamento per la realizzazione dell'implementazione di *Spectral Clustering* delineata nel Capitolo 4.

1.4 Limiti delle tradizionali tecniche di segmentazioni e come superarli con lo Spectral Clustering

Nel campo dell'analisi dei dati, la segmentazione riveste un ruolo fondamentale nel rivelare pattern, strutture nascoste e relazioni all'interno di un dataset. Tuttavia le tradizionali tecniche di segmentazione, come il clustering basato sul *k-means* o sugli algoritmi gerarchici, presentano alcune limitazioni che possono influire sulla loro efficacia in determinati contesti. È importante comprendere queste limitazioni al fine di esplorare e utilizzare approcci alternativi, come lo *Spectral Clustering*, che possano superarle e migliorare i risultati dell'analisi (Ng, Jordan, & Weiss, 2001).

Rispetto agli "algoritmi tradizionali", come il *k-means* o il *single linkage*, lo *Spectral Clustering* offre diversi vantaggi fondamentali. Questo metodo presenta risultati spesso superiori, è implementato in modo semplice e può essere risolto in modo efficiente utilizzando metodi standard di algebra lineare (von Luxburg, 2007).

Una delle principali limitazioni delle tecniche classiche è la necessità di specificare in anticipo il numero di cluster desiderato, ad esempio attraverso il

parametro "*k*" nel *k-means*. Sebbene l'algoritmo *k-means* sia efficace in molte situazioni, la scelta del numero di cluster può essere problematica e influenzare notevolmente i risultati del clustering dimostrandosi dunque una vera e propria sfida in molte applicazioni reali. Ad esempio, potrebbe essere difficile determinare il numero di cluster in un set di dati complesso o in continuo cambiamento, come i flussi di dati in tempo reale o le reti sociali. Inoltre, la scelta del numero di cluster può variare a seconda degli obiettivi dell'analisi o dell'interpretazione dei dati (Davidson, 2002; Brian S. Everitt, 2001; MacQueen, 1967).

Per superare questa limitazione, è possibile adottare un approccio basato sullo *Spectral Clustering*. Questa tecnica si basa su una decomposizione spettrale della matrice di similarità per identificare i cluster naturali presenti nei dati, senza richiedere la specifica del numero di cluster a priori. Ciò rende lo *Spectral Clustering* più adattabile a situazioni in cui il numero di cluster non è noto o può variare (Ng, Jordan, & Weiss, 2001).

Un altro limite delle tecniche classiche è la loro sensibilità alle forme dei cluster. Le tecniche di clustering come il *k-means* tendono a identificare solo cluster sferici o ellissoidali, il che può essere limitante quando si lavora con dati che contengono cluster di forme arbitrarie o con contorni complessi. Questa restrizione può ridurre la capacità delle tecniche classiche di adattarsi a situazioni in cui i cluster hanno forme non convenzionali (Davidson, 2002).

Al contrario, lo *Spectral Clustering*, grazie alla sua capacità di individuare strutture non lineari, è meno vincolato dalle forme dei cluster e può identificare cluster di forme più complesse. Questa capacità è particolarmente vantaggiosa quando si affrontano problemi di segmentazione che richiedono una maggiore flessibilità nella forma dei cluster e rende lo *Spectral Clustering* adatto a molte applicazioni in cui la forma dei cluster è un elemento cruciale per l'analisi dei dati.

Le tecniche classiche possono essere inoltre influenzate negativamente dalla presenza di rumore nei dati o da punti anomali, il che può compromettere l'accuratezza e l'affidabilità del processo di clustering. Questi punti possono alterare la stima dei centroidi o influenzare il processo di assegnazione dei punti ai cluster, portando a risultati indesiderati (Davidson, 2002).

Lo *Spectral Clustering*, grazie alla sua robustezza alle distorsioni locali, può gestire meglio il rumore e i punti anomali. La robustezza dello *Spectral Clustering* alle distorsioni locali può essere attribuita alla sua capacità di considerare l'intera struttura dei dati, invece di basarsi solo su informazioni locali limitate. Questo

approccio rende l'algoritmo meno suscettibile alle influenze locali e può produrre segmentazioni più coerenti anche in presenza di rumore (Jarman, 2020).

Un ulteriore limite delle tecniche classiche è la loro gestione di dataset di grandi dimensioni. Algoritmi tradizionali come il *k-means* possono richiedere un numero elevato di calcoli e risorse computazionali, rendendo il processo di segmentazione inefficiente per dataset di grandi dimensioni. Questa limitazione può rendere difficile l'applicazione delle tecniche classiche a problemi di clustering su larga scala (Jimenez, 1997).

Al contrario, lo *Spectral Clustering* può affrontare meglio questa sfida. Lo *Spectral Clustering* sfrutta la matrice di similarità, che può essere calcolata in modo efficiente anche per dataset di grandi dimensioni. Questa matrice può essere utilizzata per ridurre il problema del clustering a una dimensione inferiore, consentendo di lavorare su dati più compatti e facilitando il processo di segmentazione (Ng, Jordan, & Weiss, 2001).

Inoltre, lo *Spectral Clustering* può essere combinato con tecniche di riduzione della dimensionalità come la Decomposizione ai Valori Singolari (*Singular Value Decomposition*, SVD) per selezionare solo le caratteristiche più informative e ridurre ulteriormente la complessità computazionale.

Lo *Spectral Clustering* ha dunque dimostrato di superare alcune delle sfide delle tradizionali tecniche aprendo nuove prospettive nell'analisi dei dati e nella segmentazione. La sua flessibilità nella segmentazione, la gestione delle forme complesse dei cluster, la robustezza al rumore e ai dati anomali e la scalabilità per dataset di grandi dimensioni rendono lo *Spectral Clustering* un'opzione promettente per l'analisi dei dati e la segmentazione in diversi contesti, offrendo nuove opportunità per la comprensione e l'interpretazione dei pattern presenti nei dati.

Nei capitoli successivi, ci immergeremo nell'algoritmo dello *Spectral Clustering*, esplorando dettagliatamente il suo funzionamento e le sue varianti. Approfondiremo le strategie per la selezione dei parametri, in modo da ottenere una comprensione più completa e un'applicazione ottimizzata di questa avanzata tecnica di clustering.

2. Grafo di similarità

Nel contesto dell'analisi esplorativa dei dati, il clustering rappresenta una tecnica fondamentale utilizzata per identificare e raggruppare insieme dati simili. In varie discipline scientifiche, come la statistica, l'informatica, la biologia, le scienze sociali, la psicologia e il marketing, il clustering trova applicazione per scoprire pattern nascosti e per condurre analisi approfondite.

Questo capitolo rappresenta un punto di partenza cruciale per l'approfondimento dello *Spectral Clustering*, concentrandosi sulla sua fondamentale componente: il grafo di similarità. Il grafo di similarità consente di rappresentare i dati come un insieme di vertici interconnessi da archi, in cui la forza degli archi riflette la misura di similarità tra i dati corrispondenti. Questo concetto svolge un ruolo fondamentale nello *Spectral Clustering*, poiché permette di rivelare pattern nascosti e strutture complesse presenti nei dati, offrendo una maggiore flessibilità nell'individuazione dei cluster e superando le limitazioni dei metodi di clustering tradizionali

Durante il capitolo, saranno esplorati gradualmente i concetti matematici fondamentali legati allo *Spectral Clustering*, con particolare attenzione al grafo di similarità. Saranno presentate diverse modalità di costruzione del grafo, comprese le misure di similarità comunemente utilizzate nei vari ambiti scientifici. Questo permetterà di comprendere come il grafo di similarità possa catturare le relazioni tra i dati e fornire una rappresentazione comprensibile e interpretabile per le successive analisi.

L'obiettivo principale di questo capitolo è quello di fornire una base solida per la comprensione e l'applicazione dello *Spectral Clustering* utilizzando il grafo di similarità come strumento chiave. Attraverso una comprensione approfondita di questi concetti, sarà possibile apprezzare l'efficacia dello *Spectral Clustering* nelle analisi esplorative dei dati e nelle scoperte di pattern rilevanti. Ciò contribuirà a migliorare la comprensione dei fenomeni studiati in diverse discipline scientifiche, inclusa l'applicazione specifica nel campo del marketing.

2.1 Matrice di similarità

Un aspetto fondamentale nella *Cluster Analysis* è rappresentato dalla matrice di similarità, la quale gioca un ruolo cruciale nel processo di clusterizzazione e influenza direttamente i risultati ottenuti.

La matrice di similarità è una rappresentazione strutturata che cattura il grado di

similarità o dissimilarità tra le diverse coppie di unità statistiche, offrendo una visione comprensiva delle strutture sottostanti i dati. L'obiettivo primario della matrice di similarità è quello di quantificare la similarità tra le unità statistiche, consentendo di individuare gruppi omogenei e strutture nascoste all'interno del dataset.

Essa può essere rappresentata come un'equazione $n \times n$, in cui n rappresenta il numero di unità statistiche, e i suoi elementi indicano la misura di similarità tra le diverse coppie di unità statistiche.

La generazione della matrice di similarità richiede la scelta di adeguate misure di similarità o dissimilarità, a seconda delle caratteristiche delle variabili coinvolte nello studio. Esistono diverse misure di similarità comunemente utilizzate, come la distanza euclidea, la correlazione di Pearson e molte altre (Bandyopadhyay & Saha, 2013). La scelta della misura dipende dalla natura delle variabili e dagli obiettivi dell'analisi.

Il seguente paragrafo avrà come obiettivo quello di richiamare il concetto di similarità, fornendo una panoramica delle diverse definizioni e misure di similarità utilizzate nell'analisi dei cluster. Sarà dedicata particolare attenzione alla trattazione delle variabili ordinali, le quali presentano specifiche sfide in termini di misurazione di similarità.

Successivamente, esploreremo le misure di similarità più comunemente utilizzate per lo *Spectral Clustering* di variabili ordinali. Questa analisi ci permetterà di valutare quale misura di similarità sia più appropriata per il nostro contesto di studio e per gli obiettivi specifici della nostra ricerca.

La matrice di similarità, generata utilizzando le adeguate misure di similarità, sarà la base per l'applicazione degli algoritmi di clustering. Questi algoritmi utilizzeranno la matrice di similarità per identificare i gruppi omogenei all'interno dei dati analizzati. L'obiettivo principale è quello di creare cluster che siano internamente coerenti e al contempo separati da altri cluster, in modo da riflettere le vere strutture nascoste e le relazioni tra le unità statistiche.

Questa conoscenza è di grande valore per la formulazione di strategie di marketing mirate, segmentazione dei clienti, previsione dei comportamenti degli utenti e molte altre applicazioni. Inoltre, la *Cluster Analysis* offre un modo efficace per ridurre la complessità dei dati, semplificando la loro interpretazione e facilitando la presa di decisioni informate.

Tuttavia, è importante sottolineare che la matrice di similarità non è una misura assoluta della verità, ma piuttosto una rappresentazione approssimata delle

relazioni tra le unità statistiche. La sua creazione richiede una considerazione attenta delle variabili coinvolte e delle misure di similarità appropriate (Bandyopadhyay & Saha, 2013), al fine di ottenere risultati rilevanti e coerenti.

La matrice di similarità svolge dunque un ruolo centrale nell'ambito delle tecniche di *Cluster Analysis*. Essa fornisce una rappresentazione strutturata delle relazioni tra le unità statistiche, consentendo di identificare gruppi e strutture nascoste nel dataset. Sfruttando la matrice di similarità come base per le tecniche di *Cluster Analysis*, siamo in grado di estrarre informazioni preziose, facilitare la comprensione dei dati complessi e supportare la presa di decisioni informate in diversi ambiti di studio.

La rappresentazione tabulare della matrice di similarità facilita anche l'interpretazione visiva dei risultati, consentendo di individuare facilmente le relazioni di similarità o dissimilarità più evidenti tra le unità statistiche. Questa matrice può essere ulteriormente esaminata e visualizzata attraverso metodi di rappresentazione grafica, come il *Multi Dimensional Scaling* (MDS), che consente di ottenere rappresentazioni visive dei dati partendo dalla matrice di similarità. L'utilizzo di tecniche di visualizzazione, come rappresentazioni grafiche dei cluster o *heatmap*, amplifica la comprensione delle strutture presenti nel dataset, agevolando l'analisi e la scoperta di pattern nascosti.

2.2 Richiami ai concetti di distanza, similarità e dissimilarità

Supponiamo di dover affrontare una situazione in cui abbiamo *n* oggetti da raggruppare, che possono rappresentare persone, prodotti, punti vendita, paesi o qualsiasi altra entità rilevante per il contesto considerato. Inoltre, disponiamo di una struttura di dati in input che rappresenta gli oggetti attraverso un insieme di p misure o attributi. Queste misure possono essere organizzate in una matrice $n \times p$, in cui le righe (n) corrispondono agli oggetti e le colonne (p) rappresentano le variabili o le caratteristiche che descrivono gli oggetti (Kaufman & Rousseeuw, 1990). Tale matrice ha l'aspetto di:

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

(2. 1)

Ad esempio, se stiamo analizzando un insieme di prodotti, le righe potrebbero rappresentare i singoli prodotti e le colonne potrebbero rappresentare le caratteristiche dei prodotti, come il prezzo, la dimensione, il colore o altre misure

specifiche.

La distanza e i concetti di similarità e dissimilarità diventano fondamentali in questo contesto, poiché rappresentano le misure che quantificano la separazione o la relazione tra gli oggetti, consentendo di definire la struttura di connessione nel grafo di similarità.

Per esaminare i dati e identificare pattern o raggruppamenti rilevanti, diventa fondamentale introdurre una nozione di distanza tra gli oggetti. La distanza rappresenta una misura che quantifica la separazione o la relazione tra due oggetti e ci fornisce un'indicazione della loro vicinanza nel contesto del clustering.

Per ogni coppia di oggetti *i* e *j*, è necessario definire una misura di distanza che ci permetta di valutare la loro relazione. La scelta della misura di distanza appropriata dipende dal contesto di studio e dalla natura dei dati. Ogni misura di distanza può catturare aspetti specifici della relazione tra gli oggetti, consentendo di considerare diverse prospettive nella valutazione della vicinanza tra di essi. La selezione accurata di una misura di distanza è essenziale per garantire risultati accurati e rilevanti nel processo di clustering.

La scelta più diffusa è la distanza euclidea:

$$d(i,j) = \left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}},$$

(2. 2)

che corrisponde alla vera distanza geometrica tra i punti con coordinate $(x_{i1}, \dots, x_{ip})$ e $(x_{j1}, \dots, x_{jp})$.

Un'altra metrica ben nota è la distanza City Block o di Manhattan, definita da

$$d(i,j) = \sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|.$$

(2. 3)

Sia la metrica euclidea che la metrica di Manhattan soddisfano i seguenti requisiti matematici che caratterizzano una funzione di distanza:

(**D1**) $d(i,j) \geq 0$;

(**D2**) $d(i,j) = 0$;

(**D3**) $d(i,j) = d(j,i)$;

(**D4**) $d(i,j) \leq d(i,h) + d(h,j)$;

(2. 4)

per tutti gli oggetti *i, j* e *h.*

La condizione (Dl) afferma che le distanze tra due oggetti sono sempre non negative e la condizione (D2) dice che la distanza di un oggetto da se stesso è pari a zero. La condizione (D3) indica la simmetria della funzione distanza, ovvero che l'ordine con cui vengono considerati gli oggetti non influisce sulla misura di distanza, e, infine, la proprietà della disuguaglianza triangolare (D4) è necessaria per consentire un'interpretazione geometrica delle distanze e afferma sostanzialmente che la distanza tra due oggetti è sempre minore o uguale alla somma delle distanze tra uno dei due oggetti e un terzo oggetto intermediario.

Si noti che $d(i,j) = 0$ non implica necessariamente che *i* = *j*, perché potrebbe accadere che due oggetti diversi abbiano le stesse misure per le variabili considerate. Tuttavia, la disuguaglianza triangolare implica che gli oggetti i e j avranno la stessa distanza da ogni altro oggetto h, perché $d(i,h) \leq d(i,j) + d(j,h) = d(j,h)$ e allo stesso tempo $d(j,h) \leq d(j,i) + d(i,h) = d(i,h)$. Insieme queste relazioni implicano che $d(i,h) = d(j,h)$.

Abbiamo a disposizione un insieme di distanze e desideriamo organizzarle in modo sistematico. Un approccio comune a questo scopo è creare una matrice $n \times n$. Gli elementi di questa matrice possono rappresentare distanze euclidee o di Manhattan, ma va notato che esistono numerose altre possibilità. Pertanto, anziché parlare di distanze parleremo di dissimilarità, o coefficienti di dissimilarità.

Le dissimilarità sono, essenzialmente, numeri non negativi indicati con $d(i,j)$, che assumono valori piccoli (vicini a zero) quando i punti i e j sono "vicini" tra loro e crescono quando i punti i e j sono molto diversi. Solitamente, assumiamo che queste dissimilarità siano simmetriche, il che significa che $d(i,j)$ è uguale a $d(j,i)$, e che la dissimilarità di un oggetto rispetto a se stesso sia sempre zero. Tuttavia è importante notare che, in generale, la disuguaglianza triangolare, ovvero la somma delle dissimilarità tra tre punti, potrebbe non essere rispettata (D4).

Infatti, spesso si suppone che le dissimilarità soddisfino (D1), (D2) e (D3) (Bock, 1974), anche se nessuna di queste proprietà è veramente indispensabile, e ci sono metodi di clustering che possono operare senza richiedere nessuna di queste condizioni. Tuttavia, la principale differenza rispetto alle distanze è che la condizione (D4) può non sussistere.

Le dissimilarità possono essere ottenute in diversi modi. Spesso, possono essere calcolate da variabili che sono binarie, nominali, ordinali, continue o una

combinazione di queste (una descrizione di queste variabili e delle possibili formule sarà fornita in seguito in questo capitolo). Inoltre, le dissimilarità possono essere determinate da semplici valutazioni soggettive, dove uno o più osservatori forniscono giudizi su quanto due oggetti siano diversi tra loro dal loro punto di vista. Questo tipo di dati è comune nelle scienze sociali e nel marketing, in cui le opinioni e le valutazioni soggettive possono essere rilevanti per l'analisi dei cluster (Kaufman & Rousseeuw, 1990).

Quando si desidera effettuare una *Cluster Analysis* su un insieme di variabili osservate in una specifica popolazione, esistono diverse misure di dissimilarità disponibili che possono essere utilizzate per comparare le variabili tra loro. Queste misure permettono di calcolare le dissimilarità tra gli oggetti sulla base dei dati raccolti. La scelta della misura di dissimilarità dipenderà dal tipo di variabili utilizzate e dagli obiettivi specifici dell'analisi.

Ad esempio, una misura comune è la correlazione di Pearson, una misura parametrica che cerca una relazione lineare tra le variabili (Cohen, et al., 2009; Benesty, Chen, & Huang, 2008). La correlazione di Pearson tra le variabili f e g è calcolata come il rapporto tra la covarianza delle variabili e il prodotto delle loro deviazioni standard:

$$R(f,g) = \frac{\sum_{i=1}^{n}(x_{if}-m_f)(x_{ig}-m_g)}{\sqrt{\sum_{i=1}^{n}(x_{if}-m_f)^2}\sqrt{\sum_{i=1}^{n}(x_{ig}-m_g)^2}}.$$

(2. 5)

Un'altra importante misura di similarità utilizzata nella *Cluster Analysis* è la correlazione di Spearman (Zar, 2005; Sedgwick, 2014; Zar, 1972), una misura non parametrica particolarmente utile quando si desidera valutare le relazioni monotone tra le variabili, ovvero situazioni in cui il cambiamento in una variabile è associato a un cambiamento costante nell'altra, ma non necessariamente in modo lineare.[11]

A differenza della correlazione di Pearson, che si basa sui valori effettivi delle variabili, la correlazione di Spearman si concentra sui ranghi delle osservazioni. Questo rende la misura meno sensibile alle distorsioni causate da *outlier* e da distribuzioni non normali dei dati, il che la rende una scelta robusta per la *Cluster Analysis*.

La correlazione di Spearman viene calcolata seguendo questi passaggi:

[11] La correlazione di Spearman è particolarmente adatta quando i dati presentano un ordine naturale o quando si sospetta che le relazioni tra le variabili siano monotone piuttosto che lineari.

1. Assegna i ranghi alle osservazioni per ciascuna variabile separatamente. Ad esempio, per la variabile *f*, si assegnano i ranghi in base ai valori delle osservazioni da più piccolo a più grande.
2. Calcola la differenza dei ranghi per ciascuna coppia di osservazioni.
3. Calcola il coefficiente di correlazione di Pearson per i ranghi calcolati.

Nel contesto della *Cluster Analysis*, l'utilizzo della correlazione di Spearman come misura di similarità può portare a una migliore identificazione dei gruppi che si basano su relazioni monotone tra le variabili, consentendo di catturare strutture nascoste che potrebbero non emergere utilizzando altre misure di similarità.

Entrambi i coefficienti, sia la correlazione di Spearman che quella di Pearson, assumono valori compresi nell'intervallo tra -1 e $+1$ e non sono influenzati dalla scelta delle unità di misura. La principale differenza tra loro è che il coefficiente di Pearson cerca una relazione lineare tra le variabili *f* e *g*, mentre il coefficiente di Spearman cerca una relazione monotona. Questi coefficienti di correlazione sono di grande utilità nell'ambito della *Cluster Analysis* poiché quantificano il grado di correlazione tra due variabili.

I coefficienti di correlazione, che siano parametrici o non parametrici, possono essere convertiti in dissimilarità $d(f, g)$ in diversi modi. Ad esempio, una formula comune è:

$$d(f, g) = \frac{(1 - R(f, g))}{2}.$$

(2. 6)

Con questa formula, le variabili con una correlazione positiva elevata ricevono un coefficiente di dissimilarità vicino a zero, mentre le variabili con una correlazione fortemente negativa vengono considerate molto dissimili.

In alternativa, in alcune applicazioni, si potrebbe preferire utilizzare:

$$d(f, g) = 1 - |R(f, g)|.$$

(2. 7)

In questo caso, anche le variabili con una correlazione fortemente negativa riceveranno una bassa dissimilarità.

Inoltre, esistono molte altre dissimilarità ad hoc tra le variabili che possono essere considerate in base alle specifiche esigenze del problema.

In molte applicazioni, i dati di input consistono semplicemente in una matrice di

dissimilarità, senza alcuna misurazione quantitativa associata. In effetti, queste dissimilarità possono essere calcolate da attributi che non sono stati pubblicati o persino persi nel processo. Potrebbe anche verificarsi il caso in cui non ci siano mai state variabili quantitative inizialmente, poiché le dissimilarità sono state ottenute da valutazioni soggettive, dati di confusione o altre fonti. Di conseguenza, è estremamente utile disporre di algoritmi di clustering in grado di operare direttamente su una matrice di dissimilarità, senza la necessità di misure quantitative.

Se l'obiettivo è valutare non tanto quanto due oggetti siano differenti tra di loro, bensì quanto siano simili, allora è necessario introdurre il concetto di similarità

La similarità tra due oggetti *i* e *j* viene rappresentata da una funzione di similarità $s(i,j)$. Maggiore è la similarità (o vicinanza) tra i due oggetti *i* e *j*, maggiore sarà il valore di $s(i,j)$. Tipicamente, la misura di similarità $s(i,j)$ assume valori compresi tra 0 e 1, dove 0 indica che *i* e *j* non sono affatto simili e 1 rappresenta la massima similarità. Mentre i valori compresi tra 0 e 1 riflettono gradi variabili di similarità. È comune assumere le seguenti proprietà per la similarità (Bock, 1974):

(**S1**) $0 \leq s(i,j) \leq 1$

(**S2**) $s(i,j) = 0$

(**S3**) $s(i,j) = s(j,i)$

(2. 8)

per tutti gli oggetti *i* e *j*. Queste proprietà indicano che la similarità è sempre non negativa, la similarità tra un oggetto e se stesso è zero e la similarità tra i due oggetti è simmetrica.

I valori $s(i,j)$ possono essere organizzati in una matrice $n \times n$, che prende il nome di matrice di similarità. Sia le matrici di similarità che di dissimilarità sono comunemente indicate come matrici di prossimità, o talvolta matrici di somiglianza.

Le similarità possono essere ottenute in diversi modi a seconda della natura delle variabili e del contesto dell'analisi. Come le dissimilarità, possono essere il risultato di giudizi soggettivi. Inoltre, ci sono formule per calcolare le similarità tra oggetti caratterizzati da attributi, anche quando queste variabili sono di diversi tipi, come vedremo nella Sezione 2.3.5 sulle misurazioni miste.

Per definire le similarità tra variabili, possiamo nuovamente ricorrere al

coefficiente di correlazione di Pearson o di Spearman. Tuttavia, nessuna delle due misure di correlazione può essere utilizzata direttamente come coefficiente di similarità, poiché assumono anche valori negativi. Per risolvere questo problema, è necessaria una trasformazione per portare i coefficienti nell'intervallo compreso tra zero e uno. Esistono essenzialmente due approcci comuni a seconda del significato dei dati e dello scopo dell'applicazione.

Se si desidera considerare le variabili con una forte correlazione negativa come molto diverse tra loro perché sono orientate in direzioni opposte, è possibile utilizzare una formula come:

$$s(f,g) = \frac{1+R(f,g)}{2},$$

(2. 9)

che produce un coefficiente di similarità $s(f,g)$ pari a zero ogni volta che il coefficiente di correlazione $R(f,g)$ è uguale a -1 (Kaufman & Rousseeuw, 1990).

D'altra parte, se si ritiene che le variabili con una forte correlazione negativa debbano essere raggruppate perché misurano essenzialmente la stessa cosa, è possibile utilizzare una formula come:

$$s(f,g) = |R(f,g)|,$$

(2. 20)

che fornisce un coefficiente di similarità $s(f,g)$ pari a uno quando il coefficiente di correlazione $R(f,g)$ è uguale a -1.

Va notato che utilizzare i coefficienti di correlazione per valutare la similarità tra gli oggetti invertendo i ruoli degli oggetti e delle variabili nell'espressione di correlazione non è appropriato (Eades, 1965; Fleiss & Zubin, 1969). Ciò comporterebbe l'applicazione di operazioni come il calcolo della media delle misure dello stesso oggetto (in diverse unità di misura), il che non ha un significato coerente. L'uso del coefficiente di correlazione tra oggetti è stato dunque criticato per diversi motivi (Kaufman & Rousseeuw, 1990).

Supponiamo successivamente che i dati siano rappresentati da una matrice di similarità, ma che si desideri applicare un algoritmo di clustering progettato per le dissimilarità. In tal caso, è necessario convertire le similarità in dissimilarità. Maggiore è la similarità $s(i,j)$ tra i due oggetti *i* e *j*, più piccola dovrebbe essere la loro dissimilarità $d(i,j)$. Pertanto, è necessaria una trasformazione decrescente per ottenere le dissimilarità.

Una comune trasformazione consiste nell'utilizzare la formula:

$$d(i,j) = 1 - s(i,j)$$

(2. 13)

Questa formula sottrae la similarità dalla massima similarità possibile (che è 1) per ottenere la dissimilarità

Un'altra possibile trasformazione, suggerita da Gower (1966) sulla base di una discussione geometrica, consiste nel calcolare:

$$\sqrt{1 - s(i,j)}.$$

(2. 42)

Questo tipo di trasformazione rende le differenze tra le somiglianze maggiormente evidenti, ma allo stesso tempo può rendere più difficile ottenere piccole dissimilarità. Di conseguenza, la matrice di dissimilarità risultante potrebbe presentare una maggiore omogeneità e avere meno probabilità di generare raggruppamenti distinti.

Inoltre, esistono molte altre misure di similarità ad hoc che possono essere adattate a specifici contesti di studio. La scelta della misura di similarità dipenderà dalle caratteristiche dei dati, dai concetti teorici sottostanti e dagli obiettivi dell'analisi dei cluster. È importante selezionare una misura di similarità appropriata che rifletta le relazioni e le differenze significative tra gli oggetti considerati, al fine di ottenere risultati validi e coerenti nel processo di clustering.

È infine importante sottolineare che, in alcune applicazioni, la similarità tra gli oggetti può essere rilevata direttamente senza la necessità di misurare le loro caratteristiche individuali. Questo è particolarmente evidente in contesti come le indagini di mercato, in cui i consumatori vengono coinvolti nell'espressione della similarità tra diversi oggetti. In queste situazioni, spesso vengono utilizzate domande ad hoc che chiedono ai partecipanti di raggruppare gli oggetti in categorie ritenute simili. Successivamente, per ogni coppia di soggetti, la similarità può essere calcolata come la frequenza relativa dei partecipanti che hanno inserito i due oggetti nello stesso gruppo.

La scelta accurata di una misura di similarità è dunque un aspetto critico dell'analisi dei cluster, poiché può influenzare notevolmente i risultati ottenuti. Una misura di similarità adeguata contribuirà a rivelare le strutture nascoste nei dati e a identificare cluster rilevanti. Pertanto, è essenziale considerare attentamente le caratteristiche dei dati e gli obiettivi dell'analisi per selezionare la

misura di similarità più appropriata per il proprio studio.

2.3 Tipi di variabili e come trattarle

Dopo aver esaminato la costruzione della matrice di similarità e i concetti di distanza, similarità e dissimilarità, è ora fondamentale affrontare un aspetto cruciale nell'analisi dei cluster: i diversi tipi di variabili presenti nel set di dati e come trattarle correttamente, ponendo particolare attenzione sulle variabili ordinali. In particolare, le variabili possono assumere forme diverse, come variabili continue, binarie, nominali o ordinali, e ciascun tipo richiede un approccio specifico per valutare la similarità tra gli oggetti.

È importante considerare attentamente il trattamento delle variabili durante l'analisi dei cluster, poiché una valutazione accurata delle similarità tra gli oggetti dipende dal tipo di variabile e dalla natura delle relazioni che intendiamo esplorare. Pertanto, prenderemo in considerazione i diversi tipi di variabili e le relative strategie di trattamento per garantire risultati rilevanti e coerenti nella nostra analisi.

Attraverso l'applicazione di strategie adeguate a ciascun tipo di variabile, saremo in grado di ottenere una valutazione accurata delle similarità tra gli oggetti e di identificare cluster informativi. Questo ci aiuterà a comprendere meglio i pattern e le relazioni presenti nei dati e a trarre conclusioni significative dall'analisi dei cluster.

Ora procederemo ad analizzare più da vicino le diverse tipologie di dati che possono essere utilizzate, considerandole una per volta.

Le prossimità indirette utilizzate nella *Cluster Analysis* sono derivate dalla matrice dei dati, che rappresenta la nostra matrice $n \times p$. In questa sezione, illustreremo diverse misure di prossimità che possono essere applicate a ciascuna tipologia di variabile presente nel nostro set di dati:

- misure di prossimità per le variabili continue;
- misure di prossimità per le variabili binarie;
- misure di prossimità per le variabili nominali;
- misure di prossimità per le variabili ordinali;
- misure di prossimità per variabili miste.

2.3.1 Misure di similarità per variabili continue

In questa situazione, consideriamo un dataset con *n* oggetti caratterizzati da *p* misure continue. Questi valori sono numeri reali positivi o negativi che possono essere organizzate in una matrice $n \times p$, in cui le righe (*n*) corrispondono agli oggetti e le colonne (*p*) alle variabili. Tale matrice ha l'aspetto di

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

(2. 53)

È comune che le variabili utilizzate nei dati abbiano unità di misura diverse a causa della natura eterogenea delle informazioni raccolte. Tuttavia, la scelta dell'unità di misura può influenzare i risultati dell'analisi. Per evitare che le diverse scala di misura delle variabili influenzino le distanze tra i punti, si può optare per la standardizzazione o normalizzazione dei dati. La standardizzazione o normalizzazione dei dati implica la conversione delle misure originali delle variabili in nuove variabili senza unità. Ciò permette di trattare tutte le variabili su una scala comparabile, eliminando l'influenza delle diverse unità di misura.

Considerato che l'attributo *f* dell'oggetto *i*-esimo sia indicato con x_{if} (dove *i* = 1, ..., *n* e *f* = 1, ..., *p*), il primo passo nella standardizzazione consiste nel calcolare il valore medio della variabile *f*, indicato con m_f.

Successivamente, si calcola una misura della dispersione o "*spread*" per ogni variabile. Tradizionalmente, per questo viene utilizzata la deviazione media assoluta, indicata con s_f, che è calcolata come segue (Kaufman & Rousseeuw, 1990):

$$s_f = \frac{1}{n}\left(\sum_{i=1}^{n} \left| x_{if} - m_f \right|\right),$$

(2. 64)

dove, $f = 1, \ldots, p$.

La deviazione media assoluta è una misura di dispersione che tiene conto della variazione dei valori senza essere troppo sensibile a valori anomali.

Assumendo che s_f sia diverso da zero (altrimenti la variabile *f* è costante su tutti gli oggetti e deve essere rimossa), le misure standardizzate sono definite come:

$$z_{if} = \frac{x_{if} - m_f}{s_f}.$$

(2. 75)

Per definizione, le misure standardizzate z_{if} hanno valore medio pari a zero e deviazione media assoluta pari a 1. L'effetto della standardizzazione è quello di trasformare le variabili in nuove variabili senza unità, rendendo i dati comparabili tra loro e consentendo una valutazione più accurata delle relazioni e delle distanze tra gli oggetti nel processo di clustering.

Quando si applica la standardizzazione, si tralasciano i dati originali e si utilizza la nuova matrice di dati standardizzati:

$$\begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix},$$

(2. 86)

per tutti i calcoli successivi nell'analisi dei dati.

Il passo successivo consiste nel calcolare le distanze tra gli oggetti, al fine di quantificare il loro grado di dissimilarità. È necessario disporre di una distanza per ogni coppia di oggetti *i* e *j*. La scelta più diffusa è la distanza euclidea, definita come segue

$$d(i,j) = \left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}},$$

(2. 97)

che corrisponde alla vera distanza geometrica tra i punti con coordinate $(x_{i1}, \ldots, x_{ip})$ e $(x_{j1}, \ldots, x_{jp})$.

Un'altra metrica ben nota è la distanza City Block o di Manhattan, definita da

$$d(i,j) = \sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|.$$

(2. 108)

Una generalizzazione della metrica euclidea e di quella di Manhattan è data dalla distanza di Minkowski, che può essere espressa come

$$d(i,j) = \left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^r\right]^{\frac{1}{r}},$$

(2. 119)

dove *r* rappresenta un qualsiasi numero reale maggiore o uguale a 1. Questa distanza è comunemente nota come metrica L_r, poiché rappresenta un'intera

famiglia di metriche in base al valore di *r*. I casi speciali includono la metrica euclidea ($r = 2$) e la metrica di Manhattan ($r = 1$), che sono ampiamente utilizzate nelle analisi di clustering.

La distanza di Minkowski offre una maggiore flessibilità nel modellare le relazioni tra le variabili all'interno del contesto dello *Spectral Clustering*. A seconda del valore di *r* scelto, è possibile adattare la misura di distanza alle caratteristiche specifiche dei dati e agli obiettivi dell'analisi (Kaufman & Rousseeuw, 1990).

Nell'analisi dei dati, può essere comune incontrare situazioni in cui alcune misurazioni non sono effettivamente disponibili, creando dei "buchi" nella matrice dei dati. Questi buchi rappresentano valori mancanti, che possono essere causati da diversi fattori come la perdita dei dati, l'omissione accidentale o la mancanza di tempo per raccogliere le informazioni necessarie. A volte invece l'informazione non è semplicemente disponibile. A causa di ciò, è importante affrontare il problema dei valori mancanti per garantire l'accuratezza e l'affidabilità delle analisi.

La soluzione più immediata al problema dei valori mancanti consiste nell'analizzare solo gli individui che possiedono un set completo di valori disponibili (*listwise deletion*). Tuttavia, questo approccio può ridurre significativamente il numero di individui disponibili per l'analisi, portando a una perdita di informazioni preziose e potenziali distorsioni dei risultati.

Per gestire una serie di dati con valori mancanti, si può adottare un approccio che prevede l'indicazione di tali valori nella matrice dei dati le misure mancanti utilizzando un codice, come ad esempio il numero 999,99, se non è già presente, che può essere riconosciuto dal programma. In questo modo, si può mantenere l'integrità della struttura dei dati, pur segnalando esplicitamente i valori mancanti. Tuttavia, è importante notare che se esiste un oggetto per il quale mancano tutte le misurazioni, non si ha alcuna informazione su quell'oggetto e pertanto è necessario eliminarlo dall'analisi. Allo stesso modo, se una variabile presenta molti valori mancanti, essa deve essere eliminata poiché non fornisce alcuna informazione utile per l'analisi.

Questo approccio consente di gestire in modo appropriato i valori mancanti nei dati, mantenendo al contempo la coerenza e l'integrità delle analisi. È importante valutare attentamente l'impatto dei valori mancanti sulle conclusioni dell'analisi e adottare strategie adeguate per affrontare questa problematica, in modo da ottenere risultati affidabili e rilevanti.

Disponiamo dunque di un insieme di distanze basate su dati standardizzati che desideriamo organizzare in modo sistematico. Un modo comune per farlo è creare una matrice $n \times n$, chiamata matrice di dissimilarità. Questa matrice presenta alcune delle proprietà fondamentali delle funzioni di distanza. In particolare, le voci presenti sulla diagonale principale della matrice di distanza sono sempre nulle, poiché la distanza di un oggetto da sé stesso deve essere pari a zero (lo stesso vale per la distanza di Manhattan o qualsiasi altra distanza). Inoltre, a causa della simmetria della matrice, è sufficiente rappresentare solo la metà triangolare inferiore o superiore della matrice di distanza (Kaufman & Rousseeuw, 1990).

A partire da un insieme di osservazioni multivariate continue, è dunque possibile derivare una matrice di dissimilarità utilizzando una delle misure riportate nella Tabella 2.1.

Tabella 2.1 – Misure di dissimilarità per dati continui.
Fonte: Kaufman & Rousseeuw, 1990.

	Misura	Formula
D1	Distanza Euclidea	$d(i,j) = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}}$ *(2.20)*
D2	Distanza City Block (o di Manhattan)	$d(i,j) = \sum_{k=1}^{p} \left\lvert x_{ik} - x_{jk} \right\rvert$ *(2.21)*
D3	Distanza di Minkowski	$d(i,j) = \left[\sum_{k=1}^{p} \left\lvert x_{ik} - x_{jk} \right\rvert^r\right]^{\frac{1}{r}}$ *(2.22)*

È importante sottolineare che la distanza Euclidea può essere vista come un caso speciale di una misura Minkowski con $r = 2$, mentre la City Block corrisponde a una misura Minkowski con $r = 1$.

2.3.2 Misure di similarità per variabili binarie

Le variabili binarie rappresentano attributi che hanno solo due possibili esiti o

stati. Ad esempio, quando si raggruppano le persone, si possono utilizzare variabili binarie come maschio/femmina, fumatore/non fumatore, risposta sì/no a una specifica domanda, e così via. Nella matrice dei dati, queste variabili sono spesso codificate come zero o uno, dove uno indica la presenza dell'attributo e zero indica la sua assenza.

Quando si lavora con variabili binarie, è possibile trattarle come se fossero a scala di intervallo e applicare le formule tradizionali per calcolare la distanza euclidea o di Manhattan. Questo approccio può talvolta produrre risultati soddisfacenti. Tuttavia, è importante essere consapevoli che esistono approcci specificamente progettati per gestire i dati binari in modo più appropriato.

In presenza di dati binari, è consigliabile utilizzare misure di dissimilarità specifiche per catturare la natura delle variabili binarie. Utilizzare misure di dissimilarità dedicate ai dati binari può portare a una rappresentazione più accurata e adatta per l'analisi e il clustering di tali dati.

Le misure di dissimilarità per dati binari sono definite in base al calcolo delle concordanze e discordanze tra le *p* variabili per due individui. Nella Tabella 2.2, possiamo osservare le seguenti relazioni:

- Gli individui *i* e *j* assumono lo stesso valore 1 su *a* variabili, significa che concordano sulla presenza dell'attributo per quelle variabili.
- Gli individui *i* e *j* assumono lo stesso valore 0 su *d* variabili, significa che concordano sull'assenza dell'attributo per quelle variabili.
- Su *b* variabili l'individuo *i* assume il valore 0, mentre l'individuo *j* assume il valore 1. Ciò indica che vi è una discordanza tra i due individui su quelle variabili, in quanto uno ha l'attributo presente e l'altro no.
- Su *c* variabili l'individuo *i* assume il valore 1, mentre l'individuo *j* assume il valore 0. Anche in questo caso, vi è una discordanza tra i due individui su quelle variabili, ma con un diverso orientamento rispetto al caso precedente.

Tabella 2.2 – Computo dell'esito binario per due individui.
Fonte: Kaufman & Rousseeuw, 1990.

	Individuo *i*			
	Esito	1	0	Totale
Individuo *j*	1	a	b	$a+b$
	0	c	d	$c+d$
	Totale	$a+c$	$b+d$	$p=a+b+c+d$

Tabella 2.3 – Misure di similarità per dati binari.
Fonte: Kaufman & Rousseeuw, 1990.

	Misura	Formula
S1	Coefficiente di corrispondenza	$s_{ij}=\frac{a+d}{a+b+c+d}$ *(2.23)*
S2	Coefficiente di Rogers e Tanimoto	$s_{ij}=\frac{a}{a+b+c}$ *(2.24)*
S3	Coefficiente di Jaccard	$s_{ij}=\frac{a}{a+2(b+c)+d}$ *(2.25)*
S4	Coefficiente di Sokas e Sneath	$s_{ij}=\frac{a}{a+2(b+c)}$ *(2.26)*
S5	Coefficiente di Gower e Legendre	$s_{ij}=\frac{a+d}{a+\frac{1}{2}(b+c)+d}$ *(2.27)*
S6	Coefficiente di Gower e Legendre	$s_{ij}=\frac{a}{a+\frac{1}{2}(b+c)}$ *(2.28)*

La Tabella 2.3 elenca alcune delle misure di similarità che sono state proposte per dati binari. La ragione di questa varietà di misure risiede nell'incertezza su come debbano essere trattate le corrispondenze $0-0$ o assenza-assenza. In alcuni casi, le corrispondenze $0-0$ sono considerate del tutto equivalenti alle corrispondenze

$1-1$ e vengono incluse nel calcolo della misura di similarità. In altri casi, tuttavia, l'inclusione o meno delle d corrispondenze $0-0$ è problematica, specialmente quando il valore zero corrisponde all'effettiva assenza di una proprietà. La domanda da porsi è se la co-assenza contenga informazioni utili sulla similarità tra due oggetti: attribuire un alto grado di similarità ad una coppia di individui semplicemente perché entrambi mancano di alcuni attributi potrebbe non essere significativo in molte situazioni.

In tali casi, è conveniente utilizzare misure che ignorano il conteggio della co-assenza, come (S2) o (S4). Quando invece alla co-assenza di un fattore può essere associato un contenuto informativo, di solito viene utilizzato il coefficiente di corrispondenza (S1). Le misure (S3) ed (S5) sono esempi di coefficienti simmetrici che prendono in considerazione anche le corrispondenze negative, pur assegnando pesi diversi nei due casi.

Sokal e Sneath (S4) sottolineano che non esistono regole veloci e rigide per stabilire se le corrispondenze negative debbano essere incluse o meno: la decisione spetta all'analista dei dati, che deve effettuare una scelta affidandosi al proprio livello di esperienza e familiarità con il materiale trattato.

La scelta della misura di similarità è molto importante, dal momento che l'utilizzo di diversi coefficienti di similarità può condurre a risultati diversi. Pertanto, è fondamentale selezionare attentamente la misura di similarità più appropriata in base alle caratteristiche dei dati e agli obiettivi specifici dell'analisi.

2.3.3 Misure di similarità per variabili nominali

Le variabili nominali possono assumere diversi stati o categorie, senza un ordinamento specifico tra di loro. Ad esempio, la nazionalità delle persone o il loro stato civile (celibe/nubile/divorziato/vedovo) sono esempi di variabili nominali. Nella matrice dei dati, tali variabili vengono codificate assegnando loro codici numerici, come $1, 2, \ldots, M$ (a volte si usano anche i codici $0, 1, \ldots, M-1$). Tuttavia, è importante sottolineare che questi codici sono utilizzati solo per scopi pratici di gestione dei dati e non implicano un ordinamento intrinseco dei valori. In altre parole, i numeri di codice potrebbero essere sostituiti da lettere o altri simboli senza alterare la natura della variabile.

Il modo più comune di misurare la similarità o la dissimiglianza tra alcuni oggetti i e j caratterizzati da variabili nominali è quello di utilizzare l'approccio della corrispondenza semplice:

$$s(i,j) = \frac{u}{p} \quad \text{e} \quad d(i,j) = \frac{p-u}{p}.$$

(2.29)

Qui, u è il numero di corrispondenze, cioè il numero di variabili per le quali gli oggetti i e j condividono lo stesso stato. Come visto precedentemente, p è il numero totale di variabili.

In altre parole, $s(i,j)$ rappresenta la proporzione di variabili che hanno lo stesso stato per i due oggetti, mentre $d(i,j)$ rappresenta la proporzione di variabili che differiscono tra i due oggetti.

In entrambi i casi, la misura di similarità $s(i,j)$ e la misura di dissimilarità $d(i,j)$ sono normalizzate rispetto al numero totale di variabili (p). Questo assicura che le misure di similarità e dissimilarità siano comprese tra 0 e 1, dove 0 indica la massima dissimilarità e 1 indica la massima similarità.

L'approccio della corrispondenza semplice è ampiamente utilizzato per valutare la similarità o dissimilarità tra oggetti che presentano variabili nominali, consentendo di quantificare la relazione tra di loro in modo interpretabile e coerente.

2.3.4 Misure di similarità per variabili ordinali

Le variabili ordinali discrete sono una forma di variabili che presentano un'importante sequenza o ordinamento tra i loro stati. A differenza delle variabili nominali, dove i codici sono arbitrari, le variabili ordinali hanno un significato intrinseco nel loro ordine. Ad esempio, i codici da 1 a M sono assegnati in modo da riflettere un ordine significativo, dove uno stato con un codice basso è considerato inferiore a uno stato con un codice alto.[12]

Le variabili ordinali sono particolarmente utili per registrare valutazioni soggettive di qualità che non possono essere misurate oggettivamente. Ad esempio, le valutazioni di soddisfazione di un prodotto o il grado di preferenza di una marca possono essere rappresentate come variabili ordinali. Queste valutazioni sono soggettive e non possono essere misurate con precisione, ma l'ordine delle valutazioni può fornire informazioni significative sulle preferenze degli individui.

In alcuni casi, è possibile ottenere variabili ordinali discretizzando variabili continue di tipo scala di intervallo. Questo processo prevede la suddivisione dell'asse continuo delle misurazioni in un numero finito di classi.

[12] M indica il numero di modalità della variabile ordinale.

Le variabili ordinali continue sono molto simili. Si verificano quando le misurazioni sono continue, ma non si ha la certezza che seguano una scala lineare. In questa situazione, l'unica informazione attendibile è l'ordinamento delle osservazioni, mentre la distanza tra i valori potrebbe non essere rappresentata con precisione (Kaufman & Rousseeuw, 1990). Ad esempio, se le misurazioni sono soggette a una trasformazione non lineare come l'esponenziale o il logaritmo, la scala di intervallo viene persa.

Per affrontare questa sfida, una soluzione comune è sostituire le misurazioni con i loro ranghi, ogni misurazione riceve un grado di posizione da 1 a *M*, dove *M* rappresenta il numero di valori diversi assunti dalla variabile continua. (Naturalmente, due misurazioni uguali ricevono lo stesso rango). Questo approccio è particolarmente utile anche quando i dati originali presentano approssimativamente una scala per intervallo, ma possono contenere errori grossolani.

La trasformazione dei dati in ranghi offre diversi vantaggi. Innanzitutto, permette di ottenere un'informazione più robusta, riducendo l'effetto degli errori o delle misurazioni anomale sulla valutazione complessiva. Inoltre, consente di mantenere l'ordinamento delle osservazioni, che può essere un elemento chiave per la *Cluster Analysis* e altre tecniche di analisi dei dati. Trasformare i dati in ranghi fornisce una rappresentazione più stabile e affidabile delle relazioni tra le osservazioni, superando le limitazioni delle scale di intervallo non lineari (Kaufman & Rousseeuw, 1990).

Indipendentemente dall'origine delle variabili ordinali in studio, è di fondamentale importanza riconoscere che i loro stati sono ordinati in una sequenza significativa da 1 a *M*. Trattare questa variabile come nominale sarebbe uno spreco di informazioni, poiché la distanza e quindi la dissimilarità tra due stati dovrebbe aumentare man mano che i loro codici diventano più distanti. Pertanto, è consigliabile trattare i ranghi come se fossero su una scala di intervallo e applicare formule di dissimilarità appropriate, come la distanza euclidea o di Manhattan, al fine di ottenere misure di dissimilarità valide.

Tuttavia, considerando che le variabili ordinali oggetto di studio potrebbero avere diversi valori di *M*, è opportuno standardizzare tutte le variabili nell'intervallo 0 – 1 per garantire una ponderazione uniforme. Ciò può essere ottenuto sostituendo il rango r_{if}, dell'oggetto *i*-esimo nella variabile *f*, con una misura normalizzata z_{if}, calcolata come:

$$z_{if} = \frac{r_{if}-1}{M_f-1},$$

(2.30)

dove M_f rappresenta il grado più alto per la variabile *f*. In questo modo tutti gli z_{if} saranno compresi tra 0 e 1, consentendo una comparazione equilibrata tra le variabili.

Il programma DAISY costituisce una soluzione pratica per l'analisi di dati contenenti variabili ordinali, sia discrete che continue. È vantaggioso disporre di un programma di gestione dei dati per prepararsi adeguatamente alla *Cluster Analysis* vera e propria. Questa applicazione è stata denominata DAISY proprio perché la sua funzione principale consiste nel calcolare dei coefficienti di dissimilarità.

L'implementazione di questo programma è stata sviluppata in Fortran e risulta compatibile con il computer IBM PC, compresi i modelli XT e AT, oltre che con microcomputer compatibili. La caratteristica fondamentale di DAISY è la sua completa interattività.[13]

L'approccio adottato da DAISY si basa sulla trasformazione dei valori delle variabili in ranghi. Questo processo garantisce che misurazioni identiche siano associate a ranghi uguali e che ciascun rango sia rappresentato almeno una volta nel dataset.

Per utilizzare il programma DAISY, la procedura iniziale prevede la conversione di ciascuna variabile *f* nei ranghi 1, 2, ..., *M*, dove *M* rappresenta il massimo grado di variabilità nella variabile f in questione. Questa metodologia assicura che misure uguali ricevano lo stesso rango e che tutti i possibili ranghi siano inclusi nel dataset. Successivamente, i ranghi così ottenuti vengono trasformati in misure normalizzate z_{if}.

Inoltre, va evidenziato che il programma DAISY può essere agevolmente utilizzato all'interno dell'ambiente R, che costituisce il contesto in cui condurremo l'applicazione descritta nel presente studio.

La dissimilarità finale tra due oggetti *i* e *j* è calcolata utilizzando la distanza di Manhattan tra le loro misure normalizzate. La distanza di Manhattan rappresenta

[13] DAISY inoltre offre la flessibilità di selezionare le variabili da considerare, evitando l'obbligo di utilizzarle tutte. Questa caratteristica risulta spesso vantaggiosa, poiché in molte situazioni possono esserci variabili non pertinenti o variabili che potrebbero essere correlate tra di loro, limitando l'apporto di informazioni aggiuntive. Talvolta, l'utente possiede una conoscenza approfondita del dominio e può decidere in modo consapevole quale variabile impiegare. Allo stesso modo, potrebbe essere interessato solamente a un aspetto specifico (ad esempio, considerare solo le variabili economiche in un insieme di dati che comprende anche variabili demografiche). Pertanto, la scelta delle variabili rappresenta un'opzione disponibile all'interno di questo programma.

la somma delle differenze in valore assoluto tra le componenti delle due misure. Questa misura di dissimilarità tiene conto delle differenze ordinali tra gli oggetti, considerando la distanza tra i loro ranghi normalizzati.

È importante notare che, per ottenere una misura di dissimilarità standardizzata, la distanza di Manhattan viene divisa per il numero totale di variabili considerate sia per l'oggetto *i* che per l'oggetto *j*. Questa normalizzazione assicura che la dissimilarità finale sia coerente e confrontabile tra diversi oggetti e set di dati.

L'applicazione del programma DAISY consente quindi di trattare in modo appropriato le variabili ordinali, consentendo l'ordinamento delle osservazioni e ottenendo una misura di dissimilarità significativa per l'analisi dei cluster. Tale approccio favorisce la comprensione delle relazioni tra gli oggetti e la creazione di raggruppamenti coerenti e interpretabili.

2.3.5 Misure di similarità per variabili miste

Abbiamo discusso i metodi per trattare insiemi di dati di diversi tipi. Tuttavia, nelle applicazioni pratiche, è comune incontrare insieme di dati che contengono diversi tipi di variabili. In queste situazioni, è necessario adottare un approccio adeguato per trattare le diverse tipologie di variabili e ottenere risultati coerenti e rilevanti.

I dati con variabili miste possono essere trattati in diversi modi. Un primo approccio consiste nel non mescolare i diversi tipi di variabili e eseguire *Cluster Analysis* separate per ciascun tipo di variabile. Se le conclusioni di queste analisi sono più o meno concordanti, allora si può considerare il problema risolto. Tuttavia, quando si ottengono risultati discordanti, diventa difficile conciliare le diverse analisi e ottenere una visione unificata del clustering dei dati.

Pertanto, l'approccio più conveniente è quello di combinare le diverse tipologie di variabili in una singola matrice di prossimità. Questa matrice di prossimità rappresenta la similarità o dissimilarità tra gli oggetti del dataset, considerando contemporaneamente tutte le variabili, sia quantitative che qualitative. Questo approccio consente di considerare in modo integrato tutte le informazioni disponibili e di ottenere una visione più completa del clustering dei dati misti.

Tra le diverse misure di similarità proposte per dati di tipo misto, una delle più utilizzate è stata introdotta da Gower nel 1971. Questa misura di similarità di Gower tiene conto delle differenze di scala tra le variabili continue e considera correttamente le variabili qualitative. Essa calcola la similarità tra due oggetti basandosi sulle differenze assolute o relative tra le loro variabili, considerando anche le variabili qualitative come fattori di dissimilarità.

Introdurremo ora una leggera generalizzazione del metodo di Gower che tiene conto anche le variabili ordinali. Infatti, inizialmente la definizione originale di Gower si basava su un coefficiente di similarità compreso tra 0 e 1, ma noi trasformeremo questa similarità in una dissimilarità utilizzando dell'espressione:

$$d(i,j) = 1 - s(i,j).$$

(2.31)

Tuttavia, è importante notare che è sempre possibile tornare alle somiglianze calcolando alla fine:

$$s(i,j) = 1 - d(i,j).$$

(2.32)

Supponiamo quindi di avere un insieme di dati che contenga *p* variabili di natura mista, inclusi i dati ordinali (Gower, 1971a). La dissimilarità $d(i,j)$ tra gli oggetti *i* e *j* è definita come:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}},$$

(2.33)

dove l'indicatore $\delta_{ij}^{(f)}$ è posto uguale a 1 quando entrambe le misure x_{if} e x_{jf} per la variabile f non sono mancanti, mentre è posto uguale a 0 in caso contrario. Tuttavia, questa espressione non può essere calcolata quando tutti gli $\delta_{ij}^{(f)}$ sono zero. In questo caso, è necessario assegnare a $d(i,j)$ un valore convenzionale o rimuovere uno degli oggetti *i* o *j* dalla considerazione.

Il contributo di ciascuna variabile *f* alla dissimilarità tra *i* e *j* è rappresentato dal numero $d_{ij}^{(f)}$. Se la variabile *f* è di tipo nominale, allora $d_{ij}^{(f)}$ è definita come

$$d_{ij}^{(f)} = 1 \quad \text{se} \quad x_{if} \neq x_{jf},$$

(2.34)

$$d_{ij}^{(f)} = 0 \quad \text{se} \quad x_{if} = x_{jf}.$$

(2.35)

In altre parole, $d_{ij}^{(f)}$ è 1 se i valori delle variabili *f* per gli oggetti *i* e *j* sono diversi, altrimenti è 0. Questa definizione di dissimilarità per variabili nominali riflette la differenza tra le categorie delle variabili.

Se tutte le variabili nel dataset sono nominali, l'espressione diventa il numero di corrispondenze sul numero totale di coppie disponibili, che coincide con il coefficiente di corrispondenza semplice

$$s(i,j) = \frac{u}{p} \quad \text{e} \quad d(i,j) = \frac{p-u}{p},$$

(2.36)

dove *u* rappresenta il numero di corrispondenze tra gli oggetti *i* e *j* e *p* rappresenta il numero totale di variabili.

Se la variabile f è continua, allora $d_{ij}^{(f)}$ è calcolato come:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f},$$

(2.37)

dove R_f rappresenta l'intervallo della variabile *f*, definita come:

$$R_f = \max_h x_{hf} - \min_h x_{hf},$$

(2.38)

dove *h* si estende a tutti gli oggetti non mancanti per la variabile *f*. In questo caso, $d_{ij}^{(f)}$ è sempre un numero compreso tra 0 e 1, riflettendo la distanza relativa tra i valori della variabile f per gli oggetti *i* e *j*. Le variabili ordinali vengono prima convertite in ranghi e successivamente si applica l'espressione (2.37).

Quando tutte le variabili nel dataset sono per scala di intervallo, la formula di Gower per la dissimilarità diventa la distanza di Manhattan, supponendo che le variabili siano state prima divise per il loro intervallo.

Concludiamo che il metodo combinato di Gower generalizza le dissimilarità discusse in precedenza per dati omogenei. I calcoli possono essere eseguiti utilizzando il programma DAISY.

Si noti che abbiamo eseguito Gower limitando $d_{ij}^{(f)}$ all'intervallo $0-1$, in modo che ogni variabile contribuisca con un valore compreso tra 0 e 1 alla dissimilarità media. Di conseguenza, la dissimilarità risultante $d(i,j)$ è anch'essa compresa tra 0 e 1. Possiamo ritrasformata la dissimilarità in una similarità utilizzando la formula:

$$s(i,j) = 1 - d(i,j).$$

(2.39)

Come ultima osservazione, è importante notare che è possibile raggruppare oggetti caratterizzati da una combinazione di misure e prossimità. Ad esempio, se abbiamo una matrice di similarità, una matrice di dissimilarità e una collezione mista di attributi per gli stessi *n* oggetti, possiamo utilizzare DAISY per convertire la matrice di similarità in una matrice di dissimilarità e calcolare un'altra matrice di dissimilarità dagli attributi. Le tre matrici di dissimilarità risultanti possono quindi essere combinate in un'unica matrice utilizzando la formula:

$$d(i,j) = \frac{w_1 d_1(i,j) + w_2 d_2(i,j) + w_3 d_3(i,j)}{w_1 + w_2 + w_3},$$

(2.40)

dove w_1, w_2 e w_3 sono alcuni pesi positivi che possono essere scelti in modo soggettivo.

2.3.6 Misure di similarità principalmente utilizzate per lo Spectral Clustering di variabili ordinali

Prima di pensare alla costruzione di un grafo di similarità, dobbiamo definire una funzione di similarità sui dati. Poiché in seguito costruiremo un grafo dei dintorni, dobbiamo assicurarci che i dintorni locali indotti da questa funzione di similarità siano rilevanti. Ciò significa che dobbiamo essere sicuri che i punti considerati "molto simili" dalla funzione di similarità siano anche strettamente correlati nell'applicazione da cui provengono i dati.

Nel contesto dello *Spectral Clustering* per variabili ordinali, sono state sviluppate diverse misure di similarità specificamente adattate a questo tipo di dati. Queste misure mirano a catturare la relazione di ordinamento tra le variabili ordinali e a fornire una rappresentazione appropriata per l'analisi dei cluster.

La distanza euclidea, sebbene ampiamente utilizzata come misura di similarità in diversi contesti, potrebbe non essere la scelta più appropriata per le variabili ordinali nell'ambito dello *Spectral Clustering*. Ciò è dovuto alla sua sensibilità alle differenze di magnitudo tra i valori e alla sua limitata capacità di catturare le relazioni ordinate tra le categorie.

Le variabili ordinali sono caratterizzate dall'ordinamento delle categorie, ma non forniscono informazioni sulla distanza o sulla magnitudo tra i valori. La distanza euclidea, che considera le differenze quadrate tra i valori, può essere influenzata da queste differenze di magnitudo, rendendo difficile la comparazione diretta tra le variabili ordinali.

Inoltre, la distanza euclidea non tiene conto dell'ordine delle categorie e quindi può ignorare le relazioni ordinate tra di esse. Questo può portare a risultati inaccurati o non rilevanti nello *Spectral Clustering* di variabili ordinali, poiché la struttura di ordinamento delle categorie potrebbe non essere adeguatamente rappresentata. Per superare queste limitazioni, è necessario utilizzare misure di similarità specificamente progettate per le variabili ordinali.

Una funzione di similarità ampiamente utilizzata in letteratura per affrontare il problema delle variabili ordinali è il Kernel gaussiano, spesso denominato semplicemente "funzione gaussiana". Quando si lavora con variabili ordinali, è fondamentale considerare l'ordinamento dei valori e tenerne conto nell'analisi. La funzione gaussiana è una funzione di similarità che fornisce una base solida per l'implementazione di algoritmi di clustering come lo *Spectral Clustering*. Essendo in grado di catturare le relazioni tra le variabili ordinali, la funzione gaussiana consente di ottenere raggruppamenti rilevanti e coerenti, che possono aiutare a identificare pattern e strutture nascoste nei dati.

La formula generale per calcolare la funzione di similarità gaussiana tra due oggetti *i* e *j*, utilizzando variabili ordinali, è la seguente:

$$s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

$$\forall\ (i, j),$$

(2.41)

dove, il parametro di scala σ controlla il tasso di decadimento delle distanze. La scelta di tale parametro può essere cruciale. Un possibile modo per selezionare σ suggerisce di eseguire ripetutamente l'algoritmo di *Spectral Clustering* per diversi valori di σ e selezionare quello che fornisce il miglior clustering secondo una misura di qualità scelta. Oltre all'ovvio aumento del costo computazionale, lo svantaggio di questa procedura è la difficoltà di scegliere una misura di qualità affidabile. Per implementare questa tecnica, si potrebbe utilizzare una sequenza decrescente, iniziando con un valore di σ minore di $\max_{ij} d_{ij}$ (Favati & al., 2020). Tuttavia, l'inconveniente di questa procedura è la scelta della misura di qualità da utilizzare, che può essere una funzione non monotona di σ. In questo caso, non si potrebbe ottenere una determinazione affidabile di un σ accettabile in un numero esiguo di tentativi (Favati & al., 2020).

Quando si utilizza la funzione di similarità gaussiana, la scelta corretta di σ può essere fondamentale per l'efficienza dell'algoritmo di *Spectral Clustering*. Un

valore troppo piccolo di σ darebbe distanze molto vicine a 0, per cui tutti i punti apparirebbero ugualmente lontani. Al contrario, un valore troppo grande di σ darebbe distanze molto vicine a 1, per cui tutti i punti apparirebbero ugualmente vicini.

L'utilizzo della funzione gaussiana per calcolare la similarità tra gli oggetti basati su variabili ordinali, sia discrete che continue, offre dunque un metodo robusto per gestire le differenze di ordinamento tra gli oggetti e catturare le relazioni tra le variabili (Kaufman & Rousseeuw, 1990; Chen, Li, Liu, Xu, & Ying, 2017).

Un'ulteriore misura di similarità comunemente utilizzata per variabili ordinali e già introdotta precedentemente, è la correlazione di Spearman. Questa misura, chiamata così in onore di Charles Spearman, è un'analisi statistica non parametrica[14] che valuta il grado di associazione monotona tra due variabili ordinali.[15]

A differenza della correlazione di Pearson, che è adatta per dati quantitativi continui, la correlazione di Spearman si concentra su dati ordinali, dove l'importante è l'ordinamento dei valori piuttosto che i valori effettivi. Ad esempio, considera un sondaggio in cui i partecipanti sono classificati come "molto soddisfatti", "soddisfatti", "neutri", "insoddisfatti" e "molto insoddisfatti". In questo contesto, la correlazione di Spearman è utile per misurare quanto le classificazioni seguano un ordine coerente tra due variabili.

Per calcolare la correlazione di Spearman, anziché utilizzare i valori effettivi delle variabili, si utilizzano i ranghi di ciascuna variabile, rendendola adatta per dati ordinali in cui l'informazione principale è l'ordinamento. La correlazione di Spearman può essere dunque interpretata come una misura di similarità tra gli oggetti, in quanto misura quanto coerentemente due oggetti seguono lo stesso ordinamento nelle variabili.

La formula per calcolare la correlazione di Spearman tra due variabili f e g è la seguente:

[14] I test non parametrici si presentano come un'alternativa meno vincolante rispetto ai test parametrici, come ad esempio l'indice di correlazione di Pearson. Essi non impongono la necessità che le variabili seguano una distribuzione normale né che la loro relazione sia lineare. Questa flessibilità li rende adatti a una vasta gamma di situazioni. In aggiunta, i test non parametrici dimostrano maggiore robustezza di fronte a valori anomali.

[15] Mentre le relazioni lineari suggeriscono che due variabili si spostino insieme con un ritmo uniforme, le relazioni monotone esaminano la probabilità che le due variabili si muovano nella stessa direzione, anche se senza richiedere un movimento a velocità costante.

$$\rho(f,g) = \frac{\sum_{i=1}^{n}(r_{if}\bar{r}_f)(r_{ig}\bar{r}_g)}{\sqrt{\sum_{i=1}^{n}(r_{if}\bar{r}_f)^2}\sqrt{\sum_{i=1}^{n}(r_{ig}\bar{r}_g)^2}},$$

(2.42)

dove, r_{if} e r_{ig} sono rispettivamente i ranghi della variabile f e g per l'i-esima osservazione, $\bar{r}_f$ e $\bar{r}_g$ sono i ranghi medi delle variabili f e g, e n rappresenta il numero totale di osservazioni.

Proponiamo ora una matrice di similarità per l'algoritmo di *Spectral Clustering* dai dati, basata sempre su una statistica non parametrica nota come τ (tau) di Kendall (Kendall & Gibbons, 1990). Questa misura è particolarmente adatta a tutte le variabili ordinate che possono assumere diversi numeri di valori.

Iniziamo definendo il tau di Kendall come misura di similarità tra due elementi (Howell, 2012). Mentre la correlazione di Spearman si basa sulle deviazioni dei dati rispetto ai loro valori medi, il tau di Kendall adotta un approccio totalmente differente basandosi sull'accordo o disaccordo tra coppie di osservazioni.

È importante notare che, in generale, il valore dell'indice di Kendall è più piccolo rispetto a quello dell'indice di Spearman calcolato sugli stessi dati. Questo non implica però che il tau di Kendall sia meno accurato, ma riflette semplicemente il fatto che valuta la relazione tra i dati da una prospettiva diversa.

Particolarmente, il tau di Kendall è una scelta consigliata quando il campione è di piccole dimensioni, poiché la stima dell'indice di correlazione ottenuta con questa misura è generalmente più precisa rispetto a quella ottenuta con l'indice di Spearman.

Ora, supponiamo di avere le risposte ordinali di n individui su due item, identificati come j e j'. Le risposte sono rappresentate da $(R_{1j} - R_{1j'}), \dots, (R_{nj} - R_{nj'})$, dove R_{ij} è la risposta dell'i-esimo individuo all'item j e $R_{ij'}$ è la risposta dell'i-esimo individuo all'item j'. In tal caso, l'indice di Kendall per la coppia di item (j, j') è definito come:

$$\tau_{jj'} = \frac{2}{n(n-1)}\sum_{1\le i<i'\le n} sgn\{(R_{ij} - R_{i'j})(R_{ij'} - R_{i'j'})\},$$

(2.43)

dove la funzione $sgn(x)$ assume valore 1 se $x > 0$, 0 se $x = 0$ e -1 altrimenti.

L'indice di Kendall, indicato come $\tau_{jj'}$, è chiaramente compreso tra -1 e $+1$. Questa misura, ampiamente adottata nel campo delle statistiche non parametriche

(Sprent & Smeeton, 2016), è utilizzata per quantificare la forza della dipendenza tra due elementi. Un valore notevolmente diverso da 0 indica un legame significativo tra le risposte, e di conseguenza, una forte similarità tra i due elementi. Questo legame sarà tanto più forte quanto più le coppie di valori si troveranno in accordo. In particolare, la correlazione sarà perfetta quando le unità statistiche saranno completamente concordi tra di loro.

Da notare che se due item misurano lo stesso costrutto ma sono formulati in modo opposto, ciò potrebbe portare a un valore negativo di $\tau_{jj'}$. Pertanto, gli item j e j' vengono considerati altamente simili se $\tau_{jj'}$ assume un valore elevato.

È importante evidenziare che $\tau_{jj'}$ può essere stimato accuratamente anche quando i dati presentano risposte mancanti o incomplete. Questa stima rimane valida a condizione che ci siano abbastanza individui che rispondono ad entrambi gli elementi.

Per effettuare il calcolo dell'indice di Kendall, è necessario prima verificare tre criteri fondamentali:

1. Tipo di variabili e scala di misurazione: le due variabili coinvolte devono essere di natura quantitativa o qualitative ordinale. È possibile utilizzare questo indice se entrambe le variabili sono quantitative, se una è quantitativa e l'altra è qualitativa, oppure se entrambe sono di tipo qualitativo ordinale. Questa misura è particolarmente utile per analizzare le relazioni tra variabili che utilizzano scale Likert.[16]
2. Appaiamento dei dati: le due variabili devono essere misurate sugli stessi individui o casi. In altre parole, per ogni unità statistica, deve essere disponibile un valore sia per la prima variabile che per la seconda variabile. Se i dati non sono appaiati in questo modo, sarà necessario adottare un'analisi basata su campioni indipendenti per esaminare le relazioni tra le variabili.
3. Relazione monotona: è essenziale che esista una relazione monotona tra le due variabili. Una relazione è considerata monotona quando all'aumentare dei valori di una variabile, i valori dell'altra variabile aumentano (anche se non necessariamente in modo lineare). Viceversa, quando i valori di una variabile aumentano, i valori dell'altra variabile diminuiscono (ancora una

[16] La scala di Likert è una metodologia di misurazione utilizzata per valutare l'atteggiamento, l'opinione o il grado di accordo o disaccordo di una persona rispetto a un insieme di affermazioni o dichiarazioni. Questa scala prevede una serie di affermazioni o dichiarazioni alle quali i partecipanti sono invitati a rispondere indicando il loro livello di accordo o disaccordo. Le risposte vengono valutate in base a un'escalation di valori, spesso numerica, che può variare da 3 a 7 o più opzioni.

volta, non necessariamente in modo lineare). In sostanza, questa condizione richiede che ci sia una tendenza generale di movimento nella stessa direzione, sebbene non sia obbligatorio che sia un aumento costante.

Oltre alle misure di similarità discusse, esistono altre misure comunemente impiegate per trattare variabili ordinali. Tra queste rientrano la distanza di Minkowski e la distanza di Jaccard (Irani, Pise, & Phatak, 2016), solo per menzionarne alcune. L'utilizzo di queste diverse misure permette di considerare aspetti specifici delle variabili ordinali e di adattare l'analisi alle caratteristiche dei dati e agli obiettivi dell'applicazione.

Nel contesto dello *Spectral Clustering*, queste misure di similarità per variabili ordinali sono ampiamente utilizzate per costruire la matrice di similarità utilizzata come input per l'algoritmo di clustering. La scelta della misura dipende dalla natura dei dati, dall'obiettivo dell'analisi e dalle proprietà desiderate per il clustering finale.

È importante considerare che, utilizzando tali misure, è possibile ottenere una rappresentazione accurata dei dati ordinali per il clustering, consentendo di identificare gruppi omogenei sulla base dell'ordinamento delle variabili.

Tuttavia, è fondamentale sottolineare che la scelta della funzione di similarità dipende dal dominio da cui provengono i dati e dalle caratteristiche specifiche del problema in esame. Non esistono consigli generali per la scelta della misura di similarità, ma è necessario valutare attentamente le proprietà dei dati e gli obiettivi dell'analisi al fine di selezionare la misura più appropriata per il contesto specifico.

2.4 Teoria dei grafi

Dato un insieme di punti x_1, …, x_n e una similarità $s_{ij} \geq 0$ tra tutte le coppie di punti x_i e x_j, l'obiettivo fondamentale del clustering è quello di suddividere i punti in diversi gruppi in modo che i punti dello stesso gruppo siano simili tra loro, mentre i punti di gruppi diversi siano dissimili (Schaeffer, 2007; Zhou, Cheng, & Yu, 2009).

Quando ci basiamo esclusivamente sulla similarità tra i punti, una rappresentazione efficace consiste nell'utilizzare un grafo di similarità $G = (V, E)$ (Bondy & Murty, 2008; Tutte, 2001; West, 2001). In questo grafo, ogni vertice v_i rappresenta un punto x_i, mentre gli archi rappresentano le connessioni tra i punti. La presenza di un arco tra due vertici indica una similarità positiva o superiore a

una determinata soglia, e il peso dell'arco è determinato dalla misura di similarità s_{ij}.

Riformulando il problema del clustering utilizzando il grafo di similarità, l'obiettivo diventa trovare una partizione del grafo tale che gli archi tra soggetti provenienti da gruppi diversi abbiano pesi molto bassi, indicando una bassa similarità tra i punti dei diversi cluster, mentre gli archi all'interno di uno stesso gruppo abbiano pesi elevati, indicando una forte similarità tra i punti all'interno dello stesso cluster. Questa formulazione mira a creare cluster compatti e ben separati (von Luxburg, 2007).

Per poter formalizzare questa intuizione, è necessario introdurre la notazione di base sui grafi e discutere successivamente i tipi di grafi che saranno oggetto di studio nel contesto del clustering.

Sia $G = (V, E)$ un grafo non diretto, cioè simmetrico e non orientato, composto da un insieme di vertici $V = \{v_1, \dots, v_n\}$ e da un insieme di archi $E \subseteq V \times V$ che collegano coppie di vertici. I vertici che sono collegati da un arco sono detti adiacenti.

Considerando *V* ed *E*, sia $w : E \to \mathrm{R}^+$ una funzione che associa un peso $w_{ij} \geq 0$ a ogni arco. In tal caso, il grafo $G = (V, E, w)$ si dice grafo pesato.

La matrice di adiacenza $W = (w_{ij})$, con $i, j = 1, \dots, n$, è definita come la matrice che rappresenta i pesi degli archi del grafo. Se $w_{ij} = 0$ significa che i vertici v_i e v_j non sono collegati da un arco. Poiché *G* è un grafo non diretto, è necessario che $w_{ij} = w_{ji}$.

Il grado di un vertice $v_i \in V$ è definito come:

$$d_i = \sum_{j=1}^{n} w_{ij},$$

(2.44)

Questo valore indica il numero di vertici adiacenti a v_i e corrisponde alla somma dei pesi della riga i-esima della matrice di adiacenza $\boldsymbol{W}$.

La matrice dei gradi $\boldsymbol{D} = diag(d_1, \dots, d_n)$ è una matrice diagonale che rappresenta i gradi dei vertici del grafo.

Infine, la nozione di connettività e componenti connesse è fondamentale nel contesto del clustering, poiché l'obiettivo è quello di trovare gruppi di punti che siano internamente connessi e isolati dal resto dei dati. Le componenti connesse

forniscono una struttura naturale per rappresentare tali gruppi e consentono di identificare i cluster all'interno del grafo di similarità.

2.5 Differenti grafi di similarità e matrice di affinità

Lo *Spectral Clustering* è un algoritmo appartenente alla categoria *graph-based clustering*, capace di ridurre la complessità di dataset ad alta dimensionalità. Sfrutta informazioni contenute negli autovalori e autovettori di matrici particolari costruite a partire dal grafo di similarità o direttamente dal dataset, al fine di ridurre la dimensionalià dei dati e facilitare il processo di clustering.

L'idea fondamentale di questi algoritmi risiede nell'utilizzo dell'algebra lineare per costruire grafi in grado di rappresentare gli elementi del dataset in uno spazio di dimensioni inferiori. I collegamenti presenti nel grafo riflettono la similarità tra gli elementi e giocano un ruolo essenziale nell'identificazione dei cluster. Questa famiglia di algoritmi è particolarmente adatta quando si lavora con dati di elevata dimensionalità, in quanto l'approccio basato sull'algebra lineare semplifica notevolmente i calcoli e la manipolazione dei dati.

Tuttavia, la costruzione del grafo di similarità non è un processo univoco, ma dipende dalla scelta della misura di similarità utilizzata. In letteratura, sono state proposte diverse misure di similarità per la costruzione dei grafi. Quando si costruiscono i grafi di similarità, l'obiettivo è modificare le relazioni di vicinato locale tra i punti dati. Tra le tipologie di grafi abbiamo:

- *ε-neighbourhood graph*: in questo tipo di grafo vengono collegati tutti i punti la cui distanza reciproca è inferiore a una specifica soglia ε. Questo grafo rappresenta una connessione binaria tra i punti, il che significa che o c'è una connessione o non c'è. Poiché le distanze tra tutti i punti collegati nel grafo di vicinato ε sono approssimativamente della stessa scala (al massimo ε), la ponderazione degli archi non aggiungerebbe ulteriori informazioni significative al grafo. Di conseguenza, l'*ε-neighbourhood graph* viene solitamente considerato come un grafo non pesato, in cui gli archi hanno un peso costante o sono semplicemente presenti o assenti.
  Questo approccio è efficace quando si desidera identificare gruppi o cluster di punti dati che sono vicini tra loro, senza preoccuparsi delle connessioni a lunga distanza.
  Tuttavia, è importante notare che la scelta del valore di ε è cruciale per determinare la connettività del grafo e quindi l'efficacia del clustering, un aspetto che verrà esaminato in dettaglio nel paragrafo (2.5.1) "Dettagli

pratici". Un valore ε troppo piccolo potrebbe infatti portare a cluster molto piccoli o addirittura a punti isolati, mentre un valore ε troppo grande potrebbe connettere troppi punti e ridurre la discriminazione tra i cluster.
Pertanto, è fondamentale selezionare attentamente il valore di ε in base alle caratteristiche dei dati e agli obiettivi specifici dell'analisi per ottenere risultati coerenti e rilevanti nel processo di clustering utilizzando l'*ε-neighbourhood graph*. La Figura 2.4 offre un esempio visuale di un *ε-neighbourhood graph*.

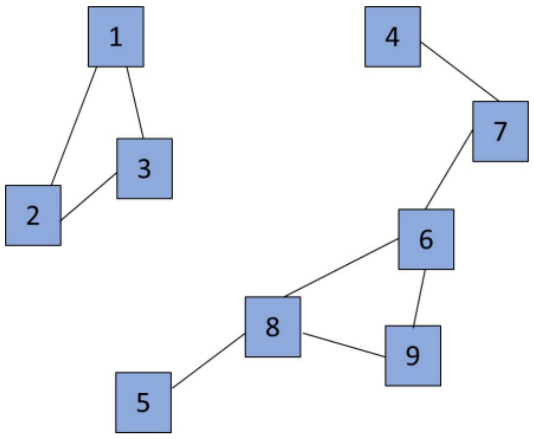


Figura 2.4 – Esempio di ε-neighbourhood graph.

- *k-nearest neighbor graph*: in questo caso l'obiettivo è collegare il vertice v_i con il vertice v_j se quest'ultimo è tra i k vicini più prossimi di v_i. Tuttavia, questa definizione iniziale porta a un grafo diretto, poiché la relazione di vicinato non è simmetrica.
  Per ottenere un grafo non diretto a partire dal *k-nearest neighbor graph*, esistono due approcci comuni. Il primo approccio consiste semplicemente nell'ignorare le direzioni degli archi, ossia collegare v_i e v_j con un arco non direzionato se v_i è tra i k vicini più prossimi di v_j o se v_j è tra i k vicini più prossimi di v_i. Il grafo risultante è quello che di solito viene chiamato *k-nearest neighbor graph*. Il secondo approccio consiste nel collegare i vertici v_i e v_j solo se v_i è tra i k vicini di v_j e v_j è tra i k vicini di v_i. Il grafo risultante è chiamato *mutual k-nearest neighbor graph*.
  In entrambi i casi, dopo aver collegato correttamente i vertici, gli archi vengono pesati in base alla similarità tra i loro punti finali.
  La scelta tra il *k-nearest neighbor graph* e il *mutual k-nearest neighbor graph* dipende dal contesto specifico e dagli obiettivi dell'analisi. Entrambi gli approcci hanno le loro peculiarità e possono essere utilizzati in base alla natura dei dati e alle relazioni che si desidera evidenziare.

Questa questione sarà esplorata in dettaglio nel paragrafo (2.5.1) “Dettagli pratici”.

È importante sottolineare che la costruzione del *k-nearest neighbor graph* richiede la scelta adeguata del valore di *k*. Un valore *k* troppo piccolo può portare a una connettività troppo ridotta, mentre un valore *k* troppo grande può causare una connettività troppo ampia e una perdita di discriminazione tra i cluster.

Pertanto, nella pratica è necessario eseguire prove empiriche con diversi valori di *k* per determinare quale valore ottimale produca i risultati desiderati e identifichi in modo accurato la struttura dei dati nel processo di clustering utilizzando il *k-nearest neighbor graph*. La Figura 2.5 illustra un esempio visuale di un *k-nearest neighbor graph*.

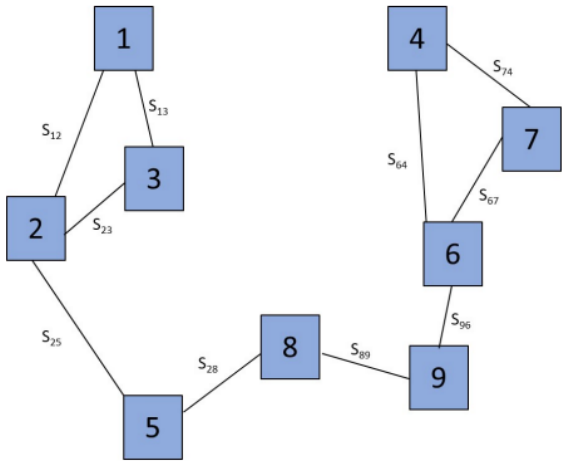


Figura 2.5 – Esempio di k-nearest neighbor graph.

- *fully connected graph*: il *fully connected graph*, noto anche come grafo completamente connesso, viene costruito collegando tra loro tutti i punti che presentano una similarità positiva e, successivamente, assegnando un peso agli archi del grafo basato sui valori di similarità s_{ij}. Questo approccio è vantaggioso quando la funzione di similarità stessa modella adeguatamente le relazioni di vicinato locale tra i punti dati.

  Un esempio comune di funzione di similarità è la funzione di similarità gaussiana, in cui il parametro σ controlla l’ampiezza dei quartieri o delle regioni di vicinato.

  Un valore di σ più grande implica una connessione più ampia tra i punti, mentre un valore di σ più piccolo limita la connessione ai punti più vicini. In questo modo, il parametro σ svolge un ruolo simile al parametro ε nel caso dell’*ε-neighbourhood graph*. Entrambi questi parametri regolano

l’ampiezza delle connessioni nel grafo e influenzano la definizione dei vicinaggi locali tra i punti.
La costruzione del grafo completamente connesso è particolarmente adatta quando la funzione di similarità rappresenta in modo accurato le relazioni locali tra i punti. Tuttavia, è importante considerare attentamente la scelta del parametro σ o dei criteri di similarità per garantire che il grafo rifletta le caratteristiche desiderate dei dati e favorisca una rappresentazione significativa per il processo di clustering. Anche quest’ultimo aspetto verrà esplorato in dettaglio nel paragrafo (2.5.1). Nella Figura 2.6 è presentato un esempio di un *fully connected graph*.

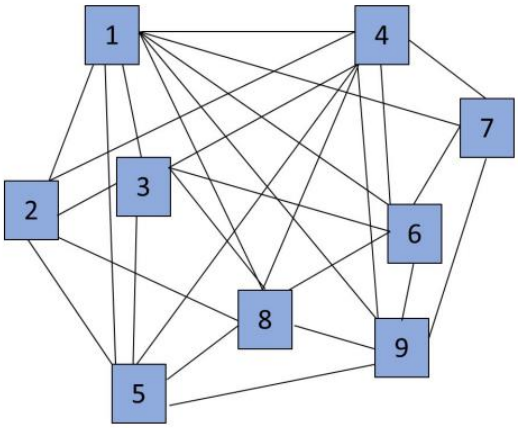


Figura 2.6 – Esempio di fully connected graph.

Tutti i grafi sopra menzionati sono ampiamente utilizzati nello *Spectral Clustering* per rappresentare i dati e identificare i cluster (von Luxburg, 2007).

Tuttavia, è importante sottolineare che attualmente non esistono risultati teorici che stabiliscano come la scelta specifica del grafo di similarità influenzi direttamente i risultati dello *Spectral Clustering*. Questo significa che non possiamo stabilire con certezza quale grafo di similarità sarà il più adatto per un determinato dataset o problema di clustering.

La scelta del grafo di similarità dipenderà quindi dalle caratteristiche dei dati, dalle conoscenze specifiche del dominio e dalle preferenze dell’utente. È una decisione che richiede una valutazione attenta e spesso empirica, basata sulla comprensione delle relazioni tra i punti dati e l’obiettivo del clustering.

Una volta costruito il grafo di similarità, si procede al calcolo della matrice di affinità, che svolge un ruolo cruciale nella rappresentazione dei dati e nell’identificazione dei cluster. Essa rappresenta in forma numerica le relazioni di vicinato tra i punti presenti nel grafo di similarità, consentendo di catturare la

struttura intrinseca dei dati e di rappresentarli in uno spazio di dimensioni inferiori.

La matrice di affinità deve soddisfare due requisiti fondamentali. In primo luogo, tutti i valori devono essere positivi, riflettendo la similarità tra i punti. Valori più alti indicano una maggiore similarità, mentre valori più bassi indicano una minor similarità. Questo permette di stabilire quanto i punti siano vicini o lontani l'uno dall'altro e influisce sulla formazione dei cluster.

In secondo luogo, la matrice di affinità deve essere simmetrica per rispecchiare la simmetria delle relazioni di vicinato. Ciò significa che se il punto A è vicino al punto B, allora il punto B sarà anche vicino al punto A. La simmetria nella matrice di affinità riflette l'equivalenza delle relazioni di vicinato tra i punti dati e contribuisce alla corretta identificazione dei cluster.

La matrice di affinità può essere calcolata in base alla costruzione del grafo di similarità utilizzando una funzione di similarità appropriata. Ad esempio, nel caso del grafo di vicinato ε, la matrice di affinità può essere ottenuta assegnando un valore costante di affinità ai punti collegati nel grafo. Nel caso del *k-nearest neighbor graph*, gli archi vengono pesati in base alla similarità tra i punti finali, mentre nel *fully connected graph*, gli archi vengono ponderati utilizzando una funzione di similarità come la funzione gaussiana.

È importante notare che la scelta del grafo di similarità influenzerà direttamente la matrice di affinità e, di conseguenza, i risultati dello *Spectral Clustering*. Pertanto, è fondamentale valutare attentamente la costruzione del grafo e il calcolo della matrice di affinità in base alle caratteristiche dei dati e agli obiettivi specifici del clustering. Esplorare diverse opzioni di costruzione del grafo e misure di similarità può portare a risultati diversi e aiutare a ottenere cluster coerenti e informativi, consentendo una migliore comprensione della struttura dei dati.

2.5.1 Dettagli pratici

La costruzione del grafo di similarità nel contesto dello *Spectral Clustering* è un compito complesso e ancora oggetto di ricerca. Attualmente, le implicazioni teoriche delle diverse costruzioni del grafo di similarità non sono completamente comprese. In questa sezione, il nostro obiettivo è quello di sensibilizzare il lettore sui problemi generali che possono sorgere e fornire un'idea generale delle considerazioni da tenere in considerazione.

Prima di procedere con la costruzione del grafo di similarità, è fondamentale definire una funzione di similarità appropriata per i dati in esame. Poiché lo *Spectral Clustering* coinvolge la costruzione di un grafo dei vicini, è importante

assicurarsi che i vicinaggi locali derivati da questa funzione di similarità siano rilevanti. Ciò implica che i punti considerati "molto simili" secondo la funzione di similarità siano anche strettamente correlati nell'applicazione specifica dei dati.

Tuttavia, l'aspetto globale a lungo raggio della funzione di similarità non è così rilevante per lo *Spectral Clustering*. Ad esempio, la differenza tra una similarità di 0,01 e 0,001 tra due punti non avrà un impatto significativo sulla costruzione del grafo di similarità, poiché non collegheremo comunque quei due punti nel grafo di similarità. In definitiva, la scelta della funzione di similarità dipende dal dominio da cui provengono i dati e non è possibile fornire consigli generali.

La scelta successiva da fare riguarda il tipo di grafo che si vuole utilizzare, come il *k-nearest neighbor graph* o l'*ε-neighbourhood graph*.

Per quanto riguarda l'*ε-neighbourhood graph*, è importante selezionare un valore adeguato per il parametro ε. Tuttavia, può essere difficile determinare un valore ottimale, soprattutto quando si lavora con dati che presentano scale diverse. Ciò significa che le distanze tra i punti dei dati possono variare significativamente in diverse regioni dello spazio. Questa variabilità può portare a problemi nella costruzione del grafo di similarità basato su ε.

Il *k-nearest neighbor graph*, d'altra parte, ha la capacità di collegare punti "su scale diverse". Questa è una caratteristica preziosa, in quanto può gestire dati con variazioni di scala e connettere punti che potrebbero essere lontani in termini di distanza euclidea ma vicini in termini di similarità. Un altro aspetto importante del *k-nearest neighbor graph* è che può essere sensibile alla densità dei dati. In presenza di regioni ad alta densità lontane l'una dall'altra, il *k-nearest neighbor graph* può frammentarsi in diverse componenti connesse (Lucińska & Wierzchoń, 2012).

Il *mutual k-nearest neighbor graph* ha la proprietà di tendere a collegare i punti all'interno di regioni a densità costante senza collegare tra loro regioni a densità diversa. Questo tipo di grafo può quindi essere considerato come una "via di mezzo" tra l'*ε-neighbourhood graph* e il *k-nearest neighbor graph*. È in grado di agire su scale diverse, ma non mescola tali scale tra loro. Pertanto, il *mutual k-nearest neighbor graph* risulta particolarmente adatto se si desidera individuare cluster di diversa densità (von Luxburg, 2007).

Il *fully connected graph* è spesso utilizzato in combinazione con la funzione di similarità gaussiana, in cui il parametro σ svolge un ruolo simile a quello del parametro ε nell'*ε-neighbourhood graph*.

Come raccomandazione generale, suggeriamo di utilizzare inizialmente il *k-nearest neighbor graph* come prima scelta (von Luxburg, 2007). Questo tipo di grafo è semplice da implementare e meno sensibile a scelte inadeguate di parametri rispetto agli altri tipi di grafi. Tuttavia, la scelta del grafo di similarità dipende sempre dalla natura dei dati, dalle caratteristiche specifiche del problema e dagli obiettivi dello *Spectral Clustering*.

Una volta deciso il tipo di grafo di similarità, si deve scegliere il suo parametro di connettività k o ε, rispettivamente. Purtroppo la scelta non è un compito semplice e non esistono risultati teorici che ci guidino in modo preciso. Tuttavia, possiamo fornire alcune linee guida generali per affrontare questa scelta.

In generale, se il grafo di similarità ha un numero di componenti connesse superiore al numero di cluster che ci aspettiamo di individuare, lo *Spectral Clustering* potrebbe restituire risultati insoddisfacenti, poiché le componenti connesse verrebbero etichettate come cluster separati. A meno che non si sia certi che tali componenti connesse corrispondano ai cluster desiderati, è importante assicurarsi che il grafo di similarità sia connesso, o che sia costituito da "poche" componenti connesse e da pochissimi o nessun vertice isolato.

Esistono diversi risultati teorici su come si possa ottenere la connettività dei grafi casuali, ma tutti questi risultati valgono solo nel limite in cui la dimensione del campione tende all'infinito ($n \to \infty$) (von Luxburg, 2007).

Nella pratica, la scelta di k o ε dipende fortemente dalla natura dei dati, dalle caratteristiche specifiche del problema e dagli obiettivi dello *Spectral Clustering*. È consigliabile eseguire diversi esperimenti con valori differenti di k o ε e valutare i risultati ottenuti. Si può anche fare affidamento su tecniche di validazione incrociata (*cross-validation*) o criteri di selezione del modello per trovare il valore ottimale del parametro di connettività che fornisce i risultati di clustering più rilevanti e coerenti con i dati.

Diamo ora alcune regole empiriche. Quando si lavora con il *k-nearest neighbor graph*, la scelta del parametro di connettività k è cruciale. Dovrebbe essere scelto in modo che il grafo risultante sia connesso o, almeno, abbia un numero di componenti connesse significativamente inferiore rispetto ai cluster che vogliamo individuare (Jonsson & Wohlin, 2004). Purtroppo, non esistono metodi statistici predefiniti per determinare il valore ottimale di k. La scelta di un valore di k molto piccolo porta a confini decisionali instabili e rende il risultato più sensibile al rumore, mentre la scelta di un valore elevato lo renderà computazionalmente costoso (Jönsson & Wohlin, 2006).

Per grafi di piccole o medie dimensioni è possibile sperimentare diversi valori di *k* in modo empirico. Tuttavia, quando si ha a che fare con grafi di dimensioni molto grandi, può essere utile seguire un'indicazione approssimativa. Una regola empirica suggerita da Duda e Hart è quella di scegliere *k* nell'ordine di $k \approx \sqrt{n}$ (Duda & Hart, 1973), dove *n* rappresenta il numero di punti dati nel set di addestramento (Jonsson & Wohlin, 2004).

Inoltre, è preferibile selezionare un valore dispari per *k*, per evitare ambiguità tra due classi di dati[17] [18]. Pertanto, la scelta suggerita è quella di esaminare $k = RoundOdd\sqrt{n}$, ovvero la radice quadrata del numero medio di casi completi dopo la rimozione dei dati, arrotondata al numero dispari più vicino (Jönsson & Wohlin, 2006).

I risultati ottenuti da Jönsson & Wohlin dimostrano che utilizzare la radice quadrata del numero di casi completi, arrotondata al numero dispari più vicino, è un modello adeguato per selezionare il valore di *k* (Jönsson & Wohlin, 2006).

Per *mutual k-nearest neighbor graph*, ci sono meno regole empiriche disponibili.

Questo tipo di grafo ha il vantaggio di non collegare regioni con diversa densità di punti. Se ci sono chiari cluster associati a regioni separate ad alta densità, questo può essere un vantaggio. Tuttavia, in situazioni meno ovvie, potrebbe portare a risultati indesiderati poiché le parti disconnesse del grafo verrebbero trattate come cluster separati dallo *Spectral Clustering*. In generale, è importante notare che il *mutual k-nearest neighbor graph* ha meno bordi rispetto al *k-nearest neighbor graph* standard per lo stesso parametro *k*. Pertanto, è consigliabile scegliere un valore di *k* significativamente più grande per lo *Spectral Clustering* con il *mutual k-nearest neighbor graph*.

Tuttavia, per sfruttare la proprietà del *mutual k-nearest neighbor* graph di non collegare regioni con densità diversa, è necessario consentire diverse parti del grafo siano "significativamente" disconnesse (von Luxburg, 2007). Purtroppo, non esistono euristiche generali per scegliere il parametro *k* in modo da ottenere questo risultato. La scelta dipende fortemente dalla struttura specifica dei dati e richiede un'analisi attenta del contesto e degli obiettivi dello *Spectral Clustering*.

Per quanto riguarda il grafo *ε-neighborhood*, una regola empirica consigliata è selezionare un valore di ε in modo che il grafo risultante sia connesso in modo robusto. Per determinare il valore più piccolo di ε in cui il grafo è connesso è

[17] https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e
[18] https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb

semplice, si può utilizzare l'approccio del *minimum spanning tree*. In particolare, si può scegliere ε come la lunghezza del bordo più lungo in un *minimum spanning tree* del grafo completamente connesso che collega tutti i punti dati. Questo *minimum spanning tree* può essere ottenuto facilmente utilizzando un algoritmo di *minimum spanning tree* (Favati & al., 2020; von Luxburg, 2007).

Tuttavia, è importante considerare alcune limitazioni di questa euristica. Se i dati contengono *outlier*, questa regola può selezionare un valore di ε troppo grande, collegando gli outlier al resto dei dati. Allo stesso modo, se i dati contengono diversi cluster stretti e distanti l'uno dall'altro, ε potrebbe essere scelto troppo grande per rappresentare correttamente la scala dei cluster più rilevanti (von Luxburg, 2007). Pertanto, è necessario valutare attentamente il contesto specifico dei dati e considerare l'impatto degli *outlier* e della distanza tra i cluster nella scelta di ε.

Infine, se si utilizza il *fully connected graph* in combinazione con una funzione di similarità scalabile, come ad esempio la funzione di similarità gaussiana, è di fondamentale importanza scegliere attentamente la scala della funzione di similarità in modo da ottenere un grafo con proprietà simili a quelle che avrebbe un grafo *k-nearest neighbor* o *ε-neighborhood* corrispondente.

In particolare, è necessario trovare un equilibrio nella scala della funzione di similarità in modo che l'insieme dei vicini con una similarità significativamente maggiore di zero non sia né troppo piccolo né troppo grande per la maggior parte dei punti dati.

Per la funzione di similarità gaussiana, esistono diverse regole empiriche comunemente adottate per selezionare il parametro σ. Ad esempio, una pratica comune è quella di impostare il parametro σ in modo che sia dell'ordine della distanza media di un punto dal suo k-esimo vicino, dove k viene scelto in modo simile a quanto discusso in precedenza. Un altro approccio possibile è determinare precedentemente il parametro ε utilizzando l'euristica del *minimum spanning tree* descritta in precedenza e successivamente impostare $\sigma = \varepsilon$ (von Luxburg, 2007; Favati & al., 2020).

Il parametro di Kernel σ ha un notevole impatto sulla struttura di una matrice di affinità e sulla qualità del partizionamento di clustering risultante, ed è generalmente difficile da determinare in modo ottimale. Gli studiosi hanno sviluppato diverse strategie per cercare di trovare il valore ottimale di σ e alcuni autori hanno proposto euristiche per calcolarlo.

Alcuni autori suggeriscono l'uso di un valore globale di σ per l'intero set di dati, come fatto da Ng et al. (2001) e Shi et al. (2009), mentre altri preferiscono l'uso di un parametro locale, come suggerito da Zelnik-Manor & Perona (2004).

La strategia di Ng et al. (2001) consiste nell'eseguire un algoritmo di clustering per diversi valori di σ (Favati & al., 2020; Fischer & Poland, 2005), esplorando sistematicamente un'ampia gamma di valori di σ per ciascun set di dati e valutando l'effetto sul grafo di similarità e sui risultati del clustering al fine di trovare la scala più appropriata per i dati specifici (Lucińska & Wierzchoń, 2012). Tuttavia, questo metodo potrebbe non essere efficace a causa della necessità di impostare parametri aggiuntivi e dei costi computazionali elevati (Nascimento & de Carvalho, 2011). Per ulteriori dettagli sull'argomento, si consiglia di fare riferimento a Favati et al. (2020).

Tuttavia, per molti set di dati, nessuna delle strategie menzionate può garantire una classificazione corretta (Lucińska & Wierzchoń, 2012), poiché tali soluzioni spesso non riescono a catturare appieno le proprietà dei set di dati del mondo reale (Xia, Cao, Zhang, & Li, 2008).

Di conseguenza, la maggior parte degli algoritmi spettrali sostituisce σ con valori approssimativi, come 0,5, 1 e 2 per semplicità (Ghojogh, Ghodsi, Karray, & Crowley, Laplacian-based dimensionality reduction including spectral clustering, Laplacian eigenmap, locality preserving projection, graph embedding, and diffusion map: Tutorial and survey, 2021), anche se tali valori potrebbero non essere adatti per molti set di dati e potrebbero influire negativamente sull'accuratezza dell'analisi spettrale. Tuttavia esistono algoritmi come quello di Bhissy et al. (2014) che cercano di stimare il valore di σ basandosi sulle caratteristiche del dataset stesso.

In conclusione, è importante sottolineare che tutte queste regole empiriche sono molto ad hoc e dipendono fortemente dalla natura specifica dei dati e dalla distribuzione delle distanze tra i punti.

In generale, l'esperienza pratica ha evidenziato che lo *Spectral Clustering* può essere molto sensibile alle variazioni del grafico di similarità e alla scelta dei suoi parametri. Purtroppo, al momento non esiste uno studio sistematico approfondito che esamini in modo completo gli effetti del grafo di similarità e dei suoi parametri sul processo di clustering, fornendo regole empiriche ben giustificate. Nessuna delle raccomandazioni di cui sopra si basa su una solida base teorica (von Luxburg, 2007).

Trovare regole che abbiano una giustificazione teorica dovrebbe essere considerato un argomento interessante e importante per la ricerca futura. L'obiettivo sarebbe quello di identificare metriche o criteri che guidino in modo più chiaro la scelta del grafo di similarità e dei suoi parametri, fornendo una solida base teorica per le decisioni di *Spectral Clustering* (von Luxburg, 2007).

3. Matrici Laplaciane e Riduzione della Dimensionalità

Questo capitolo approfondisce la definizione di matrice Laplaciana del Grafo e spiega la sua importanza nel contesto della riduzione della dimensionalità e dell'analisi dei dati mediante l'uso degli algoritmi di *Spectral Clustering*. Conoscere a fondo questi concetti riveste un ruolo cruciale per acquisire una comprensione completa delle metodologie e delle tecniche applicate all'interno dello *Spectral Clustering*.

Inizieremo esaminando la matrice Laplaciana, uno strumento chiave che ci consente di analizzare le relazioni di connettività tra i vertici del grafo. Questa matrice svolge un ruolo centrale nell'analisi degli autovettori e degli autovalori, aspetti essenziali per comprendere il funzionamento dello *Spectral Clustering*. Tuttavia, è importante notare che la matrice Laplaciana non è unica, ma esistono diverse varianti che aprono le porte alla Teoria Spettrale dei Grafi (Chung, 1997).

Per questo motivo, esamineremo la matrice Laplaciana normalizzata, una versione adattata della matrice Laplaciana non normalizzata che offre una rappresentazione geometrica dei dati indipendente dalle dimensioni del grafo. Questa forma normalizzata risulta particolarmente interessante per le nostre analisi, in quanto consente una migliore interpretazione geometrica dei dati e garantisce maggiore robustezza nelle soluzioni ottenute.

Ciascuna di queste matrici presenta un approccio diverso alla rappresentazione dei grafi e alla gestione delle informazioni di similarità tra i vertici, e la scelta tra di esse dipende dalle specifiche caratteristiche del dataset (von Luxburg, 2007).

Il calcolo della matrice Laplaciana del grafo è finalizzato all'individuazione degli autovalori e degli autovettori necessari per la proiezione dei punti dati in uno spazio a bassa dimensione. Questa proiezione rappresenta un passaggio fondamentale, in quanto permette di semplificare la complessità computazionale e di ottenere una visualizzazione dei dati più significativa (Van Der Maaten, 2009). Tale proiezione costituirà la base su cui costruiremo il nostro approccio di *Spectral Clustering*.

L'idea chiave che guida lo *Spectral Clustering* è la trasformazione dei dati da uno spazio di input a un sottospazio a bassa dimensionalità, rendendo i cluster più chiaramente distinguibili. Successivamente, applicheremo un algoritmo di clustering, in particolare l'algoritmo *k-means*, per assegnare i dati ai cluster appropriati (Ghodsi, 2021). Tuttavia, la determinazione del numero ottimale di cluster costituisce una sfida critica, poiché una scelta errata può compromettere la qualità dei risultati.

Continuando, entreremo nelle fasi conclusive dell'algoritmo di *Spectral Clustering* proposto da Ng et al. (2001), come descritto nella Sezione (1.3.1). Questo algoritmo si basa sull'intuizione che gli autovettori associati agli autovalori più piccoli racchiudano informazioni preziose sulla struttura di connettività del grafo. Sfruttando questi autovettori, l'algoritmo è in grado di individuare pattern e relazioni tra i vertici del grafo, consentendo una suddivisione rilevante del grafo in base a tali informazioni (Chung, 1997).

Questo capitolo costituirà dunque una solida base per la comprensione e l'implementazione dei metodi spettrali utilizzati per analizzare dati complessi e rappresentare informazioni di similarità in modo efficace. Attraverso questa esplorazione, miriamo a fornire una solida base per la comprensione della matrice Laplaciana e della sua relazione con la riduzione della dimensionalità nel contesto dello *Spectral Clustering*. Queste conoscenze saranno di fondamentale importanza per valutare l'efficacia dell'algoritmo di *Spectral Clustering* proposto da Ng et al. (2001) nell'applicazione pratica illustrata nel Capitolo 4.

3.1 Rappresentazioni matriciali dei grafi di similarità

Gli strumenti fondamentali per l'applicazione dello *Spectral Clustering* sono le matrici Laplaciane dei grafi. Questo concetto riveste un ruolo centrale nella Teoria Spettrale dei Grafi, un campo di studio dedicato all'esplorazione delle proprietà spettrali di una particolare matrice chiamata Laplaciana (Chung, 1997; Spielman, 2019).

La Teoria Spettrale dei Grafi offre un modo innovativo per comprendere e analizzare i dati rappresentati dai grafi (Chung, 1997; Spielman, 2019), trattandoli come una sorta di approssimazione di uno spazio topologico.[19] Attraverso l'analisi delle proprietà spettrali della matrice Laplaciana, è possibile caratterizzare i grafi e sfruttare queste informazioni per eseguire partizionamenti e raggruppamenti adeguati dei dati, un processo essenziale nell'ambito dello *Spectral Clustering*.

Il termine "*Spectral*" deriva proprio dal fatto che queste proprietà spettrali della matrice Laplaciana sono utilizzate per caratterizzare i grafi e consentono di analizzare e interpretare i dati in modo efficace (Jimenez, 1997).

[19] Lo spazio topologico è il più generale tipo di spazio con il quale, attraverso la nozione di intorno, si formalizzano relazioni di "vicinanza" e di "continuità" senza necessità d'introdurre concetti metrici quali per esempio quelli di distanza, di direzione o di angolo, che lo renderebbero una struttura più "rigida".

La possibilità di utilizzare la matrice Laplaciana per estrarre informazioni significative dai grafi offre un'importante opportunità per svolgere partizionamenti e raggruppamenti appropriati. In altre parole, le proprietà spettrali della matrice Laplaciana ci aiutano a individuare pattern, sottostrutture e relazioni tra i dati rappresentati dal grafo, fornendo una base solida per l'applicazione di tecniche di clustering che consentono di raggruppare dati simili in cluster distinti (Chung, 1997).

In questa sezione, il nostro obiettivo è quello di definire e approfondire le differenti varianti delle matrici Laplaciane del grafo. È importante sottolineare che nella letteratura scientifica non c'è una definizione univoca su quale matrice venga precisamente chiamata "matrice Laplaciana del grafo". Di fatto, ogni autore può attribuire questa designazione alla matrice che ritiene più rilevante per il proprio contesto di studio (von Luxburg, 2007). Pertanto, quando ci si avvicina alla lettura di articoli riguardanti le matrici Laplaciane dei grafi, è essenziale prestare attenzione alla definizione utilizzata in ciascun contesto specifico.

Per ogni grafo, come verrà discusso in dettaglio in questo capitolo, esiste un insieme di numeri non negativi chiamati autovalori del grafo (von Luxburg, 2007; Chung, 1997). Questi autovalori sono di importanza cruciale poiché rappresentano la "carta d'identità" del grafo stesso. Essi racchiudono gran parte delle informazioni rilevanti sulla struttura e le caratteristiche uniche del grafo, consentendo di esplorare aspetti fondamentali come la presenza di sottostrutture, componenti connesse e altre proprietà geometriche.

Gli autovalori del grafo rappresentano un modo potente per descrivere e analizzare le proprietà intrinseche del grafo. Ogni autovalore è associato a una corrispondente matrice Laplaciana del grafo e rappresenta una misura della variazione e delle interazioni dei vertici all'interno del grafo. La presenza di autovalori maggiori di zero indica l'esistenza di connessioni e relazioni tra i vertici, mentre gli autovalori nulli indicano la presenza di componenti connesse o sottografi[20] isolati all'interno del grafo.

Questi concetti rappresentano un fondamento essenziale per la comprensione del ruolo degli autovalori nell'ambito dello *Spectral Clustering* e della Teoria Spettrale dei Grafi. Gli autovalori e gli autovettori sono strumenti chiave per analizzare le proprietà strutturali e topologiche dei grafi, e la loro corretta interpretazione è cruciale per sfruttare al meglio le informazioni contenute nei dati rappresentati dai grafi (Chung, 1997).

[20] In matematica, un sottografo è un grafo ottenuto da un altro considerandone solo alcuni vertici e alcuni spigoli.

3.1.1 Matrice Laplaciana

Una volta costruita la matrice di similarità $\boldsymbol{W}_{n\times n}$, con $w_{ij} = w_{ji} \geq 0$, associata al grafo non orientato e ponderato *G*, è possibile eseguire lo *Spectral Clustering*.

La matrice di similarità $\boldsymbol{W}$ rappresenta le connessioni che caratterizzano un grafo. A partire da questa matrice $\boldsymbol{W}$ è possibile ottenere una matrice di fondamentale importanza nello *Spectral Clustering*: la matrice Laplaciana. In letteratura, si distinguono diverse definizioni della matrice Laplaciana, tra cui la matrice Laplaciana non normalizzata, la Laplaciana normalizzata simmetrica e la Laplaciana normalizzata *random walk* (Ling, 2020). Ognuna di queste fornisce informazioni diverse sul grafo e può essere impiegata per scopi specifici nello *Spectral Clustering* (von Luxburg, 2007).

La matrice Laplaciana non normalizzata del grafo è particolarmente significativa poiché riflette la struttura del grafo e mette in evidenza le differenze tra i vertici basate sul numero di vicini connessi. La sua definizione è data da:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W},$$

(3. 1)

dove $\boldsymbol{D}$ è la matrice diagonale dei gradi dei vertici del grafo *G*. La matrice diagonale $\boldsymbol{D}$ contiene i gradi $d_1, \dots, d_n$ sulla diagonale, dove:

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

(3. 2)

Nello *Spectral Clustering*, la matrice dei gradi $\boldsymbol{D}$ gioca un ruolo cruciale in quanto rappresenta una media locale della similarità tra i vertici del grafo. La matrice $\boldsymbol{D}$ viene impiegata come un fattore di normalizzazione per la matrice di similarità $\boldsymbol{W}$, come descritto da von Luxburg (2007).

La matrice dei gradi $\boldsymbol{D}$ e la matrice Laplaciana $\boldsymbol{L}$ hanno una forma esplicita data da:

$$\boldsymbol{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}, \qquad \boldsymbol{L} = \begin{bmatrix} d_1 - w_{11} & -w_{12} & \cdots & -w_{1n} \\ -w_{21} & d_2 - w_{22} & \cdots & -w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{n1} & -w_{n2} & \cdots & -w_{nn} \end{bmatrix}.$$

(3. 3)

Come è possibile notare, la matrice $\boldsymbol{D}$ è una matrice diagonale con i gradi $d_1, \dots, d_n$ come valori diagonali. Ogni d_i rappresenta la somma delle similarità

w_{ij} tra il vertice *i* e tutti i suoi vertici vicini *j*. Quindi, in sostanza, la matrice $\boldsymbol{D}$ riflette la similarità media locale tra i vertici del grafo. La matrice Laplaciana $\boldsymbol{L}$, come già menzionato, è ottenuta sottraendo la matrice $\boldsymbol{W}$ dalla matrice dei gradi $\boldsymbol{D}$. Quindi, $\boldsymbol{L}$ mostra le differenze tra i gradi dei vertici e le loro connessioni, evidenziando la struttura del grafo.

È interessante notare che la matrice Laplaciana è anche una matrice diagonale dominante, il che significa che il valore assoluto delle voci sulla diagonale è maggiore o uguale alla somma di tutte le altre voci nella stessa riga o colonna (Ling, 2020).

La diagonale dominante della matrice Laplaciana è un'importante proprietà che contribuisce alla stabilità e alla convergenza degli algoritmi di *Spectral Clustering*. Questa caratteristica garantisce che la matrice sia ben condizionata, il che è essenziale per ottenere risultati accurati e stabili durante la fase di clustering. Inoltre, la struttura specifica della matrice Laplaciana consente di estrarre informazioni rilevanti per l'aggregazione dei vertici in cluster durante il processo di *Spectral Clustering* (Ling, 2020).

La matrice Laplaciana del grafo presenta alcune proprietà importanti, le quali vengono esposte e dimostrate in riferimenti come Mohar (1991; 1997) e von Luxburg (2007), i quali offrono una solida base teorica per il concetto di matrice Laplaciana del grafo.

È importante sottolineare che la matrice Laplaciana non normalizzata del grafo dipende solo dalle connessioni tra i vertici rappresentate dalla matrice di adiacenza $\boldsymbol{W}$. Gli elementi diagonali di $\boldsymbol{W}$, che rappresentano i collegamenti di un vertice con se stesso, non influenzano la matrice Laplaciana del grafo. In altre parole, qualsiasi matrice di adiacenza che coincida con $\boldsymbol{W}$ sugli elementi extradiagonali produrrà la stessa matrice Laplaciana non normalizzata $\boldsymbol{L}$. Questo è un risultato interessante perché dimostra come questi elementi diagonali non hanno alcun effetto sulla struttura complessiva del grafo quando si utilizza la matrice Laplaciana per lo *Spectral Clustering* (von Luxburg, 2007).

3.1.2 Matrice Laplaciana normalizzata

In letteratura, troviamo due matrici importanti conosciute come Laplaciane normalizzate dei grafi: la Laplaciana normalizzata simmetrica ($\boldsymbol{L_{sym}}$) e la Laplaciana normalizzata *random walk* ($\boldsymbol{L_{rw}}$). Entrambe queste matrici sono strettamente correlate tra loro e forniscono una rappresentazione significativa della struttura dei grafi.

La Laplaciana normalizzata simmetrica, indicata con $\boldsymbol{L_{sym}}$, è ottenuta mediante una normalizzazione simmetrica dei pesi del grafo. Questa operazione coinvolge la matrice di adiacenza pesata ($\boldsymbol{W}$) e la matrice diagonale dei gradi dei vertici ($\boldsymbol{D}$). Nello specifico, per ottenere $\boldsymbol{L_{sym}}$, vengono divisi tutti gli elementi della matrice Laplaciana $\boldsymbol{L}$ per la radice quadrata del prodotto dei gradi dei vertici corrispondenti (Ling, 2020). In partica, si esprime come:

$$\boldsymbol{L_{sym}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$

(3. 4)

dove, $\boldsymbol{I}$ è la matrice identità di dimensione $n \times n$, $\boldsymbol{W}$ è la matrice di adiacenza pesata del grafo di similarità e $\boldsymbol{D}$ è la matrice diagonale contenete i gradi dei vertici del grafo (von Luxburg, 2007).

La normalizzazione simmetrica permette alla matrice Laplaciana di essere indipendente dalla dimensione del grafo e migliora la rappresentazione geometrica dei dati (Ling, 2020). Questa versione della matrice Laplaciana è particolarmente utile quando si desidera analizzare il grafo in modo invariante rispetto alla dimensione, mantenendo la sua struttura relativa.

D'altro canto, la Laplaciana normalizzata *random walk*, indicata con $\boldsymbol{L_{rw}}$, è ottenuta mediante una normalizzazione basata sul concetto di *random walk* (cammino casuale) nel grafo (Meilă M., 2001). La normalizzazione coinvolge sempre la matrice Laplaciana ($\boldsymbol{L}$) e la matrice diagonale dei gradi dei vertici ($\boldsymbol{D}$). In particolare, si esprime come:

$$\boldsymbol{L_{rw}} = D^{-1} L.$$

(3. 5)

In questa forma di normalizzazione, ogni elemento della matrice Laplaciana viene diviso per il grado del vertice corrispondente. Questo rispecchia il processo di *random walk*, dove la probabilità di spostarsi da un vertice all'altro è proporzionale ai gradi dei vertici (von Luxburg, 2007).

La Laplaciana normalizzata *random walk* è utile quando si vuole evidenziare il concetto di raggiungibilità tra i vertici del grafo. Attraverso il *random wa*lk, si possono individuare le probabilità relative di visitare diversi vertici e, di conseguenza, rivelare strutture di cluster o comunità nel grafo (Meilă M., 2001).

Queste differenti formulazioni delle matrici Laplaciane offrono, quindi, prospettive diverse sulla rappresentazione dei grafi e permettono di gestire in maniera differente le informazioni di similarità (o affinità) tra i vertici. La scelta

tra $\boldsymbol{L_{sym}}$ e $\boldsymbol{L_{rw}}$ dipenderà dagli obiettivi specifici dell'analisi, dalle caratteristiche del grafo e dalle strutture che si desidera evidenziare durante l'analisi del grafo. Entrambe le matrici Laplaciane normalizzate sono potenti strumenti per l'analisi dei grafi, lo *Spectral Clustering* e altre applicazioni in cui l'interpretazione delle relazioni tra i vertici è fondamentale (von Luxburg, 2007).

In von Luxburg (2007) e nel riferimento standard di Chung (1997) è possibile trovare i riferimenti e le dimostrazioni di alcune proprietà di $\boldsymbol{L_{sym}}$ e $\boldsymbol{L_{rw}}$, le quali offrono importanti informazioni sulle caratteristiche degli autovalori e autovettori delle matrici Laplaciane normalizzate $\boldsymbol{L_{sym}}$ e $\boldsymbol{L_{rw}}$ e costituiscono i principali risultati necessari per comprendere l'applicazione dello *Spectral Clustering*.

3.1.3 Scelta della matrice Laplaciana del grafo

La questione cruciale nello *Spectral Clustering* riguarda la scelta della matrice Laplaciana del grafo da utilizzare per calcolare gli autovettori. Prima di prendere una decisione in merito, è essenziale esaminare attentamente la distribuzione dei gradi nel grafo di similarità (von Luxburg, 2007).

Se il grafo è caratterizzato da un'elevata regolarità e la maggior parte dei vertici ha approssimativamente lo stesso grado, allora tutte le matrici Laplaciane ($\boldsymbol{L_{rw}}$, $\boldsymbol{L_{sym}}$ e la matrice Laplaciana non normalizzata) saranno molto simili tra loro e tenderanno a funzionare allo stesso modo per il clustering.

Tuttavia, se i gradi nel grafo sono distribuiti in modo ampio, con alcuni vertici che hanno gradi molto elevati e altri con gradi molto bassi, allora le matrici Laplaciane potrebbero risultare notevolmente diverse. In questo caso, la scelta della matrice Laplaciana giusta diventa cruciale poiché avrà un impatto significativo sulle prestazioni complessive dello *Spectral Clustering* (von Luxburg, 2007).

Nel nostro studio, abbiamo deciso di adottare l'algoritmo proposto da Ng et al. (2001), il quale si basa sull'utilizzo della matrice Laplaciana normalizzata simmetrica ($\boldsymbol{L_{sym}}$). Pertanto, ci concentreremo esclusivamente su questa specifica matrice durante la fase di sperimentazione. La scelta di $\boldsymbol{L_{sym}}$ è motivata dalla sua capacità di fornire una rappresentazione geometricamente significativa dei dati, indipendentemente dalle dimensioni del grafo, rendendola particolarmente adatta per affrontare una vasta gamma di strutture grafiche.

La nostra decisione di utilizzare $\boldsymbol{L_{sym}}$ è supportata dalle sue proprietà che ci consentono di interpretare i risultati in modo coerente e di estrarre informazioni rilevanti per l'analisi dei dati. Questa scelta ci permetterà di sfruttare al meglio le

potenzialità dello *Spectral Clustering* e di ottenere risultati accurati e affidabili per la nostra sperimentazione.

La comprensione del concetto di matrice Laplaciana e delle sue diverse definizioni nel contesto dello *Spectral Clustering* è essenziale per acquisire solide basi teoriche e una corretta implementazione dell'algoritmo proposto da Ng et al. (2001). Questa comprensione permette di sfruttare appieno le potenzialità dello *Spectral Clustering* per ottenere suddivisioni significative dei dati basate sulle relazioni strutturali del grafo.

Il calcolo della matrice Laplaciana del grafo è finalizzato all'individuazione degli autovalori e degli autovettori che permettono di incorporare i punti dati in uno spazio a bassa dimensione. Questa proiezione è di estrema utilità in molte applicazioni, poiché consente di ridurre la complessità computazionale e di visualizzare i dati in uno spazio più maneggevole e rilevante (Van Der Maaten, 2009).

Nel paragrafo successivo, ci concentreremo su come utilizzare gli autovalori e gli autovettori della matrice Laplaciana per ottenere una rappresentazione efficace dei dati e per eseguire lo *Spectral Clustering*. Questo passaggio è cruciale per ottenere suddivisioni coerenti e informative dei dati in base alle loro relazioni di similarità e connettività nel grafo.

3.2 Riduzione della dimensionalità

I dati del mondo reale spesso presentano una straordinaria complessità e sono caratterizzati da un elevato numero di dimensioni. Tuttavia, la gestione di dati ad alta dimensionalità può rivelarsi problematica in quanto richiede notevoli risorse computazionali e può portare a risultati poco affidabili a causa della cosiddetta "maledizione della dimensionalità" (*curse of dimensionality*) (Jimenez, 1997).

La "maledizione della dimensionalità" è un termine utilizzato per descrivere il fatto che l'aumento delle dimensioni dei dati può causare problemi di densità dei dati stessi e distribuzione sparsa, rendendo difficile la loro analisi e interpretazione. Inoltre, l'aumento delle dimensioni può portare a un eccessivo consumo di memoria e risorse, aumentando il tempo di calcolo necessario per l'esecuzione di algoritmi di *machine learning*. Infine, una maggiore dimensionalità può anche far sì che le relazioni tra le variabili diventino meno significative e meno rilevanti per la predizione o l'interpretazione dei dati (Van Der Maaten, 2009; Ghojogh B. , Data Reduction Algorithms in Machine Learning and Data Science, 2021).

Per superare queste sfide, la riduzione della dimensionalità si propone di trasformare i dati ad alta dimensionalità in una rappresentazione significativa di dimensionalità ridotta. L'obiettivo è quello di preservare al massimo le informazioni rilevanti presenti nei dati originali, mentre si riducono le dimensioni degli spazi rappresentanti (Van Der Maaten, 2009). In altre parole, la riduzione della dimensionalità mira a trovare una rappresentazione compatta e informativa dei dati, dove le caratteristiche principali sono mantenute e le caratteristiche meno rilevanti sono eliminate.

L'importanza della scelta della dimensionalità ridotta adeguata è legata al concetto di "dimensionalità intrinseca" dei dati. La dimensionalità intrinseca dei dati è il numero minimo di parametri necessari per spiegare le proprietà osservate dei dati (Fukunaga, 1990). In altre parole, rappresenta il numero effettivo di variabili o *feature* necessarie per descrivere in modo adeguato il dataset senza perdere informazioni importanti. La riduzione della dimensionalità mira a trovare questa dimensionalità intrinseca, in modo da evitare la sovra-rappresentazione o sotto-rappresentazione dei dati.

Tradizionalmente, la riduzione della dimensionalità è stata effettuata attraverso tecniche lineari come l'Analisi delle Componenti Principali (PCA) (Pearson, 1901), l'analisi dei fattori (Spearman, 1904) e la scalatura classica (*classical scaling*) (Torgerson, 1952). Queste tecniche si basano su trasformazioni lineari dei dati e sono efficaci nel catturare relazioni lineari tra le variabili. Tuttavia, possono rivelarsi limitate nel gestire dati complessi e non lineari, tipici del mondo reale.

Negli ultimi anni, sono state proposte numerose tecniche non lineari per la riduzione della dimensionalità. Queste tecniche, a differenza di quelle lineari, sono in grado di trattare dati complessi e non lineari, consentendo una rappresentazione più accurata e significativa dei dati. Ciò è particolarmente vantaggioso quando i dati del mondo reale formano una varietà altamente non lineare, dove le relazioni tra le variabili sono complesse e possono assumere forme curve o irregolari (Van Der Maaten, 2009).

Le tecniche di riduzione della dimensionalità non lineare operano in diversi modi, ma condividono l'obiettivo comune di trovare una varietà dei dati in modo non lineare (Ghojogh B. M., 2019d). Questo viene ottenuto mediante l'estrazione di caratteristiche o la visualizzazione dei dati in spazi a dimensionalità ridotta. Queste tecniche possono essere suddivise in tre categorie principali: metodi spettrali, metodi probabilistici e metodi basati su reti neurali (Ghojogh B. , Data Reduction Algorithms in Machine Learning and Data Science, 2021; Ghodsi, 2021).

I metodi spettrali per la riduzione della dimensionalità si basano sulla matrice del grafo e sulla geometria dei dati. Molte di queste tecniche si riducono a risolvere un problema di autovalori o autovalori generalizzati (Ghojogh B., 2019a). Una sottocategoria dei metodi spettrali è basata sulla matrice Laplaciana del grafo dei dati (Merris, 1994), e questa è stata una scoperta cruciale nel contesto dello *Spectral Clustering* (Weiss, 1999; Ng, Jordan, & Weiss, 2001).

Lo *Spectral Clustering* è un metodo di clustering che risolve il problema degli autovalori per identificare un sottospazio a bassa dimensionalità in cui i cluster dei dati sono più separati (Ghojogh B. , Data Reduction Algorithms in Machine Learning and Data Science, 2021). Questo approccio esegue una doppia funzione di riduzione della dimensionalità ed estrazione delle caratteristiche. L'idea principale è quella di trasformare i dati da uno spazio di input a un sottospazio a bassa dimensionalità in modo da rendere i cluster più evidenti. Successivamente, viene applicato un algoritmo di clustering per assegnare i dati ai cluster corrispondenti (Ghodsi, 2021).

I metodi di riduzione della dimensionalità, tra cui lo *Spectral Clustering*, hanno dimostrato di essere efficaci nella gestione di dati reali complessi e non lineari, contribuendo a migliorare la precisione delle analisi e delle previsioni (Ghojogh B. , Data Reduction Algorithms in Machine Learning and Data Science, 2021). Tuttavia, la scelta della tecnica di riduzione della dimensionalità più adatta dipende dalla natura specifica del dataset e dai requisiti dell'applicazione. Pertanto, la ricerca continua nel campo della riduzione della dimensionalità è essenziale per sviluppare nuove tecniche e approcci che affrontino sfide sempre più complesse nel trattamento dei dati reali.

3.2.1 Autovettori e autovalori di una matrice

Gli autovalori e gli autovettori rivestono un ruolo fondamentale nell'analisi del grafo e nel processo di *Spectral Clustering*. Ogni grafo possiede un insieme di numeri non negativi chiamati autovalori, che costituiscono una sorta di "carta d'identità" del grafo stesso, portando con sé informazioni rilevanti sulla sua struttura e caratteristiche uniche (Chung, 1997).

Gli autovalori del grafo rappresentano una misura quantitativa della sua variazione e delle interazioni tra i vertici. Ogni autovalore è associato a una corrispondente matrice Laplaciana del grafo e rappresenta una caratteristica intrinseca della sua struttura. La presenza di autovalori maggiori di zero indica l'esistenza di connessioni e relazioni tra i vertici, mentre gli autovalori nulli denotano la presenza di componenti connesse o sottografi isolati all'interno del grafo. Questa informazione consente di esplorare aspetti fondamentali, come la

presenza di sottostrutture, componenti connesse e altre proprietà geometriche del grafo.

D'altra parte, gli autovettori sono i vettori speciali associati agli autovalori e costituiscono una rappresentazione particolare del grafo. Essi rappresentano le associazioni tra i vertici del grafo e catturano la struttura relazionale dei dati. In particolare, gli autovettori associati agli autovalori più piccoli catturano le informazioni cruciali sulla struttura di connettività del grafo e vengono impiegati per il clustering dei dati. Gli autovettori permettono di esprimere i vertici in uno spazio a bassa dimensionalità, fornendo una proiezione significativa dei dati nel contesto del clustering.

Una corretta comprensione del concetto di matrice Laplaciana, autovalori e autovettori è essenziale per una solida base teorica e una corretta implementazione dell'algoritmo di *Spectral Clustering* proposto da Ng et al. (2001). Grazie all'analisi degli autovalori e degli autovettori, il processo di *Spectral Clustering* può estrarre informazioni rilevanti dai dati non etichettati e fornire una rappresentazione spettrale efficace per il clustering.

3.2.2 Definizione di autovettori e autovalori

Per ottenere una comprensione approfondita dell'importanza degli autovalori e degli autovettori nella matrice Laplaciana dei grafi, esaminiamo attentamente il loro ruolo fondamentale. Questi concetti costituiscono la chiave per rivelare in modo esaustivo le proprietà e la struttura sottostante contenute in questa matrice grafica (Chung, 1997). Iniziamo dunque con una definizione generale.

Gli autovalori e gli autovettori sono concetti fondamentali nell'analisi delle matrici quadrate. Nella matrice Laplaciana normalizzata e simmetrica $\boldsymbol{L_{sym}}$, questi concetti svolgono un ruolo di particolare rilevanza.

Gli autovalori di $\boldsymbol{L_{sym}}$ sono rappresentati da $\lambda_1, \lambda_2, \dots, \lambda_n$, dove *n* è l'ordine della matrice, ovvero il numero di vertici presente nel grafo. Tra questi n autovalori scegliamo di considerare solo gli m autovalori distinti. Di conseguenza, ci concentreremo sugli autovalori da $\lambda_1, \lambda_2, \dots, \lambda_m$ con $0 < m \leq n$, dove m è il numero di autovalori distinti. Ciascun autovalore λ_i è associato a un autovettore colonna u_i non nullo, che soddisfa l'equazione caratteristica (White, 1958):

$$\boldsymbol{L_{sym}}\lambda_i = u_i.$$

(3. 6)

In questa espressione, $\boldsymbol{L_{sym}}$ è la matrice Laplaciana normalizzata simmetrica, mentre u_i denota l'autovettore associato all'autovalore λ_i. Questa equazione rivela un concetto interessante: quando moltiplichiamo la matrice $\boldsymbol{L_{sym}}$ per il suo rispettivo autovettore u_i, il risultato ottenuto è una versione scalare dello stesso autovettore, rappresentato proprio da λ_i (Chung, 1997).

In altre parole, ogni autovalore λ_i rappresenta una scala, una costante, che governa come l'autovettore u_i interagisce con la matrice $\boldsymbol{L_{sym}}$. L'interazione tra autovalori e autovettori è di fondamentale importanza nell'analisi dei grafi, in quanto ci permette di esplorare le dinamiche e le caratteristiche intrinseche di questi grafi attraverso una prospettiva matematica (Chung, 1997).

Per calcolare gli autovalori di una matrice come $\boldsymbol{L_{sym}}$, è necessario ricercare gli zeri del suo cosiddetto "polinomio caratteristico". Per fare ciò, prendiamo la matrice Laplaciana normalizzata simmetrica $\boldsymbol{L_{sym}}$. Successivamente, sottraiamo da questa matrice il prodotto tra un parametro λ_i (che possiamo considerare come una sorta di "ipotetico" autovalore) e una matrice identità $\boldsymbol{I}$ con lo stesso ordine di λ_i. In termini matematici, questo processo è espresso dall'equazione (White, 1958):

$$det\,(\boldsymbol{L_{sym}} - \lambda_i \boldsymbol{I}) = 0.$$

(3. 7)

Se consideriamo λ_i come un'incognita, o se si preferisce come un'indeterminata, allora l'espressione $det\,(\boldsymbol{L_{sym}} - \lambda_i \boldsymbol{I})$ corrisponde a un polinomio di grado *n*, che viene chiamato per definizione "polinomio caratteristico associato alla matrice $\boldsymbol{L_{sym}}$".

La chiave per determinare gli autovalori λ_i è individuare gli zeri di questo polinomio caratteristico. Poiché il polinomio è di grado *n*, avremo esattamente *n* autovalori distinti.

Una volta ottenuti gli autovalori distinti associati alla matrice $\boldsymbol{L_{sym}}$, il passo successivo è calcolare gli autovettori associati a ciascun autovalore risolvendo, per ogni autovalore, un sistema di equazioni lineari noto come sistema lineare omogeneo (White, 1958):

$$\left(\boldsymbol{L_{sym}} - \lambda_i \boldsymbol{I}\right) u_i = 0,$$

(3. 8)

dove u_i rappresenta l'autovettore associato a λ_i. La soluzione a questo sistema lineare fornisce gli autovettori u_i relativi agli autovalori λ_i. Poiché il sistema (3.8) è un sistema omogeneo, esso avrebbe come unica soluzione $u_i = \mathbf{0}$, se non fosse per il fatto che è stato posta la condizione (3.7), il che assicura che invece il sistema abbia infinite soluzioni. Convenzionalmente tra gli infiniti vettori u_i che soddisfano la (3.8) si sceglie quello di norma unitaria, come si vedrà in seguito.

3.2.3 Scelta del numero di cluster e ottenimento della matrice U

Nel processo di implementazione dello *Spectral Clustering*, è essenziale calcolare i primi *k* autovalori, ossia $\lambda_1, \dots, \lambda_k$, insieme ai relativi autovettori $u_1, \dots, u_k$, derivati dalla matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$.

Una questione aperta di fondamentale importanza nello *Spectral Clustering* è la selezione di un numero appropriato di cluster. La scelta del numero *k* di cluster è un problema cruciale per tutti gli algoritmi di clustering, incluso lo *Spectral Clustering*. Esistono diversi metodi e approcci più o meno efficaci per affrontare questa problematica, a seconda del contesto e delle assunzioni sul modello sottostante.

Nei contesti di clustering basati su modelli, dove sono state fatte specifiche ipotesi sulle distribuzioni dei dati nei cluster, possono essere utilizzati criteri ben giustificati per determinare il numero ottimale di cluster. Questi criteri si basano generalmente sulla log-verosimiglianza dei dati e possono essere trattatati da un punto di vista frequentista o bayesiano. Ad esempio, il criterio di massima verosimiglianza o l'approccio Bayesiano possono essere utilizzati per stimare il numero di cluster ottimale (per esempi, si veda Fraley e Raftery (2002)).

Tuttavia, in contesti in cui le assunzioni sul modello sottostante sono limitate o assenti, è necessario ricorrere a metodi diversi per selezionare il numero di cluster. Esistono numerosi indici e misure ad hoc che possono essere utilizzati, il rapporto tra le similarità all'interno dei cluster e tra i cluster. Questi indici valutano la coesione all'interno di ciascun cluster e la separazione tra i cluster. Altri approcci di questo tipo includono anche criteri informativi, che si basano sulla quantità di informazione necessaria per descrivere la struttura dei dati con un determinato numero di cluster (Still S., 2004).

Altri metodi di selezione del numero di cluster includono la statistica dei gap (Tibshirani R., 2001), che confronta la similarità tra i dati nei cluster reali con quella ottenuta da cluster casuali, e approcci di stabilità (Ben-Hur A., 2002; Lange T., 2004; Ben-David S., 2006), che valutano la stabilità delle soluzioni di clustering attraverso variazioni nei dati o campioni.

Naturalmente, tutti questi metodi di selezione del numero di cluster utilizzati in algoritmi di clustering possono essere applicati anche allo *Spectral Clustering*. Tuttavia, per lo *Spectral Clustering* esiste uno strumento specifico progettato per affrontare questa problematica, noto come "euristica del gap degli autovalori", che può essere utilizzata con tutte e tre le matrici Laplaciane del grafo.

L'euristica del gap degli autovalori mira a selezionare il numero *k* in modo da ottenere una distinta separazione tra gli autovalori $\lambda_1, \dots, \lambda_k$, che dovrebbero essere molto piccoli, e l'autovalore successivo λ_{k+1}, il quale dovrebbe essere relativamente grande (von Luxburg, 2007). La misura della differenza tra questi valori, chiamata "gap", può quindi essere utilizzata come criterio per stabilire il numero appropriato di cluster *k*.

L'euristica del gap degli autovalori è uno strumento utile per determinare il numero ottimale di cluster quando si utilizza lo *Spectral Clustering*. Tuttavia, è importante notare che la sua efficacia dipende fortemente dalla struttura dei dati e dalla natura dei cluster presenti nel set di dati. In particolare, l'euristica funziona bene quando i cluster nei dati sono distinti e ben definiti, ossia quando i punti all'interno di ciascun cluster sono altamente simili tra loro e differiscono significativamente dai punti in altri cluster. In questa situazione, si osserva un chiaro divario tra gli autovalori corrispondenti ai primi *k* cluster e l'autovalore successivo, il che semplifica la scelta del numero di cluster (von Luxburg, 2007).

Tuttavia, quando i cluster sono meno ben definiti, più rumorosi o sovrapposti, l'euristica del gap degli autovalori può essere meno efficace e persino ingannevole. Nei set di dati in cui i cluster sono più "sfocati" la differenza tra gli autovalori potrebbe non essere così evidente, rendendo difficile individuare un punto di separazione chiaro tra i cluster. In queste situazioni, l'euristica potrebbe non fornire un numero di cluster appropriato e la scelta del *k* potrebbe risultare ambigua.

Se invece nel set di dati non è presente un divario ben definito tra gli autovalori, tutte le differenze tra tutti gli autovalori potrebbero risultare approssimativamente simili. Questo scenario può verificarsi quando i cluster si sovrappongono notevolmente tra loro, con punti che condividono similarità sia all'interno dello stesso cluster che con altri cluster. In queste circostanze, gli algoritmi di clustering, inclusa l'euristica del gap degli autovalori, potrebbero avere difficoltà a rilevare in modo accurato e coerente i cluster, a meno che non siano supportati da forti ipotesi sul modello sottostante (von Luxburg, 2007).

Dunque, sebbene l'euristica del gap degli autovalori sia uno strumento prezioso per la scelta del numero di cluster, è fondamentale considerare attentamente le

caratteristiche specifiche del set di dati e la struttura dei cluster presenti. In particolare, è importante valutare la chiarezza dei cluster, il livello di sovrapposizione tra i cluster e la presenza di rumore nei dati. Nel caso di cluster ben definiti, l'euristica del gap degli autovalori può fornire una stima affidabile di *k*. Tuttavia, nei casi più complessi, potrebbe essere necessario utilizzare altre tecniche di selezione del numero di cluster.

Una volta determinato il numero di cluster desiderato, passiamo ora alla costruzione della matrice $\boldsymbol{U}$. Questa matrice ha dimensioni $n \times k$, dove *n* rappresenta il numero di punti nel dataset e *k* è il numero di cluster scelto. La matrice $\boldsymbol{U}$ è composta dai primi *k* autovettori della matrice normalizzata simmetrica $\boldsymbol{L_{sym}}$ (Ghojogh B., 2019a). Ogni colonna di $\boldsymbol{U}$ rappresenta un autovettore associato a un determinato autovalore.

Un aspetto cruciale da notare è che le colonne di $\boldsymbol{U}$ sono organizzate in ordine crescente in base agli autovalori corrispondenti. In altre parole, le prime colonne di $\boldsymbol{U}$ contengono gli autovettori associati agli autovalori più piccoli, mentre le colonne successive rappresentano gli autovettori corrispondenti agli autovalori progressivamente più grandi (von Luxburg, 2007).

Questa matrice $\boldsymbol{U}$ cattura le informazioni di raggruppamento contenute negli autovettori. In pratica, otteniamo $\boldsymbol{U} \in \mathbb{R}^{n \times k}$, che rappresenta ciò che viene chiamato "*spectral embedding*" (Niu D., 2011). Questo *embedding* rappresenta una proiezione dei dati in un nuovo spazio delle caratteristiche completamente nuovo, in cui le informazioni di similarità tra i punti sono rappresentate in modo più rilevante, rendendo più agevole l'effettiva separazione e il clustering dei dati in base alle strutture intrinseche rivelate dagli autovalori e dagli autovettori.

3.2.4 Ottenimento della matrice T e algoritmo k-means

Poiché lo *spectral embedding* è continuo, per ottenere una suddivisione discreta dei dati nei *k* cluster desiderati, dobbiamo eseguire un passaggio di "arrotondamento". Un algoritmo di arrotondamento specifico, proposto da Ng et al. (2001), si basa sul ri-normalizzare ogni riga della matrice $\boldsymbol{U}$ al fine di ottenere una nuova matrice $\boldsymbol{T}$ di dimensione $n \times k$, in cui ogni riga ha norma 1 (lunghezza unitaria). Questo processo di normalizzazione viene eseguito come segue:

per ogni elemento t_{ij} in $\boldsymbol{T}$, impostiamo (von Luxburg, 2007)

$$t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{\frac{1}{2}}},$$

(3. 9)

per $i = 1, \dots, n$ e $j = 1, \dots, k$. Dove, u_{ij} è l'elemento corrispondente nella matrice $\boldsymbol{U}$.

La matrice $\boldsymbol{T}$ risultante avrà quindi le stesse dimensioni di $\boldsymbol{U}$, ma le sue righe saranno normalizzate in modo da essere lunghezza unitaria. In altre parole, $\boldsymbol{T}$ rappresenta lo *spectral embedding* discreto, dove le informazioni di similarità tra i punti sono rappresentate in un formato più adatto per la suddivisione dei dati in *k* cluster (Ng, Jordan, & Weiss, 2001).

Dopo aver trovato la matrice $\boldsymbol{T} \in \mathbb{R}^{n \times k}$, trattiamo le righe di $\boldsymbol{T}$ come nuovi punti in uno spazio *k*-dimensionale. Quindi, otteniamo un insieme di vettori $(y_i)_{i=1,\dots,n}$, dove y_i rappresentano i punti del dataset trasformati nello spazio dei *k* autovettori normalizzati. Questo nuovo spazio di incorporamento, chiamato spazio *Y*, è stato progettato per rappresentare le informazioni di similarità tra i punti in un formato più adatto per la suddivisione dei dati in *k* cluster.

Successivamente, in accordo con Ng et al., applichiamo l'algoritmo di clustering *k-means* sulle righe della matrice $\boldsymbol{T}$ normalizzata per suddividere gli *n* punti in *k* cluster distinti (Ng, Jordan, & Weiss, 2001). Ogni punto y_i sarà assegnato al cluster a cui è assegnata la riga t_i della matrice $\boldsymbol{T}$. Come risultato, si ottengono *k* cluster, indicati come $C_1, \dots, C_k$.

L'algoritmo proposto da Ng et al. (2001) fornisce come output i cluster $A_1, \dots, A_k$, dove ogni A_i contiene gli indici dei punti che appartengono al cluster C_i. In altre parole, $A_i = \{j \mid y_j \in C_i\}$ (von Luxburg, 2007). Questo ci fornisce una suddivisione discreta e ben definita dei dati in base alle loro caratteristiche e somiglianze, consentendoci di identificare gruppi di punti con connessioni forti all'interno dei cluster e connessioni più deboli tra i cluster.

Un aspetto interessante dello *Spectral Clustering* è che, a differenza dei metodi di clustering tradizionali, come il *k-means*, che operano direttamente nello spazio di input (spazio *X*), viene applicato nello spazio di incorporamento (spazio di *Y*). In altre parole, lo *Spectral Clustering* estrae prima le caratteristiche per una migliore discriminazione dei cluster e poi applica il clustering. Questo di solito porta a una migliore suddivisione dei cluster perché le caratteristiche estratte dovrebbero discriminare meglio i cluster dei dati. Questa caratteristica rende lo *Spectral Clustering* un approccio potente e flessibile per la scoperta di strutture nascoste all'interno dei dati (Ghojogh, Ghodsi, Karray, & Crowley, Laplacian-based dimensionality reduction including spectral clustering, Laplacian eigenmap, locality preserving projection, graph embedding, and diffusion map: Tutorial and survey, 2021).

Complessivamente, lo *Spectral Clustering* dimostra dunque di essere un potente approccio per il clustering dei dati, integrando efficacemente i vantaggi della rappresentazione spettrale con l'uso del metodo *k-means*. La capacità di codificare informazioni sulla similarità tra i punti nei cluster rende lo *Spectral Clustering* una tecnica di grande interesse e ampiamente utilizzata in diverse applicazioni di clustering e analisi dei dati.

4. Applicazione dello Spectral Clustering in un'analisi sulla Visitor Experience presso la Pinacoteca Tosio Martinengo

Senza dubbio, il capitolo che stiamo per affrontare è il culmine di un percorso che ci ha condotto attraverso i concetti teorici e le fondamenta dell'analisi dei dati. Qui, nell'ultimo capitolo di questa tesi, ci immergeremo completamente nel cuore pulsante della ricerca e dell'analisi, con l'obiettivo di applicare concretamente l'innovativo algoritmo di *Spectral Clustering*. Un algoritmo che, come avete appreso nei capitoli precedenti, si rivela straordinariamente potente nell'analisi dei dati complessi.

Il nostro campo di applicazione è affascinante: un dettagliato studio basato su dati concreti, raccolti durante una ricerca di mercato inerente a un'analisi di marketing sensoriale.[21] Questa indagine è stata condotta presso la Pinacoteca Tosio Martinengo di Brescia, un luogo che incarna una tradizione storica di inestimabile valore e rappresenta una delle gemme culturali della città. La ricerca si è svolta nel periodo compreso tra giugno e settembre del 2019, consentendoci di esplorare le intricate connessioni tra esperienze artistiche, emozioni e contesto culturale.

L'obiettivo principale di questo capitolo è chiaro: condurre un'analisi approfondita dei dati raccolti, concentrandoci in particolare sull'esperienza dei visitatori all'interno della Pinacoteca. Identificheremo quindi dei gruppi di visitatori con caratteristiche simili, fornendo un approfondito insight che si rivelerà di inestimabile valore per l'ottimizzazione della *visiting experience*, ovvero l'esperienza di coinvolgimento che mantiene alto il livello di attenzione, e delle strategie di marketing del museo stesso.

Durante la fase di ricerca precedente, Giulia Zanoletti, la tesista responsabile del progetto, ha condotto un approfondito sforzo per raccogliere dati preziosi, orientati principalmente verso una domanda chiave: "Quali sensazioni ed emozioni provano i visitatori durante la loro visita alla Pinacoteca Tosio Martinengo, e in che modo possiamo utilizzare queste informazioni per sviluppare nuove idee innovative in grado di soddisfare i bisogni latenti impliciti del pubblico?"

Sfruttando i dati preziosi raccolti da Giulia Zanoletti nella fase di ricerca precedente, impiegheremo l'algoritmo di *Spectral Clustering*. Questo algoritmo, basandosi sui dati raccolti, sarà il nostro strumento per rivelare connessioni

[21] Zanoletti G., *Analisi Sensoriale dell'esperienza di visita di un museo: il caso della Pinacoteca di Brescia*, Università degli Studi di Brescia, Dipartimento di Economia e Management, Anno Accademico 2018-2019.

nascoste tra i visitatori. Utilizzeremo il linguaggio di programmazione R per condurre questa analisi. Il nostro intento è quello di creare cluster di visitatori con esperienze sensoriali simili, identificando tendenze e comportamenti comuni tra i diversi gruppi. Questa suddivisione in cluster non solo agevolerà la comprensione delle preferenze e dei modelli comportamentali dei visitatori, ma aprirà anche la strada a una serie di opportunità strategiche.

La nostra ambizione in questo capitolo è duplice: innanzitutto, dimostrare la trasformazione della teoria in pratica attraverso l'applicazione di concetti precedentemente discussi nei capitoli teorici della tesi. In secondo luogo, vogliamo offrire un'illuminante finestra sulla potenza dell'analisi dei dati, evidenziando il suo valore inestimabile non solo per le istituzioni culturali, ma anche per il mondo della ricerca di mercato.

Attraverso l'applicazione dell'algoritmo di *Spectral Clustering*, miriamo a rivelare e definire in modo unico i modelli di comportamento dei visitatori, le loro preferenze e le sfumature delle loro esperienze, svelando così segreti nascosti all'interno dei dati raccolti.

Quest'applicazione pratica dimostrerà in modo tangibile che la teoria è in grado di trasformarsi in una realtà potente nell'analisi dei dati. Ci consentirà di aprire nuove prospettive nella comprensione della *visitor experience* e offrirà agli stakeholder culturali uno strumento fondamentale per ottimizzare l'esperienza dei visitatori e promuovere un coinvolgimento più profondo con l'arte e la cultura.

Nel corso di questo capitolo, esploreremo il processo dettagliato di applicazione dell'algoritmo di *Spectral Clustering* ai dati reali che sono stati accuratamente raccolti. Sarà un viaggio che ci porterà attraverso le fasi cruciali del processo, ognuna delle quali riveste un ruolo fondamentale nel determinare il successo dell'analisi.

Cominceremo con la preparazione dei dati, un passaggio essenziale per assicuraci che i dati siano pronti per l'applicazione dell'algoritmo. Questa fase include la standardizzazione delle variabili e la creazione della matrice di similarità, un elemento chiave per *Spectral Clustering*. La scelta accurata di come misurare la similarità tra le osservazioni gioca un ruolo cruciale nella riuscita del clustering.

Dopo la preparazione dei dati, affronteremo la delicata questione della scelta dei parametri dell'algoritmo. Questa fase richiede attenzione ai dettagli e l'utilizzo di tecniche di ottimizzazione per determinare il numero ottimale di cluster (k) e altri parametri rilevanti. La scelta accurata di questi parametri influenzerà direttamente la qualità della suddivisione in cluster.

Una volta effettuato il clustering, ci concentreremo sull'interpretazione dei risultati ottenuti. Analizzeremo in dettaglio i cluster risultanti, identificando le caratteristiche distintive di ciascun gruppo di visitatori. Questa fase ci permetterà di trarre conclusioni significative sull'esperienza dei visitatori e di comprendere meglio le preferenze e i modelli comportamentali all'interno del museo.

Tutto ciò che scopriremo attraverso questa analisi non sarà solo di grande interesse accademico, ma avrà anche un impatto concreto. Potremo suggerire raccomandazioni specifiche per migliorare la *visitor experience* e sviluppare strategie di marketing mirate che rispondano alle esigenze di ciascun cluster di visitatori. Questo rappresenta un passo avanti significativo per il museo, in quanto potrà offrire un'esperienza più personalizzata e coinvolgente per i visitatori.

Questo capitolo rappresenta dunque una tappa fondamentale nel nostro percorso, dove la teoria e la pratica si incontrano per produrre risultati tangibili e preziosi. Siamo pronti ad affrontare le sfide e le opportunità che ci attendono, guidati dalla promettente prospettiva dello *Spectral Clustering* e dalla passione per la ricerca.

4.1 Descrizione dei dati utilizzati

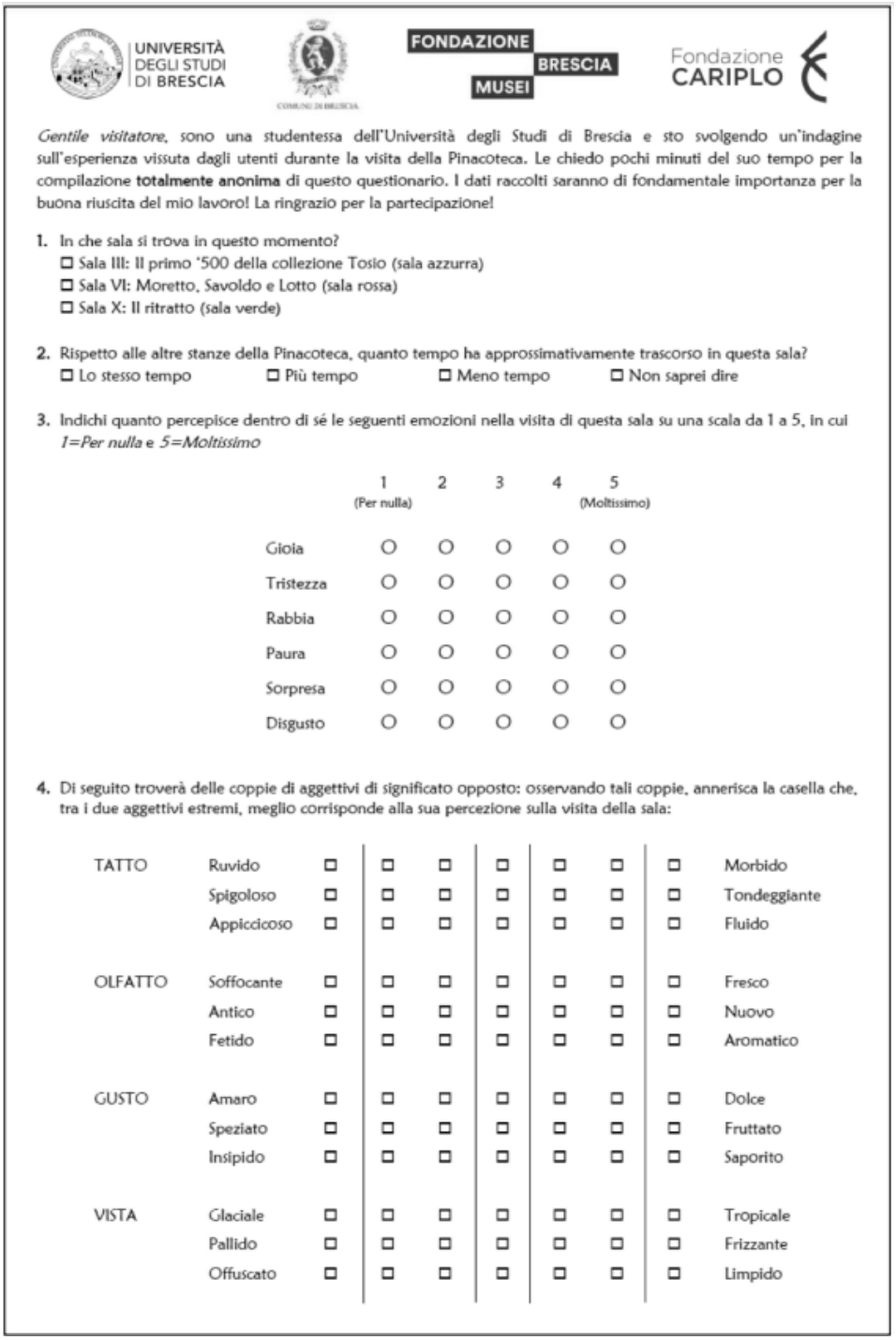
Per comprendere appieno l'ambito e il contesto della nostra analisi basata sullo *Spectral Clustering*, è essenziale immergersi nel mondo dei dati che costituiscono la base della nostra ricerca. In questa sezione, forniremo una dettagliata descrizione dei dati che abbiamo utilizzato, esaminando come sono stati raccolti, quali variabili sono state registrate e, soprattutto, focalizzandoci sulla domanda centrale che ha dato origine alle preziose variabili qualitative ordinali, oggetto del nostro studio.

Come precedentemente menzionato, i dati a nostra disposizione sono il risultato di un meticoloso processo di raccolta condotto all'interno della Pinacoteca Tosio Martinengo di Brescia. Questa raccolta di dati è stata effettuata come parte di un progetto di analisi di marketing sensoriale, con l'obiettivo di indagare le esperienze sensoriale dei visitatori all'interno del museo, in particolare durante la loro permanenza in specifiche sale espositive.

Per avere una visione più completa dei dati raccolti durante questa ricerca di mercato presso la Pinacoteca Tosio Martinengo, è importante considerare il contesto in cui sono stati acquisiti. La Pinacoteca, una cornice ricca di opere d'arte e cultura, è un luogo in cui l'estetica e l'emozione convergono. Questo contesto particolare rende i dati raccolti ancora più preziosi, poiché consentono di esplorare le reazioni dei visitatori all'arte in un ambiente culturale unico.

Per condurre questa indagine, sono state selezionate tre stanze specifiche all'interno della Pinacoteca, ognuna delle quali è stata progettata per suscitare sensazioni uniche nei visitatori. Queste stanze sono state identificate come la Sala III, con opere del primo '500 della collezione Tosio (sala azzurra), la Sala VI, dedicata a Moretto, Savoldo e Lotto (sala rossa), e infine la Sala X, incentrata sul tema del ritratto (sala verde). La scelta di queste stanze è stata motivata dalla loro decorazione con colori accesi e brillanti, sia nelle tappezzerie che nei singoli quadri esposti. Questi colori intensi erano progettati per evocare una gamma ampia di emozioni, positive e negative. È importante sottolineare che, per evitare distorsioni in sede di analisi dei dati, le sale selezionate presentano le stesse dimensioni architettoniche, garantendo così una base uniforme per il confronto tra i risultati ottenuti in diverse stanze.

I dati sono stati raccolti mediane un questionario attentamente progettato per indagare l'esperienza sensoriale dei visitatori. Esso, come si evince dalle Figure 4.1 e 4.2, è composto da 11 domande, oltre a una sezione relativa ai dati anagrafici. Le domande sono suddivise in tre macro-aree. La prima riguarda l'aspetto sensoriale dell'esperienza e include sia domande sulla percezione generica della visita sia una serie di domande che si addentrano nel nucleo centrale dell'indagine concentrandosi specificamente sulle emozioni e sulle sensazioni avvertite nella sala espositiva visitata, richiedendo pertanto un maggiore approfondimento da parte dei partecipanti. La seconda area riguarda il turismo presso la Pinacoteca e coinvolge una serie di domande volte a valutare le visite turistiche nella struttura museale. Per concludere, la terza area del questionario esamina i profili sociodemografici dei rispondenti, analizzando dati quali genere, età, livello di istruzione, luogo di residenza e occupazione.



UNIVERSITÀ DEGLI STUDI DI BRESCIA

FONDAZIONE BRESCIA MUSEI

Fondazione CARIPLO

Gentile visitatore, sono una studentessa dell'Università degli Studi di Brescia e sto svolgendo un'indagine sull'esperienza vissuta dagli utenti durante la visita della Pinacoteca. Le chiedo pochi minuti del suo tempo per la compilazione **totalmente anonima** di questo questionario. I dati raccolti saranno di fondamentale importanza per la buona riuscita del mio lavoro! La ringrazio per la partecipazione!

1. In che sala si trova in questo momento?
   - ☐ Sala III: Il primo '500 della collezione Tosio (sala azzurra)
   - ☐ Sala VI: Moretto, Savoldo e Lotto (sala rossa)
   - ☐ Sala X: Il ritratto (sala verde)

2. Rispetto alle altre stanze della Pinacoteca, quanto tempo ha approssimativamente trascorso in questa sala?
   ☐ Lo stesso tempo ☐ Più tempo ☐ Meno tempo ☐ Non saprei dire

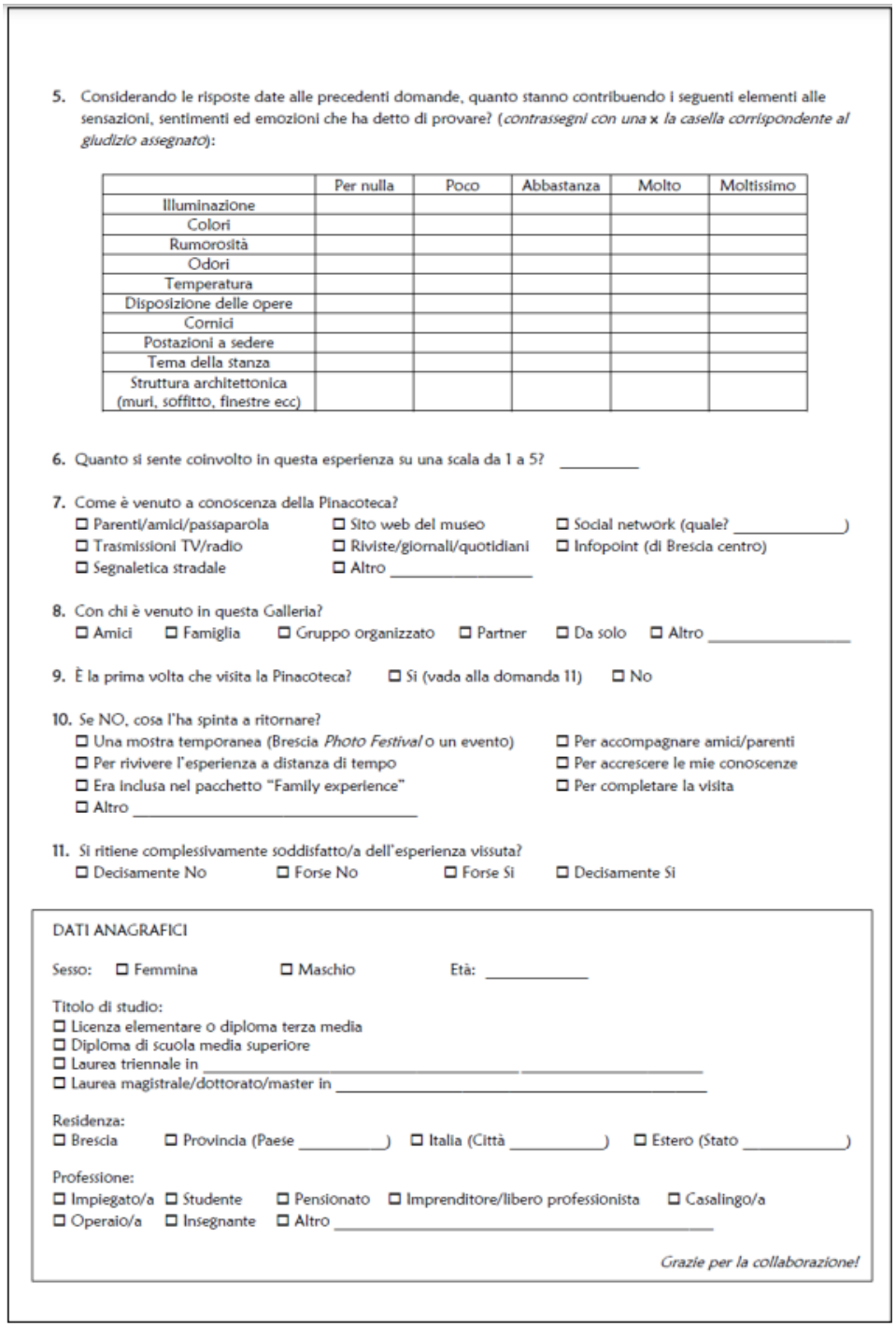
3. Indichi quanto percepisce dentro di sé le seguenti emozioni nella visita di questa sala su una scala da 1 a 5, in cui *1=Per nulla* e *5=Moltissimo*

	1 (Per nulla)	2	3	4	5 (Moltissimo)
Gioia	O	O	O	O	O
Tristezza	O	O	O	O	O
Rabbia	O	O	O	O	O
Paura	O	O	O	O	O
Sorpresa	O	O	O	O	O
Disgusto	O	O	O	O	O

4. Di seguito troverà delle coppie di aggettivi di significato opposto: osservando tali coppie, annerisca la casella che, tra i due aggettivi estremi, meglio corrisponde alla sua percezione sulla visita della sala:

TATTO	Ruvido	☐	☐	☐	☐	☐	☐	☐	Morbido
	Spigoloso	☐	☐	☐	☐	☐	☐	☐	Tondeggiante
	Appiccicoso	☐	☐	☐	☐	☐	☐	☐	Fluido
OLFATTO	Soffocante	☐	☐	☐	☐	☐	☐	☐	Fresco
	Antico	☐	☐	☐	☐	☐	☐	☐	Nuovo
	Fetido	☐	☐	☐	☐	☐	☐	☐	Aromatico
GUSTO	Amaro	☐	☐	☐	☐	☐	☐	☐	Dolce
	Speziato	☐	☐	☐	☐	☐	☐	☐	Fruttato
	Insipido	☐	☐	☐	☐	☐	☐	☐	Saporito
VISTA	Glaciale	☐	☐	☐	☐	☐	☐	☐	Tropicale
	Pallido	☐	☐	☐	☐	☐	☐	☐	Frizzante
	Offuscato	☐	☐	☐	☐	☐	☐	☐	Limpido

Figura 4.1 – Prima pagina del questionario della Pinacoteca sensoriale. Fonte: Zanoletti G., Analisi Sensoriale dell'esperienza di visita di un museo: il caso della Pinacoteca di Brescia, 2019.



5. Considerando le risposte date alle precedenti domande, quanto stanno contribuendo i seguenti elementi alle sensazioni, sentimenti ed emozioni che ha detto di provare? (*contrassegni con una* x *la casella corrispondente al giudizio assegnato*):

	Per nulla	Poco	Abbastanza	Molto	Moltissimo
Illuminazione					
Colori					
Rumorosità					
Odori					
Temperatura					
Disposizione delle opere					
Cornici					
Postazioni a sedere					
Tema della stanza					
Struttura architettonica (muri, soffitto, finestre ecc)					

6. Quanto si sente coinvolto in questa esperienza su una scala da 1 a 5? _____

7. Come è venuto a conoscenza della Pinacoteca?
☐ Parenti/amici/passaparola ☐ Sito web del museo ☐ Social network (quale? _____)
☐ Trasmissioni TV/radio ☐ Riviste/giornali/quotidiani ☐ Infopoint (di Brescia centro)
☐ Segnaletica stradale ☐ Altro _____

8. Con chi è venuto in questa Galleria?
☐ Amici ☐ Famiglia ☐ Gruppo organizzato ☐ Partner ☐ Da solo ☐ Altro _____

9. È la prima volta che visita la Pinacoteca? ☐ Si (vada alla domanda 11) ☐ No

10. Se NO, cosa l'ha spinta a ritornare?
☐ Una mostra temporanea (Brescia *Photo Festival* o un evento) ☐ Per accompagnare amici/parenti
☐ Per rivivere l'esperienza a distanza di tempo ☐ Per accrescere le mie conoscenze
☐ Era inclusa nel pacchetto "Family experience" ☐ Per completare la visita
☐ Altro _____

11. Si ritiene complessivamente soddisfatto/a dell'esperienza vissuta?
☐ Decisamente No ☐ Forse No ☐ Forse Si ☐ Decisamente Si

DATI ANAGRAFICI

Sesso: ☐ Femmina ☐ Maschio Età: _____

Titolo di studio:
☐ Licenza elementare o diploma terza media
☐ Diploma di scuola media superiore
☐ Laurea triennale in _____
☐ Laurea magistrale/dottorato/master in _____

Residenza:
☐ Brescia ☐ Provincia (Paese _____) ☐ Italia (Città _____) ☐ Estero (Stato _____)

Professione:
☐ Impiegato/a ☐ Studente ☐ Pensionato ☐ Imprenditore/libero professionista ☐ Casalingo/a
☐ Operaio/a ☐ Insegnante ☐ Altro _____

Grazie per la collaborazione!

Figura 4.2 – Seconda pagina del questionario della Pinacoteca sensoriale. Fonte: Zanoletti G., Analisi Sensoriale dell'esperienza di visita di un museo: il caso della Pinacoteca di Brescia, 2019.

La domanda 3, domanda chiave sulla quale ci concentreremo per l'applicazione dell'algoritmo di *Spectral Clustering*, riveste un ruolo di primaria importanza

nell'analisi di marketing sensoriale condotta presso la Pinacoteca Tosio Martinengo. Essa chiede ai partecipanti di indicare quanto percepiscono dentro di sé sei emozioni specifiche su una scala da 1 a 5, dove "1" indica una mancata percezione dell'emozione e "5" rappresenta una percezione totale dell'emozione, ed è formulata in modo da esplorare le emozioni percepite dai visitatori durante la visita di una sala espositiva specifica all'interno del museo. Questo tipo di scala, nota come scala di Likert, è comunemente utilizzata per misurare il grado di accordo o disaccordo con una specifica affermazione.

La scelta di includere un intervallo di valori da "*Per nulla*" a "*Moltissimo*" permette ai partecipanti di esprimere con precisione la loro percezione delle emozioni. Questa scala graduata cattura le sfumature delle risposte emotive, consentendo una maggiore profondità nell'analisi. È un'indagine che va direttamente al cuore dell'esperienza dei visitatori, cercando di catturare una diversificata gamma di reazioni e risposte emotive.

La valutazione delle emozioni riveste un ruolo di fondamentale importanza nella comprensione del comportamento e delle reazioni dei visitatori. Le emozioni giocano un ruolo cruciale nel sistema emotivo, psicologico e comportamentale degli individui, esercitando un'influenza significativa sulle loro scelte, le loro opinioni e le loro esperienze. Di conseguenza, la raccolta di dati sulla percezione delle emozioni da parte dei visitatori offre un'opportunità straordinaria per comprendere in che modo queste emozioni plasmino e guidino la loro interazione con il museo.

Le emozioni considerate in questa domanda sono sei e sono rispettivamente gioia, tristezza, rabbia, paura, sorpresa e disgusto. Ciascuna di queste emozioni rappresenta una sfaccettatura diversa delle risposte emotive umane e cattura una gamma di reazioni che possono emergere di fronte all'arte. Queste emozioni sono state scelte con cura per offrire una panoramica completa delle risposte emotive dei visitatori. Tali emozioni, conosciute come le "*Big Six*" e identificate da Charles Darwin (Darwin, 1872) nelle percezioni umane, sono universalmente riconoscibili e sono coerenti con l'obiettivo di comprendere le risposte emotive dei visitatori.

Le risposte a questa domanda generano variabili qualitative ordinali, ciascuna delle quali rappresenta il grado di percezione di una specifica emozione. Ogni risposta è codificata con un numero corrispondente al grado di percezione dell'emozione su una scala da 1 a 5. Questi dati saranno fondamentali per la nostra analisi, poiché ci aiuteranno a identificare i cluster di visitatori con esperienze emotive simili e a migliorare ulteriormente l'esperienza dei visitatori.

Inoltre, è importante sottolineare che la natura ordinale delle risposte qualitative fornisce un ulteriore livello di dettaglio nell'analisi. Le risposte rappresentano gradazioni di intensità emotiva, consentendo una comprensione più approfondita delle sfumature nelle reazioni dei visitatori.

Per illustrare ulteriormente il contesto, possiamo considerare un visitatore che si trova di fronte a un'opera d'arte in una specifica sala. Questo visitatore potrebbe rispondere "5" alla domanda sulla sorpresa, poiché l'opera lo ha completamente stupefatto, ma allo stesso tempo potrebbe rispondere "4" alla domanda sulla rabbia, perché l'opera ha evocato sentimenti di indignazione o protesta. Questi dettagli mostrano quanto sia preziosa la raccolta di dati per esplorare le reazioni individuali e collettive dei visitatori alle opere d'arte.

I dati raccolti rappresentano dunque un tesoro di informazioni che ci permetteranno di esplorare le emozioni predominanti dei visitatori nella specifica sala espositiva, gettando luce sui loro sentimenti e sensazioni durante l'esperienza di visita da una prospettiva emozionale e culturale. Questi dati possono essere utilizzati per adattare l'offerta museale, migliorare la progettazione delle mostre e sviluppare strategie di marketing mirate per migliorare ulteriormente l'esperienza dei visitatori e attirarne di nuovi.

Nel processo di preparazione dell'indagine, una fase critica è la definizione della popolazione di interesse e la selezione di un campione su cui condurre il questionario. Quando parliamo di "popolazione" in questo contesto, ci riferiamo all'insieme di tutte le unità statistiche che sono soggette all'indagine, ovvero i visitatori della Pinacoteca Tosio Martinengo. Tuttavia, dato il notevole numero totale di visitatori, è stato adottato un approccio di campionamento. In altre parole anziché coinvolgere l'intera popolazione, è stato estratto un campione rappresentativo, composto da un numero gestibile di partecipanti (n). Questo campione selezionato è stato sottoposto al questionario e ha fornito le informazioni necessarie per condurre l'analisi senza la necessità di coinvolgere ogni singolo visitatore. Tale approccio è stato dunque fondamentale per condurre un'indagine accurata senza dover gestire la complessità e i costi associati al coinvolgimento dell'intera popolazione di visitatori della Pinacoteca Tosio Martinengo.

La somministrazione del questionario è avvenuta direttamente presso la Pinacoteca nei mesi estivi del 2019, precisamente dal 21 giugno al 15 settembre. Il metodo di raccolta dati è stato implementato sia attraverso interviste dirette, in cui un intervistatore ha guidato i partecipanti nella compilazione dei questionari, sia attraverso l'auto-compilazione dei questionari cartacei nelle lingue disponibili,

tra cui italiano, inglese, tedesco, francese e spagnolo. Al termine dell'indagine, sono stati raccolti un totale di 1.024 questionari sensoriali.

Questa descrizione approfondita dei dati ci ha dunque permesso di gettare le basi per l'analisi successiva, mettendo in luce l'origine e la natura dei dati che abbiamo a disposizione. Ora siamo pronti per procedere alla prossima fase dell'analisi, che coinvolge la preparazione dei dati per l'applicazione dell'algoritmo di *Spectral Clustering*. In questo modo, il nostro lavoro avrà inizio, esplorando il profondo legame tra l'arte e l'emozione, e gettando luce sulla complessità delle reazioni dei visitatori di fronte alle opere d'arte all'interno della Pinacoteca Tosio Martinengo.

4.2 Preparazione dei dati e creazione della matrice di similarità

I dati a nostra disposizione rappresentano una preziosa fonte di informazioni che ci consente di comprendere meglio come i visitatori interagiscono con il museo e come le loro percezioni influenzino il loro coinvolgimento.

Prima di procedere con l'applicazione dell'algoritmo di *Spectral Clustering* ai dati raccolti, è fondamentale sottolineare l'importanza di una fase preliminare cruciale: la preparazione dei dati. Questa fase è essenziale per garantire che i dati siano adeguatamente strutturati e pronti per l'analisi così da essere elaborati in modo efficace e accurato.

Una volta completata la fase di raccolta dati, è stata necessaria una preparazione accurata per l'elaborazione successiva attraverso tecniche di analisi statistica. Questo processo includeva la codifica delle risposte, assegnando un codice numerico a ciascuna risposta per consentire l'interpretazione, la creazione di un dataset in formato Excel per la gestione dei dati, il controllo delle risposte per individuare eventuali dati mancanti o incoerenti e, infine, la decisione su come trattare le risposte anomale. È importante sottolineare che non è stato necessario eliminare alcun questionario, e tutti i dati raccolti sono stati inclusi nell'analisi.

I dati, una volta codificati e attentamente controllati, sono stati inseriti in una matrice dei dati dedicata. Ogni riga di questa matrice rappresenta un partecipante all'indagine, mentre ogni colonna rappresenta una variabile statistica, ovvero una delle sei emozioni (gioia, tristezza, rabbia, paura, sorpresa e disgusto) oggetto dell'analisi.

È da questa matrice che inizieremo il nostro lavoro di *Cluster Analysis*. Come precedentemente menzionato, concentreremo la nostra attenzione esclusivamente sulla terza domanda del questionario, poiché questa rappresenta la base su cui

applicheremo l'algoritmo di *Spectral Clustering*, algoritmo volto a esplorare le connessioni nascoste tra le risposte emotive dei visitatori e identificare cluster omogenei di visitatori con esperienze emotive simili.

Prima di entrare nei dettagli dell'analisi dei dati con l'algoritmo di *Spectral Clustering*, è fondamentale garantire che l'ambiente di lavoro sia correttamente configurato e che i dati siano pronti per l'elaborazione. Questo processo richiede la pulizia dell'ambiente di lavoro per garantire risultati accurati e affidabili e l'impostazione di alcuni parametri chiave. Di seguito, illustreremo le fasi iniziali di questa preparazione dei dati.

Per mantenere l'ambiente di lavoro ordinato e garantire che non ci siano interferenze o dati inutili, eseguiamo una pulizia iniziale.

```
######## Pulizia

rm(list=ls())  # Se c'è qualcosa nel workplace rimuovi tutto
graphics.off() # Se ci sono grafici aperti pulisci tutto
```

Successivamente, impostando nel nostro script i parametri necessari per l'analisi. Questi parametri includono:

```
######## Set parameters

dir <- "…"        # settare la directory di lavoro

miss <- "."       # etichetta assegnata ai missing values

colstart <- 2     # colonna in cui si trova la prima variabile su
cui effettuare l'analisi
colend <- 7       # colonna in cui si trova l'ultima variabile su
cui effettuare l'analisi
colnames <- 0     # colonna in cui si trovano i nomi dei soggetti
- se non disponibile scrivere 0

datafile <- "dataset.pinacoteca.spectral.clustering.txt"   # nome
file che contiene i dati
```

Questi parametri sono fondamentali per identificare la posizione dei dati e garantire che vengano trattati correttamente.

Con l'ambiente di lavoro pulito e i parametri configurati, siamo pronti per procedere con l'elaborazione dei dati e l'analisi mediante lo *Spectral Clustering*.

Nella fase successiva, è fondamentale assicurarsi che R sia configurato per operare nella directory corretta, affinché possa accedere ai file di dati situati nella

directory specificata (*dir*). Questo obiettivo viene raggiunto tramite il seguente comando:

```
######## Impostazione della directory di lavoro

setwd(dir)
```

Successivamente, carichiamo i dati dal file specificato utilizzando il comando "*read.table()*". Questo ci consente di lavorare con il set di dati specificato nella nostra analisi.

```
######## Caricamento dati

dataset <- read.table(datafile, na.strings=miss, header=TRUE)
```

Prima di procedere con ulteriori operazioni, è utile verificare le dimensioni del dataset. Questa verifica ci consente di comprendere meglio la struttura dei dati e garantire che non ci siano problemi nell'importazione.

```
######## Stampa dimensioni dataset

dimdati <- dim(dataset)
print(dimdati)
```

Infine, definiamo una funzione, "*varn()*", che ci sarà utile durante l'analisi dei dati. Questa funzione calcola la varianza non corretta delle variabili e sarà utilizzata successivamente nel nostro script.

```
######## Funzione per varianza non corretta

varn <- function (x) {
  x <- na.omit(x)
  varianza <- sum((x-mean(x))^2)/(length(x))
  return(varianza)
}
```

Con questi passaggi preliminari completati, possiamo ora procedere con la selezione delle variabili e l'analisi dei dati attraverso l'algoritmo di *Spectral Clustering*.

Per preparare i dati per l'analisi con *Spectral Clustering*, effettuiamo dunque una selezione mirata delle variabili. In questo contesto, creiamo un nuovo data frame in R, denominato ***X***, contenente solo le risposte alla terza domanda del questionario. Questo data frame è stato generato nel seguente modo:

```
######## Creazione data frame per lo Spectral Clustering

X <- dataset[,colstart:colend] # Estrazione dati su cui
```

```
effettuare lo Spectral Clustering

X <- data.frame(X)
n <- nrow(X)  # Numero totale dei partecipanti (righe matrice)
p <- ncol(X)  # Numero di variabili statistiche (colonne matrice)
vars <- names(X) # Nomi delle variabili nel data frame
```

Inizialmente, abbiamo estratto dal nostro dataset originale una porzione di dati rilevanti per l'analisi con lo *Spectral Clustering*. Questa operazione è stata eseguita utilizzando i parametri "*colstart*" e "*colend*" per specificare l'intervallo delle colonne di interesse nel dataset completo. I dati estratti sono stati quindi organizzati in un nuovo data frame denominato ***X***, il quale costituirà il nostro set di dati di partenza per l'analisi successiva.

Questo data frame ***X*** contiene unicamente le risposte alla terza domanda del nostro questionario, che rappresentano le emozioni percepite dai visitatori durante la loro esperienza nelle sale espositive del museo, ed è composto da un numero totale di rispondenti (righe) pari a "*n*" e da un numero di variabili statistiche (colonne) pari a "*p*". Ogni variabile nel data frame è identificata dai rispettivi nomi, che sono memorizzati nella variabile "*vars*".[22]

Nel percorso di preparazione dei dati per l'analisi di clustering, uno dei passaggi critici è la standardizzazione delle variabili (Kaufman & Rousseeuw, 1990). Questa fase è essenziale prima di applicare qualsiasi algoritmo di clustering. La standardizzazione è una pratica comune che mira a garantire che tutte le variabili abbiano un'influenza equa sui risultati del clustering, indipendentemente dalla loro scala di misurazione originale. La standardizzazione rappresenta una tappa cruciale nel processo di preparazione dei dati in vista dell'analisi tramite lo *Spectral Clustering*, un algoritmo notoriamente sensibile alle differenze di scala tra le variabili.

Il processo di standardizzazione è particolarmente utile nel processo di preparazione dei dati in quanto il clustering si basa sulla misurazione della distanza tra le osservazioni. Se le variabili non vengono standardizzane, quelle con scale di misurazione più ampie potrebbero avere un peso eccessivo nella determinazione delle distanze, distorcendo i risultati dell'analisi (Kaufman & Rousseeuw, 1990).

In pratica, la standardizzazione delle variabili assicura che tutte le variabili abbiano la stessa scala di misura, evitando che una variabile influenzi in modo sproporzionato il risultato ottenuto tramite l'applicazione dello *Spectral*

[22] Abbiamo estratto i nomi delle variabili presenti nel data frame e li abbiamo memorizzati in una lista denominata "*vars*".

Clustering a causa di differenze nelle unità di misura. Questo approccio consente di condurre un'analisi più accurata dei dati, identificando strutture e relazioni nascoste tra le unità analizzate.

Inizialmente, procediamo dunque con la standardizzazione delle variabili considerate nell'analisi mediante il software R:

```
######## Standardizzazione delle variabili

m <- colMeans(X)
s <- sqrt(apply(X,2,varn))
Xscale <- (X)  # Numero totale dei partecipanti (righe matrice)
for(i in 1:ncol(X)){
  Xscale[,i] <- (X[,i]-m[i])/s[i]
}
X <- Xscale
write.table(X, file = "matriceX.txt", row.names = FALSE)
```

Nella sezione di codice fornita, ci occupiamo del processo di standardizzazione delle variabili. Per prima cosa, calcoliamo le medie delle colonne nel data frame ***X*** e memorizziamo questi valori nella variabile "*m*". Successivamente, calcoliamo le deviazioni standard delle colonne e le immagazziniamo nella variabile "*s*". Creiamo quindi una copia del nostro data frame originale ***X*** chiamata ***Xscale***, che ci permette di seguire la standardizzazione senza modificare il dataset originale.

Infine, mediante un ciclo, standardizziamo ciascuna variabile nel data frame ***Xscale***, assicurandoci che ognuna di esse abbia una media di 0 e una deviazione standard di 1. Mediante la standardizzazione, infatti, la media di ciascuna variabile è stata centrata attorno allo zero. In secondo luogo, la deviazione standard di ciascuna variabile, dopo la standardizzazione, è stata impostata a 1. La deviazione standard misura quanto i dati si disperdono intorno alla media. Una deviazione standard di 1 indica che i dati sono distribuiti in modo uniforme intorno alla media standardizzata, senza una dispersione eccessiva o insufficiente.

La standardizzazione è particolarmente cruciale in quanto ci permette di mettere in relazione in modo coerente le diverse risposte dei visitatori, garantendo che nessuna variabile abbia un peso eccessivo sull'analisi rispetto alle altre e contribuendo a creare una base solida per l'identificazione dei cluster di visitatori con esperienze emotive simili all'interno della Pinacoteca Tosio Martinengo.

In tal modo, siamo stati in grado di individuare le somiglianze e le differenze nelle percezioni delle emozioni tra i visitatori con maggiore precisione, un passo fondamentale per l'efficacia dell'analisi di clustering.

Un'ultima fase cruciale nel processo di preparazione dei dati per l'analisi tramite lo *Spectral Clustering*, riguarda infine la creazione di una matrice di similarità. Questa matrice, come già affrontato nel Capitolo 2, è di fondamentale importanza per l'efficacia dell'algoritmo di clustering e consente di misurare la somiglianza tra le risposte al questionario dei partecipanti. La creazione di questa matrice è un passo chiave nell'applicazione dello *Spectral Clustering*, poiché offre una base solida per identificare gruppi omogenei di visitatori in base alle loro risposte simili alle domande sulle emozioni.

Per misurare la similarità tra le risposte dei partecipanti, adottiamo una specifica misura di similarità, nota come Kernel gaussiano o funzione gaussiana (Kaufman & Rousseeuw, 1990; Chen, Li, Liu, Xu, & Ying, 2017). Configuriamo quindi il parametro sigma (σ) a un valore di 1,5, prendendo questa decisione in base a considerazioni fondamentali. La nostra scelta è guidata dalla necessità di catturare le complesse relazioni non lineari tra le variabili e di ottenere una rappresentazione dei dati che sia flessibile e accurata.

Il Kernel gaussiano è stato preferito in quanto consente un'indagine approfondita della struttura dei dati, inclusa la capacità di esplorare le intricate relazioni tra le risposte alle domande sulle emozioni. Tuttavia, va notato che la scelta del parametro di scala σ è di fondamentale importanza quando si utilizza l'algoritmo di *Spectral Clustering* (Favati & al., 2020). Questo parametro regola il tasso di decadimento delle distanze tra le diverse osservazioni, influenzando direttamente la formazione dei cluster e, di conseguenza, il risultato dell'intero processo di clustering.

Una possibile strategia per selezionare il parametro σ nell'algoritmo di *Spectral Clustering* consiste nell'eseguire l'algoritmo ripetutamente per diversi valori di σ e successivamente scegliere quello che fornisce il clustering migliore, valutato attraverso una specifica misura di qualità. Tuttavia, questo approccio, sebbene promettente, presenta alcune sfide significative. Innanzitutto, comporta un aumento del carico computazionale, poiché richiede l'esecuzione dell'algoritmo multiple volte, con diverse impostazioni di σ. Inoltre, sorge la complessità di selezionare una misura di qualità affidabile, poiché questa può essere una funzione non monotona di σ (Favati & al., 2020). Di conseguenza, ottenere una determinazione affidabile di un valore σ accettabile potrebbe richiedere un numero considerevole di tentativi, comportando un'analisi complessa e dispendiosa.

È importante notare che la scelta del valore ottimale di σ può essere complessa poiché influisce in modo significativo la qualità del clustering risultante. Per semplificare il processo, molti algoritmi spettrali utilizzano valori approssimativi

come 0,5, 1 e 2, che rappresentano scelte comuni in questo contesto (Ghojogh, Ghodsi, Karray, & Crowley, 2021).

Nel nostro caso specifico, l'impostazione di σ a un valore fisso di 1,5 è stata motivata da diverse ragioni. In primo luogo, tale scelta si colloca all'interno di questa gamma di valori comuni, consentendo un bilanciamento tra la sensibilità alle distanze e la capacità di catturare relazioni non lineari tra le variabili (Ghojogh, Ghodsi, Karray, & Crowley, 2021). Un valore troppo piccolo di σ avrebbe causato distanze molto vicine a zero, portando a una percezione di similarità tra tutte le osservazioni. D'altro canto, un valore troppo grande avrebbe prodotto distanze vicine a uno, riducendo la sensibilità alle differenze sottili tra le risposte dei partecipanti (Favati & al., 2020).

Questo valore è emerso dunque come una scelta pragmatica che si adatta bene ai dati complessi, come le risposte alle domande sulle emozioni. Complessivamente, l'utilizzo di σ a 1,5 ci ha permesso di raggiungere un adeguato equilibrio tra la sensibilità alle relazioni tra le variabili e la gestione dei dati complessi, consentendo un'analisi accurata delle risposte emotive dei visitatori.

In base alle considerazioni esposte, procediamo ora a introdurre una funzione "*sim*" che svolge un ruolo essenziale nella misura di similarità tra due vettori ***x1*** e ***x2***, attraverso l'applicazione della funzione di similarità gaussiana.

```
######## Matrice di similarità

# Misura di similarità: Gaussian Kernel
sim <- function(x1, x2, sigma=1.5) {
  exp(-sum((x1-x2)^2)/(2*sigma^2))
  }
```

La similarità tra i vettori ***x1*** e ***x2*** è determinata considerando la distanza euclidea tra di essi e il parametro sigma, che controlla la larghezza del Kernel gaussiano. È fondamentale sottolineare che questa funzione è un passo cruciale nella creazione della matrice di similarità, che rappresenta un elemento fondamentale nel nostro processo di analisi dati con *Spectral Clustering*.

In seguito, definiamo una funzione denominata "*make.similarity*" la quale riceve come input la matrice ***X*** e una funzione di similarità (nel nostro caso, la funzione "*sim*").

```
make.similarity <- function(X, similarity) {
  n <- nrow(X)
  S <- matrix(rep(NA,n^2), ncol=n)
  for(i in 1:n) {
    for(j in 1:n) {
```

```
      S[i,j] <- similarity(X[i,], X[j,])
    }
  }
  S
}

S <- make.similarity(X, sim)
S[1:1024,1:1024]
```

Questa funzione ha il compito di calcolare e restituire la matrice di similarità ***S***. Il processo avviene attraverso un ciclo che scorre tutte le possibili coppie di righe nella matrice ***X*** (che contiene le risposte dei partecipanti) e calcola la similarità tra di esse utilizzando la funzione di similarità specificata.

La matrice di similarità risultante è memorizzata nella variabile ***S*** ed è un'importante componente dell'analisi con lo *Spectral Clustering*. La porzione di codice successiva, "S[1:1024, 1:1024]", ci consente di visualizzare la similarità tra le 1.024 righe e colonne di questa matrice.

La scelta del Kernel gaussiano si è dimostrata particolarmente appropriata in questo contesto, in cui le variabili coinvolte nel nostro studio erano di natura qualitativa ordinale. Le risposte dei partecipanti erano associate a livelli di percezione delle emozioni, ciascuno dei quali era rappresentato da un valore ordinale. La scelta della funzione gaussiana è stata fondamentale per catturare le relazioni intricate tra variabili ordinali. Questo aspetto è di particolare rilevanza quando ci si confronta con dati complessi, come le risposte delle domande sulle emozioni, in quanto ci ha permesso di condurre un'analisi dettagliata delle reazioni emotive dei visitatori.

Inoltre l'utilizzo del Kernel gaussiano è stato vantaggioso in quanto il parametro σ può essere ottimizzato per adattarsi meglio ai dati specifici, consentendo una maggiore flessibilità nell'analisi.

Tuttavia, la scelta del parametro di scala σ è stata un passo cruciale. Per affrontare questa sfida, è stato necessario selezionare σ con attenzione, basandosi sulla natura delle variabili e sui principi dell'algoritmo di *Spectral Clustering*. Questa selezione ponderata ha contribuito in modo significativo all'efficacia dell'analisi, consentendo di ottenere risultati accurati e rivelando strutture nascoste nei dati.

L'importanza di questa selezione accurata di σ è ancora più evidente quando consideriamo la natura delle nostre variabili, che sono qualitative ordinali. Queste variabili rappresentano i livelli di percezione delle emozioni, ognuno dei quali è associato a un valore ordinale. L'utilizzo della funzione gaussiana per calcolare la similarità tra gli oggetti basati su variabili ordinali ha dimostrato dunque di essere

un metodo robusto per affrontare le sfide legate all'ordinamento delle variabili e per catturare relazioni complesse tra di esse (Kaufman & Peter J., 1990; Chen, Li, Liu, Xu, & Ying, 2017).

Oltre all'applicazione del Kernel gaussiano, abbiamo eseguito il calcolo di ulteriori misure, tra cui la distanza Euclidea e la distanza di Gower. Tuttavia, è importante notare che quest'ultime non verranno utilizzate nella nostra analisi principale. La ragione dietro questa decisione è che abbiamo optato per l'applicazione dell'algoritmo di *Spectral Clustering* utilizzando il Kernel gaussiano, poiché è risultato essere più adeguato e efficace nel nostro contesto di studio.

```
######## Matrice di similarità con misura per variabili
quantitative (distanza Euclidea)

library(cluster)

# Costruiamo matrice di distanze con daisy
dist_euclidean <- daisy(X, metric = "euclidean")
dist_euclidean <- as.matrix(dist_euclidean)
summary(dist_euclidean)

one_matrix <- matrix(rep(1, len = n^2), nrow = n, ncol = n)
sim_euclidean <- one_matrix - dist_euclidean

# oppure:
#dist_euclidean <- dist(X, method = "euclidean", diag= T, upper =
T)
#dist_euclidean <- as.matrix(dist_euclidean)
```

In questa sezione del codice, abbiamo sfruttato la libreria "*cluster*" per calcolare diverse misure di similarità tra le righe della matrice $\boldsymbol{X}$, rappresentando le risposte dei partecipanti alle domande sulle emozioni.

Per prima cosa, abbiamo calcolato la distanza euclidea utilizzando il pacchetto "*daisy*" e la metrica "*euclidean*". Il risultato è stato memorizzato nella variabile "*dist_euclidean*" dopo aver trasformato la matrice di distanza in una matrice di similarità sottraendo i valori dalla matrice di unità ("*one_matrix*"). Questo passaggio ci ha fornito una misura di similarità euclidea tra le righe della matrice $\boldsymbol{X}$.

```
######## Matrice di similarità con misura per variabili ordinali
(distanza di Gower)

library(cluster)
```

```
# Diciamo a R che le variabili sono di tipo ordinale
X$Gioia <- factor(X$Gioia, ordered = T)
X$Tristezza <- factor(X$Tristezza, ordered = T)
X$Rabbia <- factor(X$Rabbia, ordered = T)
X$Paura <- factor(X$Paura, ordered = T)
X$Sorpresa <- factor(X$Sorpresa, ordered = T)
X$Disgusto <- factor(X$Disgusto, ordered = T)
```

Successivamente, per gestire le variabili qualitative ordinali nel nostro dataset, abbiamo effettuato una trasformazione dichiarando queste variabili come "*ordered*" (ordinali). Questo ci ha permesso di trattare in modo adeguato le relazioni di ordine tra le risposte dei partecipanti.

```
# Costruiamo matrice di distanze con daisy

dist_gower <- daisy(X, metric = "gower")
dist_gower <- as.matrix(dist_gower)
summary(dist_gower)

one_matrix <- matrix(rep(1, len = n^2), nrow = n, ncol = n)
sim_gower <- one_matrix - dist_gower
```

Infine, abbiamo calcolato la distanza di Gower utilizzando ancora una volta il pacchetto "*daisy*" e specificando la metrica "*gower*". Come prima, il risultato è stato memorizzato nella variabile "*dist_gower*", e abbiamo ottenuto la similarità di Gower sottraendo i valori dalla matrice di unità. Questa misura di similarità è utile per catturare le relazioni tra le variabili, considerando sia quelle quantitative che quelle ordinali (Kaufman & Rousseeuw, 1990).

Questi passaggi ci hanno consentito di disporre di diverse misure di similarità, inclusa la similarità euclidea e quella di Gower, sebbene successivamente abbiamo scelto di utilizzare il Kernel gaussiano per l'applicazione dell'algoritmo di *Spectral Clustering*. Nonostante non utilizzate direttamente in questa analisi, la valutazione di diverse misure di similarità ci ha permesso di esplorare le diverse prospettive con cui potremmo analizzare i dati.

In conclusione, la selezione della funzione gaussiana con parametro sigma impostato a 1,5 si è dimostrata la decisione più appropriata per affrontare le peculiarità delle nostre variabili qualitative ordinali. Questa scelta ha permesso di rivelare in modo efficace le sottili connessioni tra le risposte dei partecipanti alle domande sulle emozioni e ha contribuito significativamente a garantire una rappresentazione dei dati accurata e altamente flessibile, ottimizzando l'efficacia complessiva del processo di *Spectral Clustering*, fondamentale per la nostra analisi.

Dopo aver applicato il Kernel gaussiano alle risposte dei partecipanti, abbiamo generato una matrice di similarità in cui ciascuna cella rappresentava la similarità tra due partecipanti. Questa matrice è fondamentale per il nostro lavoro di *Cluster Analysis*, poiché servirà come punto di partenza per identificare e analizzare approfonditamente i dati al fine di individuare cluster di visitatori con profili emotivi simili.

Ora, questa matrice di similarità diventerà l'input cruciale per l'algoritmo di *Spectral Clustering*, il quale sfrutterà le relazioni di somiglianza tra i partecipanti per condurre un'analisi dettagliata e identificare i cluster di visitatori che condividono esperienze emotive simili.

La preparazione dei dati è stata dunque una fase cruciale per garantire il successo dell'analisi. Attraverso la predisposizione di un dataset, il controllo della qualità dei dati, la standardizzazione e, infine, l'organizzazione dei dati in una matrice di similarità, siamo stati in grado di creare una base solida per l'applicazione dell'algoritmo di *Spectral Clustering* e l'identificazione dei cluster di visitatori con esperienze emotive simili.

4.3 Implementazione dell'algoritmo di Spectral Clustering

Nel nostro percorso verso la comprensione e l'ottimizzazione dell'esperienza dei visitatori presso la Pinacoteca Tosio Martinengo, entriamo ora nella fase cruciale dell'applicazione dell'algoritmo di *Spectral Clustering*. Questa fase rappresenta il cuore della nostra analisi, in cui mettiamo in pratica i concetti teorici precedentemente esplorati per rivelare strutture nascoste all'interno dei dati sensoriali ed emotivi raccolti dai visitatori del museo. Nella seguente sezione, passo dopo passo, esploreremo l'implementazione dell'algoritmo di *Spectral Clustering*, offrendone una guida completa e dettagliata.

Questo processo complesso può essere suddiviso in diversi passaggi chiave, ognuno dei quali contribuisce al successo dell'algoritmo di *Spectral Clustering*. Inizieremo esaminando il calcolo del grafo di similarità e della matrice di affinità, fondamentali per la creazione della rappresentazione spettrale dei dati. Successivamente, affronteremo il calcolo della matrice Laplaciana, il calcolo degli autovettori e autovalori, e la creazione delle matrici $\boldsymbol{U}$ e $\boldsymbol{T}$. Infine, concluderemo con la creazione dei cluster, che rappresenta l'obiettivo finale dell'algoritmo di *Spectral Clustering*.

Con un focus sulla metodologia e sui passaggi chiave coinvolti, questa sezione offre un quadro completo del nostro approccio all'applicazione dello *Spectral*

Clustering all'analisi della *visitor experience* presso la Pinacoteca Tosio Martinengo. Attraverso questa analisi, cercheremo di rivelare le sfumature e le differenze nell'esperienza dei visitatori, contribuendo così a informare futuri sforzi di miglioramento e personalizzazione dell'esperienza museale.

4.3.1 Costruzione del grafo di similarità e della matrice di affinità

Nel processo di implementazione dell'algoritmo di *Spectral Clustering*, una tappa fondamentale è la costruzione del grafo di similarità, un passaggio critico che ci consente di rappresentare in modo efficace le relazioni tra i visitatori della Pinacoteca Tosio Martinengo, rivelando importanti insight sulle dinamiche delle loro risposte al questionario.

L'utilizzo di un grafo di similarità è guidato dalla necessità di cogliere in modo accurato e significativo le connessioni tra i visitatori, basandoci principalmente sui loro stimoli sensoriali e sulle esperienze artistiche vissute durante la visita al museo.

Nella sezione 2.5, abbiamo fornito una dettagliata spiegazione del processo di costruzione del grafo di similarità. In particolare, abbiamo discusso le specifiche tecniche utilizzate, sottolineando l'importanza cruciale di queste nella creazione di un grafo che sia rappresentativo delle relazioni effettive tra i visitatori. È essenziale notare che questa fase richiede una meticolosa attenzione ai dettagli ed è strettamente connessa con la scelta dalla misura di similarità da adottata nei passaggi precedenti, nel nostro caso la funzione gaussiana.

Le risposte al questionario, raccolte dai visitatori, sono state inizialmente trasformate in una matrice di similarità che ha permesso di catturare i punti di contatto tra le diverse esperienze sensoriali e artistiche dei partecipanti. La matrice di similarità risultante è stata poi utilizzata come base per la costruzione del grafo.

Il grafo di similarità è una rappresentazione visuale delle relazioni tra i visitatori del museo, in cui i vertici del grafo rappresentano i singoli visitatori e i collegamenti tra i vertici riflettono il grado di similarità tra le loro risposte al questionario. Questo grafo è fondamentale per l'algoritmo di *Spectral Clustering* in quanto fornisce una struttura di dati che prepara il terreno per l'identificazione dei cluster e l'interpretazione dei risultati ottenuti durante l'analisi di marketing sensoriale svoltasi presso la Pinacoteca.

Nel contesto della nostra ricerca, abbiamo adottato un approccio mirato per la costruzione del grafo basato sulla similarità delle risposte fornite dai visitatori al questionario somministrato. Tra le opzioni di costruzione del grafo, due approcci comuni emergono come protagonisti: il *fully connected graph* (o grafo

completamente connesso) e il *k-nearest neighbor graph* (von Luxburg, 2007). Tuttavia, la scelta tra queste due modalità di costruzione del grafo non è arbitraria ma dipende dalla natura dei dati e dagli obiettivi specifici dell'analisi.

Il *fully connected graph* è un concetto fondamentale nell'analisi dei dati. Rappresenta la massima espressione di connessione tra i punti dati, in quanto collega ogni punto con ogni altro punto sulla base di una specifica misura di similarità. Questa modalità di costruzione del grafo è cruciale quando desideriamo considerare tutte le connessioni possibili tra i dati e quando la funzione di similarità utilizzata è in grado di rappresentare accuratamente le relazioni tra i punti (von Luxburg, 2007).

D'altra parte, il *k-nearest neighbor graph* offre un approccio più selettivo per la costruzione del grafo. Invece di collegare ogni punto con tutti gli altri, questo metodo considera solo i *k* punti più vicini (*k-nearest neighbors*) per ciascun punto. L'idea è quella di creare connessioni locali, evidenziando le relazioni di prossimità nei dati. Questo approccio è particolarmente utile quando desideriamo dare maggiore rilevanza alle relazioni di vicinato locale tra i punti (von Luxburg, 2007).

Il metodo dei *k-nearest neighbors* funziona nel seguente modo: per ciascun visitatore nel nostro dataset, calcoliamo le similarità con tutti gli altri visitatori. Successivamente, selezioniamo i *k* visitatori più simili, quelli che condividono le risposte più vicine in termini di similarità. Questi *k* visitatori diventano i "vicini più prossimi" per il visitatore considerato.

Il parametro *k*, che rappresenta il numero di vicini da considerare, è una scelta critica in questo processo, come già approfonditamente esaminato nella sezione 2.5.1. Un valore troppo piccolo potrebbe portare a una rappresentazione incompleta delle connessioni, mentre un valore troppo grande potrebbe inclinare il grafo verso una completa connettività tra tutti i visitatori, perdendo così le sottili sfumature delle relazioni (Jönsson & Wohlin, 2006).

Una volta identificati i *k-nearest neighbors* per ciascun visitatore, creiamo la matrice di affinità che rappresenta le relazioni di similarità (Bhissy, Faleet, & Ashour, 2014). Questa matrice sarà successivamente utilizzata come base per l'algoritmo di *Spectral Clustering*, che sfrutterà le informazioni contenute nella matrice per identificare e analizzare i cluster di visitatori con esperienze emotive simili.

Nel nostro contesto di ricerca, adottiamo un approccio completo all'analisi dei dati, esplorando entrambe queste opzioni e valutando attentamente i risultati

ottenuti. Questo ci permetterà di valutare quale metodo sia più adatto alle peculiarità dei nostri dati e agli obiettivi specifici della ricerca. Sarà attraverso questa valutazione critica che determineremo il metodo ottimale da utilizzare nel nostro contesto di studio.

Per esaminare questa scelta in modo dettagliato, abbiamo sviluppato un codice specifico per entrambi i metodi. Inizialmente, abbiamo eseguito il grafo basato sui *k-nearest neighbor*, impostando il parametro "*n.neighbors*" con un valore specifico, nel nostro caso 32. Successivamente, abbiamo costruito il grafo completamente connesso, impostando il parametro "*n.neighbors*" al numero totale di osservazioni "*n*".

Nella nostra tesi, abbiamo affrontato la delicata decisione relativa alla scelta del parametro "*n.neighbors*" nella funzione "*make.affinity*". Come discusso nel Capitolo 2, tale parametro riveste un ruolo cruciale nella costruzione del grafo basato sul *k-nearest neighbor*, influenzando la connettività del grafo risultante e, di conseguenza, il processo di analisi dei cluster.

Abbiamo considerato le sfide legate alla determinazione del valore ottimale di *k*, tenendo presente le linee guida suggerite dalla letteratura. Una di queste linee guida empiriche è quella di selezionare *k* in modo che sia approssimativamente uguale alla radice quadrata del numero medio di casi completi, come indicato da Duda e Hart (Duda & Hart, 1973; Jonsson & Wohlin, 2004).

Nel nostro contesto specifico, abbiamo scelto di impostare "*n.neighbors*" a 32. Questa decisione è stata guidata da alcune considerazioni specifiche legate ai nostri dati e ai nostri obiettivi di analisi. Questo approccio ci ha permesso di ottenere risultati più chiari e interpretabili nel contesto della nostra analisi di *Spectral Clustering*.

Ora, procediamo alla presentazione del codice implementato per entrambe queste strategie:

```
######## Matrice di affinità con metodo k-nearest neighbors

make.affinity <- function(S, n.neighbors=32) {
  n <- length(S[,1])

  if (n.neighbors >= n) {  # completamente connesso
    W <- S
  } else {
    W <- matrix(rep(0,n^2), ncol=n)
    for(i in 1:n) { # per ogni riga
      # collega solo i punti con maggiore somiglianza
```

```
      best.similarities <- sort(S[i,],
decreasing=TRUE)[1:n.neighbors]
      for (sim in best.similarities) {
        j <- which(S[i,] == sim)
        W[i,j] <- S[i,j]
        W[j,i] <- S[i,j] # creare un grafo non direzionato, cioè
la matrice diventa simmetrica
      }
    }
  }
  W
}

W <- make.affinity(S, 33) # utilizzare 33 vicini (incluso se
stesso)
W[1:1024,1:1024]
```

Nella sezione di codice fornita, implementiamo una funzione personalizzata denominata "*make.affinity*", la quale svolge un ruolo centrale nel calcolo del grafo di similarità. Questa funzione richiede due argomenti principali: la matrice di similarità ***S*** e il parametro "*n.neighbors*", che specifica il numero di vicini da considerare durante il calcolo dell'affinità. Il risultato di questa funzione è una matrice di affinità denominata ***W***.

Per iniziare, calcoliamo la lunghezza delle righe nella matrice ***S*** ottenendo così il numero totale di osservazioni presenti nel dataset, che assegniamo alla variabile "*n"*. Questo valore è fondamentale per determinare le dimensioni della matrice di affinità ***W***.

Successivamente, effettuiamo una verifica condizionale per determinare se il numero di vicini desiderato, indicato come "*n.neighbors*", è maggiore o uguale al totale delle osservazioni "*n"*. In caso negativo, ovvero nel caso in cui il numero di vicini sia inferiore a "*n",* creiamo una matrice vuota ***W*** di dimensioni *nxn*, inizializzandola con valori zero. Per ciascuna riga del dataset, selezioniamo i "*n.neighbors*" punti con la maggiore somiglianza dalla matrice ***S*** e stabiliamo i collegamenti nella matrice ***W***. Questo processo genera un grafo non direzionato, il che significa che la matrice ***W*** deve risultare simmetrica.

Alla fine del processo, la matrice ***W*** contiene la matrice di affinità basata sul metodo dei *k-nearest neighbor*. Infine, applichiamo questa funzione alla matrice di similarità precedentemente calcolata ***S***, utilizzando un valore di "*n.neighbors*" pari a 33 (incluso anche il punto stesso).

Di conseguenza, la matrice risultante $\boldsymbol{W}$ riflette fedelmente le relazioni di affinità tra le 1.024 righe e colonne del dataset. Questo passaggio è fondamentale nell'ambito dell'analisi dei cluster.

Nel caso in cui il numero di vicini desiderato ("*n.neighbors*") sia maggiore o uguale al totale delle osservazioni, rappresentato da "*n"*, adottiamo l'approccio di un grafo completamente connesso. In questa situazione, impostiamo che la matrice di affinità $\boldsymbol{W}$ sia uguale alla matrice di similarità $\boldsymbol{S}$. Ciò significa che ogni punto nel dataset è connesso direttamente a tutti gli altri.

La decisione di avere un grafo completamente connesso è un passaggio cruciale nel contesto dell'implementazione dell'algoritmo di *Spectral Clustering*. Tale scelta può avere un impatto significativo sui risultati del clustering e, di conseguenza, deve essere valutata attentamente in base alle specifiche esigenze dell'analisi e alle caratteristiche dei dati.

Quando si desidera quindi ottenere un grafo completamente connesso, è necessario impostare il parametro "*n.neighbors*" in modo che corrisponda al numero totale di osservazioni o punti nel dataset, rappresentato in questo caso con "*n*". Ciò significa che ogni punto sarà direttamente connesso a tutti gli altri punti nel grafo, creando una rete densamente interconnessa.

Pertanto, nel nostro script, abbiamo proceduto come segue:

```
######## Matrice di affinità con grafo completamente connesso

A <- make.affinity(S,n.neighbors = n)  # Grafo completamente
connesso
```

In questo modo, abbiamo impostato il parametro "*n.neighbors*" uguale a "*n*", assicurando così la creazione di un grafo completamente connesso.

Questo approccio si rivela prezioso quando si desidera massimizzare la connettività tra le diverse osservazioni. Ad esempio, è particolarmente adatto per individuare cluster estremamente compatti o quando si presume che ciascuna osservazione possa influenzare direttamente tutte le altre.

Dopo aver ottenuto entrambi i risultati, abbiamo proceduto a una valutazione attenta dei medesimi. È importante notare che, nel contesto specifico della nostra analisi, abbiamo convenuto che l'utilizzo del *k-nearest neighbor graph* rappresenta la scelta più appropriata per la creazione della matrice di affinità. Questa decisione è giustificata dalla capacità del *k-nearest neighbor graph* di offrire un approccio più selettivo nella costruzione del grafo. Infatti, il *k-nearest neighbor graph*, considerando solo i *k* punti più vicini per ciascun punto nel

dataset, mira a creare connessioni locali e dunque a enfatizzare le relazioni di prossimità tra i dati. Tale selezione mirata dei vicini più prossimi è particolarmente utile quando si desidera dare maggiore rilevanza alle relazioni di vicinato locale tra i punti, poiché consente di identificare cluster o gruppi di punti che sono strettamente correlati tra loro in contesti specifici all'interno dei dati (von Luxburg, 2007). Pertanto, la scelta del *k-nearest neighbor graph* è coerente con l'obiettivo di catturare le relazioni di prossimità locali nei nostri dati e svolge un ruolo fondamentale nell'analisi e nella clusterizzazione dei dati.

Abbiamo dunque fatto la scelta consapevole di utilizzare il *k-nearest neighbor graph* come metodo preferito per la creazione della matrice di affinità. Abbiamo ritenuto che fosse il più idoneo per il nostro studio, in quanto questa decisione si è rivelata fondamentale per garantire una rappresentazione accurata delle relazioni tra i visitatori della Pinacoteca. Inoltre ha contribuito a ottimizzare l'efficacia complessiva del nostro processo di *Spectral Clustering*, un passaggio cruciale nella nostra analisi dei dati.

In conclusione, il calcolo del grafo di similarità rappresenta un pilastro fondamentale nell'applicazione dell'algoritmo di *Spectral Clustering*, poiché definisce le connesse tra le osservazioni e svolge un ruolo determinante nella comprensione dei cluster identificati.

L'approccio del *k-nearest neighbor graph*, associato all'uso della funzione di similarità gaussiana nella costruzione del grafo di similarità, ci ha permesso di concentrarci sulle relazioni più rilevanti tra i visitatori. Questa strategia ha creato una solida base per l'analisi dei cluster, offrendo un quadro significativo per esplorare le dinamiche delle risposte dei visitatori alla Pinacoteca e ha contribuito in modo sostanziale al successo complessivo dell'analisi.

4.3.2 Calcolo della matrice Laplaciana

Dopo aver costruito la matrice di affinità pesata $\boldsymbol{W}$, un passo cruciale nel nostro processo di analisi dei dati è il calcolo della matrice Laplaciana (Chung, 1997). Questo calcolo costituisce un punto nodale in quanto offre una misura significativa della struttura sottostante del grafo di similarità da noi generato ed è centrale per l'intera analisi dei dati.

Nel contesto della nostra ricerca sullo *Spectral Cluste*ring, seguendo l'approccio dell'algoritmo di *Normalized Spectral Clustering* proposto da Ng et al. (2001), utilizzeremo la matrice Laplaciana normalizzata simmetrica $\boldsymbol{L_{sym}}$ (von Luxburg, 2007).

La matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$ è calcolata seguendo l'Equazione (1.12) che avevamo precedentemente definito (von Luxburg, 2007). Questa matrice è una rappresentazione matematica che deriva direttamente dalla matrice di affinità pesata $\boldsymbol{W}$. La sua importanza risiede nel fatto che cattura in modo accurato la struttura delle relazioni tra i visitatori, come riflessa nel grafo di similarità.

Per comprendere meglio l'importanza di questa matrice, è utile richiamare alcune delle sue proprietà chiave. Innanzitutto, la matrice Laplaciana normalizzata simmetrica $\boldsymbol{L_{sym}}$ è una misura della connettività tra i vertici del grafo, dove ciascun vertice rappresenta un visitatore nel nostro contesto di studio. Essa tiene conto delle affinità tra i visitatori sulla base delle risposte fornite al questionario, considerando sia la similarità tra di essi che la loro relativa posizione nel grafo (Chung, 1997).

Inoltre, questa matrice assume un ruolo cruciale nell'ambito dell'analisi spettrale, una tecnica di fondamentale importanza nell'algoritmo di *Spectral Clustering*. Gli autovalori e gli autovettori associati a $\boldsymbol{L_{sym}}$ contengono informazioni preziose riguardo alla struttura dei di cluster nei dati. Essi consentono di individuare gruppi omogenei tra i visitatori, contribuendo in modo significativo all'identificazione dei pattern di comportamento e delle relazioni tra i partecipanti al questionario (Chung, 1997).

Per ottenere la matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$, sfruttiamo la potenza di una libreria R specializzata chiamata "*matrixLaplacian*". Questa libreria si è dimostrata uno strumento prezioso, semplificando notevolmente l'elaborato processo di calcolo. Il suo utilizzo evita la necessità di scrivere un codice personalizzato complesso, garantendoci efficienza e precisione nei risultati. Il procedimento di calcolo è eseguito mediante l'esecuzione del seguente codice:

```
######## Matrice Laplaciana

# install.packages("matrixLaplacian")

library(matrixLaplacian)
```

Iniziamo con il caricamento della libreria "*matrixLaplacian*", la quale è utilizzata per calcolare la matrice Laplaciana normalizzata del grafo. La libreria fornisce funzioni e strumenti specifici per questo calcolo.

Calcolando dunque la matrice Laplaciana normalizzata del grafo rappresentato dalla matrice di affinità $\boldsymbol{W}$:

```
# Calcolo della matrice laplaciana normalizzata del grafo
rappresentato dalla matice W

normalised_laplacian <- matrixLaplacian(W, plot2D=F, plot3D = F)
```

Questo passaggio coinvolge la funzione "*matrixLaplacian*", che accetta la matrice di affinità $\boldsymbol{W}$ come input. L'opzione "*plot2D = F*" e "*plot3D = F*" è utilizzata per evitare la visualizzazione dei grafici 2D e 3D associati a questa operazione, poiché ci interessa principalmente il calcolo numerico dei risultati.

Una volta calcolata la matrice Laplaciana normalizzata, estraiamo questa matrice e la assegniamo alla variabile "*normalised_laplacian_matrix*".

```
# Estrazione della matrice laplaciana normalizzata

normalised_laplacian_matrix <-
normalised_laplacian$LaplacianMatrix
```

Questa matrice sarà essenziale nelle fasi successive dell'algoritmo di *Spectral Clustering*.

Infine, estraiamo gli autovettori associati alla matrice Laplaciana normalizzata.

```
# Estrazione degli autovettori della matrice laplaciana
normalizzata

norm_lap_eigenvectors <- normalised_laplacian$eigenvector
```

Gli autovettori sono importanti per l'analisi dei cluster e ci permetteranno di scomporre i dati in uno spazio diverso per l'identificazione dei pattern nei prossimi passaggi dell'algoritmo.

Il calcolo della matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$ rappresenta dunque una tappa fondamentale nell'implementazione dell'algoritmo di *Spectral Clustering*. Questo processo riveste un ruolo cruciale nel preparare le matrici e gli autovettori essenziali per l'analisi dei dati. Nell'ambito della nostra ricerca, questi risultati saranno utilizzati per individuare e interpretare i cluster di visitatori con comportamenti simili. Inoltre ci aiuteranno a comprendere le relazioni tra questi cluster, contribuendo in modo significativo all'identificazione dei pattern di comportamento tra i visitatori della Pinacoteca. Questo passaggio costituisce dunque un solido fondamento per l'analisi approfondita dei dati e per il conseguimento dei nostri obiettivi di ricerca.

4.3.3 Calcolo degli autovettori e degli autovalori

Il prossimo passo fondamentale nel nostro processo di implementazione dell'algoritmo di *Spectral Clustering* è il calcolo degli autovettori e degli autovalori della matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$ (Chung, 1997). Questi autovettori e autovalori sono elementi cruciali per comprendere la struttura dei dati e identificare i cluster.

Di seguito mostriamo il codice che utilizziamo per eseguire la decomposizione spettrale e ottenere gli autovettori e gli autovalori:

```
######## Calcolo autovalori e autovettori

# Eseguire la decomposizione spettrale per ottenere gli
autovettori e gli autovalori

eigen_result <- eigen(normalised_laplacian_matrix)
eigenvalues <- eigen_result$values
eigenvectors <- eigen_result$vectors
```

Inizialmente, eseguiamo la decomposizione spettrale sulla matrice Laplaciana normalizzata precedentemente calcolata, denominata "*normalised_laplacian_matrix*". Questa operazione è eseguita utilizzando la funzione "*eigen*" di R, che permette di ottenere sia gli autovettori che gli autovalori associati alla matrice. Gli autovalori, contenuti nella variabile *"eigenvalues*", sono risultati preziosi per comprendere la distribuzione dei dati e rilevare eventuali strutture nascoste nei dati. Gli autovettori, invece, memorizzati nella variabile "*eigenvectors*", rappresentano i modelli principali all'interno dei dati, consentendoci di identificare cluster omogenei e relazioni rilevanti tra i visitatori nel nostro studio.

In poche righe di codice, abbiamo ottenuto un insieme di autovalori e autovettori che costituiscono un elemento centrale per il nostro algoritmo di *Spectral Clustering*.

Ora ci focalizziamo sull'ottenere i primi k autovalori $\lambda_1, \dots, \lambda_k$ insieme ai loro corrispondenti autovettori $u_1, \dots, u_k$ della matrice Laplaciana normalizzata $\boldsymbol{L_{sym}}$. Questi autovettori e autovalori rivestono un ruolo cruciale nella nostra analisi in quanto rappresentano le direzioni principali di variazione nei dati (Chung, 1997). Contenendo informazioni cruciali sulla struttura dei dati, saranno fondamentali per il processo di creazione dei cluster.

All’interno del nostro script R, includiamo una sezione dedicata al calcolo e alla visualizzazione di alcune informazioni relative agli autovalori e agli autovettori. Questa sezione è suddivisa in due parti chiave.

Inizialmente, per valutare in modo accurato il contributo degli autovalori al comprensivo spettro dei dati, adottiamo un approccio visuale. A tal scopo, creiamo un grafico degli autovalori associati ad ogni dimensione, anche chiamato *scree plot*. Tale grafico svolge un ruolo essenziale nel processo di selezione degli autovalori rilevanti per la nostra analisi. Esso offre una chiara rappresentazione visiva che ci consente di determinare quanti autovalori contribuiscono in modo sostanziale a catturare la complessità e la varianza dei nostri dati. In altre parole, ci permette di individuare quanta “informazione” ciascun autovalore aggiunge alla comprensione dei modelli nei dati.

Di seguito, mostriamo il codice che utilizziamo per creare il grafico degli autovalori con margini ridotti:

```
######## Grafico degli autovalori associati ad ogni dimensione

par(mar = c(4, 4, 2, 2))  # Imposta i margini (bottom, left, top, right)
plot(eigenvalues[1:30], type = "b", xlab = "Numero di Dimensioni",
     ylab = "Autovalori")
pdf(file=paste("ScreePlot",".pdf"),paper='special')
```

Il codice utilizza la funzione “*plot*” per visualizzare i primi 30 autovalori (questo valore è scelto per agevolare l’individuazione del numero di dimensioni da considerare) sull’asse delle ascisse e i valori degli autovalori sull’asse delle ordinate. Gli assi del grafico sono etichettati in modo chiaro per consentire una facile interpretazione dei risultati.

Questo grafico è stato progettato per visualizzare chiaramente il contributo di ciascun autovalore nell’ambito dell’analisi dei dati. Guardando il grafico, possiamo identificare rapidamente quanti autovalori contribuiscono significativamente alla struttura dei dati e quindi siano rilevanti per la nostra analisi.

Ogni autovalore rappresenta quanto un determinato autovettore incide sulla variazione dei dati (Chung, 1997). Pertanto, analizzando il grafico, possiamo identificare il punto in cui la spezzata degli autovalori presenta una brusca caduta. In altre parole, quando notiamo una “discontinuità” nell’andamento del grafico, possiamo concludere che i primi autovalori sono più rilevanti per la nostra analisi,

poiché contengono una quantità molto elevata di informazione (von Luxburg, 2007).

Nella Figura 4.3 viene presentato il grafico degli autovalori risultante dalla nostra analisi.

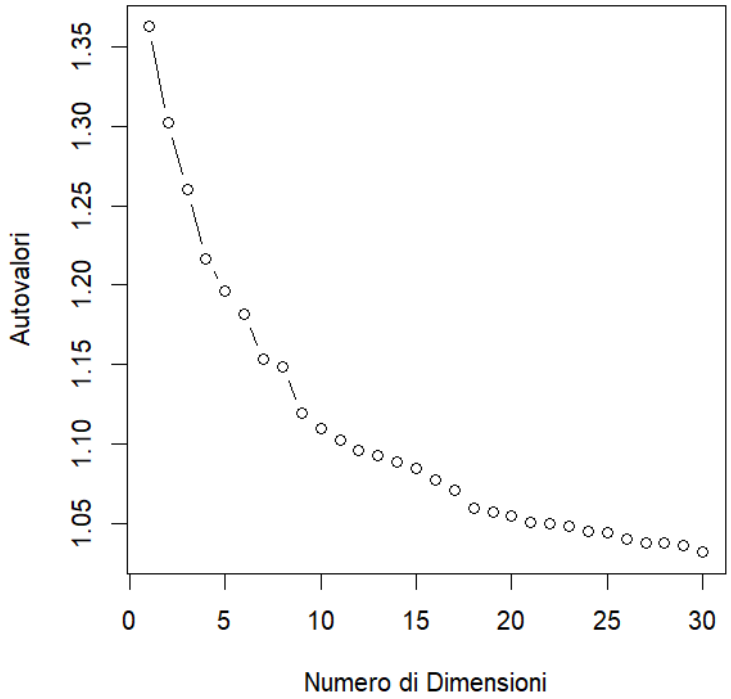


Figura 4.3 – Grafico degli autovalori associati ad ogni dimensione.
Fonte: nostre elaborazioni.

Inizialmente, possiamo osservare una rapida diminuzione della quantità di informazione, il che significa che i primi autovalori contribuiscono significativamente a catturare le principali caratteristiche dei dati. Tuttavia, man mano che ci spostiamo verso destra nel grafico, la diminuzione diventa sempre meno pronunciata, suggerendo che gli autovalori successivi apportano un contributo sempre meno significativo.

In pratica, questo grafico ci fornisce una guida visiva per selezionare il numero ottimale di autovalori da considerare nella nostra analisi. Quando notiamo che la decrescita dell'informazione rallenta notevolmente e che gran parte di questa è stata catturata dai primi autovalori, possiamo concludere che i successivi hanno un impatto marginale sulla nostra comprensione dei dati.

Di conseguenza, utilizziamo questo grafico per determinare il numero di autovalori da includere nel nostro modello e nel nostro processo di *Spectral Clustering*. La scelta di questo parametro è cruciale per ottenere una

rappresentazione rilevante dei dati e per concentrare l'attenzione sui modelli di connessione più rilevanti tra i visitatori nella nostra analisi.

Dopo aver acquisito dunque una visione d'insieme degli autovalori, diventa cruciale decidere esattamente quanti di essi, e quindi quante componenti principali, vogliamo considerare nel nostro processo di clustering (von Luxburg, 2007). Il parametro *k_eigenvectors* rappresenta il numero di autovettori che intendiamo utilizzare per il clustering. Nella nostra specifica analisi, abbiamo deciso di utilizzare i primi 8 autovettori ($k\_eigenvectors = 8$) per rappresentare i dati.

```
# Numero primi k autovettori

k_eigenvectors <- 8
```

La scelta di utilizzare i primi 8 autovettori è stata basata su considerazioni derivate dall'osservazione del grafico degli autovalori. Abbiamo scelto di considerare i primi 8 autovettori poiché sono in grado di catturare una quantità di informazione estremamente rilevante, come chiaramente dimostrato dal grafico. Questi 8 autovettori rappresentano efficacemente le principali direzioni di variazione nei dati e, di conseguenza, sono stati ritenuti sufficienti per il nostro processo di clustering. La scelta di questo valore è stata finalizzata a ottenere un equilibrio tra la comprensione delle caratteristiche dei dati e la riduzione della complessità computazionale del modello.

È fondamentale sottolineare che selezione del numero appropriato di autovettori ha un impatto significativo sulla fase successiva del nostro processo di *Spectral Clustering*. Questa selezione influisce direttamente sulla qualità e l'accuratezza del nostro clustering, e quindi deve essere effettuata con grande attenzione e considerazione. La corretta determinazione di quanti autovettori includere è cruciale per garantire che il nostro modello sia in grado di catturare in modo efficace le strutture nascoste nei dati e produrre risultati significativi nella fase di clustering.

Questi risultati rappresentano dunque un tassello cruciale per la nostra analisi dei dati. Gli autovettori e autovalori che abbiamo estratto saranno ampiamente utilizzati nelle fasi successive dell'algoritmo per effettuare la suddivisione ottimale delle osservazioni nei vari cluster. Questo passaggio è essenziale per ottenere una comprensione approfondita dei comportamenti e delle dinamiche dei visitatori della Pinacoteca.

L'utilizzo di questi autovettori e autovalori nei calcoli successivi ci consente dunque di assegnare correttamente le osservazioni ai cluster appropriati,

contribuendo in modo significativo alla riuscita dell'intero processo di *Spectral Clustering*. Grazie a questa analisi, siamo in grado di ottenere una visione dettagliata e informativa delle relazioni tra i visitatori, delle strutture di cluster presenti nei dati e delle interazioni all'interno della Pinacoteca.

4.3.4 Creazione delle matrici U e T

Dopo aver selezionato il numero desiderato di autovettori, passiamo alla creazione di una matrice speciale denominata ***U***. Questa matrice svolge un ruolo fondamentale nelle fasi successive del nostro algoritmo di *Spectral Clustering*, poiché contiene gli autovettori associati ai primi *k* autovalori che abbiamo scelto di considerare. La sua dimensione sarà quindi $n \times k$, dove *n* rappresenta il numero di osservazioni del nostro dataset e *k* è il numero di autovettori selezionati (von Luxburg, 2007).

L'implementazione di questa matrice ***U*** è effettuata mediante il seguente codice:

```
######## Matrice degli autovettori U

U <- eigenvectors[,1:k_eigenvectors]

n <- nrow(U)  # Numero di righe
k_eigenvectors <- ncol(U)  # Numero di colonne
```

La matrice ***U*** cattura le informazioni di raggruppamento contenute negli autovettori. Ogni colonna di questa matrice rappresenta un autovettore corrispondente a uno specifico autovalore. In sostanza, rappresenta una sorta di mappa delle connessioni principali nei dati, identificando le relazioni significative tra i visitatori. Questi autovettori ci consentono di esaminare la struttura sottostante dei dati, catturando le principali direzioni di variazione nei nostri dati.

La matrice ***U*** rappresenta ciò che viene comunemente definito "spettro dell'*embedding*" (o *spectral embedding*) (Niu D., 2011). Questo termine si riferisce essenzialmente a una proiezione dei dati in un nuovo spazio delle caratteristiche completamente differente. In questo nuovo spazio, le informazioni di similarità tra i punti sono rappresentate in modo più rilevante. Questo processo di proiezione rende più agevole la separazione e il clustering dei dati in base alle strutture intrinseche rivelate dagli autovalori e dagli autovettori.

In altre parole, ***U*** svolge il compito di trasformare i dati originali in un nuovo spazio in cui le relazioni tra i punti sono più evidenti, facilitando così l'analisi dei cluster e l'identificazione delle strutture nascoste all'interno dei dati. Questo rappresenta un passo fondamentale nel nostro processo di *Spectral Clustering*,

poiché ci consente di esplorare le dinamiche dei visitatori in modo più accurato e dettagliato, contribuendo così al successo complessivo dell'analisi.

Grazie a questa matrice, siamo dunque in grado di svolgere l'analisi di *Spectral Clustering* per individuare cluster omogenei tra i visitatori e rilevare relazioni tra di essi nel contesto della nostra ricerca.

È importante notare che il processo di *spectral embedding* è continuo. Questo significa che, per ottenere una suddivisione discreta dei dati nei *k* cluster desiderati, dobbiamo eseguire un passaggio di "arrotondamento". Questo passaggio è cruciale per garantire una migliore gestione dei dati e una maggiore coerenza nell'analisi.

L'algoritmo di arrotondamento si basa sulla ri-normalizzazione di ogni riga della matrice $\boldsymbol{U}$ al fine di ottenere una nuova matrice, denominata $\boldsymbol{T}$, di dimensione nxk. In questa nuova matrice $\boldsymbol{T}$, ogni riga è stata regolarizzata in modo che abbia una norma pari a 1, il che significa che le righe hanno lunghezza unitaria. In altre parole, $\boldsymbol{T}$ rappresenta ciò che potremmo chiamare uno *spectral embedding* discreto (von Luxburg, 2007). In questa rappresentazione, le informazioni di similarità tra i punti sono organizzate in un formato più adatto per suddividere i dati in *k* cluster, come descritto anche da Ng, Jordan, & Weiss (2001).

Questo processo è eseguito mediante l'implementazione del seguente codice R:

```
######## Matrice normalizzata T

# Inizializzazione della matrice T

T <- matrix(0, nrow = n, ncol = k_eigenvectors)
# Calcolo della matrice T
for (i in 1:n) {
  denominator <- 0

  for (j in 1:k_eigenvectors) {
    denominator <- denominator + U[i, j]^2
  }

  denominator_sqrt <- sqrt(denominator)

  for (j in 1:k_eigenvectors) {
    T[i, j] <- U[i, j] / denominator_sqrt
  }
}

# Visualizzazione della matrice T
```

```
T

write.table(T, file = "matriceT.txt", row.names = FALSE)
```

Questo codice calcola la matrice ***T***, in cui ogni riga è normalizzata in modo che la norma euclidea delle righe sia pari a 1.

Una volta ottenuta la matrice ***T***, trattiamo le sue righe come nuovi punti in uno spazio *k*-dimensionale. Questo ci fornisce un insieme di vettori, ognuno dei quali rappresenta uno dei punti originali del dataset trasformato nello spazio dei *k* autovettori normalizzati. Questo nuovo spazio è stato appositamente progettato per rappresentare le informazioni di similarità tra i punti in un formato più adatto per suddividere i dati in *k* cluster. In sostanza, lo spazio è ottimizzato per facilitare l'analisi dei cluster e rivelare le strutture nascoste all'interno dei dati, contribuendo in modo significativo al successo dell'intero processo di *Spectral Clustering*

4.3.5 Creazione dei cluster

Una volta ottenuta la matrice dei punti trasformati ***T*** e rappresentanti i nostri dati nello spazio dei *k* autovettori normalizzati, procediamo con il clustering dei punti, al fine di giungere a una suddivisione dei visitatori in cluster omogenei. Concordemente all'algoritmo proposto da Ng et al. (2001), l'algoritmo scelto per questa fase è il *k-means*, una tecnica di clustering ampiamente utilizzata che mira a suddividere i dati in *k* cluster in modo da minimizzare la somma dei quadrati delle distanze tra i punti e i centroidi dei rispettivi cluster (von Luxburg, 2007). Questo processo di clustering ha consentito di raggruppare i visitatori in cluster basati sulla loro similarità nei dati delle risposte al questionario.

Applichiamo dunque l'algoritmo *k-means* sulla matrice dei punti ***T*** per ottenere i cluster. Il codice esegue i seguenti passaggi:

```
######## Creazione cluster

# install.packages("stats")

library (stats)

# Esecuzione algoritmo k-means sulla matrice T

k <-8
cluster <- kmeans(T, k, 10) # 10: numero massimo di interazioni
per l'algoritmo di clustering
```

```
# Estrazione delle etichette dei cluster assegnate a ciascuna
osservazione della matrice

cluster_number <- cluster$cluster # cluster$cluster è un vettore
di numeri interi che rappresenta l'appartenenza di ciascuna
osservazione al cluster corrispondente

# Creazione di un nuovo dataset senza la colonna degli
identificativi (le colonne contengono le variabili utilizzate per
il clustering)

data_clustered <- dataset[,colstart:colend]

# Aggiungere la colonna cluster_number che contiene le etichette
dei cluser assegnate a ciscuna osservazione (ogni riga di
data_clustered viene assegnata al cluster corrispondente)

data_clustered$cluster_number <- cluster_number

write.table(cluster_number, file = "cluster.txt", row.names =
FALSE)
```

Abbiamo avviato questa fase dell'analisi importando la libreria "*stats*", fondamentale per l'esecuzione dell'algoritmo *k-means*. Il nostro obiettivo era chiaro: suddividere i dati in cluster rilevanti che potessero aiutarci a individuare i pattern nascosti nei nostri dati.

L'implementazione dell'algoritmo *k-means* è condotta sulla matrice ***T***, ottenuta dalla normalizzazione delle righe della matrice dei punti. Abbiamo incluso il parametro 10 per limitare il numero massimo di iterazioni consentite durante il processo di clustering, garantendo così una buona efficienza computazionale.

Una volta eseguito l'algoritmo *k-means*, è emersa la necessità di associare ciascuna osservazione del nostro dataset a uno dei cluster identificati. Questo compito è reso possibile grazie all'estrazione delle etichette dei cluster, raccolte nel vettore "*cluster_number*". Questo vettore assegna a ciascuna osservazione un'etichetta numerica, identificando così il cluster di appartenenza.

Creiamo quindi un nuovo dataset denominato "*data_clustered*", selezionando le colonne del dataset originale che sono state utilizzate nel processo di clustering. Questo dataset "*data_clustered*" agisce come una sorta di rappresentazione dei dati clusterizzati.

Per completare questa fase, aggiungiamo una colonna chiamata "*cluster_number*" a "*data_clustered*", assegnando a ciascuna osservazione l'etichetta del cluster corrispondente. In questo modo, otteniamo un dataset "*data_clustered*"

contenente le nostre osservazioni originali, arricchite da un'indicazione chiara e categorica del cluster di appartenenza.

L'utilizzo dell'algoritmo *k-means* ha permesso di suddividere efficacemente il nostro dataset trasformato nello spazio dei *k* autovettori normalizzati, in *k* cluster omogenei, facilitando così l'analisi dei comportamenti e delle dinamiche dei visitatori all'interno di ciascun cluster.

Nel corso di questo capitolo, abbiamo fatto ampio uso di vari pacchetti R e funzioni, alcune delle quali sono state adattate da pacchetti esistenti, mentre altre sono state personalizzate per soddisfare le esigenze specifiche della nostra analisi. Questa diversificazione di strumenti ci ha consentito di condurre un'analisi approfondita e dettagliata del nostro dataset, sfruttando appieno l'algoritmo di *Spectral Clustering*. Attraverso la costruzione di matrici, il calcolo degli autovettori e degli autovalori, la normalizzazione dei dati e l'applicazione dell'algoritmo di clustering *k-means*, siamo stati in grado di suddividere i visitatori in cluster omogenei basati sulla loro esperienza sensoriale e le risposte al questionario.

Questo passaggio rappresenta una pietra miliare del nostro processo di analisi, poiché fornisce una base solida per le fasi successive, consentendoci di esplorare e interpretare in modo efficace i cluster identificati e le dinamiche dei visitatori della Pinacoteca Tosio Martinengo.

Per fornire una visione d'insieme completa e sistematica di tutte le componenti cruciali delle nostre analisi e degli strumenti utilizzati, presenteremo una tabella di sintesi. Questa tabella rappresenterà un utile punto di riferimento, riassumendo il lavoro svolto e mettendo in evidenza i principali elementi dell'analisi dei dati e dell'algoritmo di *Spectral Clustering* impiegato. La nostra analisi, basata sull'algoritmo di *Spectral Clustering*, ha dunque contribuito significativamente a suddividere i visitatori in cluster omogenei, offrendo preziose informazioni per migliorare l'esperienza museale e le strategie di marketing del museo stesso.

Tabella 4.4 – Tabella di riepilogo degli strumenti utilizzati in R.
Fonte: nostre elaborazioni.

Pacchetto	**Funzione**	**Uso**
Funzioni di nostra elaborazione	`sim()`	Calcola la similarità utilizzando il Kernel gaussiano.
Funzioni di nostra	`make.similarity()`	Crea la matrice di

elaborazione		similarità.
`Cluster`	`daisy()`	Calcola la matrice delle distanze Euclidee per variabili quantitative.
`Cluster`	`daisy()`	Calcola la matrice delle distanze di Gower per variabili ordinali.
Funzioni di nostra elaborazione	`make.affinity()`	Crea la matrice di affinità utilizzando il metodo dei *k-nearest neighbor*.
`MatrixLaplacian`	`matrixLaplacian()`	Calcola la matrice Laplaciana di un grafo rappresentato dalla matrice di affinità $\boldsymbol{W}$.
`Stats`	`kmenas()`	Esegue l'algoritmo di *k-means* sui dati.

4.4 Risultati e interpretazione

L'interpretazione dei risultati ottenuti dal clustering riveste un ruolo fondamentale nell'intero processo di ricerca di mercato sull'esperienza dei visitatori. Tale interpretazione consente di cogliere appieno l'importanza delle informazioni emerse e di collegarle all'obiettivo iniziale della ricerca, consentendo così una comprensione più profonda dei dati raccolti.

Nel contesto di questa sezione, procederemo con un'analisi dettagliata dei cluster identificati dall'algoritmo di *Spectral Clustering*. Ogni cluster rappresenta un gruppo distinto di visitatori della Pinacoteca Tosio Martinengo, ciascuno caratterizzato da tratti e comportamenti emotivi specifici. Questa suddivisione in cluster è stata realizzata in base alle similitudini tra i visitatori, consentendo di evidenziare pattern e tendenze comuni all'interno di ciascun gruppo.

L'obiettivo principale di questa segmentazione dei dati è quello di ottenere una visione più chiara e dettagliata delle diverse tipologie di visitatori, identificando le caratteristiche che li accomunano. In questo modo, siamo in grado di comprendere meglio come i visitatori percepiscono e vivono l'esperienza museale in modo unico e differenziato.

Attraverso l'analisi dei cluster, emergono insight preziosi che possono essere sfruttati per ottimizzare l'esperienza dei visitatori e per sviluppare strategie di marketing mirate. Questi insight ci permettono di rispondere a domande cruciali, come quali tipologie di visite attraggono determinati gruppi di visitatori, come adattare le esposizioni o le attività del museo per soddisfare le diverse esigenze emotive dei visitatori e come personalizzare le strategie di marketing per ciascun cluster al fine di massimizzare l'attrattività del museo.

Nel perseguire l'obiettivo di comprendere appieno l'esperienza dei visitatori presso la Pinacoteca e migliorare la loro *visiting experience*, abbiamo applicato con successo l'algoritmo di *Spectral Clustering* ai dati raccolti. Questo approccio ci ha permesso di segmentare la nostra vasta popolazione di visitatori in 8 cluster distinti, ognuno dei quali rappresenta un gruppo unico di visitatori con caratteristiche emotive e comportamentali simili. Questi cluster non sono semplici aggregazioni di dati, ma piuttosto rappresentano un potente strumento per svelare le sfumature e le specificità dell'esperienza museale dei nostri visitatori. La composizione dei cluster rivela il numero di visitatori che è stato assegnato a ciascun gruppo. Questi numeri, specificamente 161, 100, 82, 95, 91, 191, 183 e 121 visitatori per i rispettivi cluster, indicano la dimensione di ogni segmento di pubblico. Questo significa che, ad esempio, 161 visitatori con caratteristiche simili sono stati raggruppati nel primo cluster, mentre nel secondo cluster ci sono 100 visitatori con comportamenti e reazioni emotive affini.

Nella Tabella 4.5, è fornita una visualizzazione della composizione di ciascun cluster.

Tabella 4.5 – Distribuzione assoluta per cluster.
Fonte: nostre elaborazioni.

Cluster	Numero soggetti
1	161
2	100
3	82
4	95
5	91
6	191
7	183
8	121
Totale	**1024**

Questi numeri riflettono la distribuzione dei visitatori all'interno delle categorie di clustering, evidenziando quanto ciascun gruppo sia rappresentativo all'interno della popolazione di visitatori del museo. Questa suddivisione in cluster fornisce una visione più dettagliata e segmentata della base di visitatori, consentendo al

museo di comprendere meglio le diverse tipologie di visitatori e di adattare le strategie di marketing e le iniziative di miglioramento dell'esperienza in modo più mirato. La Figura 4.6 illustra chiaramente queste distribuzioni, offrendo una rappresentazione visiva delle frequenze relative dei cluster.

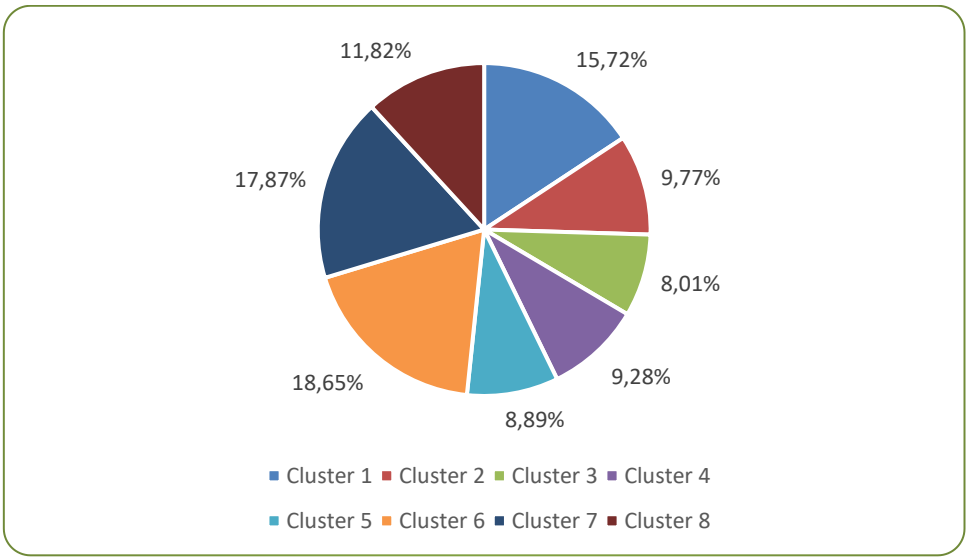


Figura 4.6 – Rappresentazione percentuale dei cluster.
Fonte: nostre elaborazioni.

Analizzeremo in questa fase i cluster identificati, ognuno dei quali rappresenta un segmento distinto e omogeneo dei visitatori della Pinacoteca Tosio Martinengo. L'obiettivo principale di questa analisi è quello di gettare luce sulle diverse sfaccettature dell'esperienza dei visitatori e di fornire una panoramica dettagliata delle caratteristiche distintive di ciascun gruppo.

L'analisi dei cluster riveste un'importanza fondamentale perché consente di andare oltre una semplice comprensione dei visitatori come un unico insieme e, invece, permette di suddividerli in categorie significative in base alle loro esperienze, percezioni ed emozioni. Questi cluster sono il risultato di complessi calcoli statistici e algoritmi di *Spectral Clustering* che hanno esaminato una vasta gamma di risposte ai questionari forniti dai visitatori.

In questo contesto, ciascun cluster assume il ruolo di una "finestra" attraverso cui possiamo esaminare da vicino le diverse reazioni e preferenze dei visitatori. Ogni finestra si apre su un mondo di caratteristiche condivise, incluse emozioni predominanti, approcci alla visita, preferenze artistiche e altro ancora. Questi cluster non solo ci aiutano a comprendere meglio i visitatori, ma forniscono anche un fondamentale strumento di supporto decisionale per la Pinacoteca e per gli operatori di marketing.

L’analisi dei cluster ci permette di rispondere a domande chiave, come quali emozioni prevalgono in ciascun gruppo e quale sia il grado di coinvolgimento emotivo dei visitatori in ciascun segmento. Oltre a ciò, offre un’opportunità unica per progettare strategie mirate che possano soddisfare le esigenze specifiche di ciascun cluster, migliorando l’esperienza complessiva dei visitatori e aumentando l’efficacia delle iniziative di marketing.

In questa sezione, esamineremo in dettaglio ciascun cluster emerso dall’analisi, descrivendo le sue caratteristiche distintive e mettendo in evidenza come queste informazioni possano essere applicate per migliorare l’esperienza dei visitatori e guidare le decisioni strategiche future. Esamineremo passo per passo ciascun cluster, fornendo così un quadro esaustivo delle diverse tipologie di visitatori della Pinacoteca e delle loro esperienze.

4.4.1 Etichettatura dei cluster

Ogni cluster rappresenta una categoria unica di visitatori, ciascuna con le proprie caratteristiche e tendenze comportamentali distintive. L’analisi dei cluster fornisce una comprensione dettagliata delle diverse tipologie di visitatori che frequentano la Pinacoteca Tosio Martinengo e delle emozioni che sperimentano durante la loro visita.

Ogni cluster verrà esaminato in modo approfondito, con un’analisi dettagliata delle sue peculiarità e dei tratti distintivi che lo contraddistinguono. Esploreremo le emozioni predominanti, il grado di coinvolgimento emotivo e altre caratteristiche rilevanti che emergono da ciascun gruppo di visitatori. Questo approfondimento ci permetterà di creare profili dettagliati dei visitatori e di comprendere meglio cosa li spinge a visitare il museo.

Inizieremo questo processo esaminando le etichette assegnate a ciascun cluster, fornendo così un quadro dettagliato delle diverse categorie di visitatori identificate.

Cluster 1 – I “Meravigliati”

Con un totale di 161 visitatori in questo cluster, è evidente che stiamo affrontando un gruppo significativo di individui che condividono esperienze emotive simili all’interno della Pinacoteca Tosio Martinengo.

Il primo cluster, denominato i “Meravigliati”, deve il suo nome alle emozioni predominanti riscontrare tra i visitatori appartenenti a questo gruppo. Questi individui, come si evince dalla Figura 4.7, vivono un’esperienza intensa

all'interno della Pinacoteca, caratterizzata da una profonda gioia e una sensazione di sorpresa mentre esplorano le opere d'arte esposte.

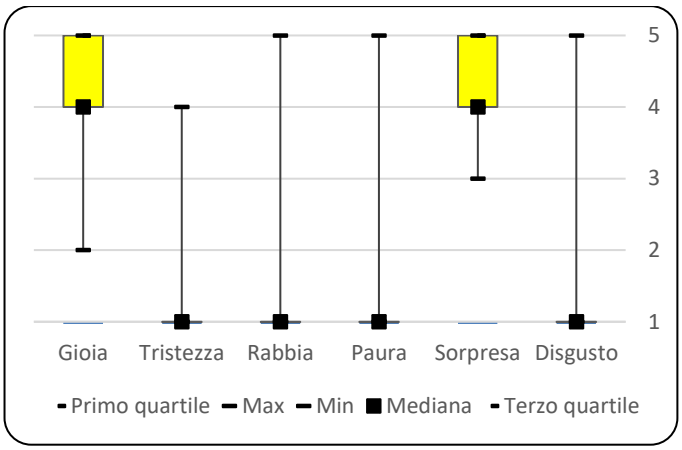


Figura 4.7 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

Dalla Figura 4.7 emerge un dato estremamente significativo: ben il 50% dei visitatori appartenenti a questo cluster ha riportato un'esperienza emotiva di notevole intensità, caratterizzata da livelli di gioia e sorpresa valutati con un'intensità pari a 4 o 5, corrispondenti alle categorie "Molto" e "Moltissimo". Questo significa che metà dei partecipanti ha vissuto un coinvolgimento profondo e una notevole sorpresa durante la visita alla Pinacoteca Tosio Martinengo.

Per comprendere meglio il contesto, è utile considerare la differenza interquartile, un concetto statistico fondamentale. Questo intervallo contiene il 50% delle osservazioni ordinate e si estende tra il primo e il terzo quartile, ognuno dei quali contiene il 25% delle osservazioni. Nel nostro caso, la differenza interquartile evidenzia la variabilità significativa nelle risposte dei visitatori, confermando l'ampia gamma di esperienze emozionali all'interno di questo cluster.

I dati rilevati trovano ulteriore conferma nell'analisi della Tabella 4.8. In particolare, è interessante osservare che la mediana relativa a queste due variabili è valutata a 4, indicando una percezione delle emozioni pari a "Molto", mentre lo scarto interquartile si estende tra le valutazioni 4 e 5, corrispondenti a "Molto" e "Moltissimo".

Questo risultato mette in luce in modo significativo l'entità delle emozioni di gioia e sorpresa sperimentate dai visitatori all'interno di questo cluster. La mediana a 4 suggerisce che la maggior parte dei partecipanti ha vissuto un'esperienza emotiva molto intensa, mentre lo scarto interquartile che si estende

fino al valore massimo di 5 evidenzia che alcuni hanno addirittura sperimentato una sorpresa estrema e una gioia molto intensa.

Questi dati offrono una visione dettagliata delle emozioni predominanti all'interno del cluster preso in esame, mettendo in luce quanto l'entusiasmo e la sorpresa sperimentati durante la visita alla Pinacoteca Tosio Martinengo siano stati vissuti in modo profondo dai visitatori.

Tabella 4.8 – Tabella dei valori quartili.
Fonte: nostre elaborazioni.

	Mediana	3° quartile	1° quartile	Scarto interquartile
Gioia	4	5	4	1
Tristezza	1	1	1	0
Rabbia	1	1	1	0
Paura	1	1	1	0
Sorpresa	4	5	4	1
Disgusto	1	1	1	0

Inoltre, l'ampiezza contenuta della differenza interquartile per entrambe le emozioni caratterizzanti questo cluster suggerisce una minore variabilità all'interno del gruppo e, di conseguenza, una maggiore coesione tra i visitatori. Questo concetto si basa sulla relazione tra la variabilità e l'ampiezza dell'intervallo interquartile, dove una minore ampiezza indica una maggiore omogeneità nelle risposte dei visitatori. In altre parole, i "Meravigliati" sembrano condividere profondamente queste emozioni, creando così un'esperienza museale uniforme e altamente coinvolgente.

Per quanto riguarda invece le altre emozioni, tra cui tristezza, rabbia, paura e disgusto, possiamo notare dalla Figura 4.7 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un'intensità per queste emozioni pari a "Per nulla". In altre parole, all'interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo risultato indica che le emozioni di tristezza, rabbia, paura e disgusto non giocano un ruolo significativo nell'esperienza emotiva del cluster dei "Meravigliati". Si può presumere che questi visitatori siano meno influenzati o coinvolti da tali emozioni durante la loro visita alla Pinacoteca, concentrandosi invece su sensazioni di gioia e meraviglia. Questo aspetto contribuisce a caratterizzare in modo distintivo questo cluster, sottolineando la centralità delle emozioni positive nella loro esperienza complessiva.

Il nome, i “Meravigliati”, è stato scelto proprio per catturare l’entusiasmo e la gioia che questi visitatori provano mentre esplorano il museo, insieme alla loro capacità di essere costantemente sorpresi e affascinati dalle opere d’arte.

Questo cluster rappresenta una parte preziosa della popolazione dei visitatori, poiché porta con sé una gioia contagiosa che può arricchire l’esperienza complessiva del museo. Le loro reazioni positive e il loro costante senso di meraviglia possono influenzare positivamente le strategie di marketing, consentendo di promuovere l’aspetto emozionale dell’esperienza museale per attirare visitatori simili.

Questi visitatori costituiscono dunque un gruppo di visitatori appassionati che possono diventare dei sostenitori entusiasti della Pinacoteca Tosio Martinengo, a condizione che l’esperienza museale continui a incantarli e a sorprenderli con la bellezza dell’arte.

Cluster 2 – I “Melanconici”

Il secondo cluster, identificato come i “Melanconici”, si distingue per l’intensità delle emozioni sperimentate durante la visita alla Pinacoteca. Questo gruppo di visitatori è caratterizzato da un’esperienza artistica ricca di emozioni profonde, con particolare enfasi sulla tristezza, accompagnata da una discreta presenza di gioia e sorpresa.

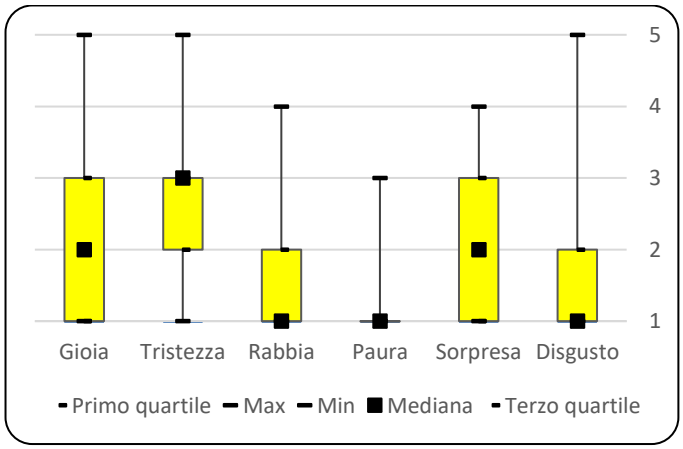


Figura 4.9 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

All’interno di questo cluster, emerge chiaramente un profilo distintivo dei visitatori, i quali sembrano essere profondamente colpiti dall’arte esposta, con una prevalenza di emozioni di tristezza che accompagnano la loro esperienza. La presenza di tristezza all’interno di questo cluster suggerisce che questi visitatori

siano in grado di vivere un profondo coinvolgimento con le opere d'arte, riflettendo su emozioni e temi complessi rappresentati nelle opere esposte presso la Pinacoteca. Questi visitatori possono essere profondamente colpiti dalle opere d'arte in modo riflessivo e contemplativo. La tristezza potrebbe derivare dalla comprensione profonda e dall'empatia nei confronti dei temi o delle emozioni rappresentate nelle opere esposte nella Pinacoteca.

Dalla Figura 4.9 emergono dati di notevole rilevanza che ci permettono di comprendere meglio l'esperienza emotiva dei visitatori all'interno di questo cluster. In particolare, è significativo notare che il 50% dei visitatori inclusi in questo gruppo ha riportato di aver vissuto un'esperienza emotiva caratterizzata da un livello di tristezza di discreta intensità. Questa tristezza è stata valutata con un'intensità compresa tra 2 e 3, rientrando quindi nelle categorie "Poco" e "Abbastanza". In altre parole, la metà dei partecipanti ha sperimentato una sensazione di tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

In aggiunta, è importante notare che la differenza interquartile associata a questa variabile è notevolmente contenuta. Questo indica che all'interno del cluster dei "Melanconici", c'è una minore dispersione nei livelli di tristezza sperimentati dai visitatori. In altre parole, i partecipanti a questo cluster condividono in modo significativo l'esperienza di emozioni, quali la tristezza, creando così un ambiente museale altamente coeso e coinvolgente.

Tuttavia, è interessante notare che, nonostante la predominanza della tristezza, permane una sottile ma riconoscibile percezione di gioia e sorpresa, aggiungendo così un ulteriore strato di complessità emotiva alla loro esperienza museale.

Nel dettaglio, dalla Figura 4.9 emerge che ben il 25% dei visitatori appartenenti a questo cluster ha riferito di provare emozioni di gioia e sorpresa valutati con un'intensità valutata tra 2 e 3, corrispondenti alle categorie "Poco" e "Abbastanza". Tuttavia, è interessante notare che la differenza interquartile associata a queste variabili presenta un'ampia dispersione. Ciò suggerisce che all'interno del cluster dei "Melanconici", esistono differenze significative nei livelli di gioia e sorpresa sperimentati dai visitatori, creando così una variabilità non indifferente all'interno di questa esperienza emotiva.

Il nome "Melanconici" è stato dunque scelto attentamente per riflettere la profondità delle emozioni sperimentate durante la visita. Questi visitatori sono in grado di affrontare emozioni intense e riflessive mentre si immergono nell'arte, traggono una comprensione significativa e una profonda riflessione dalle opere esposte.

Tali visitatori, portando con sé emozioni profonde e una capacità unica di affrontare l'arte in modo riflessivo e aperto, rappresentano un segmento prezioso della popolazione dei visitatori. Le loro esperienze possono infatti arricchire ulteriormente l'atmosfera emotiva della Pinacoteca e offrire spunti interessanti per la progettazione di mostre e strategie di coinvolgimento.

Cluster 3 – Gli "Spensierati"

Il terzo cluster, noto come gli "Spensierati", svela una prospettiva interessante sul profilo dei visitatori che lo compongono. Questi individui si contraddistinguono per l'esperienza artistica che vivono, caratterizzata principalmente dall'assenza di emozioni negative, in particolare la tristezza, e dalla presenza di emozioni positive, come la gioia e la sorpresa.

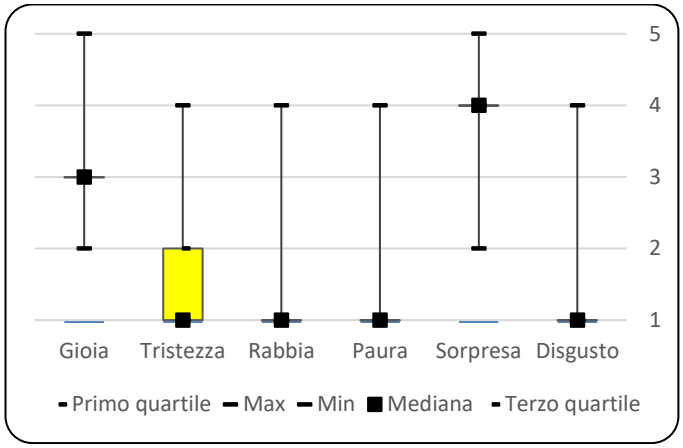


Figura 4.10 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

All'interno del Cluster 3, emerge chiaramente una notevole assenza di tristezza. Questo dato suggerisce che i visitatori all'interno di questo gruppo non sperimentano alcuna forma di tristezza durante la loro visita alla Pinacoteca Tosio Martinengo. Tale conclusione è ben supportata dai dati analizzati nella Figura 4.10, dove diventa evidente la totale assenza di tristezza tra i membri di questo cluster.

In particolare, è significativo notare che il 50% dei visitatori inclusi in questo cluster ha riportato di aver vissuto un'esperienza emotiva caratterizzata da un livello di tristezza valutato con un'intensità compresa tra 1 e 2, rientrando quindi nelle categorie "Per nulla" e "Poco". In altre parole, la metà dei partecipanti non ha sperimentato alcuna sensazione di tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

In aggiunta, l'analisi delle emozioni di gioia e sorpresa rivela un'interessante dinamica all'interno del nostro cluster. Dalla Figura 4.10 emerge chiaramente che la maggioranza dei visitatori ha riportato un'intensità per queste emozioni che si colloca nelle categorie "Abbastanza" e "Molto". Questo risultato è ulteriormente confermato dall'analisi della differenza interquartile, che rivela uno scarto interquartile schiacciato a valori rispettivamente di 3 e 4 per queste emozioni.

In termini più semplici, ciò significa che la maggior parte dei visitatori all'interno di questo cluster ha sperimentato una quantità considerevole di gioia e sorpresa durante la loro visita al museo. Le emozioni positive sembrano essere una parte significativa dell'esperienza complessiva di questo gruppo, contribuendo a creare un'atmosfera di apprezzamento e ammirazione mentre esplorano le opere d'arte esposte nella Pinacoteca Tosio Martinengo.

Per quanto riguarda invece le emozioni di rabbia, paura e disgusto, possiamo notare dalla Figura 4.10 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un'intensità per queste emozioni pari a "Per nulla". In altre parole, all'interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo risultato indica che le emozioni di rabbia, paura e disgusto non giocano un ruolo significativo nell'esperienza emotiva del cluster degli "Spensierati". Si può presumere che questi visitatori siano meno influenzati o coinvolti da tali emozioni durante la loro visita alla Pinacoteca, concentrandosi invece su sensazioni di tristezza, gioia e sorpresa. Questo aspetto contribuisce a caratterizzare in modo distintivo questo cluster.

L'appellativo "Spensierati" scelto per questo cluster cattura in modo eloquente l'esperienza dei visitatori che lo compongono. Queste persone possono immergersi nell'arte e nelle opere esposte nella Pinacoteca in uno stato di mente privo di preoccupazioni o tristezza. La loro visita è caratterizzata da un'atmosfera leggera e piacevole, dove l'arte viene vissuta in uno stato di serenità, senza ostacoli emotivi negativi.

È importante riconoscere che questi visitatori costituiscono un segmento significativo della popolazione di coloro che visitano la Pinacoteca. La loro presenza e l'esperienza positiva che vivono possono contribuire in modo sostanziale alla promozione e al successo della Pinacoteca. La disposizione "spensierata" di questi visitatori può influenzare positivamente l'atmosfera all'interno del museo, creando un ambiente accogliente e piacevole per tutti i

visitatori, che potranno godere appieno delle opere d'arte senza distrazioni negative contribuendo così a elevare l'esperienza museale complessiva.

Cluster 4 – gli "Stupiti"

Il quarto cluster, noto come gli "Stupiti", offre uno sguardo interessante sul profilo dei visitatori che lo compongono. Questo gruppo di individui si distingue per l'esperienza artistica che vicino quest'ultimi, e questa esperienza è notevolmente caratterizzata dall'assenza di emozioni intense, come tristezza, rabbia, paura o disgusto, durante la loro visita alla Pinacoteca Tosio Martinengo. Invece, l'emozione predominante all'interno di questo cluster sembra essere la sorpresa, almeno per una parte significativa dei visitatori.

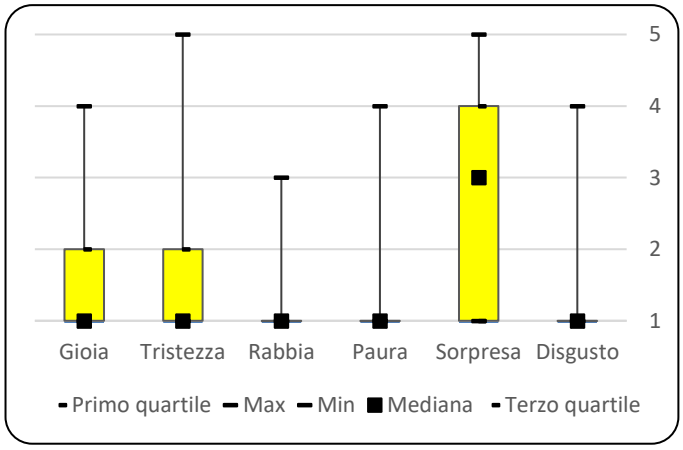


Figura 4.11 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

Questo risultato indica che le emozioni di rabbia, paura e disgusto non giocano un ruolo predominante nell'esperienza emotiva del cluster degli "Stupiti". Si può presumere che questi visitatori siano meno influenzati o coinvolti da tali emozioni durante la loro visita al museo, concentrandosi invece su sensazioni di sorpresa. La presenza prominente della sorpresa all'interno di questo cluster contribuisce in modo distintivo a caratterizzare il profilo emotivo dei suoi membri, distinguendoli dagli altri visitatori della Pinacoteca.

Dalla Figura 4.11 emergono infatti dati di notevole rilevanza che ci permettono di comprendere meglio l'esperienza emotiva dei visitatori all'interno di questo cluster. In particolare, è significativo notare che il 25% dei visitatori inclusi in questo gruppo ha riportato di aver vissuto un'esperienza emotiva caratterizzata da un livello di sorpresa di considerevole intensità. Questa sorpresa è stata valutata con un'intensità compresa tra 3 e 4, rientrando quindi nelle categorie "Abbastanza" e "Molto". In altre parole, un quarto dei visitatori ha sperimentato

una sensazione di tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

Tuttavia, è interessante notare che la differenza interquartile associata a questa variabile presenta un'ampia dispersione. Ciò suggerisce che all'interno del cluster degli "Stupiti", esistono differenze significative nei livelli di sorpresa sperimentati dai visitatori, creando così una variabilità non indifferente all'interno di questa esperienza emotiva.

Al contrario, è interessante notare che sia la gioia che la tristezza sono presenti in misura limitata all'interno di questo cluster, ma non emergono come emozioni predominanti. Come evidenziato nella Figura 4.11, il 50% dei visitatori appartenenti a questo cluster ha riportato di aver vissuto un'esperienza emotiva caratterizzata da un livello di gioia valutato con un'intensità compresa tra 1 e 2, corrispondenti alle categorie "Per nulla" e "Poco". Questo stesso trend si riscontra nella percezione della tristezza. In altre parole, la metà dei partecipanti non ha sperimentato alcuna sensazione significativa di gioia e/o tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

Ciò implica che, sebbene alcuni visitatori possano occasionalmente percepire tracce di gioia e tristezza in risposta a determinate opere d'arte, queste emozioni non hanno un ruolo rilevante nell'influenzare in modo significativo la loro esperienza complessiva all'interno della Pinacoteca.

Per quanto riguarda invece le emozioni di rabbia, paura e disgusto, possiamo notare dalla Figura 4.11 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un'intensità per queste emozioni pari a "Per nulla". In altre parole, all'interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo risultato sottolinea dunque che le emozioni di rabbia, paura e disgusto non hanno un impatto significativo nell'esperienza emotiva del cluster degli "Stupiti". Possiamo dedurre che questi visitatori non vengono influenzati o coinvolti da tali emozioni durante la loro visita alla Pinacoteca, preferendo concentrarsi invece su sensazioni di sorpresa. Questo aspetto contribuisce a caratterizzare in modo distintivo questo cluster.

Il nome gli "Stupiti" cattura quindi in modo accurato l'esperienza dei visitatori appartenenti a questo cluster. Questi individui sembrano affrontare l'arte con un atteggiamento di stupore e meraviglia, essendo particolarmente reattivi all'elemento sorprendente delle opere stesse. Questo atteggiamento riflette la loro

disposizione a essere colpiti e affascinati dall'arte in modo profondo, anche se altre emozioni come rabbia, paura o disgusto non emergono in modo rilevante durante la loro visita.

Questo segmento di visitatori porta con sé una prospettiva unica, poiché il loro focus è rivolto principalmente alla dimensione sorprendente e stimolante delle creazioni artistiche. Questa particolare inclinazione contribuisce in modo significativo a diversificare l'esperienza generale dei visitatori del museo, aggiungendo una prospettiva fresca e innovativa alla fruizione delle opere d'arte.

Cluster 5 – gli "Entusiasti"

Il quinto cluster, noto come gli "Entusiasti", offre un'interessante panoramica delle caratteristiche distintive dei visitatori in questo gruppo. Questi individui vivono un'esperienza artistica estremamente positiva all'interno della Pinacoteca Tosio Martinengo, caratterizzata da una profonda felicità, gioia costante e una continua sorpresa. Inoltre, è fondamentale evidenziare che l'assenza completa di sensazioni di tristezza durante la loro visita rappresenta un tratto distintivo e significativo del cluster degli "Entusiasti".

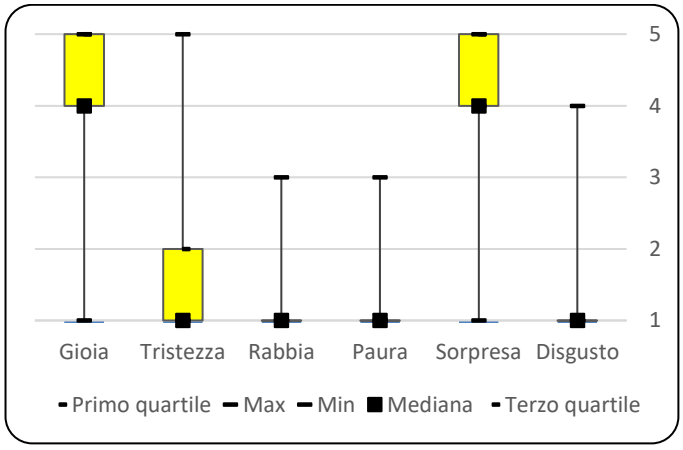


Figura 4.12 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

All'interno del Cluster 5, la gioia è un'emozione molto presente. I visitatori in questo gruppo sperimentano una profonda gioia mentre esplorano le opere d'arte esposte nella galleria. La loro risposta positiva all'arte è evidente e enfatizzata dalla presenza consistente della gioia.

La sorpresa rappresenta un altro elemento distintivo all'interno di questo cluster, suggerendo che questi visitatori rimangono costantemente sorpresi e affascinati dalle opere d'arte esposte.

In particolare, come si evince dalla Figura 4.12, è significativo notare che il 50% dei visitatori inclusi in questo cluster ha dichiarato di aver vissuto un'esperienza emotiva caratterizzata da livelli di gioia e sorpresa valutati con un'intensità compresa tra 4 e 5, rientrando quindi nelle categorie "Molto" e "Moltissimo". In altre parole, la metà dei partecipanti ha sperimentato una profonda gioia e una costante sorpresa durante la loro visita alla Pinacoteca Tosio Martinengo.

Questo risultato mette in luce in modo significativo l'entità delle emozioni sperimentate dai visitatori all'interno di questo cluster. Con una mediana a 4 per entrambe le emozioni in questione, emerge chiaramente che la maggior parte dei partecipanti ha vissuto un'esperienza emotiva estremamente intensa, come sei può osservare dalla Tabella 4.13. Inoltre, lo scarto interquartile che si estende fino al valore massimo di 5 sottolinea che alcuni di loro hanno persino sperimentato una sorpresa estrema e una gioia molto intensa durante la visita alla Pinacoteca.

Tabella 4.13 – Tabella dei valori quartili.
Fonte: nostre elaborazioni.

	Mediana	3° quartile	1° quartile	Scarto interquartile
Gioia	4	5	4	1
Tristezza	1	2	1	1
Rabbia	1	1	1	0
Paura	1	1	1	0
Sorpresa	4	5	4	1
Disgusto	1	1	1	0

All'interno del Cluster 3, emerge inoltre in modo evidente una totale assenza di tristezza. Questo suggerisce che i membri di questo gruppo non sperimentano alcuna forma di tristezza durante la loro visita al museo. Tale conclusione è solidamente supportata dai dati analizzati nella Figura 4.12, i quali chiaramente evidenziano che nessun visitatore di questo cluster ha riportato livelli significativi di tristezza.

È rilevante notare che il 50% dei membri di questo cluster ha dichiarato di aver vissuto un'esperienza emotiva caratterizzata da un livello di tristezza valutato tra 1 e 2, rientrando quindi nelle categorie "Per nulla" e "Poco". In altre parole, la metà dei partecipanti non ha sperimentato alcuna sensazione di tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

Inoltre, va notato che l'ampiezza contenuta della differenza interquartile per tutte e tre le emozioni considerate in questo cluster, suggerisce una minore variabilità tra i visitatori e, di conseguenza, una maggiore coesione all'interno del gruppo. In

altre parole, gli “Entusiasti” sembrano condividere in modo profondo queste emozioni, creando così un’esperienza museale uniforme e altamente coinvolgente.

Il nome gli “Entusiasti” cattura dunque l’esperienza di visitatori che, mentre esplorano l’arte, sono costantemente sorpresi e affascinati dalla bellezza, dall’originalità e dalla creatività delle opere. La predominanza di gioia e sorpresa unita all’assenza La partecipazione attiva degli “Entusiasti” contribuisce in modo significativo all’atmosfera positiva e alla vitalità della Pinacoteca, creando un ambiente accogliente e coinvolgente per tutti i visitatori. La loro disposizione entusiastica può influenzare positivamente il modo in cui le opere d'arte vengono percepite e apprezzate da chi visita il museo.

Cluster 6 – gli “Indifferenti”

Il sesto cluster, noto come gli “Indifferenti”, rappresenta i visitatori che vivono un’esperienza emotiva relativamente moderata durante la loro visita alla Pinacoteca Tosio Martinengo. In questo gruppo, le emozioni positive, come la gioia e la sorpresa, sono presenti, ma in misura molto limitata. I membri di questo gruppo non sperimentano quindi emozioni estremamente intense o prevalenti durante la loro visita al museo. Questo aspetto distingue quindi gli “Indifferenti” dagli altri cluster in cui le emozioni sono più forti e predominanti.

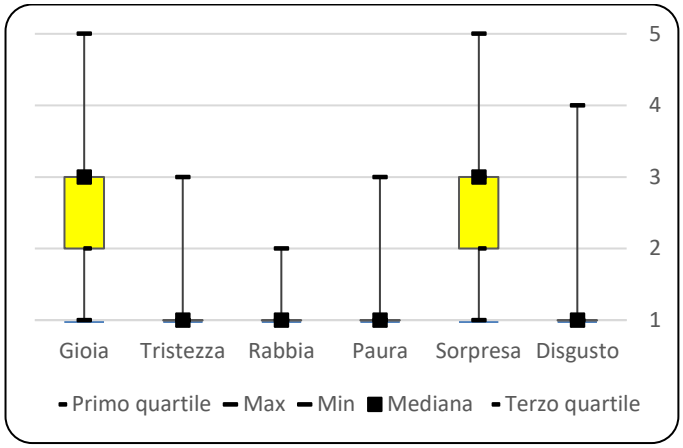


Figura 4.14 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

All’interno del Cluster 6, la gioia è presente in misura discreta. Questo suggerisce che questi visitatori sperimentano una felicità discreta durante la loro visita alla galleria, senza essere completamente sopraffatti dalle emozioni positive. La loro reazione emotiva positiva è equilibrata e moderata, contraddistinta da una gioia misurata.

La sorpresa, pur essendo anch'essa presente all'interno di questo cluster, non raggiunge livelli particolarmente elevati. Ciò indica che questi visitatori sono sì aperti a nuove esperienze, ma senza manifestare una sorpresa eccessiva o intensa.

In particolare, come evidenziato nella Figura 4.14, è significativo notare che il 50% dei visitatori appartenenti a questo cluster ha riportato di aver vissuto un'esperienza emotiva caratterizzata da livelli di gioia e sorpresa valutati con un'intensità compresa tra 2 e 3, rientrando quindi nelle categorie "Poco" e "Abbastanza". In altre parole, la metà dei partecipanti ha sperimentato una discreta gioia e una sorpresa moderata durante la loro visita alla Pinacoteca Tosio Martinengo.

Guardando più attentamente ai dati nella Tabella 4.15, emerge un quadro interessante all'interno del Cluster 6. La mediana a 3 per le emozioni di gioia e sorpresa sottolinea chiaramente che la maggior parte dei partecipanti ha vissuto un'esperienza emotiva di intensità moderata durante la loro visita alla Pinacoteca Tosio Martinengo. Questo significa che questi visitatori hanno sperimentato una felicità e una sorpresa bilanciate, senza essere completamente sopraffatti dalle emozioni positive. La loro risposta emotiva è stata misurata e controllata, caratterizzata da una gioia discreta e una sorpresa di intensità moderata.

Tabella 4.15 – Tabella dei valori quartili.
Fonte: nostre elaborazioni.

	Mediana	3° quartile	1° quartile	Scarto interquartile
Gioia	3	3	2	1
Tristezza	1	1	1	0
Rabbia	1	1	1	0
Paura	1	1	1	0
Sorpresa	3	3	2	1
Disgusto	1	1	1	0

Per quanto riguarda invece le emozioni negative di tristezza, rabbia, paura e disgusto, possiamo notare dalla Figura 4.14 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un'intensità per queste emozioni pari a "Per nulla". In altre parole, all'interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo nome, gli "Indifferenti", cattura dunque l'esperienza di visitatori che apprezzano l'arte in modo moderato. La loro reazione emotiva discretamente positiva è equilibrata e non estrema. La mancanza di emozioni negative come

tristezza, rabbia o disgusto suggerisce che il loro approccio all'arte è quindi prevalentemente positivo.

Sebbene questo cluster possa sembrare meno coinvolto emotivamente rispetto ad altri, è comunque importante considerare che questi visitatori contribuiscono al mosaico delle esperienze dei visitatori della Pinacoteca. La loro reazione moderata alla visita della galleria può riflettere una prospettiva più equilibrata o una disposizione a vivere l'arte in modo più calmo, contribuendo così a creare un'atmosfera complessivamente positiva senza emozioni eccessivamente intense.

Cluster 7 – i "Contemplativi"

Il settimo cluster, noto come i "Contemplativi", offre una prospettiva delle caratteristiche dei visitatori che lo compongono. Questi individui si distinguono per la loro esperienza artistica, che si discosta dalle emozioni tradizionali come la gioia, ma si concentra invece su una moderata sorpresa e tristezza. Mentre visitano la Pinacoteca, questi partecipanti sembrano inclini a riflettere e contemplare le opere d'arte, abbracciando un'ampia gamma di emozioni complesse anziché cercare una gioia immediata.

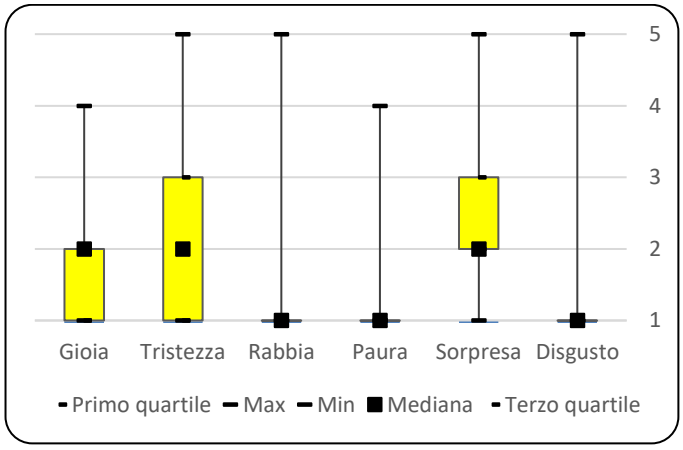


Figura 4.16 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

Questo cluster rappresenta quindi un gruppo di visitatori che possono apprezzare l'arte in un modo più profondo e riflessivo. La loro esperienza non è dominata dalla gioia, ma dalla capacità di immergersi nell'arte, connettendosi con emozioni più sfumate e complesse, come la tristezza e la sorpresa.

In particolare, come evidenziato nella Figura 4.16, è rilevante osservare che il 25% dei visitatori appartenenti a questo cluster ha segnalato di aver vissuto un'esperienza emotiva caratterizzata da un livello di tristezza valutato con

un’intensità compresa tra 2 e 3, rientrando quindi nelle categorie “Poco” e “Abbastanza”. In altre parole, un quarto dei visitatori ha sperimentato una moderata tristezza durante la loro visita alla Pinacoteca Tosio Martinengo.

Una situazione abbastanza analoga emerge anche per quanto riguarda la sorpresa. Come evidente dalla Figura 4.16, il 50% dei visitatori inclusi in questo cluster ha riportato di aver vissuto un’esperienza emotiva caratterizzata da un livello di sorpresa valutato con un’intensità compresa tra 2 e 3, rientrando quindi nelle categorie “Poco” e “Abbastanza”. Ciò significa che la metà dei partecipanti ha avvertito una moderata sensazione di sorpresa durante la loro visita alla Pinacoteca Tosio Martinengo.

In contrasto, osserviamo che il 50% dei visitatori all’interno del cluster dei “Contemplativi” ha dichiarato di aver sperimentato un’esperienza emotiva caratterizzata da un livello di gioia valutato con un’intensità compresa tra 1 e 2, rientrando quindi nelle categorie “Per nulla” e “Poco”. In altre parole, la metà dei partecipanti non ha avvertito alcuna sensazione significativa di gioia durante la loro visita alla Pinacoteca Tosio Martinengo.

Per quanto riguarda invece le emozioni di rabbia, paura e disgusto, possiamo notare dall’analisi della Figura 4.16 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un’intensità per queste emozioni pari a “Per nulla”. In altre parole, all’interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo risultato indica che le emozioni di rabbia, paura e disgusto non giocano un ruolo significativo nell’esperienza emotiva del cluster dei “Contemplativi”. Si può presumere che questi visitatori siano meno influenzati o coinvolti da tali emozioni durante la loro visita alla Pinacoteca, concentrandosi invece su sensazioni di tristezza e sorpresa. Questo aspetto contribuisce a caratterizzare in modo distintivo questo cluster.

Il nome i “Contemplativi” riflette dunque l’esperienza di visitatori che apprezzano l’arte in modo sereno e riflessivo. La mancanza di reazioni emotive intense suggerisce che il loro approccio all’arte è caratterizzato da una calma riflessione. Questi visitatori potrebbero concentrarsi sulla contemplazione e sulla riflessione profonda durante la loro visita alla galleria. La loro esperienza nella Pinacoteca è caratterizzata da un’atmosfera tranquilla e pacata, il che potrebbe tradursi in un’esperienza più riflessiva e oggettiva rispetto alle emozioni effervescenti che altri visitatori potrebbero sperimentare.

Questa prospettiva unica contribuisce quindi a diversificare l'esperienza complessiva dei visitatori all'interno della Pinacoteca, arricchendo il dialogo emotivo che l'arte può ispirare.

Cluster 8 – i "Rilassati"

Il gruppo identificato come i "Rilassati" riflette le caratteristiche distintive dei visitatori all'interno di questo gruppo. Questi individui vivono un'esperienza artistica in cui la gioia è una componente centrale, mentre la sorpresa è presente con un'intensità moderata. Questi visitatori si godono quindi la visita alla Pinacoteca con un senso di gioia ben definito, ma senza essere facilmente sorpresi o affascinati dalle opere d'arte.

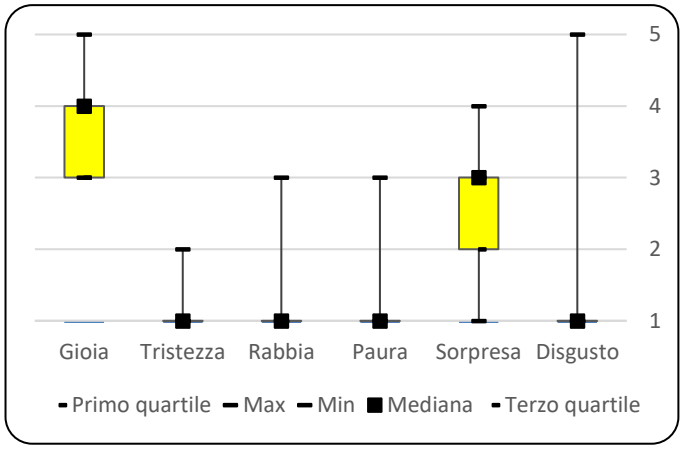


Figura 4.17 – Cluster profile boxplot.
Fonte: nostre elaborazioni.

Questa dinamica di emozioni suggerisce che i "Rilassati" possono apprezzare l'arte in modo rilassato e tranquillo, senza essere travolti da reazioni emotive intense. La presenza della gioia indica che hanno una reazione positiva all'arte, ma la moderata sorpresa suggerisce che la loro esperienza non è caratterizzata da forti sbalzi emotivi.

All'interno del Cluster 8, la gioia è dunque una delle emozioni dominanti. Questi visitatori sperimentano una notevole felicità e soddisfazione mentre esplorano le opere d'arte esposte nella galleria. La gioia è presente in misura significativa, evidenziando una reazione estremamente positiva all'arte.

In particolare, come evidenziato nella Figura 4.17, è rilevante osservare che il 50% dei visitatori appartenenti a questo cluster ha segnalato di aver vissuto un'esperienza emotiva caratterizzata da un livello di gioia valutato con un'intensità compresa tra 3 e 4, rientrando quindi nelle categorie "Abbastanza" e

“Molto”. In altre parole, la metà dei partecipanti ha sperimentato un’elevata gioia durante la loro visita alla Pinacoteca Tosio Martinengo.

Nel contesto della sorpresa, è invece rilevante notare che tra i visitatori appartenenti a questo cluster, il 50% ha segnalato di aver sperimentato una sorpresa moderata. Questo si traduce nel fatto che metà dei partecipanti ha avuto una reazione emotiva caratterizzata da un grado di sorpresa valutato come “Poco” o “Abbastanza” durante la visita alla Pinacoteca Tosio Martinengo, come indicato nella Figura 4.17.

Per quanto riguarda invece le emozioni di rabbia, paura e disgusto, possiamo notare dall’analisi della Figura 4.17 che lo scarto interquartile è schiacciato a un valore di 1, il che suggerisce che la maggioranza dei visitatori ha riportato un’intensità per queste emozioni pari a “Per nulla”. In altre parole, all’interno di questo cluster, queste emozioni sembrano essere praticamente assenti o, al massimo, percepite in modo marginale.

Questo risultato indica che le emozioni di rabbia, paura e disgusto non giocano un ruolo significativo nell’esperienza emotiva del cluster dei “Rilassati”. Si può presumere che questi visitatori siano meno influenzati o coinvolti da tali emozioni durante la loro visita alla Pinacoteca, concentrandosi invece su sensazioni di tristezza e sorpresa. Questo aspetto contribuisce a caratterizzare in modo distintivo questo cluster.

Il nome i “Rilassati” riflette dunque con precisione l’esperienza di visitatori che, mentre esplorano l’arte, si lasciano costantemente sorprendere dalla gioia significativa che provano nel contemplare le opere d’arte. La mancanza di emozioni negative suggerisce chiaramente che il loro approccio all’arte è prevalentemente positivo.

Questo segmento di visitatori potrebbe dunque preferire un approccio più calmo e riflessivo alla fruizione dell’arte, concentrandosi sulla bellezza intrinseca delle opere e godendo di un’esperienza museale rilassante. La loro disposizione “rilassata” può contribuire a creare un’atmosfera piacevole e serena all’interno della galleria, rendendo l’arte più accessibile per tutti i visitatori.

4.4.2 Heatmap

Dopo aver assegnato le etichettati ai cluster, l’analisi delle risposte ottenute in riferimento alla domanda 3 del questionario ha portato alla creazione di una “*heatmap*” o “mappa del calore”, la cui visualizzazione è consultabile nella Figura 4.18. Questa rappresentazione visuale offre un’analisi dettagliata della distribuzione assoluta delle risposte dei visitatori negli 8 cluster in relazione alle 6

emozioni principali in esame (gioia, tristezza, rabbia, paura, sorpresa, disgusto) e alle loro 5 modalità di percezione (per nulla, poco, abbastanza, molto, moltissimo). Questa visualizzazione fornisce un quadro dettagliato delle reazioni emotive dei visitatori all'interno di ciascun cluster durante la loro visita alla Pinacoteca Tosio Martinengo.

Una *heatmap* è una rappresentazione visuale dei dati che utilizza una scala di colori per evidenziare le tendenze e le differenze nei valori tra varie categorie o variabili.

Nella rappresentazione visuale, ogni cella è colorata in base al valore assoluto che contiene. Utilizziamo una variazione di colori che va da tonalità più intense a tonalità più sfumate. Ad esempio, una cella con un valore percentuale elevato avrà un colore più intenso, mentre una con un valore più basso sarà rappresentata con una tonalità più chiara.

La *heatmap* riflette il numero di visitatori di ciascun cluster che ha risposto a ciascuna delle possibili combinazioni di emozioni e modalità. Ad esempio, possiamo osservare quanti visitatori nel Cluster 1 hanno riportato di percepire "Molto" la gioia, quelli nel Cluster 2 che hanno riportato di percepire "Abbastanza" la tristezza, e così via.

Questa *heatmap* è uno strumento potente per identificare chiaramente le tendenze e le differenze significative nelle risposte emotive dei visitatori tra i vari cluster. Ciò consente alla Pinacoteca di comprendere meglio come ciascun gruppo di visitatori interagisce con le opere d'arte esposte e può informare la progettazione di mostre future, programmi educativi mirati e strategie di coinvolgimento per soddisfare le esigenze e le preferenze specifiche di ciascun cluster.

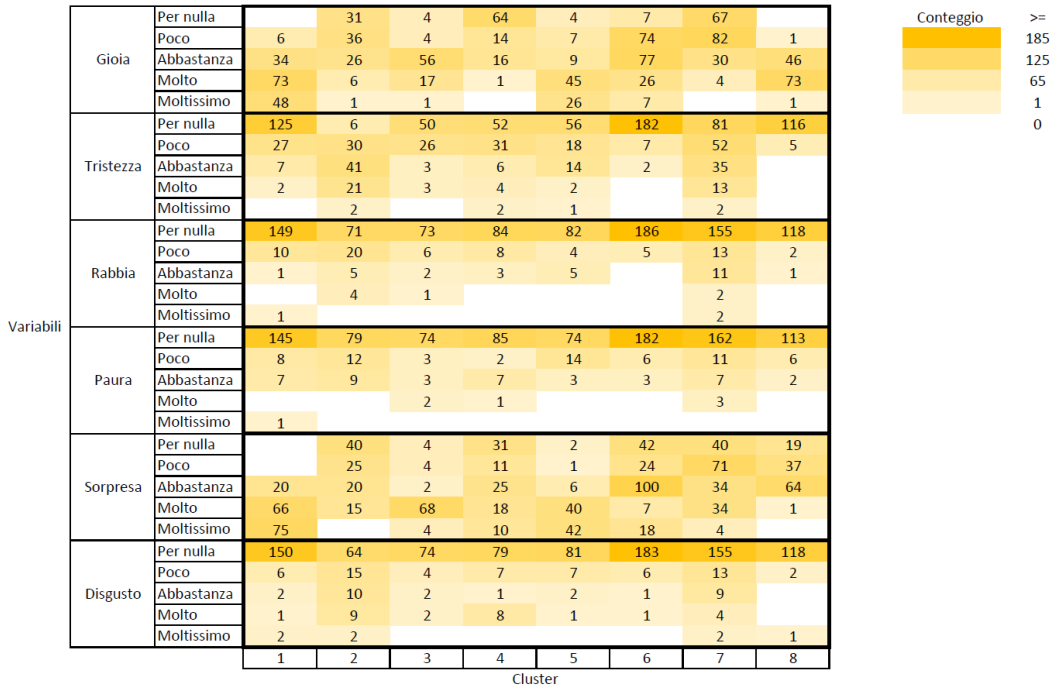


Figura 4.18 – Composizione assoluta dei cluster.
Fonte: nostre elaborazioni.

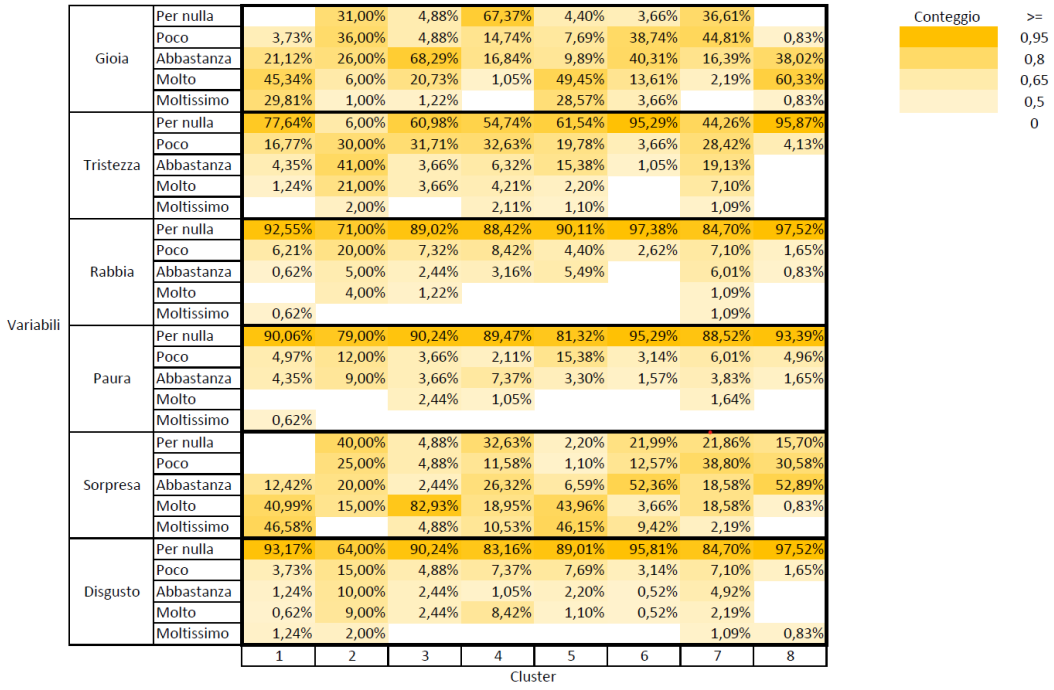
Questa graduazione di colori rende immediatamente evidenti le differenze significative nelle reazioni emotive dei vari cluster. Inoltre, facilita la comprensione delle tendenze generali, permettendo agli osservatori di individuare rapidamente le emozioni predominanti in ogni cluster.

Inoltre, l'uso di una scala di colori completa crea una rappresentazione visivamente coinvolgente delle dinamiche emotive all'interno dei cluster. Questo approccio non solo rende la *heatmap* più leggibile e interpretabile, ma aggiunge anche un elemento visivo coinvolgente che può catturare l'attenzione degli spettatori, rendendo l'analisi delle emozioni un'esperienza visivamente appagante.

Successivamente, abbiamo realizzato una seconda *heatmap* che presenta stavolta la distribuzione relativa delle risposte dei visitatori negli 8 cluster rispetto alle 6 emozioni principali (gioia, tristezza, rabbia, paura, sorpresa, disgusto) con le loro 5 modalità di percezione (per nulla, poco, abbastanza, molto, moltissimo). Questa visualizzazione, resa visibile nella Figura 4.19, offre una rappresentazione più dettagliata delle proporzioni di emozioni all'interno di ciascun cluster, consentendo una migliore comprensione delle dinamiche emotive dei visitatori.

La *heatmap* della distribuzione relativa indica quanto ogni combinazione di emozione e modalità contribuisca alle reazioni emotive complessive di ciascun cluster. Questo approccio fornisce una chiara rappresentazione delle emozioni

predominanti in ciascun gruppo di visitatori e come queste differiscono tra i vari cluster.



Variabili		1	2	3	4	5	6	7	8
Gioia	Per nulla		31,00%	4,88%	67,37%	4,40%	3,66%	36,61%	
	Poco	3,73%	36,00%	4,88%	14,74%	7,69%	38,74%	44,81%	0,83%
	Abbastanza	21,12%	26,00%	68,29%	16,84%	9,89%	40,31%	16,39%	38,02%
	Molto	45,34%	6,00%	20,73%	1,05%	49,45%	13,61%	2,19%	60,33%
	Moltissimo	29,81%	1,00%	1,22%		28,57%	3,66%		0,83%
Tristezza	Per nulla	77,64%	6,00%	60,98%	54,74%	61,54%	95,29%	44,26%	95,87%
	Poco	16,77%	30,00%	31,71%	32,63%	19,78%	3,66%	28,42%	4,13%
	Abbastanza	4,35%	41,00%	3,66%	6,32%	15,38%	1,05%	19,13%	
	Molto	1,24%	21,00%	3,66%	4,21%	2,20%		7,10%	
	Moltissimo		2,00%		2,11%	1,10%		1,09%	
Rabbia	Per nulla	92,55%	71,00%	89,02%	88,42%	90,11%	97,38%	84,70%	97,52%
	Poco	6,21%	20,00%	7,32%	8,42%	4,40%	2,62%	7,10%	1,65%
	Abbastanza	0,62%	5,00%	2,44%	3,16%	5,49%		6,01%	0,83%
	Molto		4,00%	1,22%				1,09%	
	Moltissimo	0,62%						1,09%	
Paura	Per nulla	90,06%	79,00%	90,24%	89,47%	81,32%	95,29%	88,52%	93,39%
	Poco	4,97%	12,00%	3,66%	2,11%	15,38%	3,14%	6,01%	4,96%
	Abbastanza	4,35%	9,00%	3,66%	7,37%	3,30%	1,57%	3,83%	1,65%
	Molto			2,44%	1,05%			1,64%	
	Moltissimo	0,62%							
Sorpresa	Per nulla		40,00%	4,88%	32,63%	2,20%	21,99%	21,86%	15,70%
	Poco		25,00%	4,88%	11,58%	1,10%	12,57%	38,80%	30,58%
	Abbastanza	12,42%	20,00%	2,44%	26,32%	6,59%	52,36%	18,58%	52,89%
	Molto	40,99%	15,00%	82,93%	18,95%	43,96%	3,66%	18,58%	0,83%
	Moltissimo	46,58%		4,88%	10,53%	46,15%	9,42%	2,19%	
Disgusto	Per nulla	93,17%	64,00%	90,24%	83,16%	89,01%	95,81%	84,70%	97,52%
	Poco	3,73%	15,00%	4,88%	7,37%	7,69%	3,14%	7,10%	1,65%
	Abbastanza	1,24%	10,00%	2,44%	1,05%	2,20%	0,52%	4,92%	
	Molto	0,62%	9,00%	2,44%	8,42%	1,10%	0,52%	2,19%	
	Moltissimo	1,24%	2,00%					1,09%	0,83%

Figura 4.19 – Composizione percentuale dei cluster.
Fonte: nostre elaborazioni.

La *heatmap* risulta dunque uno strumento prezioso in quanto offre una panoramica chiara e visuale delle emozioni predominanti e delle sfumature emotive all'interno di ciascun cluster e consente alla Pinacoteca di adattare le proprie strategie in modo più preciso per soddisfare le esigenze e le preferenze dei visitatori. Questa rappresentazione visuale aiuta a creare un'esperienza museale più ricca e coinvolgente per tutti i visitatori, basata sulla comprensione delle reazioni emotive dei visitatori e sulla loro distribuzione tra i diversi cluster.

4.4.3 Descrizione dei cluster

Dopo aver raggruppato i 1.024 partecipanti in 8 cluster in base alle loro risposte, abbiamo proseguito con un'analisi dettagliata delle caratteristiche demografiche che delineano distintamente ciascun cluster. Nelle Tabelle 4.20 e 4.21 presentiamo un riassunto dei risultati emersi da questa analisi, evidenziando le principali differenze tra i cluster.

Tabella 4.20 – Valori assoluti per i Cluster 1-4, Sesso e Fascia d'Età.
Fonte: nostre elaborazioni.

Cluster	1		2		3		4	
Sesso	Femmina	104	Femmina	57	Femmina	50	Femmina	54
	Maschio	57	Maschio	43	Maschio	32	Maschio	41
Fascia d'età	15-86		15-78		15-76		15-85	
Numero soggetti	161		100		82		95	

Tabella 4.21 – Valori assoluti per i Cluster 5-8, Sesso e Fascia d'Età. Fonte: nostre elaborazioni.

Cluster	5		6		7		8	
Sesso	Femmina	50	Femmina	101	Femmina	98	Femmina	74
	Maschio	41	Maschio	90	Maschio	85	Maschio	47
Fascia d'età	16-82		15-87		15-85		15-72	
Numero soggetti	91		191		183		121	

È evidente che in ogni cluster prevale il genere femminile rispetto a quello maschile. Ecco i dettagli:

- Per il Cluster 1, il 64,60 % dei partecipanti è di genere femminile, mentre il 35,40% è di genere maschile. Questi risultati sono basati su un totale di 161 partecipanti in questo cluster.
- Nel Cluster 2, il 57% dei partecipanti è di genere femminile, mentre il 43% è di genere maschile. Questi dati provengono da un campione di 100 partecipanti.
- Per il Cluster 3, il 60,98% è di genere femminile, mentre il 39,02% è di genere maschile. Questi risultati sono basati su un totale di 82 partecipanti.
- Nel Cluster 4, il 56,84% è di genere femminile, mentre il 43,16% è di genere maschile. Questi dati provengono da un campione di 95 partecipanti.
- Per il Cluster 5, il 54,95 % dei partecipanti è di genere femminile, mentre il 45,05% è di genere maschile. Questi risultati sono basati su un totale di 91 partecipanti in questo cluster.
- Nel Cluster 6, il 52,88% dei partecipanti è di genere femminile, mentre il 47,12% è di genere maschile. Questi dati provengono da un campione di 191 partecipanti.
- Per il Cluster 7, il 53,55% è di genere femminile, mentre il 46,45% è di genere maschile. Questi risultati sono basati su un totale di 183 partecipanti.
- Infine, nel Cluster 8 il 61,16 % è di genere femminile, mentre il 38,84% è di genere maschile. Questi dati provengono da un campione di 121 partecipanti.

Questi risultati mostrano chiaramente la predominanza del genere femminile in tutti i cluster considerati, con percentuali variabili tra i diversi gruppi di rispondenti.

Procediamo ora ad analizzare nello specifico la ripartizione del sesso in funzione di ogni cluster. I risultati di questa analisi sono chiaramente visualizzati nella Figura 4.22.

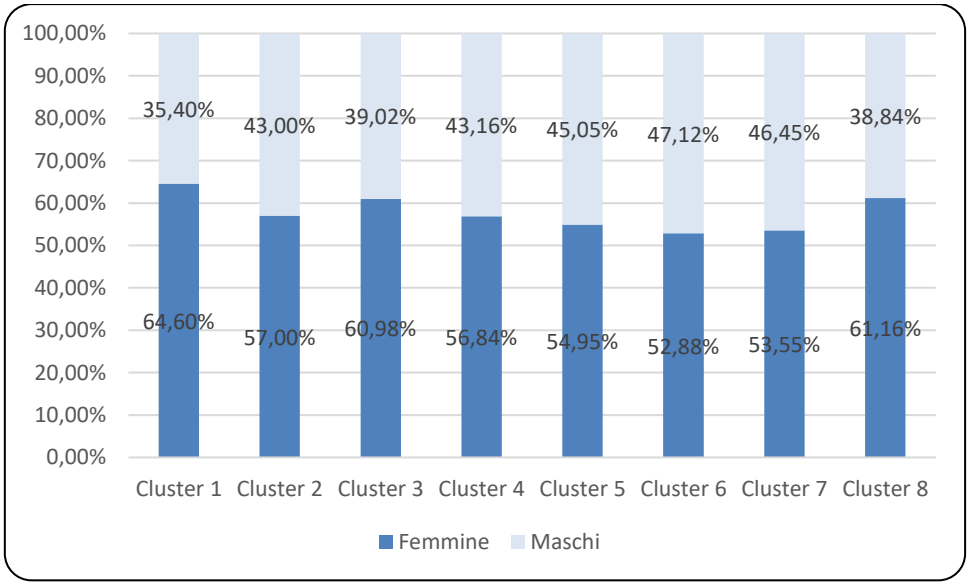


Figura 4.22 – Valori percentuali per Cluster, Sesso.
Fonte: nostre elaborazioni.

Nei Cluster 1, 3 e 8, notiamo una significativa predominanza delle partecipanti di genere femminile, con una percentuale che supera rispettivamente di circa 7, 3,5, e 4 punti percentuali la distribuzione generale nel campione, che si attesta al 57,42%. Di contro, la composizione di genere maschile in questi cluster è inferiore rispetto alla distribuzione generale nel campione. Questa differenza sottolinea una caratteristica distintiva di questi gruppi, in cui gli uomini sono meno rappresentati rispetto al campione complessivo.

Per quanto riguarda i Cluster 2 e 4, la composizione per il genere femminile è sostanzialmente in linea con la distribuzione generale nel campione, con una leggera differenza negativa di circa 40 e 60 centesimi di punti percentuali rispetto al campione complessivo. Ciò significa che in questi cluster la presenza di donne è allineata alla distribuzione del campione, senza variazioni significative.

Infine, nei Cluster 5, 6 e 7, notiamo una maggiore rappresentatività del genere maschile rispetto alla composizione generale del campione. In questi cluster, il genere maschile ha una rappresentatività superiore di circa 2,5, 5,5 e 4 punti

percentuali rispetto alla distribuzione generale nel campione, che si attesta al 42,58%. Di conseguenza, in questi cluster gli uomini sono in proporzione più rappresentativi rispetto alle donne, che risultano dunque essere meno rappresentate rispetto alla proporzione del campione complessivo.

Questa analisi dettagliata evidenzia le differenze nella distribuzione di genere tra i vari cluster e sottolinea come ciascun cluster presenti una composizione di genere unica rispetto alla distribuzione generale nel campione, con alcune variazioni più o meno significative.

Dopo aver esaminato con attenzione le differenze nella distribuzione di genere, procediamo ora ad analizzare la ripartizione per fasce d'età all'interno di ciascun cluster. I risultati di questa analisi sono chiaramente visualizzati nella Figura 4.23.

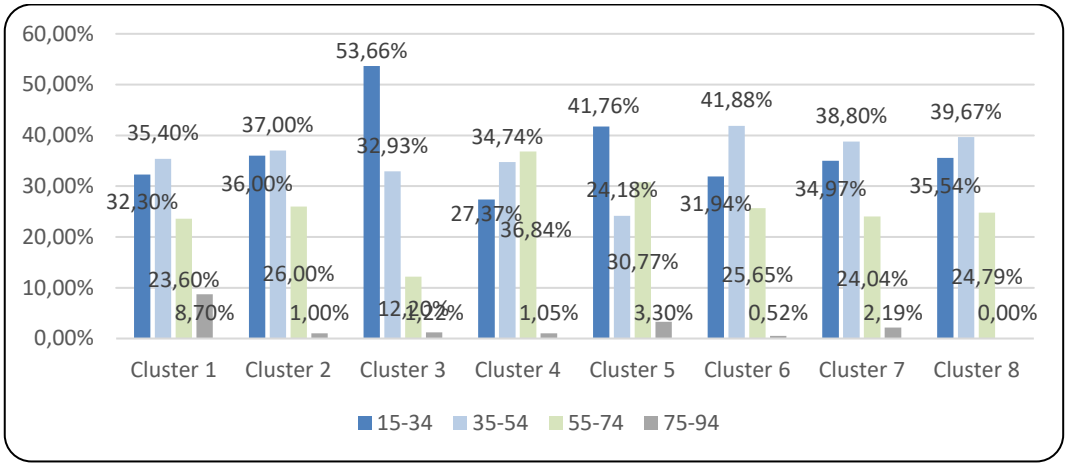


Figura 4.23 – Valori percentuali per Cluster, Fascia d'Età.
Fonte: nostre elaborazioni.

Esaminando la distribuzione per fasce d'età all'interno di ciascun cluster, notiamo:

- Nel Cluster 1, il 35,40% dei partecipanti appartiene alla fascia d'età compresa tra i 35 e i 54 anni, il 32,30% rientra nella fascia d'età compresa tra i 15 e i 34 anni, il 23,60% ha un'età compresa tra i 55 e i 74 anni, e infine l'8,70% ha più di 75 anni.
- Per quanto riguarda il Cluster 2, il 37,00% dei partecipanti è nella fascia d'età tra i 35 e i 54 anni, il 36,00% rientra nella fascia d'età tra i 15 e i 34 anni, il 26,00% ha un'età compresa tra i 55 e i 74 anni, mentre solo un 1,00% appartiene alla fascia d'età superiore ai 75 anni.

- Nel Cluster 3, il 53,66% dei partecipanti ha un'età compresa tra i 15 e i 34 anni, il 32,93% è nella fascia d'età tra i 35 e i 54 anni, il 12,20% ha un'età compresa tra i 55 e i 74 anni e, infine, l'1,22% ha più di 75 anni.
- Nel Cluster 4, il 37,74% dei partecipanti appartiene alla fascia d'età compresa tra i 35 e i 54 anni, il 36,84% rientra nella fascia d'età tra i 55 e i 74 anni, il 27,37% ha un'età compresa tra i 15 e i 34 anni e, infine, l'1,05% ha più di 75 anni.
- Nel Cluster 5, il 41,76% dei partecipanti appartiene alla fascia d'età compresa tra i 15 e i 34 anni, il 30,77% rientra nella fascia d'età compresa tra i 55 e i 74 anni, il 24,18% ha un'età compresa tra i 55 e i 74 anni, e infine il 3,30% ha più di 75 anni.
- Per quanto riguarda il Cluster 6, il 41,88% dei partecipanti è nella fascia d'età tra i 35 e i 54 anni, il 31,94% rientra nella fascia d'età tra i 15 e i 34 anni, il 25,65% ha un'età compresa tra i 55 e i 74 anni, mentre solo uno 0,52% appartiene alla fascia d'età superiore ai 75 anni.
- Per quanto riguarda il Cluster 7, il 38,80% dei partecipanti ha un'età compresa tra i 35 e i 54 anni, il 34,97% è nella fascia d'età tra i 15 e i 34 anni, il 24,04% ha un'età compresa tra i 55 e i 74 anni e, infine, il 2,19% ha più di 75 anni.
- Infine, nel Cluster 8, il 39,67% dei partecipanti appartiene alla fascia d'età compresa tra i 35 e i 54 anni, il 35,54% rientra nella fascia d'età tra i 15 e i 34 anni, il 24,79% ha un'età compresa tra i 55 e i 74 anni, mentre nessuno appartiene alla fascia d'età superiore ai 75 anni.

Nello specifico, per quanto riguarda i Cluster 3 e 5, i partecipanti appartenenti alla fascia d'età compresa tra i 15 e i 34 anni hanno una rappresentatività significativamente superiore rispetto alla composizione generale del campione, con un aumento di circa 18 e 6 punti percentuali rispetto al totale del campione, che è del 35,55%.

Tuttavia, una situazione intrigante emerge quando analizziamo il Cluster 1. Da una prima lettura, la percentuale di individui appartenenti alla fascia d'età superiore ai 75 anni sembra relativamente bassa nel grafico. Tuttavia, uno sguardo più approfondito rivela che questa fascia d'età è in realtà più rappresentativa di quanto ci si potesse aspettare. La sua rappresentatività è superiore di circa il 6% rispetto alla distribuzione generale nel campione, che si attesta al 2,44%. Questo dato sorprendente suggerisce che il Cluster 1 comprenda un numero significativo di partecipanti anziani, una caratteristica unica che lo distingue dagli altri cluster.

Questa analisi dettagliata mette quindi in luce le differenze nella distribuzione per fasce d'età tra i vari cluster. Ogni cluster presenta una composizione unica rispetto

alla distribuzione generale del campione, evidenziando variazioni più o meno significative e offrendo una panoramica dettagliata delle caratteristiche demografiche di ciascun gruppo.

Successivamente, nel contesto della nostra analisi, abbiamo dato particolare importanza a un elemento fondamentale: la prima domanda del questionario chiedeva ai partecipanti di specificare la sala del museo in cui si trovavano nel momento della compilazione del questionario. Ci siamo quindi concentrati sull'analisi delle peculiarità riscontrate nelle risposte dei partecipanti negli 8 gruppi distinti, tenendo conto delle caratteristiche della specifica sala e delle emozioni che hanno dichiarato di aver sperimentato durante la permanenza in quella precisa sala. Queste informazioni si sono rivelate cruciali poiché le risposte fornite dai visitatori alle domande successive, incentrate sulle emozioni vissute durante la loro visita al museo, sono strettamente correlate alle specificità e all'atmosfera della sala specifica in cui si trovavano. In altre parole, le emozioni espresse dai visitatori sono state influenzate dalle caratteristiche della sala, creando così un collegamento significativo tra il contesto circostante e le loro risposte al questionario. Questi risultati sono presentati in dettaglio nella Figura 4.24.

Per esaminare questa relazione, abbiamo effettuato un incrocio tra due variabili chiave:

1. Variabile "Cluster": questa variabile rappresenta i gruppi o cluster identificati tra i partecipanti. Ogni partecipante è stato assegnato a uno degli 8 cluster in base alle loro risposte, che potrebbero riflettere similitudini nelle loro esperienze o nelle loro caratteristiche.
2. Variabile "Sala": questa variabile indica la sala specifica in cui ogni partecipante si trovava quando ha completato il questionario. Ogni sala presenta delle caratteristiche uniche, come l'ambiente circostante, l'illuminazione, la temperatura, ecc.

L'obiettivo di questo incrocio tra le variabili "Cluster" e "Sala" è stato identificare se all'interno di ciascun cluster ci fossero prove di consistenza nelle emozioni riferite in relazione a una determinata sala. In altre parole, abbiamo cercato di scoprire se i partecipanti all'interno di uno stesso cluster avevano esperienze emotive simili quando si trovavano in una sala specifica.

L'analisi che abbiamo condotto, incrociando i cluster con le diverse sale, è un passo cruciale per esaminare come le condizioni ambientali influenzino le emozioni dei partecipanti. Analizzando i dati in questo modo, possiamo infatti individuare tendenze e modelli nei modi in cui le persone reagiscono

emotivamente in relazione a specifiche condizioni ambientali. Ad esempio, potremmo rivelare che alcuni cluster di visitatori manifestano emozioni più positive o negative in determinate sale o contesti. Ciò significa che potremmo scoprire che un cluster di visitatori ha una probabilità significativamente maggiore di provare emozioni specifiche in una particolare sala rispetto ad altri cluster. Queste correlazioni possono essere estremamente informative per comprendere come l'ambiente circostante influenzi le emozioni.

Infine, la presenza di differenze significative tra i vari cluster in termini di reazioni emotive può essere molto interessante. Potrebbe indicare che alcuni gruppi di partecipanti sono particolarmente sensibili a certe condizioni ambientali, mentre altri possono essere meno influenzati.

Comprendere dunque come l'ambiente circostante incide sulle emozioni può avere applicazioni pratiche in vari contesti. Ad esempio, nelle impostazioni di design degli spazi, questa conoscenza potrebbe essere utilizzata per creare ambienti che favoriscano emozioni positive o per adattare l'ambiente in modo da mitigare emozioni negative.

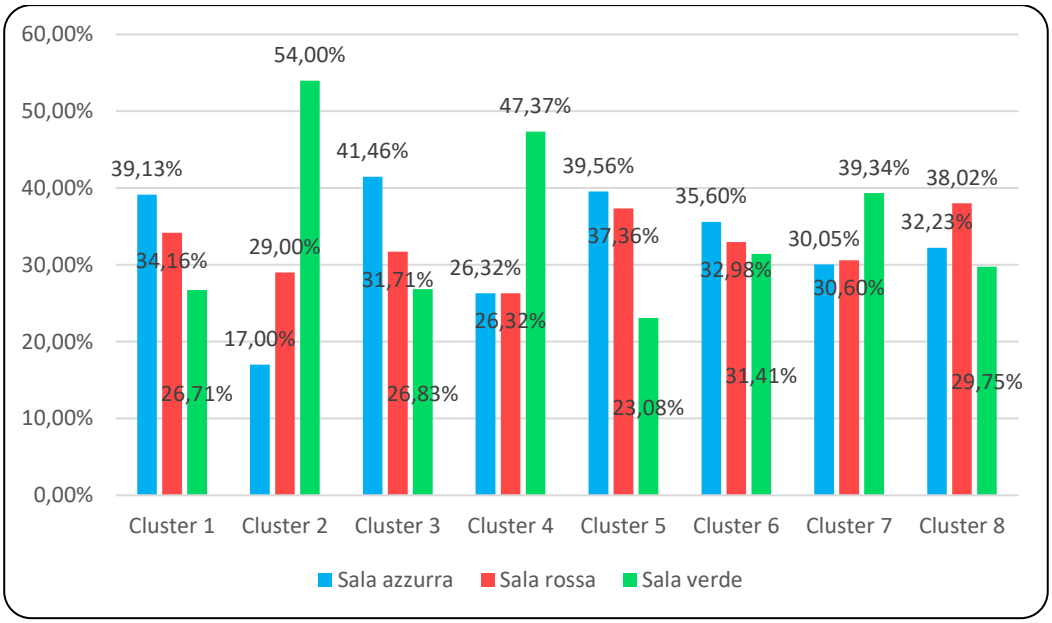


Figura 4.24 – Analisi delle Emozioni in base alle diverse Sale del museo. Fonte: nostre elaborazioni.

Nell'analisi incrociata dei cluster di visitatori con le emozioni provate nelle diverse sale del museo (azzurra, rossa e verde), emergono interessanti correlazioni tra colore e atmosfera della sala e le rispettive emozioni evocate. In particolare, è possibile notare che i cluster dei "Meravigliati", degli "Spensierati", degli

“Entusiasti” e degli “Indifferenti” hanno segnalato la sala azzurra come il luogo in cui hanno dichiarato di aver sperimentato determinate emozioni, come è chiaramente illustrato nella Figura 4.24. In questa sala, le emozioni prevalenti sembrano essere la gioia, la sorpresa e una generale sensazione di benessere. Questi visitatori sembrano dunque essere stati influenzati positivamente dal colore azzurro, che ha contribuito a creare un’atmosfera rilassante e tranquilla che stimola emozioni legate all’apprezzamento dell’arte, favorendo così la sperimentazione di emozioni positive.

D’altra parte, i cluster dei “Melanconici”, degli “Stupiti” e dei “Contemplativi” hanno occupato la sala verde durante la compilazione del questionario. La sala verde sembra offrire un ambiente che favorisce la riflessione e l’introspezione. Qui, le emozioni emerse sono principalmente la tristezza e la sorpresa. Il verde è spesso associato a sentimenti di natura, equilibrio e tranquillità, il che potrebbe spiegare perché i visitatori in questa sala abbiano sperimentato emozioni complesse come la tristezza e la sorpresa. Questa combinazione di emozioni potrebbe indicare una profonda contemplazione delle opere d’arte esposte e un’esperienza più riflessiva.

Infine, il cluster dei “Rilassati” ha trascorso il tempo nella sala rossa, dove la gioia è stata l’emozione predominante. Il colore rosso, noto per evocare sensazioni di passione ed energia, sembra aver stimolato una risposta emotiva positiva e appassionata. Questo ambiente vivace potrebbe aver contribuito a creare un’atmosfera che ha enfatizzato l’aspetto emozionale e gioioso dell’arte, portando i visitatori a sperimentare una gioia intensa e coinvolgente.

Questi risultati suggeriscono che il colore e l’atmosfera delle sale di un museo possono influenzare notevolmente le emozioni sperimentate dai visitatori, contribuendo a modellare l’esperienza complessiva della visita. Tuttavia, è importante notare che le reazioni emotive possono variare notevolmente da persona a persona, e le correlazioni osservate possono essere il risultato di diverse variabili. Ulteriori ricerche e analisi potrebbero aiutare a comprendere meglio questi legami complessi tra ambiente, emozioni e apprezzamento dell'arte.

Questa analisi rappresenta dunque un’opportunità significativa per ottenere una migliore comprensione delle dinamiche emotive e ambientali, consentendo di scoprire come le persone reagiscono alle diverse condizioni e come queste reazioni possano variare tra gruppi di individui.

In conclusione, l’analisi dei cluster non solo ci fornisce una panoramica dettagliata delle diverse tipologie di visitatori e delle loro esperienze emotive, ma rappresenta anche un passo fondamentale per orientare le azioni future e

migliorare ulteriormente l'esperienza complessiva dei visitatori presso la Pinacoteca Tosio Martinengo.

Questi risultati sono intrinsecamente legati all'obiettivo originale della ricerca di mercato sull'esperienza dei visitatori. Attraverso la segmentazione dei dati e l'analisi dei cluster, siamo così in grado di capire meglio chi sono i visitatori, cosa cercano e come possiamo migliorare l'esperienza del museo per soddisfare le loro esigenze specifiche. Inoltre, questo approccio consente di adottare decisioni informate e di personalizzare le strategie di coinvolgimento dei visitatori, contribuendo così al successo generale del museo e alla soddisfazione dei suoi visitatori.

4.4.4 Implicazioni per il marketing museale

L'analisi dei cluster dei visitatori all'interno della Pinacoteca Tosio Martinengo apre interessanti prospettive nel campo del marketing museale. Ogni cluster, caratterizzato da differenti risposte emotive e preferenze, offre opportunità uniche per adattare e migliorare l'esperienza dei visitatori. Ad esempio, per i cluster degli "Entusiasti" e dei "Meravigliati", notoriamente inclini alla gioia e alla sorpresa, il museo potrebbe considerare programmi speciali che enfatizzano le opere d'arte più innovative e stimolanti, incoraggiando così l'interazione e la partecipazione attiva dei visitatori. Questi programmi potrebbero includere visite guidate interattive o eventi di lancio di mostre per continuare a suscitare la loro costante meraviglia.

Allo stesso modo, per il cluster degli "Indifferenti", i quali sperimentano emozioni positive come gioia e sorpresa in misura però moderata, il museo potrebbe cercare di coinvolgerli in modo più attivo sempre tramite esperienze interattive e coinvolgenti. Tour guidati focalizzati sulla scoperta di dettagli nascosti nelle opere d'arte o mostre temporanee che cambiano frequentemente potrebbero suscitare il loro interesse.

Per i visitatori "Stupiti", i quali si distinguono per l'assenza di emozioni intense e per la predominante presenza di sorpresa, il museo potrebbe capitalizzare su questa costante meraviglia creando esperienze interattive o installazioni sorprendenti nelle sale. Eventi speciali, come aperture serali con luci e suoni particolari, potrebbero mantenere il loro interesse e coinvolgerli in modo significativo.

Nel caso dei visitatoti "Rilassati", che sperimentano gioia ma con una bassa intensità di sorpresa, il museo potrebbe creare un'atmosfera più rilassante e contemplativa nelle sale espositive. Sarebbe possibile realizzare spazi di relax o

zone dove i visitatori possono apprezzare le opere d'arte in un'ambiente più tranquillo.

D'altra parte, per il cluster dei "Contemplativi", ovvero coloro che non provano gioia ma sperimentano una moderata sorpresa e tristezza, il museo potrebbe enfatizzare l'aspetto della riflessione e della contemplazione. Installazioni o esposizioni che invitano alla riflessione e alla connessione emotiva potrebbero attirare e coinvolgere questi visitatori.

Allo stesso modo, il cluster dei "Melanconici", con la tendenza a provare tristezza durante la visita, potrebbe beneficiare di programmi e iniziative incentrati sul benessere emotivo. Il museo potrebbe quindi creare ambienti rassicuranti per aiutare a mitigare la tristezza e rendere l'esperienza più gratificante.

L'analisi dei cluster offre dunque una guida preziosa per personalizzare l'esperienza dei visitatori. Comprendere e adattarsi alle esigenze di ciascun cluster permette al museo di creare un'esperienza più coinvolgente e significativa per tutti i visitatori. Personalizzare le iniziative e gli eventi in base alle preferenze e alle reazioni emotive dei visitatori può trasformare il modo in cui vivono e apprezzano l'arte, migliorando la reputazione e l'attrattiva del museo. Ciascun cluster offre dunque al museo opportunità uniche per ottimizzare il coinvolgimento dei visitatori e garantire una visita emozionante.

Conclusioni

Nella presente tesi, abbiamo affrontato un percorso che ci ha condotto attraverso i concetti teorici e l'applicazione pratica dell'algoritmo di *Spectral Clustering* nell'analisi dei dati e nella segmentazione del mercato. Abbiamo iniziato introducendo la segmentazione di mercato come una strategia fondamentale nel contesto del marketing e della gestione aziendale. Successivamente, ci siamo addentrati nel mondo dello *Spectral Clustering*, una tecnica innovativa che supera le limitazioni delle metodologie tradizionali di clustering e apre nuove strade per la comprensione dei dati complessi e l'ottimizzazione delle strategie aziendali.

Abbiamo esplorato il grafo di similarità e la matrice Laplaciana come componenti chiave dello *Spectral Clustering*, approfondendo le loro basi matematiche e concettuali. Questa comprensione ha gettato le basi per l'applicazione pratica dell'algoritmo.

Nel capitolo culminante, abbiamo applicato lo *Spectral Clustering* ai dati raccolti durante uno studio sulla *Visitor Experience* presso la Pinacoteca Tosio Martinengo di Brescia. Attraverso questa applicazione, abbiamo suddiviso i visitatori in cluster con esperienze simili e identificato pattern comportamentali rilevanti. L'analisi dei cluster non è solo una metodologia statistica, ma un mezzo per comprender meglio il pubblico e personalizzare le interazioni con esso. La segmentazione dei visitatori in base alle loro preferenze e alle loro esperienze permette al museo di offrire un'esperienza più coinvolgente e significativa. Questi risultati non solo hanno contribuito a migliorare l'esperienza dei visitatori nel museo, ma hanno anche dimostrato il potenziale dell'algoritmo di *Spectral Clustering* nell'analisi dei dati complessi.

Nell'Introduzione, abbiamo sollevato delle domande chiave sul ruolo dell'analisi dei cluster e sulla sua applicazione pratica nell'ottimizzazione dell'esperienza dei visitatori presso la Pinacoteca Tosio Martinengo. Le domande chiave includevano: "In che modo l'analisi dei cluster può aiutare a comprendere meglio i modelli di comportamento dei visitatori?" e "Come l'analisi dei cluster può contribuire a migliorare l'esperienza dei visitatori nel museo?". Attraverso il nostro lavoro, abbiamo fornito risposte concrete a queste domande.

Abbiamo dimostrato che l'analisi dei cluster non solo offre una panoramica dettagliata delle diverse tipologie di visitatori e delle loro esperienze emotive, ma rappresenta anche un passo fondamentale per orientare le azioni future e per migliorare ulteriormente l'esperienza complessiva dei visitatori nel museo.

Durante il percorso di ricerca, abbiamo acquisito una profonda comprensione della potenza dell'analisi dei dati e dello *Spectral Clustering* nell'estrarre informazioni rilevanti da dati complessi. Questa esperienza ci ha fatto apprezzare l'importanza di un approccio basato sui dati nell'ottimizzazione delle strategie di marketing e nell'adattamento alle esigenze del pubblico.

Inoltre, abbiamo sperimentato direttamente come la teoria possa tradursi in risultati concreti preziosi e miglioramenti significativi, offrendo così nuove prospettive nel campo del marketing e della gestione aziendale. Questo ha rafforzato la nostra convinzione che l'analisi dei dati non sia solo una disciplina accademica, ma una potente leva di cambiamento in vari contesti applicativi.

Guardando al futuro, ci rendiamo conto che emergono diverse opportunità e potenzialità di sviluppo del lavoro svolto in questa tesi. In primo luogo, un approccio promettente per arricchire la comprensione e l'applicazione del metodo di *Spectral Clustering* potrebbe consistere nell'utilizzo di dati simulati. Questa strategia consentirebbe di superare le sfide tipiche dei dati reali, come la complessità nel determinare il numero di cluster o la presenza di osservazioni anomale. L'analisi su dati simulati offre infatti la possibilità di lavorare su un dataset pulito e controllato, in cui è possibile applicare il metodo senza tali complicazioni.

In aggiunta, uno dei punti di forza dello *Spectral Clustering* risiede nella sua capacità di individuare cluster non sferici, a differenza di approcci tradizionali come l'algoritmo *k-means*. Questa caratteristica si dimostra particolarmente vantaggiosa nell'affrontare dati complessi, in cui i cluster possono presentare forme irregolari o distribuzioni non uniformi. Tuttavia, l'interpretazione dei cluster non sferici può rappresentare una sfida significativa. Per affrontare efficacemente questa complessità e massimizzare il valore delle future analisi, potrebbe essere opportuno esplorare approfonditamente la vasta letteratura scientifica alla ricerca di metodi e tecniche specificamente consigliati da esperti per affrontare questa precisa sfida. Questi approcci potrebbero essere progettati appositamente per migliorare la chiarezza e la comprensione dei cluster non sferici, semplificando così il processo di interpretazione e agevolando l'identificazione delle caratteristiche chiave che li definiscono.

Inoltre, vale la pena considerare che lo *Spectral Clustering* trova applicazioni in una vasta gamma di discipline scientifiche, tra cui il marketing, ma anche la biologia, le scienze sociali e molte altre. L'ulteriore esplorazione di come questo metodo possa essere impiegato in contesti diversificati e la valutazione delle possibilità di collaborazioni interdisciplinari potrebbero aprire nuove prospettive

per l’analisi dei dati. Questo potrebbe portare a scoperte interessanti e contribuire alla crescita della conoscenza in un ampio spettro di campi.

Infine, con l’evolversi delle tecnologie e la crescente raccolta di dati, l’analisi dei cluster potrebbe diventare ancora più raffinata e precisa. L’integrazione di tecniche di intelligenza artificiale, come il *machine learning*, potrebbe aprire nuove prospettive nell’analisi dei dati di esperienza dei visitatori, consentendo previsioni più accurate e personalizzate.

Queste considerazioni riflettono la nostra visione di un futuro fertile e promettente per l’applicazione e lo sviluppo continuo dell’algoritmo di *Spectral Clustering* nell’analisi dei dati e nella ricerca interdisciplinare.

Speriamo che questo lavoro possa ispirare ulteriori ricerche e applicazioni innovative nel campo dell’analisi dei dati e del marketing, contribuendo a creare esperienze più coinvolgenti e soddisfacenti per il pubblico. Inoltre, auspichiamo che questa ricerca possa gettare ulteriormente luce sul potenziale di questa tecnica in vari contesti applicativi.

Elena Romano

Bibliografia

Ackerman M., S. B.-D. (2016). A Characterization of Linkage-Based Hierarchical Clustering. *Journal of Machine Learning Research 17*, 1-17.

Ahmadizadeh S., S. I. (2017). On eigenvalues of Laplacian matrix for a class of directed signed graphs. *Linear Algebra and its Applications*, 281-306.

Alpert C., K. A. (1999). Spectral partitioning with multiple eigenvectors . *Discrete Applied Mathematics 90(3)*, 3-26.

Anderson Jr., W. N. (1985). Eigenvalues of the Laplacian of a graph. *Linear and multilinear algebra 18(2)*, 141-145.

Bandyopadhyay, S., & Saha, S. (2013). Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications. In S. Bandyopadhyay, & S. Saha, *Unsupervised Classification* (p. 59-73). Berlin, Heidelberg: Springer.

Barbarito, L. (2011). *L'analisi di settore. Metodologia ed applicazioni* (Vol. 24). FrancoAngeli.

Bassi F., I. S. (2022). *Statistica per analisi di mercato. Metodi e strumenti.* Pearson.

Ben-David S., v. L. (2006). A sober look on clustering stability. *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, 5-19.

Bender, C. O. (1999). *Advanced mathematical methods for scientists and engineers I : asymptotic methods and perturbation theory.* New York: Springer.

Benesty, J., Chen, J., & Huang, Y. (2008). On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing, 16(4)*, 757-765.

Ben-Hur A., E. A. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 6-17.

Bhissy, K. E., Faleet, F. E., & Ashour, W. (2014). Spectral clustering using optimized gaussian kernel function. *International Journal of Artificial Intelligence and Applications for Smart Devices 2(1)*, 41-45.

Bock, H. H. (1974). *Automatische Klassifikation.* Gottingen: Vandenhoeck & Ruprecht.

Bolla, M. (1991). Relations between spectral and classification properties of multigraphs. *Technical Report No. DIMACS-91-27, Center for Discrete Mathematics and Theoretical Computer Science*.

Bondy, J. A., & Murty, U. S. (2008). *Graph theory.* Springer Publishing Company.

Brian S. Everitt, S. L. (2001). *Cluster Analysis.* Oxford University Press, fourth edition.

Bridges Jr, C. C. (1966). Hierarchical cluster analysis. *Psychological reports 18.3*, 851-854.

Cao, F. L. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications, 58(3)*, 474-483.

Chen, Y., Li, X., Liu, J., Xu, G., & Ying, Z. (2017). Exploratory item classification via spectral graph clustering. *Applied psychological measurement, 41(8)*, 579-599.

Chung, F. R. (1997). Spectral Graph Theory. *Regional Conference Series in Mathematics, 92*, 1-21.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., & Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, 1-4.

Cvetkovic D., D. M. (1979). *Spectra of Graphs: Theory and Application.* New York: Academic Press.

Darwin, C. (1872). *The expression of the emotions in man and animals.* London, UK: John Marry.

Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York*.

Ding C., H. X. (2001). A min-max cut algorithm. *Proceedings of the first IEEE International Conference on Data Mining (ICDM)*, 107-114.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis.* NY Engels JM: John Wiley and Sons.

Eades, D. C. (1965). The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. *Systematic Zoology*, 98-100.

Favati, P., & al., e. (2020). Construction of the similarity matrix for the spectral clustering method: Numerical experiments. *Computational and Applied Mathematics*.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math*, 298-305.

Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matricesand its application to graph theory. *Czechoslovak Mathematical Journal 25*, 619-633.

Fischer, I., & Poland, J. (2005). Amplifying the Block Matrix Structure for Spectral Clustering. *Technical Report No. IDSIA-03-05, Telecommunications Lab*.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning 2*, 139-172.

Fleiss, J. L., & Zubin, J. (1969). On the methods and theory of clustering. *Multiuariate Behaoiour* , 235-250.

Florek K., L. L. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum, 2*, 282-285.

Fraley C., R. A. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc. 97*, 611-631.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* San Diego, CA, USA: Academic Press Professional, Inc.

Ghojogh B., K. F. (2019a). Eigenvalue and generalized eigenvalue problems: Tutorial. *arXiv preprint arXiv:1903.11240*.

Ghojogh, B. (2021). *Data Reduction Algorithms in Machine Learning and Data Science.* PhD thesis, University of Waterloo.

Ghojogh, B. M. (2019d). Feature selection and feature extraction in pattern analysis: A literature review. *arXiv preprint arXiv:1905.02845*.

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Laplacian-based dimensionality reduction including spectral clustering, Laplacian eigenmap, locality preserving projection, graph embedding, and diffusion map: Tutorial and survey. *arXiv preprint arXiv:2106.02154*.

Golub, G., & Van Loan, C. (1996). Matrix Computations. *Johns Hopkins University Press*.

Gower, J. C. (1971a). A general coefficient of similarity and some of its properties. *Biomerrics*, 857-871.

Gülagiz, F. K. (2017). Comparison of hierarchical and non-hierarchical clustering algorithms. *International Journal of Computer Engineering and Information Technology*, 6-14.

H., F. D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of artificial intelligence research, 4* , 147-178.

Hall, K. M. (1970). An r-dimensional quadratic placement algorithm. *Management Science 17*, 219-229.

Hartigan, J. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society, series c (applied statistics) 28(1),* , 100-108.

Hastie T., T. R. (2001). *The Elements of Statistical.* New York: Springer.

Higgs, R. E. (1997). Experimental designs for selecting molecules from large chemical databases. *Journal of chemical information and computer sciences, 37(5)*, 861-870.

Holmes, M. (2013). *Introduction to perturbation methods.* New York: Springer.

Howell, D. C. (2012). *Statistical methods for psychology.* Belmont, CA: Cengage Learning.

Huttenhower C., F. A. (2007). Nearest neighbor networks: Clustering expression data based on gene neighborhoods. *BMC Bioinformatics 8*, 250.

Irani, J., Pise, N., & Phatak, M. (2016). Clustering techniques and the similarity measures used in clustering: A survey. *International journal of computer applications, 134(7)*, 9-14.

Jardine N., S. R. (1971). *Mathematical Taxonomy.* New York: John Wiley.

Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University, 29*.

Jimenez, L. L. (1997). e. Supervised classification in high-dimensional space: geometrical,statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man and Cybernetics, 28(1)*, 39-54.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241-254.

Jonsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *10th International Symposium on Software Metrics*, 108-118.

Jönsson, P., & Wohlin, C. (2006). Benchmarking k-nearest neighbour imputation with homogeneous Likert data. *Empirical Software Engineering, 11*, 463-489.

Kaufman, L., & Rousseeuw, P. J. (1990). *Findings Group In Data: An Introduction To Cluster Analysis.* Hoboken, New Jersey: John Wiley & Sons.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods.* London: Edward Arnold.

Kim J., C. S. (2006). Semidefinite spectral clustering. *Pattern Recognition 39 (11),* 2025–2035.

Lance G. N., W. W. (1966). Computer programs for hierarchical polythetic classification ("Similarity analyses"). *The Computer Journal 9(1)*, 60-64.

Lancia G. N., W. W. (1967). A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal 9(4)*, 373-380.

Lange T., R. V. (2004). Stability-based validation of clustering solutions. *Neural Compu 16(6)*, 1299-1323.

Likas A., V. N. (2003). The Global KMeans Clustering Algorithm. *Pattern Recognition 36(2)*, 451-461.

Ling, S. (2020). Spectral Clustering, Graph Laplacian. *MATH-SHU 236*, 1-12.

Lucińska, M., & Wierzchoń, S. T. (2012). Spectral clustering based on k-nearest neighbor graph. *Computer Information Systems and Industrial Management: 11th IFIP TC 8 International Conference, CISIM 2012, Venice, Italy, September 26-28, 2012. Proceedings 11* (p. 254-265). Springer Berlin Heidelberg.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1*(14), 281-297.

Marsden, A. (2013). Eigenvalues of the Laplacian and their relationship to the connectedness of a graph. *University of Chicago*.

McQuitty, L. L. (1964). Capabilities and Improvements of Linkage Analysis as a Clustering Method. *Educational and Psychological Measurement 24(3)*, 441-456.

Meilă M., S. J. (2001). Learning Segmentation by Random Walks. *Neural Information Processing Systems 13*, 873-879.

Meilă, M. H. (1998). An experimental comparison of several clustering and initialization methods. *Proceedings of the fourteenth conference on uncertainty in artificial intelligence,*, 386-395.

Merris, R. (1994). Laplacian matrices of graphs: a survey. *Linear algebra and its applications, 197*, 143-176.

Milligan G. W., C. M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50(2)*, 159-179.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika 45*, 325-342.

Mohar, B. (1991). The Laplacian spectrum of graphs. *Graph Theory, Combinatorics, and Applications 2*, 871–898.

Mohar, B. (1997). Some applications of Laplace eigenvalues of graphs. *Graph symmetry: Algebraic methods and applications*, 225-275.

Nascimento, M., & de Carvalho, A. (2011). Spectral methods for graph clustering – A survey. *European Journal of Operational Research 211*, 221-231.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems, 14*.

Niu D., D. J. (2011). Dimensionality reduction for spectral clustering. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 552-560.

Paul E. Green, R. E. (1967). Cluster Analysis in Test Market Selection 13(8). *Management Science*, 387-400.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philiosophical Magazine, 2*, 559–572.

Pena J.M., L. J. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters, 20*, 1027-1040.

Pérez, A. M. (2018). A Density-Sensitive Hierarchical Clustering Method. *Journal of Classification 35*, 481-510.

Polito M., P. P. (2002). Grouping and dimensionality reduction by locally linear embedding. *Advances in neural information processing systems*, 1255-1262.

Schaeffer, S. E. (2007). Graph clustering. *Computer science review, 1(1)*, 27-64.

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *Bmj*, 349.

Selim S. Z., I. M. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, 81-87.

Shi J., M. J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 888-905.

Shi, T., Belkin, M., & Yu, B. (2009). Data spectroscopy: eigenspace of convolution operators and clustering. *The Annals of Statistics, 37*(6B), 3960–3984.

Smart, J. (2005). Statistics in Market Research. *Journal of yìthe Royal Statistics Society Series A: Statistics in Society 168(3)*, 630-631.

Snarey, M. T. (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling, 15(6)*, 372-385.

Sneath, P. H. (1957). The application of computers to taxonomy. *Journal of General Microbiology, 17*, 201-226.

Sokal R. R., M. C. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Scientific Bulletin, 28*, 1409-1438.

Sokal, R. S. (1963). *Principles of Numerical Taxonomy,*. San Francisco: WH Freeman.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 206–222.

Spielman, D. (2019). Spectral graph theory. *Lecture notes*.

Sprent, P., & Smeeton, N. C. (2016). *Applied nonparametric statistical methods.* Boca Raton, FL: CRC Press.

Still S., B. W. (2004). How many clusters? An information-theoretic perspective. *Neural Comput 16(12)*, 2483-2506.

Stoer M., W. F. (1997). A simple min-cut algorithm. *J. ACM 44(4)*, 585-591.

Tibshirani R., W. G. (2001). Estimating the number of clusters in a dataset via the gap statistic. *J. Roy. Stat. Soc. B 63(2)*, 411-423.

Torgerson, W. (1952). Multidimensional scaling I: Theory and method. *Psychometrika, 17*, 401-419.

Tryon, R. C. (1939). *Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality,* . Ann Arbor, Mich: Edwards brothers, inc., lithoprinters and publishers.

Tutte, W. T. (2001). *Graph theory* (Vol. 21). Cambridge university press.

Van Der Maaten, L. P. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*, 66-71.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*, 395-416.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *American Statistical Association 58(301)*, 236-244.

Weiss, Y. (1999). Segmentation using eigenvectors: a unifying view. *In Proceedings of the seventh IEEE international conference on computer vision, 2*, 975-982.

West, D. B. (2001). *Introduction to graph theory* (Vol. 2). Upper Saddle River: Prentice hall.

White, P. A. (1958). The computation of eigenvalues and eigenvectors of a matrix. *Journal of the Society for Industrial and Applied Mathematics, 6(4)*, 393-437.

Xia, T., Cao, J., Zhang, Y., & Li, J. (2008). On defining affinity graph for spectral clustering through ranking on manifolds. *Neurocomputing, 72*, 3203-3211.

Zanoletti G., *Analisi Sensoriale dell'esperienza di visita di un museo: il caso della Pinacoteca di Brescia*, Università degli Studi di Brescia, Dipartimento di Economia e Management, Anno Accademico 2018-2019.

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association, 67(339)*, 578-580.

Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. *Proc. of NIPS'04*, 1601–1608.

Zhou, Y., Cheng, H., & Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment, 2(1)*, 718-729.

Zou, H. (2020). Clustering Algorithm and Its Application in Data Mining. *Wireless Pers Commun, 110*, 21-30.

Sitografia

John Clements, Introduction to Hierarchical Clustering, *Towards Data Science*, *11 novembre 2019*.
https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e#:~:text=Average%2Dlinkage%20is%20where%20the,distance%20metrics%20in%20hierarchical%20clustering

Linkage: Agglomerative hierarchical cluster tree, *MathWorks*.
https://it.mathworks.com/help/stats/linkage.html

Dhilip Subramanian, A Simple Introduction to K-Nearest Neighbors Algorithm, *Towards Data Science*, *8 giugno 2019*.
https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e

Amey Band, How to find the optimal value of K in KNN?, *Towards Data Science*, *23 maggio 2020*.
https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb

Ringraziamenti

Desidero innanzitutto ringraziare la Professoressa Paola Zuccolotto per la sua disponibilità, il tempo dedicato e la gentilezza dimostrata durante l'intero processo di scrittura di questa tesi.

Un sentito ringraziamento va altresì al Dottor Matteo Ventura, la cui preziosa disponibilità e attenzione costante hanno saputo guidarmi con competenza lungo il percorso di questo lavoro di ricerca.

Un ringraziamento particolare va alla mia famiglia per il loro sostegno, sia economico che morale, e per avermi incoraggiata a perseverare nello studio, sempre con costanza e dedizione.

Vorrei inoltre ringraziare Stefano, che con amore e pazienza mi ha sempre spronata a dare il massimo.

Un pensiero speciale va infine a tutti i miei amici e compagni di studio che hanno condiviso con me questo percorso, offrendomi il loro sostegno e la loro preziosa amicizia.

Questo percorso di studio è stato ricco di sfide, ma ha anche saputo regalarmi grandi soddisfazioni che vorrei condividere con ognuno di voi.

Grazie.