UNIVERSITÀ DELLA VALLE D'AOSTA
UNIVERSITÉ DE LA VALLÉE D'AOSTE

DIPARTIMENTO DI SCIENZE ECONOMICHE E POLITICHE

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

ANNO ACCADEMICO 2025/2026

TESI DI LAUREA

Segmentazione data-driven dei flussi turistici mediante tecniche di clustering: un'analisi dei Comuni della Valle d'Aosta

DOCENTE relatore: Prof.ssa Consuelo R. Nava

STUDENTE: 23 C06 064, Manuel Guerini

Indice

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE

INTRODUZIONE

Contesto e motivazioni della ricerca

Negli ultimi anni, la crescente disponibilità di dati statistici e il progressivo sviluppo delle metodologie quantitative hanno ampliato significativamente le possibilità di analisi dei fenomeni economici e territoriali. Anche nell'ambito degli studi sul turismo, la disponibilità di *database* sempre più dettagliati ha favorito la diffusione di approcci *data-driven*, con l'obiettivo di estrarre conoscenza direttamente dai dati osservati e di supportare l'interpretazione di fenomeni caratterizzati da una crescente complessità.

In questo contesto, la segmentazione rappresenta uno degli strumenti maggiormente impiegati per analizzare la struttura dei sistemi turistici territoriali. A differenza degli approcci fondati su classificazioni a priori, la segmentazione *data-driven* consente di individuare gruppi omogenei emergenti direttamente dai dati, offrendo una rappresentazione sintetica della struttura del fenomeno analizzato e permettendo di evidenziare relazioni che difficilmente emergerebbero attraverso l'osservazione separata delle singole variabili.

Tale prospettiva risulta particolarmente interessante anche nello studio dei sistemi turistici territoriali. Le destinazioni turistiche presentano infatti caratteristiche spesso molto differenziate in termini di intensità dei flussi turistici, distribuzione della domanda e modalità di utilizzo delle strutture ricettive. Comprendere tali differenze è fondamentale sia al fine di approfondire la conoscenza della struttura dei sistemi turistici sia per supportare, attraverso strumenti quantitativi, le attività di analisi territoriale e i processi decisionali in ambito turistico.

Il presente lavoro si inserisce in tale contesto, assumendo come caso di studio il sistema turistico della Valle d'Aosta. Nonostante la consolidata rilevanza del turismo per l'economia valdostana, l'applicazione di approcci di segmentazione *data-driven* a questo contesto risulta ancora poco esplorata. Ciò potrebbe essere attribuibile alla difficoltosa reperibilità di informazioni statistiche dettagliate sui flussi turistici comunali. Il presente elaborato si propone pertanto di verificare il contributo che le tecniche di *clustering* possono offrire alla comprensione della struttura del sistema turistico valdostano, applicando metodologie consolidate della letteratura internazionale a un contesto territoriale specifico.

Obiettivo e domande di ricerca

Con riferimento al contesto introdotto nei paragrafi precedenti, il presente lavoro si propone di analizzare il sistema turistico della Valle d'Aosta attraverso un approccio di segmentazione *data-driven*, applicando tecniche di *clustering* ai dati comunali relativi ai flussi turistici e alla distribuzione della domanda tra le diverse tipologie di strutture ricettive oggetto di analisi.

L'obiettivo della ricerca è quello di verificare in che misura le tecniche di *clustering* consentano di individuare profili territoriali omogenei emergenti direttamente dai dati e di interpretarne le principali caratteristiche, valutando al tempo stesso il contributo che tali tecniche possono offrire alla comprensione della struttura del sistema turistico valdostano e al supporto delle attività di analisi territoriale finalizzate alla gestione e alla promozione dell'offerta turistica regionale.

Più in generale, il lavoro intende discutere le potenzialità e i limiti dell'impiego delle tecniche di *clustering* come strumento quantitativo di supporto all'analisi dei sistemi turistici territoriali, evidenziandone il valore conoscitivo nell'ambito di un approccio *data-driven*. Coerentemente con tale obiettivo, la ricerca si propone di rispondere ai seguenti quesiti:

i. In che misura le tecniche di *clustering* consentono di individuare profili territoriali omogenei a partire dai dati relativi ai flussi turistici e alla distribuzione della domanda tra le diverse tipologie di strutture ricettive nei comuni della Valle d'Aosta?
ii. Quali caratteristiche emergono dalla segmentazione ottenuta e cosa rivelano sulla struttura del sistema turistico valdostano?
iii. Quale contributo possono offrire le tecniche di *clustering*, come strumento quantitativo complementare all'analisi e alla pianificazione dei sistemi turistici territoriali?

Attraverso la risposta a tali quesiti, il presente lavoro intende contribuire alla comprensione delle dinamiche territoriali del sistema turistico valdostano, verificando se un approccio di segmentazione *data-driven* possa rappresentare un utile strumento quantitativo per analizzarne la struttura e supportarne l'interpretazione mediante profili territoriali emergenti direttamente dai dati osservati.

Struttura della tesi

Il presente elaborato si articola in tre capitoli principali, oltre all'introduzione, alle conclusioni, ai riferimenti bibliografici e sitografici e all'appendice.

Il primo capitolo presenta una review della letteratura dedicata al tema della segmentazione nel settore turistico. Dopo aver introdotto il concetto di segmentazione e i principali obiettivi perseguiti attraverso tale approccio, vengono analizzate le diverse tipologie di segmentazione, con particolare riferimento alla distinzione tra approcci *a priori* e approcci *data-driven*. Il capitolo approfondisce inoltre il ruolo della *cluster analysis* quale principale tecnica di segmentazione *data-driven*, esaminandone le applicazioni alla segmentazione dei turisti, delle destinazioni e dei territori, fino a collegare le più recenti evidenze empiriche sul turismo alpino al caso di studio della Valle d'Aosta.

Il secondo capitolo è dedicato alla metodologia statistica adottata nell'elaborato. Dopo l'introduzione ai principi dell'analisi dei dati e della segmentazione *data-driven*, vengono presentate le principali misure di distanza utilizzate nelle tecniche di *clustering*. Successivamente vengono illustrati i metodi di *clustering* gerarchico e non gerarchico, approfondendo il funzionamento del metodo del legame completo (*Complete Linkage*), del metodo di *Ward* e dell'algoritmo *K-means*. Il capitolo si conclude con un confronto tra le principali metodologie analizzate e con alcune considerazioni critiche riguardanti i limiti e le implicazioni metodologiche.

Il terzo capitolo costituisce la parte empirica dello studio. Dopo la presentazione del dataset, delle variabili considerate e delle operazioni di preparazione e pulizia dei dati, vengono illustrate le analisi descrittive preliminari e l'applicazione delle tecniche di *clustering* ai dati relativi ai comuni della Valle d'Aosta. I risultati ottenuti mediante i diversi algoritmi vengono poi confrontati e interpretati, al fine di individuare gruppi omogenei di comuni caratterizzati da simili profili turistici e di discuterne le principali implicazioni per il caso di studio.

L'elaborato si conclude con un capitolo dedicato alle conclusioni, nel quale vengono sintetizzati i principali risultati della ricerca, vengono fornite risposte alle domande di ricerca e vengono discussi il valore, il contributo e i limiti della ricerca. Il capitolo prosegue con alcune riflessioni sui futuri sviluppi della ricerca e si chiude con alcune considerazioni finali sul lavoro svolto.

Dati e metodologia

La ricerca è stata sviluppata adottando un approccio quantitativo di tipo *data-driven*, applicato all'analisi del sistema turistico della Valle d'Aosta. L'analisi empirica si basa su un dataset costruito a partire dai dati relativi ai flussi turistici registrati nel periodo compreso tra gennaio e settembre 2025 nei singoli comuni valdostani, comprendente informazioni sugli arrivi, sulle presenze e sulla distribuzione della domanda tra le diverse tipologie di strutture ricettive. Dopo

una fase preliminare di selezione, organizzazione e preparazione dei dati, è stato costruito un dataset finale costituito da 74 unità statistiche, corrispondenti ai comuni valdostani, e da 258 variabili quantitative rappresentative delle principali caratteristiche della domanda turistica osservata.

L'analisi è stata interamente sviluppata mediante il software statistico R ed è stata articolata in più fasi. In primo luogo, le variabili sono state sottoposte a un processo di standardizzazione, al fine di renderle confrontabili e di evitare che differenze di scala influenzassero i risultati del *clustering*. Successivamente sono state applicate differenti tecniche di *clustering*, appartenenti sia ai metodi gerarchici sia ai metodi non gerarchici. Nello specifico, sono stati applicati il metodo del legame completo (*Complete Linkage*), il metodo di *Ward* e l'algoritmo *K-means*, al fine di confrontare le segmentazioni ottenute e determinare quella maggiormente coerente con l'obiettivo della ricerca.

Per tale ragione, la scelta della soluzione interpretativa finale non è stata effettuata esclusivamente sulla base di un unico criterio, ma attraverso una valutazione congiunta dei risultati ottenuti mediante i diversi algoritmi di *clustering* e le differenti configurazioni del numero di *cluster*, integrata da considerazioni di natura interpretativa, coerentemente con quanto suggerito dalla letteratura metodologica sulla segmentazione *data-driven*. L'intero percorso di analisi è stato pertanto sviluppato privilegiando criteri di trasparenza, replicabilità e interpretabilità, documentando le principali decisioni metodologiche adottate nelle diverse fasi della ricerca.

Principali risultati e contributi ottenuti dalla ricerca

L'analisi condotta ha evidenziato come l'applicazione di tecniche di *clustering* ai dati relativi ai flussi turistici dei comuni della Valle d'Aosta consenta di individuare profili territoriali omogenei emergenti direttamente dai dati osservati, senza ricorrere a classificazioni definite a priori. Il confronto tra differenti algoritmi e diverse configurazioni di segmentazione ha permesso di individuare, quale soluzione interpretativamente più coerente con gli obiettivi della ricerca, la segmentazione ottenuta mediante l'algoritmo *K-means* con cinque *cluster*, successivamente approfondita attraverso l'analisi dei comuni appartenenti ai diversi *cluster*.

I risultati ottenuti mettono in evidenza come il sistema turistico valdostano presenti una struttura complessivamente piuttosto omogenea, all'interno della quale emergono alcuni comuni caratterizzati da una marcata specializzazione turistica. La segmentazione ottenuta ha consentito di sintetizzare in un numero limitato di profili territoriali interpretabili la complessità

delle informazioni contenute nel dataset, offrendo una rappresentazione integrata delle principali caratteristiche della domanda turistica regionale.

Oltre a fornire una fotografia della struttura del sistema turistico valdostano nel periodo oggetto di analisi, il presente studio contribuisce a discutere il ruolo delle tecniche di *clustering* come strumento quantitativo complementare per l'analisi dei sistemi turistici territoriali. I risultati confermano infatti come un approccio di segmentazione *data-driven*, se applicato con rigore metodologico e interpretato criticamente alla luce delle caratteristiche dei dati disponibili, possa rappresentare un valido supporto alla comprensione delle dinamiche dei sistemi turistici territoriali e alle attività di analisi e pianificazione delle destinazioni turistiche.

1. REVIEW DELLA LETTERATURA

Introduzione

Il presente capitolo ha l'obiettivo di delineare il quadro teorico e scientifico entro il quale si inserisce il presente elaborato. Per tale ragione, viene proposta una review della letteratura dedicata alla segmentazione nel settore turistico, con particolare attenzione agli approcci di segmentazione *data-driven* e al ruolo delle tecniche di *clustering* nell'analisi dei fenomeni turistici.

Dopo aver introdotto il concetto di segmentazione e i principali obiettivi perseguiti attraverso tale approccio, il capitolo analizza le differenti tipologie di segmentazione illustrate dalla letteratura, soffermandosi in particolare sulla distinzione tra approcci a priori e approcci a posteriori (*data-driven*). In seguito, viene approfondito il ruolo della *cluster analysis* quale principale metodologia impiegata nella costruzione di segmenti emergenti direttamente dai dati, esaminandone le principali applicazioni alla segmentazione dei turisti, delle destinazioni e dei territori.

La review della letteratura si conclude esaminando le evidenze recenti nell'ambito dell'applicazione di tecniche di *clustering* nel turismo alpino, contesto analogo al caso della Valle d'Aosta, evidenziando come le più recenti applicazioni delle tecniche di *clustering* forniscano il quadro teorico e metodologico di riferimento per l'analisi empirica sviluppata nel terzo capitolo.

1.1 La segmentazione di mercato nel settore turistico

Nel contesto economico e manageriale odierno, i mercati sono caratterizzati da un'elevata eterogeneità della domanda, determinata dalla presenza al loro interno di consumatori che differiscono per preferenze, comportamenti di acquisto, bisogni e motivazioni di consumo. Tale eterogeneità rende spesso inefficace l'adozione di strategie di marketing indifferenziate. Diventa quindi necessario individuare gruppi relativamente omogenei di soggetti a cui poter destinare specifiche politiche di prodotto, prezzo, comunicazione e distribuzione, i cosiddetti segmenti *target*[1] (obiettivo).

Il concetto moderno di segmentazione di mercato viene generalmente ricondotto al contributo di Smith (1956), il quale definisce la segmentazione come un processo attraverso il quale un

[1] Per segmento *target* si intende la porzione specifica di mercato alla quale l'impresa decide di rivolgere la propria offerta e le proprie strategie di marketing.

mercato eterogeneo viene suddiviso in sottoinsiemi di consumatori caratterizzati da esigenze e comportamenti simili. Secondo l'autore, la segmentazione rappresenta una strategia alternativa rispetto alla semplice differenziazione del prodotto[2] che consente alle imprese di adattare la propria offerta alle caratteristiche specifiche dei diversi gruppi di domanda.

Nel settore turistico, la necessità di segmentare il mercato risulta ancora più evidente. La domanda turistica, infatti, presenta un elevato grado di variabilità dovuto a fattori quali le motivazioni di viaggio, la stagionalità della domanda[3], le preferenze individuali, il livello di spesa, la durata del soggiorno e le modalità di fruizione della destinazione. Ad esempio, turisti apparentemente simili dal punto di vista sociodemografico possono manifestare comportamenti di consumo estremamente differenti, rendendo spesso insufficiente l'utilizzo di criteri di classificazione tradizionali (Dolnicar, 2002).

Proprio per rispondere a una domanda così dinamica e mutevole, la letteratura scientifica ha progressivamente orientato l'attenzione verso approcci di segmentazione basati sull'analisi empirica dei dati osservati. In particolare, Dolnicar (2002), attraverso una revisione sistematica degli studi di segmentazione turistica, evidenzia come le tecniche statistiche multivariate[4] e i metodi di *clustering*[5] abbiano progressivamente assunto un ruolo predominante nella costruzione dei segmenti di mercato, consentendo di individuare gruppi (segmenti) omogenei senza imporre classificazioni predefinite.

La segmentazione di mercato può pertanto essere interpretata come un processo finalizzato a ridurre la complessità di un fenomeno eterogeneo attraverso l'individuazione di gruppi relativamente omogenei al proprio interno e sufficientemente distinti rispetto agli altri gruppi. In altri termini, l'obiettivo della segmentazione consiste nel massimizzare l'omogeneità interna ai gruppi e l'eterogeneità tra gruppi differenti. Tale principio costituisce il fondamento teorico

[2] Smith (1956) definisce, in termini semplici, la differenziazione del prodotto come piegare la domanda alla volontà dell'offerta. L'autore descrive quindi la differenziazione del prodotto come un tentativo di influenzare la struttura della domanda a favore dell'offerta, riducendo la sensibilità dei consumatori ai prodotti concorrenti. Tale approccio deriva dall'obiettivo di stabilire una sorta di equilibrio di mercato, apportando un adeguamento della domanda alle condizioni di offerta favorevoli al venditore.

[3] Per stagionalità della domanda turistica si intende la concentrazione ciclica dei flussi turistici (presenze di turisti) in determinati periodi dell'anno.

[4] Per tecniche statistiche multivariate si intendono le tecniche che permettono di esaminare simultaneamente più variabili riferite alle medesime unità di analisi, con l'obiettivo di identificare relazioni e schemi non immediatamente osservabili attraverso l'analisi separata delle singole variabili.

[5] Le caratteristiche e il funzionamento dei principali metodi di *clustering* saranno approfonditi nella sezione 2.3 del capitolo 2 relativa alle tecniche di *clustering*.

delle moderne applicazioni della *cluster analysis* nel settore turistico, dove le tecniche di segmentazione vengono sempre più frequentemente impiegate non solo per classificare i turisti, ma anche per individuare differenti tipologie di destinazioni e territori caratterizzati da specifici profili di sviluppo turistico.

L'evoluzione delle metodologie statistiche e la crescente disponibilità di dati hanno ulteriormente ampliato le potenzialità della segmentazione, favorendo lo sviluppo di approcci cosiddetti *data-driven*, nei quali i segmenti emergono direttamente dall'analisi dei dati anziché essere definiti a priori dal ricercatore. Tali approcci, oggi ampiamente diffusi nella ricerca turistica, saranno approfonditi nei paragrafi successivi.

1.2 Obiettivi e vantaggi della segmentazione

La segmentazione rappresenta uno strumento fondamentale per supportare le decisioni strategiche nell'ambito del marketing turistico e della gestione delle destinazioni. La possibilità di individuare segmenti specifici, ovvero gruppi di utenti caratterizzati da esigenze, preferenze e comportamenti simili, consente agli operatori turistici e ai responsabili delle destinazioni di adattare la propria offerta alle caratteristiche dei potenziali visitatori, aumentando l'efficacia delle attività di promozione, comunicazione e sviluppo turistico (Smith, 1956; Dolnicar, 2002).

Uno dei principali vantaggi della segmentazione consiste nel migliorare l'efficienza dei processi decisionali. La disponibilità di informazioni più dettagliate sui diversi gruppi che compongono il mercato consente infatti di orientare in modo più efficace gli investimenti e le strategie di sviluppo, favorendo un utilizzo più razionale delle risorse disponibili. In tale prospettiva, la segmentazione non costituisce solamente uno strumento di analisi della domanda, ma rappresenta anche un supporto operativo per la definizione delle politiche di marketing e delle strategie di gestione delle destinazioni.

Nella letteratura turistica, la segmentazione assume inoltre un ruolo rilevante quale strumento di pianificazione e gestione delle destinazioni. L'individuazione di gruppi caratterizzati da profili turistici simili consente infatti di evidenziare differenze e analogie tra mercati, destinazioni e territori, favorendo una migliore comprensione delle dinamiche turistiche locali e contribuendo alla definizione di strategie di sviluppo coerenti con le specificità dei differenti contesti territoriali (Dolnicar et al., 2018).

Un ulteriore vantaggio della segmentazione riguarda la possibilità di monitorare e interpretare i cambiamenti che interessano la domanda turistica nel tempo. L'evoluzione delle preferenze dei visitatori, delle modalità di fruizione delle destinazioni e delle caratteristiche dell'offerta rende infatti necessario disporre di strumenti in grado di cogliere tali trasformazioni e supportare l'adozione di interventi coerenti con le nuove esigenze del mercato. In questo senso, la segmentazione consente di ottenere una rappresentazione più articolata della domanda turistica rispetto a quella derivante dall'analisi del mercato considerato come un insieme omogeneo (Dolnicar, 2002).

In sintesi, la segmentazione permette di semplificare l'analisi di fenomeni complessi, caratterizzati da un'elevata eterogeneità, favorendo una maggiore comprensione del mercato e supportando i processi decisionali sia a livello operativo sia a livello strategico. Per tali ragioni, essa rappresenta uno degli strumenti maggiormente utilizzati nella ricerca turistica e costituisce il presupposto teorico delle diverse metodologie di segmentazione che saranno approfondite nelle sezioni successive.

1.3 Tipologie di segmentazione

Come emerso dalle sezioni precedenti, il turismo rappresenta un fenomeno particolarmente complesso rispetto ad altri prodotti. Un viaggio è infatti il risultato della combinazione di molteplici elementi che può essere influenzato da un'ampia gamma di motivazioni, preferenze e modalità di consumo. La domanda turistica è infatti caratterizzata da un elevato grado di eterogeneità e variabilità, derivante dai fattori precedentemente discussi. A ciò si aggiunge il fatto che il processo decisionale associato alla pianificazione di un viaggio coinvolge spesso più soggetti decisionali, contribuendo ad incrementare ulteriormente la complessità del fenomeno turistico. Tale complessità ha favorito nel tempo lo sviluppo di numerosi approcci alla segmentazione del mercato turistico (Dolnicar et al., 2018).

Le tipologie di segmentazione possono essere classificate secondo differenti criteri. Una prima distinzione riguarda le variabili utilizzate per suddividere il mercato in segmenti, mentre una seconda distinzione riguarda il processo attraverso cui i suddetti segmenti vengono individuati. In questa prospettiva, la letteratura distingue tra approcci di segmentazione *a priori* (*commonsense segmentation*) e *a posteriori*, nota anche come *data-driven segmentation* o *post-hoc segmentation* (Dolnicar, 2014).

1.3.1 Criteri di segmentazione

Nell'ambito della segmentazione di mercato in senso lato, una delle modalità più diffuse per classificare le tipologie di segmentazione consiste nel considerare la natura delle variabili utilizzate per distinguere i consumatori. Nel settore turistico, la letteratura individua quattro categorie (o tipologie) principali di segmentazione, o *segmentation bases* (Dolnicar, 2014): geografica, sociodemografica, psicografica e comportamentale.

La segmentazione geografica suddivide i turisti secondo un criterio di provenienza territoriale, come ad esempio il paese d'origine. Tale approccio è il più comunemente diffuso nelle organizzazioni turistiche nazionali, perché consente di distinguere con molta facilità e praticità quali segmenti sono più facili da raggiungere (Dolnicar, 2014), ad esempio consente di distinguere facilmente tra bacini d'utenza nazionali e internazionali, rendendo possibile lo sviluppo di strategie promozionali differenziate per ciascun mercato.

La segmentazione sociodemografica utilizza criteri o variabili, quali età, genere, professione, reddito e livello di istruzione. Anche in questo caso, la semplicità di raccolta delle informazioni unita alla facilità di interpretazione dei risultati rende questo approccio uno dei più utilizzati.

La segmentazione psicografica considera invece aspetti afferenti alla sfera personale e psicologica degli individui, quali motivazioni di viaggio, stili di vita, interessi, valori personali e percezione delle destinazioni turistiche. Essa consente, rispetto ai criteri precedenti, di cogliere dimensioni e sfumature altrimenti non percettibili e che influenzano profondamente la domanda turistica.

La segmentazione comportamentale si basa invece sulle effettive modalità di consumo turistico, quali le attività svolte durante la vacanza, la frequenza di viaggio, la spesa sostenuta o le modalità di utilizzo dei servizi turistici.

1.3.2 Segmentazione "a priori" e "a posteriori"

Oltre alla distinzione basata sulle variabili utilizzate, la letteratura propone anche una classificazione che si basa sul processo di costruzione dei segmenti. In tale prospettiva, Dolnicar (2004) distingue tra approcci di segmentazione *a priori* e segmentazione *a posteriori*. Questa distinzione non riguarda la natura delle variabili utilizzate, ma il momento in cui i segmenti vengono definiti rispetto all'analisi dei dati. Si tratta di un aspetto fondamentale e particolarmente rilevante ai fini del presente elaborato, in quanto verrà adottato un approccio di segmentazione *a posteriori* o *segmentazione data-driven*.

Nella segmentazione a priori, nota anche come *commonsense segmentation* (Dolnicar, 2014), i segmenti vengono stabiliti ex ante sulla base delle conoscenze e delle ipotesi del ricercatore. In tale prospettiva, la classificazione dei soggetti avviene utilizzando uno o più criteri considerati rilevanti per gli obiettivi dell'analisi. Ad esempio, una destinazione turistica può decidere di segmentare i visitatori in funzione della loro provenienza geografica o di criteri sociodemografici. I segmenti sono pertanto definiti prima dell'analisi e gli individui vengono successivamente assegnati alle categorie prestabilite, definite appunto *a priori* dal ricercatore. Il principale vantaggio di questo approccio è dato dalla sua semplicità applicativa e interpretativa. Generalmente i segmenti risultano intuitivi e facilmente utilizzabili nelle attività di pianificazione e promozione turistica. Tuttavia, la definizione a priori dei gruppi può comportare una rappresentazione semplificata della realtà perché individui appartenenti alla stessa categoria possono manifestare comportamenti o preferenze profondamente differenti.

Al contrario, la segmentazione a posteriori non prevede la definizione preventiva dei gruppi. Tale approccio prevede che i segmenti emergano direttamente dai dati attraverso l'applicazione di procedure statistiche finalizzate a individuare osservazioni caratterizzate da profili simili. L'obiettivo consiste nell'individuare strutture latenti presenti nei dati senza imporre classificazioni ex ante.

I gruppi vengono quindi costruiti sulla base delle effettive caratteristiche simili riscontrate tra gli individui o tra le unità oggetto di studio. Secondo Dolnicar (2004) tale approccio di segmentazione, noto anche come *data-driven segmentation*, risulta particolarmente adatto e utile nei contesti caratterizzati da elevata eterogeneità, all'interno dei quali le differenze tra i soggetti non possono essere rappresentate adeguatamente mediante l'utilizzo di semplici classificazioni definite a priori. In tali contesti, l'utilizzo di tecniche statistiche permette di identificare segmenti potenzialmente più omogenei e maggiormente rappresentativi della realtà.

Tuttavia, pur garantendo una maggiore flessibilità, questo approccio è anche caratterizzato da una maggiore complessità metodologica. La definizione dei segmenti dipende infatti da numerose decisioni analitiche, quali la scelta delle variabili, delle misure di distanza e delle procedure di classificazione adottate. Inoltre, è rilevante ricordare che, sebbene questo approccio sia spesso percepito come "più sofisticato" (Dolnicar, 2014), l'elevata complessità metodologica aumenta conseguentemente il rischio di commettere errori di analisi e interpretazione dei dati.

1.3.3 Confronto e considerazioni finali

Tabella 1.1 Segmentazione "a priori" VS "a posteriori"

Caratteristica	Segmentazione a priori	Segmentazione a posteriori
Definizione dei gruppi	ex ante	ex post
Approccio	deduttivo	induttivo
Complessità	ridotta	Elevata
Flessibilità	limitata	maggiore
Ruolo dei dati	secondario	centrale

Fonte: Elaborazione personale dell'autore

La crescente disponibilità di dati e lo sviluppo delle tecniche di analisi statistica degli ultimi decenni hanno favorito una progressiva diffusione degli approcci di segmentazione a posteriori. Con particolare riferimento alla segmentazione *data-driven*, hanno assunto un ruolo fondamentale le tecniche di *clustering*, poiché consentono di individuare, a partire dalle caratteristiche osservate delle unità analizzate, gruppi omogenei dotati di caratteristiche simili al loro interno. Per tale ragione, la *cluster analysis* rappresenta oggi uno degli strumenti maggiormente utilizzati nella ricerca turistica e verrà approfondita nella sezione successiva.

1.4 La *cluster analysis* come strumento di segmentazione *data-driven* nel settore turistico

Come è emerso nelle sezioni precedenti, la segmentazione a posteriori si differenzia dagli approcci tradizionali poiché i segmenti non vengono definiti a priori dal ricercatore, ma emergono direttamente dai dati osservati. Nell'ambito della ricerca turistica, lo sviluppo di tale approccio ha favorito una crescente diffusione delle tecniche statistiche multivariate.

In un contesto come quello turistico, caratterizzato da un'elevata eterogeneità della domanda, gli approcci *data-driven* si sono diffusi sempre di più come strumenti capaci di individuare segmenti maggiormente aderenti alla reale struttura del mercato. Tra gli strumenti maggiormente utilizzati negli studi di segmentazione a posteriori assume particolare rilevanza la *cluster analysis*. Essa consente di raggruppare osservazioni caratterizzate da profili simili, individuando segmenti caratterizzati da elevata omogeneità interna ed eterogeneità esterna. Secondo Dolnicar (2002), la *cluster analysis* rappresenta la procedura impiegata con maggiore frequenza negli studi di segmentazione turistica.

La *review* della letteratura condotta da Dolnicar (2002) mostra con particolare evidenza come la *cluster analysis* sia stata applicata a una vasta gamma di contesti turistici, che includono differenti tipologie di destinazioni, prodotti e segmenti di mercato.

Le applicazioni della *cluster analysis* nel settore turistico si sono concentrate inizialmente sulla segmentazione dei visitatori, per poi estendersi progressivamente anche alla classificazione delle destinazioni e dei territori. Nelle sezioni successive verranno presentati alcuni tra i principali contributi della letteratura scientifica, attraverso i quali verrà evidenziata l'evoluzione delle applicazioni della segmentazione *data-driven* nel settore turistico.

1.4.1 Applicazione della *cluster analysis* alla segmentazione dei turisti

Nel settore turistico, le prime applicazioni della *cluster analysis* si sono concentrate prevalentemente sulla segmentazione della domanda. L'obiettivo principale consisteva nell'individuare gruppi omogenei di visitatori caratterizzati da preferenze, motivazioni e comportamenti simili. Per tale ragione, la segmentazione della domanda è considerata uno strumento fondamentale in grado di sviluppare prodotti e strategie promozionali coerenti con le esigenze dei differenti gruppi di consumatori, e conseguentemente di aumentare l'efficacia delle attività di comunicazione e l'efficienza nell'allocazione delle risorse.

Nella letteratura scientifica dei primi anni Duemila emerge chiaramente l'importanza attribuita alla segmentazione nel settore turistico. Dolnicar e Leisch (2003), con riferimento a una meta-analisi[6] condotta da Baumann (2000), evidenziano una crescita significativa nel corso degli anni Novanta del Novecento dell'interesse verso le tecniche di segmentazione. La maggior parte degli studi analizzati aveva come obiettivo principale l'identificazione di gruppi omogenei di consumatori, a conferma nuovamente del ruolo centrale della segmentazione nell'ambito del marketing turistico. La diffusione di questo strumento riflette la crescente necessità di comprendere e soddisfare la domanda di un settore caratterizzato da elevata eterogeneità e da comportamenti difficilmente riconducibili a classificazioni tradizionali.

Gli stessi autori osservano tuttavia come gran parte della letteratura precedente tende ad utilizzare una sola tipologia di variabili per la segmentazione. In particolare, molti studi utilizzano esclusivamente criteri psicografici oppure comportamentali, trascurando la natura multidimensionale del comportamento turistico. Dolnicar e Leisch (2003) sostengono infatti che tale approccio rischia di fornire una rappresentazione parziale del mercato turistico, perché le scelte di viaggio derivano dall'interazione tra più fattori, quali motivazioni, preferenze, attività svolte e caratteristiche individuali del singolo visitatore.

[6] La meta-analisi è una tecnica statistica che combina dati qualitativi di molteplici studi indipendenti condotti su un determinato argomento, al fine di ottenere una stima complessiva dell'effetto di un intervento o di un fattore di rischio. Si basa sull'idea che l'aggregazione dei dati provenienti da diversi studi possa fornire una visione più completa e accurata del fenomeno di interesse (Analisi-Statistiche.it).

Per superare tale limite, Dolnicar e Leisch (2003) propongono il concetto di *vacation styles*, secondo cui i segmenti non dovrebbero essere costruiti sulla base di una singola dimensione, ma dovrebbero riflettere contemporaneamente aspetti comportamentali e psicografici. L'obiettivo di tale approccio consiste nell'individuare gruppi di turisti caratterizzati da stili di vacanza distinti, in grado di rappresentare in modo più completo la complessità del comportamento turistico. Applicando tecniche di *clustering* ai dati dell'*Austrian National Guest Survey*, gli autori identificano diversi profili turistici invernali, evidenziando come l'utilizzo di approcci che prevedono l'interazione tra più dimensioni informative consenta di ottenere segmenti maggiormente interpretabili e utili ai fini delle strategie di marketing.

Un ulteriore contributo dello studio condotto da Dolnicar e Leisch (2003) riguarda il tema della stabilità dei risultati. Gli autori evidenziano infatti come solo una piccola parte degli studi di segmentazione turistica affronti esplicitamente il problema della robustezza delle soluzioni ottenute. Tale aspetto assume particolare rilevanza, poiché i risultati della *cluster analysis* possono essere influenzati dalle scelte metodologiche adottate, rendendo necessario valutare con attenzione la coerenza e l'interpretabilità dei segmenti individuati. Inoltre, l'affidabilità della segmentazione costituisce un elemento chiave per la successiva definizione di strategie operative e di marketing.

Un diverso esempio di applicazione della *cluster analysis* alla segmentazione della domanda turistica è fornito da Park e Yoon (2009), che analizzano il turismo rurale in Corea del Sud. Gli autori utilizzano come principale base segmentativa le motivazioni di viaggio, partendo dal presupposto che esse rappresentino uno dei fattori più significativi per comprendere il comportamento dei turisti. Attraverso la raccolta di dati relativi alle motivazioni turistiche e l'applicazione di tecniche statistiche multivariate, individuano differenti dimensioni motivazionali utilizzate successivamente per la costruzione dei segmenti.

Nello studio di Park e Yoon (2009) l'applicazione della *cluster analysis* consente di identificare quattro segmenti distinti di turisti rurali, caratterizzati da differenti combinazioni di motivazioni e aspettative. I risultati evidenziano come visitatori apparentemente simili possano presentare esigenze molto differenti in termini di ricerca di nuove esperienze, apprendimento, relax e relazioni familiari. Lo studio dimostra pertanto come la segmentazione basata su criteri motivazionali possa costituire uno strumento efficace per identificare nicchie di mercato e sviluppare un'offerta turistica maggiormente coerente con le aspettative dei turisti.

In generale, gli studi analizzati evidenziano come la *cluster analysis* abbia avuto un ruolo significativo nell'evoluzione degli approcci di segmentazione della domanda turistica. Rispetto agli approcci tradizionali basati esclusivamente su variabili socio-demografiche, le tecniche *data-driven* consentono di cogliere la multidimensionalità del comportamento turistico e di identificare segmenti maggiormente omogenei e interpretabili. Tali caratteristiche hanno favorito la diffusione della *cluster analysis* come uno degli strumenti più utilizzati nella segmentazione dei turisti e hanno gettato le basi metodologiche per le successive applicazioni rivolte non solo ai turisti, ma anche alle destinazioni e ai territori, oggetto della sezione successiva.

1.4.2 Applicazione della *cluster analysis* alla segmentazione delle destinazioni e dei territori

L'evoluzione della letteratura scientifica ha progressivamente ampliato l'ambito di applicazione delle tecniche di *clustering*, estendendolo dalla segmentazione della sola domanda turistica anche a quella delle destinazioni e dei territori turistici. La diffusione degli studi finalizzati alla classificazione delle destinazioni e dei territori turistici è stata favorita dalla crescente disponibilità di dati territoriali ad elevata risoluzione[7] e dallo sviluppo di tecniche computazionali e dei sistemi informativi geografici. In tale contesto, la *cluster analysis* non viene utilizzata per identificare gruppi omogenei di visitatori, ma per individuare territori caratterizzati da profili turistici simili.

Batista et al. (2021) osservano come la crescente disponibilità di *Big Data*[8] e statistiche ufficiali[9] abbia aperto nuove opportunità per l'analisi territoriale in ambito turistico. Gli autori sottolineano tuttavia come l'integrazione tra le tipologie di fonti informative sopracitate risulti ancora poco esplorata su scala europea. Partendo dall'ipotesi che la localizzazione delle strutture ricettive costituisca un indicatore della reale distribuzione spaziale delle attività

[7] Per dati territoriali ad elevata risoluzione si intendono dati georeferenziati disponibili ad un livello di dettaglio spaziale particolarmente elevato, tali da consentire l'analisi di fenomeni localizzati su porzioni limitate di territorio.

[8] I *Big Data* sono insiemi di dati caratterizzati da elevato volume, elevata velocità di generazione e notevole varietà delle fonti informative, la cui raccolta, gestione e analisi richiedono strumenti computazionali avanzati e tecniche analitiche specifiche (Li et al., 2018).

[9] Per statistiche ufficiali si intendono dati prodotti e diffusi da istituzioni statistiche nazionali o sovranazionali mediante metodologie standardizzate e procedure finalizzate a garantirne qualità, affidabilità e comparabilità nel tempo e nello spazio.

turistiche, gli autori propongono una classificazione delle regioni europee basata sulla distribuzione e sulla capacità ricettiva delle strutture alberghiere.

Lo studio condotto da Batista et al. (2021) propone un'analisi su 1165 regioni NUTS3[10] appartenenti all'Unione Europea, considerando quattro differenti contesti geografici: aree costiere, aree montane e naturali, città e aree rurali. Attraverso l'utilizzo, come variabili di input, della quota di capacità ricettiva localizzata in ciascuna tipologia territoriale, gli autori applicano una procedura di *clustering* gerarchico, ottenendo una classificazione in cinque differenti tipologie di regioni turistiche, ossia cinque *cluster*. I risultati mostrano come la *cluster analysis* possa essere impiegata efficacemente al fine di individuare modelli territoriali omogenei e analizzare le relazioni tra caratteristiche geografiche, offerta ricettiva e dinamiche della domanda turistica. Inoltre, gli autori evidenziano come tale classificazione possa supportare la previsione della domanda, lo studio delle dinamiche spazio-temporali[11] del turismo e la definizione di strategie di pianificazione territoriale.

Un ulteriore contributo della letteratura scientifica in questa direzione è rappresentato dallo studio di Glyptou et al. (2022), che propongono un approccio alla segmentazione delle destinazioni turistiche basata sulle performance di sostenibilità. Gli autori osservano come la maggior parte degli studi di segmentazione si concentri sulla domanda turistica o sulle caratteristiche naturali e culturali delle destinazioni, mentre risultano ancora poco esplorate le applicazioni finalizzate alla costruzione di tipologie territoriali fondate sui processi di sviluppo e sulle performance gestionali[12] delle destinazioni.

Lo studio considera undici destinazioni costiere appartenenti a diversi paesi del Mediterraneo e costituisce un insieme di indicatori articolato in tre dimensioni principali: impronta ambientale, dipendenze economiche dal turismo e prosperità della popolazione residente. Attraverso un approccio integrato basato su analisi fattoriale[13] e *clustering* gerarchico, gli autori

[10] NUTS3 (*Nomenclature of Territorial Units for Statistics – livello 3*) indica una suddivisione territoriale adottata dall'Unione Europea a fini statistici, generalmente corrispondente a province, dipartimenti o distretti amministrativi (Eurostat, 2024).

[11] Per dinamiche spazio-temporali del turismo si intendono le variazioni nella distribuzione geografica, nell'intensità e nella stagionalità dei flussi turistici osservate nel corso del tempo.

[12] Per performance gestionali delle destinazioni si intendono i risultati conseguiti da una destinazione turistica in termini economici, sociali e ambientali, nonché l'efficacia delle strategie adottate per il suo sviluppo, la sua promozione e la gestione delle risorse disponibili.

[13] Per analisi fattoriale si intende una tecnica statistica multivariata finalizzata a ridurre il numero di variabili osservate mediante l'individuazione di un insieme più ristretto di fattori latenti in grado di spiegare le correlazioni esistenti tra le variabili originarie.

identificano differenti gruppi di destinazioni (*cluster*) caratterizzati da specifiche combinazioni di risorse naturali e culturali, stagionalità dell'offerta, tipologie ricettive prevalenti e profilo della domanda turistica.

Nel complesso, gli studi analizzati evidenziano come la *cluster analysis* possa costituire uno strumento particolarmente efficace per la segmentazione delle destinazioni e dei territori turistici. Rispetto ai tradizionali approcci costruiti a priori, le tecniche di segmentazione *data-driven* consentono di individuare strutture territoriali emergenti a partire dai dati osservati, favorendo una migliore comprensione delle dinamiche turistiche e supportando i processi di pianificazione e gestione delle destinazioni turistiche. Tali applicazioni risultano particolarmente rilevanti anche nel contesto del turismo montano, oggetto di recenti studi e del presente elaborato.

1.4.3 Evidenze recenti nel turismo alpino e collegamento con il caso di studio

Negli ultimi anni, l'applicazione delle tecniche di segmentazione e classificazione statistica al turismo alpino ha assunto sempre più rilevanza, anche in risposta alla scarsa disponibilità di informazioni sistematiche e comparabili sui flussi turistici nelle aree montane. La letteratura internazionale evidenzia infatti come le statistiche ufficiali, generalmente basate sulle strutture ricettive, tendano a sottostimare l'effettiva intensità dei flussi turistici in quanto non considerano fenomeni quali l'escursionismo giornaliero, i soggiorni nelle seconde case e le forme di ospitalità informale. Tale problematica risulta particolarmente ricorrente nel contesto alpino italiano, dove una quota significativa dei flussi turistici non viene intercettata dai tradizionali sistemi di monitoraggio.

In questo contesto diventa particolarmente rilevante lo studio condotto da Visintin et al. (2026), il quale ha l'obiettivo di approfondire le caratteristiche della domanda turistica nelle Alpi italiane attraverso un'indagine campionaria condotta tra dicembre 2023 e gennaio 2024 su 1218 residenti dell'Italia nord-occidentale, nord-orientale e del Friuli-Venezia-Giulia. Lo studio si propone di colmare il deficit informativo esistente analizzando i flussi turistici, le caratteristiche sociodemografiche, le preferenze e i comportamenti dei turisti delle aree alpine italiane. In tale prospettiva, gli autori combinano statistiche descrittive e analisi inferenziali con tecniche di *clustering*, adottando un approccio di segmentazione dei visitatori basato sulle attività svolte e sulle motivazioni di viaggio.

Dal punto di vista metodologico, Visintin et al. (2026) applicano l'algoritmo *K-means* testando differenti configurazioni, con un numero di *cluster* compreso tra due e cinque. La selezione

delle variabili avviene distinguendo tra fattori *push* e *pull*, considerando quindi sia la loro rilevanza teorica che la loro capacità di descrivere efficacemente i comportamenti turistici. Dopo aver effettuato la standardizzazione delle variabili, gli autori confrontano diversi modelli di *clustering* e individuano nella soluzione a tre gruppi (k=3) quella maggiormente soddisfacente in termini di convergenza, distanza tra i centroidi e riduzione della variabilità interna ai *cluster*. La robustezza statistica della soluzione individuata viene inoltre verificata attraverso un'analisi della varianza (ANOVA) e procedure post-hoc[14]. L'applicazione del *K-means* consente di identificare tre profili distinti di turisti alpini: gli *Active Young Enthusiasts*, caratterizzati da un'elevata partecipazione ad attività outdoor e motivati dalla ricerca di arricchimento culturale e benessere psicologico; i *Well-being-Oriented Walkers*, maggiormente orientati ad attività ricreative a bassa intensità e al rilassamento; e gli *Hiking-Oriented Explorers*, contraddistinti da una spiccata propensione per le escursioni e per attività associate al benessere psico-fisico. Inoltre, dallo studio emergono significative differenze nei comportamenti di visita in funzione dell'età, del reddito e dell'area geografica di residenza, confermando l'elevata eterogeneità della domanda turistica nelle regioni alpine italiane.

Nonostante l'obiettivo di Visintin et al. (2026) sia principalmente quello di segmentare i visitatori piuttosto che classificare le destinazioni o i territori, il contributo del loro studio è particolarmente rilevante nel contesto del presente elaborato. I risultati ottenuti confermano infatti la natura multidimensionale del fenomeno turistico alpino ed evidenziano l'utilità delle tecniche di *clustering* nell'individuazione di configurazioni omogenee all'interno di sistemi territoriali complessi, come quelli montani. In questa prospettiva, lo studio effettuato nel presente elaborato si differenzia dall'approccio adottato da Visintin et al. (2026) poiché assume come unità di analisi i comuni della Valle d'Aosta e utilizza indicatori relativi agli arrivi e alle presenze per tipologia di struttura ricettiva, con l'obiettivo di individuare gruppi territoriali accomunati da simili caratteristiche di domanda turistica.

Conclusione

La review della letteratura ha evidenziato come la segmentazione rappresenti uno degli strumenti maggiormente impiegati per analizzare la complessità dei fenomeni turistici, consentendo di individuare gruppi omogenei caratterizzati da profili simili e di supportare i processi di analisi e pianificazione delle destinazioni turistiche. L'evoluzione della ricerca

[14] Una procedura post-hoc consiste in un insieme di test statistici effettuati successivamente all'analisi della varianza (ANOVA) per individuare le coppie di gruppi che presentano differenze statisticamente significative tra loro.

scientifica ha inoltre mostrato una progressiva diffusione degli approcci di segmentazione *data-driven*, nei quali i segmenti emergono direttamente dai dati osservati attraverso l'impiego di tecniche di *clustering*.

In sintesi, la letteratura esaminata mostra come la *cluster analysis* abbia trovato applicazione sia nella segmentazione dei visitatori sia nella classificazione delle destinazioni e dei territori. La *review* ha inoltre consentito di mettere in luce le potenzialità di tale strumento a supporto della comprensione dei sistemi turistici. In particolare, la letteratura più recente evidenzia un crescente interesse verso il turismo alpino e la necessità di strumenti statistici in grado di coglierne l'eterogeneità.

Il collegamento con tale contesto ha pertanto consentito di collocare il caso di studio della Valle d'Aosta all'interno di un filone di ricerca attuale, nel quale le tecniche di *clustering* assumono un ruolo sempre più rilevante nell'interpretazione delle dinamiche turistiche territoriali. In tale prospettiva, il caso di studio analizzato nel presente elaborato si propone di fornire una migliore comprensione delle dinamiche dei flussi turistici della Valle d'Aosta, attraverso l'applicazione di tecniche di *clustering* ai comuni valdostani.

2. METODOLOGIA STATISTICA

Introduzione

Il presente capitolo illustra i principali fondamenti teorici e metodologici delle tecniche statistiche adottate nel presente elaborato. Dopo una breve introduzione ai principi dell'analisi dei dati e dell'approccio di segmentazione *data-driven*, vengono presentati i concetti necessari per comprendere il funzionamento delle tecniche di *clustering* impiegate nell'analisi empirica.

In particolare, il capitolo approfondisce il ruolo delle misure di distanza, quale strumento per quantificare il grado di similarità tra le osservazioni, descrive il funzionamento dei principali metodi di *clustering* gerarchico e non gerarchico, con particolare riferimento al metodo del legame completo (*Complete Linkage*), al metodo di *Ward* e all'algoritmo *K-means*, evidenziandone i principali vantaggi, i limiti e le implicazioni metodologiche. Vengono inoltre messi a confronto i diversi approcci analizzati, al fine di evidenziarne le caratteristiche distintive e di motivare le scelte metodologiche adottate nel presente elaborato.

Le conoscenze sviluppate nel corso del capitolo costituiscono il quadro di riferimento teorico necessario per comprendere le procedure statistiche applicate nel capitolo successivo, dedicato all'analisi del sistema turistico della Valle d'Aosta.

2.1 Introduzione all'analisi di dati: segmentazione data driven

Nel presente elaborato, l'analisi dei dati viene condotta con l'obiettivo di individuare gruppi omogenei di unità statistiche sulla base delle loro caratteristiche, attraverso un approccio di tipo *data-driven.*

Per approccio *data-driven* si intende un metodo di analisi basato sull'utilizzo diretto dei dati per individuare strutture e relazioni significative, senza fare affidamento su criteri stabiliti a priori (Dolnicar et al., 2018). In questo contesto, i dati rappresentano la principale fonte informativa per guidare il processo decisionale e supportare l'individuazione di gruppi omogenei riducendo il ricorso a valutazioni puramente intuitive (Provost & Fawcett, 2013).

La segmentazione di mercato rappresenta uno strumento fondamentale per il marketing che mira a raggruppare i consumatori in gruppi con bisogni o comportamenti simili. Smith (1956) definisce la segmentazione di mercato come il considerare un mercato eterogeneo (caratterizzato da una domanda a sua volta eterogena) come un insieme di mercati più piccoli e omogenei. Nel presente elaborato si tratta più precisamente di un processo volto a suddividere un insieme eterogeneo di unità statistiche in gruppi più omogenei, detti segmenti o *cluster*, in

modo tale che le unità appartenenti allo stesso *cluster* siano simili tra loro e differenti rispetto a quelle appartenenti ad altri gruppi.

Nel contesto della presente analisi tale approccio risulta particolarmente adatto in quanto consente di analizzare i dati relativi ai flussi turistici e alle strutture ricettive nei comuni della Valle d'Aosta mediante l'impiego di metodi statistici. Attraverso l'uso delle tecniche di *clustering*, sarà infatti possibile individuare gruppi omogenei di comuni caratterizzati da dinamiche turistiche simili, fornendo una rappresentazione sintetica e interpretabile delle caratteristiche del fenomeno analizzato.

2.2 Misure di distanza

Nel contesto dell'analisi dei flussi turistici è utile individuare gruppi di comuni caratterizzati da dinamiche simili. I comuni della Valle d'Aosta presentano infatti caratteristiche differenti in termini di presenze turistiche e offerta ricettiva, rendendo necessario individuare eventuali similarità tra di essi.

Al fine di raggruppare i comuni in base a tali caratteristiche, è necessario disporre di una misura che consenta di quantificare il grado di somiglianza o dissimilarità tra le unità considerate. In ambito statistico, tale concetto è formalizzato attraverso l'utilizzo delle misure di distanza.

È quindi necessario, prima di affrontare le tecniche di *clustering*, approfondire il concetto di misura di distanza nell'analisi dei dati. Nel presente elaborato viene fatto particolare riferimento alla distanza euclidea, alla quale sarà dedicata maggiore attenzione, in quanto tra le diverse misure di distanza esistenti essa è particolarmente adatta all'analisi di variabili quantitative.

Una matrice di dati rappresenta la struttura tipica con cui vengono organizzate le informazioni nell'analisi dei dati. In essa, ogni riga rappresenta un'osservazione, mentre ogni colonna rappresenta una variabile. Dal punto di vista matematico, tale struttura può essere rappresentata mediante una matrice $(n \times p)$, dove (n) indica il numero di osservazioni e (p) indica il numero di variabili.

A ciascuna riga della matrice corrisponde un vettore di osservazioni che descrive i valori assunti dalle diverse variabili per una determinata unità statistica. L'insieme di tutti i vettori rappresenta quindi l'insieme delle osservazioni considerate.

Per poter confrontare tra loro le diverse unità statistiche, è necessario introdurre una misura che consenta di quantificare quanto due osservazioni siano simili o differenti. A tal fine, si utilizzano le cosiddette misure di distanza, che rappresentano una funzione $d(\cdot,\cdot)$ con due argomenti, in

grado di associare a ogni coppia di osservazioni un valore numerico non negativo, interpretabile come il grado di dissimilarità tra esse.

Un modo intuitivo per comprendere il concetto di distanza può essere quello di fare riferimento alla distanza tra due città. Essa può essere interpretata, ad esempio, come la lunghezza del percorso che le separa. Analogamente, in statistica, la distanza tra due osservazioni misura quanto esse siano "lontane" nello spazio delle variabili considerate.

Una misura di distanza deve soddisfare alcune proprietà fondamentali, al fine di garantire una corretta interpretazione del concetto di somiglianza tra le osservazioni. In particolare, essa deve essere simmetrica, ovvero la distanza tra due osservazioni non dipende dall'ordine in cui queste vengono considerate; tale proprietà è necessaria affinché il confronto tra le unità statistiche sia coerente.

$$d(x, y) = d(y, x)$$

Inoltre, la distanza deve essere nulla solo nel caso in cui le due osservazioni coincidano, poiché ciò implica che non esista alcuna differenza tra di esse.

$$d(x, y) = 0 \Leftrightarrow x = y$$

Questa proprietà consente di distinguere in modo chiaro tra osservazioni identiche e osservazioni differenti.

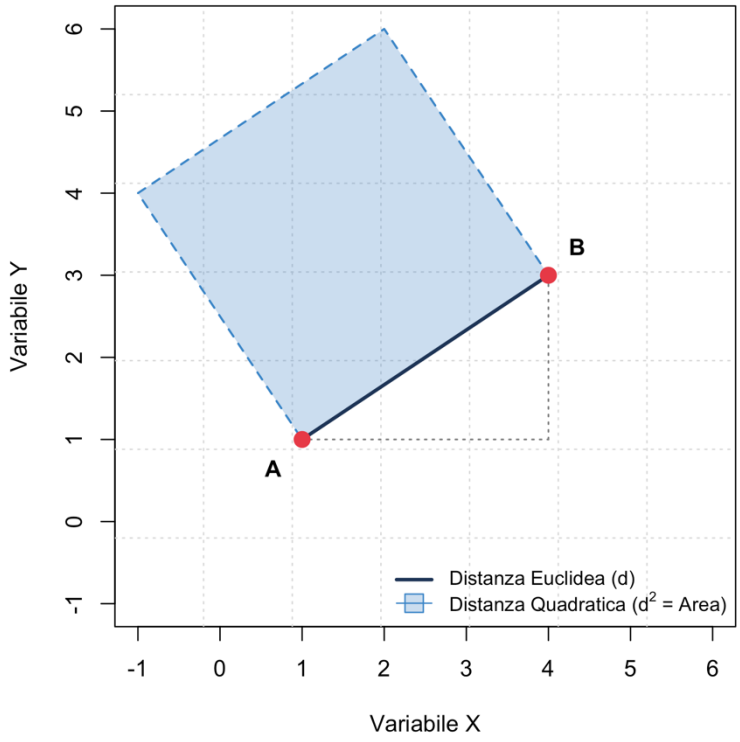
2.2.1 Distanza euclidea

La distanza euclidea rappresenta la misura più utilizzata nell'ambito delle tecniche di *clustering* e corrisponde alla radice quadrata della somma dei quadrati delle differenze tra i valori delle variabili delle osservazioni considerate.

$$d(x, y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2} \tag{2.1}$$

Essa può essere interpretata come la distanza "in linea retta" tra due punti nello spazio delle variabili, in particolare in uno spazio bidimensionale (Figura 2.1). La distanza euclidea misura quindi quanto due osservazioni siano "lontane" considerando tutte le variabili simultaneamente. In questo modo, la distanza riflette le differenze complessive tra le osservazioni, anziché concentrarsi esclusivamente su una singola variabile.

Figura 2.1 Confronto geometrico tra distanza euclidea (d) e distanza euclidea quadratica (d^2) in uno spazio bidimensionale. La linea continua indica la distanza lineare tra i punti A e B, mentre la superficie poligonale tratteggiata rappresenta l'area corrispondente alla distanza quadratica.



Fonte: Elaborazione personale dell'autore

2.2.2 Distanza euclidea quadratica

In alcuni metodi di *clustering*, come il metodo di *Ward* che verrà trattato nelle sezioni successive, viene utilizzata la distanza euclidea quadratica, ottenuta considerando il quadrato della distanza euclidea tradizionale:

$$d^2(x,y) = \sum_{j=1}^{p} (x_j - y_j)^2 \tag{2.2}$$

Tale formulazione attribuisce maggiore peso alle osservazioni caratterizzate da distanze più elevate ed è particolarmente utile nei metodi che hanno l'obiettivo di minimizzare la variabilità[1] interna ai *cluster* (Everitt et al., 2011; Dolnicar et al., 2018).

È importante sottolineare che, nel caso in cui le variabili considerate presentino scale di misura differenti, il calcolo della distanza può risultare influenzato dalle variabili caratterizzate da valori più elevati, le quali tendono a dominare il calcolo della distanza. Per tale motivo, si rende

[1] La variabilità esprime il grado di dispersione dei valori osservati rispetto a una misura di posizione centrale, ad esempio la media, o nel caso del *clustering*, il centroide dei *cluster*; tanto maggiore è la variabilità, tanto più le osservazioni risultano differenti tra loro e viceversa.

spesso necessario procedere a una standardizzazione dei dati, al fine di rendere confrontabili le variabili e garantire un corretto calcolo delle distanze prima di applicare le tecniche di *clustering* (Dolnicar et al., 2018).

La standardizzazione consiste nel trasformare le variabili rendendole confrontabili tra loro, generalmente attraverso la sottrazione della media e la divisione per la deviazione standard, così da ottenere variabili caratterizzate dalla medesima scala di misura.

Nel caso specifico dell'analisi dei flussi turistici, variabili quali il numero di arrivi, il numero di presenze e le caratteristiche delle strutture ricettive possono presentare scale di misura differenti. Di conseguenza, si rende necessario procedere alla standardizzazione dei dati, al fine di evitare che le variabili con valori più elevati influenzino in maniera predominante il calcolo delle distanze.

2.3 Tecniche di *clustering*

Le tecniche di *clustering* rappresentano strumenti di analisi statistica utilizzati per suddividere un insieme eterogeneo di osservazioni in gruppi omogenei, detti *cluster*, sulla base delle caratteristiche considerate. L'obiettivo principale consiste nel raggruppare nello stesso *cluster* le osservazioni maggiormente simili tra loro e separare quelle caratterizzate da differenze più marcate. Tali tecniche rientrano nell'ambito dell'apprendimento non supervisionato (*unsupervised learning*), poiché i gruppi vengono individuati automaticamente a partire dalla struttura dei dati e non sono definiti a priori (Dolnicar et al., 2018).

Nel presente elaborato, le tecniche di *clustering* verranno utilizzate per individuare i gruppi omogenei di comuni della Valle d'Aosta, caratterizzati da dinamiche turistiche (flussi turistici) e caratteristiche ricettive simili. In particolare, l'analisi consentirà di evidenziare eventuali similitudini tra i comuni in termini di presenze, arrivi turistici e struttura dell'offerta ricettiva, fornendo una rappresentazione sintetica e interpretabile del fenomeno analizzato.

Le principali tecniche di *clustering* possono essere suddivise in due grandi macrocategorie: il *clustering* gerarchico e il *clustering* non gerarchico.

2.4 *Clustering* gerarchico (*hierarchical methods)*

Il *clustering* gerarchico rappresenta una delle metodologie maggiormente utilizzate nell'ambito dell'analisi esplorativa dei dati[2]. Tale approccio costruisce progressivamente una struttura gerarchica di *cluster* attraverso un processo iterativo di aggregazione delle osservazioni.

Esso può essere suddiviso in approcci agglomerativi e divisivi. Nel *clustering* divisivo, si parte da un unico *cluster* contenente tutte le osservazioni, che viene successivamente suddiviso in gruppi più piccoli. Nel *clustering* agglomerativo, invece, inizialmente ogni osservazione costituisce un *cluster* autonomo. Successivamente, ad ogni iterazione[3], vengono aggregati i due *cluster* caratterizzati dalla minore distanza reciproca. Il procedimento continua fino alla formazione di un unico *cluster* contenente tutte le osservazioni del dataset. Nei metodi di *clustering* gerarchico, il processo di aggregazione segue una sequenza deterministica di operazioni. Ciò significa che, applicando l'algoritmo[4] allo stesso dataset e utilizzando il medesimo criterio di aggregazione, si ottiene sempre la stessa struttura gerarchica di *cluster*. Nel presente elaborato verrà utilizzato principalmente l'approccio gerarchico agglomerativo, in quanto rappresenta la metodologia maggiormente utilizzata nelle applicazioni statistiche (Dolnicar et al., 2018).

Alla base del *clustering* gerarchico, vi è la matrice delle distanze, costruita a partire dalle distanze calcolate tra tutte le coppie di osservazioni. Sia n il numero di osservazioni presenti nel dataset e k il numero di *cluster* che si desidera individuare. Le tecniche di *clustering* consentono di suddividere le n osservazioni in k gruppi omogenei sulla base delle caratteristiche considerate. Nel caso del *clustering* gerarchico, siano X e Y due insiemi di osservazioni appartenenti a *cluster* differenti: la distanza tra i *cluster* viene determinata a partire dalle distanze calcolate tra le osservazioni appartenenti ai due gruppi. Nel presente elaborato verrà utilizzata principalmente la distanza euclidea quadratica illustrata nella sezione 2.2.2 del capitolo 2.

---

[2] Per analisi esplorativa dei dati (*Exploratory Data Analysis,* EDA) si intende l'insieme di tecniche statistiche e grafiche utilizzate per sintetizzare, visualizzare ed esplorare le principali caratteristiche dei dati, al fine di individuare pattern, relazioni e anomalie, prima dell'applicazione di modelli statistici più complessi (Komorowski et al., 2016)

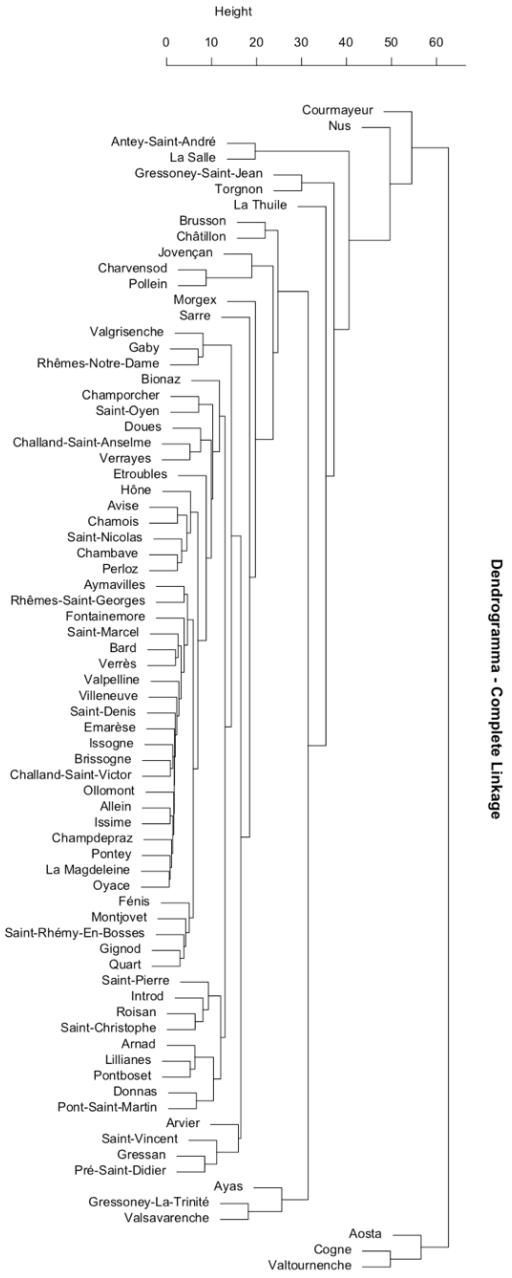
[3] Per iterazione si intende ciascun passaggio del procedimento algoritmico nel quale vengono eseguite specifiche operazioni e aggiornati i risultati ottenuti nella fase precedente. Nel *clustering* gerarchico agglomerativo, ogni iterazione corrisponde all'aggregazione dei due *cluster* caratterizzati dalla minore distanza reciproca.

[4] Per algoritmo si intende una procedura automatica composta da una sequenza di operazioni matematiche e statistiche che consentono di raggruppare progressivamente le osservazioni sulla base delle loro similarità.

2.4.1 Dendrogramma

Il risultato del procedimento iterativo dei metodi di *clustering* gerarchico viene generalmente rappresentato attraverso un dendrogramma, ovvero una rappresentazione grafica ad albero che mostra le fusioni effettuate tra i *cluster* nei diversi passaggi dell'algoritmo. Il dendrogramma consente di visualizzare il livello di similarità tra le osservazioni e può risultare utile nell'individuazione del numero di *cluster* appropriato.

Figura 2.2 Dendrogramma – Complete Linkage



Fonte: Elaborazione personale dell'autore

Esistono molteplici criteri atti a definire la distanza tra *cluster* e, di conseguenza, molteplici metodi di *clustering* gerarchico. Tra i principali criteri di aggregazione utilizzati vi sono il metodo del legame completo (*Complete Linkage*) e il metodo di *Ward*.

2.4.2 Metodo del legame completo (*Complete Linkage*)

Nel *clustering* gerarchico, la distanza tra *cluster* viene determinata attraverso una funzione di *linkage* $l(X, Y)$ che definisce il criterio di aggregazione tra gruppi di osservazioni (Everitt et al., 2011). Il metodo del legame completo definisce la distanza tra due *cluster* come la massima distanza tra le osservazioni appartenenti ai due gruppi:

$$l(X, Y) = \max_{x \in X, y \in Y} d(x, y) \tag{2.3}$$

Dove $d(x, y)$ rappresenta la distanza tra le osservazioni x e y, mentre X e Y rappresentano due *cluster* distinti. Tale approccio tende generalmente a produrre *cluster* relativamente compatti e ben separati tra loro (Dolnicar et al., 2018).

Nel procedimento gerarchico agglomerativo, il metodo del legame completo opera attraverso fusioni successive dei *cluster*. Ad ogni iterazione, vengono confrontati tutti i *cluster* presenti e vengono aggregati i due gruppi caratterizzati dalla minore distanza massima reciproca. Successivamente, la matrice delle distanze viene aggiornata e il procedimento continua fino al raggiungimento della struttura gerarchica finale (Everitt et al., 2011).

2.4.3 Metodo di *Ward* (*Ward* method)

Il metodo di *Ward*, proposto da *Ward* (1963), rappresenta uno dei criteri di aggregazione maggiormente utilizzati nell'ambito del *clustering* gerarchico. Tale metodo si basa sulla minimizzazione della variabilità interna ai *cluster* attraverso l'utilizzo delle distanze euclidee quadratiche. In particolare, esso aggrega i gruppi di osservazioni in modo tale da ridurre il più possibile la dispersione interna ai *cluster*, favorendo la formazione di gruppi relativamente omogenei e compatti. Ciò significa che l'algoritmo tende ad aggregare osservazioni tra loro molto simili, cercando di ottenere *cluster* quanto più possibile omogenei al loro interno (Everitt et al., 2011; Dolnicar et al., 2018).

Ad ogni iterazione, vengono aggregati i due *cluster* che determinano il minore incremento possibile della variabilità complessiva. Il metodo considera la distanza tra i centroidi dei *cluster*, ovvero i punti centrali ottenuti, calcolando la media delle osservazioni appartenenti a ciascun gruppo. I centroidi possono quindi essere interpretati come rappresentativi dei *cluster* stessi

(Dolnicar et al., 2018), perché sintetizzano le caratteristiche medie delle osservazioni appartenenti a ciascun gruppo.

Il procedimento del metodo di *Ward* si sviluppa attraverso una sequenza iterativa di aggregazioni successive. Inizialmente, ogni osservazione costituisce un *cluster* autonomo (*singleton cluster*). Ad ogni iterazione vengono valutate tutte le possibili fusioni tra *cluster* e viene selezionata quella che determina il minore incremento della variabilità interna complessiva.

Successivamente, le distanze vengono ricalcolate e il procedimento continua fino alla formazione della struttura gerarchica finale (Everitt et al., 2011; Dolnicar et al., 2018).

2.4.4 Vantaggi, svantaggi e limiti del *clustering* gerarchico

Le tecniche di *clustering* gerarchico presentano diversi vantaggi nell'ambito dell'analisi esplorativa dei dati. Uno dei principali punti di forza di tali metodi consiste nella capacità di rappresentare in modo intuitivo la struttura dei dati attraverso il dendrogramma, che consente di visualizzare graficamente il processo di aggregazione dei *cluster* e il livello di similarità tra le osservazioni. Inoltre, i metodi gerarchici non richiedono la definizione preliminare del numero di *cluster* da individuare, permettendo una maggiore flessibilità nell'analisi dei dati.

Un ulteriore vantaggio riguarda il carattere deterministico degli algoritmi gerarchici: applicando lo stesso metodo al medesimo dataset e mantenendo invariati i criteri di distanza e aggregazione, si ottiene sempre la stessa struttura gerarchica di *cluster*. Ciò favorisce la replicabilità dell'analisi e rende i risultati maggiormente interpretabili (Dolnicar et al., 2018), aspetto fondamentale in un protocollo sperimentale.

Tuttavia, i metodi di *clustering* gerarchico presentano anche alcuni limiti. In primo luogo, tali tecniche possono risultare computazionalmente onerose in presenza di data set particolarmente estesi, perché richiedono il calcolo e l'aggiornamento continuo della matrice delle distanze. Inoltre, le fusioni effettuate durante il procedimento sono irreversibili: una volta aggregati, i *cluster* non possono più essere separati nelle iterazioni successive. Di conseguenza, eventuali errori nelle prime fasi del procedimento possono influenzare la struttura finale dei *cluster* (Everitt et al., 2011).

Con riferimento specifico al metodo del legame completo, uno dei principali vantaggi consiste nella capacità di produrre *cluster* relativamente compatti e ben separati tra loro, riducendo il rischio di formazione di gruppi eccessivamente allungati o poco omogenei (Dolnicar et al.,

2018). Tuttavia, tale metodo può risultare sensibile alla presenza di osservazioni anomale (*outlier*), poiché la distanza tra *cluster* viene determinata considerando le osservazioni maggiormente distanti appartenenti ai due gruppi (Everitt et al., 2011).

Per quanto riguarda il metodo di *Ward*, esso rappresenta invece uno dei criteri di aggregazione maggiormente utilizzati grazie alla sua capacità di minimizzare la variabilità interna ai *cluster* e di produrre gruppi generalmente omogenei e relativamente equilibrati in termini di numerosità. Tale approccio risulta particolarmente efficace nell'analisi di variabili quantitative e nell'utilizzo della distanza euclidea quadratica (Dolnicar et al., 2018). D'altra parte, il metodo di *Ward* può risultare influenzato dalla presenza di variabili caratterizzate da scale di misura differenti, rendendo spesso necessaria una preventiva standardizzazione dei dati. Inoltre, come altri metodi gerarchici esso può essere sensibile alla presenza di *outlier* e alla scelta della misura di distanza adottata (Everitt et al., 2011).

2.5 *Clustering* non gerarchico (*partitioning methods*)

Il *clustering* non gerarchico, detto anche *partitional clustering*, rappresenta un'alternativa al metodo gerarchico quando l'obiettivo consiste nel suddividere le osservazioni in un numero di *cluster* fissato a priori. A differenza delle tecniche gerarchiche, tali metodi non producono una struttura gerarchica o un dendrogramma, ma mirano direttamente all'individuazione di una partizione dei dati in k gruppi omogenei.

I metodi non gerarchici risultano particolarmente adatti all'analisi di dataset caratterizzati da un'elevata numerosità di osservazioni, poiché richiedono un numero inferiore di calcoli rispetto ai metodi gerarchici. Infatti, mentre i metodi gerarchici richiedono il calcolo delle distanze tra tutte le coppie di unità statistiche presenti nel dataset, i metodi non gerarchici operano generalmente considerando le distanze tra le osservazioni e i centri dei *cluster*. Ad esempio, un dataset composto da 1.000 osservazioni richiederebbe il calcolo di $(1000 \cdot 999) \div 2 = 499.500$ distanze per la costruzione della matrice completa delle distanze utilizzate dai metodi gerarchici. Un algoritmo di partizionamento, vale a dire un algoritmo non gerarchico, finalizzato all'individuazione di cinque *cluster*, invece, dovrebbe calcolare ad ogni iterazione un numero di distanze compreso tra 5 e 5.000, a seconda della specifica implementazione adottata. Questo evidenzia come i metodi non gerarchici risultino molto più efficienti dal punto di vista computazionale e maggiormente adatti all'analisi di *dataset* di grandi dimensioni (Dolnicar et al., 2018).

2.5.1 Metodo *K-means*

Tra i metodi di *clustering* non gerarchico basati sui centroidi (*centroid-based clustering methods*), il più noto e diffuso è il metodo *K-means*, originariamente proposto da MacQueen (1967). Tale algoritmo appartiene alla categoria dei metodi di partizionamento (*partitioning methods*) e si pone l'obiettivo di suddividere un insieme di n osservazioni in un numero prefissato di k *cluster*, massimizzando l'omogeneità all'interno dei gruppi e la differenziazione tra gruppi distinti. A differenza dei metodi gerarchici, il *K-means* non costruisce una struttura gerarchica di *cluster*, bensì ricerca direttamente una partizione dei dati mediante un procedimento iterativo basato sull'utilizzo dei centroidi.

Sia $X = \{x_1, \dots, x_n\}$ l'insieme delle osservazioni presenti nel dataset. L'obiettivo del metodo consiste nel suddividere tali osservazioni in (k) *cluster* $(S_1, \dots, S_k)$, in modo tale che le osservazioni appartenenti allo stesso gruppo risultino quanto più possibile simili tra loro, mentre quelle appartenenti a gruppi differenti risultino quanto più possibile dissimili tra loro.

Il rappresentante di ciascun *cluster* viene definito centroide (c_j). Nel caso del *K-means* basato sulla distanza euclidea quadratica, il centroide corrisponde al vettore delle medie calcolate per ciascuna variabile, considerando tutte le osservazioni appartenenti al *cluster* (Dolnicar et al., 2018).

$$c_j = \frac{1}{n_j} \sum_{i \in S_j} x_i \qquad j = 1, \dots, k \tag{2.4}$$

Dove:

- S_j è il j-esimo *cluster*;
- $n_j = |S_j|$ è la sua cardinalità;
- x_i è il vettore delle variabili dell'osservazione i.

Tale vettore rappresenta il punto che minimizza la somma delle distanze euclidee quadratiche tra il centroide e le osservazioni del *cluster*. Esso può essere pertanto interpretato come il profilo medio o rappresentativo del gruppo individuato (Dolnicar et al., 2018).

Il metodo *K-means* può quindi essere interpretato come una procedura euristica[5] finalizzata alla risoluzione del problema di ottimizzazione sottostante all'analisi di *clustering*. L'algoritmo

[5] Per euristica si intende una procedura semplificata utilizzata per affrontare problemi complessi mediante regole pratiche che consentono di ottenere rapidamente una soluzione soddisfacente senza esaminare tutte le possibili alternative. Tale concetto è strettamente collegato alla teoria della razionalità limitata (*bounded rationality*)

ricerca progressivamente una partizione dei dati caratterizzata da elevata omogeneità interna ai *cluster* e da una marcata differenziazione tra gruppi distinti. Attraverso una sequenza di passaggi iterativi, esso migliora progressivamente la soluzione ottenuta fino al raggiungimento di una configurazione stabile dei *cluster*. Pur convergendo sempre verso una soluzione finale, il procedimento non garantisce necessariamente il raggiungimento dell'ottimo globale.

Il funzionamento dell'algoritmo può essere descritto attraverso cinque passaggi principali:

1. Definizione del numero di *cluster*

   In primo luogo, viene specificato il numero di *cluster* k che si desidera individuare. Tale scelta rappresenta una delle principali differenze rispetto ai metodi gerarchici, nei quali il numero di *cluster* può essere determinato anche successivamente attraverso l'analisi del dendrogramma.

2. Selezione dei centroidi iniziali

   Successivamente vengono selezionati k centroidi iniziali, generalmente attraverso una procedura casuale che individua k osservazioni del dataset da utilizzare come rappresentanti temporanei dei *cluster*.

   $$C = \{c_1, \ldots, c_k\}$$

   I centroidi iniziali non rappresentano necessariamente una soluzione ottimale, ma costituiscono il punto di partenza necessario per avviare il procedimento iterativo dell'algoritmo (Dolnicar et al., 2018)

3. Assegnazione delle osservazioni ai *cluster*

   In questa fase ciascuna osservazione viene assegnata al *cluster* il cui centroide risulta più vicino secondo la misura di distanza adottata. Formalmente, il *cluster* (S_j) è definito come:

   $$S_j = \{x \in X | d(x, c_j) \leq d(x, c_h), 1 \leq h \leq k\} \quad (2.5)$$

elaborata da Simon (1955), secondo cui individui e organizzazioni, a causa di vincoli cognitivi e informativi, tendono a ricercare soluzioni soddisfacenti (*satisficing*) piuttosto che la soluzione ottimale in senso assoluto. Analogamente, il metodo *K-means* migliora progressivamente la partizione dei dati attraverso una procedura iterativa, pur non garantendo necessariamente il raggiungimento dell'ottimo globale (Simon, 1955; Tversky & Kahneman, 1974; Dolnicar et al., 2018).

Ciò significa che ogni osservazione viene confrontata con tutti i centroidi disponibili e assegnata al gruppo caratterizzato dalla distanza minima. Al termine di questa fase si ottiene una prima partizione del dataset, generalmente ancora lontana dalla soluzione finale.

4. Aggiornamento dei centroidi

   Una volta assegnate le osservazioni ai *cluster*, i centroidi vengono ricalcolati sulla base della nuova composizione dei gruppi. Formalmente, il nuovo centroide viene determinato individuando il punto che minimizza la distanza complessiva tra le osservazioni appartenenti al *cluster* e il rispettivo rappresentante:

$$c_j = arg \min_{c} \sum_{x \in S_j} d(x, c) \tag{2.6}$$

   Nel caso della distanza euclidea quadratica, il nuovo centroide coincide con il vettore delle medie delle osservazioni appartenenti al *cluster*.

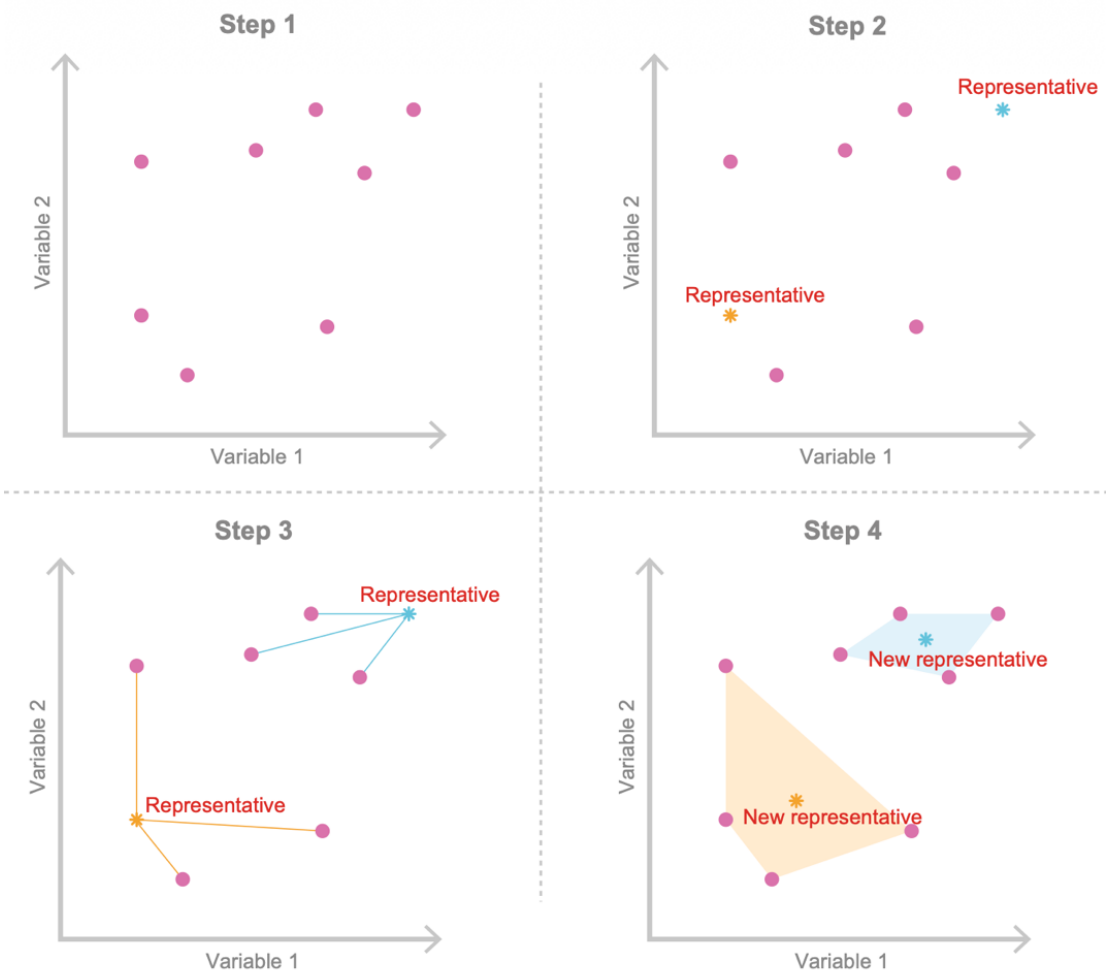
   In termini più semplici, questa operazione consente di individuare un nuovo rappresentante che descriva nel modo più accurato possibile le caratteristiche medie del gruppo.

5. Ripetizione del procedimento

   I passaggi di assegnazione delle osservazioni e aggiornamento dei centroidi vengono ripetuti iterativamente fino al raggiungimento della convergenza. L'algoritmo termina quando le assegnazioni delle osservazioni ai *cluster* non subiscono ulteriori variazioni significative oppure quando viene raggiunto un numero massimo di iterazioni prestabilito. La partizione ottenuta viene quindi considerata la soluzione interpretativa finale del procedimento.

Una rappresentazione schematica del funzionamento del metodo *K-means* è riportata nella Figura 2.3. Come evidenziato da Dolnicar et al. (2018), la figura sintetizza visivamente i primi quattro passaggi dell'algoritmo, mentre il quinto consiste nella ripetizione iterativa delle fasi di assegnazione delle osservazioni e aggiornamento dei centroidi fino al raggiungimento della convergenza.

Figura 2.3 Rappresentazione schematica del metodo K-means



Fonte: Dolnicar et al., 2018, p. 91

2.5.2 Cluster Plot

La rappresentazione grafica dei risultati ottenuti attraverso il metodo *K-means* costituisce uno strumento utile per facilitare l'interpretazione della segmentazione e per valutare visivamente la struttura dei gruppi individuati. In presenza di dataset caratterizzati da un elevato numero di variabili, la visualizzazione diretta delle osservazioni nello spazio multidimensionale originario risulta infatti difficoltosa; per tale ragione si ricorre spesso a tecniche di riduzione della dimensionalità e a rappresentazioni sintetiche dei *cluster* (Dolnicar et al., 2018).

Uno degli strumenti più utilizzati è il *cluster plot* (Figura 3.15; Figura 3.16), o grafico a dispersione dei *cluster*, nel quale le osservazioni vengono proiettate su un piano bidimensionale, i cui assi x e y sono definiti dalle prime due componenti principali ottenute mediante l'Analisi delle Componenti Principali (*Principal Component Analysis* – PCA).

La PCA è una tecnica statistica finalizzata alla sintesi dell'informazione contenuta in un insieme di variabili attraverso la costruzione di nuove variabili, denominate componenti principali, ottenute come combinazioni lineari delle variabili originarie. Le componenti vengono ordinate in funzione della quota di variabilità complessiva dei dati che riescono a spiegare, consentendo di rappresentare le osservazioni in uno spazio di dimensioni ridotte ma preservando una parte consistente dell'informazione presente nel dataset iniziale (Johnson e Wichern, 2007).

Nel *cluster plot*, ciascun punto rappresenta un'unità statistica, mentre il colore identifica il *cluster* di appartenenza. La distanza tra i punti riflette il grado di somiglianza tra le osservazioni: gruppi ben separati indicano una segmentazione più distinta, mentre eventuali sovrapposizioni tra *cluster* possono suggerire la presenza di osservazioni caratterizzate da profili intermedi. La presenza di punti particolarmente distanti gli uni dagli altri può inoltre evidenziare possibili osservazioni atipiche note anche come *outlier* (Dolnicar et al., 2018).

Un ulteriore strumento di supporto all'interpretazione dei risultati ottenuti è rappresentato dai grafici a barre dei profili medi dei *cluster* (Figura 3.17). Tali rappresentazioni consentono di confrontare i valori medi assunti dalle variabili considerate all'interno dei diversi gruppi (*cluster*) individuati. Attraverso il confronto tra le altezze delle barre è possibile identificare le caratteristiche che contraddistinguono ciascun *cluster* ed evidenziare eventuali differenze nell'intensità dei fenomeni osservati, nella loro distribuzione temporale o nella composizione delle categorie analizzate. I grafici a barre risultano pertanto particolarmente utili nella fase di descrizione e denominazione dei segmenti ottenuti, permettendo di associare ai *cluster* individuati specifici profili interpretabili (Dolnicar et al., 2018).

Infine, può essere utilizzato uno *Scree plot* relativo alla PCA (Figura 3.9). Questo grafico riporta, per ciascuna componente principale, la quota di varianza totale spiegata e consente di valutare quanta informazione del dataset originario venga preservata nella rappresentazione bidimensionale impiegata per visualizzare i *cluster* (*cluster plot*). L'osservazione dell'andamento decrescente della varianza spiegata permette inoltre di individuare eventuali punti di flesso della curva, oltre i quali ciascuna componente aggiuntiva contribuisce spiegando solo una quota marginale della variabilità complessiva dei dati.

Lo scree plot consente quindi di valutare se poche componenti siano sufficienti a rappresentare in modo soddisfacente la struttura del dataset originario (Jolliffe, 2002). Inoltre, nonostante non costituisca una rappresentazione specifica del metodo *K-means*, lo *Scree plot* fornisce indicazioni utili per quanto riguarda l'adeguatezza della proiezione ottenuta mediante la PCA e

anche per quanto riguarda la capacità delle prime componenti principali di rappresentare efficacemente la struttura complessiva dei dati.

Nel loro complesso, i suddetti strumenti consentono di integrare l'analisi numerica dei risultati con una rappresentazione visiva della segmentazione prodotta, facilitando e rendendo più immediata l'interpretazione dei gruppi individuati e delle relazioni esistenti tra le osservazioni (Dolnicar et al., 2018).

2.5.3 Vantaggi, svantaggi e limiti del *clustering* non gerarchico

Dal punto di vista computazionale, il metodo *K-means* risulta particolarmente efficiente nell'analisi di dataset caratterizzati da un elevato numero di osservazioni. Infatti, mentre i metodi gerarchici richiedono il calcolo delle distanze tra tutte le coppie di unità statistiche presenti nel dataset, il *K-means* considera principalmente le distanze tra le osservazioni e i centroidi del *cluster*. Ciò rende i metodi non gerarchici particolarmente adatti all'analisi di *dataset* di grandi dimensioni (Dolnicar et al., 2018).
È opportuno rimarcare nuovamente che il metodo *K-means* converge sempre verso una soluzione finale, ma tale soluzione non coincide necessariamente con l'ottimo globale. Dal momento che i centroidi iniziali vengono generalmente selezionati in modo casuale, differenti punti di partenza possono condurre soluzioni di *clustering* differenti. Per tale ragione, nelle applicazioni empiriche è spesso consigliabile eseguire più volte l'algoritmo e confrontare i risultati ottenuti, al fine di individuare la soluzione maggiormente stabile e interpretabile.

2.6 Confronto tra i principali metodi di *clustering*

Le tecniche di *clustering* gerarchico e non gerarchico perseguono il medesimo obiettivo, ovvero l'individuazione di gruppi omogenei di osservazioni sulla base delle caratteristiche considerate. Tuttavia, esse differiscono per modalità operative, complessità computazionale e caratteristiche dei risultati ottenuti.

I metodi gerarchici, come il metodo del legame completo (*Complete Linkage*) e il metodo di *Ward*, costruiscono progressivamente una struttura gerarchica di *cluster* rappresentabile attraverso un dendrogramma. Tale approccio consente di analizzare le relazioni tra le osservazioni a differenti livelli di aggregazione e di individuare il numero di *cluster* anche in una fase successiva all'esecuzione dell'algoritmo. I metodi non gerarchici, al contrario, non producono una struttura gerarchica bensì mirano direttamente alla suddivisione delle osservazioni in un numero prefissato di gruppi attraverso procedure iterative di partizionamento

(Dolnicar et al., 2018) e i *cluster* vengono rappresentati da un grafico a dispersione (*Cluster Plot*).

La Tabella 2.1 riassume le principali caratteristiche dei metodi di *clustering* impiegati nell'analisi empirica sviluppata nel presente elaborato.

Tabella 2.1 Confronto tra clustering gerarchico e clustering non gerarchico

Caratteristica	***Clustering* gerarchico** (*Complete Linkage* e *Ward*)	***Clustering* non gerarchico** (*K-means*)
Numero di *cluster*	determinato successivamente attraverso il dendrogramma	definito a priori
Struttura dei risultati	gerarchica (dendrogramma)	partizione diretta dei dati (*Cluster Plot*)
Approccio	aggregazione o suddivisione progressiva dei *cluster*	assegnazione iterativa delle osservazioni ai *cluster*
Complessità computazionale	relativamente elevata	generalmente più contenuta
Dataset di grandi dimensioni	meno adatto	maggiormente adatto
Interpretabilità	elevata	dipende dalla qualità della partizione ottenuta
Rappresentazione grafica	dendrogramma	grafico a dispersione (o *cluster* plot)
Dipendenza dalle condizioni iniziali	assente	presente
Flessibilità esplorativa	elevata	limitata

Fonte: Elaborazione personale dell'autore

Nel presente elaborato i due approcci verranno utilizzati in maniera complementare. In particolare, il *clustering* gerarchico consentirà di esplorare la struttura dei dati e di supportare l'individuazione del numero appropriato di *cluster*, mentre il metodo *K-means* verrà impiegato per ottenere la classificazione finale delle unità territoriali oggetto di analisi.

2.7 Considerazioni critiche, limiti e implicazioni metodologiche delle tecniche di *clustering*

Le tecniche di *clustering* rappresentano uno degli strumenti più utilizzati nell'ambito della segmentazione *data-driven* grazie alla loro capacità di individuare gruppi omogenei direttamente a partire dai dati osservati. Tuttavia, come evidenziato dalla letteratura scientifica, tali tecniche non conducono a una classificazione definitiva, univoca e oggettiva della realtà, ma producono segmentazioni la cui qualità dipende in larga parte dalle decisioni assunte durante il processo di analisi (Dolnicar, 2002). Per questo motivo, l'applicazione del *clustering* richiede particolare attenzione nella definizione delle scelte metodologiche e nell'interpretazione dei risultati ottenuti.

Tra i contributi metodologici più significativi in questo ambito si colloca la review sugli studi di segmentazione turistica *data-driven* proposta da Dolnicar (2002), in cui l'autrice evidenzia come molte delle differenze riscontrate tra le segmentazioni non dipendano esclusivamente dalle caratteristiche del fenomeno analizzato, bensì anche dalle scelte effettuate dal ricercatore. Tra gli aspetti più critici vengono indicati la selezione delle variabili, la scelta dell'algoritmo di *clustering*, la determinazione del numero di *cluster* e le procedure di validazione adottate. Dal momento che non esiste una procedura universalmente riconosciuta come ottimale, la qualità di una segmentazione dipende dalla trasparenza metodologica con cui tali decisioni vengono motivate e documentate.

Le criticità evidenziate dalla letteratura hanno costituito un riferimento nella progettazione dell'analisi sviluppata nel presente elaborato. Alla luce di tali considerazioni, si è infatti cercato di adottare un approccio metodologico il più possibile trasparente e replicabile.

In tale prospettiva, le principali decisioni assunte durante l'analisi, dalla selezione delle variabili alla scelta dell'algoritmo di *clustering* e alla determinazione del numero finale di *cluster*, sono state motivate sulla base delle indicazioni presenti in letteratura e successivamente discusse nel capitolo dedicato all'analisi dei dati, così da consentire al lettore di ricostruire l'intero processo analitico e valutarne criticamente le implicazioni. In particolare, la segmentazione non è stata costruita facendo riferimento ad un'unica soluzione, ma mediante il confronto tra differenti metodologie e configurazioni, così da valutare la stabilità e la coerenza interpretativa dei risultati.

Un ulteriore elemento evidenziato dalla letteratura riguarda la necessità che i *cluster* ottenuti risultino non soltanto statisticamente coerenti, ma anche sostanzialmente interpretabili. Una segmentazione difficilmente spiegabile o scarsamente riproducibile rischia, infatti, di ridurre la propria utilità sia in ambito scientifico sia nelle possibili applicazioni pratiche.

Per tale ragione, la valutazione dei risultati dovrebbe sempre integrare indicatori quantitativi con una lettura critica delle caratteristiche dei gruppi individuati, evitando interpretazioni automatiche o esclusivamente basate sugli output prodotti dagli algoritmi di *clustering* (Dolnicar, 2002).

Dall'analisi della review proposta da Dolnicar (2002) emerge come non sia possibile individuare un algoritmo di *clustering* universalmente migliore degli altri, in quanto l'adeguatezza di una determinata tecnica dipende dagli obiettivi della ricerca, dalle caratteristiche dei dati disponibili e dalle scelte metodologiche adottate dal ricercatore.

Le considerazioni esposte evidenziano in maniera piuttosto marcata come il *clustering* debba essere interpretato non come uno strumento in grado di individuare una classificazione definitiva della realtà osservata, bensì come una metodologia esplorativa finalizzata ad individuare strutture latenti nei dati. La consapevolezza dei limiti e delle criticità metodologiche costituisce pertanto un elemento essenziale per un'applicazione rigorosa delle tecniche di segmentazione e rappresenta il presupposto per la corretta interpretazione dei risultati che verranno presentati nel capitolo successivo.

Conclusione

La trattazione sviluppata nel presente capitolo ha illustrato i principali fondamenti teorici e metodologici delle tecniche di *clustering* impiegate nel presente elaborato. Dopo aver introdotto il concetto di segmentazione *data-driven* e il ruolo delle misure di distanza nell'analisi della similarità tra le osservazioni, sono stati approfonditi i principali metodi di *clustering* gerarchico e non gerarchico, evidenziandone il funzionamento, le caratteristiche distintive, i principali vantaggi e i relativi limiti applicativi.

L'analisi metodologica ha inoltre evidenziato come la scelta dell'algoritmo di *clustering* e delle relative procedure operative rappresenti una fase determinante del processo di segmentazione, in quanto i risultati ottenuti dipendono non solo dalle caratteristiche dei dati disponibili, ma anche dalle decisioni metodologiche adottate durante l'intero percorso di analisi. In tale prospettiva, le tecniche di *clustering* devono essere interpretate come strumenti di analisi esplorativa, la cui efficacia è strettamente connessa al rigore metodologico con cui vengono applicate e alla corretta interpretazione dei risultati.

Sulla base di tali presupposti teorici e metodologici, il capitolo successivo presenta l'applicazione empirica delle tecniche illustrate ai dati relativi ai flussi turistici dei comuni della Valle d'Aosta, descrivendo il dataset utilizzato, le procedure di preparazione dei dati e il processo di segmentazione sviluppato attraverso i diversi algoritmi di *clustering*.

3. ANALISI DEI DATI

Introduzione

Il presente capitolo costituisce la parte empirica del presente elaborato ed è dedicato all'applicazione delle tecniche di *clustering* illustrate nel capitolo precedente ai dati relativi ai flussi turistici dei comuni della Valle d'Aosta.

L'analisi di un sistema turistico territoriale come quello valdostano presenta infatti un elevato grado di complessità, in quanto ciascun comune è descritto simultaneamente da numerose variabili riguardanti gli arrivi, le presenze, la loro distribuzione temporale e le differenti tipologie di strutture ricettive. In un contesto caratterizzato da una tale multidimensionalità, l'osservazione separata delle singole variabili risulta spesso insufficiente a cogliere le relazioni complessive esistenti tra i territori. L'impiego di tecniche di *clustering* risulta pertanto particolarmente utile, poiché consente di sintetizzare tale complessità individuando gruppi di comuni caratterizzati da profili turistici omogenei emergenti direttamente dai dati osservati.

Per tale ragione, il capitolo presenta inizialmente il *dataset* utilizzato, le variabili considerate e le operazioni di preparazione e pulizia dei dati necessarie per l'analisi. Successivamente vengono illustrate alcune analisi descrittive preliminari, finalizzate a fornire un primo inquadramento delle caratteristiche della domanda turistica regionale, per poi procedere all'applicazione delle tecniche di *clustering*, al confronto tra i diversi algoritmi impiegati e all'interpretazione delle segmentazioni ottenute.

Il capitolo si conclude con una discussione delle principali implicazioni emerse dal caso di studio, evidenziando il contributo delle tecniche di *clustering* alla comprensione della struttura del sistema turistico valdostano e riflettendo sui vantaggi e limiti dell'approccio adottato.

3.1 Presentazione dei dati e variabili analizzate

Per l'applicazione delle tecniche di *clustering* è stato utilizzato un *dataset* costruito a partire dalle statistiche regionali sui flussi turistici comunali forniti dall'ufficio regionale competente. L'analisi è stata condotta considerando il periodo compreso tra gennaio e settembre 2025.

La scelta di limitare l'analisi ai soli primi nove mesi del 2025 è frutto di due motivazioni principali. In primo luogo, al momento dello svolgimento della ricerca, i dati disponibili e completi relativi ai flussi turistici comunali risultavano riferiti unicamente al periodo oggetto d'analisi. In secondo luogo, l'entrata in vigore della Legge Regionale n. 11 del 2023 ha introdotto l'obbligo di registrazione dei movimenti turistici, anche per le strutture ricettive

extralberghiere quali affittacamere e appartamenti ad uso privato. L'entrata in vigore di tale Legge Regionale ha comportato una modifica normativa che ha determinato una discontinuità nella rilevazione statistica dei flussi turistici, rendendo non direttamente confrontabili i dati antecedenti con quelli successivi all'applicazione dei nuovi provvedimenti introdotti.

L'unità statistica considerata nell'analisi è costituita dal singolo Comune della Valle d'Aosta. Sono stati analizzati i 74 comuni della Regione, descritti attraverso un insieme di variabili relative agli arrivi e alle presenze turistiche, distinti per tipologia di struttura ricettiva e per mese di osservazione.

Le tipologie di strutture ricettive considerate sono le seguenti: Affittacamere/Chambres d'hôtes; Agriturismi; Alberghi; Alloggi ad uso turistico; Aree di sosta; Bed and Breakfast; Campeggi; Campeggi sociali; Case e appartamenti per vacanze; Case per ferie; Ostelli per la gioventù; Posti tappa/Dortoir; Residenze Turistico-Alberghiere; Rifugi alpini; Villaggi turistici.

Per ciascuna tipologia sono stati rilevati il numero di arrivi e il numero di presenze registrati mensilmente nel periodo compreso tra gennaio e settembre 2025.

Complessivamente, il *dataset* utilizzato per la *clustering analysis* è costituito da 74 unità statistiche e da 258 variabili quantitative, rappresentative sia dell'intensità dei flussi turistici sia della distribuzione della domanda tra le diverse tipologie di strutture ricettive nel corso del periodo oggetto di analisi.

3.2 Preparazione e pulizia dei dati

Prima di procedere all'analisi descrittiva e all'applicazione delle tecniche di *clustering*, il *dataset* originario è stato sottoposto a una fase preliminare di preparazione e pulizia dei dati. Tale passaggio è stato necessario affinché i dati fossero coerenti con la struttura richiesta dalle successive elaborazioni statistiche effettuate attraverso il software R.

Il *dataset* di partenza conteneva informazioni sui flussi turistici dei comuni della Valle d'Aosta disaggregate per mese, tipologie di struttura ricettiva e indicatore osservato, distinguendo tra arrivi e presenze. Ai fini dell'analisi, i dati sono stati riorganizzati in modo tale che ciascuna riga corrispondesse a un comune della Valle d'Aosta e ciascuna colonna rappresentasse una specifica combinazione tra indicatore turistico (arrivo o presenza), tipologia di struttura ricettiva e mese di osservazione.

Tale riorganizzazione ha consentito di ottenere una matrice dei dati adatta all'applicazione delle tecniche di *clustering*, nelle quali le unità statistiche devono essere confrontate sulla base di un insieme comune di variabili quantitative.

Successivamente, il *dataset* è stato importato in R e sottoposto ad alcuni controlli preliminari con l'obiettivo di verificarne la corretta struttura e l'idoneità alle successive elaborazioni statistiche. In particolare, è stata verificata la dimensione della matrice dei dati, controllando il numero di osservazioni e di variabili disponibili ed è stata effettuata una ricerca di eventuali valori mancanti (*missing values*), ossia informazioni non disponibili per una o più combinazioni tra Comune, tipologia di struttura ricettiva e mese di osservazione. Nel caso oggetto d'analisi, dal controllo effettuato non è emersa la presenza di valori mancanti nel dataset utilizzato.

Il nome dei comuni, inizialmente presente come prima colonna del *dataset*, è stato separato dalle variabili numeriche e utilizzato come identificativo delle osservazioni (Figura 3.1), in modo da consentire una più agevole interpretazione dei risultati ottenuti. Tutte le variabili quantitative relative agli arrivi e alle presenze turistiche, distinte per tipologia di struttura ricettiva e per mesi di osservazione, sono invece state mantenute nella matrice dei dati impiegata per l'analisi, poiché considerate potenzialmente informative ai fini dell'individuazione di gruppi omogenei di comuni.

In seguito, sono state eliminate eventuali variabili costanti, vale a dire variabili che presentavano lo stesso valore per tutti i comuni osservati. Tali variabili non mostrano alcuna variabilità tra le unità statistiche e pertanto non contribuiscono alla distinzione tra i comuni e possono dunque essere escluse dall'analisi di *clustering*.

Figura 3.1 Importazione del dataset e controlli preliminari in R

```
# 1. Importo il dataset

df <- read_excel("dataset_R_mensile_pronto.xlsx",
                 sheet = "Dataset_R")

View(df)


# 2. Separo il nome dei comuni dalle variabili numeriche

comuni <- df$Comune

df_num <- df[, -1]
df_num <- as.data.frame(df_num)

rownames(df_num) <- comuni

View(df_num)


# 3. Effettuo controlli preliminari

dim(df_num)
sum(is.na(df_num))


# 4. Elimino eventuali variabili costanti

sd_var <- apply(df_num, 2, sd)

df_num2 <- df_num[, sd_var > 0]

dim(df_num2)

View(df_num2)
```

Fonte: Elaborazione personale dell'autore

Una volta completata la fase di pulizia dei dati, le variabili sono state standardizzate. La standardizzazione è stata necessaria poiché le variabili considerate presentano scale e livelli di intensità differenti: ad esempio, alcune tipologie di strutture ricettive registrano valori di arrivi e presenze molto più elevati rispetto ad altre. Se non si fosse proceduto a una standardizzazione delle variabili, quelle caratterizzate da valori numericamente maggiori avrebbero potuto influenzare in modo predominante il calcolo delle distanze tra i comuni. La standardizzazione ha quindi reso possibile il confronto tra le variabili e l'attribuzione a ciascuna di esse di un peso più equilibrato nell'analisi.

Figura 3.2 Standardizzazione delle variabili in R

```
# 5. Standardizzo le variabili

X <- scale(df_num2)
```

Fonte: Elaborazione personale dell'autore

Infine, a partire dal dataset preparato sono state costruite alcune tabelle di sintesi utili per l'interpretazione dei risultati, tra cui i profili medi dei *cluster* e le medie sintetiche relative agli arrivi e alle presenze mensili. Tali tabelle non sostituiscono il dataset utilizzato per il *clustering*,

ma rappresentano uno strumento di supporto per commentare più chiaramente le caratteristiche dei *cluster* individuati.

3.2.1 Utilizzo di strumenti di IA nelle fasi preliminari dell'analisi

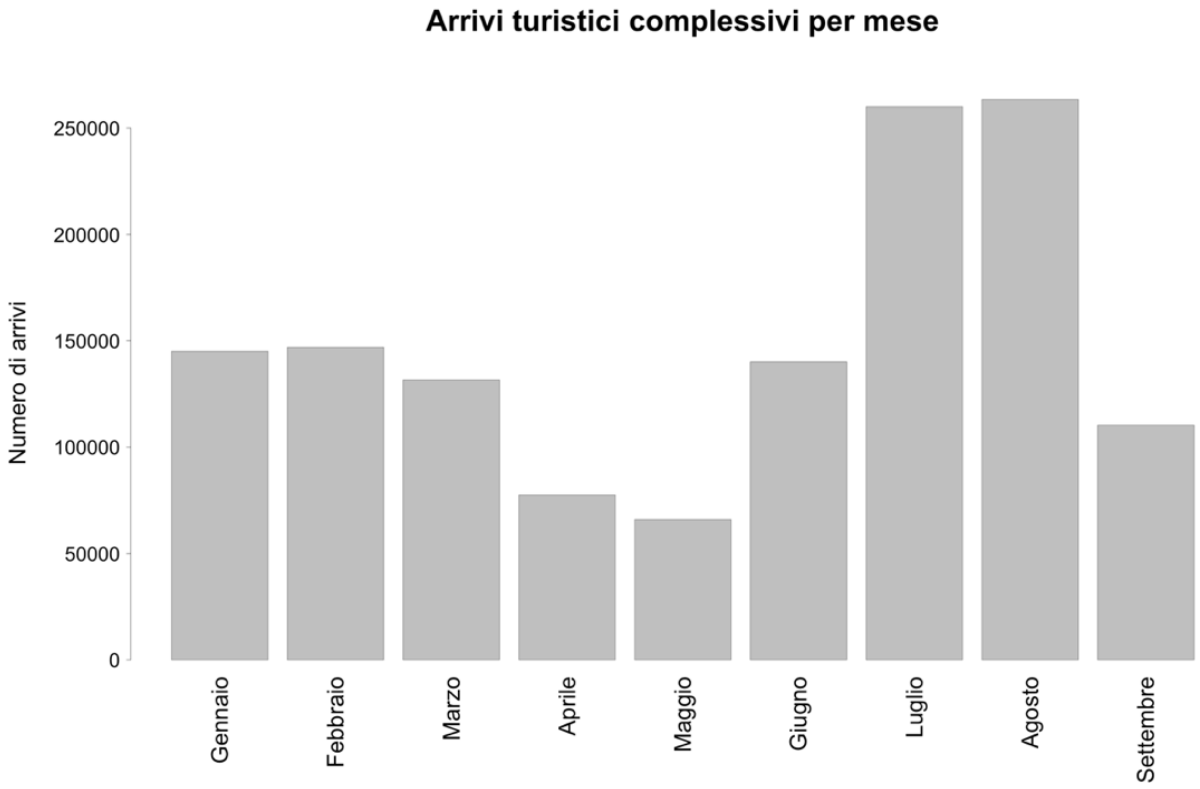
Nelle fasi preliminari dell'analisi è stato utilizzato uno strumento di intelligenza artificiale generativa (ChatGPT) come supporto operativo per la riorganizzazione e la preparazione del dataset da importare nel software R. In particolare, l'assistente è stato impiegato per automatizzare alcune operazioni di trasformazione dei dati, predisporre una struttura informativa coerente con le esigenze delle tecniche di *clustering* e verificare la correttezza del codice implementato. La scelta della tipologia di dataset da costruire per la successiva importazione in R, le scelte metodologiche, l'interpretazione dei risultati e le decisioni relative alla selezione delle procedure statistiche sono state invece assunte in completa autonomia dall'autore della ricerca.

3.3 Analisi descrittive dei dati

Per poter fornire un primo inquadramento delle caratteristiche del *dataset* oggetto di analisi, sono state effettuate alcune analisi descrittive preliminari riguardanti la distribuzione temporale dei flussi turistici e la composizione dell'offerta ricettiva regionale.

La Figura 3.3 riporta il numero complessivo di arrivi registrati nei comuni della Valle d'Aosta nel periodo oggetto di analisi compreso tra gennaio e settembre 2025. Dall'analisi del grafico emerge una marcata stagionalità dei flussi turistici, caratterizzata da valori relativamente elevati nei mesi invernali di gennaio, febbraio e marzo, seguita da una successiva contrazione nel periodo primaverile, in particolare con un minimo registrato nel mese di maggio, e infine una successiva crescita nel corso della stagione estiva, che raggiunge i valori più alti nei mesi di luglio e agosto.

Figura 3.3 Grafico a Barre (bar plot) degli arrivi complessivi registrati nei comuni della Valle d'Aosta tra gennaio e settembre 2025

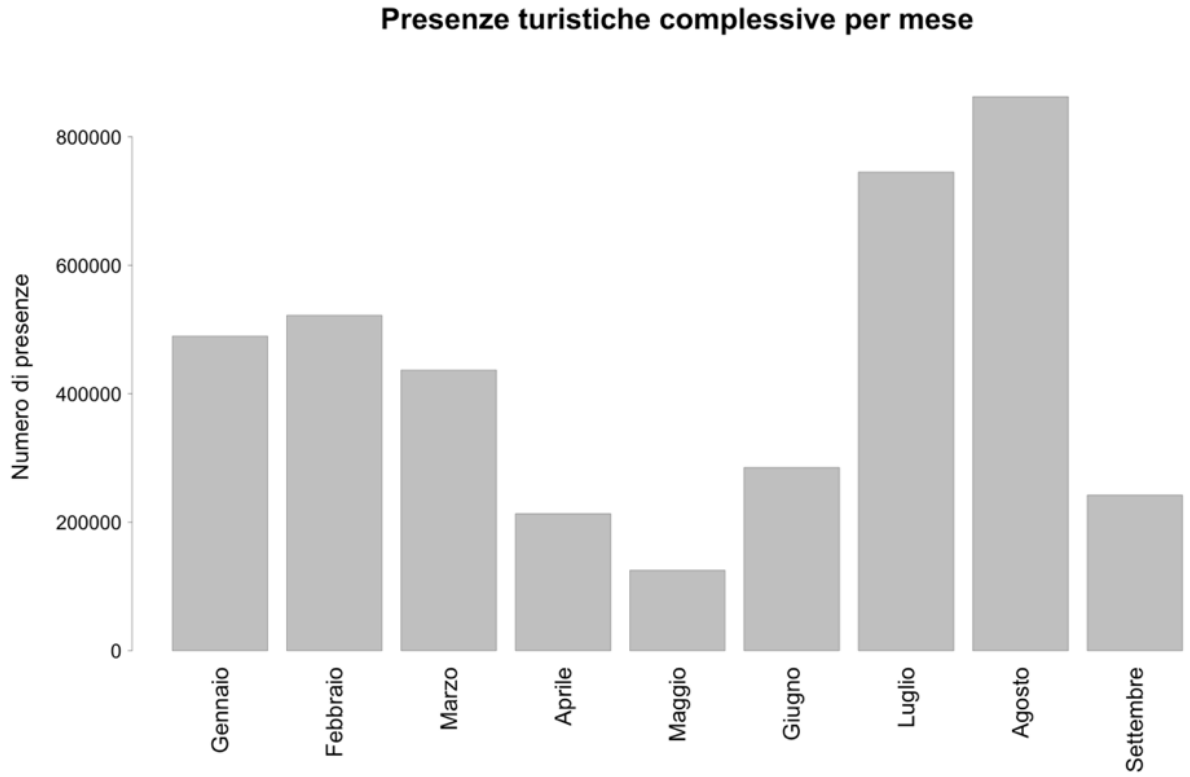


Fonte: Elaborazione personale dell'autore

Considerando il numero complessivo di presenze turistiche (Figura 3.4) si osserva un andamento analogo a quello osservato nel grafico precedente. Anche in questo caso, i valori massimi si registrano nel periodo estivo, mentre i mesi primaverili presentano livelli significativamente inferiori.

La distribuzione delle presenze suggerisce pertanto l'esistenza di una forte componente stagionale che influenza la domanda turistica valdostana, probabilmente riconducibile sia alla stagione sciistica e degli sport invernali, sia alla stagione estiva legata alle attività escursionistiche e naturalistiche.

Figura 3.4 Grafico a Barre (bar plot) delle presenze complessive registrate nei comuni della Valle d'Aosta tra gennaio e settembre 2025

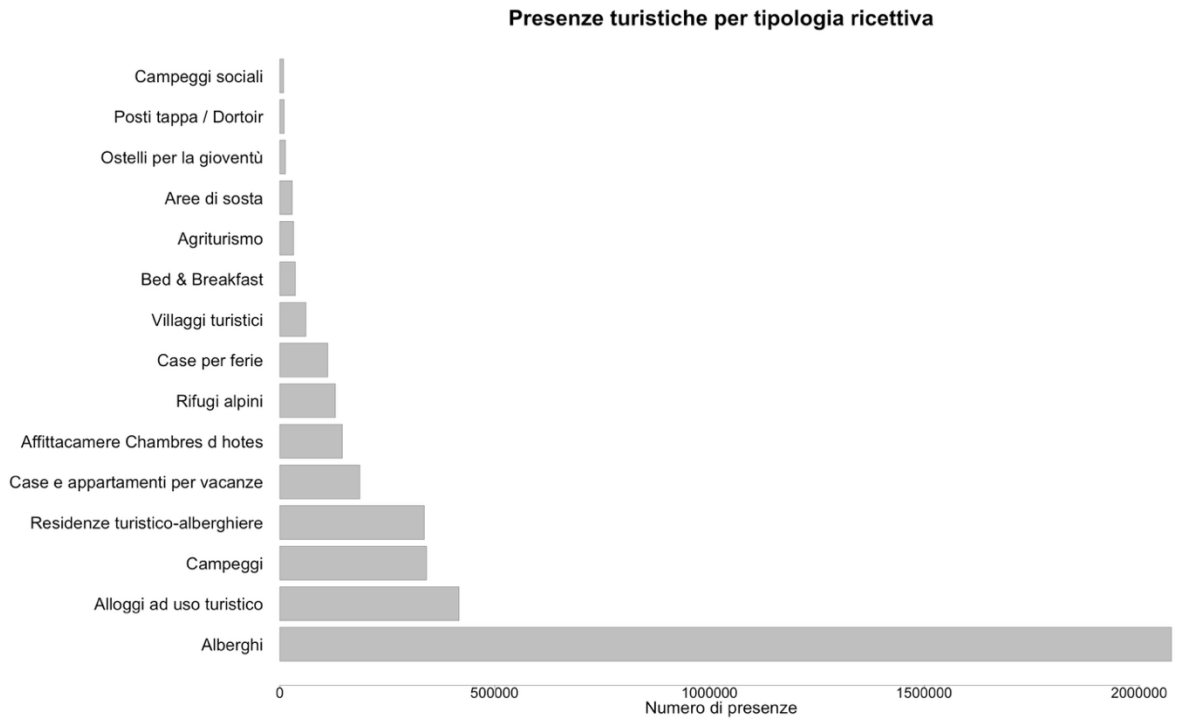


Fonte: Elaborazione personale dell'autore

Per quanto riguarda la struttura dell'offerta ricettiva, la Figura 3.5 mostra la distribuzione delle presenze complessive per tipologia di struttura ricettiva. Dall'analisi emerge una netta predominanza della componente alberghiera, che concentra la quota maggiore delle presenze registrate nel periodo osservato. Si osservano valori rilevanti anche per gli alloggi ad uso

turistico, i campeggi e le residenze turistico-alberghiere, mentre le restanti categorie di strutture ricettive presentano livelli di domanda decisamente più contenuti.

Figura 3.5 Grafico a barre (bar plot) delle presenze complessive per tipologia di struttura ricettiva registrate nei comuni della Valle d'Aosta tra gennaio e settembre 2025

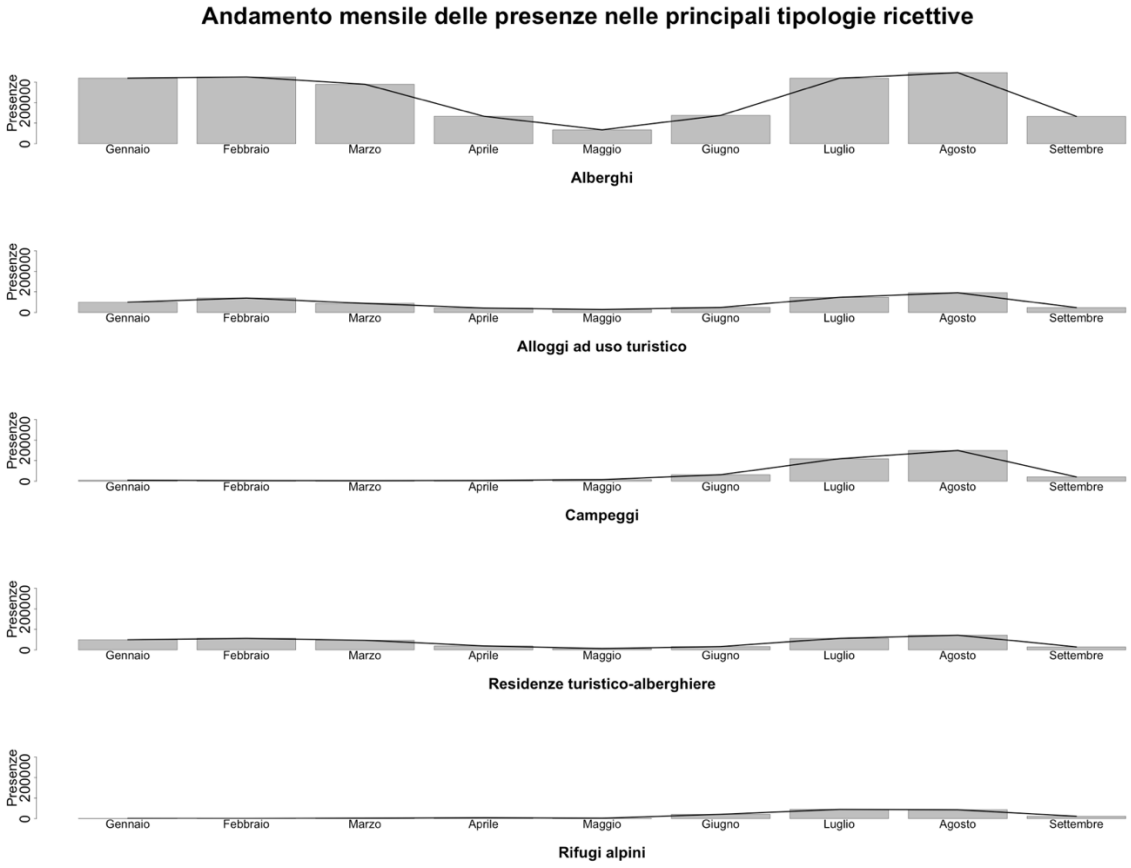


Fonte: Elaborazione personale dell'autore

L'analisi per tipologia è stata condotta esclusivamente sulle presenze e non sugli arrivi, in quanto esse rappresentano una misura più completa dell'intensità della domanda turistica, poiché tengono conto non soltanto del numero di turisti ospitati, ma anche del numero complessivo di pernottamenti effettuati. Gli arrivi descrivono invece esclusivamente il numero di ingressi registrati nelle strutture ricettive e forniscono pertanto informazioni più limitate riguardo all'effettivo utilizzo delle diverse tipologie di strutture.

Per approfondire la componente stagionale della domanda turistica, la Figura 3.6 riporta l'evoluzione mensile delle presenze registrate nelle principali tipologie ricettive.

Figura 3.6 Andamento mensile delle presenze nelle principali tipologie ricettive della Valle d'Aosta tra gennaio e settembre 2025



Fonte: Elaborazione personale dell'autore

Dall'analisi della suddetta evoluzione mensile, emerge come la componente alberghiera mantenga livelli di domanda superiori rispetto alle altre tipologie durante tutto il periodo osservato, e mostri una marcata doppia stagionalità, con livelli elevati sia nei mesi invernali che in quelli estivi, confermando il ruolo centrale degli alberghi all'interno del sistema ricettivo regionale. Gli alloggi ad uso turistico e le residenze turistico-alberghiere presentano un andamento analogo, seppur con volumi inferiori. I campeggi e i rifugi alpini evidenziano invece una forte concentrazione delle presenze nei mesi estivi, coerente con la fruizione prevalentemente legata alle attività all'aria aperta e all'escursionismo. Nel complesso, il grafico conferma come la domanda turistica regionale presenti una stagionalità differenziata anche in funzione della tipologia ricettiva considerata.

Per esigenze di sintesi, sono state rappresentate solo le categorie di struttura ricettiva determinate da consistenti livelli di presenze.

3.4 Segmentazione *data-driven* attraverso tecniche di *clustering*

3.4.1 Applicazione del *clustering*

Al fine di individuare gruppi omogenei di comuni caratterizzati da profili turistici simili, è stata condotta una *cluster analysis* utilizzando sia approcci gerarchici sia approcci non gerarchici. L'obiettivo della procedura è stato quello di segmentare i Comuni Valdostani sulla base delle principali variabili relative ai flussi turistici, evidenziando eventuali analogie e differenze nei modelli di domanda e di utilizzo dell'offerta ricettiva.

Come già anticipato nelle sezioni precedenti, poiché le variabili considerate presentavano ordini di grandezza differenti, prima di procedere con la clusterizzazione è stata effettuata una standardizzazione dei dati mediante la funzione "scale ( )" del software R (Figura 3.2). Tale operazione ha consentito di attribuire uguale peso a tutte le variabili analizzate evitando che quelle caratterizzate da valori assoluti più elevati influenzassero in misura predominante il calcolo delle distanze tra le osservazioni. Per la misurazione della dissimilarità tra i comuni è stata adottata la distanza euclidea, il cui uso è ampiamente diffuso nelle applicazioni di *clustering* su dati quantitativi standardizzati. Sono stati quindi applicati due algoritmi di *clustering* gerarchico, corrispondenti ai criteri di aggregazione *Complete Linkage* e *Ward*, e l'algoritmo partizionale *K-means*. Con particolare riferimento al metodo di *Ward* è stata impiegata la distanza euclidea quadratica, coerentemente con il criterio di aggregazione basato sulla minimizzazione dell'incremento della varianza interna ai gruppi.

Il metodo del *Complete Linkage* (legame completo) determina la distanza tra due gruppi come la massima distanza osservata tra le unità appartenenti ai rispettivi *cluster*, tendendo a produrre gruppi relativamente compatti ma potenzialmente sensibili alla presenza di osservazioni estreme. Il metodo di *Ward*, invece, aggrega progressivamente le osservazioni minimizzando l'incremento della varianza interna ai *cluster*, risultando generalmente più efficace nella formazione di gruppi omogenei e bilanciati.

Figura 3.7 Applicazione dei metodi di clustering gerarchico in R – Criterio del Complete Linkage e Ward

```r
# 6. Calcolo la matrice delle distanze euclidee

D <- dist(X, method = "euclidean")


# 7. Applico il metodo del legame completo

hc_complete <- hclust(D, method = "complete")


# 8. Applico il metodo di Ward

hc_ward <- hclust(D, method = "ward.D2")
```

Fonte: Elaborazione personale dell'autore

L'algoritmo *K-means* è stato utilizzato come tecnica complementare di segmentazione, al fine di confrontare le partizioni ottenute attraverso i metodi gerarchici con una procedura di ottimizzazione iterativa basata sulla minimizzazione della variabilità interna ai *cluster*.

Figura 3.8 Applicazione dell'algoritmo K-means in R

```r
# 16. Applico K-means con k = 5 e k = 6

set.seed(123)

km5 <- kmeans(X,
              centers = 5,
              nstart = 50)

km6 <- kmeans(X,
              centers = 6,
              nstart = 50)

km5$size
km6$size

table(km5$cluster)
table(km6$cluster)

split(rownames(X), km5$cluster)
split(rownames(X), km6$cluster)
```

Fonte: Elaborazione personale dell'autore

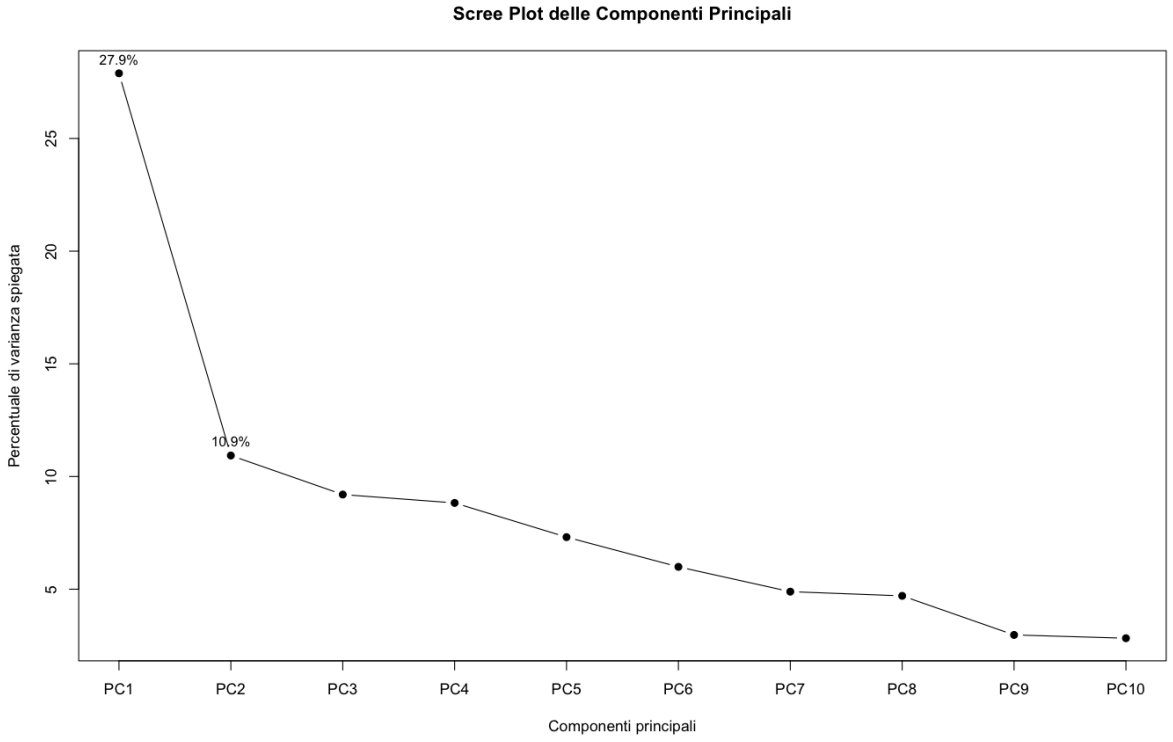
Per agevolare l'interpretazione grafica dei risultati, è stata inoltre effettuata un'Analisi delle Componenti Principali (PCA), impiegata esclusivamente ai fini esplorativi e di visualizzazione. La rappresentazione dei comuni nello spazio definito dalle prime due componenti principali consente infatti di osservare la distribuzione delle osservazioni e la separazione tra i gruppi individuati dagli algoritmi di *clustering*.

La Figura 3.9 riporta lo *Scree Plot* relativo alla PCA. Le prime due componenti principali spiegano complessivamente circa il 38,8% della variabilità totale del dataset, di cui il 27,9%

attribuibile alla prima componente e il 10,9% alla seconda, fornendo una rappresentazione bidimensionale sufficientemente informativa delle relazioni esistenti tra i comuni analizzati.

Tale informazione risulta particolarmente rilevante in quanto i *Cluster Plot* (Figura 3.15; Figura 3.16) analizzati nelle sezioni successive dell'analisi sono costruiti proprio utilizzando le coordinate delle osservazioni sulle prime due componenti principali. Lo *Scree Plot* consente quindi di valutare quanta parte della struttura informativa originaria venga preservata nella rappresentazione bidimensionale utilizzata per visualizzare i *cluster* derivanti dall'algoritmo *K-means*.

Figura 3.9 Scree Plot della PCA



Fonte: Elaborazione personale dell'autore

3.4.2 Identificazione dei segmenti

L'analisi delle configurazioni ottenute attraverso i diversi algoritmi di *clustering* è stata condotta con l'obiettivo di individuare una segmentazione dei comuni valdostani che risultasse al contempo statisticamente equilibrata e che garantisse facilità di interpretazione dei risultati. Per tale ragione sono state inizialmente considerate partizioni comprese tra tre e sei gruppi; tuttavia, per esigenze di sintesi e chiarezza nell'interpretazione dei risultati, nel presente capitolo verranno presentate esclusivamente le soluzioni con cinque e sei *cluster*, mentre le ulteriori configurazioni verranno riportate in appendice.

Nei dendrogrammi, l'altezza dei rami rappresenta il livello di dissimilarità al quale le osservazioni o i gruppi vengono aggregati. Rami che si uniscono a livelli più bassi indicano comuni più simili tra loro, mentre fusioni a livelli più elevati segnalano profili turistici più distanti. I rettangoli rossi evidenziano i tagli dell'albero gerarchico in corrispondenza del numero di *cluster* considerato, ottenuti mediante la funzione "cutree" del software statistico R (Figura 3.10).

Figura 3.10 Estrazione delle partizioni dai dendrogrammi ottenuti mediante Metodo di Ward e Complete Linkage in R

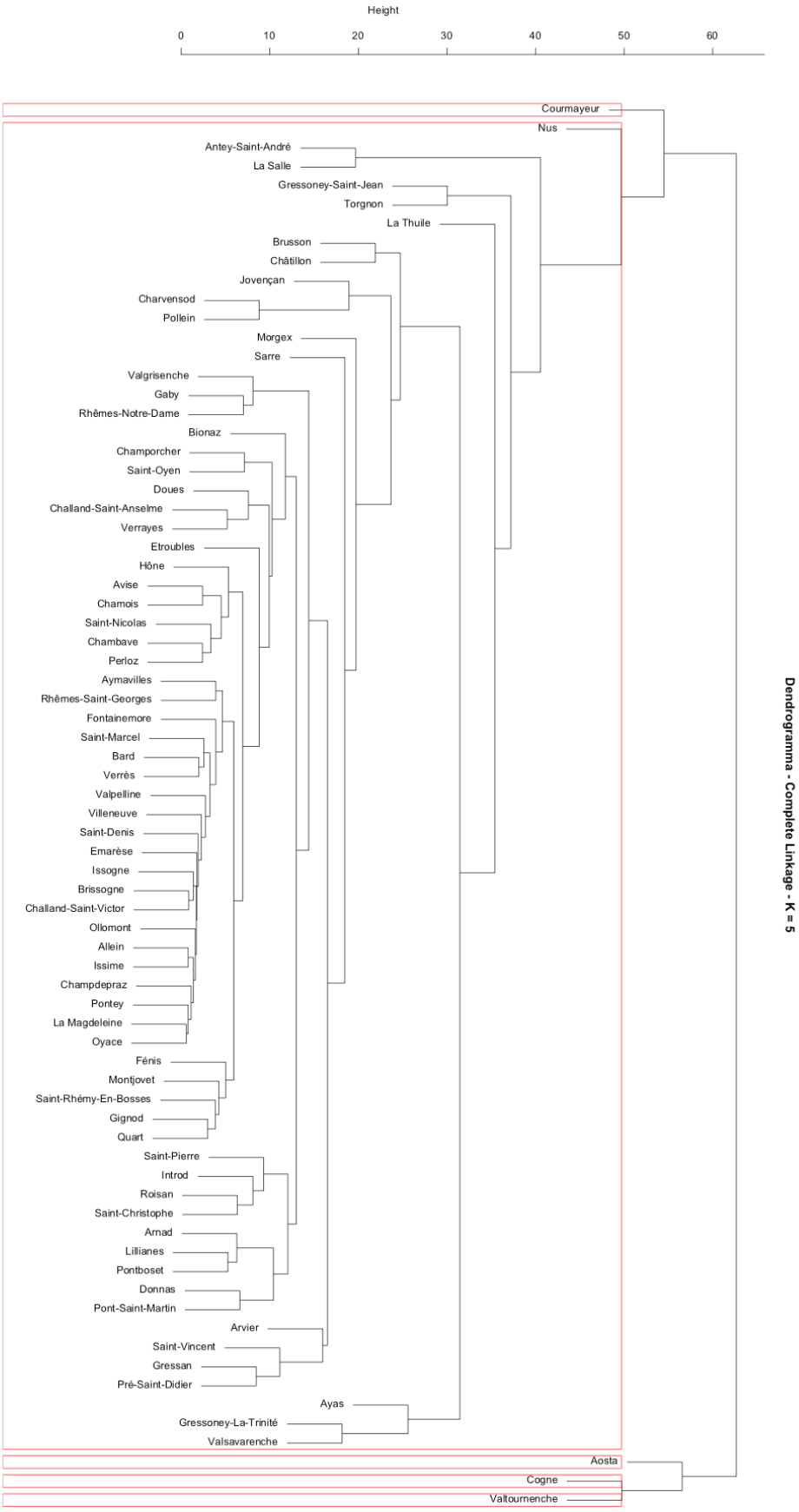
```
# 15. Estraggo le partizioni Ward

cluster3 <- cutree(hc_ward, k = 3)
cluster4 <- cutree(hc_ward, k = 4)
cluster5 <- cutree(hc_ward, k = 5)
cluster6 <- cutree(hc_ward, k = 6)

table(cluster3)
table(cluster4)
table(cluster5)
table(cluster6)

split(rownames(X), cluster3)
split(rownames(X), cluster4)
split(rownames(X), cluster5)
split(rownames(X), cluster6)


# 15-bis. Estraggo ed esamino anche le partizioni Complete Linkage

cluster3_complete <- cutree(hc_complete, k = 3)
cluster4_complete <- cutree(hc_complete, k = 4)
cluster5_complete <- cutree(hc_complete, k = 5)
cluster6_complete <- cutree(hc_complete, k = 6)

table(cluster3_complete)
table(cluster4_complete)
table(cluster5_complete)
table(cluster6_complete)

split(rownames(X), cluster3_complete)
split(rownames(X), cluster4_complete)
split(rownames(X), cluster5_complete)
split(rownames(X), cluster6_complete)
```

Fonte: Elaborazione personale dell'autore

Metodo del *Complete Linkage*

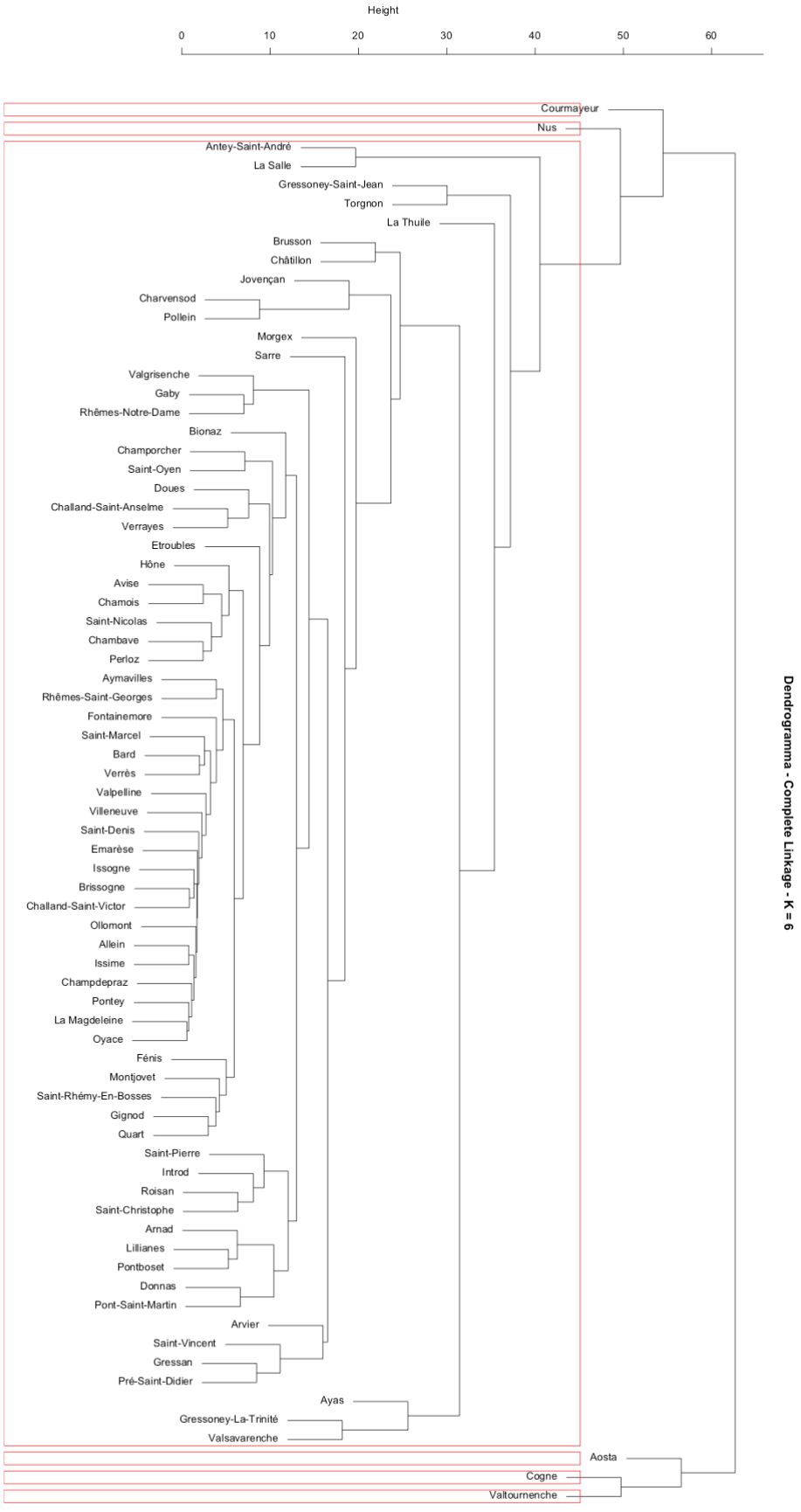
La Figura 3.11 e la Figura 3.12 rappresentano rispettivamente le segmentazioni ottenute mediante il metodo del *Complete Linkage* nel caso di una suddivisione in cinque ($K = 5$) e sei ($K = 6$) *cluster*. Dall'osservazione dei dendrogrammi emerge una certa tendenza del metodo a generare gruppi di dimensioni ridotte e ad isolare progressivamente alcune osservazioni caratterizzate da profili turistici particolarmente distintivi. Sebbene tale approccio consenta di individuare *cluster* relativamente compatti, la presenza di gruppi costituiti da un numero limitato di comuni può rendere più complessa l'interpretazione complessiva della segmentazione e ridurre la capacità di generalizzazione dei risultati, e quindi rendere meno agevole l'individuazione di profili turistici interpretabili dal punto di vista territoriale e riconducibili a categorie di destinazione più ampie.

Figura 3.11 Dendrogramma – Complete Linkage k=5 (clusters)



Fonte: Elaborazione personale dell'autore

Figura 3.12 Dendrogramma – Complete Linkage k=6 (clusters)

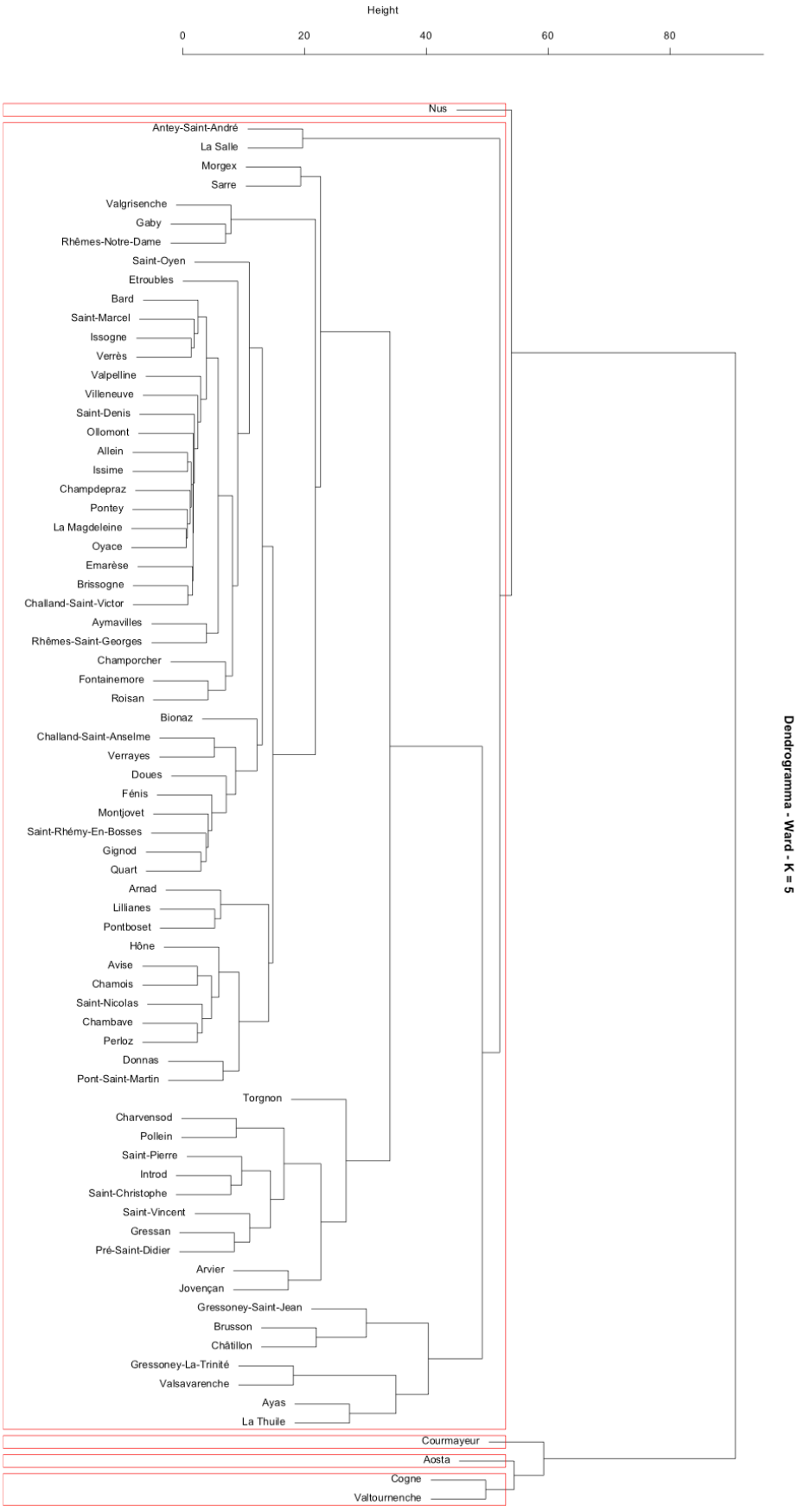


Fonte: Elaborazione personale dell'autore

Metodo di *Ward*

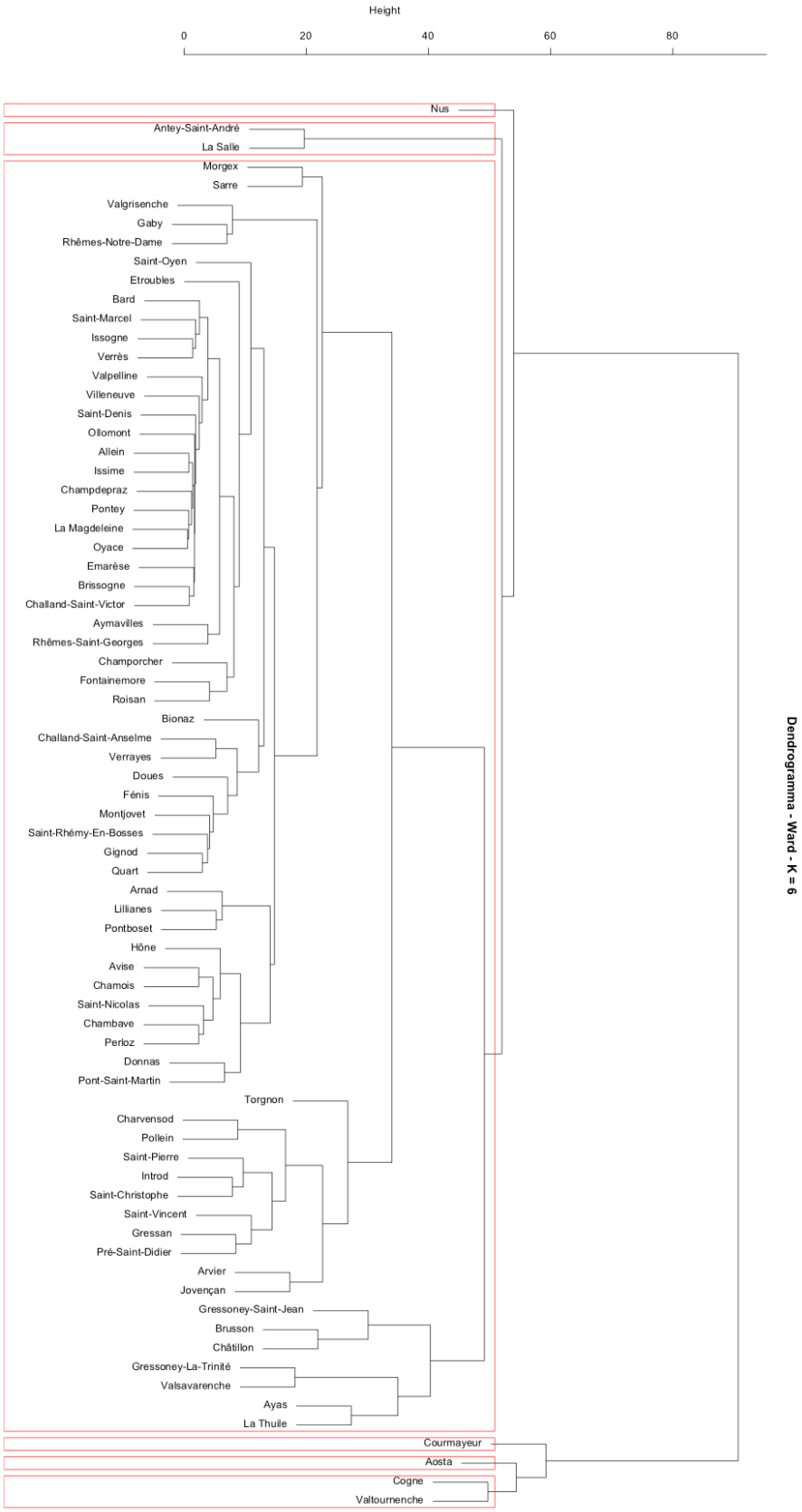
La Figura 3.13 e la Figura 3.14 rappresentano invece le segmentazioni ottenute mediante il metodo di *Ward*. Rispetto al *Complete Linkage*, il criterio di *Ward* sembra produrre una struttura gerarchica più equilibrata, con gruppi generalmente più omogenei e caratterizzati da una minore variabilità interna ai *cluster*. Tale comportamento risulta coerente con l'utilizzo della distanza euclidea quadratica prevista dal metodo di *Ward* e implementata in R attraverso il metodo "*Ward.D2*" (Figura 3.7), la quale attribuisce un peso relativamente maggiore alle osservazioni più distanti dal centroide del *cluster*, favorendo la formazione di gruppi caratterizzati da una ridotta variabilità o dispersione interna. In particolare, la soluzione con cinque *cluster* sembra in grado di distinguere segmenti sufficientemente differenti tra loro senza determinare un'eccessiva frammentazione delle osservazioni, mentre la partizione in sei gruppi genera la formazione di *cluster* più piccoli e maggiormente specializzati.

Figura 3.13 Dendrogramma - Metodo di Ward K=5 (clusters)



Fonte: Elaborazione personale dell'autore

Figura 3.14 Dendrogramma – Metodo di Ward k=6 (clusters)



Fonte: Elaborazione personale dell'autore

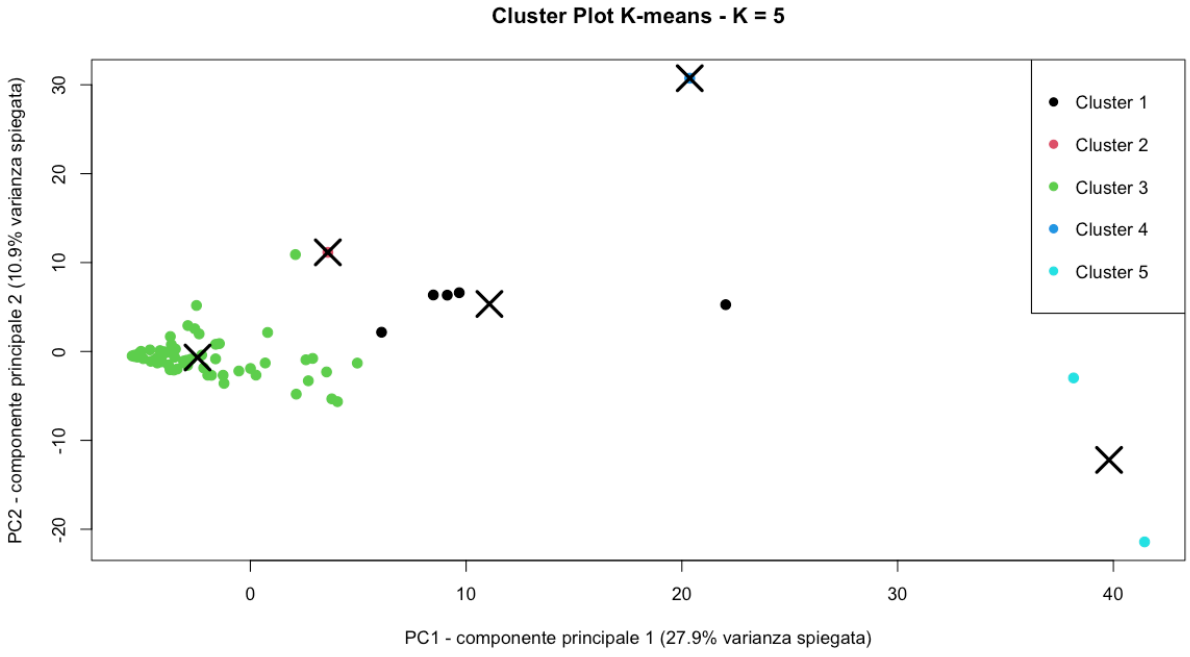
Algoritmo *K-means*

Successivamente è stato applicato l'algoritmo *K-means* con l'obiettivo di verificare la stabilità delle partizioni ottenute mediante le tecniche di *clustering* gerarchico, considerando anche in questo caso soluzioni con 5 e 6 *cluster*.

La Figura 3.15 e la Figura 3.16 mostrano la distribuzione dei comuni nello spazio definito dalle prime due componenti principali e consentono di valutare visivamente il grado di separazione tra i gruppi individuati. Com'è illustrato nelle sezioni precedenti, la scelta di utilizzare le prime due componenti principali per la costruzione dei *cluster plot* è supportata dai risultati dello *Scree plot* relativo alla PCA, dal quale emerge che tali componenti spiegano complessivamente il 38,8% della variabilità complessiva del *dataset*.

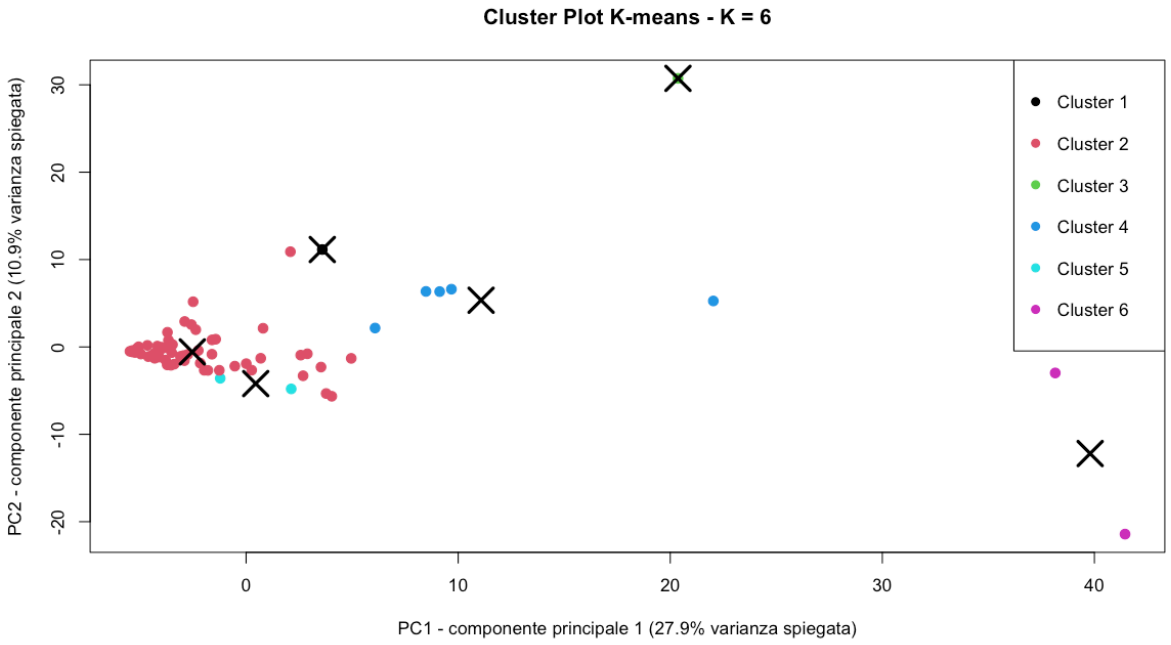
Nei *cluster plot*: ciascun punto rappresenta un comune proiettato sulle prime due componenti principali della PCA; i colori indicano il *cluster* di appartenenza; le "X" nere rappresentano i centroidi dei *cluster*, ovvero i punti medi dei gruppi nello spazio bidimensionale considerato. La distanza visiva tra punti e centroide fornisce quindi una lettura esplorativa della compattezza e della separazione dei *cluster*.

Figura 3.15 Grafico a dispersione – Cluster Plot K-means – K=5 (clusters)



Fonte: Elaborazione personale dell'autore

Figura 3.16 Grafico a dispersione – Cluster Plot K-means – K=6 (clusters)



Fonte: Elaborazione personale dell'autore

Nel caso della segmentazione con cinque *cluster* (Figura 3.15), nel complesso i gruppi appaiono ben distinti tra loro, pur in presenza di alcune aree di parziale sovrapposizione che possono essere attribuite a comuni caratterizzati da profili turistici intermedi. Nel caso della soluzione con sei *cluster* (Figura 3.16), si evidenzia invece una maggiore frammentazione della struttura dei dati e la presenza di gruppi costituiti da un numero più limitato di osservazioni.

In entrambi i *cluster plot* si osservano inoltre alcuni comuni relativamente isolati rispetto al *macrocluster* che costituisce il nucleo principale delle osservazioni. Ciò suggerisce la presenza di profili turistici particolarmente distintivi e peculiari all'interno del contesto regionale. Tali osservazioni non devono tuttavia essere interpretate come *outlier* o valori anomali in senso strettamente statistico, bensì come possibili indicazioni di destinazioni caratterizzate da modelli di sviluppo turistico significativamente differenti rispetto alla maggior parte dei comuni valdostani.

Nel complesso, le segmentazioni ottenute con cinque *cluster* sembrerebbero rappresentare il miglior compromesso tra capacità descrittiva, omogeneità interna ai gruppi e possibilità di individuare profili turistici chiaramente distinguibili. Tuttavia, la scelta della soluzione maggiormente rappresentativa del sistema turistico valdostano richiede che si proceda ad un ulteriore approfondimento delle caratteristiche dei segmenti individuati, il quale verrà sviluppato nelle sezioni successive attraverso l'analisi dei profili medi e delle principali variabili descrittive associate a ciascun *cluster*.

3.4.3 Analisi delle caratteristiche dei *cluster* e individuazione dell'algoritmo di *clustering* più adatto

L'applicazione dei diversi algoritmi di *clustering* ha consentito di ottenere segmentazioni caratterizzate da differenti livelli di dettaglio, omogeneità interna ai *cluster* e interpretabilità territoriale. Al fine di individuare la soluzione maggiormente idonea agli obiettivi del presente lavoro, è stato effettuato un confronto tra i risultati ottenuti mediante il metodo gerarchico di *Ward*, l'algoritmo *K-means* e il metodo del *Complete Linkage*, considerando sia le soluzioni a cinque *cluster* sia quelle a sei.

Sebbene il metodo del *Complete Linkage* sia stato incluso nella fase esplorativa dell'analisi al fine di ampliare il confronto tra differenti criteri di aggregazione gerarchica, esso non è stato successivamente considerato per l'analisi dettagliata delle caratteristiche dei *cluster*. Le segmentazioni ottenute mediante tale approccio risultano infatti caratterizzate da una forte concentrazione delle osservazioni in un unico gruppo principale e da numerosi *cluster* costituiti da un numero estremamente ridotto di comuni, limitando la possibilità di individuare i profili turistici territorialmente interpretabili e potenzialmente generalizzabili. Per tale ragione, il metodo è stato escluso dalla fase finale di selezione dell'algoritmo di *clustering* ritenuto maggiormente adatto al caso oggetto di studio.

Una prima valutazione delle diverse segmentazioni può essere effettuata analizzando la numerosità dei *cluster* ottenuti mediante il metodo di *Ward* e l'algoritmo *K-means* nelle configurazioni a cinque e sei *cluster* (Tabella 3.1).

Tabella 3.1 Numerosità dei cluster metodo di Ward e algoritmo K-means (k=5 e k=6)

Metodo	**k**	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**	***Cluster* 5**	***Cluster* 6**
Ward	5	69	1	2	1	1	
Ward	6	67	2	1	2	1	1
K-means	5	5	1	65	1	2	
K-means	6	1	63	1	5	2	2

Fonte: Elaborazione personale dell'autore

L'analisi della distribuzione delle osservazioni evidenzia come le soluzioni ottenute mediante il metodo di *Ward* tendono a concentrare la maggior parte dei comuni valdostani in un unico *cluster* di grandi dimensioni, isolando allo stesso tempo alcune località caratterizzate da profili turistici particolarmente distintivi. In particolare, la segmentazione a cinque *cluster* individua un gruppo principale costituito da 69 comuni e quattro *cluster* di dimensione molto ridotta,

comprendenti rispettivamente il Comune di Aosta, i Comuni di Cogne e Valtournenche, il Comune di Courmayeur e il Comune di Nus.

Una configurazione di questo tipo consente di descrivere in modo accurato le peculiarità delle principali destinazioni turistiche regionali, ma risulta meno efficace nell'identificazione di categorie territoriali più ampie. In altre parole, una segmentazione costituita prevalentemente da *cluster* contenenti uno o due comuni presenta una limitata capacità di generalizzazione dei risultati, poiché tende a rappresentare soprattutto casi eccezionali piuttosto che profili turistici maggiormente rappresentativi del sistema turistico regionale nel suo complesso.

Per approfondire ulteriormente le caratteristiche delle diverse segmentazioni, sono stati costruiti profili sintetici dei *cluster* utilizzando le variabili relative agli arrivi e alle presenze turistiche. Per ciascun comune sono stati prima calcolati i valori complessivi registrati nel periodo compreso tra gennaio e settembre 2025, e successivamente sono stati calcolati i valori osservati nei singoli mesi. In seguito, per ogni *cluster*, è stata calcolata la media aritmetica delle variabili considerate tra tutti i comuni appartenenti al medesimo gruppo.

Ad esempio, il valore relativo alle presenze registrate nel mese di luglio per il *macrocluster* principale ottenuto attraverso il metodo di *Ward* con cinque *cluster* rappresenta la media aritmetica delle presenze complessivamente osservate nel mese di luglio nei 69 comuni appartenenti a tale *cluster*. Analogamente, le tabelle relative alle tipologie ricettive riportano il numero medio di presenze registrate nei comuni appartenenti a ciascun *cluster*, ottenuto sommando prima le presenze mensili per singola tipologia ricettiva e calcolando successivamente la media all'interno di ciascun gruppo.

Tabella 3.2 Profili medi dei cluster ottenuti mediante Ward (K=5)

Cluster	**Arrivi totali**	**Presenze totali**	**Arrivi gennaio**	**Presenze gennaio**
1	11254	33025	1224	3859
2	121478	261403	11669	28449
3	99118	389422	14043	63416
4	231129	573326	19330	65072
5	13998	30801	1505	2969

Cluster	**Arrivi febbraio**	**Presenze febbraio**	**Arrivi marzo**	**Presenze marzo**
1	1258	4114	1069	3334
2	11320	29468	10126	23364
3	14088	67744	14598	63420
4	19195	70843	16975	54111
5	1454	2584	1597	2691

Cluster	**Arrivi aprile**	**Presenze aprile**	**Arrivi maggio**	**Presenze maggio**
1	618	1532	610	1145
2	10394	18740	11918	22083
3	9338	38596	3308	7469
4	4861	10132	3933	6905
5	1023	1667	1446	2200

Cluster	**Arrivi giugno**	**Presenze giugno**	**Arrivi luglio**	**Presenze luglio**
1	1175	2468	2182	6786
2	13594	25808	19234	40399
3	6954	18294	15150	51978
4	30395	50137	57689	125680
5	1144	2425	2260	7130

Cluster	**Arrivi agosto**	**Presenze agosto**	**Arrivi settembre**	**Presenze settembre**
1	2248	7824	869	1962
2	20106	47188	13117	25904
3	16614	64948	5025	13556
4	52506	138749	26245	51697
5	2533	6837	1036	2298

Fonte: Elaborazione personale dell'autore

Tabella 3.3 Presenze medie per tipologia ricettiva nei cluster ottenute mediante Ward (K=5)

Cluster	Affittacamere Chambres d'hôtes	Agriturismo	Alberghi	Alloggi ad uso turistico
1	1481	297	15544	3258
2	10336	3577	123106	68275
3	12676	3317	246030	45512
4	5683	0	382401	27063
5	1777	1326	5139	6026

Cluster	Aree di sosta	Bed & Breakfast	Campeggi	Campeggi sociali
1	156	419	3179	74
2	0	3842	0	0
3	9083	1068	16234	1806
4	0	710	89353	0
5	0	462	0	0

Cluster	Case e appartamenti per vacanze	Case per ferie	Ostelli per la gioventù	Posti tappa / Dortoir
1	1398	1110	97	82
2	39656	8713	0	0
3	19200	9036	926	16
4	6091	6446	0	2418
5	5421	1793	4444	2010

Cluster	Residenze turistico-alberghiere	Rifugi alpini	Villaggi turistici
1	3887	1236	880
2	3898	0	0
3	22253	4068	0
4	19714	33447	0
5	0	2403	0

Fonte: Elaborazione personale dell'autore

Dall'analisi dei profili associati alla soluzione ottenuta mediante il metodo di *Ward* con cinque *cluster* emerge come il *macrocluster* principale presenti livelli relativamente contenuti di arrivi e presenze turistiche e una distribuzione stagionale moderatamente concentrata nei mesi estivi, rappresentando verosimilmente il profilo turistico prevalente dei comuni valdostani. Alcuni *cluster* di dimensioni molto ridotte evidenziano invece volumi di domanda significativamente

superiori alla media regionale, suggerendo la presenza di località caratterizzate da una forte specializzazione turistica e da una capacità attrattiva particolarmente elevata.

Un'analisi analoga è stata effettuata, considerando la segmentazione ottenuta attraverso l'algoritmo *K-means* con cinque *cluster*.

Tabella 3.4 Presenze medie per tipologia ricettiva nei cluster ottenute mediante algoritmo K-means (K=5)

Cluster	**Affittacamere Chambres d'hôtes**	**Agriturismo**	**Alberghi**	**Alloggi ad uso turistico**
1	6070	1123	81724	10075
2	1777	1326	5139	6026
3	1193	304	11724	2929
4	5683	0	382401	27063
5	15007	2648	258520	71676

Cluster	**Aree di sosta**	**Bed & Breakfast**	**Campeggi**	**Campeggi sociali**
1	3849	757	11190	0
2	0	462	0	0
3	45	393	2833	78
4	0	710	89353	0
5	3378	2814	5878	1806

Cluster	**Case e appartamenti per vacanze**	**Case per ferie**	**Ostelli per la gioventù**	**Posti tappa / Dortoir**
1	4203	9037	370	111
2	5421	1793	4444	2010
3	1279	667	103	79
4	6091	6446	0	2418
5	35190	7408	0	10

Cluster	**Residenze turistico-alberghiere**	**Rifugi alpini**	**Villaggi turistici**
1	41136	4021	0
2	0	2403	0
3	1233	1057	934
4	19714	33447	0
5	15363	2288	0

Fonte: Elaborazione personale dell'autore

Tabella 3.5 Profili medi dei cluster ottenuti mediante algoritmo K-means (K=5)

Cluster	Arrivi totali	Presenze totali	Arrivi gennaio	Presenze gennaio
1	49295	173667	6455	22962
2	13998	30801	1505	2969
3	9121	24773	911	2633
4	231129	573326	19330	65072
5	128426	420179	16376	67814

Cluster	Arrivi febbraio	Presenze febbraio	Arrivi marzo	Presenze marzo
1	6524	24903	5437	19732
2	1454	2584	1597	2691
3	940	2747	793	2215
4	19195	70843	16975	54111
5	16308	72903	17160	68794

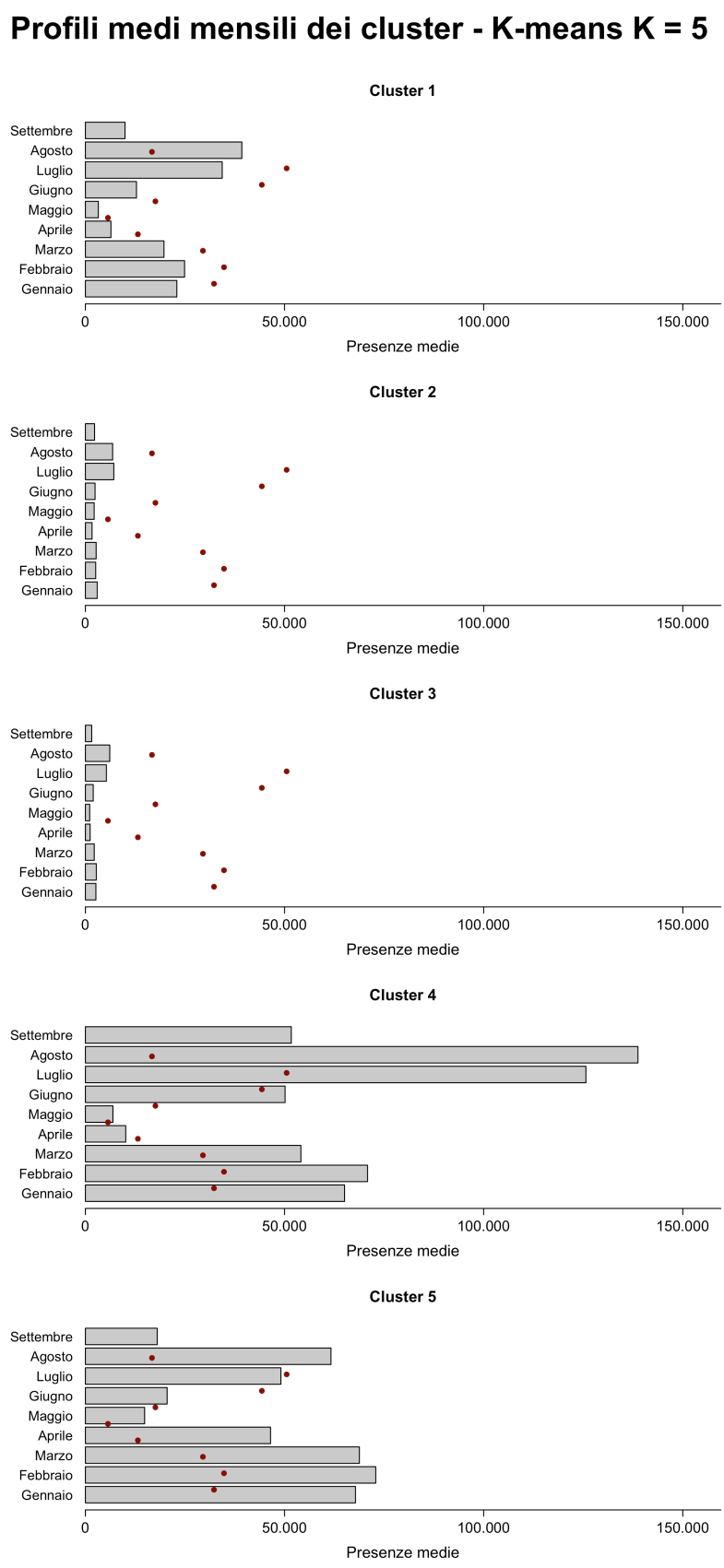
Cluster	Arrivi aprile	Presenze aprile	Arrivi maggio	Presenze maggio
1	1706	6413	1664	3245
2	1023	1667	1446	2200
3	547	1179	578	1078
4	4861	10132	3933	6905
5	13788	46486	7394	14868

Cluster	Arrivi giugno	Presenze giugno	Arrivi luglio	Presenze luglio
1	5084	12826	9380	34351
2	1144	2425	2260	7130
3	979	1962	1780	5271
4	30395	50137	57689	125680
5	9776	20522	18752	49093

Cluster	Arrivi agosto	Presenze agosto	Arrivi settembre	Presenze settembre
1	9394	39283	3651	9953
2	2533	6837	1036	2298
3	1856	6111	738	1578
4	52506	138749	26245	51697
5	20404	61665	8468	18034

Fonte: Elaborazione personale dell'autore

Figura 3.17 Profili mensili delle presenze nei cluster ottenuti mediante l'algoritmo K-means (k=5). I punti rossi rappresentano la media complessiva delle presenze mensili calcolata sull'intero dataset.



Fonte: Elaborazione personale dell'autore

Risulta particolarmente interessante la segmentazione ottenuta mediante l'algoritmo *K-means* con cinque *cluster*. In questo caso, oltre ai *cluster* associabili a località con profili turistici fortemente specifici, emerge un *cluster* composto da Ayas, Cogne, Brusson, Gressoney-Saint-Jean e La Thuile. Si tratta di comuni appartenenti alle vallate laterali della regione, accomunati da una consolidata vocazione turistica montana e dalla presenza di importanti poli di attrazione sia durante la stagione invernale sia in quella estiva.

La capacità dell'algoritmo di partizionamento di individuare questo insieme di località suggerisce una maggiore sensibilità nel cogliere alcune specificità territoriali che risultano meno evidenti nelle segmentazioni ottenute attraverso il *clustering* gerarchico. Inoltre, la presenza di un *cluster* composto da cinque comuni consente di individuare una categoria di destinazioni turistiche maggiormente rappresentativa e interpretabile rispetto alle soluzioni ottenute mediante il metodo di *Ward*, caratterizzate prevalentemente dall'isolamento di singole località con profili eccezionali.

Nel complesso i risultati ottenuti mostrano un'elevata convergenza tra le diverse metodologie applicate, confermando l'esistenza di un nucleo consistente di comuni caratterizzati da profili turistici relativamente omogenei e di un numero limitato di località che si distinguono per intensità dei flussi turistici, composizione dell'offerta ricettiva e grado di specializzazione turistica.

Con il termine grado di specializzazione turistica si intende il livello di concentrazione dell'attività turistica di una località rispetto a particolari segmenti di domanda, periodi dell'anno o modalità di fruizione dell'offerta ricettiva. Ad esempio, località quali Cogne, Ayas, Gressoney-Saint-Jean, La Thuile e Brusson possono essere considerate maggiormente specializzate rispetto al profilo turistico medio regionale, non tanto per la presenza di attività riconducibili al turismo montano, diffuso in tutta la regione, quanto per la capacità di attrarre consistenti flussi turistici e per il ruolo che svolgono in quanto località appartenenti a vallate laterali tradizionalmente caratterizzate da una marcata vocazione turistica.

Sebbene l'approccio gerarchico di *Ward* consenta di individuare *cluster* caratterizzati da un'elevata omogeneità interna, favorita dall'impiego della distanza euclidea quadratica che attribuisce maggiore peso alle osservazioni caratterizzate da distanze più elevate, la presenza di gruppi costituiti da un numero estremamente limitato di comuni può ridurre la capacità di generalizzazione dei risultati.

Nonostante tale limite, il metodo di *Ward* conserva però un elemento di particolare interesse dal punto di vista interpretativo, come illustrato nel paragrafo dedicato al *clustering* gerarchico nel capitolo 2. Infatti, a differenza dei metodi non gerarchici o di partizionamento (*partitioning methods*), il *clustering* gerarchico non restituisce esclusivamente una segmentazione finale, ma rappresenta attraverso il dendrogramma l'intero processo di aggregazione progressiva delle osservazioni. L'analisi del dendrogramma completo (Figura 3.18) consente pertanto di approfondire la struttura delle relazioni esistenti tra i comuni valdostani, evidenziando come il *macrocluster* principale individuato nella soluzione a cinque *cluster* sia, a sua volta, costituito da sottogruppi caratterizzati da differenti livelli di similarità.

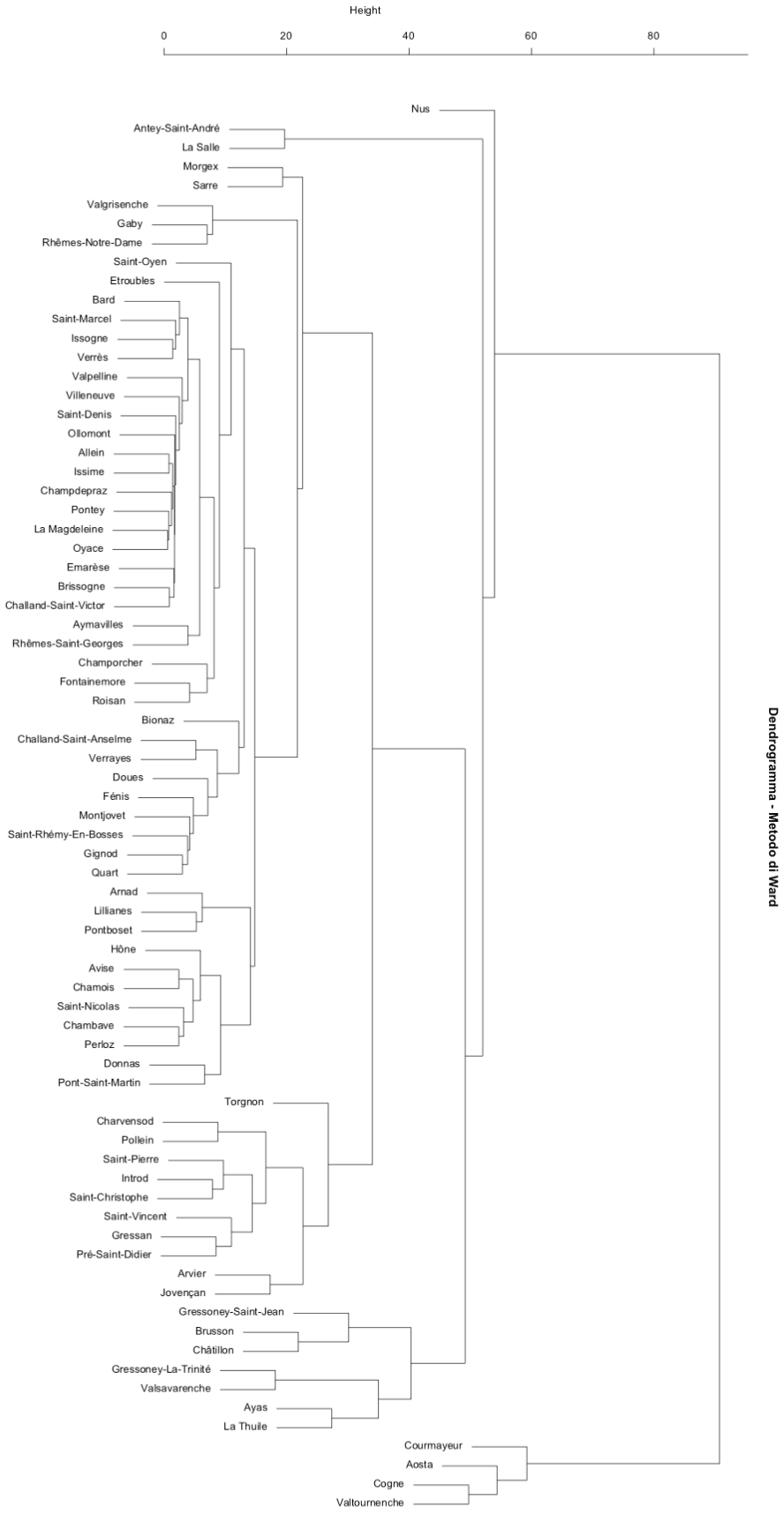
Tabella 3.6 Composizione dei cluster ottenuti mediante il metodo di Ward (k=5)

Cluster	**Comuni**
1	Allein, Antey-Saint-André, Arnad, Arvier, Avise, Ayas, Aymavilles, Bard, Bionaz, Brissogne, Brusson, Challand-Saint-Anselme, Challand-Saint-Victor, Chambave, Chamois, Champdepraz, Champorcher, Charvensod, Châtillon, Donnas, Doues, Emarèse, Etroubles, Fontainemore, Fénis, Gaby, Gignod, Gressan, Gressoney-La-Trinité, Gressoney-Saint-Jean, Hône, Introd, Issime, Issogne, Jovençan, La Magdeleine, La Salle, La Thuile, Lillianes, Montjovet, Morgex, Ollomont, Oyace, Perloz, Pollein, Pont-Saint-Martin, Pontboset, Pontey, Pré-Saint-Didier, Quart, Rhêmes-Notre-Dame, Rhêmes-Saint-Georges, Roisan, Saint-Christophe, Saint-Denis, Saint-Marcel, Saint-Nicolas, Saint-Oyen, Saint-Pierre, Saint-Rhémy-En-Bosses, Saint-Vincent, Sarre, Torgnon, Valgrisenche, Valpelline, Valsavarenche, Verrayes, Verrès, Villeneuve
2	Aosta
3	Cogne, Valtournenche
4	Courmayeur
5	Nus

Fonte: Elaborazione personale dell'autore

La Tabella 3.6 evidenzia come la soluzione ottenuta mediante il metodo di *Ward* con cinque *cluster* sia costituita da un ampio *macrocluster* comprendente la maggior parte dei comuni valdostani e da quattro *cluster* di dimensioni ridotte, corrispondenti ai Comuni di Aosta, Courmayeur, Nus e alla coppia di Cogne e Valtournenche. Come evidenziato in precedenza, la simile configurazione risulta meno adatta quale soluzione interpretativa finale di segmentazione, ma offre l'opportunità di analizzare nel dettaglio le relazioni gerarchiche presenti all'interno del *macrocluster* principale.

Figura 3.18 Dendrogramma completo – Metodo di Ward



Fonte: Elaborazione personale dell'autore

L’osservazione del dendrogramma completo ottenuto mediante il metodo di *Ward* mostra come il *macrocluster* non costituisca un insieme omogeneo e indistinto di comuni, bensì presenti una struttura interna articolata. In particolare, è possibile riconoscere alcuni nuclei di comuni che si aggregano a livelli di distanza relativamente contenuti, suggerendo l’esistenza di sottogruppi caratterizzati da una maggiore similarità reciproca.

Un primo comprende i comuni di Antey-Saint-André, La Salle, Morgex e Sarre ai quali si associa successivamente un secondo piccolo raggruppamento costituito da Valgrisenche, Gaby, Rhêmes-Notre-Dame e Saint-Oyen. Il nucleo centrale del *macrocluster* raccoglie invece numerosi comuni caratterizzati da profili turistici complessivamente simili, ai quali si aggregano progressivamente ulteriori sottogruppi, tra cui quello comprendente Bionaz, Challand-Saint-Anselme, Verrayes, Doues, Fénis, Montjovet, Saint-Rhémy-en-Bosses, Gignod e Quart e quello formato da Arnad, Lillianes, Pontboset, Hône, Avise, Chamois, Saint-Nicolas, Chambave, Perloz, Donnas e Pont-Saint-Martin.

Andando verso i livelli di aggregazione più elevati emerge inoltre un ulteriore gruppo costituito da Charvensod, Pollein, Saint-Pierre, Introd, Saint-Cristophe, Saint-Vincent, Gressan, Pré-Saint-Didier, Arvier, Jovençan, Châtillon, Gressoney-La-Trinité, Valsavarenche, Ayas e La Thuile. Quest’ultimo appare particolarmente interessante in quanto raggruppa prevalentemente comuni appartenenti ad alcune delle principali vallate laterali della regione, caratterizzata da una consolidata vocazione turistica e da una marcata specializzazione nell’offerta ricettiva montana.

Nonostante non costituiscano una nuova segmentazione del *dataset*, tali diramazioni consentono di cogliere livelli intermedi di similarità che la semplice partizione finale non è in grado di rappresentare. Il dendrogramma completo ottenuto mediante il metodo di *Ward* aggiunge pertanto un ulteriore livello interpretativo all’analisi, evidenziando come il *macrocluster* principale sia costituito a sua volta da nuclei di comuni progressivamente più simili tra loro, e fornendo inoltre una rappresentazione più articolata della struttura del sistema turistico valdostano.

Tali considerazioni non modificano tuttavia la scelta della soluzione interpretativa finale adottata nello studio, che rimane quella ottenuta mediante l’algoritmo *K-means* con cinque *cluster*, ritenuta la più adatta per l’analisi dettagliata dei profili turistici comunali, e quindi maggiormente coerente con gli obiettivi della ricerca, che verrà sviluppata nel paragrafo successivo.

3.4.4 Descrizione dei *cluster* ottenuti mediante algoritmo *K-means* (k=5)

In seguito alle considerazioni sviluppate nel paragrafo precedente, la segmentazione ottenuta mediante l'algoritmo *K-means* con cinque *cluster* (Tabella 3.1) è stata ritenuta la soluzione più coerente con gli obiettivi conoscitivi della ricerca.

La Tabella 3.7 riporta la composizione dei *cluster* individuati, mentre la Figura 3.15 evidenzia visivamente una buona separazione tra i *cluster* individuati mediante l'algoritmo *K-means*, mostrando in particolare la presenza di alcuni gruppi chiaramente distinti rispetto al nucleo principale costituito dalla maggior parte dei comuni valdostani. Tale configurazione suggerisce l'esistenza di differenti profili turistici territoriali, descritti nei paragrafi successivi.

Tabella 3.7 Composizione dei cluster ottenuti mediante l'algoritmo K-means (k=5)

Cluster	**Comuni**
1	Ayas, Brusson, Cogne, Gressoney-Saint-Jean, La Thuile
2	Nus
3	Allein, Antey-Saint-André, Arnad, Arvier, Avise, Aymavilles, Bard, Bionaz, Brissogne, Challand-Saint-Anselme, Challand-Saint-Victor, Chambave, Chamois, Champdepraz, Champorcher, Charvensod, Châtillon, Donnas, Doues, Emarèse, Etroubles, Fontainemore, Fénis, Gaby, Gignod, Gressan, Gressoney-La-Trinité, Hône, Introd, Issime, Issogne, Jovençan, La Magdeleine, La Salle, Lillianes, Montjovet, Morgex, Ollomont, Oyace, Perloz, Pollein, Pont-Saint-Martin, Pontboset, Pontey, Pré-Saint-Didier, Quart, Rhêmes-Notre-Dame, Rhêmes-Saint-Georges, Roisan, Saint-Christophe, Saint-Denis, Saint-Marcel, Saint-Nicolas, Saint-Oyen, Saint-Pierre, Saint-Rhémy-En-Bosses, Saint-Vincent, Sarre, Torgnon, Valgrisenche, Valpelline, Valsavarenche, Verrayes, Verrès, Villeneuve
4	Courmayeur
5	Aosta, Valtournenche

Fonte: Elaborazione personale dell'autore

Cluster 1 – Destinazioni turistiche delle vallate laterali ad alta attrattività

Il *cluster* comprende Ayas, Brusson, Cogne, Gressoney-Saint-Jean e La Thuile. Si tratta di località appartenenti a vallate laterali della Regione tradizionalmente caratterizzate da una marcata vocazione turistica. Pur presentando specificità territoriali differenti, tali comuni condividono livelli relativamente elevati di domanda turistica e svolgono un ruolo rilevante nel sistema turistico regionale, sia durante la stagione invernale, sia in quella estiva.

La capacità dell'algoritmo *K-means* di individuare autonomamente questo insieme di località suggerisce l'esistenza di un profilo turistico intermedio tra le principali destinazioni regionali e il gruppo prevalente dei comuni valdostani. In particolare, tali località appaiono accomunate da una capacità attrattiva relativamente elevata ed una consolidata funzione turistica, pur non raggiungendo i livelli osservati nelle destinazioni maggiormente specializzate rappresentate dai *cluster* 4 e 5.

Cluster 2 – Comune con profilo peculiare

Il *cluster* comprende esclusivamente il Comune di Nus, il quale pur non presentando livelli di intensità dei flussi turistici comparabili a quelli delle principali destinazioni regionali (Tabella 3.5), mostra una configurazione delle presenze turistiche (Figura 3.17) sufficientemente distinta rispetto agli altri comuni tanto da determinare la formazione di un *cluster* autonomo (*singleton cluster*).

La presenza di un *cluster* costituito da una sola osservazione suggerisce l'esistenza di una configurazione turistica distinta rispetto a quella osservata nel restante sistema regionale. Un approfondimento condotto attraverso il confronto tra i dati comunali di Nus e quelli degli altri comuni valdostani non ha evidenziato una singola variabile in grado di spiegare, da sola, l'isolamento del comune. Tale risultato sembra essere piuttosto riconducibile a una particolare combinazione di caratteristiche relative sia alla distribuzione temporale dei flussi turistici sia alla distribuzione delle presenze tra le diverse tipologie ricettive.

In assenza di un'evidenza empirica riconducibile ad un unico fattore dominante, si ritiene dunque opportuno mantenere un'interpretazione prudente del risultato, limitandosi a evidenziare l'esistenza di un profilo turistico differenziato che potrebbe essere oggetto di futuri approfondimenti.

Cluster 3 – Sistema turistico diffuso valdostano

Il *cluster* comprende 65 comuni e rappresenta il profilo turistico prevalente della Valle d'Aosta. I comuni appartenenti a questo *cluster* presentano livelli relativamente contenuti di arrivi e presenze turistiche (Tabella 3.5) e una struttura dell'offerta ricettiva (Tabella 3.4) generalmente meno specializzata rispetto ai *cluster* caratterizzati dai principali poli turistici regionali. Inoltre, mostrano una distribuzione delle presenze tra le diverse tipologie ricettive generalmente più equilibrata.

La presenza di un numero elevato di osservazioni suggerisce l'esistenza di un modello turistico relativamente omogeneo e diffuso sul territorio valdostano, caratterizzato da una funzione turistica presente ma meno intensiva rispetto alle località maggiormente orientate all'attività turistica.

Cluster 4 – Destinazione internazionale ad elevata intensità turistica

Il *cluster* comprende esclusivamente il Comune di Courmayeur. La presenza di un *cluster* autonomo conferma la forte specificità del profilo turistico della località rispetto al resto del sistema turistico valdostano.

Come evidenziato dai profili medi riportati nella Tabella 3.5, il *cluster* presenta i più elevati livelli di domanda turistica osservati nel campione, sia in termini di arrivi sia di presenze complessive nel periodo compreso tra gennaio e settembre 2025. Inoltre, l'analisi dei flussi turistici mensili nel periodo oggetto d'analisi (Tabella 3.8) mostra una marcata capacità attrattiva durante entrambe le principali stagioni turistiche regionali (invernale ed estiva), con consistenti volumi di domanda sia nei mesi invernali, in particolare tra gennaio e marzo, sia nel periodo estivo compreso tra giugno e settembre. Poiché il *cluster* è costituito da una sola osservazione, i valori medi riportati nella tabella coincidono con quelli effettivamente registrati dal Comune di Courmayeur.

Tabella 3.8 Flussi turistici del Comune di Courmayeur nel periodo tra gennaio e settembre 2025

Indicatore	**Valore**
Arrivi totali	231129
Presenze totali	573326
Arrivi gennaio	19330
Presenze gennaio	65072
Arrivi febbraio	19195
Presenze febbraio	70843
Arrivi marzo	16975
Presenze marzo	54111
Arrivi giugno	30395
Presenze giugno	50137
Arrivi luglio	57689
Presenze luglio	125680
Arrivi agosto	52506
Presenze agosto	138749
Arrivi settembre	26245
Presenze settembre	51697

Fonte: Elaborazione personale dell'autore

L'analisi delle presenze medie per tipologia di struttura ricettiva (Tabella 3.4) evidenzia inoltre una maggiore concentrazione della domanda turistica in alcune specifiche categorie di strutture, suggerendo un elevato grado di specializzazione turistica[1] della località.

Tali evidenze appaiono coerenti con il ruolo svolto a Courmayeur, in quanto si tratta di uno dei principali poli turistici della Valle d'Aosta, e contribuiscono a spiegare la formazione di un *cluster* autonomo nell'ambito della segmentazione ottenuta mediante l'algoritmo *K-means*.

Cluster 5 – Destinazioni ad elevata centralità turistica
Il *cluster* comprende i Comuni di Aosta e Valtournenche. Nonostante le due località presentino caratteristiche territoriali differenti, entrambe registrano consistenti volumi di domanda turistica e svolgono una funzione centrale all'interno del sistema turistico regionale.

Nel caso di Aosta, si ipotizza che tale ruolo sia riconducibile alla presenza del capoluogo regionale e alla prossimità con il comprensorio sciistico di Pila, mentre Valtournenche beneficia della presenza del comprensorio di Breuil-Cervinia, tra le principali destinazioni sciistiche della regione e non solo, in quanto entrambi i comprensori sfruttano gli impianti di risalita anche nel periodo estivo per ragioni dedicate all'escursionismo e alla mountain bike.

L'accorpamento di questi due comuni all'interno di un medesimo *cluster* suggerisce la presenza di caratteristiche comuni in termini di intensità dei flussi turistici e capacità attrattiva complessiva.

3.5 Implicazioni del caso di studio

3.5.1 Considerazioni critiche, limiti e vantaggi del *clustering*

L'analisi condotta ha evidenziato come il sistema turistico della Valle d'Aosta presenti una struttura complessivamente piuttosto omogenea, all'interno della quale emergono tuttavia alcune destinazioni caratterizzate da una marcata specializzazione turistica. Indipendentemente dalla metodologia adottata e dal numero di *cluster* considerato, tutte le segmentazioni ottenute hanno infatti individuato un *macrocluster* principale comprendente oltre l'85% dei comuni valdostani, affiancato da un numero limitato di *cluster* di dimensioni ridotte associati ai principali poli turistici regionali.

[1] In questo contesto, con il termine grado di specializzazione turistica si intende la presenza di caratteristiche distintive dei flussi turistici e delle modalità di utilizzo delle diverse tipologie ricettive, tali da rendere una destinazione maggiormente riconoscibile rispetto al profilo turistico regionale prevalentemente osservato.

Tale risultato suggerisce come, pur esistendo differenze significative tra alcune destinazioni ad elevata attrattività, il sistema turistico regionale sia costituito prevalentemente da comuni che condividono caratteristiche relativamente simili in termini di domanda turistica e composizione dell'offerta ricettiva.

Sotto questo profilo, il confronto tra il metodo gerarchico di *Ward* e l'algoritmo *K-means*, già discusso nel capitolo 3.4.3, ha rappresentato anche un utile strumento di validazione dei risultati ottenuti. Nonostante le due metodologie producano segmentazioni non perfettamente coincidenti nella composizione dei *cluster* minori, entrambe convergono nell'individuare una struttura generale sostanzialmente analoga, caratterizzata da un ampio gruppo di comuni con profili turistici simili e da pochi comuni nettamente differenziati.

Tale convergenza non costituisce una dimostrazione dell'esistenza di una classificazione univoca, ma rappresenta un elemento che rafforza la robustezza dell'analisi, suggerendo che le principali configurazioni individuate riflettano caratteristiche effettivamente presenti nei dati piuttosto che essere esclusivamente il risultato delle scelte metodologiche adottate.

Allo stesso tempo, è importante sottolineare come il *clustering* costituisca una metodologia di natura esplorativa e descrittiva, i cui risultati dipendono inevitabilmente dalle decisioni assunte durante il processo di analisi. Infatti, la scelta delle variabili considerate, la standardizzazione dei dati, la misura di distanza adottata, il criterio di aggregazione utilizzato nel *clustering* gerarchico e la definizione del numero finale di *cluster* rappresentano fattori in grado di influenzare sensibilmente la segmentazione ottenuta. Di conseguenza, i *cluster* individuati non devono essere interpretati come categorie assolute o definitive, bensì come una rappresentazione sintetica della struttura dei dati osservati nel periodo oggetto di analisi.

Un ulteriore limite riguarda la natura stessa delle tecniche di *clustering* impiegate nel presente studio. Sia il metodo di *Ward* sia l'algoritmo *K-means* prevedono che ogni osservazione venga assegnata esclusivamente ad un unico *cluster*. Tale approccio facilita l'interpretazione dei risultati e consente di ottenere segmentazioni chiaramente distinguibili, ma può risultare riduttivo nel caso di sistemi territoriali complessi, nei quali alcuni comuni potrebbero condividere caratteristiche riconducibili a più profili turistici contemporaneamente. In quest'ottica, un possibile sviluppo futuro potrebbe consistere nell'applicazione di metodologie di *fuzzy clustering*, che consentono di attribuire ad ogni osservazione differenti gradi di appartenenza ai diversi *cluster*, rappresentando in modo più flessibile situazioni caratterizzate da confini meno definiti tra i gruppi (Bezdek et al., 1984).

Occorre inoltre considerare i limiti connessi ai dati utilizzati. L'analisi è stata sviluppata esclusivamente sulla base delle informazioni disponibili relative agli arrivi, alle presenze turistiche e alla loro distribuzione tra le diverse tipologie di strutture ricettive nel periodo oggetto di analisi. Nonostante tali variabili costituiscano indicatori particolarmente significativi dell'intensità della domanda e delle caratteristiche dell'offerta, esse non sono sufficienti a spiegare la complessità del fenomeno turistico nella sua interezza. L'integrazione di ulteriori informazioni, quali indicatori socioeconomici, infrastrutturali, ambientali, dati relativi alla spesa dei visitatori o alla permanenza media, potrebbe contribuire a una descrizione ancora più completa delle differenze esistenti tra i diversi comuni, consentendo l'individuazione di profili turistici maggiormente articolati. Tale considerazione risulta coerente con quanto riportato nella review della letteratura, che evidenzia come la qualità di una segmentazione dipenda in larga parte dalla selezione delle variabili utilizzate per descrivere il fenomeno oggetto di studio.

Nel complesso, i risultati ottenuti mostrano come le tecniche di *clustering* rappresentino uno strumento efficace per supportare l'analisi esplorativa dei sistemi turistici territoriali, consentendo di sintetizzare grandi quantità di informazioni in un numero limitato di profili facilmente interpretabili. La loro applicazione richiede allo stesso tempo un'attenta valutazione critica delle scelte metodologiche adottate e delle caratteristiche dei dati disponibili, in modo tale che le segmentazioni ottenute vengano interpretate come strumenti di supporto all'analisi e non come classificazioni definitive della realtà osservata.

Infine, è particolarmente rilevante ricordare che le tecniche adottate nel presente lavoro rientrano nell'ambito dell'apprendimento non supervisionato (*unsupervised learning*), in quanto consentono di individuare strutture latenti nei dati a partire dai dati disponibili. Pertanto, una naturale evoluzione del presente lavoro potrebbe consistere inizialmente nel confronto con ulteriori algoritmi di apprendimento non supervisionato e, successivamente, nello sviluppo di modelli di apprendimento supervisionato (*supervised learning*) in maniera tale da poter svolgere un'analisi predittiva dei flussi turistici regionali.

3.5.2 Considerazioni sulle tipologie di segmentazione adottate

Come evidenziato nella review della letteratura, gli studi sulla segmentazione turistica possono essere ricondotti a due principali approcci metodologici: la segmentazione a priori, nella quale i gruppi vengono definiti sulla base di criteri stabiliti prima dell'analisi, e la segmentazione a posteriori o *data-driven*, nella quale i segmenti emergono direttamente dai dati attraverso l'applicazione di tecniche statistiche o di *Machine Learning*.

Nel presente elaborato è stato adottato un approccio di segmentazione *data-driven*, ritenuto più coerente con gli obiettivi della ricerca. Infatti, l'intento non era quello di verificare l'esistenza di categorie turistiche già note o costruite sulla base di conoscenze preliminari del territorio, bensì quello di individuare eventuali profili omogenei emergenti direttamente dalle caratteristiche osservate nei dati relativi ai flussi turistici e alla distribuzione della domanda tra le diverse tipologie di strutture ricettive nei comuni valdostani.

L'applicazione delle tecniche di *clustering* ha consentito di identificare una struttura della domanda turistica che, pur evidenziando la presenza di alcune destinazioni caratterizzate da una marcata specializzazione, ha mostrato come gran parte dei comuni presenti caratteristiche relativamente simili. Tale evidenza rappresenta uno dei principali risultati emersi dalla ricerca, in quanto suggerisce che il sistema turistico regionale risulti complessivamente più omogeneo di quanto si potrebbe ipotizzare sulla base di una classificazione costruita esclusivamente mediante criteri geografici, o di altra natura.

L'approccio adottato ha inoltre evidenziato come alcune segmentazioni risultino immediatamente interpretabili, mentre altre richiedono una maggiore attenzione e cautela nell'interpretazione. In questo senso, il *clustering* non attribuisce automaticamente un significato turistico ai gruppi individuati, ma fornisce una classificazione statistica che necessita di essere interpretata e quindi letta e contestualizzata alla luce delle caratteristiche del territorio e della conoscenza del fenomeno analizzato. La fase interpretativa rappresenta quindi un elemento fondamentale ed essenziale dell'intero processo di segmentazione ed è la naturale conseguenza di un'analisi quantitativa.

La classificazione proposta può rappresentare un utile strumento conoscitivo a supporto della comprensione della struttura del sistema turistico valdostano e delle attività di pianificazione, promozione e gestione territoriale. Tuttavia, essa non deve essere considerata uno strumento sostitutivo dell'esperienza degli operatori del settore turistico o delle valutazioni dei decisori pubblici, bensì deve essere un supporto quantitativo in grado di integrare ulteriori elementi conoscitivi e favorire decisioni maggiormente basate sull'evidenza empirica. Si tratta quindi di uno strumento quantitativo complementare.

Concludendo, è opportuno sottolineare come la segmentazione ottenuta rappresenti una fotografia del sistema turistico valdostano limitata al periodo oggetto di analisi. Come osserva Dolnicar (2014), una soluzione di segmentazione costituisce infatti uno *"snapshot of the market at the time of analysis"*, poiché riflette le caratteristiche del mercato nel momento in cui i dati

vengono raccolti e analizzati. La stessa autrice evidenzia inoltre l'importanza di utilizzare dati il più possibile recenti, sottolineando come l'evoluzione del mercato possa rendere rapidamente superate le segmentazioni costruite su informazioni non aggiornate. In questo senso, la scelta di utilizzare i dati riferiti al periodo più recente disponibile e statisticamente maggiormente omogeneo, già motivata nel capitolo 3.1, risulta coerente con le indicazioni metodologiche della letteratura scientifica.

Ciononostante, i *cluster* individuati nel presente studio non devono essere interpretati come una definitiva classificazione dei comuni valdostani, ma come una rappresentazione della struttura del sistema turistico regionale nel periodo considerato. Eventuali cambiamenti nei flussi turistici, nell'offerta ricettiva o nelle dinamiche della domanda potrebbero modificare nel tempo la configurazione dei *cluster*, rendendo opportuno l'aggiornamento periodico dell'analisi attraverso dati più recenti, così da monitorarne l'evoluzione nel tempo e verificare l'eventuale presenza di cambiamenti strutturali.

Conclusione

La trattazione sviluppata nel presente capitolo ha consentito di applicare le tecniche di *clustering* illustrate nel capitolo precedente all'analisi dei dati relativi ai flussi turistici dei comuni della Valle d'Aosta. Dopo la presentazione del dataset e delle principali operazioni di preparazione e standardizzazione dei dati, sono state svolte alcune analisi descrittive preliminari e successivamente applicati differenti algoritmi di *clustering*, appartenenti sia ai metodi gerarchici sia ai metodi non gerarchici.

Il confronto tra le diverse segmentazioni ottenute ha permesso di individuare una soluzione interpretativa coerente con gli obiettivi della ricerca, evidenziando come il sistema turistico valdostano presenti una struttura complessivamente omogenea, all'interno della quale emergono alcuni comuni caratterizzati da profili turistici particolarmente distintivi.

Alla luce delle valutazioni quantitative e interpretative sviluppate nel corso dell'analisi, la segmentazione ottenuta mediante l'algoritmo *K-means* con cinque *cluster* è stata ritenuta la soluzione maggiormente rappresentativa della struttura del sistema turistico oggetto di studio. L'analisi dei *cluster* ha inoltre mostrato come le tecniche di *clustering* consentano di sintetizzare un insieme complesso di informazioni multidimensionali in un numero limitato di profili territoriali interpretabili, facilitando la lettura delle principali caratteristiche del sistema turistico valdostano.

I risultati ottenuti costituiscono la base per le considerazioni sviluppate nel capitolo conclusivo, nel quale vengono sintetizzati i principali risultati della ricerca, fornite risposte alle domande di ricerca e discussi il valore, il contributo e i limiti della ricerca. Il capitolo prosegue con alcune riflessioni sui futuri sviluppi della ricerca e si chiude con alcune considerazioni finali sul lavoro svolto.

CONCLUSIONI

Principali risultati ottenuti

Il presente lavoro ha affrontato il tema della segmentazione dei sistemi turistici territoriali attraverso l'applicazione di tecniche di *clustering* ai dati comunali relativi ai flussi turistici della Valle d'Aosta. A partire dalle principali evidenze emerse nella letteratura sulla segmentazione turistica *data-driven*, la ricerca ha proposto un percorso metodologico finalizzato all'individuazione di profili territoriali omogenei emergenti direttamente dai dati osservati, evitando il ricorso a classificazioni definite a priori.

L'analisi è stata sviluppata utilizzando informazioni relative agli arrivi, alle presenze turistiche e alla distribuzione della domanda tra le diverse tipologie di strutture ricettive nei comuni valdostani nel periodo compreso tra gennaio e settembre 2025. Attraverso il confronto tra differenti tecniche di *clustering* e diverse configurazioni di segmentazione, è stato possibile individuare una soluzione interpretativamente coerente, successivamente analizzata alla luce delle caratteristiche dei comuni appartenenti ai diversi *cluster*.

I risultati ottenuti hanno consentito non soltanto di descrivere la struttura del sistema turistico valdostano nel periodo considerato, ma anche di riflettere sul contributo che un approccio di segmentazione *data-driven* può offrire all'analisi quantitativa dei sistemi turistici territoriali.

L'analisi condotta consente innanzitutto di affermare che l'approccio metodologico adottato si è dimostrato coerente con l'obiettivo della ricerca. L'applicazione delle tecniche di *clustering* ha infatti permesso di costruire una segmentazione del sistema turistico valdostano basata esclusivamente sulle informazioni contenute nei dati osservati, senza ricorrere a classificazioni definite a priori. Tale risultato conferma come un approccio di segmentazione *data-driven* sia in grado di individuare configurazioni territoriali interpretabili, lasciando emergere direttamente dai dati profili caratterizzati da differenti livelli di intensità della domanda turistica e da diverse modalità di distribuzione delle presenze tra le tipologie ricettive.

Dall'analisi empirica è emersa una rappresentazione del sistema turistico valdostano caratterizzata da una prevalente omogeneità territoriale, all'interno della quale si distinguono tuttavia alcuni comuni con una marcata specializzazione turistica. In particolare, tutte le configurazioni analizzate, indipendentemente dall'algoritmo impiegato e dal numero di *cluster* considerato, hanno evidenziato la presenza di un ampio *macrocluster* principale comprendente la maggior parte dei comuni, affiancato da un numero limitato di *cluster* di dimensioni ridotte riconducibili ai principali poli turistici regionali. Tale evidenza suggerisce che le differenze più

marcate riguardino un insieme circoscritto di comuni, mentre gran parte del territorio regionale presenta caratteristiche relativamente simili in termini di intensità della domanda e distribuzione delle presenze tra le diverse tipologie ricettive.

Questo risultato assume particolare rilevanza se confrontato con una lettura esclusivamente descrittiva dei dati. L'analisi delle singole variabili consente infatti di evidenziare differenze nei livelli di arrivi e presenze o nella loro distribuzione tra le diverse tipologie ricettive, ma difficilmente consente di cogliere in maniera sintetica le relazioni che caratterizzano contemporaneamente tali dimensioni. La segmentazione ottenuta mediante *clustering* ha invece consentito di integrare tali informazioni in un numero limitato di profili territoriali interpretabili, vale a dire i *cluster* ottenuti, offrendo una rappresentazione complessiva della struttura del sistema turistico regionale.

In tale prospettiva, i risultati ottenuti evidenziano inoltre come le tecniche di *clustering* possano rappresentare uno strumento quantitativo complementare per l'analisi dei sistemi turistici territoriali. Anziché fornire classificazioni definitive dei territori, tali metodologie consentono di sintetizzare grandi quantità di informazioni in profili interpretabili, offrendo un supporto alla comprensione delle dinamiche territoriali e ai processi di pianificazione e gestione delle destinazioni turistiche. Il loro impiego richiede tuttavia una costante attenzione alle scelte metodologiche adottate, all'aggiornamento costante e continuativo dei dati impiegati nell'analisi e, conseguentemente, ad un'interpretazione critica dei risultati ottenuti.

Risposte alle domande di ricerca

Alla luce dei risultati ottenuti, è possibile rispondere in modo diretto ai quesiti di ricerca formulati nell'introduzione.

Con riferimento al primo quesito, l'analisi ha mostrato come le tecniche di *clustering* consentano effettivamente di individuare profili territoriali omogenei emergenti direttamente dai dati, sintetizzando in un numero limitato di *cluster* la complessità delle informazioni relative ai flussi turistici della Valle d'Aosta e alla distribuzione della domanda tra le diverse tipologie di strutture ricettive nel periodo oggetto di analisi.

Per quanto riguarda il secondo quesito, la segmentazione ottenuta ha evidenziato un sistema turistico regionale caratterizzato da un ampio nucleo di comuni con profili tra loro simili, affiancato da un numero limitato di comuni con una marcata specializzazione turistica, riconducibile principalmente all'elevata intensità dei flussi turistici. L'analisi del dendrogramma ottenuto mediante il metodo di *Ward* ha permesso inoltre di individuare,

all'interno del *macrocluster* principale, ulteriori sottogruppi interpretabili anche dal punto di vista territoriale, suggerendo l'esistenza di differenti configurazioni turistiche che non emergono immediatamente da una semplice classificazione in cinque *cluster*.

Concludendo, con riferimento al terzo quesito, il lavoro conferma che le tecniche di *clustering* possono rappresentare di fatto un valido strumento quantitativo complementare all'analisi descrittiva tradizionale. Pur non sostituendo il giudizio dell'analista, esse consentono infatti di sintetizzare grandi quantità di informazioni, mettere in evidenza strutture latenti nei dati e fornire un supporto ai processi di analisi, pianificazione e gestione dei sistemi turistici territoriali.

Valore e contributo della ricerca

Alla luce dei risultati ottenuti, il presente studio consente di evidenziare alcuni contributi che, pur inserendosi in quelli già forniti dalla letteratura esistente, risultano significativi con riferimento al caso di studio analizzato. Tali contributi possono essere ricondotti a quattro principali dimensioni:

i. Un contributo metodologico, rappresentato dall'applicazione di un approccio di segmentazione *data-driven* ai dati sui flussi turistici dei comuni della Valle d'Aosta;
ii. Un contributo empirico, legato alla ricostruzione della struttura del sistema turistico regionale nel periodo oggetto di analisi;
iii. Un contributo interpretativo, derivante dalla possibilità di sintetizzare la complessità dei dati in un numero limitato di profili territoriali interpretabili;
iv. Un contributo applicativo, connesso al potenziale impiego della segmentazione come strumento quantitativo di supporto all'analisi e alla pianificazione dei sistemi turistici territoriali.

Il principale valore della presente ricerca non risiede nell'introduzione di nuove metodologie di segmentazione, bensì nell'applicazione rigorosa di strumenti consolidati ad un contesto territoriale specifico, verificandone la capacità di offrire una lettura sintetica e interpretabile della struttura del sistema turistico valdostano. In tale prospettiva, la ricerca si inserisce coerentemente nel filone degli studi sulla segmentazione turistica *data-driven*, fornendo un'ulteriore evidenza empirica a supporto delle potenzialità che la letteratura attribuisce alle tecniche di *clustering*.

Tale aspetto assume particolare rilevanza se si considera che la ricerca quantitativa non avanza esclusivamente attraverso l'introduzione di nuove metodologie, ma anche attraverso la loro

applicazione critica e verifica empirica in contesti differenti. Nel caso della Valle d'Aosta, l'impiego delle tecniche di *clustering* ha consentito di evidenziare come, al di là delle differenze esistenti tra alcuni comuni ad elevata attrattività turistica, il sistema turistico regionale presenti una struttura complessivamente piuttosto omogenea. Tale evidenza dimostra come un approccio di segmentazione *data-driven* possa contribuire non soltanto a individuare le differenze tra i territori, ma anche a mettere in evidenza gli elementi di omogeneità che caratterizzano un sistema turistico nel suo complesso.

Uno degli elementi di maggiore interesse emersi dall'analisi riguarda la possibilità di affiancare all'analisi descrittiva tradizionale una lettura integrata delle informazioni statistiche. Tale elemento costituisce probabilmente uno dei principali apporti conoscitivi del presente lavoro. L'analisi descrittiva rimane infatti uno strumento indispensabile per rappresentare l'andamento delle singole variabili, ma quando il fenomeno oggetto di studio è descritto da un elevato numero di indicatori, come nel caso del *dataset* analizzato nel presente elaborato, la loro interpretazione congiunta può risultare complessa. La segmentazione ottenuta attraverso il *clustering* consente invece di cogliere contemporaneamente le relazioni tra le diverse variabili considerate, sintetizzandole in una rappresentazione complessiva della struttura territoriale del fenomeno turistico.

In tale prospettiva, il *clustering* non sostituisce la statistica descrittiva tradizionale ma amplia le potenzialità interpretative, offrendo una lettura multidimensionale delle informazioni disponibili. In altre parole, il valore aggiunto del *clustering* non consiste esclusivamente nell'individuazione di gruppi omogenei, ma nella possibilità di trasformare una molteplicità di informazioni statistiche in una rappresentazione sintetica del sistema turistico, favorendone una lettura complessiva e maggiormente interpretabile.

Il presente studio mostra inoltre come le metodologie consolidate nella letteratura internazionale possano essere efficacemente applicate anche all'analisi di sistemi turistici territoriali di dimensioni più contenute, come quelle del caso oggetto di studio. Il caso della Valle d'Aosta evidenzia infatti come, anche in contesti territoriali relativamente contenuti, un approccio di segmentazione *data-driven* possa contribuire a mettere in luce configurazioni non immediatamente osservabili attraverso la sola analisi delle statistiche descrittive. A dimostrazione di quanto appena affermato, la *cluster analysis* condotta nel presente elaborato ha contribuito a mettere in evidenza dinamiche dei flussi turistici non immediatamente osservabili mediante il solo svolgimento dell'analisi descrittiva di base. Ciò dimostra come il

contributo del *clustering* risieda non soltanto nella costruzione della segmentazione, ma anche nella capacità di evidenziare relazioni strutturali tra le variabili considerate che difficilmente emergerebbero attraverso l'osservazione separata dei singoli indicatori.

Un esempio significativo è rappresentato dalla notevole stabilità della struttura generale della segmentazione ottenuta. Il fatto che differenti algoritmi di *clustering* e differenti configurazioni del numero di *cluster* abbiano individuato sistematicamente un ampio *macrocluster* principale, affiancato da un numero limitato di comuni caratterizzati da una marcata specializzazione turistica, suggerisce infatti che tale configurazione rappresenti una caratteristica strutturale del sistema turistico valdostano nel periodo oggetto di analisi e non il semplice risultato delle specifiche scelte metodologiche adottate nel presente lavoro. Pur senza attribuire a tale evidenza un valore definitivo, essa contribuisce significativamente a rafforzare l'affidabilità interpretativa della segmentazione proposta.

Complessivamente, tali elementi contribuiscono a rafforzare il valore conoscitivo del presente lavoro, evidenziando come l'impiego delle tecniche di *clustering* possa effettivamente costituire un valido strumento quantitativo complementare, e non sostitutivo, per l'analisi dei sistemi turistici territoriali. La ricerca conferma pertanto come un approccio di segmentazione *data-driven*, se applicato con rigore metodologico e interpretato criticamente alla luce del contesto territoriale e delle caratteristiche dei dati disponibili, possa contribuire significativamente alla comprensione della struttura dei sistemi turistici, offrendo una prospettiva complementare rispetto agli strumenti di analisi tradizionalmente impiegati.

Limiti della ricerca

La cornice interpretativa dei risultati ottenuti nel presente studio è delineata da alcuni limiti che lo caratterizzano e ne definiscono il corretto ambito di validità. Tali limiti non influenzano negativamente il valore delle evidenze empiriche emerse, ma rappresentano elementi essenziali per una loro corretta interpretazione. Come evidenziato nel corso dell'elaborato, le tecniche di *clustering* costituiscono infatti strumenti di analisi di natura esplorativa, il cui risultato dipende sia dalle caratteristiche dei dati disponibili sia dalle scelte metodologiche adottate durante il processo di ricerca.

Il primo limite riguarda l'orizzonte temporale oggetto di analisi. La segmentazione proposta rappresenta infatti una fotografia del sistema turistico valdostano riferita ai dati disponibili per il periodo compreso tra gennaio e settembre 2025. Come evidenziato anche dalla letteratura sulla segmentazione turistica, i profili individuati mediante approcci *data-driven* non devono

essere considerati come strutture univoche e definitive, ma come rappresentazioni riferite allo specifico momento in cui l'analisi viene condotta. L'evoluzione della domanda turistica e le trasformazioni del contesto economico, sociale e normativo possono infatti modificare nel tempo la composizione dei *cluster*, rendendo opportuno un costante e continuativo aggiornamento delle analisi.

Un secondo limite riguarda la natura delle informazioni impiegate. Pur rappresentando indicatori significativi dell'intensità della domanda e della distribuzione dei flussi tra le diverse tipologie ricettive, le variabili considerate non esauriscono la complessità del fenomeno turistico. Elementi quali le caratteristiche socioeconomiche dei territori, gli aspetti ambientali, la dotazione infrastrutturale, la spesa dei visitatori, la permanenza media e ulteriori indicatori potrebbero contribuire a fornire una rappresentazione ancora più articolata e rappresentativa delle differenze esistenti tra i comuni della Valle d'Aosta.

Infine, occorre ricordare che le segmentazioni ottenute mediante tecniche di *clustering* in senso lato non sono mai definitive. Tale considerazione si applica chiaramente anche al caso di studio dei comuni valdostani. I *cluster* individuati costituiscono infatti il risultato dell'applicazione di specifiche tecniche di *clustering* ai dati disponibili, e devono per tale ragione essere interpretati come una rappresentazione sintetica della struttura osservata nel periodo analizzato. La loro utilità risiede quindi soprattutto nella capacità di supportare l'analisi e l'interpretazione dei fenomeni territoriali, piuttosto che nell'attribuzione di classificazioni definitive ai territori.

Futuri sviluppi della ricerca

Le considerazioni emerse nel presente elaborato suggeriscono numerose possibili prospettive di approfondimento, sia sotto il profilo metodologico sia sotto quello applicativo. L'analisi sviluppata rappresenta infatti uno dei primi contributi all'impiego di tecniche di segmentazione *data-driven* nello studio del sistema turistico valdostano, inserendosi in un filone di ricerca che, con riferimento al sistema turistico valdostano risulta ancora poco esplorato e, proprio per la natura esplorativa dell'approccio adottato, apre la strada a ulteriori sviluppi di ricerca.

Un primo possibile sviluppo riguarda l'aggiornamento periodico della segmentazione proposta mediante l'integrazione dei dati che diverranno progressivamente disponibili. Come evidenziato dalla letteratura sulla segmentazione turistica, i *cluster* individuati attraverso approcci *data-driven* rappresentano una fotografia del fenomeno osservato nel momento in cui l'analisi viene condotta e possono quindi modificarsi nel tempo in conseguenza dell'evoluzione della domanda turistica o di cambiamenti di natura economica, sociale, normativa o ambientale.

L'aggiornamento periodico dell'analisi consentirebbe pertanto di monitorare nel tempo l'evoluzione della struttura del sistema turistico valdostano, verificando la stabilità dei profili individuati oppure l'eventuale comparsa di nuove configurazioni territoriali.

Un secondo sviluppo potrebbe consistere nell'ampliamento della base informativa utilizzata per la segmentazione. L'integrazione di ulteriori variabili, quali indicatori socioeconomici, infrastrutturali, ambientali, dati relativi alla spesa dei visitatori, alla permanenza media o ad altri aspetti che caratterizzano il fenomeno turistico, potrebbe contribuire a una rappresentazione ancora più articolata delle differenze esistenti tra i comuni, consentendo l'individuazione di profili territoriali maggiormente dettagliati e una più approfondita comprensione delle dinamiche turistiche regionali.

Dal punto di vista metodologico, un ulteriore sviluppo della ricerca potrebbe riguardare il confronto tra differenti tecniche di apprendimento non supervisionato (*unsupervised learning*). Pur avendo prodotto risultati coerenti con l'obiettivo della ricerca attraverso l'impiego del metodo gerarchico di *Ward* e dell'algoritmo *K-means*, sarebbe interessante valutare il comportamento di ulteriori metodologie di *clustering,* quali il *fuzzy clustering*, che consente a una stessa unità territoriale di appartenere contemporaneamente a più *cluster* con differenti gradi di appartenenza. In questo senso, il confronto tra differenti approcci potrebbe contribuire a valutare ulteriormente la robustezza della segmentazione proposta e ad approfondire la comprensione della struttura dei dati.

L'impiego delle tecniche di *clustering* nel presente elaborato rientra nell'ambito dell'apprendimento non supervisionato, nel quale l'obiettivo consiste nell'individuazione di strutture latenti direttamente a partire dai dati disponibili. Una naturale evoluzione del presente lavoro potrebbe pertanto consistere nell'integrare tale approccio con metodologie di apprendimento supervisionato (*supervised learning*), orientate non più alla segmentazione dei territori, bensì alla previsione dei flussi turistici.

La previsione della domanda turistica rappresenta infatti uno degli ambiti di maggiore interesse nell'ambito della ricerca quantitativa applicata al turismo, ma costituisce allo stesso tempo una delle problematiche metodologicamente più complesse. L'evoluzione dei flussi turistici dipende infatti da una molteplicità di fattori economici, sociali, geopolitici e climatici, spesso caratterizzati da relazioni non lineari e da una significativa variabilità nel tempo. La disponibilità di serie storiche sufficientemente estese o di un insieme più ampio di variabili esplicative potrebbe consentire l'applicazione di modelli di *Machine Learning* finalizzati alla

previsione della domanda turistica, integrando così l'approccio descrittivo ed esplorativo sviluppato nel presente lavoro con un approccio di natura predittiva.

Il crescente interesse della ricerca verso l'impiego di metodologie quantitative avanzate e strumenti di *Machine Learning* applicati ai fenomeni turistici, suggerisce come l'integrazione tra tecniche di segmentazione *data-driven* e modelli predittivi possa rappresentare uno dei principali ambiti di sviluppo futuro.

Complessivamente, tali prospettive evidenziano come la segmentazione proposta nel presente studio non rappresenti un punto di arrivo, bensì un possibile punto di partenza per ulteriori approfondimenti metodologici ed empirici. L'integrazione di nuovi dati, l'impiego di metodologie differenti e lo sviluppo di modelli previsionali potranno infatti contribuire ad ampliare ulteriormente la comprensione delle dinamiche turistiche regionali e a rafforzare il ruolo degli strumenti quantitativi a supporto dell'analisi e della pianificazione dei sistemi turistici.

Considerazioni finali

La crescente disponibilità di dati statistici territoriali e il continuo sviluppo delle metodologie quantitative stanno progressivamente ampliando la possibilità di analisi dei sistemi turistici. In questo contesto, tecniche come il *clustering* rappresentano strumenti in grado di affiancare gli approcci statistici tradizionali, consentendo di esplorare strutture e relazioni che difficilmente emergono dall'osservazione separata delle singole variabili.

Il presente elaborato si inserisce in questa prospettiva, proponendo l'applicazione di un approccio di segmentazione *data-driven* al caso dei comuni della Valle d'Aosta. Più che individuare una classificazione definitiva dei territori, il presente lavoro intende mostrare come un percorso metodologico rigoroso, fondato sulla trasparenza delle scelte analitiche e sull'interpretazione critica dei risultati, possa contribuire ad ampliare gli strumenti disponibili per lo studio dei sistemi turistici territoriali.

In questo senso, il principale messaggio che emerge dal presente lavoro non riguarda tanto la specifica segmentazione ottenuta, quanto il valore del metodo adottato. La possibilità di integrare grandi quantità di informazioni in profili territoriali interpretabili rappresenta infatti un'opportunità significativa per la ricerca quantitativa applicata al turismo e, più in generale, per lo sviluppo di analisi sempre più orientate ai dati, capaci di supportare in modo rigoroso la conoscenza dei fenomeni turistici territoriali.

“Essentially, all models are wrong, but some are useful.”

– George E.P. Box

RIFERIMENTI BIBLIOGRAFICI

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means *clustering* algorithm. Computers & geosciences, 10(2-3), 191-203.

Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, p. 424.

Data, M. C. (2016). Secondary *analysis* of electronic health records (Vol. 183). Cham: Springer International Publishing.

Dolnicar S., Grün B., Leisch F. (2018). *Market Segmentation Analysis*, Singapore, Springer.

Dolnicar, S. (2014). Market segmentation approaches in tourism. In The Routledge handbook of tourism marketing (pp. 197-208). Routledge.

Dolnicar, S. (2004). Beyond "commonsense segmentation": A systematics of segmentation approaches in tourism. Journal of travel research, 42(3), 244-250.

Dolnicar, S., & Leisch, F. (2003). Winter tourist segments in Austria: Identifying stable vacation styles using bagged *clustering* techniques. Journal of Travel Research, 41(3), 281-292.

Dolnicar, S. (2002). A review of *data-driven* market segmentation in tourism. Journal of travel & tourism marketing, 12(1), 1-22.

E Silva, F. B., Barranco, R., Proietti, P., Pigaiani, C., & Lavalle, C. (2021). A new European regional tourism typology based on hotel location patterns and geographical criteria. Annals of Tourism research, 89, 103077.

Everitt B. S., Landau S., Leese M., Stahl D. (2011). *Cluster Analysis* (5th ed.). Wiley.

Glyptou, K., Kalogeras, N., Skuras, D., & Spilanis, I. (2022). *Clustering* Sustainable Destinations: Empirical Evidence from Selected Mediterranean Countries. Sustainability 2022, 14, 5507.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York, NY: Springer.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism management, 68, 301-323.

MacQueen, J. (1967). Some methods for classification and *analysis* of multivariate observations. In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press.

Park, D. B., & Yoon, Y. S. (2009). Segmentation by motivation in rural tourism: A Korean case study. Tourism management, 30(1), 99-108.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Simon, H. A. (1955). *A Behavioral Model of Rational Choice*. Quarterly Journal of Economics, 69(1), 99–118.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of marketing, 21(1), 3-8.

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases*. Science, 185(4157), 1124–1131.

Ward JH (1963). Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244

RIFERIMENTI SITOGRAFICI

Meta analisi statistica: cos'è e come si fa
https://www.analisi-statistiche.it/meta-analisi/
consultato il 03/07/2026

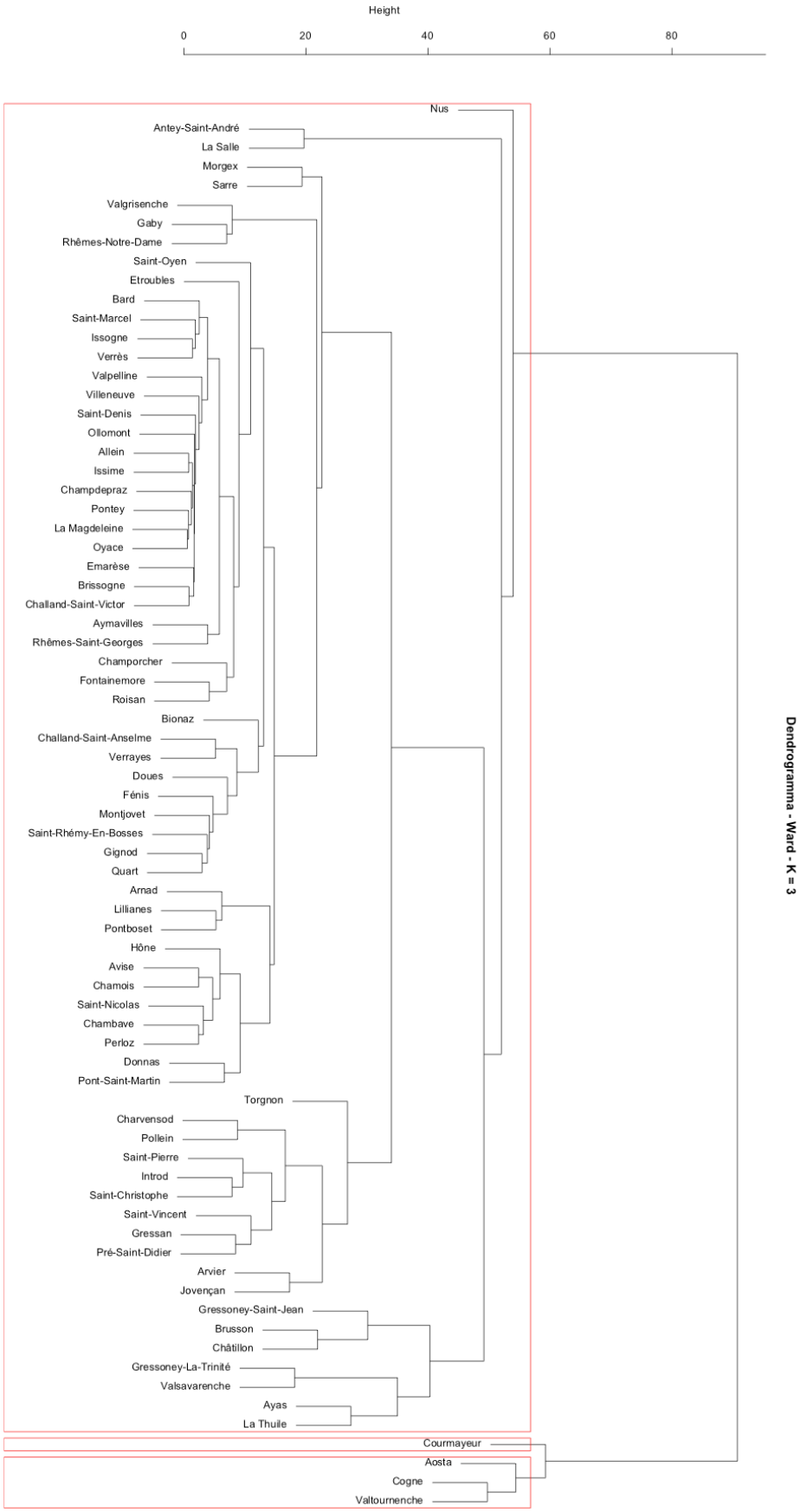
Eurostat – NUTS classification
https://ec.europa.eu/eurostat/web/nuts

consultato il 03/07/2026

A. APPENDICE

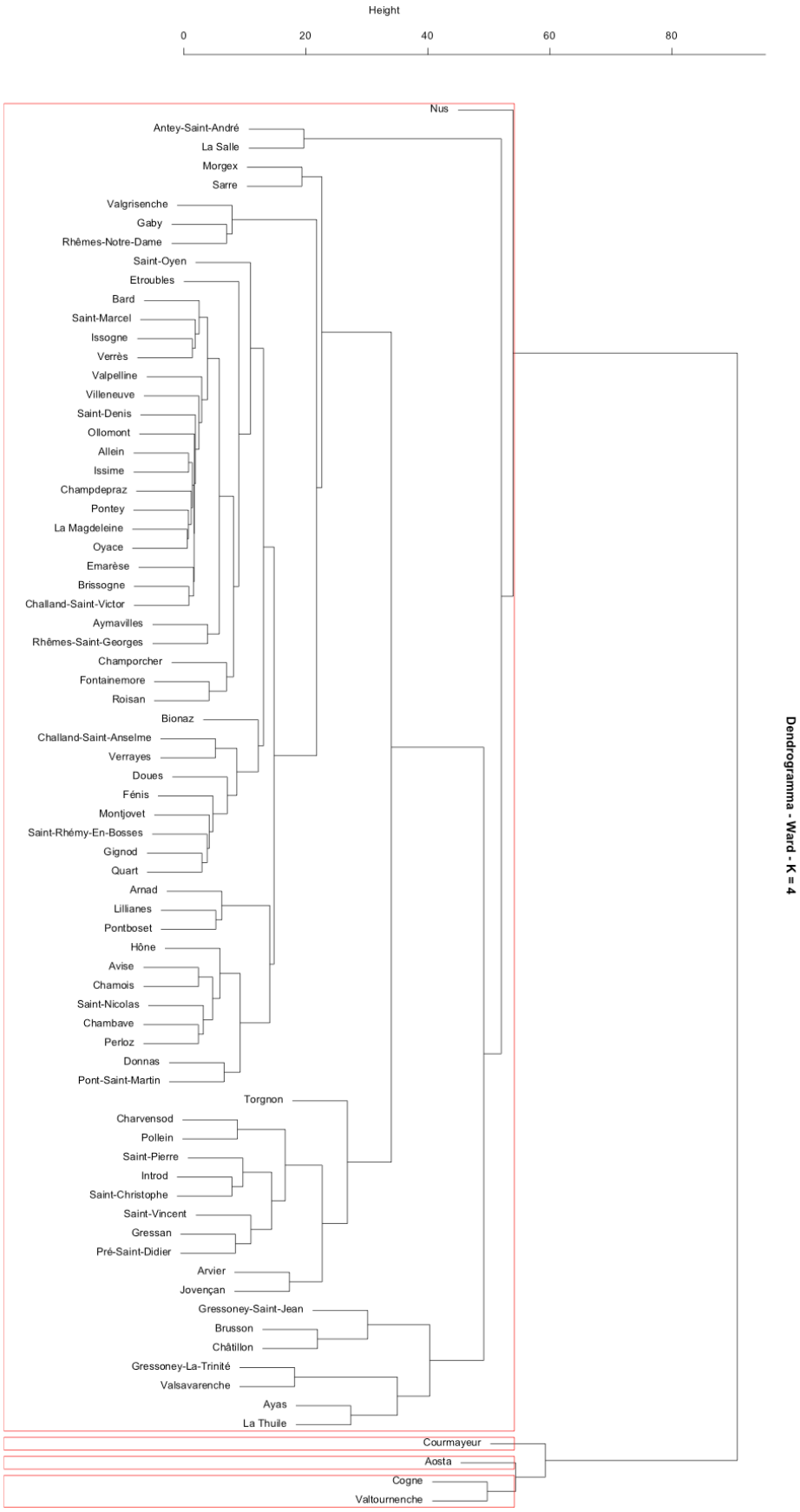
Lo script R utilizzato per lo sviluppo di rappresentazioni grafiche a supporto del capitolo metodologico e dell'intera analisi empirica non è riportato integralmente nel presente elaborato in ragione della sua estensione, ma è disponibile su richiesta presso l'autore.

Figura A.1 Dendrogramma – Metodo di Ward k=3 (clusters)



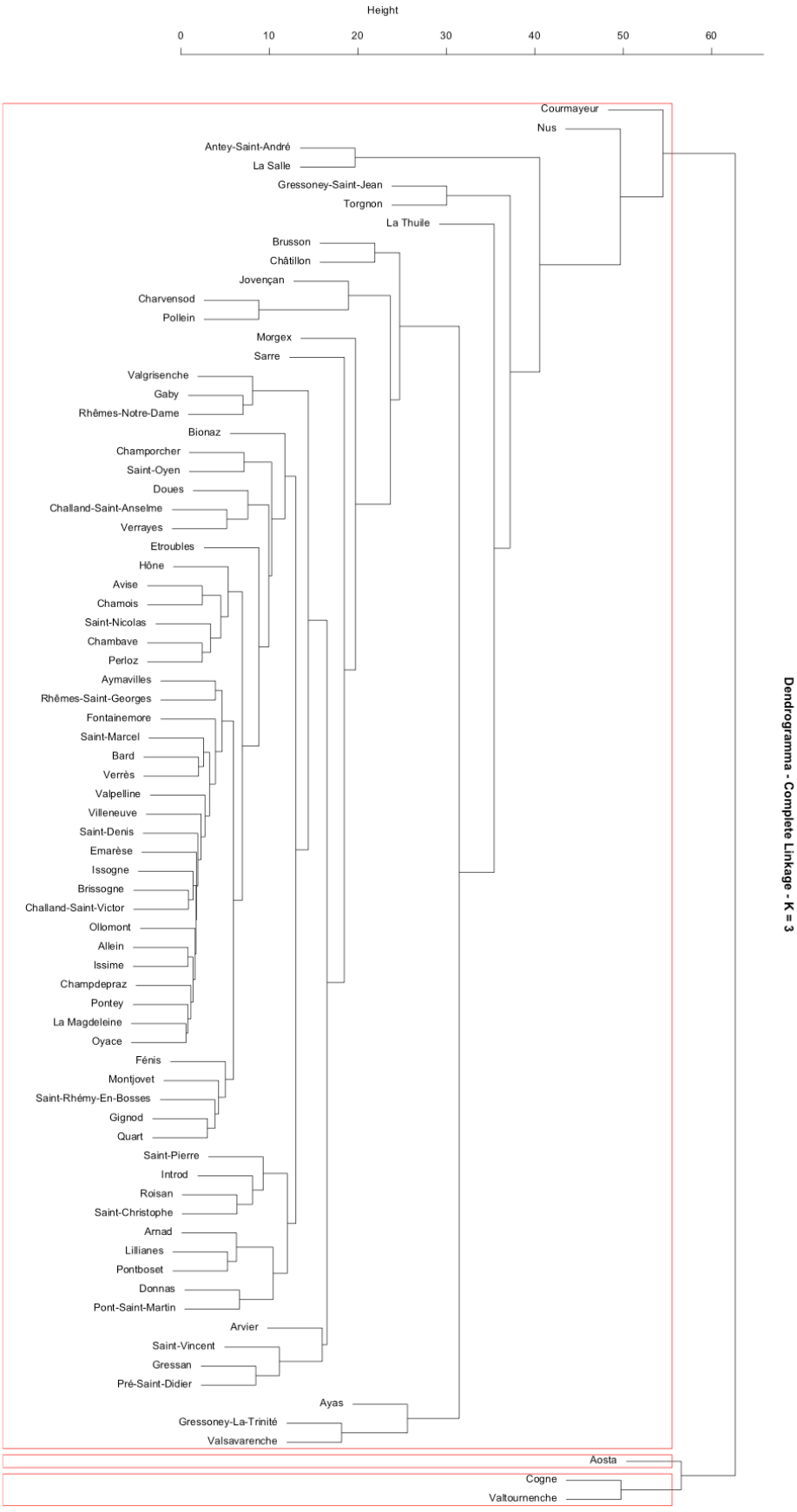
Fonte: Elaborazione personale dell'autore

Figura A.2 Dendrogramma – Metodo di Ward k=4 (clusters)



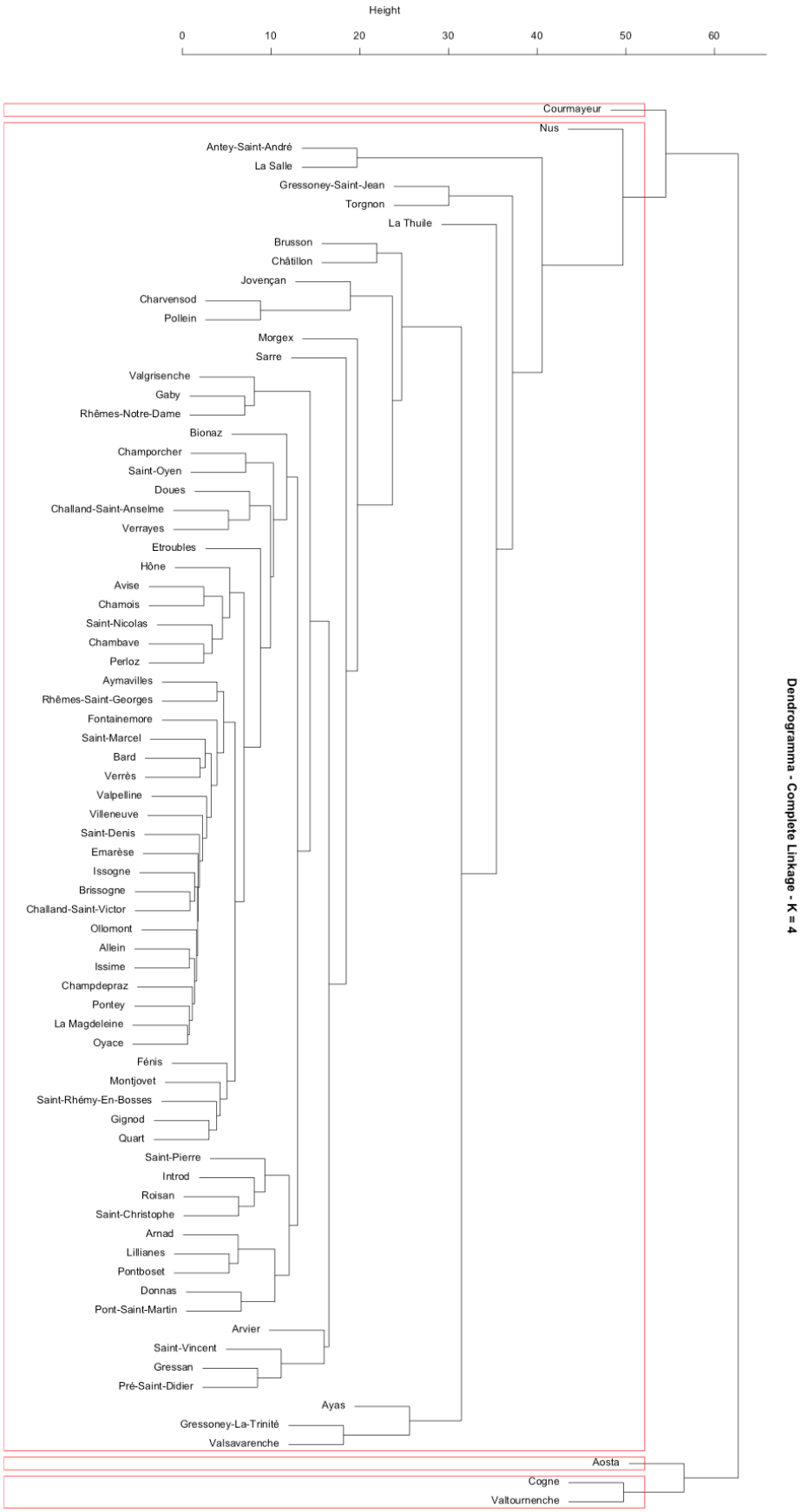
Fonte: Elaborazione personale dell'autore

Figura A.3 Dendrogramma – Complete Linkage k=3 (clusters)



Fonte: Elaborazione personale dell'autore

Figura A.4 Dendrogramma – Complete Linkage k=4 (clusters)



Fonte: Elaborazione personale dell'autore