



Dipartimento di Impresa e Management

Cattedra di Statistica

Analisi dell'attrattività turistica  
mediante la cluster analysis

RELATORE

Prof.ssa Livia De Giovanni

CANDIDATO

Nicola Violante matr. 183091

ANNO ACCADEMICO 2016/2017

## SOMMARIO

Introduzione .....	1
Capitolo I: Il Turismo .....	3
Cenni storici.....	3
L'industria turistica.....	4
Importanza del fenomeno .....	8
Influenza del turismo sull'economia .....	9
Conclusioni.....	13
Capitolo II: Cluster analysis.....	15
Definizione di distanza .....	15
Alcuni tipi di distanza.....	16
Misure di similarità .....	18
Raggruppamento delle unità.....	21
Metodi di classificazione.....	21
Metodi gerarchici agglomerativi.....	23
Metodi gerarchici divisivi.....	27
Metodi non gerarchici .....	27
Capitolo III: Travel & Tourism Competitiveness Report.....	30
T&T Competitiveness Report.....	30
TTCI.....	31
Capitolo 4: La segmentazione dei paesi con la cluster analysis .....	39
Metodo .....	39
Analisi.....	40
Capitolo 5: La cluster analysis con "R" .....	51
Introduzione.....	51
L'utilizzo di R per la cluster analysis.....	52
Bibliografia.....	53

Il presente elaborato mira a dimostrare l'efficacia della cluster analysis applicata a scelte e valutazioni di carattere economico.

Di solito, tale tecnica statistica, è utilizzata in campo economico, per la segmentazione di gruppi d'aziende, in base al loro fatturato, e di gruppi di clienti, in base alle loro scelte di consumo.

In questo lavoro, invece, si effettua un'applicazione della cluster analysis al caso di 141 Paesi in base alla loro attrattività turistica.

Nel primo capitolo viene presentato il settore turistico, descrivendone l'industria e soffermandosi sull'importanza del fenomeno e sulle dinamiche che influenzano l'economia in generale.

Nel secondo capitolo viene descritta la cluster analysis, partendo con delle indispensabili definizioni delle misure di distanza e similarità. In seguito vengono descritti i metodi di classificazione, iniziando con quelli "gerarchici", presentandone le varie tipologie di legami, e terminando con quelli "non gerarchici", con una breve descrizione dell'algoritmo k-means.

Nel terzo capitolo viene presentata la fonte dei dati che ha reso possibile tale tipo di lavoro, il "Travel & Tourism Competitiveness Report". Si

inizia con una breve descrizione del report per poi elencare nel dettaglio le variabili che hanno dato corpo all'indice vero e proprio. In fine la descrizione del metodo utilizzato, durante la costruzione dell'indice, per far assumere ai punteggi delle diverse variabili un valore compreso tra 1 e 7.

Nel quarto capitolo viene dapprima descritto il metodo utilizzato per l'analisi dei dati, poi presentata, passo dopo passo, l'analisi vera e propria. Vengono illustrate le composizioni dei vari gruppi emersi e ne viene fornita una descrizione per ognuno.

L'ultimo capitolo dell'elaborato descrive la piattaforma utilizzata per l'analisi dei dati (R) e le varie funzioni utilizzate.

## IL TURISMO

Nell'era della globalizzazione, il turismo è diventato uno stile di vita, quasi una parte della nostra vita quotidiana. In tutto il mondo, il volume del fenomeno è in fase di lancio, e con la nascita delle economie emergenti sembra plausibile che sia il turismo domestico che quello internazionale continueranno ad essere uno dei settori industriali più dinamici.

- **Cenni storici**

Facendo un passo indietro nel tempo è facile notare che il turismo, così come lo conosciamo, è un fenomeno del tutto nuovo. Nel XV secolo iniziarono i pellegrinaggi medioevali e le esplorazioni, che vedevano come mete principali Roma, Gerusalemme e Santiago de Compostela. Verso queste tre località si snodava una rete di santuari, ostelli, ospedali e conventi che assistevano i viandanti ed i pellegrini. L'epoca dei grandi viaggi iniziò con la prima rivoluzione industriale, la popolazione si spostò dalle campagne alle città per trovare lavoro. Molte volte gli abitanti lavoravano in un posto diverso rispetto a quello in cui vivevano

e iniziò il pendolarismo, da qui nacque la necessità di servizi pubblici di trasporto più efficienti. Thomas Cook, un imprenditore inglese noto per aver fondato la prima agenzia di viaggio, allora, colse questa necessità dei mezzi di trasporto per sfruttarli non solo a scopo lavorativo ma anche a scopo di svago. Un tipo di viaggiatore più simile al turista di oggi era il giovane che, a partire dal XVII secolo, si impegnava nel Grand Tour, pratica di educazione all' "uso del mondo" sempre più frequente per i figli della borghesia ricca, soprattutto inglese. Ma si trattava ancora di viaggiatori d'élite. In seguito con la moltiplicazione dei mezzi di trasporto, la maggiore facilità di accesso alle informazioni, grazie alla nascita dei mass media, e all'incremento dei redditi nel mondo occidentale, molte più persone hanno acquisito la capacità di viaggiare ed è nato il fenomeno turistico vero e proprio.

- **L'industria turistica**

Al giorno d'oggi, l'industria turistica è un sistema, locale e globale, complesso e sfaccettato, con molti attori e luoghi coinvolti. Per sistema turistico si intende il complesso dei fattori che, direttamente o indirettamente, in misura prevalente o accessoria, compongono, attengono, influenzano il fenomeno turistico.

Tali fattori possono essere riuniti in tre gruppi:

### 1. Strutture

### 2. Infrastrutture

### 3. Interdipendenze:

- la situazione politica
- il clima
- il sistema sociale (tempo libero, reddito, organizzazione sociale).

Con il termine **strutture** si intendono gli impianti costruiti per il turismo e destinati in via esclusiva o principale ad esso. Spesso si definiscono strutture turistiche solamente le opere destinate alla ricettività primaria, ma oggi il concetto di turismo coincide con il concetto di “vacanza”, cioè il complesso dei servizi, delle motivazioni, delle attese che spingono il turista a muoversi e a scegliere in un particolare momento uno specifico complesso di servizi. Le principali componenti strutturali sono tre: le imprese ricettive, le agenzie di viaggio e i tour operator e i servizi complementari. L'attività delle prime consiste nel produrre e scambiare servizi di ospitalità dietro pagamento di un

determinato prezzo. Le seconde costituiscono l'attività d'intermediazione turistica, che oggi va oltre il semplice ruolo di favorire l'incontro tra la domanda e l'offerta per svolgere soprattutto un'importantissima funzione di ricerca, sperimentazione, creazione dei prodotti. Tour operator ed agenzie dettaglianti rappresentano il fulcro e determinano, in buona parte, l'orientamento del mercato turistico.

Per servizi complementari si intendono la ristorazione, lo spettacolo, l'animazione ( intesa come servizio appositamente previsto ma anche come espressione caratteristica dei luoghi ), le attività culturali, etnografiche, paesaggistiche. Questi sono servizi importantissimi e spesso indispensabili, ma il più delle volte comuni anche ad altri settori.

Le **infrastrutture** sono tutte quelle opere utili e spesso indispensabili al turismo ma che vengono condivise con altri settori dell'organizzazione economico-sociale o che costituiscono un complemento alla vacanza.

Le **interdipendenze** rappresentano la situazione politica, il clima e il sistema sociale presenti sul territorio.

Oggi l'industria del turismo costituisce forse l'attività più dipendente dalla politica. È merce di scambio fra le nazioni e costituisce per i cosiddetti Paesi emergenti il più facile e più sicuro mezzo per ottenere

valuta pregiata.

Stabilità politica e sicurezza interna sono presupposti fondamentali per un normale e continuo flusso di turismo: anche il solo dubbio relativo alla stabilità di un Paese o di un regime è infatti in grado di distruggere la più forte ed efficiente organizzazione turistica.

L'importanza relativa del fattore clima nel turismo si è oggi molto ridimensionata.

È molto forte per certi tipi di turismo, soprattutto quelli detti appunto “climatici”, mentre è solo “relativa” e comunque non condizionante per altri, quali per esempio il turismo d'arte, il turismo congressuale, il turismo di studio, spesso anche il turismo sportivo. Il clima è solo una variabile alla segmentazione della domanda.

È quasi ovvio, poi, affermare che senza tempo libero non può esserci turismo. Gli aspetti però che caratterizzano maggiormente il turismo oggi sono di tipo sociale: la qualità della vita, l'edonismo, la ricerca del benessere e di un ambiente aperto alla ricezione dei messaggi.

Il reddito è chiaramente uno dei fattori maggiormente considerati nella fase di impostazione di qualsiasi discorso sul turismo, infatti esso è una variabile fondamentale per il manifestarsi del fenomeno.

L'ultimo fattore infine che influenza il settore turistico è l'organizzazione sociale, intesa sia come apparato economico-assistenziale, sia come influenza della società e dei suoi valori civili e religiosi sulla famiglia, gli usi, le abitudini, le regole, la morale, con i conseguenti condizionamenti, per esempio verso la donna o sul grado di libertà dei giovani.

- **Importanza del fenomeno**

Il turismo rispecchia la diversità negli stili di vita in una società moderna; alcuni turisti vogliono solo rilassarsi, mentre altri vogliono divertirsi e godere di un'esperienza senza precedenti. In tutti i casi, si può dire di assistere ad una nuova tendenza nel settore: il turista evoluto, ben preparato e ben informato, vuole ricevere valore in cambio di quello che paga.

Il turismo è un valore aggiunto importante all'economia di una città, di una regione e quindi di tutto un Paese. Di conseguenza lo sviluppo di questo settore può essere determinante al miglioramento delle condizioni esistenti in un luogo.

Non esiste Paese al mondo che non abbia un potenziale turistico siccome vi sono tanti motivi per fare turismo e tanti modi per pubblicizzarlo. Dal turismo balneare a quello montano, da quello religioso a quello culturale, il turismo scolastico, congressuale o enogastronomico. Le forme di turismo, specialmente nell'era moderna, sono pressochè infinite, così come lo sono le forme per promuoverlo. È quindi chiaro che lo sviluppo del turismo è praticamente possibile ovunque.

- **Influenza del turismo sull'economia**

Partendo da un punto di vista economico il turismo gioca un ruolo importante nel contribuire alla crescita economica, creare posti di lavoro, aumentare la produttività e il reddito.

Per quanto riguarda la capacità di generare posti di lavoro il settore turistico ne possiede una spiccata propensione. Secondo quanto riportato dal WTTC il settore ha generato 107.833.000 posti di lavoro diretti nel 2015 (il 3,6% dell'occupazione totale) e ci si aspetta un

incremento dell'1,9% nel 2016 che porterebbe a 109.864.000 posti di lavoro, e con un incremento stimato del 2,1% annuo a 135.884.000 nel 2026. Se però consideriamo l'impatto sull'occupazione derivante anche dagli effetti degli investimenti, dalla catena di offerta e dall'impatto del reddito indotto, il dato diventa ancora più interessante. In questo caso si tratta di 283.578.000 posti nel 2015, e si stimano 370.204.000 posti entro il 2026 (11% dell'occupazione totale).

Interessante si presenta poi la dinamica della natura dell'occupazione che vede impegnati nel settore turistico, più che negli altri settori, le donne e i giovani, riguardo le prime inoltre la percentuale è destinata a crescere nei prossimi anni, la seconda categoria invece subirà un leggero decremento dovuto all'aumento stimato dell'età occupazionale generale.

Altro beneficio che merita attenzione è quello della “diversificazione” che consiste nella presenza di più beni e/o servizi offerti.

Come già detto, le vie per attrarre turisti al giorno d'oggi sono pressochè infinite, per cui qualsiasi Paese al mondo ha la possibilità di aggiungere al proprio corredo di attività economiche quella turistica e può addirittura diversificare tra le attrazioni data la loro varietà. La diversificazione gioca un ruolo fondamentale per la stabilità

economica di qualsiasi sistema economico siccome può garantire entrate per la bilancia di un Paese anche in un momento di recessione negli altri settori.

L'impatto del turismo produce una varietà di effetti. I turisti contribuiscono alle vendite, ai profitti, ai posti di lavoro, alle entrate fiscali e al reddito in una certa area.

Altre conseguenze, meno dirette, della presenza di turisti sono:

- Cambiamenti nei prezzi: i turisti contribuiscono a gonfiare il costo degli alloggi e dei prezzi di vendita al dettaglio nella zona, spesso su base stagionale.

- Cambiamenti nella qualità e nella quantità dei beni e dei servizi: il turismo può portare a una gamma più ampia di beni e servizi disponibili in un area (di qualità superiore o inferiore che in assenza di turismo).

- Variazioni nelle imposte di proprietà e altre imposte: le imposte per coprire il costo dei servizi locali possono essere più o meno onerose in presenza di attività turistica.

In alcuni casi, le imposte riscosse direttamente o indirettamente dai turisti possono produrre una riduzione delle imposte locali per le scuole, strade, ecc. In altri casi, i locali possono essere tassati più pesantemente per coprire i costi delle infrastrutture e dei servizi aggiunti.

- Dimensioni economiche degli impatti sociali e ambientali: ci sono

delle conseguenze economiche anche ai maggiori impatti sociali e ambientali. Questi possono essere positivi o negativi. Per esempio la formazione del traffico stradale aumenterà i costi di trasporto sia per le famiglie che per le imprese, oppure il miglioramento nei servizi e quindi la maggiore attrattività turistica, possono incoraggiare il business di pensionati e imprese individuali della zona.

Importante variabile economica è poi la spesa turistica il cui effetto si propaga dal settore turistico al resto dell'economia.

L'impatto della spesa turistica produce tre tipologie di effetti: diretti, indiretti e indotti.

Gli effetti diretti derivano dalla spesa per consumi direttamente effettuata dai turisti, che ha l'effetto, per esempio, di aumentare le entrate per una struttura ricettiva e di conseguenza il suo reddito e il salario dei suoi dipendenti.

Gli effetti indiretti si riferiscono al ciclo di spesa dei proventi derivanti dalla spesa turistica diretta e quindi dal reddito prodotto dal settore turistico e impiegato in altri settori strettamente collegati ad esso. In definitiva consistono nei cambiamenti di produzione nei settori che servono in modo diretto le strutture e le imprese del settore turistico.

In fine gli effetti indotti rappresentano i cambiamenti nell'attività economica generale conseguente sia ai maggiori proventi e salari

prodotti dal settore turistico che ai cambiamenti di reddito e salari realizzati dai settori collegati ad esso.

La somma di questi tre effetti, più la spesa statale per il settore, costituiscono la domanda finale di beni e servizi turistici. In quasi tutti i casi i consumatori finali sono costituiti dalle famiglie.

- **Conclusioni**

Si può concludere con una considerazione sullo sfruttamento del settore da parte dell'Italia. Solitamente chi pensa all'Italia pensa a una storia lunga migliaia di anni, monumenti meravigliosi che tutto il mondo ci invidia e alla famosa “ospitalità italiana”.

Tuttavia, nonostante il nostro Paese abbia un eccellente potenziale turistico, una recente indagine condotta da Confturismo, ha evidenziato come proprio il settore turistico sia un tesoro mal sfruttato nella nostra penisola.

Secondo quanto emerso dall'analisi, pur essendo aumentati, i turisti spendono meno di quanto facciano in Francia e in Spagna, causando una perdita sulle entrate che si stima essere di ben 14 miliardi di euro (pari a un punto percentuale del PIL).

La soluzione, ovviamente, è tutt'altro che semplice e richiede un importante impegno per modificare e migliorare l'offerta turistica italiana.

Oltre ad alcuni provvedimenti specifici come rilanciare l'attrattività del meridione (visitato soltanto dal 12% dei turisti), fornire un'adeguata governance per limitare la diatriba Stato-Regioni, aggiungere finanziamenti per il settore e prestare maggiore attenzione ai provvedimenti europei per migliorare la rappresentanza dell'Italia nell'Unione Europea, è importante cercare di incentivare forze sinergiche che realmente abbiano come obiettivo quello di prolungare il soggiorno dei turisti e di migliorare l'offerta.

Si tratta quindi di un importante sfida per il futuro del turismo in Italia, che andrebbe strutturato e sfruttato molto meglio di quanto non si faccia attualmente.

## CLUSTER ANALYSIS

Con il termine “cluster analysis” si intende un insieme di tecniche di analisi multivariate con l’obiettivo di effettuare raggruppamenti di unità statistiche mediante la similarità dei vari attributi.

In questo tipo di analisi due attributi vengono uniti in un cluster mediante la similarità tra gli stessi, ovvero la distanza degli altri attributi o cluster.

Prima di proseguire è necessario dare una definizione di distanza e similarità in termini statistici.

- **Definizione di distanza**

Il concetto di distanza in statistica è mutuato dalla geometria ove si fa riferimento alla distanza tra due punti.

Una misura di distanza deve godere delle seguenti proprietà:

1.  $d_{ij} \geq 0$  (non negatività)
2.  $d_{ii} = 0$  (identità)
3.  $d_{ij} = d_{ji}$  (simmetria)
4.  $d_{ij} \leq d_{ir} + d_{rj}$  (diseguaglianza triangolare)

Uno spazio con riferimento al quale si sia definita una misura di distanza che rispetta le quattro proprietà appena enunciate è detto spazio metrico.

- **Alcuni tipi di distanza**

### **Distanza Euclidea**

Questo è il tipo di distanza tra due punti più noto in assoluto, infatti nel caso di due sole variabili, la distanza euclidea è rappresentata dal segmento che unisce i due punti sul piano.

Siano  $X_i$  e  $X_j$  due vettori contenenti il profilo di due unità misurato su  $p$  attributi, la distanza euclidea è calcolata nel modo seguente:

$$d_{ij} = |x_i - x_j| = [(x_i - x_j)'(x_i - x_j)]^{1/2} = [\sum_{k=1}^p (x_{ik} - x_{jk})^2]^{1/2}$$

### **Distanza Euclidea ponderata**

Sia  $W$  una matrice diagonale contenente i coefficienti di ponderazione

$$W = \text{diag}(w_1, w_2, \dots, w_p).$$

$$d_{ij} = [(x_i - x_j)'w(x_i - x_j)]^{1/2} = [\sum_{k=1}^p (x_{ik} - x_{jk})^2 w_k]^{1/2}$$

## Distanza della città a blocchi

Essa corrisponde alla lunghezza che si deve percorrere per spostarsi da  $X_i$  a  $X_j$  qualora sia consentito muoversi solo nelle direzioni parallele agli assi.

$$d_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}|$$

I due tipi di distanza appena descritti possono ottenersi entrambi da una formula più generale:

## Distanza di Minkowski

Si dice distanza di Minkowski di ordine  $K$  tra le unità  $i$  e  $j$  l'espressione seguente:

$$d_{ij}^k = \left[ \sum_{s=1}^p |x_{is} - x_{js}|^k \right]^{1/k} \quad k \geq 1$$

Si ricava che la distanza euclidea può interpretarsi come la metrica di Minkowski per  $k=2$  e la distanza della città a blocchi per  $k=1$ .

## Confronto tra distanza euclidea e distanza della città a blocchi.

La distanza euclidea è influenzata più fortemente dalle differenze elevate tra i valori (poiché essa è in funzione del quadrato delle stesse), mentre la distanza della città a blocchi attua una compensazione, su un piano paritetico, tra differenze <<grandi>> e differenze <<piccole>>.

⇒ Consideriamo un esempio numerico:

CASO A:  $X_1 = (10,5)$  d città a blocchi = 4

$X_2 = (12,7)$  d euclidea =  $\sqrt{8}$

CASO B:  $X_1 = (10,5)$  d città a blocchi = 4

$X_2 = (11,8)$  d euclidea =  $\sqrt{10}$

In conclusione nella distanza della città a blocchi due differenze uguali a 2 equivalgono ad una differenza uguale a 1 e una uguale a 3. Questo non accade con la distanza euclidea, poiché, considerando i quadrati, le differenze più grandi incidono maggiormente e non vengono compensate dalle differenze più piccole.

- **Misure di similarità**

Una misura complessiva di similarità è data dall'indice di similarità di Gower:

$$C_{ij} = \frac{\sum_{k=1}^p c_{ij,k}}{\sum_{k=1}^p \delta_{ij,k}}$$

Dove “C” è una misura di similarità.

“ $\delta$ ” invece rappresenta una variabile nominale che assume valore 1 se le variabili sono confrontabili rispetto al k-esimo attributo, 0 altrimenti, denota quindi l’ammissibilità del confronto.

La definizione di tali indici varia a seconda della tipologia delle variabili.

### Variabili quantitative

$$C_{ij,k} = 1 - d_{ij,k} = 1 - \frac{|X_{ij} - X_{jk}|}{R_k}; \quad \delta_{ij,k} = 1$$

$R_k$  = campo di variazione della variabile k.

### Variabili qualitative politomiche

$C_{ij,k}$  assume valore 1 se le unità presentano la stessa modalità, 0 altrimenti.  $\delta_{ij,k} = 1$ , quindi il confronto è sempre ammissibile.

### Variabili qualitative dicotomiche

$$C_{ij,k} =$$

	$i = 1$	$i = 0$
$j = 1$	1	0
$j = 0$	0	0

La misura di similarità assume valore 1, se entrambe le variabili presentano la modalità 1. 0, nei restanti casi.

$$\delta_{ij,k} =$$

	$i = 1$	$i = 0$
$j = 1$	1	1
$j = 0$	1	0

Il confronto è ammesso in tutti i casi tranne quando entrambe le variabili assumono la modalità 0.

Come si può dedurre dalle formule l'importanza relativa di ciascuna variabile nella formazione dei grappoli è direttamente collegata alla varianza delle diverse variabili. Per questo hanno maggiore impatto sulla misura della distanza le variabili caratterizzate da un più elevato grado di

dispersione.

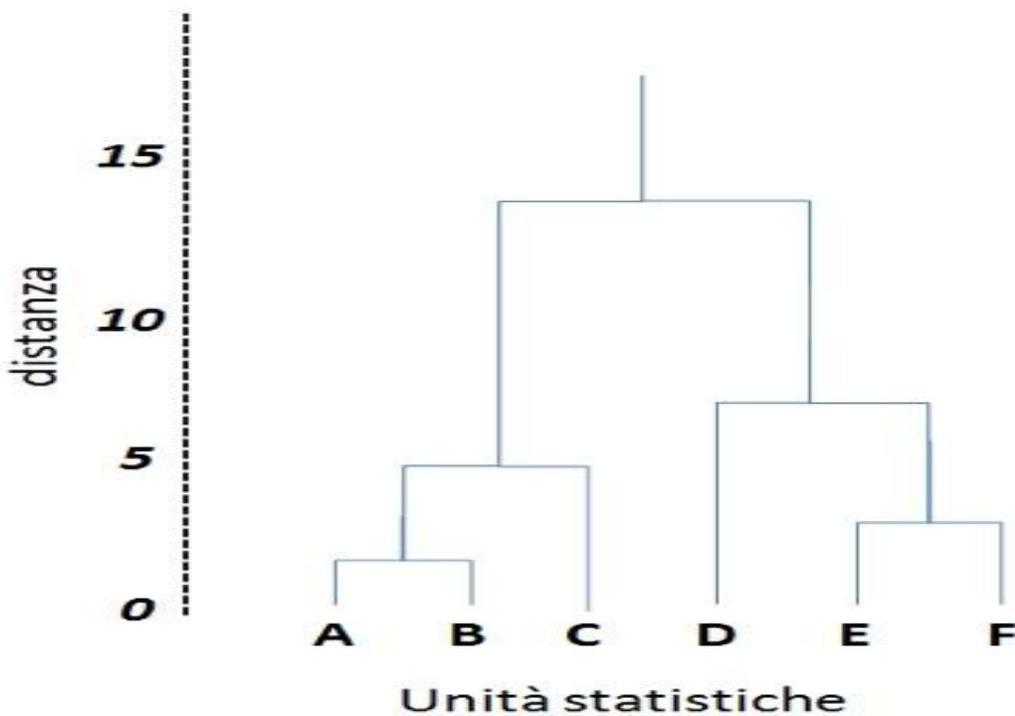
- **Raggruppamento delle unità**

Effettuata la scelta della misura di diversità da utilizzare, si pone la scelta del metodo o algoritmo di classificazione e dell'eventuale criterio di aggregazione/suddivisione.

- **Metodi di classificazione**

Una prima distinzione è quella tra i metodi gerarchici e non gerarchici; i primi seguono una sequenza ordinata di operazioni della stessa natura e la caratteristica principale che li distingue dai metodi non gerarchici è che l'assegnazione di un oggetto ad un cluster è irrevocabile; inoltre il risultato finale di un metodo gerarchico non è una singola partizione delle unità, ma una serie di partizioni nidificate che possono essere rappresentate graficamente attraverso un "dendogramma" nel quale sull'asse delle ordinate viene riportato il livello di distanza, mentre sull'asse delle ascisse vengono riportate le singole unità.

Ogni ramo del diagramma (linea verticale) corrisponde ad un grappolo (gruppo o cluster).



La linea di congiunzione (orizzontale) di due o più rami individua il livello di distanza al quale i grappoli si fondono.

Questi possono dividersi in:

**Agglomerativi:** procedono per aggregazioni successive delle unità partendo da  $n$  gruppi formati da un solo individuo.

**Divisivi:** partono da un solo gruppo formato da tutte le unità e procedono a partizioni successive fino a giungere a gruppi formati da una sola unità.

I secondi sono solo di tipo agglomerativo e producono una sola partizione. Procedono a riallocazioni successive delle unità tra i gruppi definite a priori, fino alla partizione giudicata “ottima” sulla base di un criterio predefinito.

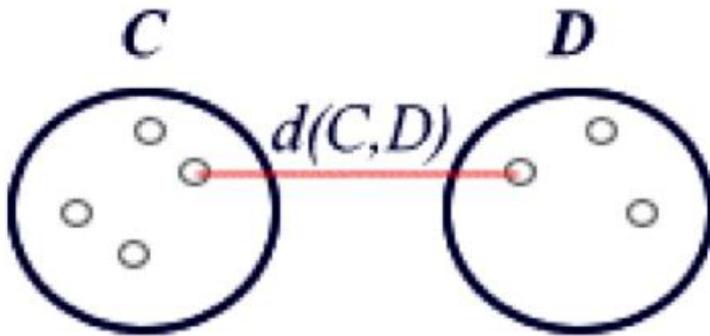
- **Metodi gerarchici agglomerativi**

Un algoritmo di cluster analysis di tipo agglomerativo può essere descritto mediante le seguenti fasi:

- a. Calcolo della matrice delle distanze tra tutti i punti;
- b. Analisi della matrice delle distanze e fusione delle sue unità aventi la distanza più piccola;
- c. Calcolo della nuova matrice delle distanze (utilizzando il metodo del legame prescelto);
- d. Si ritorna al punto b. ripetendo il processo fino a comprendere tutti gli elementi in un'unica classe.

### **Metodo del legame singolo**

Nel metodo del legame singolo, detto anche del “vicino più prossimo”, la distanza tra i gruppi è posta pari alla più piccola delle distanze esistenti tra gli elementi appartenenti ad un gruppo e quelli appartenenti ad un altro.



### Metodo del legame completo

Nel metodo del legame completo la distanza tra i gruppi è definita come la massima distanza esistente tra gli individui componenti.

Si avrà quindi:

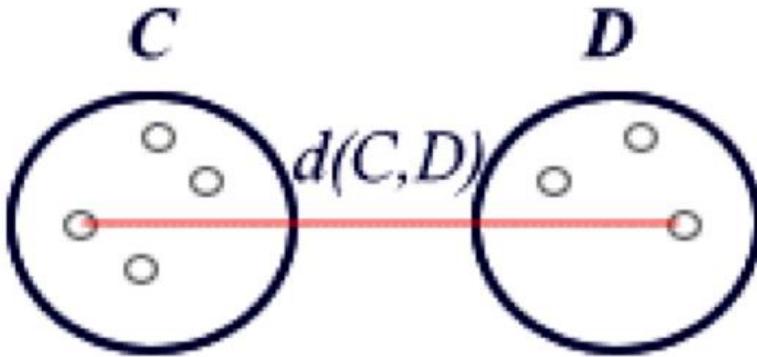
$$d(C,D) = \max(d_{ij})$$

Per ogni  $i$  appartenente a  $C$  e  $j$  appartenente a  $D$ .

Questo algoritmo di aggregazione evidenzia in maniera netta le differenze tra elementi: privilegia la differenza tra i gruppi piuttosto che l'omogeneità degli elementi di ogni gruppo.

Il dendrogramma costruito su questa matrice ha i rami molto più lunghi, i gruppi (e soprattutto i rami) si formano a distanze

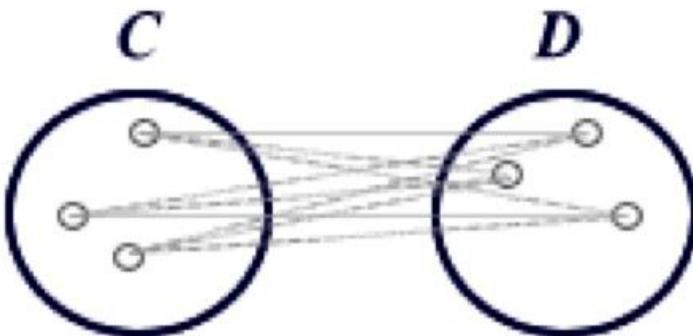
maggiori. In uno stesso range di valori, rispetto al legame singolo, gli elementi sono molto meno compatti e più diluiti.



### Metodo del legame medio

La distanza tra i gruppi è calcolata come media aritmetica semplice delle distanze tra tutte le unità che compongono i due gruppi.

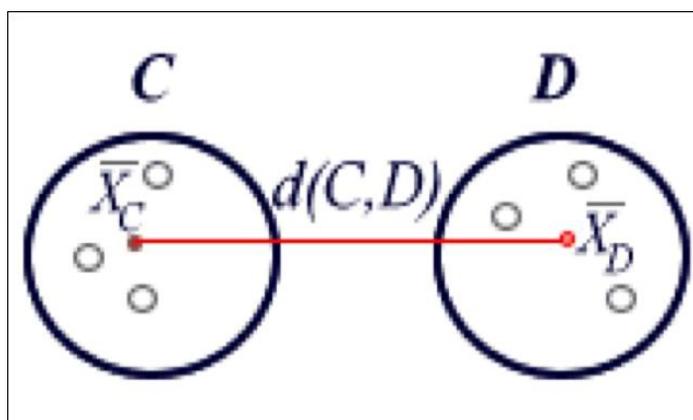
Essendo basato sulla media delle distanze, i risultati sono più attendibili e i gruppi risultano più omogenei e ben differenziati tra di loro.



## Metodo del centroide

La distanza tra i gruppi è determinata dalla distanza tra i rispettivi centroidi (o baricentri).

Vale a dire i valori medi calcolati sugli individui appartenenti ai gruppi.



## Metodo di Ward

Il metodo di Ward segue un approccio differente da quelli precedenti.

Secondo il metodo di Ward, ad ogni passo della costruzione agglomerativa del dendrogramma, si uniscono i gruppi dalla cui “fusione” deriva il minimo incremento possibile della devianza “entro”.

Quindi tra tutte le possibili coppie di cluster viene fusa la coppia per cui la varianza entro i gruppi risulta minima.

- **Metodi gerarchici divisivi**

Nei metodi gerarchici divisivi la configurazione iniziale prevede che tutte le unità siano raggruppate in un unico cluster. Nei metodi “nodali”, si procede alla scelta delle due unità più distanti chiamate nodi (o fuochi) e tutte le altre unità vengono allocate ai due gruppi in ragione della loro vicinanza rispetto ai nodi. Successivamente vengono suddivisi i due grappoli con lo stesso criterio e si continua il processo finchè ciascuna unità costituisce un gruppo a parte. L’algoritmo consiste di  $n-1$  divisioni successive, e può essere così descritto:

- a. Si individua una coppia di punti nodali, (punti che presentano distanza massima);
- b. Si attribuiscono le unità rimanenti ai due gruppi corrispondenti ai punti nodali, in base alla distanza minima da questi;
- c. Si iterano i passi precedenti finchè si avranno  $n$  gruppi.

- **Metodi non gerarchici**

Richiedono che il numero di cluster sia determinato a priori e generalmente forniscono una sola partizione finale. Ciò avviene attraverso una procedura iterativa che permette di ottenere, quale risultato finale, un'unica partizione (mentre nei metodi gerarchici si ottiene una gerarchia innestata di partizioni).

Sulla base delle scelte iniziali, l'algoritmo partiziona le unità in un numero predefinito di gruppi basandosi sulla ottimizzazione di un criterio (es. massimizzazione dell'omogeneità all'interno dei gruppi).

L'inizializzazione dell'algoritmo avviene indicando  $G$  centri di partenza intorno a cui aggregare le unità.

A differenza dei metodi gerarchici, l'assegnazione di un oggetto ad un cluster non è irrevocabile. Ovvero le unità vengono riassegnate ad un diverso cluster se l'allocazione iniziale risulta inappropriata. La configurazione finale e la velocità di convergenza dell'algoritmo dipendono dalla scelta del numero di cluster che si desidera ottenere, per cui, è consigliabile applicare l'algoritmo con valori iniziali diversi, e controllare la stabilità della soluzione ottenuta. Questa si rivelerà molto instabile quando la popolazione è omogenea e non ammette partizioni.

### **Algoritmo k-means**

L'algoritmo k-means permette di suddividere un insieme di oggetti in  $k$

gruppi sulla base dei loro attributi. Tale algoritmo si prepone di minimizzare la varianza intra-gruppo seguendo una procedura iterativa: inizialmente crea  $k$  partizioni e calcola il centroide di ogni gruppo, costruisce quindi una nuova partizione associando ogni oggetto al cluster il cui centroide è più vicino ad esso, quindi vengono calcolati i centroidi dei nuovi gruppi e così via, fino a quando l'algoritmo non converge.

Il pregio di questo algoritmo è la velocità con cui converge. Infatti, si è osservato che il numero di iterazioni è, di solito, minore del numero di punti.

In termini di qualità della soluzione, però, l'algoritmo non garantisce sempre il raggiungimento dell'ottimo globale. Inoltre, un altro svantaggio, è che esso richiede che siano scelti a priori il numero di cluster ( $k$ ) da trovare.



**TRAVEL & TOURISM  
COMPETITIVENESS INDEX**

- **Travel & Tourism competitiveness report**

Il “Travel & Tourism competitiveness Report” analizza le performance di 141 economie a livello globale attraverso l’indice di competitività (TTCI).

Esso mostra la natura e l’estensione dell’elasticità del settore turistico agli shock, così come mostra la criticità dello stesso nella creazione di posti di lavoro, nella crescita e nello sviluppo economico, sia nelle economie sviluppate che in quelle in via di sviluppo.

Pubblicato ogni due anni, il report, fornisce uno strumento strategico sia per il business che per il Governo dei vari Stati; offre infatti la possibilità di evidenziare, ai responsabili politici delle varie Nazioni, le sfide per la competitività del settore che richiedono l’attenzione della politica. È capace quindi di realizzare un dialogo, tra tutti coloro che hanno interesse nel settore, sulla formulazione delle azioni e delle politiche appropriate.

- **TTCI (indice)**

La struttura del TTCI è basata su 4 subindici, che si dividono in 14 pilastri, a loro volta derivati dalla media dei valori attribuiti a specifiche variabili, di seguito mostrati:

I. **CONTESTO AMBIENTALE:** cattura le impostazioni generali necessarie per operare in un paese.

pillar 1. **Contesto imprenditoriale**

1.1. Diritti di proprietà

1.2. Impatto delle norme in materia di investimenti diretti esteri

1.3. Efficienza delle leggi nella risoluzione delle controversie

1.4. Efficienza delle leggi nei regolamenti impegnativi

1.5. Tempo richiesto per permessi di costruzione

1.6. Costo dei permessi di costruzione

1.7. Misura della posizione dominante sul mercato

1.8. Tempo richiesto per intraprendere un business

1.9. Costo per intraprendere un business

1.10. Effetto e portata della tassazione sugli incentivi al lavoro

1.11. Effetto e portata della tassazione sugli incentivi all'investimento

1.12. Aliquota fiscale totale

## pillar 2. **Sicurezza e protezione**

- 2.1. Costi aziendali del crimine e della violenza
- 2.2. Affidabilità del servizio di polizia
- 2.3. Costi aziendali del terrorismo
- 2.4. Indice degli incidenti terroristici
- 2.5. Tasso di omicidi

## pillar 3. **Salute ed igiene**

- 3.1. Densità di medici
- 3.2. Accesso ai servizi igienici
- 3.3. Accesso all'acqua potabile
- 3.4. Letti ospedalieri
- 3.5. Prevalenza di HIV
- 3.6. Incidenza di malaria

## pillar 4. **Risorse umane e mercato del lavoro**

*(Qualifica della forza lavoro)*

- 4.1 Tasso d'iscrizione all'istruzione primaria
- 4.2 Tasso d'iscrizione all'istruzione secondaria
- 4.3 Entità della formazione del personale
- 4.4 Trattamento dei clienti

*(Mercato del lavoro)*

- 4.5 Pratiche di assunzione e licenziamento
- 4.6 Facilità reperimento dipendenti qualificati
- 4.7 Facilità assunzione manodopera straniera
- 4.8 Paga e produttività
- 4.9 Partecipazione femminile alla forza lavoro

pillar 5. **Dotazione tecnologie informatiche e di comunicazione**

- 5.1. Uso ICT per transazioni B2B
- 5.2. Utilizzo di internet per transazioni B2C
- 5.3. Individui che utilizzano internet
- 5.4. Abbonamenti internet a banda larga
- 5.5. Abbonamenti telefonia mobile
- 5.6. Abbonamenti alla banda larga mobile
- 5.7. Copertura di rete mobile
- 5.8. Qualità delle forniture di energia elettrica

**II. POLITICHE E CONDIZIONI FAVOREVOLI AL**

**SETTORE TURISTICO:** Cattura le specifiche politiche e aspetti strategici che impattano direttamente l'industria turistica.

## pillar 6. **Priorità del settore turistico**

- 6.1. Priorità del governo nel settore turistico
- 6.2. Spesa pubblica per il turismo
- 6.3. Efficacia del marketing per attrarre i turisti
- 6.4. Completezza dei dati annuali sul turismo
- 6.5. Tempestività di fornitura dati mensili/trimestrali
- 6.6. Valutazione della strategia di marca del paese

## pillar 7. **Apertura internazionale**

- 7.1. Obbligo di visto
- 7.2. Apertura agli accordi bilaterali sui servizi aerei
- 7.3. Numero di accordi commerciali regionali in vigore

## pillar 8. **Competitività del prezzo**

- 8.1. Tasse, biglietti e diritti aereoportuali
- 8.2. Indice dei prezzi alberghieri
- 8.3. Parità di potere d'acquisto
- 8.4. Livelli di prezzo del carburante

## pillar 9. **Sostenibilità ambientale**

- 9.1. Rigore dei regolamenti ambientali

- 9.2. Applicazione della regolamentazione ambientale
- 9.3. Sostenibilità dello sviluppo dell'industria turistica
- 9.4. Concentrazione di particolato
- 9.5. Numero di omologazioni dei trattati ambientali
- 9.6. Stress idrico di riferimento
- 9.7. Specie minacciate
- 9.8. Cambiamento nella copertura forestale
- 9.9. Trattamento delle acque reflue
- 9.10. Pressione della pesca costiera

**III. INFRASTRUTTURE:** Cattura la disponibilità e qualità delle infrastrutture fisiche di ciascuna economia

**pillar 10. Infrastrutture trasporto aereo**

- 10.1. Qualità delle infrastrutture di trasporto aereo
- 10.2. Posti disponibile per km, interni
- 10.3. Posti disponibili per km, internazionali
- 10.4. Partenze aeree
- 10.5. Densità aeroporti
- 10.6. Numero di compagnie aeree operanti

## pillar 11. **Infrastrutture terrestri e portuali**

11.1. Qualità delle strade

11.2. Qualità delle infrastrutture ferroviarie

11.3. Qualità delle infrastrutture portuali

11.4. Qualità della rete di trasporto a terra

11.5. Densità ferroviaria

11.6. Densità stradale

11.7. Densità strade asfaltate

## pillar 12. **Infrastrutture servizio turistico**

12.1. Camere

12.2. Estensione dei viaggi di lavoro consigliata

12.3. Presenza di grandi società di noleggio auto

12.4. Sportelli bancomat che accettano carte visa

## **IV. RISORSE NATURALI E CULTURALI**

### pillar 13. **Risorse naturali**

13.1. Numero di siti naturali (patrimonio mondiale)

13.2. Totale specie conosciute

13.3. Totale aree protette

13.4. Domanda digitale turismo naturale

13.5. Qualità dell'ambiente naturale

pillar 14. **Risorse culturali e viaggi d'affari**

14.1. Numero di luoghi d'interesse culturale (patrimonio mondiale)

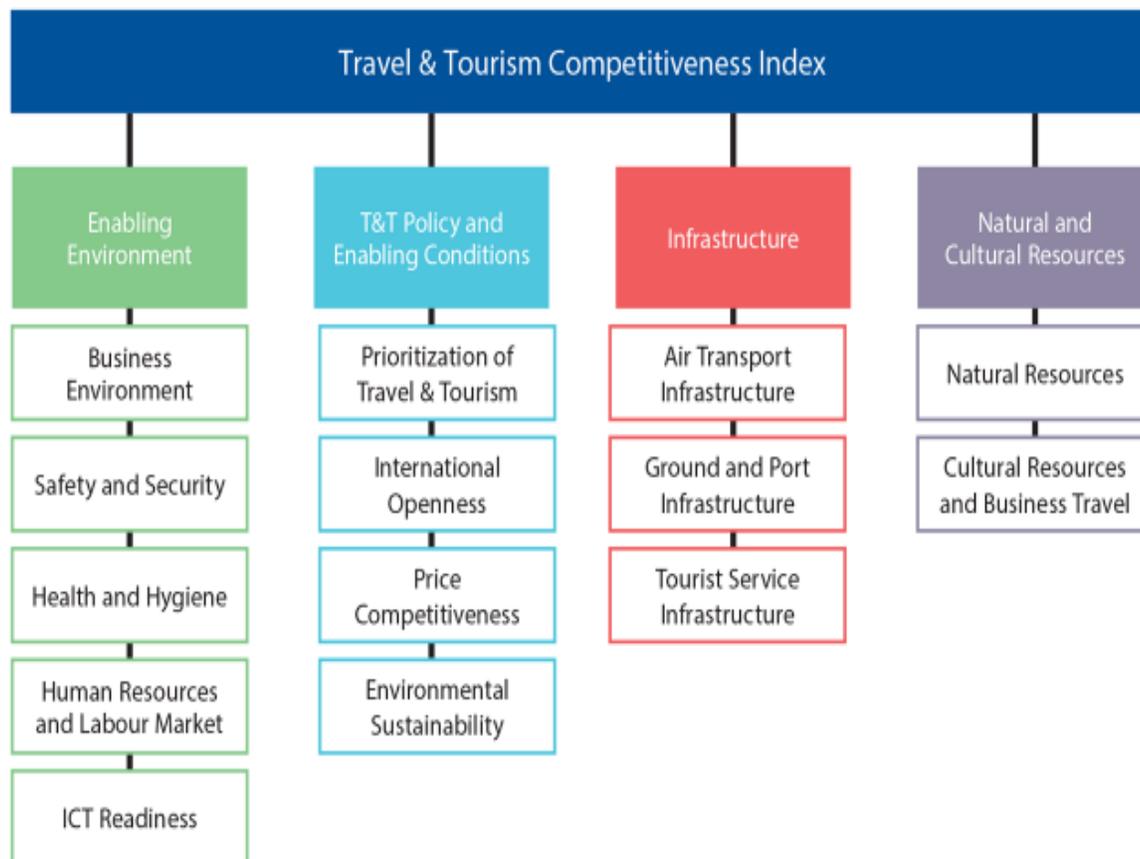
14.2. Numero di espressioni culturali orali e immateriali

14.3. Numero di stadi

14.4. Numero di meeting di associazioni internazionali

14.5. Domanda digitale turistica culturale e di intrattenimento

Figure 1: The T&T Competitiveness Index 2015 framework



Il risultato dei vari indicatori sono stati modificati affinché assumano un valore compreso tra 1 e 7, mediante la seguente formula:

$$6 * \left( \frac{\text{punteggio nazione} - \text{punteggio minimo}}{\text{punteggio massimo} - \text{punteggio minimo}} \right) + 1$$

Così facendo alla Nazione col punteggio maggiore corrisponderà un valore di 7, a quella col punteggio minore un valore di 1.



## **LA SEGMENTAZIONE DEI PAESI CON LA CLUSTER ANALYSIS**

In questa fase del lavoro viene presentata l'applicazione della cluster analysis ai paesi trattati nel "Travel & Tourism Competitiveness report"

- **Metodo**

L'obiettivo è di segmentare i Paesi in base al loro livello di attrattività turistica ed analizzare il rapporto tra i gruppi emersi dall'analisi e i continenti delle varie Nazioni.

Le variabili utilizzate in questo caso, per semplicità, sono:

- Contesto ambientale
- Politiche e condizioni favorevoli al settore turistico
- Infrastrutture
- Risorse naturali e culturali

Ma, per approfondire l'analisi, si potrebbero usare come variabili i 14 sottopilastrini illustrati in precedenza.

La misura di distanza e l'algoritmo che hanno generato i risultati migliori sono: la distanza euclidea e il metodo del legame medio.

La scelta della cluster analysis avviene quando non ci sono sufficienti informazioni preliminari. La prima operazione da compiere è la scelta del numero di cluster in cui conviene suddividere la popolazione, per ottenere con precisione questa importante informazione si procede nel modo seguente:

- Si avvia un processo di clustering gerarchico che mira a raggruppare l'intera popolazione in un singolo gruppo;

- Si analizza il dendrogramma che mostra la distanza che intercorre tra i gruppi durante il processo di clustering.
- Si taglia il dendrogramma all'altezza ritenuta opportuna

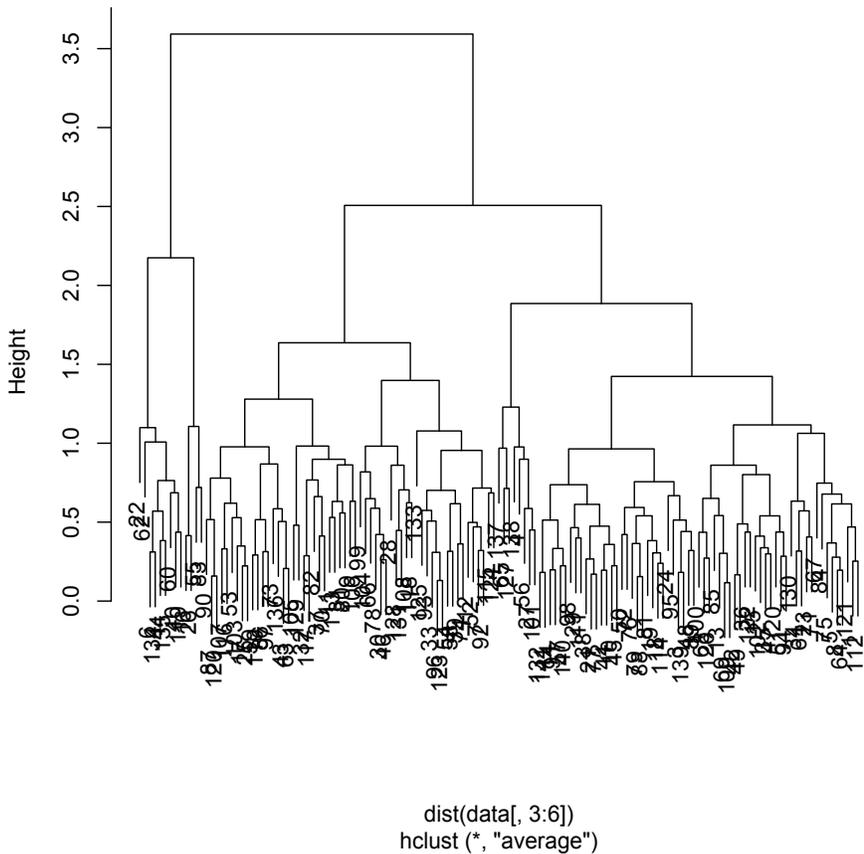
Una volta ottenuto il numero ottimale di gruppi si procede a effettuare una cluster analysis non gerarchica, nel nostro caso utilizzando l'algoritmo kmeans, ricavando la composizione di ogni gruppo.

- **Analisi**

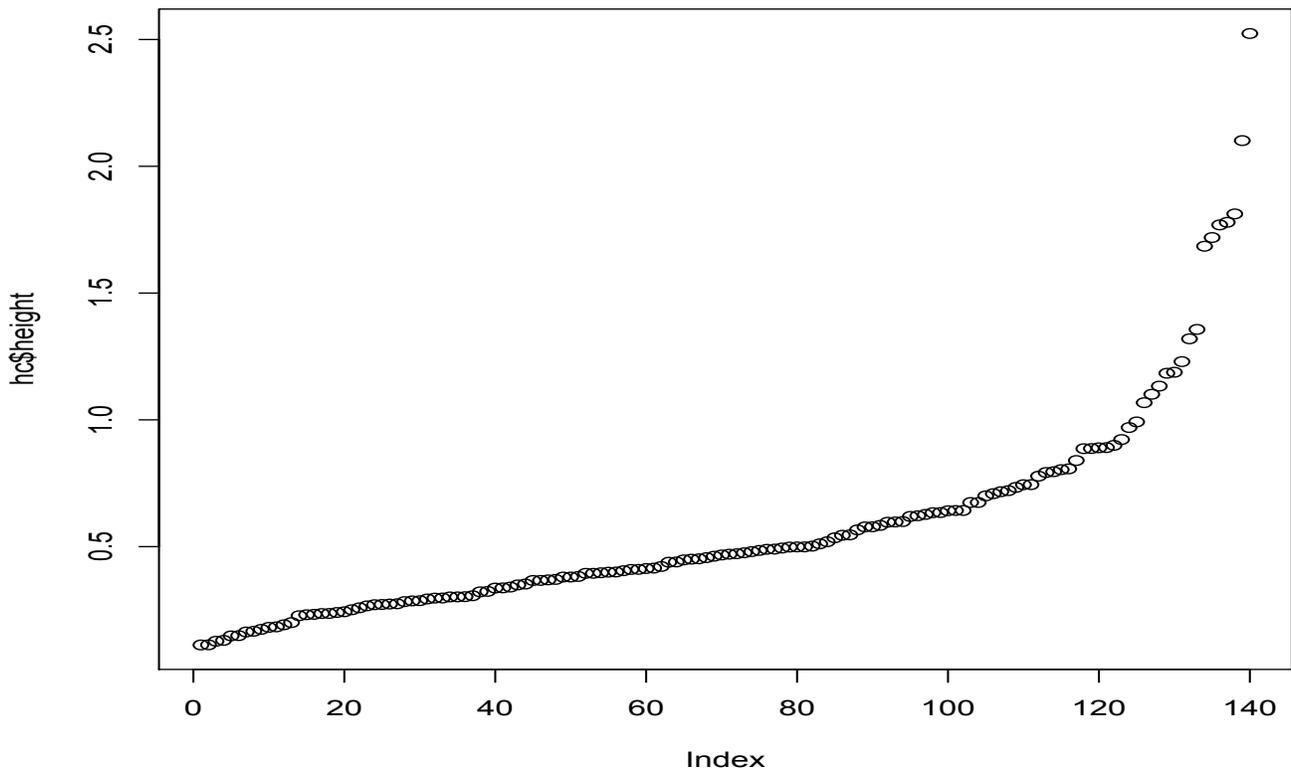
Come spiegato al punto precedente il primo passo è stato effettuare un clustering non gerarchico, e ottenere il dendrogramma come output.

Il dendrogramma del processo di clustering è il seguente:

Cluster Dendrogram



Per stabilire il miglior numero di gruppi da considerare si poteva analizzare il dendrogramma, partendo dal basso, fino a quando non si sarebbe registrato un grande aumento della distanza: in quel caso il numero di gruppi del passo precedente sarebbe stato ottimale perché si utilizzano il numero di cluster la cui formazione comporta la minor distanza di fusione. Per maggiore precisione, però, si è scelta la distanza di taglio analizzando la seguente figura che mostra graficamente la distanza che intercorre tra il numero di gruppi.



Come si nota dalla figura, all'altezza di 1.5, si registra un importante incremento di distanza. Il grande aumento di distanza, indica quindi, che si andrebbero ad unire gruppi molto distanti fra loro. Questa valutazione ci permette di capire che il numero ottimale di cluster è 6.

A questo punto del lavoro è stata praticata una cluster analysis non gerarchica mirata all'ottenimento di 6 gruppi, impostandola con i dati a disposizione e utilizzando la funzione "k-means".

Sono stati ottenuti i 6 cluster con:

- 14 Paesi per il cluster 1;
- 10 Paesi per il cluster 2;
- 21 Paesi per il cluster 3;
- 37 Paesi per il cluster 4;
- 24 Paesi per il cluster 5;
- 35 Paesi per il cluster 6.

Di seguito viene rappresentata la composizione di ogni cluster:

CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6
Argentina	Australia	Austria	Algeria	Bahrain	Albania
Brazil	Canada	Belgium	Angola	Barbados	Armenia
Colombia	China	Croatia	Bangladesh	Bulgaria	Azerbaijan
Costa_Rica	France	Denmark	Bolivia	Chile	Bhutan
India	Germany	Finland	Burkina_Faso	Cyprus	Botswana
Indonesia	Italy	Greece	Burundi	Czech_Republic	Cape_Verde
Kenya	Japan	Hong_Kong_SAR	Cambodia	Estonia	Rep.Dominican
Mexico	Spain	Iceland	Cameroon	Hungary	Egypt
Peru	UK	Ireland	Chad	Israel	El_Salvador
Russian_Fed	United_States	Korea_Rep.	Cote_dIvoire	Latvia	Gambia
South_Africa		Malaysia	Ethiopia	Lithuania	Georgia
Thailand		Netherlands	Gabon	Luxembourg	Guatemala
Turkey		New_Zealand	Ghana	Malta	Guyana
Vietnam		Norway	Guinea	Mauritius	Honduras
		Panama	Haiti	Montenegro	Jamaica
		Portugal	Iran,Islamic Rep.	Oman	Jordan
		Singapore	Kyrgyz.Rep.	Poland	Kazakhstan
		Sweden	Lesotho	Puerto_Rico	Kuwait
		Switzerland	Madagascar	Qatar	Lao_PDR
		Taiwan	Malawi	Saudi_Arabia	Lebanon
		China	Mali	Seychelles	Macedonia
		United.Arab.Emirates	Mauritania	Slovak_Republic	Moldova
			Mozambique	Slovenia	Mongolia
			Myanmar	Trinidad e Tobago	Morocco
			Nepal		Namibia
			Nigeria		Nicaragua
			Pakistan		Philippines
			Paraguay		Romania
			Senegal		Rwanda
			Sierra_Leone		Serbia
			Tajikistan		Sri_Lanka
			Tanzania		Suriname
			Uganda		Swaziland
			Venezuela		Tunisia
			Yemen		Uruguay
			Zambia		
			Zimbabwe		

A questo punto sono stati calcolati i valori centrali di ogni variabile:

	Contesto ambientale culturali	politiche condizioni fav T.T	infrastrutture	risorse naturali &
1	4.391429	4.095000	3.615714	3.925714
2	5.434000	4.247000	5.188000	5.273000
3	5.620000	4.432857	4.956190	3.125238
4	3.664865	3.641892	2.351622	2.097838
5	5.236250	4.285417	4.317500	2.159583
6	4.529429	4.088286	3.198000	1.998857

Per un confronto viene fornita di seguito la media di ogni variabile sul totale della popolazione:

contesto ambientale	politiche e condiz. T.T	infrastrutture	risorse naturali e clturali
4,635744681	4,067943262	3,610921986	2,643475177

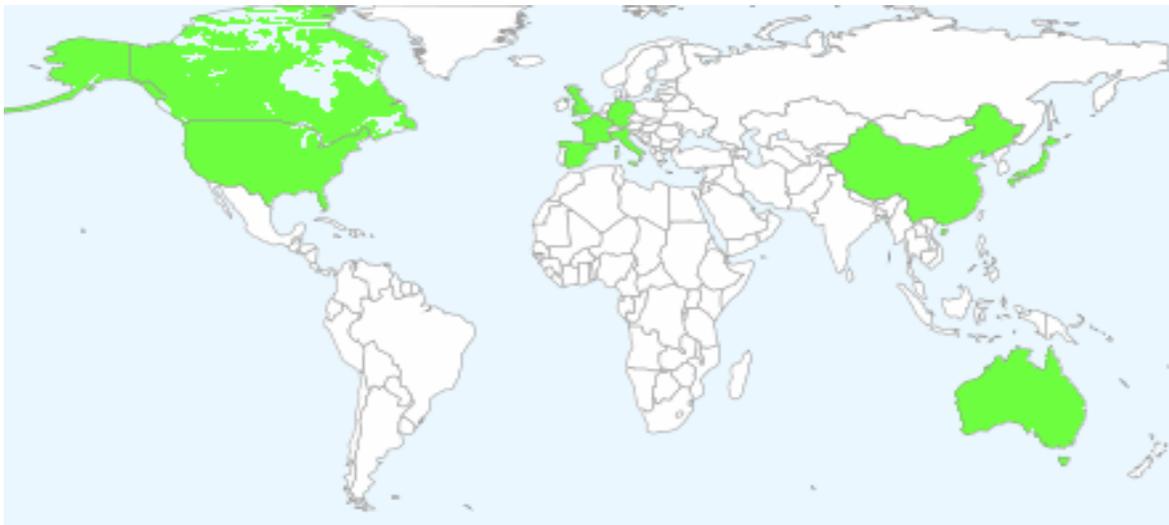
A questo punto è possibile fornire una breve descrizione per ogni

gruppo:

- Il primo possiede un buon punteggio relativamente sia alle impostazioni generali per operare nel Paese che alle politiche e agli aspetti strategici che influenzano il settore turistico, un punteggio medio invece, per quanto riguarda la qualità delle infrastrutture e la presenza di risorse naturali e culturali. La popolazione del cluster, ad eccezione della Russia, è composta da paesi in via di sviluppo, nei quali è evidente lo sforzo sia dei Governi che dell'economia in generale per lo sviluppo del settore, ma che presentano ancora carenze, sul lato delle infrastrutture e, quindi, sul lato degli investimenti.



- Il secondo è composto dai principali Paesi sviluppati. Esso presenta valori alti per tutte le variabili. La grande presenza di risorse culturali e naturali, e lo stato avanzato delle economie di tali Paesi, rendono facile attribuire a tale cluster l'etichetta di “gruppo leader” nel campo dell'attrattività turistica.



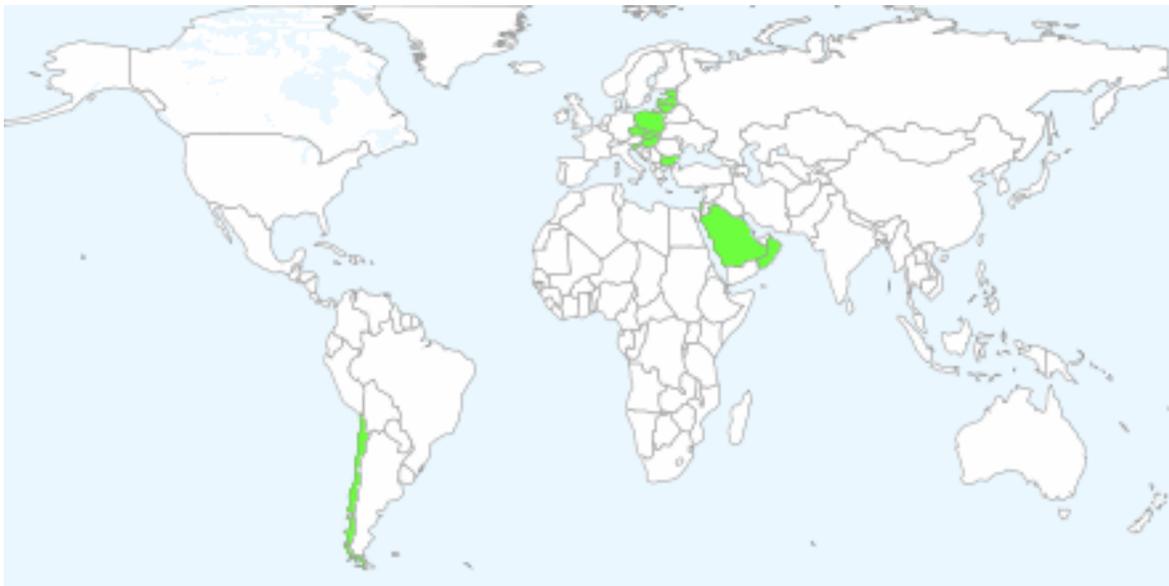
- Il terzo è composto prevalentemente da paesi del nord Europa e da qualche paese asiatico. Presenta valori alti per le prime tre variabili, ma la debole presenza di risorse culturali e naturali è tale da non permettergli il confronto con il cluster precedente.



- Il quarto gruppo è ultimo per le prime tre variabili: insufficienti si presentano quindi il contesto ambientale, il supporto politico al settore, gli aspetti strategici che impattano sull'attività turistica e le infrastrutture. Medio invece il risultato della variabile “risorse culturali e naturali”, che però presenta un indubbio valore per l'attrattività turistica di un Paese, soltanto se supportata da un buon livello delle altre variabili considerate. Il cluster risulta composto quasi per intero da paesi dell'Africa e dell'Asia mediorientale: i cosiddetti “Paesi arretrati”. In tali aree l'economia è ancora in fase iniziale, e mancano un adeguato supporto di regolamenti e politiche, prima ancora che un adeguato livello di investimenti.



- Nel quinto gruppo prevalgono i Paesi dell'est Europa, ma non mancano Nazioni del Medio Oriente e del Sud America. Questi, pur avendo un ottimo contesto ambientale e livelli medio-alti per le politiche intraprese e le infrastrutture, soffrono sul lato culturale e naturale, registrando uno tra i più bassi punteggi in tale variabile.



- Il sesto gruppo è il più eterogeneo a livello geografico. Si distingue dagli altri per il più basso punteggio in riferimento alla variabile “risorse culturali e naturali” e una mediocre qualità delle infrastrutture. Medio invece si presenta il risultato delle prime due variabili.



## LA CLUSTER ANALYSIS CON “R”

### • **Introduzione**

Praticare la cluster analysis senza l'aiuto di strumenti software appropriati renderebbe il processo estremamente lento e l'elaborazione dei dati e la rappresentazione dei risultati molto difficili.

In tempi relativamente recenti, la cluster analysis, così come molte altre discipline statistiche, hanno beneficiato del supporto di hardware e software utili a trattare una grossa mole di dati.

Il presente lavoro è stato reso possibile dall'utilizzo di “R”.

Tale software è utile per la gestione e l'analisi statistica di dati e la produzione di grafici. In R l'analisi è fatta attraverso una serie di passi, con risultati intermedi che sono immagazzinati in “oggetti”. Rispetto a programmi come SAS o SPSS, R fornisce un output minimo. I rimanenti risultati possono, poi, essere richiamati con gli opportuni comandi. R è stato pensato per essere utilizzato tramite riga di comando, infatti ogni istruzione va inserita tramite comandi di linea. Esistono anche interfacce grafiche, ma sono limitate alle semplici operazioni ricorrenti (caricare dati nel programma, ecc).

- **L'utilizzo di R per la cluster analysis**

Per prima cosa sono stati importati i dati in R, che devono essere strutturati in casi per riga e variabili su colonne, mediante la funzione “*read.csv*”.

La prima operazione è stata una clustering gerarchica, utile per analizzare i dati e stabilire il numero di cluster, richiamata con la funzione “*hclust*”.

In seguito, per la rappresentazione del dendrogramma, è stata utilizzata la funzione “*plot*” e, una volta decisa l'altezza alla quale praticare il taglio, è stata utilizzata la funzione “*cutree*”.

Mediante la funzione “*table*” è stato possibile capire il numero ottimale di gruppi, che è 6.

A questo punto è stata utilizzata la funzione “*kmeans*”, che prende il nome dall'omonimo algoritmo, e si sono formati i 6 cluster oggetto di analisi.

La funzione “*kmeans\$centers*” ha permesso di evidenziare le medie, delle variabili considerate, per ogni gruppo.

## BIBLIOGRAFIA

Analisi di mercato, Tommaso Proietti

Economic impacts of Tourism, Daniel J. Stynes

Puretourism.it

Quantitative methods in tourism economics, Alvaro Matias, P.N. , M.S.

Storia economica, Ennio De Simone

The World Economic Forum

WTTC-Travel e Tourism economic impact 2016 world

Rizzi, 1985, p. 52 ; Everitt, 1993, p

