Università
Ca'Foscari
Venezia

Corso di laurea magistrale in
Marketing e Comunicazione

Tesi di laurea

Cà Foscari
Dorsoduro 3246
30123 Venezia

Strumenti per la gestione ambientale nel settore dell'agriturismo in Veneto: un'analisi empirica

Relatore

Ch.ma Prof.ssa Isabella Procidano

Laureando

Alan De Renzis

Matricola 821643

Anno accademico

2013/2014

INDICE

Introduzione *……1*

Capitolo 1

Agriturismi e sostenibilità

1.1 Turismo e sostenibilità……3

1.1.1 Impatti del turismo sull'ambiente……3

1.1.2 Sostenibilità turistica……...7

1.1.3 I promotori del turismo sostenibile……10

1.1.4 I documenti più significativi a favore del turismo e dello sviluppo sostenibile…….12

1.1.5 Obiettivi di sostenibilità a livello europeo…….18

1.1.6 Gli indicatori di sviluppo sostenibile per le destinazioni turistiche……21

1.2 Agriturismi e sostenibilità ambientale……..25

1.2.1 L'attività agrituristica……....25

1.2.2 L'Agri-turista…….28

1.2.3 Gli impatti ambientali di un agriturismo……28

1.2.4 Verso un agriturismo sempre più "ecologico"……..30

1.2.5 Le principali certificazioni ambientali……...38

1.2.5.1 Norma ISO 14001……39

1.2.5.2 Emas……42

1.2.5.3 Ecolabel……...44

Capitolo 2

Verso un approccio fuzzy al clustering

2.1 Due logiche a confronto……49

2.1.1 La logica dicotomica……..50

2.1.2 La logica fuzzy……...57

2.2 La cluster analysis……64

2.2.1 Introduzione……...64

2.2.2 Classificazione delle tecniche di clustering……65

2.2.3 Classificazione classica e classificazione sovrapposta……...68

2.2.4 Partizioni hard, fuzzy e possibilistiche……...73

Capitolo 3

Fuzzy clustering

3.1 Introduzione……..79

3.2 Fuzzy clustering gerarchico……..80

3.2.1 Metodo della sintesi di più partizioni……81

3.2.2 Metodo dei ricoprimenti sfocati……85

3.2.3 Metodo del legame medio sfocato……..91

3.3 Fuzzy clustering non gerarchico……...94

3.3.1 Metodo delle k-medie sfocato……94

3.3.2 Miglioramenti del metodo delle k-medie sfocato……107

3.3.2.1 Metodi di classificazione semisfocata: i metodi di Selim, Ismail e Kamel……108

3.3.2.2 Metodo delle k-medie semisfocato……110

3.3.2.3 Algoritmo di Gustafson-Kessel……112

3.3.3 Altri metodi non gerarchici……..116

3.3.3.1 Il metodo FANNY……117

3.3.3.2 Il metodo MND2……118

3.4 Conclusioni……119

Capitolo 4

Un'analisi empirica

4.1 Premessa……121

4.2 Analisi del questionario……..122

4.2.1 Informazioni generali sugli agriturismi……124

4.2.2 Focus sulla sostenibilità……...136

4.3 Analisi dei cluster……...154

4.3.1 Analisi introduttiva……..155

4.3.2 Classificazione classica……...163

4.3.2.1 Clustering gerarchico……164

4.3.2.2 Clustering non gerarchico……..169

4.3.3 Fuzzy clustering……...174

4.3.4 Valutazione della composizione dei cluster……184

4.3.5 Profilazione dei gruppi……195

Conclusioni……197

Appendice

Appendice A: Tabella corrispondenze domande questionario <--> variabili……201

Appendice B: Tabelle di contingenza……...207

Appendice C: Gradi di appartenenza ai cluster……211

Bibliografia e sitografia……..213

Introduzione

L'obiettivo del presente lavoro è quello di fornire un confronto il più possibile completo fra due tecniche antagoniste di analisi di cluster, o per meglio dire due "famiglie" antagoniste di tecniche di analisi dei cluster.
Da un lato troviamo infatti le diverse tecniche di classificazione classica, detta hard, il cui obiettivo è quello di classificare un insieme di unità statistiche in gruppi ben definiti, in cui le unità appartengono o non appartengono ad un dato gruppo.
Dall'altro lato invece troviamo tutte quelle tecniche appartenenti alla recente classificazione sfocata, detta fuzzy clustering, introdotta da James C. Bezdek nel 1981.
Il confronto verrà effettuato utilizzando un data set contenente le risposte date ad un questionario on-line rivolto agli agriturismi del Veneto, avente l'obiettivo di raccogliere informazioni circa gli strumenti adottati dagli stessi per rendere la propria attività sostenibile.
È da dire fin da ora che inizialmente l'obiettivo di questo lavoro era quello di operare un'analisi dei cluster volta a identificare gruppi di agriturismi a livello regionale che presentassero caratteristiche comuni in riferimento agli strumenti adottati per la salvaguardia e la sostenibilità dell'ambiente, il tutto con l'intento di definire un quadro generale sulla situazione nella regione Veneto. Purtroppo il riscontro al questionario da parte degli agriturismi è stato deludente.
Una minima parte infatti ha risposto alle domande proposte nel questionario on-line, nonostante le ripetute sollecitazioni via mail e in molti anche via telefono. Dunque si è dovuto optare per sviluppare il lavoro nel modo sopra indicato poiché, evidentemente, un obiettivo come quello inizialmente pensato non avrebbe avuto senso.

La presentazione del lavoro è suddivisa in quattro capitoli:

1. *Primo capitolo*: si tratta di un capitolo teorico contenente nella prima parte i concetti principali riguardanti il turismo sostenibile e nella seconda parte un focus sugli agriturismi e sui relativi strumenti adottabili per rendere la loro attività sostenibile.

2. *Secondo capitolo*: contiene un confronto fra la logica dicotomica e la logica sfumata, logiche sulle quali di fondano le diverse tecniche di analisi dei cluster. Inoltre, a titolo di introduzione e confronto con il successivo capitolo, sarà presente un quadro generale delle principali tecniche di analisi classica dei cluster. Infine, utilizzando la funzione di appartenenza, verranno analizzate le differenze fra le partizioni hard e le partizioni fuzzy, in modo da completare il percorso di analisi che permetterà di affrontare l'analisi vera e propria della fuzzy clustering.

3. *Terzo capitolo*: rappresenta il capitolo teorico centrale del presente lavoro. Viene qui infatti affrontata un'analisi completa e dettagliata delle principali tecniche di analisi sfocata dei cluster, distinguendo in tecniche gerarchiche e tecniche non gerarchiche.

4. *Quarto capitolo*: capitolo applicativo contenente nella prima parte l'analisi delle domande del questionario e nella seconda parte il confronto "operativo" fra una tecnica di analisi dei cluster classica ed una corrispondente tecnica di analisi dei cluster sfocata. Le diverse analisi verranno effettuate utilizzando il programma di statistica R.

Capitolo 1

AGRITURISMI E SOSTENIBILITA'

1.1 Turismo e sostenibilità

1.1.1 Gli impatti del Turismo sull'ambiente

Il turismo è sempre stato considerato un settore industriale che genera ricchezza pulita, tale da essere definito "smokeless industry", cioè industria senza ciminiere. Dagli anni 60 però si è presa coscienza del fatto che il turismo può determinare degrado ed inquinamento ambientale e sociale se non opportunamente gestito.
Il turismo, legato ad una particolare destinazione, fonda la sua esistenza e la sua ricchezza su alcuni elementi fondamentali:

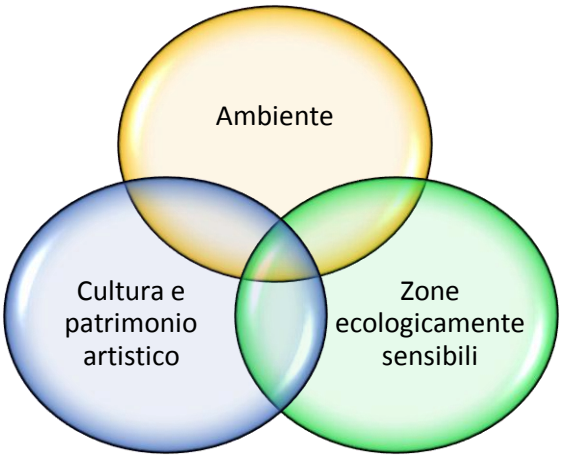


Figura 1.1 - Elementi fondanti del turismo

Di conseguenza le imprese del settore, le amministrazioni pubbliche, i turisti e i residenti devono contribuire, in rapporto alle proprie possibilità, a proteggere l'ambiente e le

risorse naturali, a preservare la cultura e le tradizioni, in quanto fattori fondamentali che alimentano il turismo.

In particolar modo è necessario salvaguardare l'ambiente: il turismo infatti può causare un forte inquinamento e degrado ambientale se non opportunamente gestito.

Gli impatti ambientali collegati all'attività turistica possono essere classificati come segue:

Impatto sul paesaggio: consiste nell'alterazione del contesto ambientale preesistente dovuta all'urbanizzazione e alla costruzione delle diverse strutture ricettive
Alterazione degli ecosistemi marini e terrestri con effetti su specie vegetali ed animali e conseguente perdita di biodiversità
Consumo di risorse naturali (acqua, energia, suolo)
Inquinamento dell'acqua, del suolo ed atmosferico
Inquinamento da rifiuti ed acque reflue
Congestione ed inquinamento acustico da traffico
Perdita dell'identità culturale dovuta all'eccessiva commercializzazione e modificazione dell'arte e dei modelli locali.

Tali effetti incrementano il loro impatto nel momento in cui l'attività turistica è accompagnata da un'elevata stagionalità, in molti casi ciò spinge gli operatori turistici a puntare su economie di scala, superando così i limiti di carico ambientale delle località interessate.

Come detto in precedenza oggi esiste una maggiore attenzione ai rischi ambientali portati da uno sviluppo non sostenibile del turismo. Nel fragile rapporto tra ambiente, tradizioni

locali e turismo possono inserirsi alcuni elementi negativi: la forte attrazione che molte località suscitano nelle persone può determinare un eccesso di presenze di turisti, con aumenti incontrollabili dei flussi e dei fenomeni di congestione. Questo può portare ad un superamento della capacità di carico sociale, ecologica e culturale delle mete turistiche e compromettere in maniera definitiva il patrimonio naturale dell'area.

Per questo è necessario definire uno sviluppo sostenibile del turismo in modo da convertirlo man mano in elemento di valorizzazione e non di degrado del territorio.

Esistono diversi strumenti per misurare l'equilibrio tra conservazione dell'ambiente e sviluppo del turismo; uno di questi è la "***capacità di carico turistica (CCT)***" definita dall'Organizzazione mondiale del Turismo come "*il numero massimo di persone che nello stesso periodo visitano una determinata località turistica senza comprometterne le caratteristiche ambientali, economiche, fisiche e socio-culturali e senza provocare una riduzione della soddisfazione dei turisti*" (www.svilupposostenibile.it).

Tale strumento è legato ad un insieme di problematiche riguardanti la politica locale, la pianificazione e il processo decisionale. Obiettivo primario della CCT è quello di realizzare scenari di turismo da poter perseguire senza che nel lungo termine vadano a compromettere le risorse naturali, l'economia e il tessuto sociale di una determinata località.

Una metodologia molto interessante di CCT sviluppata dall'UNEP (Unite Nations Environmental Programme) in collaborazione con il Centro di Attività Regionali del Piano d'Azione per il Mediterraneo consiste in una tecnica di analisi (Carrying Capacity Assesment) diventata parte integrante del processo di pianificazione e gestione del turismo (A.r.p.a.v., 2012).

Lo strumento della CCA considera tutte le componenti fondamentali dello sviluppo turistico locale:

1. Tipologia ed ampiezza delle attività turistiche;
2. Preferenze della popolazione locale, degli operatori e dei turisti;
3. Politiche turistiche a livello micro e macro e come queste entrano in relazione fra loro;
4. Caratteristiche ambientali e urbane dell'area.

Senza scendere nei dettaglio di questo strumento, la cui applicazione deve essere adattata di volta in volta alla realtà in esame, esso si articola essenzialmente in cinque fasi:

Figura 1.2 - Fasi di applicazione CCA. Fonte: elaborazione personale.

Di esempi applicativi di questa metodologia ne possiamo trovare moltissimi, sia a livello italiano che a livello internazionale. Vedi per esempio la "Knowledge Resource Guide (KRG)[1]" realizzata dalla provincia di Rimini nell'ambito del progetto SUVOT (Sousteinable and Vocational Tourism) volta ad analizzare tutte le esperienze e pratiche più significative di turismo sostenibile realizzate nelle destinazioni costiere del Mediterraneo.

A livello internazionale possiamo invece citare il progetto CAMP[2] (Costal Area Management Programme) per l'area di Fuka Matrouh in Egitto.

[1] http://www.a21italy.it/medias/9A52DFBBB778E3F0.pdf

[2] http://www.pap-thecoastcentre.org/itl_public.php?public_id=46&lang=en

1.1.2 Sostenibilità turistica

E' tempo di vacanze o di un weekend lontano dalla città? Cosa c'è di meglio che trovarsi imbottigliati per ore sulla via del mare, non trovare una spiaggia libera ma solo accessi a pagamento e, infine, ammirare una lunga distesa di opere edilizie costruite a pochi metri dal mare?

Ovviamente quanto appena detto è ironico ed è diventato un vero problema per molte località turistiche.

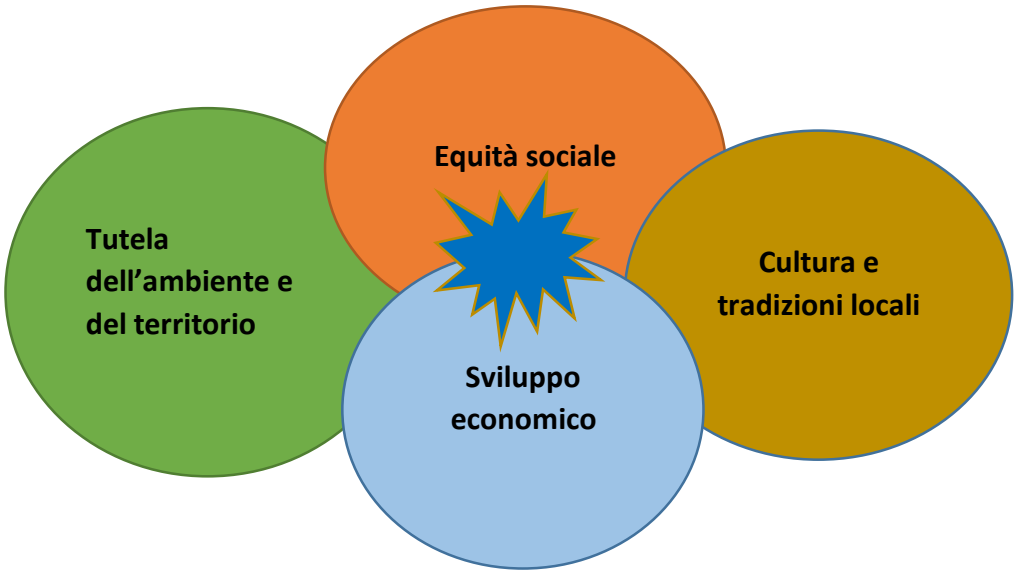
Il turismo sostenibile nasce anche in risposta a queste problematiche, considerate sempre più importanti dai turisti di tutto il mondo.

Il concetto di turismo sostenibile è stato definito per la prima volta nel 1988 dall'Organizzazione Mondiale del Turismo, la cosiddetta OMT. Le attività turistiche si definiscono sostenibili nel momento in cui si sviluppano in modo da mantenersi vitali in un'area turistica per un periodo di tempo illimitato, non alterano l'ambiente (naturale, artistico, sociale) e non inibiscono o ostacolano lo sviluppo di altre attività sociali ed economiche.

Oltre a questa definizione ce ne sono molte altre, suggerite da vari enti ed organizzazioni. Per esempio il W.W.F definisce il turismo sostenibile come un turismo, con associate infrastrutture che, ora e nel futuro:

Poi c'è il Centro VIA (Italia) che definisce il turismo sostenibile come un turismo capace di durare nel tempo mantenendo inalterati i suoi valori qualitativi e quantitativi, un turismo capace di far coincidere nel breve e nel lungo termine le aspettative dei residenti con quelle dei turisti senza compromettere il livello qualitativo dell'esperienza turistica e senza danneggiare il territorio interessato dal fenomeno.

I pilastri della sostenibilità possono essere riassunti nel seguente grafico:



Cosa indicano questi quattro macro gruppi?

Con equità sociale indichiamo la ricerca di soluzioni adeguate per i turisti in termini di costi, qualità e tipologia dei servizi offerti, e la ricerca di condizioni di lavoro accettabili per gli operatori.

Con cultura e tradizioni locali facciamo riferimento al rispetto dello stile di vita e della cultura delle popolazioni locali. Nello sviluppo economico rientrano invece i seguenti obiettivi: la ricerca di effetti positivi sulle economie locali e regionali, l'aumento della qualità nella produzione di beni e servizi, l'utilizzo di nuove tecnologie (nuove fonti energetiche, prodotti a basso consumo energetico), la creazione di nuovi posti di lavoro.

Infine con tutela dell'ambiente e del territorio intendiamo la salvaguardia e la promozione dell'ambiente naturale, del patrimonio artistico e architettonico e del paesaggio delle varie località turistiche. L'obiettivo della sostenibilità è quello di unire questi quattro gruppi.

Infine, analizzando nel dettaglio le caratteristiche del turismo sostenibile possiamo identificarne sette:

a) ***Dimensionato nel tempo***, allo scopo di ridurne il più possibile gli effetti legati alla stagionalità;
b) ***Dimensionato nello spazio***, individuando la capacità massima di accoglienza del territorio e limitando l'affluenza dei turisti in base alle caratteristiche fisiche dei luoghi. Viene definita in questo modo una quantità di visitatori che permetta di garantire la conservazione degli spazi e la qualità dell'esperienza turistica;
c) ***Durevole***: l'attenzione è verso il lungo termine. L'obiettivo è l'adozione di un modello turistico che cerchi di mettere in armonia la crescita economica e la conservazione dell'ambiente e dell'identità locale, fattori che rappresentano il principio attivo dello sviluppo turistico nel tempo;
d) ***Integrato e diversificato***: il turismo deve essere un elemento in armonia con l'identità del luogo, integrato alla ricchezza economica e culturale dello stesso. L'offerta turistica deve essere dunque il risultato naturale delle risorse locali (feste tradizionali, gastronomia, patrimonio artistico e architettonico e molto altro). Per diversificato intendiamo dire che la "*monocultura turistica*" dovrebbe essere sostituita con dei modelli diversificati di turismo in cui esso occupi una parte importante dell'intera struttura economica;
e) ***Economicamente vitale***: l'obiettivo non è la rapida crescita dei redditi turistici, ma la durabilità nel tempo degli investimenti fatti. La costruzione di un modello alternativo che dia importanza alla ricerca del benessere sociale;
f) ***Pianificato***: per garantire un esito il più possibile positivo e durevole è necessaria un'analisi attenta delle condizioni presenti e delle prospettive future, tenendo in considerazione anche le variabili che nel momento della decisione intervengono nel processo turistico;
g) ***Partecipato***: è necessario coinvolgere tutti i soggetti nei processi decisionali e attuativi che riguardano lo sviluppo del territorio.

1.1.3 I promotori del turismo sostenibile

In questo breve paragrafo verranno presentati i principali promotori[3] del turismo sostenibile a livello mondiale, europeo e nazionale. Tali "organi promotori" hanno nel corso degli anni elaborato documenti, sottoscritto carte e adottato iniziative a sostegno del turismo sostenibile allo scopo di renderlo, come dire, un po' meno "di massa".

A livello mondiale i principali promotori degni di nota sono:

a) ***L'Organizzazione Mondiale del Turismo (WTO/OMT):*** è un'organizzazione intergovernativa a cui partecipano 138 paesi e che si pone come punto di riferimento globale per le politiche turistiche e le questioni ad esse correlate. E' promotrice di moltissimi convegni internazionali, di svariate dichiarazioni sul turismo sostenibile e ha come compito quello di promuovere ed incentivare il turismo quale strumento per favorire la comprensione internazionale e lo sviluppo economico.

b) ***La United Nations Environment Programme (UNEP):*** è la principale organizzazione di assistenza all'attuazione dell'Agenda 21 sulle tematiche riguardanti il turismo. Il suo compito è quello di assistere i governi nella creazione e applicazione dei piani politici ambientali.

c) ***La Commissione delle Nazioni Unite sullo Sviluppo Sostenibile (CSD):*** tale commissione è nata con lo scopo di garantire l'attuazione dell'Agenda 21 e la Dichiarazione di Rio sullo Sviluppo e l'Ambiente (2002). Inoltre svolge il ruolo di guida politica a livello locale, nazionale e mondiale di supporto al Piano d'implementazione di Johannesburg del 2002.

d) ***Il World Travel and Tourism Council (WTTC):*** rappresenta a livello mondiale il settore industriale e del turismo. Il suo scopo è quello di incoraggiare i governi ad adottare una politica di sviluppo sostenibile del turismo, cercando di aumentare la consapevolezza degli impatti positivi che un turismo ben gestito può creare per la società (prosperità e lavoro).

[3] http://www.sinanet.isprambiente.it/gelso/tematiche/buone-pratiche-per-il-turismo/principali-promotori-del-turismo-sostenibile

e) ***Il Dipartimento per lo Sviluppo Sostenibile della FAO (SD):*** nasce con l'obiettivo di favorire un approccio strategico integrato allo sviluppo sostenibile. È responsabile della coordinazione delle azioni della FAO in seguito alle decisioni prese durante la Conferenza di Rio del 1992.

f) ***L'Istituto Internazionale per l'Ambiente e lo Sviluppo (IIED) e l'Istituto Internazionale per lo Sviluppo Sostenibile (IISD):*** questi due istituti si occupano di ricerche i cui risultati diano un aiuto allo sviluppo sostenibile globale oltre che essere di aiuto alle decisioni dei governi.

A livello Europeo i principali promotori sono:

a) ***L'Unità Turismo della Commissione Europea:*** svolge attività di supporto al settore turistico. Fornisce un contributo molto importante alla crescita dello sviluppo e alla stabilità dell'occupazione in Europa.

b) ***L'Agenzia Europea per l'Ambiente (EEA):*** è un Agenzia dell'Unione Europea composta da 32 paesi membri. Rappresenta una fonte di informazione di primaria importanza per coloro che si occupano dello sviluppo, dell'implementazione, dell'adozione e della valutazione delle politiche ambientali. Ha realizzato il portale Destinet, destinato a tutti coloro che hanno un ruolo nella gestione delle destinazioni turistiche a livello europeo.

Infine, a livello nazionale hanno rilievo:

a) ***L'Associazione Cultura Turismo Ambiente (ACTA):*** si occupa della ricerca e della sperimentazione di servizi e prodotti innovativi per il settore del turismo. Inoltre supporta e promuove l'applicazione dell'Ecolabel europeo.

b) ***L'Associazione Italiana del Turismo Responsabile (AITR):*** sostiene e promuove il turismo sostenibile, responsabile ed etico. È costituita da una serie di organizzazioni, associazioni e cooperative che svolgono attività nel campo del turismo responsabile.

1.1.4 I documenti più significativi a favore del turismo e dello sviluppo sostenibile

In questo paragrafo verranno esposti in modo cronologico e sintetico i principali documenti a livello nazionale ed internazionale sullo sviluppo sostenibile e sul turismo sostenibile allo scopo di mostrare quelli che sono stati i passi fondamentali nella loro promozione. Saranno descritti i punti fondamentali alla comprensione di ciascun documento e indicate nella sitografia di questo lavoro le fonti da poter consultare per un approfondimento degli stessi.

Conferenza di Stoccolma (1972)

È stata la prima importante tappa dello Sviluppo Sostenibile a livello mondiale. Durante gli anni 70 nasce infatti la consapevolezza che le risorse naturali della Terra devono essere tutelate. Inoltre vi è l'affermazione dei principi di libertà, uguaglianza e del diritto di tutti ad adeguate condizioni di vita.

Parteciparono a questa conferenza 113 nazioni le quali redissero un piano d'azione con 109 raccomandazioni, oltre ad una Dichiarazione contenente 26 principi sui diritti e responsabilità dell'uomo in relazione all'ambiente.

È con tale conferenza che nasce UNEP, il Programma per l'Ambiente delle Nazioni Unite.

Agenda 21 (1992)

Figura 1.3 - Logo Agenda 21

Nei giorni 3 e 4 giugno 1992 a Rio de Janeiro ha avuto luogo la "United Nations Conference on Environment and Development[4]" (UNCED), meglio conosciuta come Earth Summit.

Durante la conferenza, passata alla storia come la più grande Conferenza per numero di partecipanti, si sono discussi i principali problemi ambientali del pianeta e le loro relazioni con i problemi dello sviluppo economico e sociale.

[4] http://sustainabledevelopment.un.org/content/documents/Agenda21.pdf

Più di 150 paesi hanno firmato due Convenzioni internazionali: la prima riguardante i mutamenti climatici, la seconda la protezione della biodiversità.
Tutte le delegazioni presenti hanno invece approvato:

- ***La dichiarazione di Rio***, un documento che sancisce l'impegno internazionale sulla tutela ambientale e lo sviluppo sostenibile;
- ***Una Dichiarazione di Principi*** senza valore legale sulla conservazione, gestione e sviluppo sostenibile delle foreste;
- ***L'agenda 21***, un vero e proprio piano d'Azione ampio e articolato delle cose da fare nel corso del 21° secolo includendo anche il ruolo del turismo nella salvaguardia delle risorse ambientali e naturali. L'Agenda 21 rappresenta il Programma d'Azione della Comunità internazionale (ONU, Governi, Stati, Ngo, settori privati) in materia di ambiente e sviluppo, una sorta di manuale per lo sviluppo sostenibile del pianeta nel corso del 21° secolo. Il documento, composto da 800 pagine organizzate in 40 capitoli divisi suggerisce che è ormai essenziale cambiare strada migliorando gli standard di vita per tutti, gestendo e proteggendo l'ambiente e le risorse naturali allo scopo di garantire un futuro più sano per l'intera umanità. I capitoli sono stati divisi in quattro parti incentrate sui seguenti temi: le dimensioni sociali ed economiche (ambiente, povertà, produzione, sanità, aspetti demografici e molto altro), la gestione e conservazione delle risorse (deserti, foreste, acqua, montagne, atmosfera, prodotti chimici, rifiuti), il rafforzamento del ruolo dei gruppi più significativi (giovani, donne, Ngo, sindacati, agricoltori), i metodi di esecuzione delle azioni.

Il capitolo 28 dell'agenda invitava le autorità locali di tutto il mondo ad avviare un dialogo con le loro popolazioni allo scopo di cercare il consenso per una Agenda 21 locale entro il 1996. Il dialogo con i cittadini, le imprese private e le associazioni locali è necessario allo scopo di acquisire le informazioni necessarie alla formulazione di efficienti strategie. In pratica l'Agenda 21 locale può essere vista come uno sforzo comune per raggiungere il massimo consenso tra i vari attori sociali di una determinata località riguardo la definizione e l'attuazione di un Piano d'Azione ambientale da avviare entro il 2000 e che sia rivolto a tutto il 21° secolo.

L'Agenda 21 locale è un processo partecipativo e democratico, alla cui definizione ed attuazione partecipano tutti i settori e che prevede la definizione degli obiettivi ambientali e la realizzazione delle condizioni per metterli in atto, cioè consenso, interesse, risorse finanziarie ed umane, sinergie. È inoltre necessario definire dei criteri e dei metodi per la valutazione dei risultati allo scopo di monitorare il perseguimento del piano nel corso del tempo.

La carta di Lanzarote (1995)

La Carta di Lanzarote, scritta dalla Conferenza Mondiale sul Turismo Sostenibile nel 1995, rappresenta un punto fondamentale nella storia.
Il documento individua i 18 punti per la creazione di un nuovo modello turistico in grado di rispettare il patrimonio ambientale e le culture locali. Esso stabilisce delle linee d'azione concrete e fa appello alla comunità internazionale, ai governi locali, alle autorità pubbliche, ai professionisti del settore del turismo e a tutti gli attori le cui attività sono in relazione con il turismo, di adottare specifiche misure per favorire l'integrazione del turismo nella strategia complessiva di sviluppo sostenibile.

La dichiarazione di Berlino (1997)

In tale dichiarazione, composta da 21 principi, viene evidenziata la responsabilità a livello locale dello sviluppo durevole del turismo e la necessità di un controllo e di una gestione costante e meticolosa dello stesso.

La dichiarazione di Calvià (1997)

Come per la dichiarazione di Berlino, anche in questa si indicano una serie di intenti per sostenere il rispetto dell'ambiente naturale ed artistico. Vengono suggerite delle misure per la riduzione delle risorse energetiche ed idriche, per la gestione dei rifiuti, per la produzione e per il rispetto delle tradizioni artistiche e culturali delle località che ospitano le attività turistiche.

La dichiarazione di Manila (1997)

La dichiarazione di Manila, del 1997, contiene una riflessione sugli aspetti sociali rilevanti del turismo sostenibile e indica una serie di azioni da perseguire al fine di combattere le iniquità sociali e lo sfruttamento indotte dalle attività turistiche.

La carta Europea del Turismo sostenibile nelle aree protette (1999)

Figura 1.4 - Logo Europarc

Tale carta elaborata nel 1999 da un gruppo di rappresentanti europei delle aree protette, del settore turistico e dei loro partner, si propone di favorire la concreta applicazione del concetto di sviluppo durevole, ossia uno sviluppo in grado di soddisfare le esigenze delle generazioni attuali senza compromettere la capacità delle generazioni future di rispondere alle proprie. Per sviluppo durevole si intende uno sviluppo sostenibile.

Al fine di perseguire questo obiettivo è necessario proteggere le risorse, attivare uno sviluppo economico vitale e socialmente equo.

La carta Europea manifesta la volontà delle istituzioni che gestiscono le aree protette e dei professionisti del turismo di favorire un turismo compatibile con i principi dello sviluppo sostenibile, e li impegna ad attuare una strategia a livello locale in favore di esso.

La carta di Rimini per il turismo sostenibile (2001)

Nel 2001 è stata emessa dalla Conferenza Internazionale sul Turismo Sostenibile la Carta di Rimini, nella quale si afferma la centralità delle strategie e delle azioni specifiche sulla mobilità delle destinazioni turistiche di massa per consolidare nuovi stili di vita e di frequentazione del territorio orientati alla sostenibilità sociale, ambientale ed economica dello sviluppo locale.

I principali obiettivi contenuti nella Carta sono i seguenti:

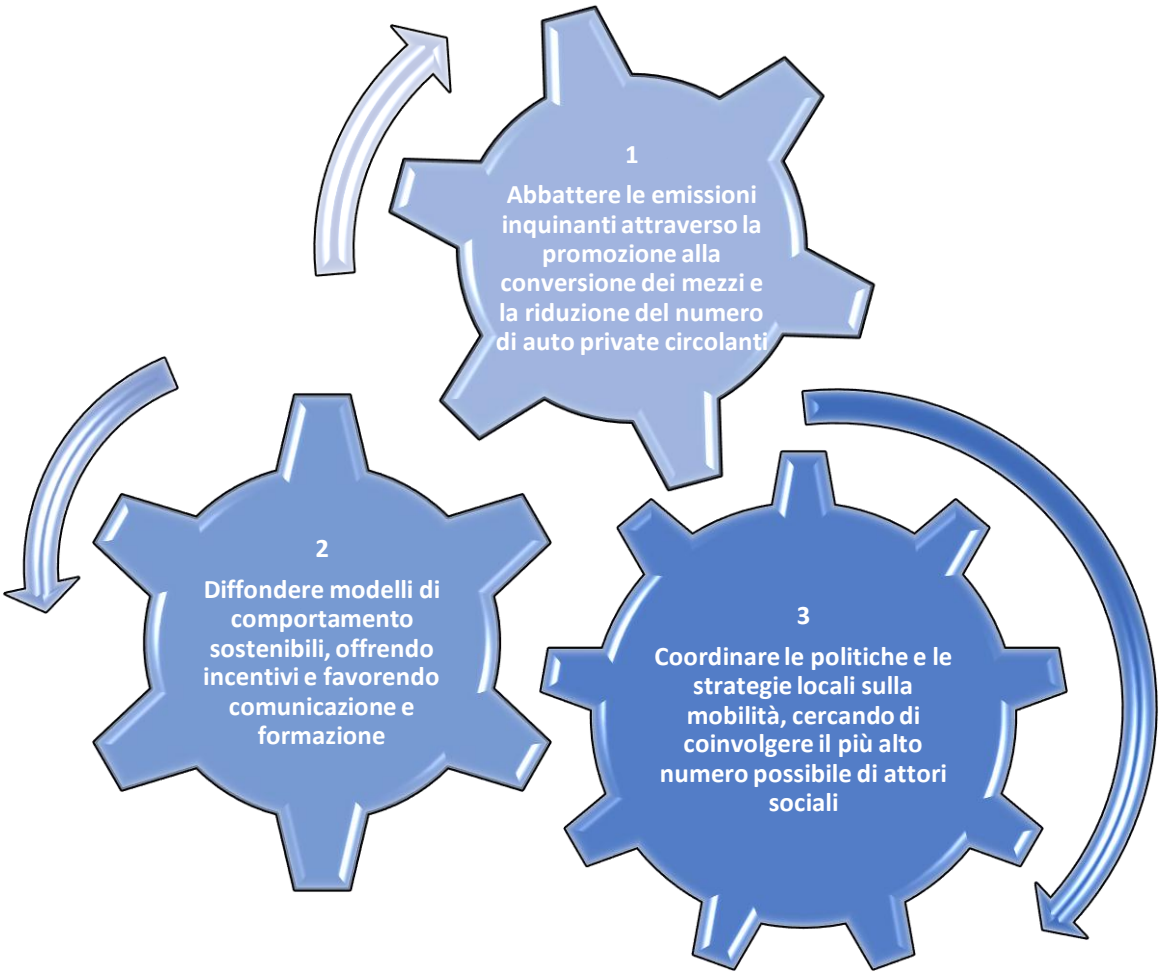


Figura 1.5 - Principali obiettivi Carta di Rimini del 2001. Fonte: elaborazione personale.

Al seguito di questi obiettivi vengono poi indicate una serie di raccomandazioni e di azioni da intraprendere al fine di perseguirli nel modo migliore possibile.
Nel 2008 si è svolta a Riccione una seconda Conferenza Internazionale sul Turismo Sostenibile nella quale sono stati ripresi ed aggiornati gli obiettivi e le relative azioni indicate nella prima Carta di Rimini.

Vertice mondiale sullo sviluppo sostenibile, Johannesburg (2002)

Nel 2002 a Johannesburg ha avuto luogo il summit mondiale sullo sviluppo sostenibile a cui hanno partecipato numerosi capi di Stato e di governo, rappresentanti delle Organizzazioni Non Governative (ONG) e del settore privato.

Lo scopo della riunione è stato quello discutere e cercare una soluzione ai principali problemi che affliggono il nostro pianeta e pianificare un futuro migliore per tutti.
Tale vertice ha costituito un'occasione per riflettere su quanto iniziato alla conferenza di Rio del 1992.
L'attenzione è stata posta alle nuove sfide da affrontare per realizzare uno sviluppo più sostenibile, cioè uno sviluppo capace di mettere in armonia gli aspetti economici, sociali e ambientali e in grado di assicurare una società equa e prospera alle generazioni future.
Il risultato più importante del summit è stato la redazione di un piano d'azione, articolato in 37 punti, nel quale sono stati individuati e descritti i temi chiave per il successivo decennio.

Conferenza di Rio (2012)

Nel 2009 l'Assemblea Generale delle Nazioni Unite ha stabilito di organizzare nel 2012 la conferenza di Rio sullo sviluppo sostenibile, chiamata Rio+20 a indicare che cade 20 anni dopo il Vertice Mondiale di Rio del 1992.
Tale conferenza è incentrata intorno a due temi principali:

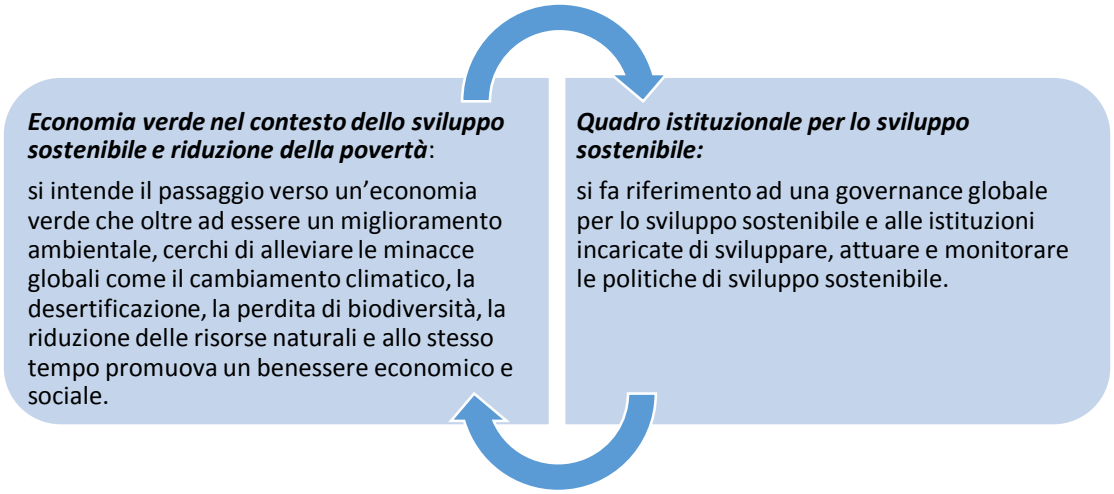


Figura 1.6 - Temi principali Conferenza di Rio 1992. Fonte: elaborazione personale.

L'obiettivo della conferenza è stato quello di promuovere nuovi traguardi considerando soprattutto i risultati raggiunti e riflettere sugli ostacoli nell'affrontare le nuove sfide in linea con le raccomandazioni emerse durante i vertici passati sullo sviluppo sostenibile.

1.1.5 Obiettivi di sostenibilità a livello europeo

Lo sviluppo sostenibile costituisce un obiettivo prioritario e comunemente accettato dall'Unione Europea.

Esso viene considerato molto importante al fine di progettare e gestire il nostro modo di vivere attuale ma soprattutto al fine di garantire un futuro alle prossime generazioni.

Attuare uno sviluppo sostenibile significa salvaguardare la capacità del nostro pianeta di sostenere la vita in tutta la sua diversità, significa preservare le risorse naturali, ma significa anche considerare l'aspetto sociale e la qualità della vita, promuovendone l'uguaglianza a livello universale e infine, cosa non meno importante, rispettare la diversità culturale.

L'Europa è da molto tempo che dedica attenzione ed energie a questo tema, ma anche al tema del turismo sostenibile considerato essenziale e particolarmente influente al fine di perseguire uno sviluppo che sia il più possibile sostenibile.

Le principali conferenze europee in tema di sviluppo e turismo sostenibile che si sono susseguite nel corso degli anni possono essere riassunte nel seguente elenco, escludendo quelle già menzionate nel precedente paragrafo:

a) ***1993: Risoluzione di Bruxelles.*** Gli Stati partecipanti alla riunione si accordano sull'incontrarsi ogni due anni per discutere delle politiche e delle normative ambientali a livello europeo.
b) ***1994: prima Conferenza Europea delle Città Sostenibili.*** 600 rappresentanti di 30 Stati si riuniscono per dare inizio alla Campagna europea delle città sostenibili. Il risultato della conferenza è stato la realizzazione di una Carta, nota come Carta di Aalborg, contenente gli impegni e i programmi d'azione da realizzare e i principi che devono guidare le città europee verso la sostenibilità. Lo scopo principale della carta è stato quello di costruire un consenso generale fra i vari Stati. Il documento definisce il concetto di sostenibilità, individua le responsabilità ambientali delle città e le impegna a sviluppare un piano d'azione per andare verso lo sviluppo sostenibile delle stesse, prevedendo l'attuazione di un'agenda 21 locale.
c) ***1996: seconda Conferenza Europea delle Città Sostenibili, Lisbona.*** L'obiettivo di questa seconda conferenza è stato quello di passare dalle dichiarazioni della

Carta di Aalborg all'azione. Sono stati infatti promossi degli strumenti operativi (indicatori, strumenti di gestione ambientale, EMAS e molti altri) e socio-politici (consenso, partecipazione, cooperazione) al fine di mettere in atto quanto discusso durante la prima conferenza.

d) ***2000: terza Conferenza Europea delle Città Sostenibili, Hannover.*** Tra il 1998 e il 1999 ci sono state una serie di Conferenze Regionali (una in Finlandia, una in Bulgaria, una in Spagna, una in Olanda e infine una a Ferrara) in preparazione alla terza Conferenza Europea delle Città Sostenibili. Tale conferenza è stata organizzata per fare il punto sui risultati conseguiti e per concordare un piano d'azione comune da attuare nel corso del 21° secolo. Appoggiandosi al piano di Lisbona sono stati identificati i punti chiave della gestione urbana in ottica di sviluppo sostenibile: sviluppo urbano compatto, trasporti locali, gestione energetica, gestione efficiente del territorio e delle risorse naturali, lotta contro la povertà e l'emarginazione sociale.

e) ***2004: quarta Conferenza Europea delle Città Sostenibili, Aalborg.*** Durante questa conferenza, svoltasi nel giugno 2004, 110 comuni di 46 paesi diversi confermarono una visione condivisa per un futuro urbano sostenibile. La conferenza, chiamata Aalborgh10+ a sottolineare i 10 anni trascorsi dalla precedente conferenza di Aalborgh (1994), ha costituito un momento di riflessione sull'impegno dedicato fino ad allora nella realizzazione di azioni locali a favore della sostenibilità ed è stata sede di definizione di impegni più precisi da assumersi e di target qualitativi e quantitativi per il perseguimento dei principi di sostenibilità. Il risultato di questa riflessione è stata la stesura dei cosiddetti "Commitments Aalborgh +10", cioè un insieme di impegni condivisi finalizzati a concretizzare a livello locale la visione comune. Lo strumento dei Commitments è flessibile e adattabile alle diverse situazioni e si integra perfettamente con l'Agenda 21 Locale.

f) ***2007: quinta Conferenza Europea delle Città Sostenibili, Siviglia.*** Durante questa quinta conferenza oltre a rinnovare l'impegno verso "un'Europa più sostenibile", si sono unite agli impegni definiti nella prima carta di Aalborgh oltre 20 città europee. Infine viene fatto un appello affinché le istituzioni europee lavorino insieme per promuovere ed attuare gli obiettivi comuni.

g) ***2010: sesta Conferenza Europea delle Città sostenibili, Dunkerque (Francia).*** Durante questa conferenza si è visto come la sostenibilità locale potesse rappresentare una risposta alle sfide economiche, sociali e climatiche di allora. L'intento della conferenza, come per le altre, è stato quello di mettere al centro dello scenario europeo l'azione della sostenibilità. Inoltre ha costituito un momento di riflessione sui progressi conseguiti dai governi locali sullo sviluppo sostenibile.

h) ***2013: settima Conferenza Europea delle Città sostenibili, Ginevra.*** Durante la conferenza è stato dimostrato che con oltre metà della popolazione mondiale che vive nelle aree urbane, le città e l'urbanizzazione svolgono un ruolo fondamentale per il raggiungimento degli obiettivi ambientali, economici e sociali all'insegna della sostenibilità. È stata dunque un'occasione per individuare strumenti pratici per realizzare piani d'azione e politiche concrete. Tale conferenza, intitolata "Un'economia economicamente e socialmente responsabile: una soluzione alla crisi?", ha avuto come centro della discussione il rapporto fra amministrazioni pubbliche e finanza nell'ottica del raggiungimento di uno sviluppo sostenibile in grado di far fronte alla finanziaria ed ecologica.

Nel 2005 è stata creata un'agenda, da parte dell'UNWTO e dell'UNEP, contenente 12 obiettivi per rendere il turismo più sostenibile.

Tale strategia, denominata SDS (Strategia UE per lo sviluppo sostenibile)[5], si propone di raggiungere 3 obiettivi principali:

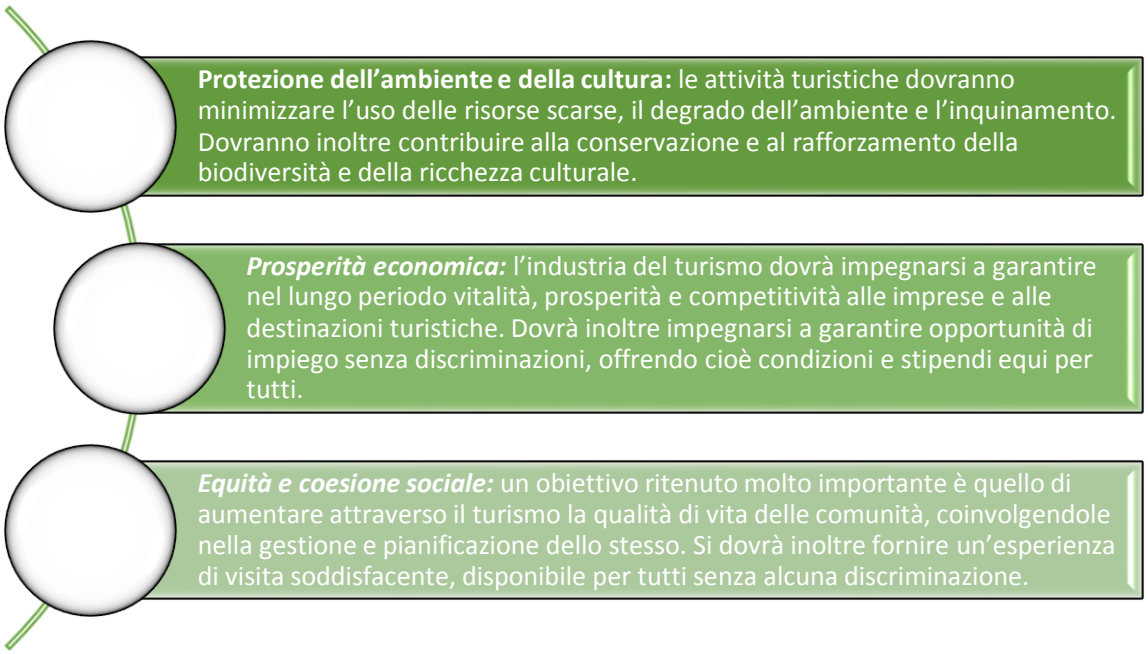


Figura 1.7 - Obiettivi principali della strategia SDS. Fonte: elaborazione personale.

1.1.6 Gli indicatori di sviluppo sostenibile per le destinazioni turistiche

Gli indicatori costituiscono uno strumento fondamentale di misura a supporto del turismo. Vengono infatti utilizzati dagli amministratori delle destinazioni turistiche come strumento di risposta alle questioni più importanti.

Essi forniscono una serie di informazioni selezionate da poter utilizzare per misurare i cambiamenti rilevanti per lo sviluppo e la gestione del turismo.

Possono misurare tre aspetti fondamentali:

1. I cambiamenti nelle strutture del turismo e nei fattori interni;
2. I cambiamenti che avvengono esternamente e che influenzano il turismo stesso;
3. Gli impatti causati dal turismo.

[5] *Making Tourism More Sustainable, A guide for policy makers* UNWTO/UNEP, 2005

Gli indicatori per essere considerati utili e rilevanti devono poter essere efficacemente usati per affrontare gli aspetti chiave della gestione e pianificazione di una meta turistica; inoltre devono essere facili da analizzare, da raccogliere e usare. I migliori indicatori, per qualsiasi meta turistica, sono quelli che rispondono in modo appropriato alle preoccupazioni e ai rischi chiave sulla sostenibilità del turismo e forniscono informazioni importanti per chiarire la situazione e misurare le risposte.

Essi normalmente forniscono informazioni che rispondono a questioni riguardanti le risorse naturali, l'ambiente di una meta turistica, le preoccupazioni relative alla sostenibilità economica, gli aspetti di gestione e organizzazione sia del turismo che della meta turistica nel suo complesso.

L'obiettivo nell'utilizzo degli indicatori è quello di ridurre i rischi futuri inerenti l'industria turistica e le destinazioni turistiche.

Alcuni vantaggi legati al loro uso sono:

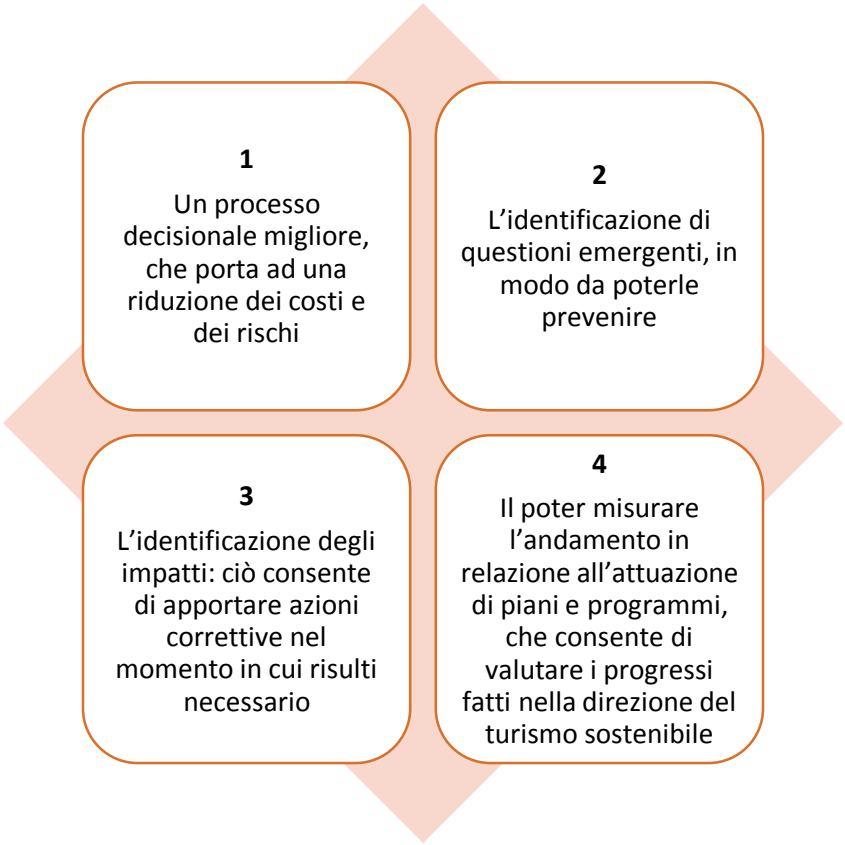


Figura 1.8 - Principali vantaggi legati all'uso degli indicatori di sviluppo sostenibile. Fonte: elaborazione personale.

Esistono diverse tipologie di indicatori, ognuno con una diversa utilità per i decisori. Per esempio ci sono indicatori:

I. Di avvertimento iniziale (es: riduzione dei turisti che intendono ritornare);
II. Di criticità sul sistema (es: carenza idrica, indici di criminalità);
III. Dello stato attuale del settore (es: soddisfazione del turista);
IV. Dell'impatto del turismo sul sistema naturale e socioeconomico (es: livello di deforestazione, cambiamenti nei modelli di consumo);
V. Misure dei risultati della gestione (es: risultati nella riduzione dell'inquinamento).

Oltre a questi ce ne sono molti altri, i quali vengono utilizzati per supportare i processi decisionali ai diversi livelli di pianificazione e gestione del turismo (nazionale, regionale, specifica destinazione turistica, singola struttura turistica, siti turistici ad elevata fruizione, grandi aziende turistiche).
Molti indicatori si riferiscono a questioni di sostenibilità comuni alla maggior parte delle destinazioni turistiche. È necessario però considerare che gli indicatori possono e devono essere adattati alla specifica destinazione turistica, alle sue particolari condizioni e devono essere costantemente rivisti e aggiornati allo scopo di garantire il perseguimento degli obiettivi e incoraggiare continui miglioramenti.

Nel 2004 è stata creata da parte della WTO (Organizzazione Mondiale del Turismo) una guida[6] che fornisce gli elementi di base e i riferimenti per sviluppare indicatori che aiutino a rispondere efficacemente alle questioni e sfide politiche e gestionali di qualsiasi meta turistica.

[6] http://www.turismosostenibile.provincia.rimini.it/documenti/archivio/Guida%20Indicatori%20UNWTO/Ita/Guida%20Indicatori%20UNWTO.pdf

Tale guida fornisce:

1. **Informazioni su come sviluppare gli indicatori nel contesto dei processi di pianificazione turistica;**
2. **Un'analisi completa degli aspetti ambientali, socioeconomici e gestionali che possono riguardare le destinazioni turistiche;**
3. **Una lista di indicatori e temi comuni ad alcune tipologie di destinazioni (piccole isole, aree naturali, siti culturali);**
4. **Un aiuto su come applicare gli indicatori nella pianificazione e nelle politiche del turismo.**
5. **Una serie di esempi applicativi ad alcune destinazioni turistiche.**

1.2 Agriturismi e sostenibilità ambientale

1.2.1 L'attività agrituristica

L'agriturismo consiste in una forma di turismo extraurbano riservato agli imprenditori agricoli.

L'attività agrituristica è nata come risposta alla sempre meno redditizia attività agricola durante gli anni 80: si ebbe infatti l'idea di utilizzare gli spazi ambientali per la ricezione turistica, nel rispetto dei principi di sostenibilità, in modo da realizzare un reddito aggiuntivo di supporto alla classica e sola attività di produzione agricola.

Figura 1.9 - Alcuni prodotti tipici di un agriturismo.

Un agriturismo infatti permette di produrre un certo reddito nel rispetto della salvaguardia del patrimonio ambientale e culturale delle campagne. Da ora in poi l'attività agricola viene dunque considerata, oltre per la sua funzione produttiva, anche per la sua capacità di fornire beni e servizi secondari.

Agriturismo è sinonimo di turismo rurale, cioè di quell'insieme di attività turistiche svolte in ambiente rurale.

Il turismo rurale è legato principalmente alle strutture agrituristiche e in Italia risulta essere un fenomeno in forte sviluppo. Basta guardare i dati degli ultimi 10 anni: tra il 2003 e il 2013 si è verificato un +57% di aumento del numero di agriturismi. In Italia la maggior concentrazione di agriturismi si registra al Nord (48% sul totale), seguono poi il Centro (35%) ed infine il Sud (19%) del paese (Coldirettti, 2013). Trentino Alto Adige e Toscana sono le regioni leader come presenza di agriturismi, questo dovuto al fatto che l'ospitalità agricola risulta più radicata per ragioni storiche.

Per agriturismo si intende qualsiasi attività di ospitalità, di ricezione, messa in atto da imprenditori agricoli attraverso l'utilizzo delle proprie aziende agricole. È da sottolineare che l'attività agrituristica è un'attività strettamente connessa a quella agricola e quindi può esistere soltanto in presenza di un'azienda agricola e quindi di un imprenditore agricolo che decide di adibire la propria azienda all'accoglienza di turisti.

L'agricoltura trae un forte vantaggio dall'attività agrituristica, consistente nell'aumento delle sue entrate. Ma i vantaggi non sono solo per l'azienda agricola dell'imprenditore ma anche per il turista: esso infatti trae vantaggio economico poiché i prezzi di un agriturismo sono solitamente più bassi ed accessibili rispetto ad una struttura ricettiva come un albergo. È necessario però tenere presente che l'offerta agrituristica può essere molto varia: ci si può infatti trovare davanti a semplici alloggi in fattorie isolate in mezzo alla campagna, come davanti a ville ben curate dove i prezzi sono ovviamente più alti.
L'agriturismo dunque è una struttura ricettiva rurale, adibita cioè all'interno di fabbricati rurali, nei quali possono alloggiare i turisti. Non possono essere utilizzati edifici non strettamente legati all'azienda agricola e ubicati in un luogo diverso dal fondo in cui si vuole esercitare l'attività di agriturismo. Non è poi possibile utilizzare costruzioni ex novo ma soltanto abitazioni ristrutturate e recuperate.

Il termine "agriturismo" viene utilizzato per indicare le seguenti tipologie di aziende agricole:

- *Quelle che offrono esclusivamente alloggio*;
- *Quelle che offrono alloggio e ristorazione*;
- *Quelle che oltre all'alloggio e alla ristorazione offrono altre attività legate al sociale, all'ambiente, allo sport, al gioco.*

L'agriturismo differisce dalla formula del "*bed and breakfast*" poiché quest'ultimo offre un servizio di ricettività nella propria abitazione con una disponibilità massima di tre camere, ognuna con una capacità massima di quattro posti letto.
In Italia l'attività agrituristica è regolamentata dalla legge 96/06, il cui articolo 2 ne definisce il contesto: "Le attività agrituristiche consistono in attività di ricezione e ospitalità esercitate dagli imprenditori agricoli di cui all'art. 2135 del Codice Civile, anche nella forma di società di capitali o di persone, oppure associati fra loro, attraverso l'utilizzazione della propria azienda in rapporto di connessione con le attività di coltivazione del fondo, di silvicoltura e di allevamento di animali" (l. 96/06, Codice Civile).

Per quanto riguarda le attività tipiche di un agriturismo possiamo identificarne quattro:

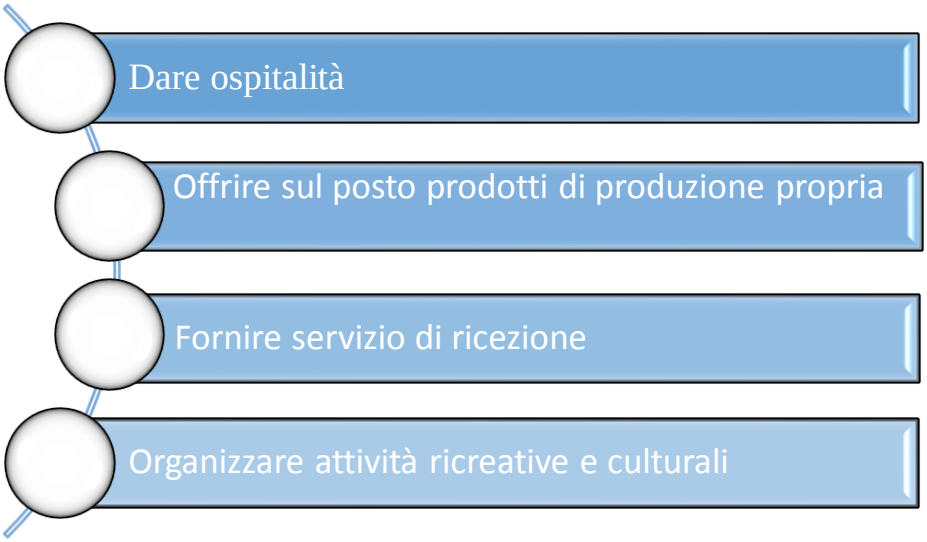


Figura 1.10 - Attività tipiche di un agriturismo. Fonte: elaborazione personale.

Sono considerati prodotti di produzione propria i cibi e le bevande prodotti e lavorati all'interno dell'azienda agricola, ma anche quei prodotti ricavati da materie prime dell'azienda lavorati esternamente.

Le attività ricreative devono essere connesse e complementari all'agricoltura. Fra queste possiamo comprendere:

Figura 1.11 - Esempi di attività ricreative di un agriturismo. Fonte: elaborazione personale.

1.2.2 L'agri-turista

Il soggiorno in agriturismo consente al turista di avere un contatto con l'ambiente circostante con in più la possibilità di partecipare alla vita agricola, svolgendo alcuni lavori in fattoria e avendo cosi la possibilità di conoscere una realtà diversa e molto spesso ignorata da chi vive in città.

Molto spesso consiste proprio in questo la motivazione per cui il turista richiede di fare una vacanza in agriturismo. Il suo desiderio è quello di riscoprire la cultura agricola, la cultura rurale, che ormai è diventata quasi un miraggio in una società sempre più consumistica e raccolta all'interno dei grossi centri urbani.

Voglia di nuove esperienze come motivazione all'agriturismo, ma oltre a questo si può individuare la curiosità e l'interesse ad avventurarsi in zone di campagna allo scopo di vivere esperienze insolite e difficilmente raggiungibili all'interno dei centri urbani, come per esempio il contatto con la tradizione contadina e i frutti della terra.

C'è poi la ricerca di un luogo tranquillo dove passare qualche giornata e non da meno la ricerca del benessere e della qualità della vita, nutrendosi in modo sano consumando prodotti genuini che vanno dai prodotti DOP, agli IGP, DOC e DOCG.

Infine quello che spinge il turista a soggiornare in un agriturismo è anche la ricerca di un buon rapporto fra costi e servizi offerti.

1.2.3 Gli impatti ambientali di un agriturismo

Considerando che il settore agrituristico è in continua crescita è facile ipotizzare che anche il corrispondente impatto sull'ambiente sia in costante crescita.

Per quanto riguarda gli impatti di un agriturismo sull'ambiente si possono essenzialmente distinguere due tipologie:

Quelli derivanti dalle attività svolte nelle strutture ricettive

Quelli derivanti dalle presenze turistiche

Riguardo alla prima tipologia di impatto, legata agli input e agli output del processo produttivo dell'azienda agrituristica, vengono elencati nella seguente tabella quelli tipici di una struttura che offre servizio di ristorazione e di pernottamento, vendita di prodotti tipici ottenuti dall'attività stessa dell'azienda (come formaggi, vini, salumi) e attività ricreative.

Impatti tipici di un'azienda agrituristica		
Attività	**Input**	**Output**
Pernottamento	Detergenti e disinfettanti, acqua, imballaggi, energia elettrica, combustibili	Scarichi idrici, emissioni in atmosfera, rifiuti non pericolosi, imballaggi
Ristorazione e preparazione prodotti tipici dell'azienda	Prodotti alimentari, detergenti e disinfettanti, acqua, combustibili, energia elettrica, carta, imballaggi	Rumore, imballaggi, rifiuti non pericolosi, odori, scarichi idrici, emissioni in atmosfera
Gestione ed amministrazione	Toner, combustibili, acqua, carta e cancelleria, energia elettrica	Rifiuti pericolosi e non pericolosi, emissioni in atmosfera, scarichi idrici, imballaggi
Attività ricreative	Acqua, energia elettrica, combustibili	Rifiuti non pericolosi, rumore, imballaggi, scarichi idrici

Tabella 1.1 - Impatti tipici di un'azienda agrituristica. Fonte: elaborazione personale.

1.2.4 Verso un agriturismo sempre più "ecologico"

Nel corso del 2012 l'Osservatorio nazionale del turismo ha condotto un sondaggio sulla "Customer care" dei turisti in Italia, facendo emergere un importante risultato dalle risposte date al questionario: la sensibilità ecologica degli ospiti è in continua crescita, soprattutto quando la destinazione di vacanza è un luogo di rilevante interesse naturalistico.

Tale aspetto risulta molto importante per le aziende agrituristiche le quali, attuando precise scelte organizzative e iniziative concrete, possono cogliere le opportunità di una domanda turistica che dimostra sempre più interesse e importanza in tema di riduzione degli sprechi e difesa dell'ambiente.

Il sondaggio ha evidenziato che il rispetto dell'ambiente e il contenimento dell'inquinamento costituisce il quarto aspetto considerato importante dai turisti per il gradimento del soggiorno (23,3 %), preceduto dalla simpatia e ospitalità delle persone (il 55,6%), dalla qualità della ristorazione (il 45,7%) e dall'offerta di attività per l'intrattenimento (il 30,6%).

Un agriturismo è considerato di per se una proposta di vacanza naturale dal turista, per cui può cogliere questa grande opportunità rispondendo in modo adeguato alle richieste e aspettative ecologiche dei suoi ospiti.

Tutto questo lo può fare semplicemente mettendo in atto determinate scelte organizzative e iniziative concrete, in modo coerente e visibile, e soprattutto informando in modo adeguato i turisti delle proprie scelte ecologiche.

Dunque, quali sono le possibili azioni da mettere in atto per un agriturismo più "ecologico"?

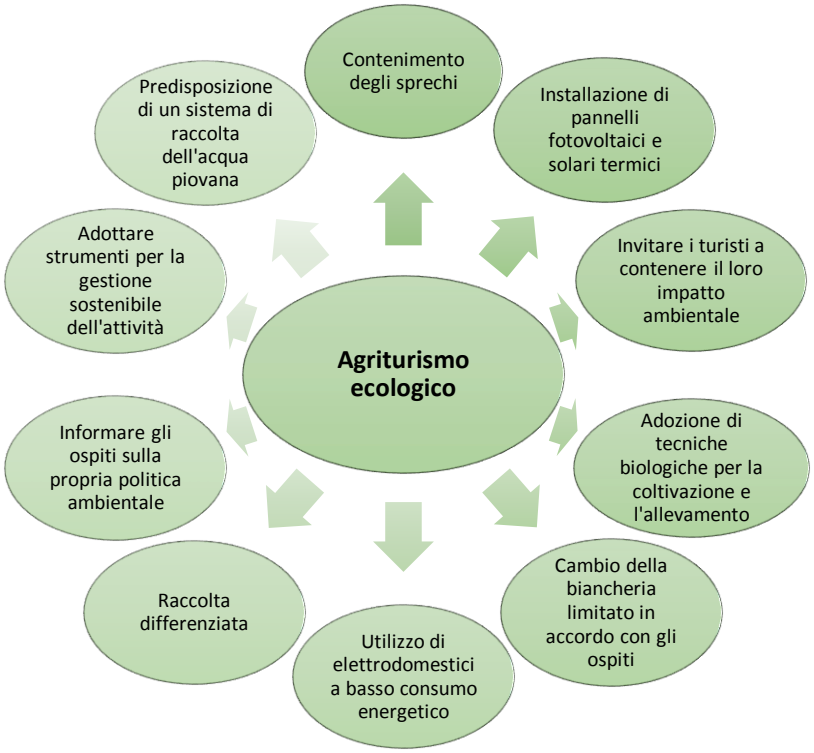


Figura 1.12 - Azioni per un agriturismo più ecologico. Fonte: elaborazione personale.

Il contenimento degli sprechi, o per meglio dire il contrasto agli sprechi, è un primo impegno che un agriturismo ecologico deve prendere con serietà.

Si tratta di fare tutto il possibile per limitare l'uso "non corretto" di energia elettrica e acqua. Una prima azione da adottare consiste nel dare agli ospiti precise informazioni su come evitare gli sprechi.

Le soluzioni che possono essere adottate sono svariate. Per esempio, si potrebbe mettere accanto ad ogni interruttore per l'illuminazione dell'alloggio un cartellino o un quadretto con scritto: "Accendi la luce solo quando serve!". Oppure vicino al lavandino un cartellino con scritto: "Fai scorrere l'acqua moderatamente!". In tal senso la stessa azienda dovrebbe procedere all'installazione di riduttori del flusso d'acqua e magari predisporre al WC un sistema di doppio scarico (poca o tanta acqua) con la possibilità di fermare l'erogazione dell'acqua nel caso serva. Per l'illuminazione invece è d'obbligo dotare tutte le stanze e i locali dell'azienda di lampadine a basso consumo energetico.

Continuando, vicino alle finestre un quadretto con scritto: "Ricordati di chiudere la finestra se il riscaldamento o il condizionatore è acceso". Poi vicino al sistema di

riscaldamento e di condizionamento un altro quadretto con scritto: “Chiudi le finestre quando accendi il riscaldamento o l’aria condizionata!”. E così via.
Un’altra cosa molto importante è applicare all’ingresso dell’agriturismo e di ogni alloggio un quadro in cui inserire tutte le informazioni su come contenere gli sprechi e perché no, informare gli ospiti sulle misure adottate dalla struttura per la salvaguardia dell’ambiente e i risultati che si possono ottenere adottando un comportamento eco-compatibile. Inoltre questa stessa informativa dovrà essere contenuta all’interno della carta dei servizi messa a disposizione degli ospiti di ogni alloggio.
L’installazione di pannelli fotovoltaici per la produzione di energia elettrica e di impianti termodinamici per il riscaldamento rappresenta una testimonianza visibile dell’impegno da parte della struttura a favore dell’ambiente. Inoltre consente un notevole risparmio nel corso del tempo per lo stesso agriturismo, considerando che l’installazione di questi impianti è agevolata dagli incentivi statali.
Tali impianti possono essere posizionati sopra il tetto delle strutture, oppure è possibile sfruttare la copertura per il parcheggio riservato ai clienti dell’agriturismo.

Figura 1.13 – Agriturismo ecologico. Fonte: www.soluzionifotovoltaico.it

Per chi ne avesse la possibilità, e in relazione ovviamente al luogo in cui è ubicato l’agriturismo, può optare anche per l’installazione di un mini impianto eolico, in modo da incrementare la produzione di energia.
Si può inoltre installare un impianto di trattamento delle biomasse allo scopo di produrre energia utilizzando la frazione biologica dei rifiuti e i residui di origine biologica provenienti dall’attività dell’azienda agrituristica.

Un’altra azione molto importante e soprattutto semplice da mettere in atto è quella di invitare gli ospiti della struttura a contenere l’impatto ambientale derivante dalla loro presenza.

Quindi si tratta di invitarli a limitare l'uso della macchina mettendogli a disposizione biciclette, proponendogli escursioni a piedi o a cavallo, ma soprattutto fornendogli tutte le indicazioni necessarie per muoversi con i mezzi pubblici. Queste informazioni, oltre ad indicazioni su mete vicine e interessanti da raggiungere in bicicletta o a piedi, possono essere integrate all'interno della carta servizi dell'azienda agrituristica.

Vi è poi l'adozione di tecniche biologiche per la coltivazione e l'allevamento che permettono all'azienda di qualificare i propri prodotti e il proprio servizio di ristorazione, offrendo prodotti buoni e salutari. Per fertilizzare il terreno si può ricorrere per esempio al compostaggio effettuato attraverso l'installazione di un apposito contenitore (composter) dove far fermentare la frazione organica dei rifiuti dell'agriturismo. Si ottiene così un ottimo compost organico da poter spargere sul terreno.

Il cambio della biancheria può essere limitato, in accordo con gli ospiti, a quanto veramente necessario. In questo modo si riduce il consumo dell'acqua e l'inquinamento da detergenti.

L'utilizzo di elettrodomestici a basso consumo energetico (classe A o superiore) è un requisito fondamentale per un agriturismo che vuole qualificarsi come "ecologico".

Risulta poi importante mettere a disposizione negli alloggi e fuori dall'agriturismo dei contenitori per la raccolta differenziata dei rifiuti, applicandoci sopra un cartellino con un elenco dei principali rifiuti conferibili in ogni tipologia di contenitore.

Predisporre un sistema di raccolta dell'acqua piovana consente di risparmiare centinaia di litri di acqua potabile. Essa infatti può essere utilizzata per l'irrigazione delle piante, ma anche come acqua per gli scarichi dei WC.

Infine molto importante è l'adozione di strumenti per la gestione sostenibile dell'attività. Fra questi possiamo individuare:

1 *Gli acquisti verdi*

Acquistare prodotti sostenibili e scegliere i fornitori sulla base di criteri ecologici permette di ridurre gli impatti ambientali derivanti dall'attività turistica, diretta conseguenza del consumo di prodotti. Ovviamente più grande sarà l'attività (ad esempio, un albergo) maggiori saranno i benefici in termini ambientali che si potranno apportare seguendo alcuni semplici accorgimenti.

Non esistono prodotti che siano sostenibili al 100%. Al fine di scegliere quelli più ecologici è necessario conoscere gli aspetti ambientali collegati al loro ciclo di vita. Ovviamente una buona fonte di informazione che può garantirci che il prodotto che stiamo scegliendo è un prodotto di ottima qualità e allo stesso tempo con un ridotto impatto ambientale è rappresentata dai fornitori di fiducia, dalle certificazioni ambientali e dai marchi.

Di seguito vengono indicate alcune linee guida su come scegliere i prodotti e i servizi seguendo criteri ecologici.

2 *I sistemi di gestione ambientale*

Un Sistema di Gestione Ambientale, SGA, consiste nella modalità di gestione di un organizzazione con cui il titolare realizza una politica ambientale allo scopo di garantire nel tempo il controllo e il miglioramento delle prestazioni ambientali della sua attività. Tale sistema può essere applicato a tutte le organizzazioni e alle strutture ricettive.

Nel presente lavoro non verranno trattati il regolamento, la struttura e le fasi di attuazione di un sistema di gestione ambientale, le quali possono essere sicuramente oggetto di un lavoro a parte. Per un approfondimento è possibile consultare la norma ISO 14001 e il Regolamento EMAS.

In breve, fra i benefici principali di un SGA possiamo comprendere:

1. La facoltà di identificare, prevedere, analizzare, controllare e prevenire gli effetti ambientali;
2. La capacità dell'impresa di svolgere in modo responsabile la propria attività attraverso modalità che rispettino l'ambiente;
3. La possibilità di modificare e aggiornare in modo continuo l'organizzazione, migliorando le prestazioni ambientali in relazione ai cambiamenti dei fattori esterni ed interni;
4. La capacità di interagire con tutti i soggetti esterni interessati alle prestazioni ambientali dell'impresa;
5. La facoltà di attivare, valorizzare e motivare l'iniziativa di tutti gli attori interni all'organizzazione.

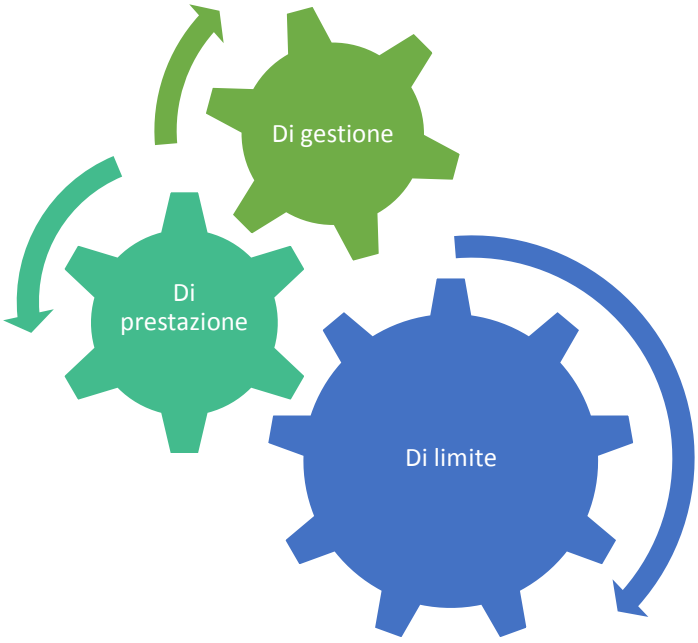
L'implementazione di un Sistema di Gestione Ambientale per un'impresa turistica può creare notevoli vantaggi competitivi, soprattutto nei confronti dei concorrenti meno dinamici. Può infatti valorizzare l'immagine dell'azienda, rispondere alla sempre maggiore domanda ambientale del turista, contribuire alla tutela ambientale della località turistica, sensibilizzare gli ospiti ma anche migliorare la gestione dell'attività riducendone i costi.

3 *I marchi ecologici*

I marchi ecologici applicati al turismo rappresentano un fenomeno in forte espansione. Sono sempre di più le aziende turistiche che decidono di dotarsi di qualche certificazione ambientale, allo scopo di concretizzare il loro impegno nei confronti della salvaguardia

dell'ambiente, per sensibilizzare i turisti ma anche per l'esigenza di valorizzarsi e differenziarsi sul mercato.
In Europa gli esempi più significativi di etichette ecologiche sono quelli rappresentati dalle etichette comprese negli standard determinati dal progetto europeo *"VISIT"* (Volountary Iniziatives for Sustainability in Tourism) e primo fra tutti, l'Ecolabel.

In breve tutti i marchi ecologici del settore turistico che sono stati censiti dal World Tourism Organisation (WTO) si basano sui seguenti tipi di criteri:



I criteri di gestione riguardano la formazione del personale, informazione alla clientela, il monitoraggio dei consumi e delle emissioni e molti altri.
I criteri di prestazione comprendono ad esempio l'installazione di riduttori di flusso per il risparmio dell'acqua, l'utilizzo di lampadine a basso consumo energetico e l'utilizzo di fonti energetiche rinnovabili.
Infine i criteri di limite comprendono lo stabilire il divieto di utilizzare disinfettanti, l'applicare un limite massimo di flusso di acqua alle docce, e molti altri.
Nel prossimo paragrafo verranno analizzate le tre principali certificazioni ambientali, l'ISO 14001, l'EMAS e il marchio ECOLABEL, cercando di fornire un quadro breve ma completo delle loro caratteristiche.

1.2.5 Le principali certificazioni ambientali

In Europa, ma soprattutto in Italia, il turismo rurale legato in particolar modo alle strutture agrituristiche, rappresenta un fenomeno in continuo sviluppo.

Per favorirlo è necessario fornire servizi di qualità e rispettosi dell'ambiente, in sintonia con la sempre maggior attenzione dei turisti verso quest'ultimo. La produzione di servizi di qualità da parte delle strutture agrituristiche passa attraverso la valorizzazione del territorio rurale in cui esse operano. Per tutelarlo è necessario attuare strumenti specifici che consentono di ridurre il più possibile gli impatti delle attività economiche, in questo caso quelle agrituristiche, e promuovere un turismo più sostenibile attraverso una costante attività di comunicazione e informazione.

Fra gli strumenti che consentono di raggiungere tali obiettivi ci sono le certificazioni ambientali ISO 14001 ed EMAS, molto simili tra loro, e il marchio europeo di qualità ecologica ECOLABEL.

Le prime due sono applicabili in modo volontario alle organizzazioni e consentono di ottenere una certificazione riferibile alle modalità e ai processi di gestione degli aspetti ambientali da parte delle organizzazioni. Il marchio ECOLABEL è invece applicabile ai prodotti e ai servizi e ha la funzione di attestare l'eccellenza ambientale di prodotti e servizi in base alla loro capacità di soddisfare determinati requisiti di prestazione ambientale.

Ma in cosa consiste una certificazione? E chi si occupa della sua emissione?

La certificazione è un atto volontario il cui ottenimento rappresenta un indice di eccellenza ed è intesa come l'atto mediante il quale una terza parte indipendente, come un ente di certificazione, dichiara che un prodotto, un servizio o un sistema di gestione è conforme a determinati requisiti stabiliti da norme specifiche. Un Ente di Certificazione per poter svolgere la sua funzione deve prima essere accreditato, cioè deve ricevere l'attestazione e il riconoscimento, da parte di un organismo nazionale o internazionale di accreditamento, di integrità, indipendenza (nessun conflitto di interesse con l'organizzazione da certificare), competenza e mezzi e strutture per assolvere a determinati compiti stabiliti da precise norme. In Italia l'unico ente di accreditamento è ACCREDIA.

Tutto questo sistema si basa sul "Mutuo Riconoscimento", cioè un accordo fra tutti gli Enti mondiali di Accreditamento per il riconoscimento delle procedure di accreditamento e degli organismi competenti.
In Europa il sistema del mutuo riconoscimento è retto dall'EA (European co-operation for Accreditation), formato da tutti gli enti nazionali di accreditamento con lo scopo di riconoscere l'affidabilità e l'equivalenza delle certificazioni rilasciate a livello europeo.

Figura 1.14 - Sistema di certificazione e accreditamento.

Verranno ora analizzate, senza entrare troppo nei dettagli, le principali certificazioni ambientali riconosciute a livello internazionale.

1.2.5.1 Norma ISO 14001

La norma ISO 14001 è uno strumento internazionale volontario che può essere applicato a tutte le imprese.
La norma specifica i requisiti di un sistema di gestione ambientale (SGA) per le imprese e i siti di produzione, viene rilasciato da un organismo indipendente accreditato che verifica il concreto impegno da parte dell'organizzazione in esame nel minimizzare l'impatto ambientale dei processi, prodotti e servizi.
All'impresa, prodotto o servizio che soddisfa i requisiti della norma viene rilasciato il marchio ISO 14001 a garanzia dell'affidabilità del sistema di gestione ambientale applicato.
Realizzare e mantenere attivo un SGA per l'organizzazione significa stabilire una propria politica ambientale, fissare degli obiettivi di miglioramento e le modalità per perseguirli; inoltre permette di gestire in modo appropriato tutti gli aspetti ambientali collegati alle proprie attività.

La norma contiene requisiti generali e riconducibili a livello gestionale al modello basato sul ciclo di Deming[7] del miglioramento continuo, il quale è composto dalle seguenti fasi:

1. ***Pianificazione:*** l'organizzazione stabilisce gli obiettivi ambientali e pianifica i processi necessari al raggiungimento dei risultati individuati dall'organizzazione stessa nella sua politica ambientale;
2. ***Attuazione e funzionamento:*** mediante adeguate procedure l'organizzazione mette in atto i processi stabiliti;
3. ***Controllo:*** l'organizzazione deve controllare costantemente i propri processi in modo da monitorare i risultati raggiunti nel tempo e consentire interventi di miglioramento;
4. ***Attuazione del miglioramento:*** l'organizzazione cerca di migliorare il proprio SGA attraverso una serie di interventi.

Figura 1.15 - Modello di una sistema di gestione ambientale.
Fonte: www.greensga.it/Sistemi/ConcettiGenerali.

L'organizzazione che vuole ottenere il marchio ISO 14001 fa richiesta ad un organismo di certificazione il quale verifica la completezza e la conformità dei documenti ai requisiti legislativi e procede poi con l'analisi delle informazioni rilevanti dell'attività dell'organizzazione. Per il mantenimento della certificazione nel tempo vengono fatte ispezioni a cadenza annuale fino al terzo anno quando il sistema viene revisionato in modo completo attraverso la "Verifica di rinnovo della certificazione".

[7] http://www.aicqer.it/download/11_La%20filosofia%20di%20Deming%20e%20il%20ciclo%20PDCA.pdf

Nel febbraio 2012 la norma è stata revisionata: fra le principali novità vi è l'adozione di un testo comune per tutte le norme dei sistemi di gestione ISO con lo scopo di facilitare l'integrazione degli aspetti legati all'ambiente, alla sicurezza e alla qualità delle imprese. La nuova norma verrà pubblicata nel 2015 e avrà l'obiettivo di supportare le organizzazioni ad affrontare le sfide ambientali future.

Di seguito verrà illustrata a scopo informativo attraverso una serie di grafici la situazione attuale delle certificazioni ISO 14001 a livello italiano, con particolare attenzione al settore alberghiero e agrituristico. I dati sono riferiti a giugno 2012. Purtroppo non è stato possibile reperire dati aggiornati riferiti ai soli agriturismi in Veneto.

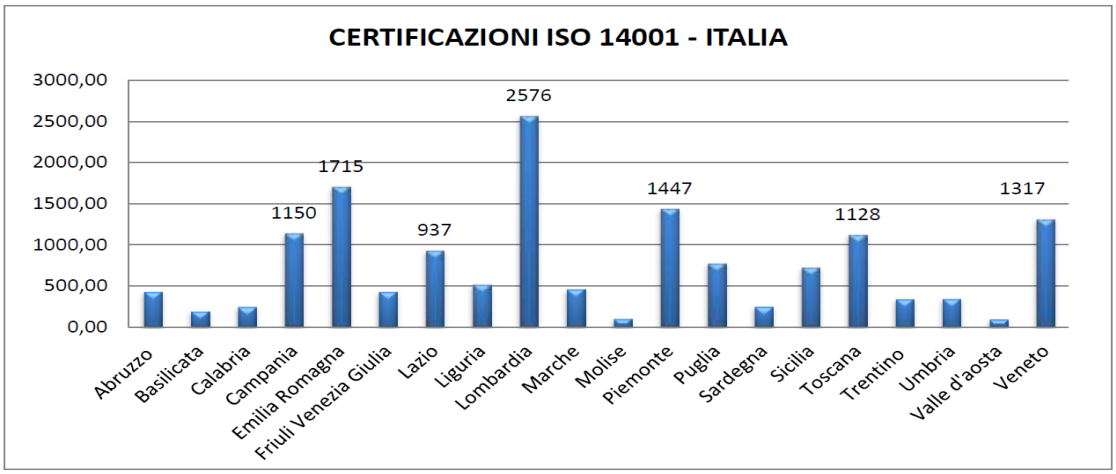


Figura 1.16 - Numero di certificazioni ISO 14001 in Italia suddivise per regione. Fonte: banca dati Acedia 12.06.2012

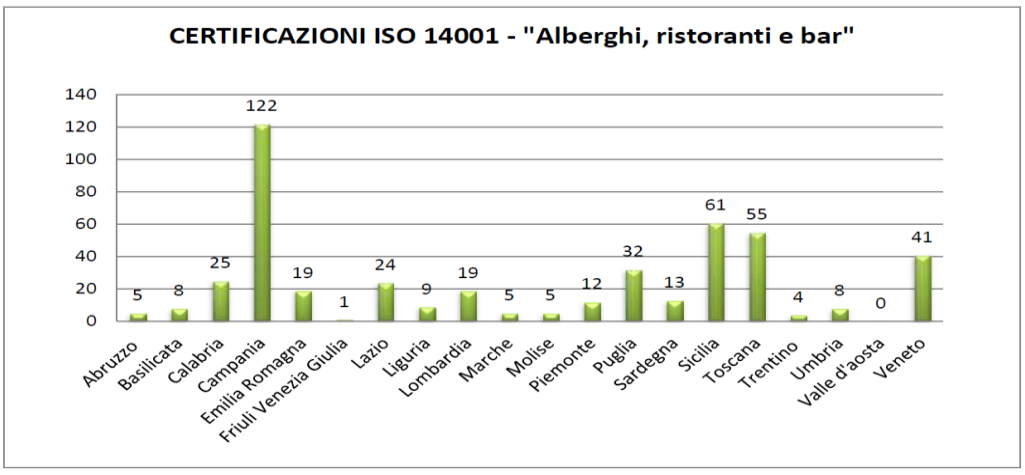


Figura 1.17 - Numero di certificazioni ISO 14001 in Italia nella categoria "Alberghi, agriturismi, ristoranti e bar". Fonte banca dati Accredia 12.06.2012

Delle 15.254 organizzazioni italiane operanti nell'ambito turistico, 447 (il 3%) appartengono al settore "Alberghi, agriturismi, ristoranti e bar". Il Veneto risulta in quarta posizione con 41 certificazioni, delle quali 29 rientrano nella categoria "Attività turistico ricettive" (Agriturismi, campeggi).

1.2.5.2 Emas

Figura 1.18 - Logo EMAS

EMAS è il sistema di Ecogestione ed audit della Comunità europea (Echo-Managemet and Audit-Scheme).

Il riferimento legislativo è nel regolamento comunitario 1221/2009 entrato in vigore nel 2010, il quale va a sostituire il regolamento 761 del 2001.

Rappresenta uno strumento al quale possono aderire volontariamente tutte le organizzazioni, sia quelle all'interno della Comunità europea sia quelle al di fuori di essa, e che certifica il raggiungimento di risultati di eccellenza nel miglioramento ambientale.

Come l'ISO 14001, anche questo strumento certifica il processo di produzione di un bene oppure la fornitura di un servizio.

EMAS mira a promuovere il miglioramento continuo delle prestazioni ambientali di qualsiasi tipo di organizzazione attraverso:

- ✓ L'analisi ambientale di ogni aspetto della loro attività;
- ✓ L'istituzione e l'applicazione di sistemi di gestione ambientale (SGA);
- ✓ La valutazione periodica di tali sistemi;
- ✓ Un dialogo aperto con tutte le parti interessate;
- ✓ La formazione e il coinvolgimento costante del personale;
- ✓ L'offerta di informazioni sulle prestazioni ambientali.

Dunque anche il certificato EMAS prevede per le organizzazioni l'implementazione di un Sistema di Gestione Ambientale, rispondente ai requisiti della norma ISO 14001.

EMAS prevede però alcuni elementi aggiuntivi che lo differenziano dalla norma ISO 14001. Per ottenere tale certificazione le organizzazioni devono infatti predisporre il documento di Dichiarazione Ambientale, strumento di informazione generale e comunicazione al pubblico e a tutte le parti interessate alle attività dell'organizzazione, alla sua politica ambientale e al suo sistema di gestione ambientale, agli impatti ambientali derivanti dalla sua attività, al programma, obiettivi e traguardi ambientali da raggiungere, al suo rispetto degli obblighi legislativi in tema di ambiente.
Nella seguente tabella vengono messe a confronto le principali caratteristiche che differenziano la certificazione EMAS dalla norma internazionale ISO 14001.

ISO 14001	EMAS
Convalida da parte di un organismo privato	Convalida da parte di un ente pubblico
Validità internazionale	Validità Europea
Nessuna dichiarazione (consigliata una strategia di comunicazione)	Dichiarazione ambientale
Rispetto delle leggi richiesto indirettamente	Rispetto delle leggi richiesto esplicitamente

Tabella 1.2 - Principali differenze fra EMAS e ISO 14001. Fonte: http://www.keisdata.it

Un ulteriore differenza con la norma di ISO 14001 è l'utilizzo del logo EMAS da parte dell'organizzazione una volta che ha ottenuto la certificazione. Dopo la registrazione l'organizzazione deve svolgere annualmente un audit interno allo scopo di verificare costantemente le prestazioni ambientali e il rispetto degli obblighi legislativi e procede con l'aggiornamento della dichiarazione ambientale. Il rinnovo della registrazione avviene, come nel caso dell'ISO 14001, ogni tre anni.

1.2.5.3 Il marchio europeo ECOLABEL

Figura 1.19 - Logo Ecolabel

L'Ecolabel, istituito per la prima volta con il regolamento 880 del 1992 e aggiornato poi nel 2000 con il regolamento 1980, è il marchio europeo di qualità ecologica.

Similmente ad EMAS è uno strumento ad adesione volontaria emanato a livello europeo, riconosciuto a quei prodotti e servizi che rispettano determinati criteri ecologici e prestazionali. Il marchio Ecolabel rappresenta un attestato d'eccellenza, rilasciato solo a quei prodotti e servizi che per tutto il loro ciclo di vita mantengono un ridotto impatto ambientale e garantiscono almeno le stesse prestazioni dei prodotti convenzionali.

Il marchio Ecolabel si basa sull'analisi del ciclo di vita del prodotto, vengono infatti analizzati tutti gli aspetti relativi al suo intero ciclo di vita: produzione, utilizzo e smaltimento. Una volta definiti tutti gli aspetti importanti di una categoria di prodotto i criteri vengono adottati a livello europeo su decisione della Commissione Europea. Dopo la loro stesura sono resi pubblici tramite la Gazzetta Ufficiale della Comunità europea e vengono costantemente revisionati per renderli sempre più selettivi.

Con la revisione del 2000 il marchio Ecolabel è stato esteso anche ai servizi. La categoria "ricettività turistica" è stata la prima per la quale sono stati definiti i criteri poiché possiede le caratteristiche che il regolamento richiede al fine che un prodotto/servizio possa ottenere il marchio, cioè:

1. Rappresenti un volume significativo di vendite e scambi nel mercato europeo;
2. Comporti impatti ambientali significativi;
3. Le scelte del consumatore abbiano una significativa capacità potenziale di indurre miglioramenti ambientali.

Possono ottenere il marchio Ecolabel tutti i servizi di ricettività turistica che abbiano come servizio base il pernottamento (Agriturismi, alberghi, B&B)
Per ottenere l'Ecolabel è necessario il rispetto di 37 criteri obbligatori e un certo numero di criteri facoltativi (in totale sono 47) tali da raggiungere un punteggio minimo di 16,5 punti nel caso in cui la struttura eroghi solo il servizio di pernottamento. Se invece la struttura offre altri servizi come la ristorazione, attività ricreative e molte altre, si dovrà aggiungere per ogni servizio offerto un punto, al punteggio minimo da raggiungere con il rispetto dei criteri facoltativi.

I vantaggi che si possono ottenere attraverso l'ottenimento del marchio Ecolabel sono numerosi. Rispetto alle diverse parti coinvolte possiamo per esempio individuare:

1. ***Vantaggi per gli operatori del settore:*** riduzione del consumo di risorse naturali e conseguentemente riduzione dei costi dell'attività, miglioramento delle azioni ambientali già messe in atto, riconoscimento a livello europeo della struttura ricettiva, acquisizione di una valore aggiunto da parte della struttura poiché il marchio Ecolabel è garanzia in tutta Europa di elevate prestazioni ambientali;
2. ***Vantaggi per i turisti:*** garanzia di trovare lo stesso comfort e gli stessi servizi offerti dalle altre strutture della stessa tipologia, maggiore attenzione per l'ambiente e la salute;
3. ***Vantaggi per l'ambiente:*** una struttura eco-certificata garantisce l'uso di prodotti chimici meno dannosi, ridotte emissioni, minor consumo di risorse, minore produzione di rifiuti e imballaggi.

Schematicamente, il procedimento di richiesta e registrazione del marchio Ecolabel con le rispettive parti coinvolte è il seguente:

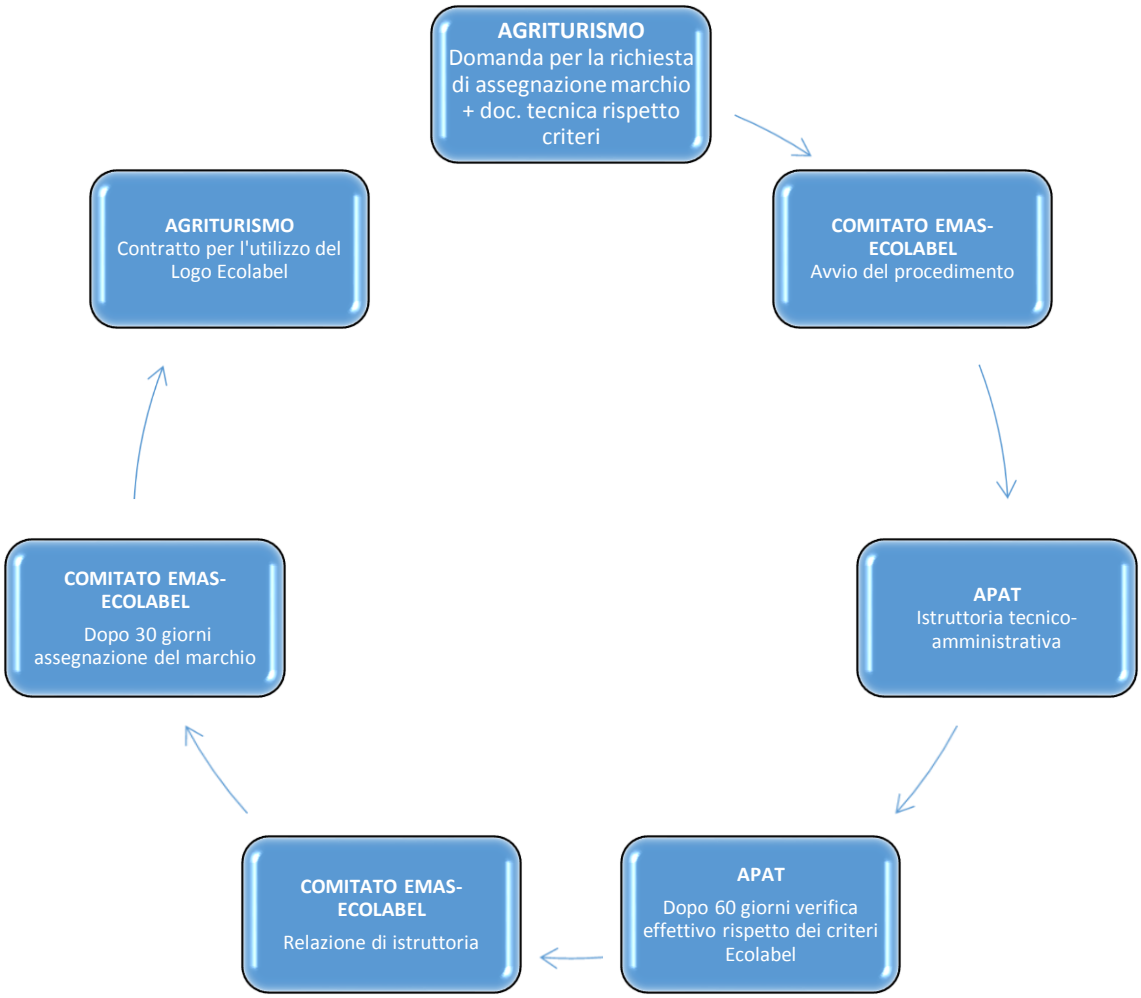


Figura 1.20 - Procedimento di richiesta e registrazione marchio ECOLABEL. Fonte: elaborazione personale.

I prodotti certificabili con il marchio europeo Ecolabel possono essere classificati in 14 gruppi di beni e 2 gruppi di servizi. In Italia attualmente ci sono 115 prodotti e 209 servizi Ecolabel. Come è possibile vedere dalle tabelle riportate nella pagina successiva, il gruppo di prodotti con il maggior numero di licenze Ecolabel in Italia è quello del servizio di ricettività turistica con un totale di 188 licenze distribuite fra alberghi, agriturismi, B&B.

Gruppo di prodotti	Numero licenze Ecolabel
Ammendanti	1
Calzature	4
Carta per copia e carta grafica	1
Carta stampata	1
Coperture dure per pavimenti	12
Detergenti multiuso	15
Detersivi per bucato	2
Detersivi per piatti	10
Mobili in legno	1
Prodotti tessili	8
Prodotti verniciati per interni ed esterni	12
Saponi, shampoo e balsami per capelli	15
Tessuto, carta	31
Substrati di coltivazione	2
Totale	115

Tabella 1.3 - Totale prodotti Ecolabel. Fonte: http://www.arpa.veneto.it/, dati gennaio 2014.

Servizi	Numero di licenze Ecolabel
Campeggio	21
Ricettività turistica	188
Totale	209

Tabella 1.4 - Campeggi e servizi di ricettività turistica con certificazione Ecolabel. Fonte: http://www.arpa.veneto.it/, dati gennaio 2014.

Capitolo 2

VERSO UN APPROCCIO "FUZZY" AL CLUSTERING

2.1 Due logiche a confronto

Fin dai tempi di Aristotele la matematica, la logica, la scienza e la cultura si sono basate su una concezione del mondo abbastanza semplice, fatta di affermazioni vere o false, di cose belle o brutte, nere o bianche, di oggetti classificabili in insiemi ben definiti e con confini precisi.

Si tratta di una logica bivalente che domina e persuade la mente degli scienziati sempre più convinti di poter spiegare la natura intrinseca delle cose, gli enigmi e i misteri più oscuri della scienza, tutto con due semplici valori, 0 e 1. Tutto è riconducibile ad essi, senza vie di mezzo. La verità si fonda su questi due piccoli e unici valori, capaci di spiegare e semplificare i mondi in cui le varie scienze affondano le loro radici.

La logica duale (binaria), nata grazie ad Aristotele, si riduce fondamentalmente ad un'unica legge: A o non-A. Il cielo è blu o non è blu: non può essere contemporaneamente blu e non blu, non può essere A e non-A.

Nonostante ciò tale logica ha sempre sollevato dubbi, ha sempre prodotto reazioni e critiche da parte di coloro che guardavano la natura sotto altri punti di vista allo scopo di far emergere le innumerevoli contraddizioni nascoste in essa, fatte di cose e non cose, di A e non-A.

La filosofia ci racconta però che questa contrapposizione alla logica binaria iniziò prima di Aristotele, circa due secoli avanti con Budda, il quale diffuse in Oriente la sua dottrina fondata essenzialmente su un unico obiettivo: andare oltre il concetto di bivalenza e vedere il mondo così com'è.

Gli strumenti logici a disposizione degli studiosi sono il frutto della logica bivalente, cioè aristotelica. Da oltre duemila anni i problemi sono affrontati e risolti con la logica del "Si o No", "Bianco o Nero", "Caldo o Freddo", "Vero o Falso", mentre i problemi che ci pone la vita reale hanno ben poco di bivalente.
È necessario per cui introdurre una nuova logica che permetta di andare oltre la classica contrapposizione fra "Bianco e Nero" e prevedere quindi anche un intervallo, come dire, di "grigi" fra i due opposti estremi.
Nel prossimo paragrafo verrà introdotta e spiegata brevemente la logica classica, cioè quella bivalente, chiamate anche dicotomica, poiché è il modo migliore per poter spiegare successivamente cosa sia la logica fuzzy (sfocata), quella logica in contrapposizione alla tradizionale logica duale introdotta da Aristotele, e analizzare in che misura le due si discostano.
Tali logiche sono alla base dei due principali metodi di classificazione, hard clustering e fuzzy clustering, trattati nei paragrafi successivi.

2.1.1 La logica dicotomica

La logica classica, sistematizzata inizialmente da Aristotele nel suo Organon, profondamente innovata successivamente da Leibniz e da Boole, è giunta alla sua forma moderna di logica simbolica, definita come la logica che "…*studia le regole che stabiliscono le condizioni in cui la verità di una proposizione può essere dedotta dalla verità di una serie di proposizioni in virtù della loro sola forma logica*…" (Garzia e Ravelli 1985, p.81).
Nella logica simbolica ci sono dei termini che possono essere predicati (aggettivi, verbi, nomi comuni) o termini individuali (nomi propri); ci sono degli enunciati, cioè delle asserzioni, che possono essere false o vere, quindi avere un valore di verità di zero o uno; dei connettivi che mettono in relazione fra loro le asserzioni sulla base di regole di composizione, e dei quantificatori, che dichiarano a quanti casi (alcuni, molti, tutti) di un insieme si applica un'espressione (Bruschi, 1993).
Gli enunciati possono essere privi di connettivi, quindi semplici, oppure composti, e vengono solitamente rappresentati tramite lettere maiuscole, come D, A, Q; i connettivi, come per esempio "*non, e, se e solo se"*, possono essere rappresentati da simboli e

vengono definiti tavole di verità. Esse mostrano il valore di verità assunto dagli enunciati quando vengono composti mediante i connettivi.

A titolo di esempio vengono ora esposte le principali operazioni che caratterizzano gli insiemi nella logica classica, senza però scendere nel dettaglio delle stesse. Queste premesse sono necessarie al fine di capire tale logica e operare successivamente un confronto con la logica fuzzy. Logica e teoria degli insiemi sono infatti reciprocamente traducibili: è la stessa cosa esprimersi in logica o in teoria degli insiemi.

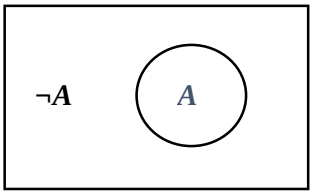
Per un approfondimento sull'argomento è possibile consultare Garzia e Ravelli 1985.

Nome	Connettivo	Simbolo
Negazione	non (not)	¬
Congiunzione	e (and, et)	∧
Disgiunzione	o (or, vel)	∨
Disgiunzione esclusiva	o (xor, aut)	+
Implicazione semplice	se… allora… (if… then…)	→
Equivalenza (doppia implicazione)	se e solo se (if and only if)	↔

Tabella 2.1 - Simboli logica dicotomica. I simboli utilizzati non sono universali, possono esserci convenzioni diverse. Fonte: elaborazione personale.

Negazione:

A	¬*A*
V	F
F	V



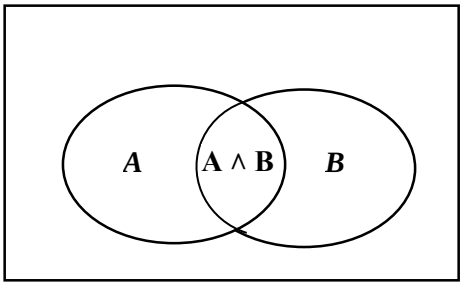
Esempio:

Proposizione *A*: “Questo cane è marrone”. Proposizione ¬*A* (NOT *A*): “Questo cane non è marrone”.

Secondo la tavola di verità della negazione, se la proposizione *A* è vera (il cane è veramente marrone) la proposizione ¬*A* è falsa, e viceversa.

Congiunzione:

A	*B*	*A* ∧ *B*
V	F	F
F	V	F
V	V	V
F	F	F



Esempio:

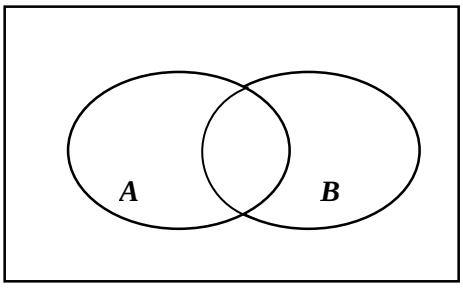
Proposizione *A*: “Questo cane è marrone”. Proposizione *B*: “Questo cane è femmina”.

Proposizione *A* ∧ *B*: “Questo cane è marrone e femmina”.

Secondo la tavola di verità della congiunzione, *A* ∧ *B* è vera solo se sono vere entrambe le proposizioni. Lo stesso ragionamento viene fatto se i soggetti delle due proposizioni sono diversi (Es: cane e gatto).

Disgiunzione:

A	*B*	*A* ∨ *B*
V	F	V
F	V	V
V	V	V
F	F	F



Proposizione *A*: "Questo cane è marrone". Proposizione *B*: "Questo cane è femmina". Secondo la tavola di verità della disgiunzione, $A \vee B$ è sempre vera, ad eccezione del caso in cui sia *A* che *B* siano falsi.

Disgiunzione esclusiva:

A	*B*	*A* + *B*
V	F	V
F	V	V
V	V	V
F	F	F

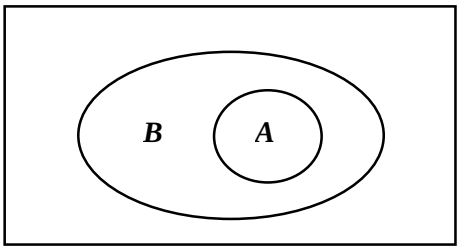
Esempio:

Proposizione *A*: "Questo cane è marrone". Proposizione *B*: "Questo cane è femmina". Secondo la tavola di verità della disgiunzione esclusiva, *A*+*B* è vera solo nel caso in cui una sola tra *A* e *B* sia vera, mentre se *A* e *B* sono entrambe vere o entrambe false *A*+*B* sarà falsa.

La disgiunzione esclusiva XOR corrisponde all' "aut…aut" dei latini: o l'una o l'altra cosa.

Implicazione semplice:

A	*B*	$A \rightarrow B$
V	F	F
F	V	V
V	V	V
F	F	V



Secondo la tavola di verità dell'implicazione semplice, "*A→B"* è sempre vera tranne quando *A* è vera e *B* è falsa.

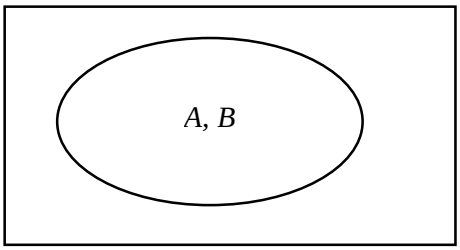
Proposizione *A*: "Questo animale è un cane". Proposizione *B*: "Questo animale è un canide". Sia *A* che *B* sono vere; "*A→B"* ("se questo animale è un cane allora è un canide") è vera se ho di fronte un cane (che in quanto tale è un canide); è vera se ho di fronte un leone (animale che non è né un cane né un canide, rende *A* falsa, *B* falsa, ma "*A→B"* rimane vera); è vera se ho di fronte un lupo (che non è un cane ma è un canide, rende *A* falsa, *B* vera e "*A→B"* rimane vera).

Dal diagramma di Eulero-Venn associato a questo connettivo è possibile osservare che la proposizione composta è vera solo se scelgo *A* come sottoinsieme di *B* (i cani sono infatti un sottoinsieme dei canidi).

Non necessariamente deve esserci un rapporto di consequenzialità tra le proposizioni *A* e *B*, come avviene nel linguaggio naturale. L'implicazione può infatti essere utilizzata per connettere tra loro proposizioni senza rapporti di causazione. Per esempio è possibile connettere tra loro la proposizione *A* "Il cane è marrone" con la proposizione *B* "La Terra è un pianeta". Nel linguaggio naturale la proposizione risultante "Se il cane è marrone allora la Terra è un pianeta" non avrebbe senso, per cui sarebbe falsa, mentre dalla tavola di verità del connettivo questa proposizione risulta vera.

Equivalenza o doppia implicazione:

A	*B*	*A↔B*
V	F	F
F	V	F
V	V	V
F	F	V



L'equivalenza è vera se entrambe le proposizioni sono vere o false.

Esempio:
Proposizione *A*: "Il coccodrillo è un rettile"; proposizione *B*: "Il gatto è un mammifero". Connettendo le sue proposizioni mediante l'equivalenza "Il coccodrillo è un rettile *se e solo se* il gatto è un mammifero" risulterà vera, essendo entrambe vere.
Anche per l'equivalenza valgono le stesse cose dette per l'implicazione semplice, cioè il fatto che il valore di verità di una proposizione composta mediante tale connettivo possa suggerire una logica di consequenzialità senza riscontro. Il valore di verità dipende, secondo la tavola di verità, soltanto dai valori di verità delle singole proposizioni di partenza.

Partendo da enunciati semplici, attraverso i connettivi, è possibile formulare enunciati composti.
Le *tautologie* "*sono affermazioni vere per qualunque valore di verità dei loro enunciati. Ciò però non dice niente sul contenuto di verità degli enunciati semplici composti in una tautologia, ma ci dice soltanto che data la struttura della tautologia essa risulta sempre vera indipendentemente dalla verità degli enunciati semplici*" (Wikipedia.it).
Le principali tautologie, principi, che sottostanno alla logica dicotomica sono:

- Principio di non contraddizione;
- Principio del terzo escluso.

Dato che è la stessa cosa esprimersi in logica o in teoria degli insiemi, tali principi possono essere spiegati attraverso il concetto di insieme.
Il primo principio afferma che un qualsiasi elemento *x* non può appartenere contemporaneamente ad un insieme *A* e al suo complemento $\tilde{A}$; quindi una mela non può essere contemporaneamente sia rossa che non rossa.

Come diceva Aristotele, "[…] *è impossibile che la stessa cosa insieme inerisca e non inerisca alla medesima cosa e secondo il medesimo rispetto; e si aggiungano tutte le altre determinazioni che si potranno aggiungere per evitare difficoltà di carattere dialettico. […] nessuno può ritenere che la medesima cosa sia e non sia, come alcuni credono che dicesse Eraclito"* (www.filosofico.net/frasarist.html).

Il secondo principio invece afferma che l'operazione di unione tra un insieme e il suo complemento genera l'insieme universo. Quindi un oggetto appartiene o non appartiene a questo insieme, senza vie di mezzo; una mela o è rossa o non lo è. Questi sono gli insiemi classici, detti anche "*crisp"*.

Nella logica classica possiamo quindi affermare che un enunciato può appartenere o no all'insieme degli enunciati veri o a quello degli enunciati falsi, considerando i due insiemi come mutuamente esclusivi. Un enunciato non può appartenere simultaneamente all'insieme degli enunciati veri e a quello degli enunciati falsi. Una persona viene considerata adulta, quindi fa totalmente parte dell'insieme degli adulti, quando compie venti anni per esempio, altrimenti fa parte dell'insieme dei giovani.

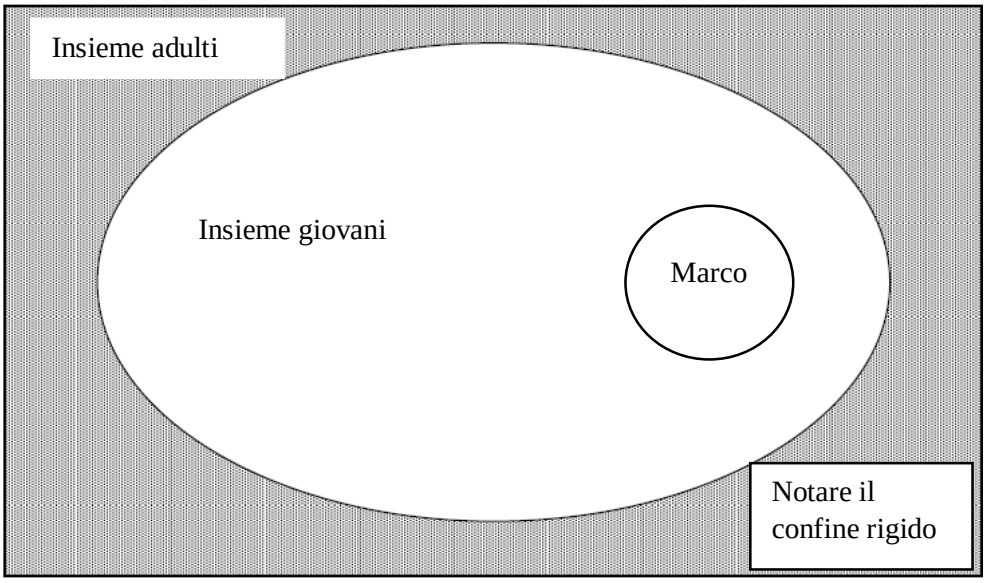


Figura 2.1 - Appartenenza di Marco all'insieme dei giovani. Fonte: elaborazione personale.

2.1.2 La logica fuzzy

Il principio di non contraddizione che caratterizza la logica tradizionale, precedentemente accennato da Parmenide, è stato contestato a più riprese.

Sia Hegel che Kant, Marx ed Engels (La dialettica della natura, Engels 1925) hanno in vario modo contestato non solo il principio di non contraddizione, ma anche quello di identità e quello del terzo escluso.

Kant mostrò i limiti e le contraddizioni della logica tradizionale; Hegel elaborò un approccio dinamico alla logica che induceva movimento e contraddizione (La scienza della logica, Hegel 1812-1816).

I primi passi importanti verso una logica non aristotelica vennero compiuti in chiave di logica multivaloriale.

Ad esempio Bertrand Russell contestò il principio del terzo escluso (Russell 1923) basandosi sull'intrinseca vaghezza di ogni simbolo e di ogni termine linguistico.

È Lukasiewicz che con la pubblicazione del saggio "Del principio di contraddizione in Aristotele, 1910" apre gli sbocchi alla logica contemporanea, di cui l'autore è universalmente riconosciuto come uno dei principali esponenti. Egli elaborò una logica trivalente, nella quale ai valori tradizionali di verità "vero e falso" aggiunse un valore di incertezza. Lukasiewicz ci fa vedere come sia perfettamente possibile pensare rigorosamente senza far valere il principio di non contraddizione.

Ma è con Lofti Zadeh, creatore della teoria degli insiemi fuzzy e della connessa logica fuzzy, che si ha il contributo più significativo. Professore all'università della California di Berkeley, nei primi anni 60 Zadeh cominciò ad avvertire che le tecniche tradizionali di analisi dei sistemi erano eccessivamente ed inutilmente accurate per i problemi tipici del mondo reale.

La pubblicazione dell'articolo "Fuzzy sets" del 1965 determina la nascita della logica fuzzy. Zadeh ebbe il merito di aver riorganizzato e rielaborato in un articolo originale tutte le intuizioni avute nel passato da altri autori, come appunto Bertrand e Lukasiewicz, Questa nuova logica ovviamente provocò reazioni e critiche da parte del mondo accademico, ancora legato al principio di non-contraddizione e del terzo escluso. Alcuni autori rimasero entusiasti delle nuove idee, altri rimasero ostili e scettici. Nonostante ciò la logica fuzzy trovò un certo successo e svariate applicazioni.

Nel 1974 E.H. Mandami applicò la logica fuzzy nell'ambito del controllo per un motore a vapore (Mandami, 1981). Da allora gli studi sulle possibili applicazioni della logica fuzzy sono cresciuti molto, in particolar modo in Giappone dove le idee di Zadeh trovarono molti sostenitori. Negli anni 80 venne infatti realizzato un sistema di controllo per la metropolitana di Sendai, dove accelerazione e frenatura erano ottenuti in modo più "morbido" dell'intervento umano, mentre la Fuji Electric applicò la logica fuzzy ad un sistema per il trattamento delle acque di scarico. Queste ed altre applicazioni condussero ad un vero boom della logica fuzzy, grazie anche ad una stretta collaborazione fra industria e università (Vedi il LIFE "Laboratory for International Fuzzy Engineering Research").

In ambito domestico la logica fuzzy venne impiegata per la realizzazione di lavatrici, aspirapolvere, frigoriferi, videocamere, macchine fotografiche (per migliorare la stabilità e l'autofocus) e per molti altri prodotti. Questo ebbe come risultato l'esplodere di una vera e propria mania per tutto quello che era etichettato come "fuzzy".

I successi giapponesi suscitarono un forte interesse per questa tecnologia anche in Corea, in Europa e, in misura minore, negli Stati Uniti.

La logica fuzzy trovò applicazione anche in ambito finanziario, dove venne applicata per la prima volta ad un sistema per le compravendite azionarie, lo Yamaichi Fuzzy Fund. Tale sistema consiste in circa 800 regole, determinate con cadenza mensile, che hanno determinato un aumento in termini di rendimento di oltre il 20% rispetto all'indice Nikkei Average.

Dunque nella logica classica una frase può essere vera o falsa, senza vie di mezzo. Un uomo può essere alto oppure no, un oggetto è verde oppure no.

Un tentativo di risolvere questo limite fu appunto proposto da Zadeh, il padre della logica sfumata.

La logica fuzzy è il tentativo di generalizzare la logica classica avvicinandola al modo di pensare umano: non si specifica che un uomo è alto 1.80 metri, ma che è alto o basso, o meglio, che è abbastanza alto. Questa logica si avvicina molto di più al ragionamento umano rispetto a quella tradizionale.

Avendo infatti a disposizione un computer che ragiona in questo modo è più semplice per chiunque dare ordini, perché per farlo basterebbe conoscere il linguaggio naturale e non i linguaggi di programmazione.
In nome fuzzy deriva dal suo significato, "sfumato, sfocato", per indicare appunto che tra 0 e 1, fra vero e falso, fra alto e basso, esiste un indefinito numero di alternative intermedie.
Passando dalla logica tradizionale a quella che utilizza gli insiemi fuzzy, sfocata, possiamo definire quest'ultima come un insieme di oggetti il cui confine è indefinito in termini di appartenenza e non appartenenza. Il concetto di appartenenza ad un insieme viene quindi ridefinito.

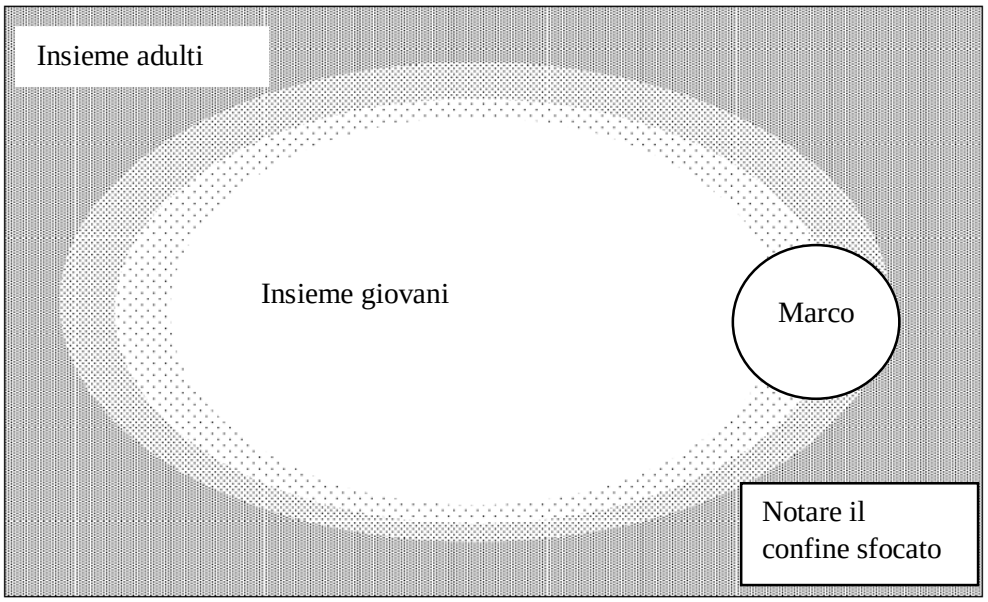


Figura 2.2 - Appartenenza di Marco all'insieme dei giovani. Fonte: elaborazione personale.

Nel 1965 Zadeh fornisce una definizione di insieme fuzzy: "Un insieme fuzzy è una classe di oggetti con un continuum di gradi di appartenenza. Tale insieme è caratterizzato da una funzione di appartenenza che assegna ad ogni oggetto un grado di appartenenza compreso fra 0 e 1. Le nozioni di unione, inclusione, complemento, intersezione e tutte le altre applicate agli insiemi tradizionali, sono estese agli insiemi fuzzy, e diverse proprietà di queste nozioni sono stabilite nel contesto degli insiemi fuzzy" (Zadeh 1965, p. 338).
Nella logica fuzzy il concetto di appartenenza è ridefinito in modo quantitativo, associando ad ogni elemento il grado di appartenenza a quell'insieme.

Un enunciato non ha più un valore di verità di 0 o 1, ma può assumere un grado di verità compreso tra 0 e 1; quindi può appartenere parzialmente all'insieme degli enunciati veri e parzialmente a quello degli enunciati falsi. In altre parole, i due insiemi non hanno dei confini ben definiti, ma sfumati, confusi, confini appunto fuzzy.
Riprendendo l'esempio dei giovani, un enunciato come "Marco è giovane" non è più esclusivamente vero o falso, non assume più solo valore di verità 1 (vero) o valore di verità 0 (falso), ma può essere vero con evidenza 0,4 oppure con evidenza 0,6 o con qualsiasi valore compreso nell'intervallo [0,1].
Come ulteriore esempio molto esaustivo è possibile considerare i paradossi del "sorite", i quali mettono in evidenza la non accuratezza dei termini del linguaggio nel descrivere il mondo reale. Per esempio, se abbiamo un mucchio di sassi e vi sottraiamo un sasso alla volta, quand'è che il mucchio cessa di essere un mucchio? Oppure se aggiungiamo mano a mano un anno di età ad una persona, questa quando non viene più considerata giovane? Certamente non quando viene raggiunto il diciottesimo anno di età, come il mucchio non smette di essere tale solo quando rimane un solo sasso. Il giudizio come si può notare diventa soggettivo. La soggettività è un elemento che prevale nella logica fuzzy, dove appunto la scelta degli intervalli e relativi insiemi resta un fatto del tutto soggettivo e dipende da colui che effettua l'analisi (Kosbo B. 1995, *Il fuzzy Pensiero. Teoria e applicazioni della Logica Fuzzy,* Baldini e Castaldi, Milano).
Supponiamo di voler considerare l'insieme delle donne giovani: cosa significa giovane? Per esempio potrei decidere di includere in questo insieme le donne fino ad un'età di vent'anni, ma un'altra persona potrebbe ritenere giovane anche una donna che ha quarant'anni, quindi intervalli diversi a seconda delle persone.

È da sottolineare che i gradi di verità della logica fuzzy non hanno niente a che fare con le funzioni di probabilità. Affermare che l'enunciato "Marco è giovane" è vero con evidenza 0,6 non significa che preso un Marco a caso questi abbia 0,6 di probabilità su 1 di essere giovane. Se fosse così la logica fuzzy non avrebbe differenze con il calcolo preposizionale. Dire che l'enunciato "Marco è giovane" è vero con evidenza 0,6 significa che in riferimento alla funzione di appartenenza all'insieme dei "giovani", Marco appartiene con una buona evidenza a questo insieme, ma appartiene anche all'insieme dei "non-giovani", con evidenza di 0,4.

I rapporti tra logica fuzzy e teoria della probabilità sono controversi e hanno determinato polemiche e confusione fra le due, in parte perché entrambe cercano di fornire un modello per trattare l'incertezza e in parte per la coincidenza dell'intervallo dei valori che possono assumere la funzione di appartenenza in un caso e la probabilità di un evento nell'altro.
La differenza sostanziale con la teoria della probabilità consiste nella sua non applicabilità ad eventi casuali bivalenti, cioè che prevedono un valore vero o falso, senza intermedi. Entrambe assumono valori variabili fra 0 e 1, ma la probabilità calcola la possibilità che si verifichi un evento futuro, mentre la logica fuzzy calcola l'ambiguità di un evento che si è già verificato. Inoltre l'incertezza della logica probabilistica muta con l'osservazione mentre quella della logica fuzzy resta immutata.
Esemplificando, se in una cassetta con 100 mele ce ne sono 7 marce, per la probabilità abbiamo una possibilità pari al 93% di prendere una mela buona, mentre per la logica fuzzy, prendendo una mela che sia per il 93% buona e per il 7% guasta, otterremo un valore di 0,93 riferito alla bontà della mela.

Le funzioni di appartenenza sono funzioni matematiche che descrivono le curve percorse dai valori della grandezza trattata in modo "fuzzy" rispetto ai gradi di verità. La funzione di appartenenza consente di attribuire il grado di verità di un enunciato in logica fuzzy e di attribuire il grado di appartenenza di un elemento ad un insieme nella teoria degli insiemi fuzzy.
Verrà ora proposto un esempio per meglio comprendere come operino le funzioni di appartenenza sugli insiemi fuzzy, riprendendo l'uso dei connettivi precedentemente esposti nel paragrafo sulla logica tradizionale, in modo da facilitare la comprensione di analogie e differenze. Per una completa trattazione sull'argomento è possibile consultare Zadeh 1965.
Consideriamo nuovamente l'esempio dei giovani e degli adulti. Mettiamo di dover dividere un certo numero di individui in tre gruppi: giovani, adulti, anziani. Nel caso della logica classica si otterrebbero tre insiemi mutuamente esclusivi.
Supponiamo di dividerli come segue:

- Minori di 20 anni: giovani;
- Tra 20 e 60 anni: adulti;
- Superiori a 60 anni: anziani.

I relativi grafici sono:

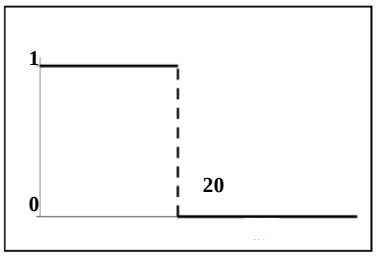


Figura 2.3 - Insieme GIOVANI

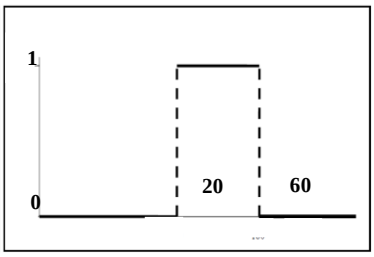


Figura 2.4 - Insieme ADULTI

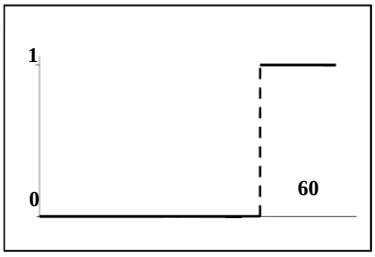


Figura 2.5 - Insieme ANZIANI

Come si può notare il valore di verità riportato sull'asse delle ordinate può assumere solo valore 0 (falso) o 1 (vero). Quindi per l'insieme adulti l'appartenenza di un elemento è vera tra i 20 e i 60 anni, è falsa negli altri casi.

Se invece si opera con la logica fuzzy, i relativi grafici verranno rappresentati come segue:

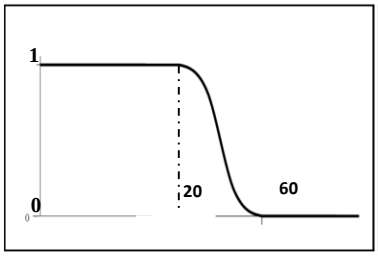


Figura 2.6 - Insieme GIOVANI

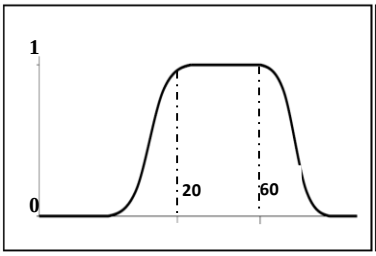


Figura 2.7 - Insieme ADULTI

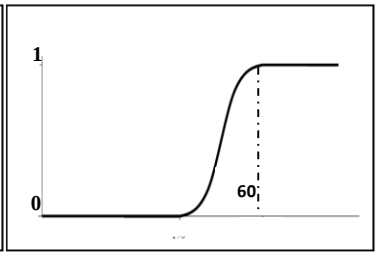


Figura 2.8 - Insieme ANZIANI

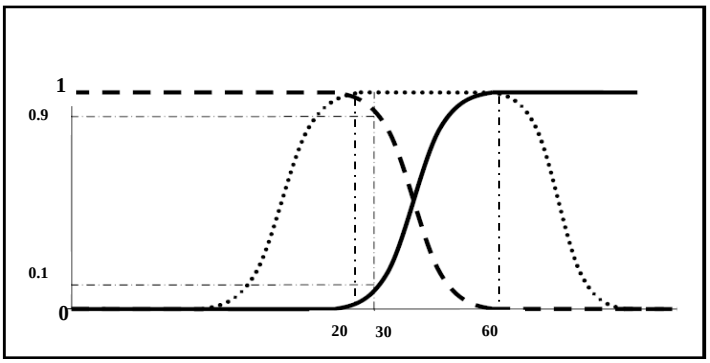


Figura 2.9 - Appartenenza di un elemento a più insiemi fuzzy

Le funzioni di appartenenza, esprimibili anche sotto forma di funzioni algebriche, mostrano per ogni valore della grandezza considerata il relativo grado di verità.
Dall'ultimo grafico si può notare che una persona di 30 anni appartiene all'insieme giovani con evidenza 0.9, all'insieme adulti con evidenza 1 e all'insieme anziani con evidenza 0.1.

Come già detto in precedenza in logica fuzzy vige un certo grado di soggettività da parte di chi effettua l'analisi. La funzione di appartenenza disegnata è infatti arbitraria e associa ad ogni età un grado di appartenenza sulla base delle decisioni di chi ha disegnato la funzione. Tale curva può essere disegnata in una pluralità di modi e questo porterebbe ad assegnare differenti gradi di appartenenza dei soggetti agli insiemi considerati.
Dunque la funzione di appartenenza è una formalizzazione, matematica o geometrica, di un insieme fuzzy e mostra il grado col quale ogni elemento vi appartiene.
Ritorniamo ora all'uso dei connettivi precedentemente esposti in modo da evidenziare un'ulteriore differenza fra le due logiche.
In che modo operano i connettivi e come si fa ad operare composizioni tra enunciati in questa logica?
È necessario evidenziare che nella logica fuzzy non esiste una modo univoco di concepire le regole di composizione. Per esempio la regola per il connettivo AND (congiunzione) non è universalmente accettata, a differenza di quanto avviene per le tabelle di verità nella logica classica.

Una determinazione comunemente accettata di alcuni connettivi fuzzy è la seguente:

Connettivo	Determinazione in logica fuzzy
Congiunzione	$A \wedge B = \min(A,B)$
Negazione	$\neg A = 1 - A$
Disgiunzione semplice	$A \vee B = \max(A,B)$

Tabella 2.2 - Connettivi fuzzy. Fonte: elaborazione personale.

Ad esempio per il connettivo AND di procede in questo modo: il grado di verità di un enunciato composto è uguale al minimo tra i valori di verità degli enunciati che lo compongono; quindi se un enunciato è composto da un enunciato con grado di verità 0,6 e da un altro con grado di verità 0,8, il valore di verità dell'enunciato composto sarà di 0,6.
Per quanto riguarda il principio di non contraddizione che caratterizza la logica tradizionale, in logica fuzzy non è più così: un enunciato può essere contemporaneamente vero e falso. Possiamo infatti avere il caso in cui l'enunciato *A* è vero con evidenza 0,4 e $\neg A$ è vero con evidenza 0,6. Stesso ragionamento per il principio del terzo escluso: mentre

nella logica classica questo afferma che $A \wedge \neg A = 0$, in logica fuzzy possiamo avere casi in cui A è vero con evidenza 0,4 e quindi $A \wedge \neg A = \min(0.4, 0.6) = 0.4$.
Di esempi così se ne potrebbero fare tantissimi e ci sarebbero ulteriori proprietà che caratterizzano la logica fuzzy ma che in tale sede non verranno trattati poiché esulano lo scopo di questa trattazione che ha avuto l'obiettivo di fornire un quadro d'insieme sulla logica fuzzy e sulle analogie e differenze con la logica tradizionale, al fine di introdurre un'importante uso di questa logica, cioè la costruzione di algoritmi di clustering.

2.2 La cluster analysis

2.2.1 Introduzione

I metodi di classificazione, detti metodi di cluster analysis o semplicemente clustering, hanno l'obiettivo di raggruppare le unità statistiche in base alla similarità del loro profilo descritto da un insieme di variabili, in modo da operare una sintesi delle informazioni contenute all'interno di un dataset.
Le unità statistiche vengono suddivise in un determinato numero di cluster in base al loro livello di somiglianza valutato a partire dai valori assunti da un certo numero di variabili in ciascuna unità.
Tali metodi, già noti alla fine del XIX secolo, hanno incominciato a suscitare interesse verso gli statistici a partire dagli anni 60. Il crescente interesse verso questi algoritmi riguarda il fatto che le tecniche di analisi dei gruppi possono essere utilizzate in moltissimi campi di ricerca (economia, fisica, medicina, ricerca sociale); per esempio in ambito economico-aziendale possono essere applicate per identificare gruppi di consumatori in base al loro comportamento di consumo o all'importanza assegnata a varie caratteristiche di un prodotto. Inoltre l'evoluzione degli strumenti automatici di calcolo ha permesso di affrontare la complessità computazionale di alcuni metodi di classificazione fino ad allora trascurati.
L'effetto di questa evoluzione è stata la creazione di algoritmi di classificazione sempre più complessi e più efficienti nell'estrapolare informazioni dai dati attraverso una loro opportuna classificazione.

Esistono varie definizioni di clustering provenienti da svariati autori.
Secondo il fisico statunitense Sokal, un processo di clustering "*consiste nel ripartire un insieme di unità elementari in modo che la suddivisione risultante goda di alcune proprietà considerate desiderabili*"; secondo altri studiosi invece consiste nel formare dei gruppi di unità in modo che tali gruppi siano il più possibile omogenei al loro interno ed eterogenei tra loro (Gordon, 1988).
In un processo di clustering le unità statistiche vengono confrontate le une con le altre in base a delle misure di distanza, valide per le variabili di tipo quantitativo, o di similarità, nel caso di variabili di tipo qualitativo.

2.2.2 Classificazione delle tecniche di clustering

Un metodo di classificazione è caratterizzato da due elementi:

- Una misura del grado di diversità tra le coppie di unità;
- Un algoritmo di individuazione dei cluster.

Una prima classificazione che si può fare dei metodi di clustering è quella che distingue fra metodi hard e metodi fuzzy a seconda che si utilizzino insiemi classici (hard) oppure insiemi sfuocati (fuzzy).
Modificando l'uno o l'altro dei due elementi indicati precedentemente si possono ottenere diverse classificazioni, alcune basate sul tipo di risultato prodotto, altre sul tipo di algoritmo adottato.
La classificazione basata sul tipo di algoritmo distingue fra metodi gerarchici e metodi non gerarchici.
I metodi gerarchici producono raggruppamenti successivi ordinabili in base a livelli crescenti o decrescenti della distanza o della similarità. Essi vengo ulteriormente distinti in metodi agglomerativi e scissori.
I primi, partendo da N unità distinte, producono via via cluster di ampiezza superiore fino ad arrivare ad un unico cluster contenente tutte le unità statistiche. Quelli scissori invece ripartiscono le unità statistiche, inizialmente unite in un unico cluster, in gruppi sempre più piccoli, fino a che il numero di cluster corrisponde al numero delle unità statistiche.

I metodi gerarchici agglomerativi sono quelli che hanno riscontrato maggior successo poiché sono più facili da programmare e comportano un rischio minore di arrivare a suddivisioni delle unità statistiche che non rispecchiano l'effettiva struttura dei dati. Infatti i metodi gerarchici scissori possono condurre più facilmente a suddivisioni errate delle unità che non vengono corrette a meno che in seguito non siano previste delle procedure di aggiustamento.
I metodi non gerarchici invece sono solo di tipo aggregativo: producono un'unica suddivisione dell'insieme di partenza procedendo a riallocazioni successive delle unità statistiche tra i gruppi definiti a priori, fino a che non si ottiene la partizione ottimale sulla base di un criterio predefinito. Questi metodi dipendono essenzialmente da due fattori: la presenza o assenza di centri e l'esistenza o meno di una funzione obiettivo.

Oltre ad una suddivisione basata sul tipo di algoritmo utilizzato è possibile operare una suddivisione sulla base del risultato prodotto.
Introduciamo ora il concetto di funzione di appartenenza. I risultati di una classificazione delle unità possono essere rappresentati in una matrice con tante righe quante sono le unità statistiche e tante colonne quanti sono i gruppi, contenente i valori della funzione di appartenenza. Tale funzione indicata con $\mu_{i,j}$ è una funzione a K valori, dove K è il numero di gruppi della partizione, che associa ad ogni unità K numeri ognuno dei quali esprime il grado di appartenenza dell'unità *i*-esima al *j*-esimo gruppo (con i = 1,2,….N e j = 1,2,…K).
La suddivisione risultante dei metodi di classificazione è la seguente:

		Metodi classici	Metodi sfocati
	Grado di appartenenza	{0,1}	[0,1]
Partizioni	$S = \sum_{j=1}^{K} \mu_{i,j} = 1$	Classificazione classica	Classificazione sfocata
Ricomprimenti	$S = \sum_{j=1}^{K} \mu_{i,j} \geq 1$	Classificazione sovrapposta	Classificazione sovrapposta sfocata

Tabella 2.3 - Classificazione dei metodi di clustering in base al risultato prodotto. Fonte: Ricolfi, 1992.

L'intervallo di definizione della funzione di appartenenza ci consente di distinguere tra metodi classici e metodi non classici, detti sfocati.

Nei primi la funzione di appartenenza è definita nell'insieme {0,1}, quindi può assumere solo i valori 0 e 1 che indicano rispettivamente se un'unità appartiene o meno all'insieme (1 = appartiene, 0 = non appartiene).
Nei metodi sfocati invece la funzione è definita nell'intervallo [0,1]; in questo caso la funzione indica il grado di appartenenza di ogni unità a ciascun gruppo.
È inoltre possibile distinguere fra partizioni e ricomprimenti. "*Una partizione è una suddivisione dell'insieme originario in K sottoinsiemi a due a due disgiunti e tali da esaurire l'insieme originario. Ogni unità appartiene ad uno e un solo sottoinsieme*" (Ricolfi 1992, p. 31).
Vale quindi il vincolo $S = \sum_{j=1}^{K} \mu_{i,j} = 1$.
"*Un ricoprimento invece è una suddivisione dell'insieme originario in K sottoinsiemi che esauriscono l'insieme originario ma possono anche essere più o meno sovrapposti fra loro*" (Ricolfi 1992, p.31).
Il vincolo diventa: $S = \sum_{j=1}^{K} \mu_{i,j} \geq 1$.

I metodi di classificazione classica sono quei metodi che forniscono una partizione classica delle unità statistiche, cioè una suddivisione in gruppi tra loro disgiunti e la cui unione fornisce l'insieme di tutte le unità. Rientrano fra questi metodi anche quelli di classificazione sovrapposta, i quali forniscono una suddivisione delle unità in gruppi non disgiunti: una stessa unità può cioè appartenere a più cluster.
I metodi di classificazione sfocata invece assegnano le unità per una parte ad un cluster e per la restante parte agli altri cluster.
Poi ci sono i metodi di classificazione sovrapposta sfocata, derivanti dall'unione dei due metodi precedenti, i quali forniscono dei ricoprimenti sfocati.
Quello che è stato detto in precedenza a proposito dei metodi gerarchici e non gerarchici vale anche all'interno di quest'ultima suddivisione. Esistono infatti metodi gerarchici e non gerarchici di classificazione classica, sovrapposta, sfocata e sovrapposta sfocata.
Infine esistono dei metodi di classificazione che sostituiscono i dati originali con dei dati sfocati, attuano cioè un'iniziale sfocatura dei dati in possesso e successivamente operano su di questi (Ponsard, 1985). Tali metodi utilizzano una funzione caratteristica che misura la quantità di carattere posseduta da un'unità rispetto a quella posseduta dalle altre unità. La classificazione che si ottiene è una classificazione classica o al massimo sovrapposta

poiché la funzione di appartenenza finale ammette valori 0 e 1. Sono dunque delle tecniche di classificazione che utilizzano dati sfocati ma che presentano la stessa funzione di appartenenza dei metodi classici e che quindi non devono essere confuse con le tecniche sfocate.

2.2.3 Classificazione classica e classificazione sovrapposta

Questo paragrafo ha lo scopo di fornire un quadro generale sulle tecniche che caratterizzano l'analisi dei cluster classica, detta hard clustering, senza scendere nel dettaglio delle stesse. Inoltre si accennerà brevemente anche la classificazione sovrapposta. Per una trattazione completa di come si effettua un'analisi dei cluster tradizionale, di come si calcola la distanza o la dissimilarità fra unità statistiche e per una spiegazione tecnica di come si applicano i vari metodi di classificazione è possibile consultare diversi autori tra cui Rizzi A. (1978) e Zani S. e Cerioli A. (2007).

I passi fondamentali di un'analisi dei cluster possono essere sintetizzati come segue:

- Scelta delle variabili di classificazione: tale scelta rispecchia per lo più le idee e le convinzioni di chi sta effettuando la ricerca, quindi implica un certo grado di soggettività. Le variabili devono essere espresse nella stessa unità di misura, altrimenti vanno standardizzate;
- Scelta di una misura di distanza/similarità fra le unità statistiche: le più note sono la distanza euclidea, euclidea al quadrato, di Manhattan, di Chebychev, di Mahalanobis e la distanza di Gower;
- Scelta dell'algoritmo di raggruppamento: algoritmi gerarchici e non gerarchici;
- Scelta del numero ottimale di cluster e valutazione della partizione ottenuta: per scegliere il numero ottimale di cluster è possibile utilizzare il grafico del dendrogramma il quale ci mostra graficamente l'intero processo di classificazione e ci permette di capire il numero opportuno di cluster da scegliere. Per valutare la partizione ottenuta si può ricorrere invece al grafico della Silhouette;
- Valutazione dei risultati ottenuti (connotazione dei cluster).

I metodi di classificazione classica, come già detto in precedenza, producono una partizione classica delle unità, cioè insiemi disgiunti e tali che la loro unione fornisca l'insieme di tutte le unità.
I metodi di classificazione possono essere distinti:

- In base al risultato fornito: agglomerativi e divisivi;
- In base all'algoritmo: gerarchici e non gerarchici.

I metodi gerarchici agglomerativi partono da una situazione iniziale in cui ogni unità rappresenta un gruppo a se stante e, attraverso successive fusioni dei grappoli, giungono ad un unico cluster contenente tutte le unità.
Tali metodi si basano su un procedimento composto dalle seguenti fasi:

1) Inizialmente ogni unità costituisce un gruppo. Si calcola quindi la matrice delle distanze;
2) Vengono fusi i due gruppi che presentano la distanza minima. Si registra quindi la distanza a cui avviene la fusione;
3) Si calcola la distanza tra i gruppi esistenti e il gruppo appena formato. Si eliminano quindi le due righe e le due colonne dalla matrice delle distanze in corrispondenza dei gruppi fusi che vengono sostituite da una sola riga e da una sola colonna contenente la nuova distanza;
4) Si ripetono i passi 2 e 3 fino a quando non si ottiene un unico gruppo formato da tutte le unità statistiche. Ci saranno quindi n-1 iterazioni.

I metodi gerarchici che vengono di seguito accennati differiscono per la modalità di calcolo della distanza tra i gruppi descritta nella terza fase. Essi sono:

- **Metodo del legame singolo**: le unità statistiche che sono tra loro a distanza minima vengono assegnate al medesimo gruppo. La distanza fra questo gruppo e le altre unità viene calcolata come la minima distanza fra le unità del gruppo e le altre unità. Il procedimento viene ripetuto fino a che tutte le unità sono nello stesso gruppo;

- **Metodo del legame medio**: è identico a quello precedente, soltanto che la distanza tra un gruppo e un'unità o un altro gruppo viene calcolata come la media tra tutte le distanze fra gli elementi. Il risultato di questo metodo è più attendibile del metodo del legame singolo, e quindi i gruppi risultano più omogenei al loro interno e differenziati tra loro;
- **Metodo del legame completo**: stessa procedura dei due metodi precedenti, solo che la distanza tra il gruppo appena formato e i gruppi o unità rimanenti è calcolata come la massima distanza tra le unità del gruppo e le altre unità;
- **Metodo del centroide**: in questo caso bisogna determinare i vettori contenenti i valori medi delle K variabili in ogni gruppo (centroidi). Le distanze fra i gruppi corrispondono alla distanza fra i rispettivi centroidi;
- **Metodo di Ward**: si basa sulla scomposizione della devianza totale tra i grappoli ed entro i grappoli. Ad ogni tappa del processo vengono uniti i due gruppi che, in seguito alla fusione, determinano il minimo incremento della varianza entro i gruppi.

Questi appena elencati sono i principali metodi gerarchici di classificazione. Oltre a questi ce ne sono molti altri che in questa sede non verranno accennati.
I metodi gerarchici scissori invece, a differenza di quelli aggregativi, partono da un solo gruppo formato da tutte le unità statistiche e procedono per partizioni successive fino ad ottenere gruppi formati da una sola unità.
A scopo informativo, tra i metodi gerarchici scissori ricordiamo il metodo di Edward e Cavalli-Sforza e i metodi di Friedman e Rubin (Rizzi A. 1985).

I metodi non gerarchici sono caratterizzati da una procedura che mira a ripartire direttamente le unità in K grappoli, fornendo come risultato una sola partizione (Matthews, 1979).
Tali metodi partono da una suddivisione iniziale delle unità e procedono spostandole da un gruppo all'altro fino a che non viene raggiunta una situazione ottimale tale che non sono consentiti ulteriori spostamenti.
Quelli più utilizzati sono i metodi di suddivisione iterativa e i metodi di programmazione matematica.

I primi eseguono spostamenti effettivi delle unità e contemplano il calcolo dei centroidi dei gruppi. Il più famoso fra questi metodi è quello di McQueen, detto algoritmo delle K medie (K-means). Tale metodo è composto dalle seguenti fasi:

1) Si scelgono *g* centri iniziali, detti anche seeds;
2) Ogni unità viene assegnata ai *g* punti in relazione alla distanza più piccola;
3) Si calcola la media, centroide, per ciascuno dei *g* gruppi;
4) Si calcola la distanza di ciascuna unità da ogni centroide. Se la distanza minima non si registra in corrispondenza del centroide del gruppo di appartenenza, l'unità viene riassegnata al gruppo con centroide più vicino;
5) Si ricalcolano i centroidi;
6) Si ripetono i passi 4 e 5 fino a che non si ottiene una soluzione stabile, cioè fino a quando si trova una soluzione per cui la devianza all'interno dei cluster è la più piccola possibile.

La misura normalmente utilizzata per calcolare la distanza fra l'unità e il centroide è quella euclidea poiché garantisce la convergenza della procedura.
Questo metodo è dotato anche di una funzione obiettivo che valuta la bontà di una partizione in modo che sia possibile scegliere lo spostamento più conveniente tra quelli possibili.
I metodi di programmazione matematica invece non prevedono il calcolo dei centroidi ed eseguono spostamenti virtuali delle unità, sulla base della soluzione di un problema di minimo o massimo vincolato.
Un vantaggio dei metodi non gerarchici consiste nella velocità di esecuzione dei calcoli. Mentre nei metodi gerarchici l'algoritmo cerca ad ogni passo la migliore aggregazione o scissione tra i cluster, nei metodi non gerarchici l'algoritmo partiziona le unità statistiche in un predefinito numero di gruppi basandosi sull'ottimizzazione di una certa funzione obiettivo che, solitamente, consiste nella minimizzazione della varianza interna. Inoltre, a differenza dei metodi gerarchici, l'assegnazione di un'unità ad un cluster non è irrevocabile, cioè le unità possono essere riassegnate ad un altro cluster se l'allocazione iniziale risulta inappropriata.

Il punto debole di questi metodi consiste nella determinazione della partizione inziale che andrà ad influenzare la classificazione finale, e nella scelta della funzione obiettivo.
Inoltre i diversi metodi di classificazione portano solitamente a soluzioni diverse. Per superare questi inconvenienti alcuni autori hanno proposto di costruire un'unica classificazione aggregando le classificazioni che derivano dall'applicazione di metodi diversi a partire dagli stessi dati. Si tratta delle cosiddette tecniche del "consenso", le quali producono un'unica classificazione a partire da più classificazioni. Queste tecniche si stanno sviluppando sempre di più grazie all'applicazione allo studio delle matrici a tre vie, cioè matrici formate da unità, variabili e occasioni, poiché sono ottimi strumenti di sintesi dell'informazione. Per un approfondimento sull'argomento consultare Gordon e Vichi, 1997.

Nella seguente tabella vengono riassunte le principali caratteristiche e differenze fra i metodi gerarchici e i metodi non gerarchici:

METODI GERARCHICI	METODI NON GERARCHICI
Possono essere sia aggregativi che scissori	Sono solo di tipo aggregativo
L'assegnazione ad un cluster è irrevocabile	Le unità possono essere riallocate
Producono diverse partizioni	Producono un'unica partizione
Sono sensibili alla presenza di outliers	Richiedono una scelta opportuna dei seeds

Tabella 2.4 – Confronti fra metodi gerarchici e metodi non gerarchici di classificazione. Fonte: elaborazione personale.

Per concludere la panoramica sui metodi di classificazione classica non si può non accennare la classificazione sovrapposta, considerata parte di quella tradizionale.
I metodi di classificazione sovrapposta, nati nel contesto della teoria dei grafi, cercano di risolvere il problema che si presenta quando ci sono unità con caratteristiche intermedie a due o più gruppi e quindi risulta opportuno assegnare le unità ad entrambi i gruppi. Il risultato è un ricoprimento e i gruppi vengono chiamati "clump" (Gordon, 1981) e non cluster. Tali metodi hanno ricevuto minore attenzione rispetto a quelli classici. Questo è dovuto in parte al fatto che alcuni autori fanno confusione tra l'approccio sovrapposto e l'approccio sfocato, per cui alcuni algoritmi che vengono definiti di classificazione sfocata in realtà sono algoritmi di classificazione sovrapposta. Inoltre alcuni metodi

vanno a sovrapporre i gruppi in maniera eccessiva per cui si ottiene un ricoprimento di difficile interpretazione. Ad ogni modo per questo tipo di problema si sono cercate delle soluzioni, come per esempio l'individuazione di vincoli per limitare il numero di sovrapposizioni consentite.
Tra i vari metodi l'unico che vale la pena evidenziare è il metodo delle piramidi, un algoritmo gerarchico di classificazione sovrapposta. Per un approfondimento è possibile consultare Diday, 1986.

2.2.4 Partizioni hard, fuzzy e possibilistiche

Dopo aver introdotto il concetto di funzione di appartenenza e utilizzato tale funzione per distinguere fra metodi di classificazione classici e metodi di classificazione sfocati, in questo paragrafo si utilizzerà tale funzione per fornire una spiegazione matematica che permetterà di evidenziare le differenze fra partizioni hard (classiche), partizioni fuzzy (sfocate) e partizioni possibilistiche. In questo modo si completerà il percorso che permetterà di affrontare l'analisi vera e propria della fuzzy clustering.
Il concetto di partizione fuzzy è essenziale per l'identificazione delle tecniche basate sulla fuzzy clustering. Le partizioni possibilistiche e fuzzy possono essere viste come una generalizzazione della partizione classica.

Dato un data set Z e supposto che il numero di cluster K sia noto, una partizione classica di Z può essere definita come una famiglia di sottogruppi $\{A_j \mid 1 \leq j \leq K\}$ che presenta le seguenti proprietà:

(1) $\sum_{j=1}^{K} A_j = Z$ (2) $A_j \cap A_p = \emptyset \rightarrow 1 \leq j \neq p \leq K$ (3) $\emptyset \subset A_j \subset Z \rightarrow 1 \leq j \leq K$

Si ricorda che *j* indica il *j*-esimo gruppo della partizione A, mentre *i* indica l'*i*-esima unità del data set Z.
La prima equazione indica che l'unione di tutti i sottoinsiemi contiene l'intero data set Z.

I sottoinsiemi devono essere disgiunti come specificato dalla seconda equazione e non devono essere vuoti o contenere tutte le unità come indicato dall'ultima proprietà.
In termini di funzione di appartenenza la partizione può essere rappresentata attraverso una matrice delle partizioni che chiameremo $U = [\mu_{ij}]_{K \times N}$.

	μ_{ij}	μ_{ij}	μ_{ij}
A_1	μ_{11}	μ_{21}	μ_{N1}
A_j	μ_{12}	…	…
A_3	μ_{1K}	…	μ_{NK}

Ciascuna riga della matrice contiene i gradi di appartenenza di ogni unità statistica al *j*-esimo gruppo; per cui nella prima riga ci sono i gradi di appartenenza di ciascuna unità al primo cluster A_1, nella seconda al secondo cluster A_2, e così via.

Gli elementi della matrice devono soddisfare le seguenti proprietà:

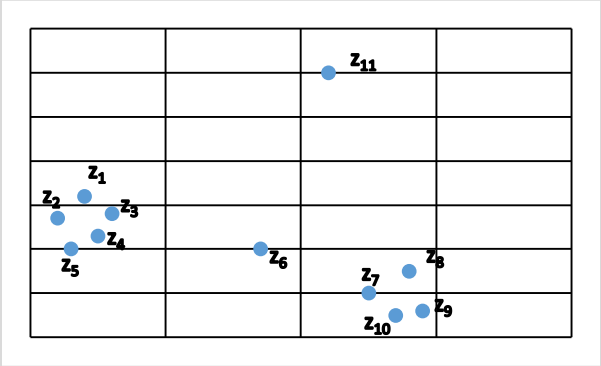
1. $\mu_{ij} \in \{0, 1\} \rightarrow 1 \leq j \leq K, \ 1 \leq i \leq N$
2. $\sum_{j=1}^{K} \mu_{i,j} = 1 \rightarrow 1 \leq i \leq N$
3. $0 \leq \sum_{i=1}^{N} \mu_{i,j} \leq N \rightarrow 1 \leq j \leq K$

Lo spazio di tutte le possibili matrici di partizione classica di Z, chiamato "*spazio di partizione hard*" (Bezdek, 1981) è definito come segue:

$$M_{hk} = \{ U \in R^{KxN} \mid \mu_{ij} \in \{0, 1\}, \forall i,j; \ \sum_{j=1}^{K} \mu_{i,j} = 1; \ 0 \leq \sum_{i=1}^{N} \mu_{i,j} < N, \forall j \}$$

Esempio:

Consideriamo un data set Z = {z_1, z_2, z_3,……., z_8}, rappresentato in questo modo:



Graficamente si può notare che ci sono due cluster separati, uno formato dai punti z_1, z_2, z_3, z_4, z_5, e un altro formato dai punti z_7, z_8, z_9, z_{10}. Inoltre il punto z_6 collocato a metà via fra i due cluster e infine il punto z_{11} che rappresenta il tipico caso di "outlier".

Una partizione possibile di questo data set può essere la seguente:

$$U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

La prima riga della matrice mostra i gradi di appartenenza di ciascuna unità al primo cluster, mentre la seconda riga i gradi di appartenenza al secondo cluster. Essendo insiemi classici, le unità possono appartenere o non appartenere ad un dato insieme, senza vie di mezzo, quindi possono assumere solamente i valori 1, appartiene, e 0, non appartiene.

In questo caso i punti z_6 e z_{11} sono stati assegnati al secondo cluster, ma potevano benissimo anche essere assegnati al primo cluster.

Come si può dedurre una partizione classica di questo data set non conduce ad un risultato efficiente. Le due unità z_6 e z_{11} devono per forza essere assegnate ad uno o l'altro cluster. Tale problema può essere superato usando un partizionamento sfocato per posizionare meglio l'unità z_6 o un partizionamento possibilistico per assegnare più efficientemente l'outlier z_{11}.

Una partizione fuzzy è ottenuta consentendo alla funzione di appartenenza μ_{ij} di assumere valori reali nell'intervallo [0, 1].

Gli elementi della matrice devono soddisfare le seguenti proprietà:

1. $\mu_{ij} \in [0,1] \rightarrow 1 \leq j \leq K,\ 1 \leq i \leq N$
2. $\sum_{j=1}^{K} \mu_{i,j} = 1 \rightarrow 1 \leq i \leq N$
3. $0 \leq \sum_{i=1}^{N} \mu_{i,j} \leq N \rightarrow 1 \leq j \leq K$

	μ_{ij}	μ_{ij}	μ_{ij}
A_1	μ_{11}	μ_{21}	μ_{N1}
A_j	μ_{12}	…	…
A_3	μ_{1K}	…	μ_{NK}

La somma di ciascuna colonna, per la seconda proprietà, deve essere 1.
Lo spazio di partizione fuzzy risulta:

$$M_{hk} = \{ U \in R^{KxN} \mid \mu_{ij} \in [0, 1], \forall i,j;\ \sum_{j=1}^{K} \mu_{i,j} = 1;\ \ 0 \leq \sum_{i=1}^{N} \mu_{i,j} < N, \forall j \}$$

Esempio:
Consideriamo lo stesso data set dell'esempio precedente sulla partizione classica.
Una delle infinite partizioni fuzzy che si possono ottenere è la seguente:

$$U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0{,}5 & 0 & 0 & 0 & 0 & 0{,}5 \\ 0 & 0 & 0 & 0 & 0 & 0{,}5 & 1 & 1 & 1 & 1 & 0{,}5 \end{bmatrix}$$

Il punto z_6 ha in questo caso una valore di appartenenza ad entrambi i cluster di 0,5 che riflette la sua posizione intermedia fra i due gruppi.
L'outlier z_{11}, nonostante sia più distante dai due cluster rispetto al punto z_6, ha gli stessi gradi di appartenenza di quest'ultimo. Questo è dovuto al fatto che la seconda proprietà richiede che la somma dei gradi di appartenenza ai vari cluster di ciascuna unità sia uguale a 1.
Inoltre è da evidenziare che in questo caso un partizionamento con tre differenti cluster sarebbe più appropriato proprio per la presenza dell'outlier.
Una partizione di tipo possibilistico può risolvere questo inconveniente che si verifica con una partizione fuzzy e assegnare all'outlier z_{11} dei gradi di appartenenza che rispecchino in modo più realistico la sua posizione rispetto a quella delle altre unità.

Una partizione possibilistica è ottenuta introducendo una modifica alla seconda proprietà della matrice delle partizioni fuzzy:

$$\exists\, j,\ \mu_{ij} > 0,\ \forall i$$

Lo spazio delle partizioni possibilistiche diventa:

$$M_{hk} = \{\ U \in \mathbb{R}^{KxN} \mid \mu_{ij} \in [0, 1],\ \forall\, i,j;\quad \forall i,\ \exists\, j,\ \mu_{ij} > 0;\quad 0 \leq \sum_{i=1}^{N} \mu_{i,j} < N,\ \forall j\ \}$$

Le partizioni possibilistiche sono state introdotte da Krishnapuram e Keller, 1993.

Esempio:

Considerando sempre lo stesso data set degli esempi precedenti, una partizione possibilistica può essere la seguente:

$$U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0{,}5 & 0 & 0 & 0 & 0 & 0{,}2 \\ 0 & 0 & 0 & 0 & 0 & 0{,}5 & 1 & 1 & 1 & 1 & 0{,}2 \end{bmatrix}$$

In questo caso, non essendoci più il vincolo per cui la somma di ogni colonna deve essere uguale a 1, all'outlier z_{11} è stato assegnato un grado di appartenenza ad entrambi i cluster di 0,2. Tale valore ci comunica che il punto z_{11} è meno tipico ai due cluster rispetto al punto z_6 e quindi rispecchia in maniera più naturale la struttura del data set.

Capitolo 3

FUZZY CLUSTERING

3.1 Introduzione

"*Probabilmente solo un modello impreciso della realtà è in grado di rappresentarla più esattamente di quanto possa fare un modello preciso*" (Ponsard 1985).
Questa frase rende perfettamente l'idea di quale sia l'utilità dei metodi di classificazione sfocata: trattare l'imprecisione che spesso, nella realtà, è presente nei dati.

Figura 3.1 - Insieme di N unità statistiche.

Le unità statistiche, come si può vedere dal disegno, molto spesso hanno una distribuzione per cui risulta difficile attribuirle ad un cluster piuttosto che ad un altro.
I metodi fuzzy hanno come punto di forza il fatto di assegnare ogni singola unità in parte a ciascun gruppo, in questo modo forniscono informazioni sul grado di appartenenza di una unità a ciascun cluster, consentendo di stabilire una gerarchia di gruppi a cui può appartenere l'unità.
Questi metodi fanno ricorso alla teoria degli insiemi fuzzy e consentono di assegnare ciascuna unità ai diversi cluster, attraverso una determinata funzione che associa un certo grado di appartenenza ad ogni singola unità compreso nell'intervallo [0,1].
In questo modo l'assegnazione ad un gruppo non risulta mai forzata come poteva accadere nella classificazione classica, mentre nel caso in cui un'unità non venga assegnata ad un cluster siamo certi che questa non fa parte di quel gruppo.

Tra i metodi di classificazione sfocata si distinguono quelli che producono cluster sfocati da quelli che producono ricomprimenti (clump). Vale poi la stessa distinzione adottata con i metodi classici fra metodi gerarchici e metodi non gerarchici, dove i primi sono spesso delle estensioni al caso fuzzy dei metodi classici.
Nel prossimo paragrafo verranno presentati i principali metodi gerarchici di classificazione sfocata, ognuno dei quali avrà una trattazione tecnica e un esempio per comprenderne meglio il funzionamento.
Seguiranno poi nel successivo paragrafo i più comuni metodi non gerarchici di classificazione.

3.2 Fuzzy clustering gerarchico

I metodi gerarchici di classificazione sfocata si caratterizzano per la presenza di due fasi:

1. Calcolo di una misura di similarità fra coppie di unità statistiche;
2. Applicazione alla matrice di similarità ottenuta di una procedura di classificazione che termina con l'assegnazione ad ogni unità di una funzione di appartenenza ai gruppi che si sono formati.

La prima fase è simile a quella della classificazione classica, cambia solo il modo in cui vengono attribuite le funzioni di appartenenza ai cluster. Nell'attribuzione delle distanze è possibile trovare infatti delle similitudini con i metodi del legame singolo e del legame completo.
Come nella classificazione classica, anche in questo caso si potrebbero studiare un gran numero di metodi di classificazione, per cui i metodi che verranno presentati in questo paragrafo possono essere modificati al fine di renderli più adatti ai casi reali che si presentano di volta in volta.

3.2.1 Metodo della sintesi di più partizioni

Il metodo della sintesi di più partizioni (Zani, 1989) è un procedimento che parte da C partizionamenti delle unità e arriva ad una classificazione sfocata di essi.

Supponiamo di avere N unità statistiche e di aver rilevato C caratteri, sia quantitativi che qualitativi. La prima cosa da fare è effettuare un iniziale partizionamento delle unità per ciascuno dei caratteri, ottenendo un totale di D partizioni (cluster). Questo ci permetterà di determinare la similarità tra una coppia di unità assumendo come indice di similarità tra le due unità della coppia, la frequenza relativa delle D partizioni in cui le due unità sono incluse in uno stesso gruppo.

Si applica poi agli indici di similarità ottenuti una procedura di classificazione che si articola in tre fasi, come verrà spiegato successivamente.

Tale procedura, come si noterà, risulta molto simile alle procedure classiche e produce una gerarchia di partizioni. Questo significa che, ad ogni livello di similarità, viene prodotta una partizione sfocata delle unità statistiche i cui gruppi contengono quelli formati ai livelli inferiori di similarità.

A ciascun livello di similarità viene assegnata ad ogni unità la sua funzione di appartenenza alla partizione, con il vincolo che la somma dei gradi di appartenenza per ciascuna unità sia uguale a 1 (Bezdek, 1981).

Dunque il primo problema da risolvere è la scelta della partizione iniziale per ciascun carattere.

Per i caratteri sconnessi, per quelli ordinali e per quelli quantitativi discreti con un numero ridotto di modalità, la scelta delle partizioni è abbastanza semplice poiché le stesse modalità solitamente comportano per ogni carattere la partizione migliore.

Per i caratteri quantitativi continui e per quelli discreti che presentano un elevato numero di modalità, per cui è necessaria una suddivisione in classi, la partizione presenta un certo grado di soggettività.

S. Zani consiglia di ripetere la procedura di classificazione considerando più partizioni iniziali con riferimento al medesimo carattere, confrontando poi i risultati che si ottengono.

Ci sono diverse modi per ricercare la partizione iniziale nel caso di caratteri quantitativi. I più utilizzati sono:

- Suddivisione in base ai quartili;
- Metodo della minima varianza (Spath, 1985);
- Metodo di ottimizzazione di Butler (Butler, 1988);
- Metodo delle classi naturali di Mineo (Mineo, 1978).

Utilizzando il primo di questi criteri, i partizionamenti si ottengono assegnando al primo gruppo le unità per le quali la modalità del carattere è compresa tra il minimo e il primo quartile, al secondo gruppo quelle con modalità del carattere compresa tra il primo e il secondo quartile, al terzo gruppo le unità con modalità del carattere compresa tra il secondo e il terzo quartile, e al quarto gruppo le unità con modalità del carattere compresa fra il terzo e il massimo valore assunto da quel carattere.
Vengono così individuati i C partizionamenti indotti dalle C variabili (caratteri).
Siano g_q i gruppi della partizione indotta dal q-esimo carattere, con $2 \leq g_q \leq N-1$ e $q = 1,2,\ldots C$.
A questo punto è possibile definire il grado di appartenenza congiunto, cioè la similarità di due unità statistiche a_i e a_j, che corrisponde alla frequenza relativa delle partizioni in cui le due unità sono incluse in uno stesso gruppo.
Per esempio, se ci sono 5 caratteri, dunque 5 differenti partizionamenti, avrò un totale di 25 partizioni (cluster), per cui il denominatore della frazione per calcolare la frequenza relativa sarà 25, mentre il numeratore corrisponderà al numero di volte in cui le due unità in questione appartengono allo stesso cluster.
Formalmente il grado di appartenenza congiunto sarà:

$$z_{i,j} = \frac{1}{C} \sum_{q=1}^{C} \delta_q (i,j)$$

dove $\delta_q (i,j)$ vale 1 se a_i e a_j appartengono allo stesso gruppo nel partizionamento q-esimo, 0 in caso contrario.
$z_{i,j}$ è un indice di similarità che assume valori compresi nell'intervallo [0,1] e gode delle seguenti proprietà:

1. Simmetrico: $z_{i,j} = z_{i,j}$
2. Proprietà riflessiva: $z_{i,i} = z_{j,j}$

3. Il suo massimo valore corrisponde alla completa appartenenza delle due unità allo stesso gruppo in ciascuno dei C partizionamenti.

Utilizzando i valori dell'indice di similarità $z_{i,j}$ si procede alla costruzione della matrice delle similarità, strutturalmente uguale alla matrice delle distanze nella classificazione classica.

Unità statistiche	a	b	c	d
a	0	$z_{a,b}$	$z_{a,c}$	$z_{a,d}$
b	$z_{b,a}$	0	$z_{b,c}$	$z_{b,d}$
c	$z_{c,a}$	$z_{c,b}$	0	$z_{c,d}$
d	$z_{d,a}$	$z_{d,b}$	$z_{d,c}$	0

Tabella 3.1 - Matrice delle similarità. Fonte: elaborazione personale.

Costruita la matrice delle similarità si può procedere alla classificazione sfocata delle unità statistiche.

Il grado di appartenenza di una unità a ciascun gruppo sfocato può essere definito ricorrendo o al criterio del minimo delle similarità tra l'unità e ciascuna delle altre unità incluse nel gruppo oppure al criterio del massimo delle similarità tra l'unità e ciascuna delle altre unità del gruppo. Il primo presenta analogie con il metodo del legame singolo nella classificazione classica poiché un'unità viene inclusa in un gruppo sfocato, ad un certo livello, quando presenta un valore della similarità con almeno un elemento del gruppo pari al valore del livello. Il secondo metodo presenta invece analogie con il metodo del legame completo.

Il metodo della sintesi di più partizioni si articola nelle seguenti fasi:

1. Le eventuali coppie di unità con similarità uguale a 1 vengono riunite tra loro, formando dunque una partizione. Queste coppie vengono definite "*nuclei*" (Rolland-May, 1985) e possono essercene più di una;
2. Si individuano le coppie di unità che presentano un grado di similarità pari a $\alpha = (C-1)/C$. Due unità costituenti una coppia con similarità α andranno a formare un nuovo gruppo con grado di appartenenza α se queste non erano già state assegnate precedentemente ad un gruppo. Tali unità individuano un nucleo al livello α. Se

invece una delle due unità era già stata inserita in un gruppo, l'altra viene unita a quel gruppo con grado di appartenenza α. Si può presentare anche il caso in cui entrambe le unità sono state già inserite in gruppi diversi. Si assegna dunque ciascuna delle unità all'altro gruppo con grado di appartenenza α, rispettando il vincolo che la somma dei gradi di appartenenza di ciascuna unità sia uguale a 1. Nel caso in cui il valore α superi questo vincolo, si attribuisce all'unità il valore massimo attribuibile. Qualora fosse 0, l'unità non può essere assegnata a nessun gruppo.

3. Si ripete il punto due considerando livelli via via decrescenti del grado di appartenenza: α = (C-2)/C, (C-3)/C, (C-4)/C,…

Il risultato di questo algoritmo è una successione gerarchica di cluster sfocati in corrispondenza di livelli decrescenti della similarità.

Esempio

L'esempio che ho deciso di riportare, così come quelli che verranno riportati per ogni metodo analizzato, è puramente a scopo indicativo per fornire un'idea di massima circa i risultati che si ottengono applicando i diversi metodi di classificazione sfocata. Per una visione completa dei metodi è possibile consultare M. A. Milioli 1994.

Nell'esempio che segue le regioni italiane vengono classificate con il metodo della sintesi di più partizioni in base ai valori assunti da 6 indicatori demografici:

X_1	Incidenza della popolazione infantile
X_2	Incidenza della popolazione senile
X_3	Indice di vecchiaia
X_4	Indice di dipendenza dei giovani
X_5	Indice di fecondità
X_6	Indice di mortalità

Figura 3.2 - Indicatori demografici. Fonte: Istat, 1990

Fra le varie partizioni sfocate ottenute a diversi livelli di similarità, è stato scelto un livello α = 0,5 poiché le regioni che rimanevano isolate, cioè che non partecipavano alla formazione di nessun gruppo, erano in numero trascurabile.

La classificazione finale ottenuta è quella seguente:

Gruppi	Regioni
1	Val d'Aosta (1), Lombardia (0.83), Umbria (0.83), Marche (1)
2	Trentino A. A. (1), Lazio (1), Abruzzi (0.83), Molise (0.83), Basilicata (0.33), Sicilia (0.17), Sardegna (0.5)
3	Piemonte (0.83), Friuli (1), Liguria (1), Emilia Romagna (1), Toscana (1)
4	Campania (1), Puglia (1), Basilicata (0.67), Calabria (1), Sicilia (0.83), Sardegna (0.5)

Figura 3.3 - Classificazione ottenuta al livello di similarità 0,5.
Fonte: M. A. Milioli (1994), "Confronto fra partizioni sfocate nell'analisi di dati territoriali".

Come si può notare nei gruppi 1 e 3 ci sono solo regioni del centro-nord, mentre il gruppo 4 è formato solo da regioni del sud. Il gruppo 2 è invece formato sia da regioni del nord che da regioni del sud, e questo ci suggerisce che ci sono caratteristiche che si trovano in tutto il paese anche se, come si può notare, i gradi di appartenenza (indicati tra parentesi a fianco di ogni regione) diminuiscono man mano che si passa dalle regioni del nord alle regioni del sud a dimostrazione del fatto che si tratta di caratteristiche presenti maggiormente al nord che al sud.

3.2.2 Metodo dei ricomprimenti sfocati

La procedura di classificazione che caratterizza questo metodo è simile al metodo del legame completo nella classificazione classica, per cui un'unità può entrare a far parte di un certo gruppo se la similarità tra essa e tutte le unità di quel gruppo è almeno pari ad un certo livello determinato di similarità. Inoltre, a differenza del metodo della sintesi di più partizioni, la somma dei valori di appartenenza di una determinata unità ai cluster può essere maggiore di 1.

Anche in questo caso è necessario partire da una matrice di similarità.

L'indice di similarità usato è quello che si adatta al minimo livello di misura dei caratteri/variabili a disposizione, in questo modo non sarà necessario né utilizzare diverse misure di similarità a seconda del tipo di carattere né effettuare modifiche dei dati che comportano solitamente variazioni degli indici di similarità.

Date N unità statistiche e C caratteri, la matrice di similarità viene costruita come segue:

1. Se tutti i caratteri sono quantitativi si calcola la distanza relativa tra le coppie di unità per ogni carattere e si fa il complemento ad uno. La distanza relativa è la distanza rapportata alla distanza massima assunta dal carattere considerato. Dunque per il carattere *q*-esimo (con q = 1,2,3,…C) si avrà:

$$V_{i,j}\,(q) = 1 - \frac{d_{i,j}(q)}{max_{i,j}[d_{i,j}\,(q)]}$$

$d_{i,j}\ (q)$ è l'indice di distanza chiamato "indice di Hamming generalizzato", calcolato tra l'unità *i*-esima e l'unità *j*-esima in riferimento al carattere *q*-esimo. L'indice di similarità totale fra le due unità sarà:

$$S(i,j) = \sum_{q=1}^{C} V_{i,j}(q) * p(q)$$

$p(q)$ è il peso, non negativo, che viene assegnato al *q*-esimo carattere sotto il vincolo che la sommatoria dei pesi sia uguale a 1:

$$\sum_{q=1}^{C} p(q) = 1$$

L'indice $S(i,j)$ è noto come indice di Gower (Gordon, 1981);

2. Se i caratteri sono ordinali la distanza tra le unità si calcola come distanza relativa tra le loro posizioni in "graduatoria" e se ne fa il complemento ad 1. Le unità devono quindi essere ordinate, per ogni carattere, in senso decrescente. Si applica poi la stessa formula $V_{i,j}$ precedente con la differenza che ora la distanza è una distanza tra le posizioni in graduatoria delle unità invece che tra le modalità assunte dalle medesime;
3. Se sono presenti dei caratteri sconnessi non è più possibile utilizzare la distanza di Hamming generalizzata, per cui come indice di similarità viene adottato quello utilizzato per la sintesi di più partizioni. Per ciascun carattere si individua una partizione iniziale delle unità statistiche e si calcola l'indice $\delta_q\ (i,j)$ che vale 1 se le unità *i* e *j* appartengono allo stesso gruppo, 0 in caso contrario. Questo

indice va sostituito a $V_{i,j}$ e di conseguenza la similarità tra le due unità statistiche è rappresentata dalla frequenza relativa delle partizioni in cui le due unità si trovano incluse nello stesso gruppo.

Una volta costruita la matrice di similarità è possibile procedere con la classificazione delle unità.
Le fasi della procedura sono le seguenti:

1. Si cercano nella matrice delle similarità le eventuali coppie di unità che presentano un indice di similarità pari a 1. Esse andranno a formare un gruppo non sfocato chiamato nucleo o core. Questa fase determina quindi una partizione non sfocata delle unità poiché la funzione di appartenenza può assumere solo i valori tipici delle partizioni classiche;
2. Nella matrice delle similarità si individuano le coppie di unità con l'indice di similarità massimo, ma minore di 1, che chiameremo α. Esse andranno a formare tanti gruppi quante sono le coppie che presentano lo stesso valore massimo di similarità. Le unità che appartengono a questi gruppi avranno grado di appartenenza α (che corrisponde anche al livello di aggregazione), quelle che non appartengono ai gruppi avranno grado di appartenenza pari al minimo delle similarità tra ciascuna unità e tutte le unità comprese nel gruppo;
3. Si cerca ora nella matrice delle similarità il valore massimo di similarità non ancora considerato fra due unità *i* e *j*, che chiameremo β. Esse andranno a formare un nuovo gruppo con livello di aggregazione β. Nel caso in cui una delle due unità, per esempio *i*, appartiene già ad un altro gruppo formato in precedenza e l'altra unità *j* presenta con le unità di questo gruppo similarità inferiori a β, quest'ultima verrà aggregata al gruppo a cui appartiene l'unità *i*. Una volta formati tutti i gruppi a questo livello di aggregazione, le funzioni di appartenenza verranno determinate allo stesso modo del passo precedente;
4. Il terzo passo viene ripetuto per valori decrescenti dell'indice di similarità finché non vengono considerati tutti i possibili livelli.

Tale procedura che produce ricomprimenti sfocati differisce da quella che produce ricoprimenti classici per due motivi: primo, una unità già assegnata ad un gruppo può essere assegnata anche ad un altro gruppo purché presenti il livello necessario di similarità con le unità di quel gruppo. Secondo, attraverso questo metodo viene prodotta una funzione di appartenenza che fornisce l'insieme dei gradi di appartenenza di ogni unità a ciascun gruppo: il grado di appartenenza di un'unità ad un determinato cluster corrisponde alla similarità minima tra l'unità considerata e ciascuna delle unità del gruppo.

Esempio

In questo esempio il metodo dei ricomprimenti sfocati viene utilizzato per classificare 22 paesi europei in base alle cause di mortalità.

Codice	Nome	Codice	Nome
AUS	Austria	YUG	Iugoslavia
BEL	Belgio	NOR	Norvegia
BUL	Bulgaria	NET	Olanda
CZE	Cecoslovacchia	POL	Polonia
DEN	Danimarca	POR	Portogallo
FIN	Finlandia	FRG	Repubblica Federale Tedesca
FRA	Francia	ROM	Romania
UNK	Gran Bretagna	SPA	Spagna
GRE	Grecia	SWE	Svezia
IRE	Irlanda	SWI	Svizzera
ITA	Italia	HUN	Ungheria

Figura 3.4 - Paesi Europei considerati.
Fonte: M. A. Milioli (1994), "Confronto fra partizioni sfocate nell'analisi di dati territoriali".

Tabella 2.4. *Cause di morte considerate.*	
Codice	Descrizione
BPI	Influenza, bronchite e polmonite
INF	Malattie infettive ed altre malattie dell'apparato respiratorio
DIG	Malattie dell'apparato digerente
ID	Malattie mal definite
DC	Tumore dell'apparato digerente
BC	Tumore della mammella
UC	Tumore dell'utero
HID	Malattie ischemiche del cuore
RA	Incidenti stradali
SUI	Suicidi
OC	Altri tumori
OCD	Altre malattie cardiovascolari
REM	Restanti cause

Figura 3.5 - Cause di morte considerate.
Fonte: M. A. Milioli (1994), "Confronto fra partizioni sfocate nell'analisi di dati territoriali".

L'esempio trattato nel libro di Milioli (M.A. Milioli, 1994, "Confronto fra partizioni sfocate nell'analisi di dati territoriali") analizza i tassi di mortalità delle donne di età compresa tra 30 e 64 anni nel 1980 per le cause di morte indicate nella tabella sopra.
I livelli di aggregazione scelti sono due: il primo, in cui la similarità minima tra le unità è di 0.80, è tra i livelli maggiori di aggregazione ottenuti e consente di individuare i paesi europei che figurano una struttura della mortalità simile tra loro; il secondo livello di aggregazione β, la cui similarità minima è di 0.70, rappresenta il primo livello nel quale non si trovano unità isolate in quanto tutte partecipano alla formazione di almeno un cluster.

La classificazione risultante che si ottiene ai due livelli è la seguente:

GRUPPI	PAESI EUROPEI
1	AUS BEL ITA FRG
2	AUS CZE ITA FRG
3	AUS ITA YUG
4	BEL ITA FRG SWE SWI
5	FIN GRE ITA NOR NET SWE SWI
6	FIN GRE ITA SPA
7	FRA ITA SPA
8	UNK IRE
9	ITA NOR FRG SWE SWI
10	YUG POR SPA
Isolati	BUL DEN POL RUM HUN

Figura 3.6 - Classificazione a livello di similarità 0.80.
Fonte: M. A. Milioli (1994), "Confronto fra partizioni sfocate nell'analisi di dati territoriali".

GRUPPI	PAESI EUROPEI
1	AUS BEL FIN FRA ITA NOR NET FRG SPA SWE SWI
2	AUS BEL CZE FIN UNK ITA NOR FRG SWE
3	AUS BEL FIN UNK ITA NOR FRG SPA SWE
4	AUS BEL CZE FIN ITA NOR FRG SPA SWE SWI
5	BEL FIN FRA GRE ITA NOR NET FRG SPA SWE SWI
6	BEL FIN FRA GRE ITA YUG NOR FRG SPA SWE
7	UNK IRE
8	AUS CZE ITA YUG POR FRG SPA
9	AUS BEL DEN UNK NET FRG SWE
10	AUS CZE HUN
11	BEL FIN UNK ITA NOR NET FRG SWE
12	BUL CZE POL POR
13	CZE DEN UNK FRG SWE
14	CZE ITA YUG POL POR FRG SPA
15	CZE RUM HUN
16	FRA GRE ITA YUG POR FRG SPA
17	YUG NOR POL FRG SPA SWE
18	YUG RUM

Figura 3.7 - Classificazione a livello di similarità 0.70
Fonte: M. A. Milioli (1994), "Confronto fra partizioni sfocate nell'analisi di dati territoriali".

Come si può notare, e ciò è ovvio, al livello di similarità 0.70 le unità coinvolte nella formazione dei gruppi sono aumentate notevolmente.

Premettendo che un'analisi completa dei risultati esula dallo scopo di questo capitolo, possiamo però anticipare che la classificazione sfocata, nel momento in cui andiamo ad analizzare i risultati ottenuti, offre alcune possibilità.
Per esempio è possibile individuare le unità tipo, cioè quelle unità che compaiono nei vari gruppi con la massima frequenza e che quindi possono essere considerate simili a buona parte delle unità in esame.
In questo caso si nota subito che al livello di similarità α l'Italia è l'unità tipo, attorno alla quale si vanno formando i gruppi. Essa infatti è presente in 8 gruppi su 10, dunque con una frequenza relativa pari a 0.8.
Se poi i caratteri sono tutti dello stesso tipo e variano all'interno dello stesso intervallo (è necessario standardizzarli) si può effettuare una classificazione in cui i caratteri sono considerati come unità e le unità sono considerate come variabili, allo scopo di vedere come i caratteri si riuniscono rispetto alle unità. Infatti si può ritenere che i caratteri che hanno una elevato grado di similarità tra loro contribuiscono insieme alla similarità tra le unità. Utilizzando i gruppi di caratteri invece che i singoli caratteri ci consente di studiare la formazione dei cluster in modo più semplice.

3.2.3 Metodo del legame medio sfocato

Si tratta di un metodo gerarchico che produce partizioni sfocate delle unità statistiche.
Anche in questo caso è necessario prima costruire una matrice delle similarità adottando lo stesso indice S*(i,j)*, valido quindi per i caratteri di tipo quantitativo.
Successivamente si procede alla classificazione delle unità nel modo seguente:

1. Nella matrice delle similarità si individuano le coppie di unità con indice di similarità massimo, le quali andranno a formare i gruppi iniziali, detti nuclei. Si calcolano poi per ogni gruppo i relativi centroidi. I centroidi rappresentano nuovi individui con caratteristiche intermedie agli individui presenti nei gruppi e quindi possono essere considerati rappresentativi dei gruppi stessi. A questo punto si calcolano i gradi di appartenenza di ogni unità a ciascun gruppo in proporzione inversa alla distanza di tale unità dal centroide del gruppo, cioè più lontana sarà l'unità dal centroide, minore sarà il suo grado di appartenenza al gruppo. La

somma dei gradi di appartenenza di ogni unità a ciascun gruppo deve essere pari a 1 in modo da ottenere una partizione e non un ricoprimento;

2. Si individua nella matrice il livello di similarità più alto tra le coppie di unità non ancora considerato. Al tale livello di similarità, che chiamiamo α, se una delle due unità della coppia (*i*, *j*), per esempio *i*, è già inclusa in un gruppo, allora l'unità *j* verrà inserita nel medesimo gruppo solo nel caso in cui la similarità media tra *j* e le altre unità che compongono il gruppo è maggiore di α. Altrimenti le unità *i* e *j* formeranno un nuovo gruppo. Una volta formati tutti i gruppi si calcolano i centroidi e i gradi di appartenenza;
3. Si ripete il punto due fino a quando tutte le unità formano un unico gruppo.

Esempio

L'esempio che segue è riferito ad un'indagine condotta nel 1997 dall'IRP (Istituto di ricerca pediatrica) circa le opinioni degli italiani sulle tendenze demografiche.

Il metodo del legame medio sfocato è stato applicato per studiare le intenzioni riproduttive degli italiani. I dati dell'indagine sono stati trasformati in dati quantitativi servendosi di alcune variabili strutturali scelte tra quelle a disposizione: gli individui che presentavano le stesse modalità di tali variabili sono stati riuniti nello stesso gruppo, costituendo una tipologia semplice (strato).

Le variabili utilizzate sono le seguenti: sesso (maschio o femmina), età (20-29, 30-39, 40-59), numero di figli (nessuno, uno, due o più).

La trasformazione in dati quantitativi è stata ottenuta calcolando le frequenze relative con cui gli individui hanno risposto alle domande circa il sesso, l'età e il numero di figli.

Combinando le modalità di queste tre variabili strutturali si sono ottenute 16 tipologie di individui, così come indicato nella tabella successiva.

Strato	Sesso	Età	N. figli
1	Maschio	Ventenni (20/29)	Nessuno
2	Maschio	Ventenni (20/29)	Uno o più
3	Maschio	Trentenni (30/39)	Nessuno
4	Maschio	Trentenni (30/39)	Uno
5	Maschio	Trentenni (30/39)	Due o più
6	Maschio	Quarantenni (40/49)	Nessuno
7	Maschio	Quarantenni (40/49)	Uno
8	Maschio	Quarantenni (40/49)	Due o più
9	Femmina	Ventenni (20/29)	Nessuno
10	Femmina	Ventenni (20/29)	Uno o più
11	Femmina	Trentenni (30/39)	Nessuno
12	Femmina	Trentenni (30/39)	Uno
13	Femmina	Trentenni (30/39)	Due o più
14	Femmina	Quarantenni (40/49)	Nessuno
15	Femmina	Quarantenni (40/49)	Uno
16	Femmina	Quarantenni (40/49)	Due o più

Figura 3.8 - Strati considerati. Fonte: elaborazione dati IRP.

A tali strati, considerati come unità statistiche, e alle frequenze relative (indicate tra le parentesi) è stato applicato il metodo del legame medio sfocato, ottenendo la classificazione seguente:

Gruppo	Num. strato	Descrizione			Gda
1	1	Maschio	Ventenni (20/29)	Nessuno	1
	3	Maschio	Trentenni (30/39)	Nessuno	1
	9	Femmina	Ventenni (20/29)	Nessuno	1
	11	Femmina	Trentenni (30/39)	Nessuno	1
2	8	Maschio	Quarantenni (40/49)	Due o più	0.4
	13	Femmina	Trentenni (30/39)	Due o più	1
	16	Femmina	Quarantenni (40/49)	Due o più	0.59
3	4	Maschio	Trentenni (30/39)	Uno	1
	5	Maschio	Trentenni (30/39)	Due o più	1
	8	Maschio	Quarantenni (40/49)	Due o più	0.23
	10	Femmina	Ventenni (20/29)	Uno o più	1
	12	Femmina	Trentenni (30/39)	Uno	1
4	7	Maschio	Quarantenni (40/49)	Uno	1
	8	Maschio	Quarantenni (40/49)	Due o più	0.37
5	15	Femmina	Quarantenni (40/49)	Uno	1
	16	Femmina	Quarantenni (40/49)	Due o più	0.41
Isolati	2	Maschio	Ventenni (20/29)	Uno o più	
	6	Maschio	Quarantenni (40/49)	Nessuno	
	14	Femmina	Quarantenni (40/49)	Nessuno	

Figura 3.9 - Classificazione al livello di similarità 0.78. Fonte: elaborazione dati IRP.

Al di là di tutte le considerazioni che si possono fare ma che esulano dallo scopo di questo esempio che si propone soltanto di fornire un'indicazione di massima sul risultato che si può ottenere applicando il metodo del legame medio sfocato, possiamo notare che all'interno di questa classificazione gli strati 8 e 16 compaiono in più gruppi a indicazione

del fatto che, probabilmente, le persone che li compongono mostrano dei comportamenti in relazione alle intenzioni riproduttive che non possono essere riferiti ad un solo schema.

3.3 Fuzzy clustering non gerarchico

I metodi di classificazione sfocata non gerarchici si caratterizzano per fornire una partizione sfocata delle unità costituita da un determinato numero di gruppi stabilito a priori. È attraverso un processo iterativo con il quale si cerca di ottimizzare una data funzione obiettivo che viene prodotta la classificazione finale.

La funzione obiettivo considerata rappresenta in genere una misura della dispersione dei punti dai centri dei cluster e differenzia i vari metodi non gerarchici di fuzzy clustering poiché, per ognuna di queste, sarà diverso il processo iterativo che porta all'attribuzione dei gradi di appartenenza delle unità ai vari gruppi.

La funzione obiettivo determina per ogni soluzione una misura dell'errore, espressa in termini di efficienza, facendo riferimento alla distanza tra le unità statistiche e le unità rappresentative di ciascun cluster.

In questa parte del capitolo verrà esposto il principale e più diffuso metodo non gerarchico di classificazione, cioè il metodo delle k-medie sfocato. Successivamente verranno illustrati dei perfezionamenti recenti di questo e infine troverà posto una breve descrizione di alcuni altri metodi di classificazione che evidenzierà le differenze nella funzione obiettivo di ciascuno di essi rispetto al metodo delle k-medie sfocato.

3.3.1 Metodo delle k-medie sfocato

Il metodo delle k-medie sfocato (Bezdek, 1981) risulta essere il metodo non gerarchico di classificazione sfocata più conosciuto e utilizzato.

Si tratta di una generalizzazione del metodo classico delle k-medie particolarmente adatto a trattare matrici con un notevole numero di dati, poiché viene solitamente raggiunta in breve tempo la convergenza verso la classificazione finale. È proprio questo il motivo per cui tale metodo è largamente utilizzato.

È stato dimostrato da alcuni studi (Bezdek e Hathaway, 1988) che le soluzioni finali ottenute attraverso questo metodo corrispondono ad un punto di minimo locale o globale delle funzioni obiettivo, o nel peggiore dei casi, ad un punto sella. Naturalmente un punto di ottimo globale risulta più efficiente di un punto di ottimo locale e soprattutto di un punto sella. Questo però non rappresenta un problema particolarmente importante poiché, sempre in questi studi, è stato fatto notare che le classificazioni corrispondenti ai tre diversi punti di ottimo sono quasi del tutto identiche tra loro nella maggior parte dei casi. Prima di applicare l'algoritmo che permette di arrivare alla classificazione sfocata delle unità è necessario specificare alcuni elementi:

- Il numero di cluster K in cui vogliamo dividere le N unità statistiche rilevate su C caratteri;
- Il partizionamento iniziale delle unità $U^{(0)}$. Esso può essere effettuato in modo casuale oppure basandosi su conoscenze a priori. Da questo partizionamento si otterrà successivamente, attraverso una procedura di minimizzazione delle funzione obiettivo, una classificazione sfocata nella quale per ciascuna unità viene determinato il grado di appartenenza ad ogni gruppo;
- Il parametro *m*. Esso risulta molto importante perché determina il grado di sfocatura (*fuzziness*) della classificazione finale;
- Il grado di errore massimo δ;
- La matrice delle norme A;
- La funzione obiettivo.

Analizziamo ora uno alla volta gli elementi appena elencati.

Numero di cluster

Il parametro K rappresenta il più importante fra i parametri da scegliere per effettuare una classificazione sfocata con il metodo delle k-medie poiché esso ha molta più influenza di qualsiasi altro parametro sulla partizione risultante.

La scelta del numero di cluster K della partizione è per lo più soggettiva, cioè dipende dalla sensibilità di colui che sta effettuando l'analisi e delle supposizioni e conoscenze che ha del fenomeno.
Gli studi di Milligan (Milligan 1985) hanno dimostrato che tutti i test provati per la scelta del numero di cluster non sono esenti da qualche difetto, per cui non esiste un criterio oggettivo per la scelta di K che vada bene per qualsiasi situazione.
Un metodo che può essere preso in considerazione per semplificare la scelta di K è quello di provare alcuni valori di K e confrontare poi le diverse classificazioni ottenute, evitando però di scegliere un numero troppo elevato di cluster poiché, dato che le unità verranno ripartite in parte nei vari gruppi, si finirebbe con l'ottenere una classificazione troppo sfocata delle unità e quindi sarebbe difficile interpretarla.
Esistono poi due approcci per determinare un appropriato numero di cluster all'interno del data set:

1. ***Misure di validità***. Le misure di validità sono degli indici che permettono di valutare la bontà della partizione ottenuta. Si possono verificare dei casi in cui il numero di cluster scelto corrisponde più o meno al numero di gruppi presenti nel data set, in questo modo possiamo aspettarci che l'algoritmo di clustering identifichi correttamente i relativi cluster. In caso contrario invece si ottiene una classificazione errata delle unità e di conseguenza i cluster non risulteranno ben separati tra di loro e compatti al loro interno. Sono stati di conseguenza formulate diverse misure di validità da utilizzare per poter identificare la migliore partizione possibile (Bezdek 1981, Gath and Geneva 1989, Pal and Bezdek 1995). Fra le tante, una misura largamente utilizzata è l'indice di Xie-Beni (Xie and Beni 1991):

$$\chi\,(Z;\,U;\,V) = \frac{\sum_{j=1}^{K}\sum_{i=1}^{N}\mu_{ij}^{m}\left\|z_i - v_j\right\|^2}{K * min_{j\neq q}\left(\left\|v_j - v_q\right\|^2\right)}$$

   z_i indica l'unità *i*-esima, v_j e v_q rispettivamente i centri dei cluster *j*-esimo e *q*-esimo, *m* il grado di sfocatura che verrà spiegato successivamente, Z il data set, U la matrice della partizione inziale e V il vettore contenente i centroidi dei K cluster. Tale indice può essere interpretato come il grado della varianza intra-

gruppo totale e come grado di separazione dei centri dei cluster. La partizione migliore minimizza il valore di questo indice.

2. ***Fusione o inserimento iterativo di cluster.*** Nel caso di fusione iterativa di cluster si parte con un numero sufficientemente grande di cluster e successivamente lo si riduce unendo i cluster simili attraverso alcuni criteri definiti (Krishnapuram and Freg 1992). Nel caso invece di inserimento iterativo di cluster si parte con un piccolo numero di cluster e si inseriscono poi, iterativamente, i cluster nelle regioni dove le unità hanno un basso grado di appartenenza ai gruppi esistenti. Per un approfondimento consultare Gath and Geva 1989.

Scelta della partizione iniziale $U^{(0)}$

Alcuni studi hanno dimostrato che il risultato fornito dal metodo delle k-medie sfocato è nella maggior parte dei casi lo stesso indipendentemente da quale sia la partizione inziale che può essere così causale o scelta da colui che sta effettuando l'analisi in base a precedenti conoscenze del fenomeno (Bezdek e Hathaway 1988).
Il metodo infatti converge quasi sempre allo stesso punto della funzione obiettivo.
Ovviamente più la partizione scelta si avvicina a quella finale, più velocemente sarà raggiunta la convergenza. Un modo per ridurre i tempi di elaborazione è quello di scegliere U mediante un altro metodo di classificazione, classico o sfocato, e procedere poi all'applicazione dell'algoritmo delle k-medie sfocato.
Ad ogni modo, solitamente, quello che viene fatto è inizializzare $U^{(0)}$ scegliendo casualmente i centri dei cluster v_j e utilizzando poi la seguente formula per determinare i gradi di appartenenza delle unità ai cluster:

$$\mu_{ij} = \frac{1}{\sum_{q=1}^{K} \left(\frac{d_{ij}}{d_{iq}}\right)^{\frac{2}{(m-1)}}}$$

μ_{ij} indica il grado di appartenenza dell'unità *i-esima* al gruppo *j*-esimo. *q* indica il gruppo *q*-esimo.

μ_{ij} deve sottostare ai seguenti vincoli:

1. $0 \leq \mu_{ij} \leq 1$ con $i = 1,\ldots,N$; $j = 1,\ldots,K$
2. $\sum_{j=1}^{K} \mu_{ij} = 1$ con i = 1,…N

Il primo vincolo indica l'insieme di definizione della funzione di appartenenza, mentre il secondo indica che la somma dei gradi di appartenenza di ogni unità ai diversi gruppi deve essere uguale a 1.
La distanza d_{ij} dell'unità *i*-esima dal centroide del gruppo *j*-esimo è calcolata come:

$$d_{ij} = \|z_i - v_j\|$$

dove ||... ... ... || è un'opportuna norma su ℝ, per esempio la norma euclidea.

Livello di sfocatura "m"

La scelta del parametro *m* determinerà il grado di sfocatura della classificazione finale.
Può assumere valori nell'intervallo [1, ∞).
Attraverso diverse applicazioni del metodo delle k-medie sfocato è stato dimostrato che non esiste un valore ottimale per *m*, in quanto esso varia in riferimento alle singole applicazioni. Tutto ciò rappresenta un difetto del metodo poiché la scelta di *m* può essere effettuata solamente osservando le diverse classificazioni finali ottenute ai differenti valori stabiliti per *m*. In ogni caso risulta difficile scegliere il livello ottimale di *m* se non si hanno precedenti informazioni sul fenomeno.

Fra tutti i possibili valori che *m* può assumere due sono i casi principali che si possono verificare:

- ***m* = 1**: la classificazione finale è una classificazione classica in quanto il grado di appartenenza assume solamente i valori 0 e 1. In questo caso il metodo delle k-medie sfocato corrisponde a quello delle k-medie classico, attribuendo le unità totalmente ai cluster da cui hanno distanza minore;

- ***m* > 1**: tanto maggiore sarà il valore di m, tanto più il grado di appartenenza di ciascuna unità tenderà ad assumere valori sempre più distanti dai casi limite 0 e 1. Il grado di sfocatura massimo che la classificazione può assumere corrisponde al valore 1/K, dove le unità sono equamente distribuite fra i vari cluster.

Come è possibile notare, *m* risulta molto importante al fine di ottenere un buon risultato della classificazione. In molti casi però, come già detto in precedenza, in assenza di informazioni sul fenomeno da analizzare la scelta di *m* risulta difficile.
Per risolvere il problema sono state comunque proposte alcune soluzioni, tra le quali un indice che misuri il grado di sfocatura delle varie classificazioni, considerando che una classificazione si dice totalmente sfocata se tutti i valori di appartenenza per ogni unità assumono lo stesso valore 1/K e si dice classica se ciascuna unità appartiene ad un solo cluster.
Da tutto questo è possibile notare che il concetto di sfocatura corrisponde a quello di eterogeneità: una classificazione è detta più o meno sfocata in base a quanto essa è eterogenea.
Dunque, un indice di eterogeneità può essere considerato anche un indice di sfocatura che permette di misurare il grado di sfocatura delle classificazioni. Per esempio si può utilizzare l'indice relativo di eterogeneità di Gini:

$$I = \frac{K}{K-1}\left[1 - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\mu_{ij}^2\right]$$

L'indice di Gini permette di misurare l'effetto dei diversi valori del parametro *m* sulle classificazioni corrispondenti; esso varia nell'intervallo [0,1], assumendo valore 1 nel caso di massima sfocatura e valore 0 nel caso di una classificazione classica. Sulla base di alcuni esperimenti è stato dimostrato che il valore migliore di *m* è quello ottenuto in corrispondenza del valore dell'indice *I* compreso tra 0,4 e 0,5.
Normalmente il valore di *m* scelto è pari a 2.

Il grado di errore ϵ

Il grado di errore massimo è utilizzato come criterio per porre un limite all'iterazione dell'algoritmo delle k-medie. L'algoritmo non viene più reiterato quando la norma della differenza tra le matrici dei gradi di appartenenza U in due iterazioni successive è minore del parametro di terminazione δ.

$$max_{ij}\left(\left|\mu_{ij}^{l} - \mu_{ij}^{l-1}\right|\right)$$

Solitamente viene scelto δ = 0,001, ma nella maggior parte dei casi si preferisce usare δ = 0,01 poiché, oltre a funzionare bene, riduce drasticamente il numero di iterazioni.

Matrice delle norme A

La scelta della matrice delle norme A determina la forma risultante dei cluster.
Un scelta diffusa è quella di porre A = I (matrice identità), ottenendo la norma Euclidea:

$$D_{ij}^2 = \left(z_i - v_j\right)^T \left(z_i - v_j\right)$$

Un'altra scelta è quella di porre A uguale alla matrice diagonale che rappresenta le diverse variazioni nelle direzioni degli assi coordinati del data set Z:

$$\begin{bmatrix} \left(\frac{1}{\sigma_1}\right)^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\frac{1}{\sigma_n}\right)^2 \end{bmatrix}$$

Infine è possibile porre A uguale alla matrice inversa delle covarianze Z: $A = R^{-1}$

$$R = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\,(z_i - \bar{z})^T$$

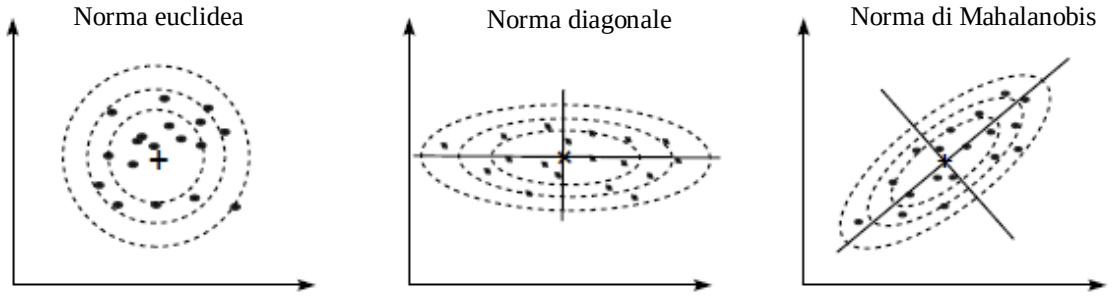
$\bar{z}$ indica la media dei dati in Z.

In questo caso la matrice A contiene le norme di Mahalanobis.

La norma influenza il metodo delle k-medie in quanto cambia la misura di dissimilarità fra le unità statistiche e la forma dei cluster.

Per esempio la norma Euclidea determina un forma sferica dei cluster, la norma diagonale e la norma di Mahalanobis un forma ellittica.

La differenza fra queste ultime due consiste nella direzione degli assi degli ellissi: mentre con la norma diagonale sono paralleli agli assi cartesiani, con la norma di Mahalanobis l'orientamento degli assi dell'ellisse è arbitraria.



Una limitazione degli algoritmi basati su una norma-distanza prefissata sta nel fatto che tale norma forza la funzione obiettivo a preferire cluster con determinate forme anche se queste non sono rappresentative dei dati in esame.

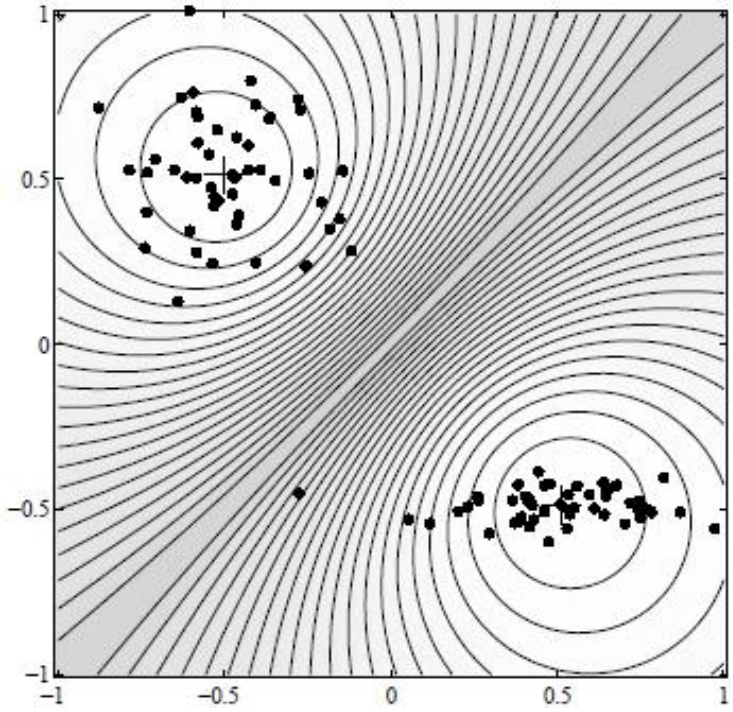


Figura 3.10 - Come si vede dalla figura, il primo cluster in alto a sinistra presenta una distribuzione sferica delle unità mentre il secondo ha una forma allungata. L'applicazione dell'algoritmo delle k-medie, ponendo A = I per entrambi i cluster, impone una forma sferica ad entrambi i gruppi (indicata dai cerchi).

Normalmente sarebbero richieste differenti matrici A per operare un confronto fra le diverse forme risultanti dei cluster, ma non esiste comunque un modo per poter scegliere tale matrice a priori.

È stato formulato però un algoritmo che permette di "adattare" le matrici A delle norme allo scopo di identificare cluster con differenti forme geometriche all'interno di un data set. Esso prende il nome di "Algoritmo di Gustafson-Kessel" (Gustafson and Kessel, 1979) e verrà esposto successivamente all'interno di questo capitolo.

La funzione obiettivo

La funzione obiettivo che deve essere minimizzata, detta funzione di ottimizzazione e indicata con J_m, è utilizzata allo scopo di determinare i valori ottimi del grado di appartenenza ed è funzione del quadrato della distanza d_{ij} tra l'unità *i*-esima e il centroide

del gruppo *j*-esimo. Inoltre dipende dal parametro *m* che, come specificato in precedenza, può assumere valori reali ≥ 1.

La funzione è la seguente:

$$J_m(Z, U, v) = \sum_{j=1}^{K} \sum_{i=1}^{N} (\mu_{ij})^m (d_{ij})^2$$

$(d_{ij})^2 = |z_i - v_j|^2$;

|... ... | è un'opportuna norma su ℝ, per esempio la norma euclidea.

$v_j \in \mathbb{R}$ rappresenta la componente *j*-esima del vettore dei centroidi v = ($v_1, \ldots v_j$).

$z_i \in \mathbb{R}$ rappresenta la componente *i*-esima del vettore delle unità x = (x_1,...x_n).

La funzione obiettivo ha la seguente interpretazione: per ogni cluster il suo centroide è la migliore rappresentazione delle unità che lo compongono poiché esso minimizza la somma dei quadrati degli errori x_i-v_j. Quindi la funzione J_m misura l'errore quadratico totale che deriva dal rappresentare le N unità statistiche con i K centroidi dei cluster.
Di conseguenza J_m dipende da come le unità sono raggruppate nei cluster e ci fornisce una misura della loro dispersione intorno ai centri dei cluster.
La partizione ottima è quella che minimizza il valore della funzione obiettivo e viene chiamata "partizione di minima varianza".

Dunque partendo dalla partizione iniziale delle unità, attraverso una serie di iterazioni che hanno lo scopo di minimizzare la funzione obiettivo, si ottiene una classificazione sfocata nella quale ad ogni unità viene assegnato il grado di appartenenza ai K gruppi.
L'algoritmo delle k-medie sfocato è il seguente e viene ripetuto fino a che non si raggiunge il livello di errore stabilito inizialmente:

1. Si fissano prima di tutto il parametro *m*, il numero di cluster K in cui vogliamo suddividere le N unità statistiche, la partizione iniziale $U^{(0)}$ (l'esponente indica il numero di iterazione) e il livello di errore massimo δ;

2. Si calcolano i K centroidi v_j dei cluster usando la seguente formula:

$$V_j^{(0)} = \frac{\sum_{i=1}^{N} (\mu_{ij})^m z_i}{\sum_{i=1}^{N} (\mu_{ij})^m}$$

3. Si calcola la nuova matrice dei gradi di appartenenza U(1), che rappresenta il risultato della prima iterazione. Le regole da seguire per la costruzione della matrice sono le seguenti:
   I. Se c'è qualche gruppo p in cui si ha che $d_{ip} = 0$ si pone $\mu_{ip} = 1$ e $\mu_{ij} = 0$ per tutti i $j \neq p$;
   II. Altrimenti si utilizza la seguente formula:

$$\mu_{ij} = \frac{1}{\sum_{q=1}^{K} \left(\frac{d_{ij}}{d_{iq}}\right)^{\frac{2}{(m-1)}}}$$

4. Usando un'opportuna norma, si calcola la differenza fra quanto ottenuto all'ultima e alla penultima iterazione. Se $\left|\mu_{ij}^{l} - \mu_{ij}^{l-1}\right| \leq \delta$ ci si può fermare e si considera come classificazione finale quella ottenuta alla penultima iterazione; in caso contrario si ripetono i punti 2 e 3 fino a che la condizione 4 non è soddisfatta.

Esempio

Il metodo delle k-medie sfocato viene ora applicato ad un gruppo di 22 unità statistiche, rappresentate nelle seguente figura:

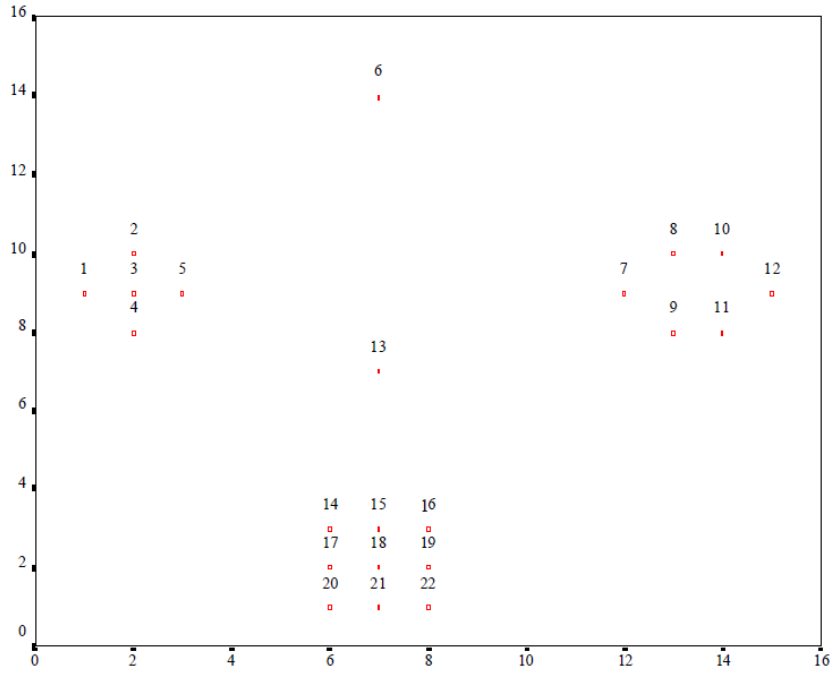


Figura 3.11 - Rappresentazione unità statistiche. Fonte: elaborazione personale.

Per il calcolo delle distanza è stata usata la norma euclidea, il parametro δ posto uguale a 0.001 e il numero di cluster scelto è stato 3 dato che le unità, come si può facilmente vedere, sono disposte in tre gruppi ben distinti.

Sono state effettuate tre diverse prove con differenti livelli del parametro *m*. I risultati della classificazione sono i seguenti:

	grado di appartenenza		
Unità	Cluster 1	Cluster 2	Cluster 3
1	1.00	0.00	0.00
2	1.00	0.00	0.00

	grado di appartenenza		
Unità	Cluster 1	Cluster 2	Cluster 3
1	0.86	0.06	0.08
2	0.89	0.05	0.06

3	1.00	0.00	0.00
4	1.00	0.00	0.00
5	1.00	0.00	0.00
6	1.00	0.00	0.00
7	0.00	1.00	0.00
8	0.00	1.00	0.00
9	0.00	1.00	0.00
10	0.00	1.00	0.00
11	0.00	1.00	0.00
12	0.00	1.00	0.00
13	1.00	0.00	0.00
14	0.00	0.00	1.00
15	0.00	0.00	1.00
16	0.00	0.00	1.00
17	0.00	0.00	1.00
18	0.00	0.00	1.00
19	0.00	0.00	1.00
20	0.00	0.00	1.00
21	0.00	0.00	1.00
22	0.00	0.00	1.00
I = 0			

Figura 3.12 - Classificazione classica con m = 1.
Fonte: Iacovacci, 1995.

3	0.97	0.01	0.02
4	0.85	0.05	0.09
5	0.90	0.04	0.06
6	0.43	0.35	0.22
7	0.08	0.82	0.10
8	0.05	0.89	0.06
9	0.06	0.86	0.08
10	0.06	0.88	0.06
11	0.06	0.86	0.08
12	0.07	0.83	0.10
13	0.36	0.26	0.38
14	0.11	0.08	0.81
15	0.07	0.06	0.87
16	0.09	0.09	0.82
17	0.08	0.06	0.86
18	0.00	0.00	1.00
19	0.06	0.07	0.87
20	0.10	0.08	0.82
21	0.07	0.06	0.87
22	0.08	0.09	0.83
I = 0.411			

Figura 3.13 - Classificazione con m = 2.7
Fonte: Iacovacci, 1995.

	grado di appartenenza		
Unità	Cluster 1	Cluster 2	Cluster 3
1	0.39	0.30	0.31
2	0.39	0.30	0.31
3	0.62	0.19	0.19
4	0.38	0.30	0.32
5	0.39	0.30	0.31
6	0.34	0.34	0.32
7	0.30	0.38	0.32
8	0.30	0.39	0.31
9	0.30	0.39	0.31
10	0.31	0.38	0.31
11	0.30	0.38	0.32
12	0.31	0.38	0.31
13	0.33	0.33	0.34
14	0.32	0.31	0.37
15	0.31	0.31	0.38
16	0.31	0.31	0.38
17	0.31	0.30	0.39
18	0.17	0.17	0.66
19	0.31	0.31	0.38
20	0.31	0.31	0.38
21	0.31	0.30	0.39
22	0.31	0.31	0.38
I = 0.976			

Figura 3.14 - Classificazione con m = 20
Fonte: Iacovacci, 1995.

Come si può notare dalla rappresentazione grafica delle unità statistiche le unità 6 e 13 non possono essere assegnate completamente ad un solo cluster poiché si trovano in una posizione intermedia rispetto ai gruppi presenti.
Scegliendo $m = 1$, quindi utilizzando il metodo delle k-medie classico, le due unità appena menzionate vengono attribuite completamente al primo cluster formato dalle unità {1,2,3,4,5} ottenendo di conseguenza una classificazione che non rispecchia la distribuzione reale delle unità.
Osservando le tabelle si può osservare che la classificazione migliore è quella ottenuta con m = 2.7 poiché le unità 6 e 13 vengono attribuite ai cluster con gradi di appartenenza che rispecchiano la loro distanza dai diversi cluster. Anche questa classificazione però non rispecchia pienamente la situazione reale. Infatti a nessuna delle unità che appartengono con certezza ad un dato cluster viene attribuito il massimo valore di appartenenza a meno che il valore di *m* non si avvicini ad 1, ma in questo caso si otterrebbe una classificazione poco sfocata che mal rappresenterebbe la situazione delle unità 6 e 13.
Per risolvere questo problema si può ricorrere al metodo delle k-medie semisfocato il quale limita la sfocatura alle sole unità che la necessitano veramente. Lo stesso esempio qui proposto verrà poi ripreso nel paragrafo del metodo delle k-medie semisfocato allo scopo di evidenziare il diverso risultato che si ottiene.

3.3.2 Miglioramenti del metodo delle k-medie sfocato

Nonostante il metodo delle k-medie sia considerato uno dei migliori metodi di classificazione sfocata, anch'esso presenta alcuni punti deboli e limitazioni.
Per esempio in alcuni casi può fornire un'eccessiva sfocatura della classificazione. In particolar modo può accadere che:

1. Se il numero di cluster K è troppo elevato la classificazione risulterà troppo frastagliata e di difficile interpretazione, per il fatto che ogni unità deve appartenere almeno in parte ad ogni cluster;
2. Se i cluster sono ben separati tra loro non esiste sfocatura che non fornisca un risultato distorto.

C'è poi una limitazione di questo metodo basato, come specificato nel paragrafo precedente, su una norma-distanza prefissata, e cioè il fatto che tale norma forza la funzione obiettivo a preferire cluster con determinate forme anche se queste non sono rappresentative dei dati in esame.
Per far fronte a questi problemi e limitazioni sono stati introdotti successivamente alcuni perfezionamenti, alcuni dei quali verranno esposti all'interno di questo paragrafo.

3.3.2.1 Metodi di classificazione semisfocata: i metodi di Selim, Ismail e Kamel

Nel 1984 Ismail e Selim (Ismail e Selim, 1984) hanno introdotto tre varianti al metodo delle k-medie sfocato arrivando ad ottenere dei metodi che forniscono come risultato delle classificazioni semisfocate, allo scopo di risolvere il problema di interpretazione che si presentava con un numero elevato di cluster.
Le variazioni introdotte sono le seguenti:

a) Imposizione di un vincolo sul numero massimo di cluster.
   All'inizio del procedimento, oltre al numero K di cluster, si sceglie un numero $\pi < K$ di cluster e si impone che le unità possano appartenere al massimo a π cluster. Questo risultato si ottiene ordinando, per ogni unità statistica, i primi π valori della funzione di appartenenza in ordine decrescente e fissando a 0 i restanti K-π valori. È necessario poi normalizzare la funzione di appartenenza allo scopo di rispettare il vincolo che impone che per ogni unità la somma dei gradi di appartenenza deve essere uguale a 1. Il valore per la *i*-esima unità del j-esimo cluster della funzione di appartenenza normalizzata è il seguente:

$$w_{ij} = \frac{\mu_{ij}}{\sum_{j=1}^{K} \mu_{ij}}$$

   Per applicare questo vincolo è però necessario possedere a priori precise informazioni per poter scegliere in modo appropriato il numero π di cluster. Infatti limitando il numero di cluster ai quali le unità possono appartenere si possono

creare delle forti distorsioni nel calcolo dei gradi di appartenenza delle unità stesse.

b) Imposizione di un valore massimo sopra il quale non viene attribuito nessun grado di appartenenza a quella unità la cui distanza dal centro del gruppo supera tale valore.

   Dunque, fissato γ come valore massimo, se la distanza d_{ij} tra l'unità *i*-esima e il centro del *j*-esimo cluster è maggiore di γ, si pone $\mu_{ij} = 0$. Si procede poi alla normalizzazione della funzione di appartenenza come nel caso precedente.

   È necessario porre attenzione nella scelta di γ poiché, se tale valore viene scelto troppo piccolo, alcune unità potrebbero venire escluse da qualsiasi cluster;

c) Imposizione di un valore soglia di accettazione per i gradi di appartenenza delle unità ai gruppi.

   Questo vincolo è stato introdotto per evitare che la funzione di appartenenza possa assumere valori troppo bassi. Per cui, stabilito il valore di β come soglia di accettazione, se $\mu_{ij} < \beta$ si pone $\mu_{ij} = 0$.

   Si procede poi alla normalizzazione della funzione di appartenenza.

   Tale valore, se scelto in modo appropriato, consente di determinare in modo naturale per ogni unità il numero di cluster ai quali appartengono, superando in questo modo la difficoltà di scelta a priori del numero π di cluster. Anche in questo caso però è necessario porre molta attenzione nella scelta di β poiché se viene scelto un valore troppo elevato alcune unità potrebbero non appartenere ad alcun gruppo o, nel peggiore dei casi, alcuni cluster potrebbero risultare vuoti.

Nel 1991 Kamel e Selim (Kamel e Selim, 1991) cercano di migliorare il terzo vincolo allo scopo di superare il problema della scelta del parametro β.

Il metodo introdotto da loro e chiamato "*Thresholded Fuzzy C-Means* (TFCM)" prevede di determinare un valore soglia β appropriato del grado di appartenenza delle unità ai vari gruppi, al fine di evitare che alcune unità non appartengano ad alcun gruppo.

Una volta calcolati per ogni unità i gradi di appartenenza definitivi attraverso il metodo delle k-medie sfocato non modificato, si sceglie il parametro β in base all'analisi dei risultati ottenuti e considerando che il massimo valore che β può assumere è:

$$\beta_{max} = min\left\{\min_i \max_j \mu_{ij}, \min_j \max_i \mu_{ij}\right\}$$

Dopo aver stabilito il valore di β tramite la formula $w_{ij} = \frac{\mu_{ij}}{\sum_{j=1}^{K} \mu_{ij}}$ esposta precedentemente, si procede alla normalizzazione della funzione di appartenenza ottenendo così i nuovi valori dei gradi di appartenenza.

Il miglioramento nella scelta di β introdotto da Kamel e Selim consente di risolvere il problema della scelta di un valore troppo elevato del parametro in questione. Inoltre è possibile confrontare le diverse classificazioni ottenute considerando diversi valori di β, poiché quest'ultimo viene utilizzato soltanto alla fine del procedimento e quindi non è necessario ripetere per ogni valore di β l'intera procedura. Ciò consente un notevole risparmio di tempo.

3.3.2.2 Il metodo delle k-medie semisfocato

I metodi precedentemente esposti costituiscono dei miglioramenti al metodo delle k-medie sfocato che consentono di superare alcune note problematiche. Tuttavia non riescono a risolvere un altro importante difetto: il metodo delle k-medie non considera il caso in cui un'unità non appartiene per niente ad uno o più cluster o al massimo appartiene soltanto ad uno di essi. Infatti per ciascuna unità viene determinato il grado di appartenenza ad ogni cluster che non può essere nullo.

Si può presentare però il caso, anche se raramente, che una unità coincida con il centro di un cluster. L'unità viene attribuita interamente al cluster in questione e il grado di appartenenza agli altri cluster viene posto uguale a 0 al fine di rispettare il vincolo secondo il quale per ciascuna unità la somma dei gradi di appartenenza deve essere pari a 1.

Il miglioramento che verrà esposto in questo paragrafo considera, oltre al caso di unità coincidenti con il centro del cluster, anche gli altri casi di unità che potrebbero essere classificate come totalmente appartenenti ad un solo cluster.

Il metodo delle k-medie semisfocato (Iacovacci, 1997) permette di classificare come totalmente appartenenti ad un gruppo quelle unità che hanno rispetto al centro del cluster

stesso una distanza ragionevolmente piccola, ma che hanno anche una distanza ragionevolmente grande da tutti gli altri cluster.

Sarebbe infatti errato attribuire una unità totalmente ad un cluster nel caso in cui tale unità fosse vicina al centro del cluster in questione, il quale è a sua volta vicino al centro di un altro cluster. Ragionevolmente sarebbe più giusto attribuire l'unità per una parte al primo cluster e per la parte restante all'altro cluster.

In base al metodo delle k-medie semisfocato l'unità *i*-esima è assegnata totalmente al cluster *j*-esimo se la loro distanza d_{ij} è minore di (1/α)*D, dove α è un parametro maggiore di 1 determinato a priori e D è la distanza tra il centro del cluster *j*-esimo e il centro del cluster ad esso più vicino. Anche in questo caso per scegliere α è utile utilizzare l'indice di sfocatura *I* utilizzato precedentemente

L'algoritmo delle k-medie semisfocato è il seguente:

1. Si sceglie il numero K di cluster in cui suddividere le unità, i parametri *m* e α. Si suddividono successivamente le unità statistiche ottenendo in questo modo la matrice iniziale $U^{(0)}$;
2. Si calcolano i centri dei cluster utilizzando la formula:

$$V_j^{(0)} = \frac{\sum_{i=1}^{N} (\mu_{ij})^m z_i}{\sum_{i=1}^{N} (\mu_{ij})^m}$$

3. A questo punto si costruisce la matrice $D_1^{(0)} = [d_{ij}]$ delle distanze delle unità dai centri dei cluster, con $i = 1 \ldots n$ e $j = 1 \ldots K$, e la matrice $D_2^{(0)} = [d_{jq}]$ delle distanze tra i centri dei cluster, con $j,q = 1 \ldots K$ e $j \neq q$.
4. Successivamente si calcola la nuova matrice $U^{(1)}$ utilizzando le seguenti regole:
   I. Se c'è qualche gruppo Y per cui $d_{iy} = 0$, si pone $\mu_{iy} = 1$ e $\mu_{ij} = 0$ per tutti i $j \neq y$;
   II. Se invece c'è qualche gruppo Y per cui $d_{iy} < (1/\alpha)*d_{yj,}$ con $y \neq j$, si pone $\mu_{iy} = 1$ e $\mu_{ij} = 0$ per tutti i $j \neq y$;
   III. Nel caso in cui nessuna delle due precedenti condizioni è soddisfatta, si utilizza la formula seguente:

$$\mu_{ij} = \frac{1}{\sum_{q=1}^{K}\left(\frac{d_{ij}}{d_{iq}}\right)^{\frac{2}{(m-1)}}}$$

5. A questo punto si calcola la differenza $|U^{(1)} - U^{(0)}|$: se risulta minore di δ si interrompe la procedura, altrimenti si ritorna al punto 2.
6. Se infine si desidera eliminare eventuali valori troppo bassi del grado di appartenenza allo scopo di rendere più facilmente interpretabile la classificazione finale, è possibile ricorrere al metodo TFCM spiegato precedentemente. Dunque scelto un opportuno valore β, se $\mu_{ij} < \beta$, si pone $\mu_{ij} = 0$ allo scopo di rispettare il vincolo secondo il quale la somma dei gradi di appartenenza per ciascuna unità deve essere pari a 1. Successivamente di esegue la normalizzazione della funzione di appartenenza.

3.3.2.3 Algoritmo di Gustafson-Kessel

L'algoritmo di Gustafson-Kessel (Gustafson and Kessel, 1979) rappresenta un'estensione dell'algoritmo standard delle k-medie sfocato. Esso prevede l'introduzione di una "norma-distanza adattata" in modo da identificare i cluster con differenti forme geometriche all'interno del data set.

Ogni cluster ha una sua matrice delle norme A_j. Tali matrici sono utilizzate come variabili di ottimizzazione permettendo ad ogni cluster di adattare la norma-distanza alla struttura locale dei dati.

La matrice delle norme secondo l'algoritmo classico era ottenuta semplicemente utilizzando la formula:

$$D_{ij}^2 = \left(z_i - v_j\right)^T \left(z_i - v_j\right)$$

La funzione obiettivo dell'algoritmo di Gustafson-Kessel da minimizzare è la seguente:

$$J_m(Z, U, v, \{A_j\}) = \sum_{j=1}^{K}\sum_{i=1}^{N}\left(\mu_{ij}\right)^m \left(d_{ij}\right)^2$$

Come si può notare, rispetto alla funzione obiettivo classica compare anche il parametro A_j.
La funzione obiettivo però non può essere minimizzata direttamente utilizzando le matrici A_j così come sono, ma devono prima essere limitate in qualche modo.

Solitamente il determinante delle matrici A_j viene limitato ad un valore prefissato:

$$|A_j| = \rho_j > 0, \text{ con } \rho_j > 0 \quad \forall_j$$

In assenza di conoscenza sul fenomeno da analizzare ρ_j viene semplicemente posto uguale a 1 per goni cluster.
Far variare le matrici A_j mantenendo costante il loro determinante corrisponde ad ottimizzare la forma dei cluster mantenendo costante il loro volume.
Utilizzando il metodo dei moltiplicatori di Lagrange, le matrici A_j possiamo calcolarle in questo modo:

$$A_j = \left[\rho_j det(F_j)\right]^{1/N} F_j^{-1}$$

F_j rappresenta la matrice sfumata di covarianza del *j*-esimo cluster ed è calcolata come segue:

$$F_j = \frac{\sum_{j=1}^{N} (\mu_{ij})^m (z_i - v_j)(z_i - v_j)^T}{\sum_{i=1}^{N} (\mu_{ij})^m}$$

Sostituendo queste due ultime formule alla formula classica per calcolare la matrice delle distanze si ottiene una distanza di Mahalanobis generalizzata dove la covarianza è pesata attraverso i gradi di appartenenza nella matrice U.
L'algoritmo di Gustafson-Kessel come è possibile notare risulta abbastanza più complicato dal punto di vista computazionale poiché ad ogni iterazione è necessario ricalcolare la matrice delle covarianze dei cluster.

Una volta scelto il numero di cluster $1 < K < N$, l'esponente *m*, il grado di errore massimo δ, il parametro ρ_j e inizializzata la matrice della partizione iniziale $U^{(0)}$, si procede con l'algoritmo, formulato come segue:

1. Si calcolano i centri dei cluster utilizzando la formula:

$$V_j^{(l)} = \frac{\sum_{i=1}^{N}\left(\mu_{ij}^{(l-1)}\right)^m z_i}{\sum_{i=1}^{N}\left(\mu_{ij}^{(l-1)}\right)^m} \quad \text{con } 1 \leq j \leq K$$

2. Si calcola la matrice di covarianza F_j per ogni cluster:

$$F_j = \frac{\sum_{j=1}^{N}\left(\mu_{ij}^{(l-1)}\right)^m \left(z_i - v_j^l\right)\left(z_i - v_j^l\right)^T}{\sum_{i=1}^{N}\left(\mu_{ij}^{(l-1)}\right)^m}$$

3. Si calcolano le distanze $D_{ijA_j}^2$:

$$D_{ijA_j}^2 = \left(z_i - v_j^{(l)}\right)^T \left\{\left[\rho_j det\left(F_j\right)\right]^{1/N} F_j^{-1}\right\}\left(z_i - v_j^{(l)}\right)$$

4. Si calcola la nuova matrice $U^{(l)}$ della partizione utilizzando le stesse regole del metodo delle k-medie sfocato:

   I. Se c'è qualche gruppo Y per cui $d_{iy} = 0$, si pone $\mu_{iy} = 1$ e $\mu_{ij} = 0$ per tutti i $j \neq y$;

   II. Se la precedente condizione non è soddisfatta si utilizza la formula seguente:

$$\mu_{ij} = \frac{1}{\sum_{q=1}^{K}\left(\frac{d_{ijA_j}}{d_{iqA_j}}\right)^{\frac{2}{(m-1)}}}$$

5. Si calcola la differenza fra quanto ottenuto all'ultima e alla penultima iterazione utilizzando un'opportuna norma:

$$|U^{(l)} - U^{(0)}| \leq \delta$$

Se tale differenza è minore del parametro δ stabilito a priori ci si ferma, altrimenti si ripete l'algoritmo.

Prima di concludere il discorso sull'algoritmo di Gustafson-Kessel è utile spendere alcune parole sulla matrice di covarianza dei cluster.
F_j ci fornisce informazioni riguardo la forma e l'orientamento del cluster *j*-esimo.
Il rapporto tra le lunghezze degli assi dell'elisse è dato dal rapporto fra le radici quadrate degli autovalori di F_j. L'orientamento degli assi è dato invece dagli autovettori di F_j.

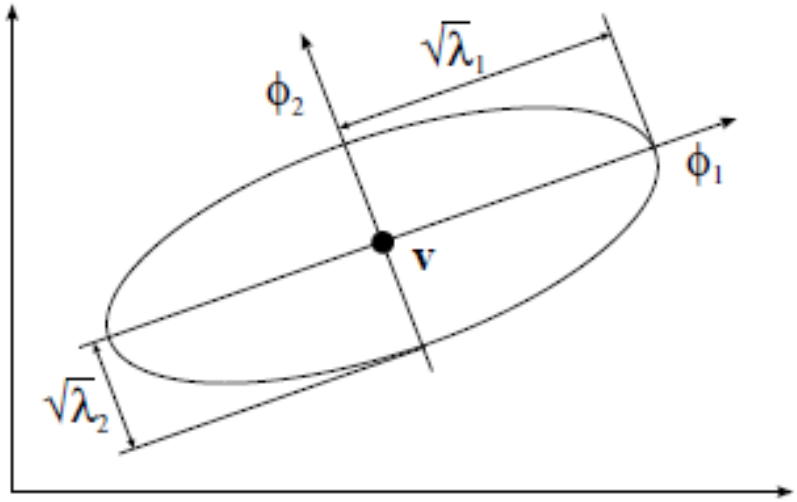


Figura 3.15 - λ_1 e λ_2 rappresentano i due autovalori, mentre $Ø_1$ e $Ø_2$ sono gli autovettori del j-esimo cluster.

L'equazione che definisce l'ellisse è la seguente:

$$(z - v)^T F^{-1}(x - v) = 1$$

L'algoritmo di Gustafson-Kessel può essere impiegato per individuare i cluster lungo sottoinsiemi lineari dello spazio dei dati. Tali cluster sono rappresentati attraverso ellissi schiacciati che possono essere considerati come degli iperpiani.
Utilizzando lo stesso data set dell'esempio riportato per l'algoritmo delle k-medie standard, è possibile notare dalle figure che seguono come l'algoritmo appena analizzato, una volta applicato, identifichi correttamente la forma e l'orientamento dei cluster all'interno del data set in esame.

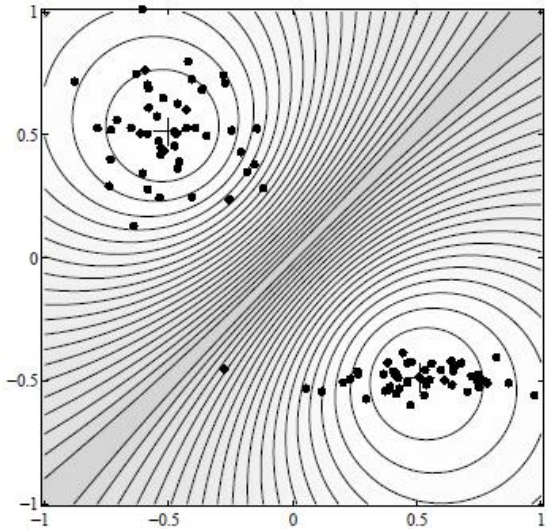


Figura 3.16 - Algoritmo standard

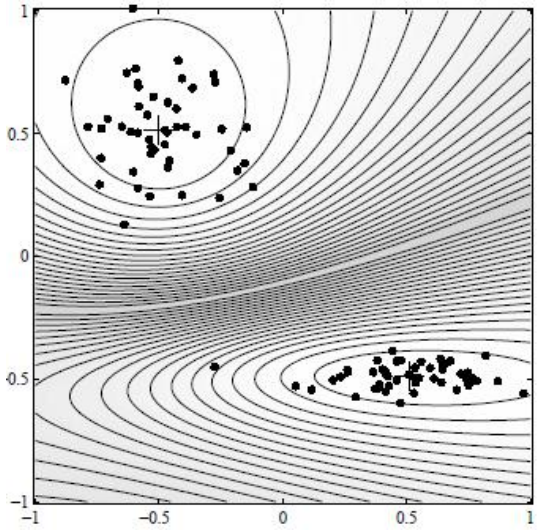


Figura 3.17 - Algoritmo di Gustafson-Kessel

3.3.3 Altri metodi non gerarchici

Oltre al metodo delle k-medie sfocato esistono molti altri metodi non gerarchici di classificazione sfocata. Si possono trovare infatti molti miglioramenti dell'algoritmo delle k-medie, alcuni dei quali sono stati esposti nei paragrafi precedenti, e alcuni metodi simili a quest'ultimo. Nei successivi paragrafi verranno presentati brevemente e senza scendere nel dettaglio dei relativi algoritmi due di questi metodi.

3.3.3.1 Il metodo FANNY

In metodo qui presentato è un algoritmo utilizzato all'interno del programma R (ambiente di sviluppo per l'analisi statistica). Esso produce delle classificazioni sfocate utilizzando qualsiasi tipo di dati, quindi sia numerici che misure di dissimilarità.

Tale metodo assomiglia all'algoritmo delle k-medie sfocato: fissato il numero di gruppi in cui suddividere le unità e fornita una partizione iniziale assegnando a priori i valori della funzione di appartenenza, la classificazione finale si ottiene attraverso una procedura di ottimizzazione della funzione obiettivo.

La differenza con il metodo delle k-medie sfocato sta proprio nella funzione obiettivo che nel metodo FANNY (Kaufmann,1990) diventa la seguente:

$$C = \sum_{j=1}^{K}\left[\sum_{i=1}^{N}\sum_{s=1}^{N}(\mu_{ij})^2(\mu_{sj})^2(d_{is})/2\sum_{s=1}^{N}(\mu_{sj})^2\right]$$

d_{is} rappresenta la distanza fra l'unità *i*-esima e l'unità **s**-esima, mentre μ_{ij} è il grado di appartenenza dell'unità *i*-esima al gruppo *j*-esimo.

La classificazione finale si ottiene minimizzando la funzione obiettivo sopra indicata, applicando il metodo dei moltiplicatori di Lagrange sotto le condizioni di Kuhn e Tucker e considerando i vincoli:

I. $\mu_{ij} \geq 0$ con $i = 1,\ldots N$; $j = 1\ldots K$

II. $\sum_{j=1}^{K} \mu_{ij} = 1$ con i = 1,…N

La soluzione si ottiene attraverso un processo di ottimizzazione e corrisponde solitamente ad una soluzione di ottimo locale.

Un'altra differenza con il metodo delle k-medie sfocato consiste nel fatto che il metodo FANNY pone l'esponente uguale a 1 nel calcolo della distanza, mentre nel metodo delle k-medie sfocato viene considerato il quadrato di questa.

Inoltre il parametro *m* viene qui posto uguale a 2, mentre nel primo caso poteva variare ($m > 1$) ed era possibile quindi sceglierne il valore migliore. Tale parametro viene posto uguale a questo valore poiché secondo l'autore è ponendo $m = 2$ che si ottengono i risultati

migliori, anche se comunque variando l'esponente in questione si ottengono delle classificazioni più o meno sfocate come avveniva per il metodo delle k-medie sfocato.
A titolo di esempio vengono ora presentati i risultati ottenuti applicando prima il metodo delle k-medie sfocato e poi il metodo FANNY. Dalla figura sottostante si nota subito che entrambi i metodi producono, a parità di *m*, le stesse classificazioni.

Unità	K-medie sfocato			FANNY		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
1	0.96	0.01	0.03	0.87	0.06	0.07
2	0.98	0.01	0.01	0.88	0.05	0.07
3	0.98	0.01	0.01	0.93	0.03	0.04
4	0.96	0.01	0.03	0.86	0.06	0.08
5	0.98	0.01	0.01	0.87	0.06	0.07
6	0.50	0.35	0.15	0.42	0.35	0.23
7	0.02	0.96	0.02	0.08	0.82	0.10
8	0.01	0.98	0.01	0.06	0.87	0.07
9	0.01	0.97	0.02	0.06	0.86	0.08
10	0.01	0.98	0.01	0.06	0.87	0.07
11	0.01	0.97	0.02	0.06	0.86	0.08
12	0.02	0.96	0.02	0.07	0.84	0.09
13	0.37	0.22	0.51	0.36	0.27	0.37
14	0.03	0.02	0.95	0.12	0.08	0.80
15	0.01	0.01	0.98	0.08	0.07	0.85
16	0.02	0.03	0.95	0.10	0.10	0.80
17	0.01	0.01	0.98	0.08	0.06	0.86
18	0.00	0.00	1.00	0.04	0.04	0.92
19	0.01	0.01	0.98	0.07	0.07	0.86
20	0.03	0.02	0.95	0.10	0.08	0.82
21	0.01	0.01	0.98	0.07	0.06	0.87
22	0.02	0.02	0.96	0.09	0.09	0.82

Figura 3.18 - Classificazioni ottenute con i metodi delle k-medie sfocato e FANNY.
Fonte: Iacovacci, 1995.

3.3.3.2 Il metodo MND2

L' MND2 (Roubens 1978) è un altro metodo di classificazione sfocata non gerarchico, simile ai metodi precedenti, ma che si differenzia da questi proprio per la seguente funzione obiettivo:

$$R = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{s=1}^{N} (\mu_{ij})^2 (\mu_{sj})^2 (d_{is})$$

Tale funzione è simile a quella adottata nel metodo FANNY ma, come si può notare, la principale differenza consiste nel non considerare il denominatore. Il metodo MND2 nel caso di molte unità statistiche tende a distorcere i risultati a causa della funzione obiettivo adottata non assicurando di conseguenza una convergenza ottimale. Quindi tale metodo risulta molto meno efficiente e preferibile rispetto agli altri metodi considerati sino ad ora.

3.4 Conclusioni

I metodi di classificazione sfocata descritti all'interno di questo capitolo sono tutte tecniche di cui il ricercatore si dovrebbe avvalere nel corso delle sue indagini. Queste tecniche, comprese quelle che non sono state trattate all'interno di questo lavoro, dovrebbero essere accostate insieme alle tecniche classiche, in modo da prendere da ciascuna di esse quello che meglio possono offrire in uno specifico contesto di ricerca.
I metodi sfocati, a differenza di quelli tradizionali, risultano di grande aiuto nel caso di situazioni in cui sono presenti unità statistiche non perfettamente includibili in un determinato gruppo o unità affette da possibili errori di misurazione. In tale situazione sarebbe infatti errato forzare una unità ad appartenere ad un cluster piuttosto che ad un altro, questo perché otterremmo un risultato non perfettamente corrispondente alla situazione che stiamo analizzando. Con una classificazione classica, in sede di analisi dei risultati, potremmo dedurre soltanto che se una unità è stata assegnata ad un determinato cluster vuol dire che possiede per la maggior parte caratteristiche corrispondenti a quel gruppo, ma non sapremmo mai in che parte. Una classificazione sfocata, assegnando ad ogni unità il suo grado di appartenenza a ciascun gruppo, ci permette di superare questo inconveniente ed ottenere così un risultato che ci fornisca più informazioni circa la classificazione finale.
Oltre agli indubbi vantaggi i metodi sfocati presentano anch'essi dei limiti come per esempio l'elevato grado di sfocatura che cresce all'aumentare del numero di gruppi, il raggiungimento di un punto di ottimo locale e non globale della funzione obiettivo, l'assenza di una base teorica per la scelta del parametro *m* e molti altri.
Per risolvere alcuni di questi problemi nel corso del tempo sono stati proposti dei miglioramenti, alcuni dei quali presentati nei paragrafi precedenti, ma come si può

dedurre gli ambiti di sviluppo di questa tipologia di analisi dei dati sottostante ad una logica sostanzialmente recente sono ampi.

Capitolo 4

UN'ANALISI EMPIRICA

4.1 Premessa

Il seguente capitolo è sviluppato a partire da un questionario sottoposto agli agriturismi del Veneto dal titolo *"Agriturismi e sostenibilità nella regione Veneto",* contente un focus specifico volto ad identificare i principali strumenti adottati dalle aziende agrituristiche per la gestione sostenibile della loro attività allo scopo di salvaguardare l'ambiente.

Il questionario è stato sottoposto attraverso lo strumento web "Lime Survey", il quale permette di sottoporre questionari on-line in modo completamente gratuito.

Inizialmente l'intenzione del questionario era quella di definire quale fosse la situazione attuale in termini di sostenibilità ambientale dell'attività degli agriturismi nella regione Veneto, suddividendoli in un determinato numero di gruppi in base a determinate caratteristiche e secondo un determinato criterio, con l'obiettivo di trarre delle conclusioni su di essi.

Purtroppo il riscontro è risultato essere molto deludente: dei circa 1178 agriturismi presenti attualmente[8], soltanto l'8,6 % ha risposto alle domande del questionario, cioè appena 101 agriturismi. Di conseguenza un'analisi volta a definire la situazione a livello regionale sulla sostenibilità ambientale degli agriturismi risulta essere priva di senso.

Si è deciso quindi di effettuare un'analisi di classificazione volta a identificare gruppi di agriturismi in base alla similarità del loro profilo descritto da un insieme di caratteri, in questo caso caratteri inerenti l'aspetto della sostenibilità.

Verranno infatti utilizzate le risposte ad alcune domande incentrate sugli strumenti adottati dagli agriturismi per la salvaguardia e la gestione sostenibile dell'ambiente.

[8] http://www.agriturismo.st/it/Italia/Veneto/. Situazione ad agosto 2014.

L'obiettivo vero e proprio di questo capitolo è quello di mostrare l'iter di svolgimento di un'analisi dei cluster; in particolar modo verranno messe a confronto due tipologie di classificazione, quella classica e quella sfocata, evidenziandone pregi e difetti di una e dell'altra.

Il capitolo è sviluppato in due parti distinte:

- La prima contenente un'analisi completa delle risposte alle domande del questionario sulla sostenibilità sottoposto agli agriturismi del Veneto;
- La seconda contenente un percorso di analisi dei cluster che inizierà dall'approccio classico per poi passare all'approccio "sfumato". Sarà proprio quest'ultimo ad essere al centro dell'attenzione in questa parte del capitolo.

Tutto questo verrà effettuato con il supporto del programma "R", un ambiente di sviluppo per l'analisi statistica che permetterà di svolgere in automatico tutti i calcoli necessari per l'applicazione degli algoritmi di clustering e in parte con l'ausilio di un altro software di statistica, cioè SPSS.

4.2 Analisi del questionario

Nella prima parte di questo capitolo verranno analizzate le domande presenti nel questionario per costruire un profilo che contenga gli elementi più importanti e interessanti alla completa descrizione di ciascuna di esse e che servirà come base alla successiva analisi dei cluster.
Le domande saranno analizzate una ad una singolarmente e quando possibile ed interessante messe in relazione tra di loro.
Come già detto nella premessa, il questionario è stato sottoposto a tutti gli agriturismi del Veneto nel periodo febbraio-aprile, attraverso lo strumento web "Lime Survey". Nonostante i numerosi solleciti inviati agli agriturismi la percentuale di risposte è risultata molto bassa: appena l'8,6%, escludendo il fatto che alcuni agriturismi non hanno risposto ad alcune domande della parte generale del questionario.

Il questionario, composto da un totale di 49 domande, è così strutturato:

Informazioni generali sugli agriturismi (24 domande)

- Genere, età e titolo di studio del titolare
- Dimensione economica, percentuale dei ricavi derivanti da attività agrituristica
- Attività integrative del reddito aziendale
- Numero addetti e media giornate lavorative
- Stima presenze in un anno
- Spesa media per persona
- Servizio di ristorazione, coperti registrati in un anno
- Tipo di turista che frequenta l'agriturismo
- Canali di promozione e vendita dei prodotti/servizi offerti dall'agriturismo
- Eventi, attività proposte dall'agriturismo
- Importanza degli eventi e delle attività nella promozione dell'azienda agrituristica
- Esigenze di finanziamento

Focus sulla sostenibilità (25 domande)

- Pratiche per la riduzione dei rifiuti
- Raccolta differenziata
- Adozione di strutture e facilitazioni per l'accesso e l'ospitalità di persone disabili
- Misure per la diminuzione del consumo idrico, elettrico e termico
- Misure per la prevenzione degli sprechi
- Misure per la gestione ambientale
- Accesso alla struttura attraverso i mezzi del trasporto pubblico
- Informazioni agli ospiti su come raggiungere la struttura e le altre destinazioni con i mezzi pubblici
- Presenza di impianti per la produzione di energia rinnovabile
- Misure di sostegno allo sviluppo rurale
- Certificazioni ambientali

Di seguito l'analisi completa di tutte le domande ad esclusione delle domande relative alle giornate medie lavorate durante l'anno, al numero di clienti/coperti registrati all'anno e all'andamento dei clienti/coperti negli ultimi 3 anni poiché tali domande hanno ricevuto risposta da appena 1/3 degli agriturismi che hanno risposto al questionario.

4.2.1 Informazioni generali sugli agriturismi

1. Genere ed età media del/i titolare/i

Analizzando il genere del titolare degli agriturismi che hanno risposto al questionario, dal grafico si può vedere subito che quelli intestati a soli uomini sono di gran lunga di più di quelli intestati a sole donne: il 62,9% contro il 37,1%.

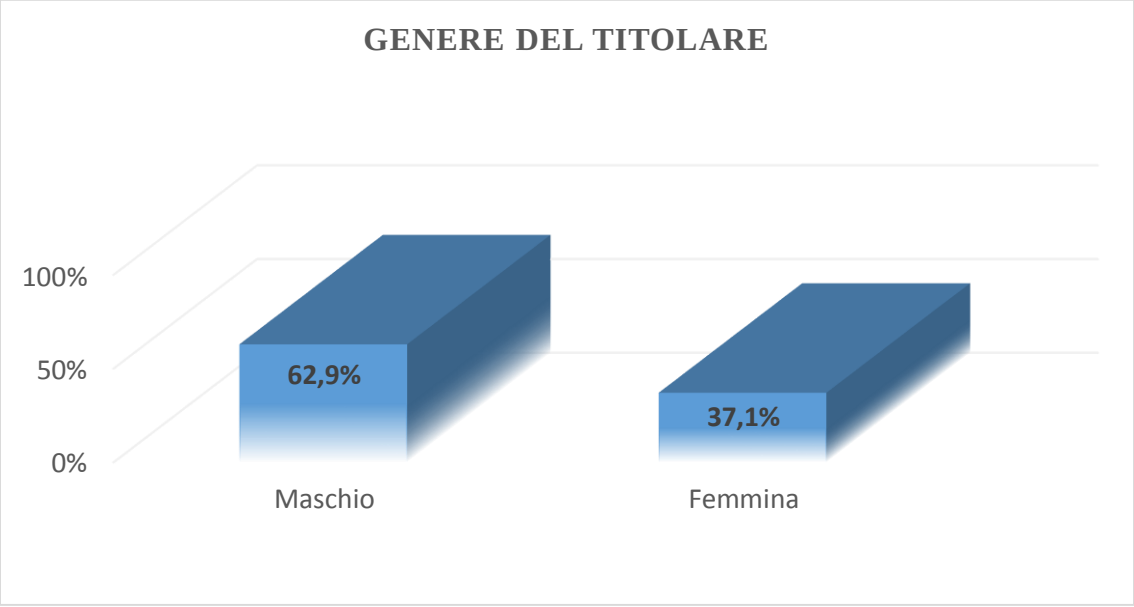


Figura 4.1 - Genere del titolare. Fonte: elaborazione personale.

Confrontando questi dati con quelli forniti dall' Istat relativamente all'anno 2012, per quanto possibile visto l'esiguo numero di agriturismi che hanno risposto al questionario, si è potuto osservare che la situazione si avvicina abbastanza. Secondo quanto dichiarato nella relazione sul convegno del 1 aprile 2014 "Donne e agricoltura, un connubio di

qualità", presente sul sito dell'Istat, l'incidenza delle aziende agrituristiche intestate alle donne sul totale nel Veneto al 2012 era del 30%, in aumento di anno in anno di circa il 3,6%.
Per cui secondo quanto raccolto attraverso gli agriturismi che hanno risposto al questionario, i dati sembrano rispecchiare quelli forniti dall'analisi Istat, considerando in questi due ultimi anni un trend costante di aumento dell'incidenza delle donne sugli uomini.

Per quanto riguarda l'età dei titolari i dati raccolti sono i seguenti:

Età dei titolari							
	Età media (mediana)	Età Min	1° quartile (25%)	2° quartile (50%)	3° quartile (75%)	Età Max	Moda
Maschi	50	27	42	50	60	72	50
Femmine	46	26	37	46	55	77	51

Tabella 4.1 - Età dei titolari. Fonte: elaborazione personale.

A livello di età media, età minima ed età modale non ci sono differenze sostanziali tra maschi e femmine. Una differenza si verifica invece per l'età massima registrata che, per le femmine, raggiunge i 77 anni, contro i 72 anni dei maschi.
È facile notare dall'analisi dei quartili che relativamente ai titolari maschi che hanno risposto al questionario:

- il 25% ha un'età compresa fra i 27 e i 41 anni;
- il 25% ha un'età compresa fra i 42 e i 49 anni;
- il 25% ha un'età compresa fra i 50 e i 59 anni;
- il 25% ha un'età compresa fra i 60 e i 72 anni.

2. Titolo di studio del/i titolare/i

Per quanto riguarda il titolo di studio dei titolari maschi (114 titolari maschi in totale), come si può notare dal grafico sottostante, prevalgono quelli in possesso di licenza superiore.

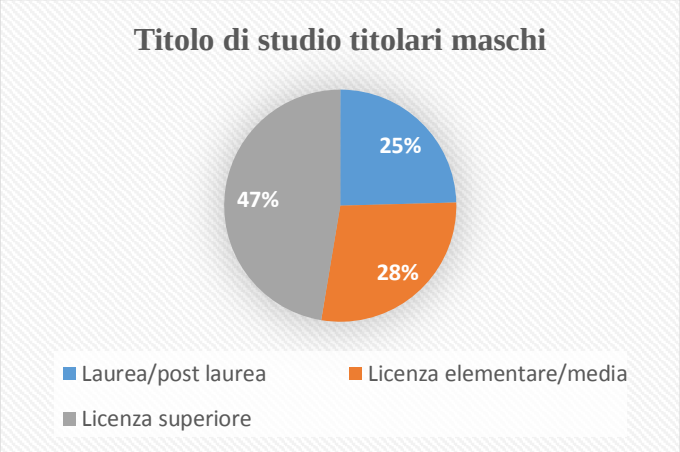


Figura 4.2 - Titolo di studio titolari maschi. Fonte: elaborazione personale.

Per i titolari donna (48 titolari femmina) invece la situazione è la seguente:

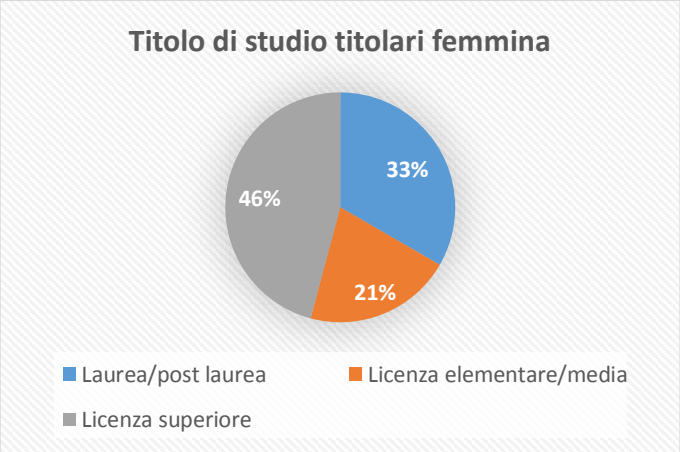


Figura 4.3 - Titolo di studio titolari femmina. Fonte: elaborazione personale.

3. Dimensione economica complessiva

La distribuzione dei ricavi delle aziende agrituristiche in Veneto che hanno risposto al questionario risulta la seguente:

- Il 12,9% realizza un fatturato annuo inferiore ai 20.000 euro;
- Il 20,8% realizza un fatturato annuo compreso fra i 20.000 e i 40.000 euro;
- Il 25,7% realizza un fatturato annuo compreso fra i 40.000 e i 100.000 euro;
- Il 32,7% realizza un fatturato annuo compreso fra i 100.000 e i 500.000 euro;
- Circa il 5% realizza un fatturato annuo superiore ai 500.000 euro.

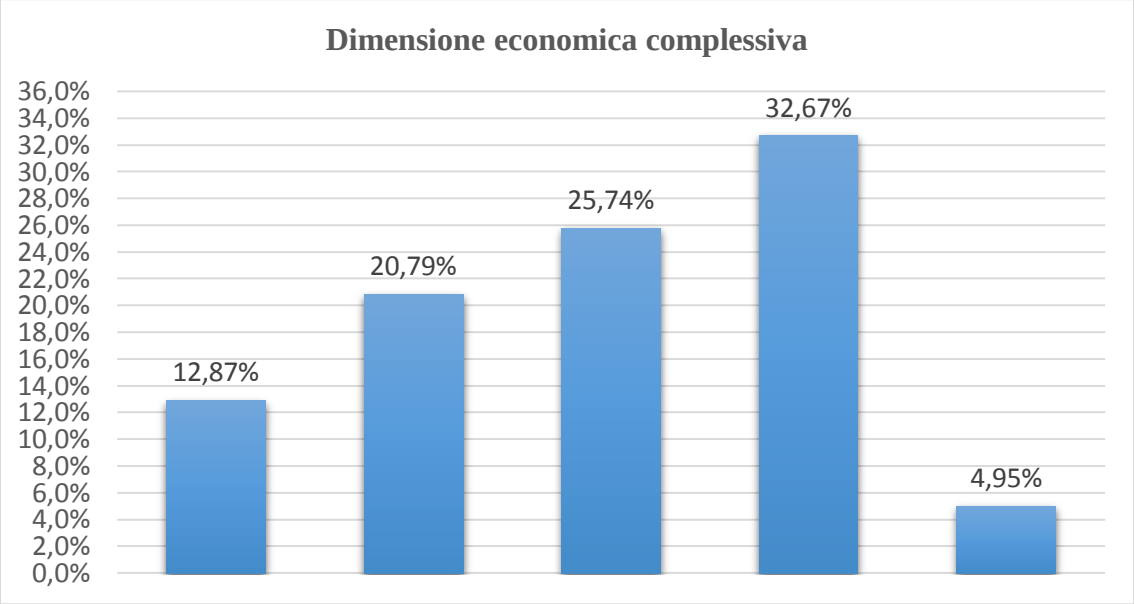


Figura 4.4 - Dimensione economica complessiva. Fonte: elaborazione personale.

4. Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?

Come si può notare dal grafico sottostante, per quanto riguarda la percentuale di ricavi aziendali realizzata attraverso l'attività di agriturismo, fra le aziende intervistate prevalgono quelle che tramite l'attività di agriturismo realizzano oltre il 40% di ricavi. Quindi, possiamo dire che per la metà delle aziende intervistate tutte le attività classificabili come "agrituristiche" sono molto importanti al fine di generare un certo reddito. Per la restante metà invece risultano poco importanti se non indifferenti al fine

della generazione del reddito (c'è infatti il 15% delle aziende che realizza il 5% dei ricavi attraverso l'attività di agriturismo).

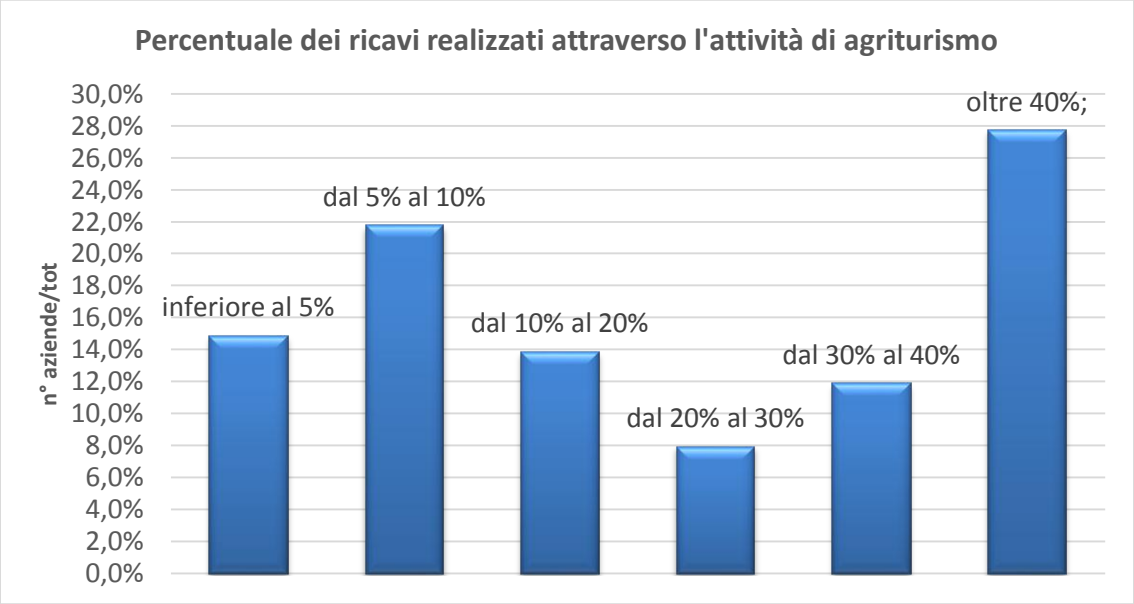


Figura 4.5 - Percentuale dei ricavi realizzati attraverso l'attività di agriturismo. Fonte: elaborazione personale.

5. Attività integrative del reddito

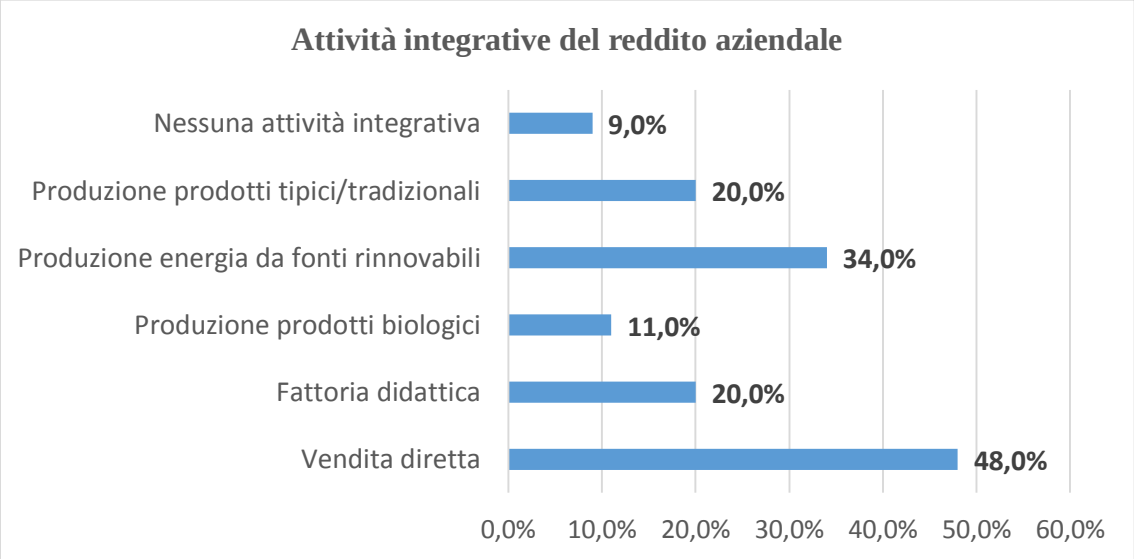


Figura 4.6 - Attività integrative del reddito aziendale. Fonte: elaborazione personale.

Fra le attività integrative del reddito aziendale quella più adottata dagli agriturismi che hanno risposto al questionario risulta essere la vendita diretta di prodotti (48%). Segue poi con il 34% la produzione di energia da fonti rinnovabili. A pari merito troviamo invece la prodizione di prodotti tipici/tradizionali e la fattoria didattica, attività adottate dal 20% degli agriturismi. Quest'ultima attività che ha lo scopo di avvicinare giovani e adulti alla natura risulta essere in continua crescita negli ultimi anni[9].

6. La struttura offre servizio di alloggio?

Come si può osservare dal grafico a torta sottostante il 67% degli agriturismi intervistati offre servizio di alloggio.

Figura 4.7 - Predisposizione del servizio di alloggio. Fonte: elaborazione personale.

[9] http://www.fattoriedidattiche.biz/

7. La struttura offre servizio di ristorazione?

Figura 4.8 – Predispozione del servizio di ristorazione. Fonte: elaborazione personale.

8. Che tipo di turista/cliente frequenta la sua azienda?

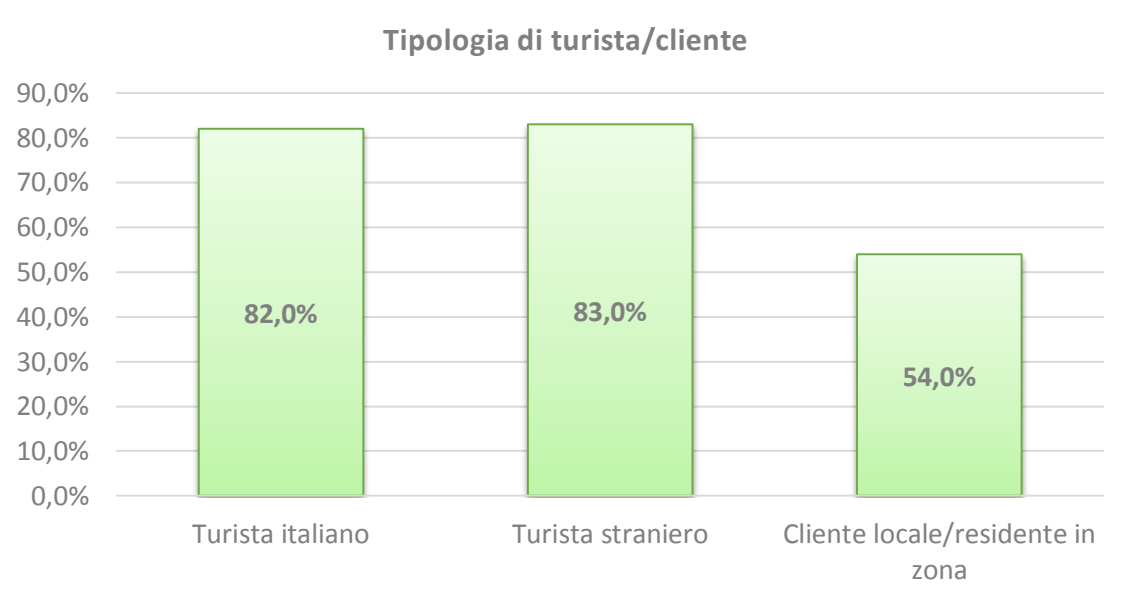


Figura 4.9 - Tipologia di turista/cliente ospitato. Fonte: elaborazione personale.

Come si può facilmente osservare, degli agriturismi intervistati, l'82% registra la presenza di turisti/clienti italiani. La presenza di turisti stranieri si attesta invece intorno all'83%.

Al di là di questi dati però, la cosa interessante da osservare è l'incidenza media della tipologia di turista/cliente sul totale:

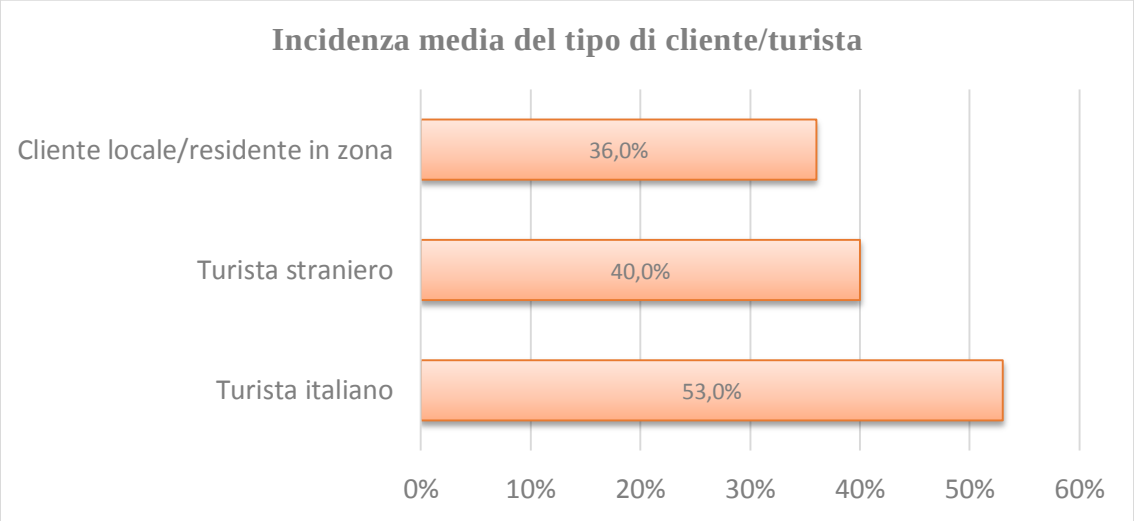


Figura 4.10 - Incidenza media del tipo di cliente/turista. Fonte: elaborazione personale.

Mediamente, per gli agriturismi intervistati, il cliente che incide di più sulle presenze totali registrate è quello italiano (53%); segue poi il turista straniero con il 40% e infine il cliente residente in zona con il 36%.

9. Quali canali utilizza maggiormente per la promozione e la vendita dei prodotti/servizi offerti dall'Agriturismo?

È possibile osservare dalle risposte date dagli agriturismi che i due canali indiscussi utilizzati per la promozione e la vendita dei prodotti e dei servizi sono quello del "passaparola" e quello del "sito internet dell'Agriturismo". Subito dopo troviamo l'utilizzo di altri siti Internet (Agenzie turistiche on-line, …), dépliant e volantini e infine, con percentuali di utilizzo molto basse, troviamo altri canali di comunicazione come le agenzie di viaggio, le manifestazioni e le guide turistiche. Oltre alle tipologie presenti nel grafico sottostante si sono verificati 3 casi di agriturismi che hanno utilizzato altri canali di comunicazione, cioè agenzie immobiliari, sms e portali di prenotazione.

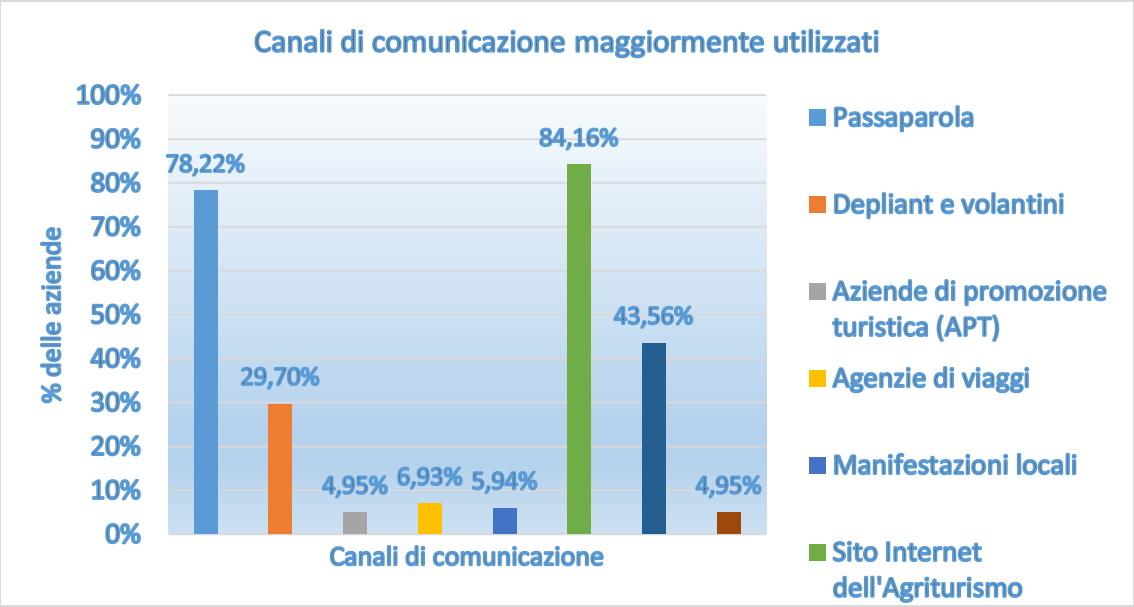


Figura 4.11 - Canali di comunicazione maggiormente utilizzati. Fonte: elaborazione personale.

È da sottolineare che gli Agriturismi mediamente utilizzano più canali di comunicazione contemporaneamente. Il grafico sottostante mostra che ben il 66% degli agriturismi intervistati utilizzano solitamente 3 diversi canali di comunicazione per promuovere e vendere i proprio prodotti/servizi:

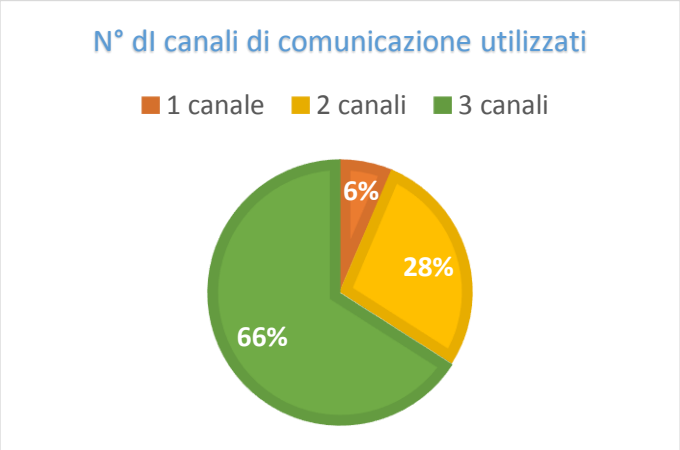


Figura 4.12 - Numero di canali di comunicazione utilizzati. Fonte: elaborazione personale.

10. La sua azienda propone eventi, attività, serate a tema? Se si, quali tipi di eventi/attività e quanto ritiene siano importanti nella promozione della sua azienda?

Ben il 61% degli agriturismi intervistati dichiara di non svolgere attività integrative per coinvolgere gli ospiti.

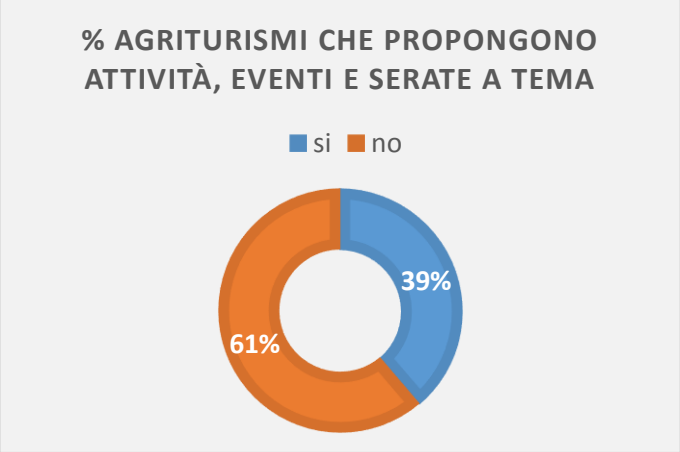


Figura 4.13 - Percentuali agriturismi che propongono attività, eventi e serate a tema. Fonte: elaborazione personale.

Per il restante 39% di agriturismi, come si può notare dal grafico sottostante, non c'è un'attività "preferita" che prevale sulle altre in modo significativo. Le attività che vengono proposte più spesso sono ovviamente le escursioni, i laboratori didattici e corsi di vario genere (es: cucina). Da notare che le attività sportive vengono proposte in pochissimi casi (solo 5 agriturismi su 40 propongono attività del genere); a mio avviso attività come tennis, calcetto, beach volley, minigolf e molte altre, se venissero proposte all'interno degli agriturismi potrebbero essere fonte di attrazione per molte persone interessate a fare un po' di movimento anche in vacanza, oltre a costituire un momento di svago e socializzazione.

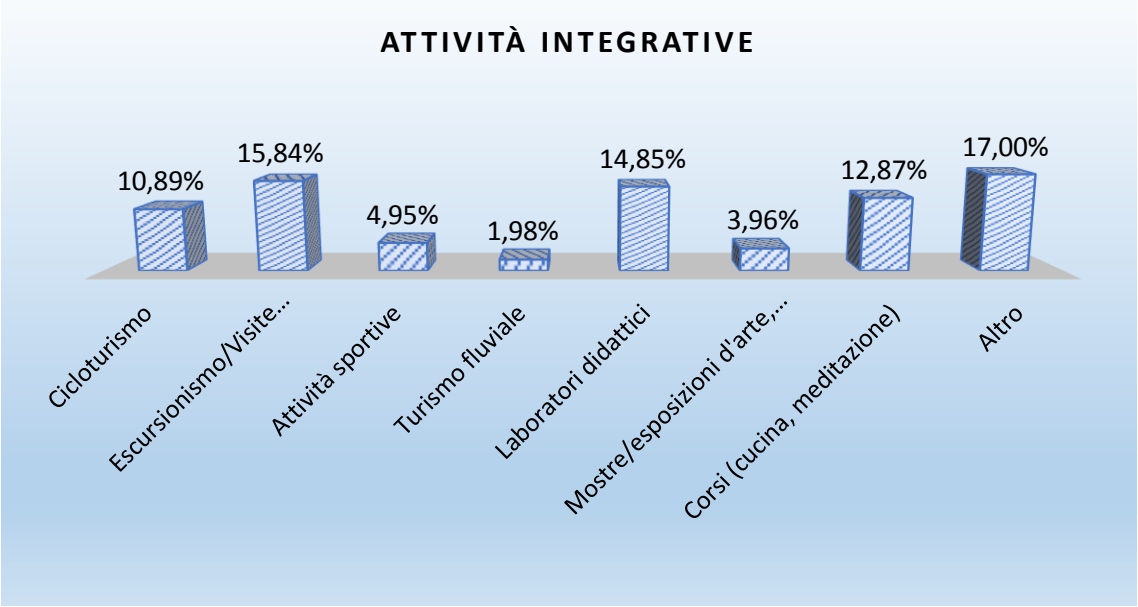


Figura 4.14 - Attività integrative. Fonte: elaborazione personale.

Fra le altre attività proposte dagli agriturismi troviamo pranzi con alimenti stagionali, musica, teatro, degustazioni guidate, fattorie didattiche, cene tradizionali sulla cultura culinaria del Veneto.

Infine, come mostrato dal seguente grafico a torta, gli agriturismi che propongono eventi ed attività integrative di coinvolgimento dichiarano di avere mediamente un buon ritorno in termini di promozione dell'agriturismo stesso. Infatti solo il 10 % ritiene tali attività non influenti nella promozione del proprio agriturismo.

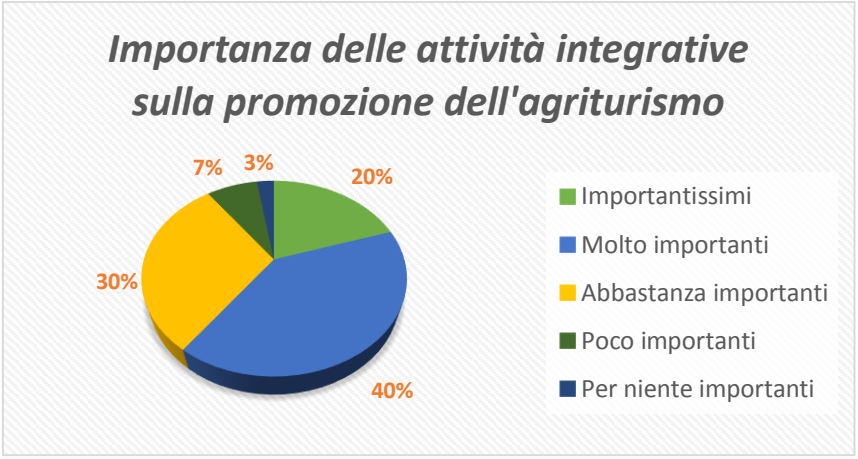


Figura 4.15 - Importanza delle attività integrative sulla promozione dell'agriturismo. Fonte: elaborazione personale.

11. Secondo la sua esperienza, quali sono le esigenze di finanziamento/sovvenzioni più urgenti?

Per gli agriturismi intervistati l'aspetto più importante per sostenere la propria attività consiste nelle attività di promozione. Un aspetto in contrasto, a mio parere, con quanto dichiarato nella precedente domanda.

Se infatti il 65% degli agriturismi dichiara che la promozione della propria attività risulta un'esigenza urgente per garantire una certa sostenibilità alla propria azienda e considerando che il 90% degli agriturismi che già propongono tali attività dichiara che queste sono importanti nella promozione della propria attività, perché il 61% degli agriturismi non svolge nessuna attività integrativa o evento al fine appunto di promuovere il proprio agriturismo?

Una risposta possibile potrebbe essere la poca informazione da parte delle istituzioni pubbliche preposte al turismo. Una maggiore informazione infatti porterebbe le aziende agrituristiche ad acquisire maggior consapevolezza circa l'importanza di tutte quelle attività integrative alla pura attività agrituristica al fine di promuovere l'azienda stessa.

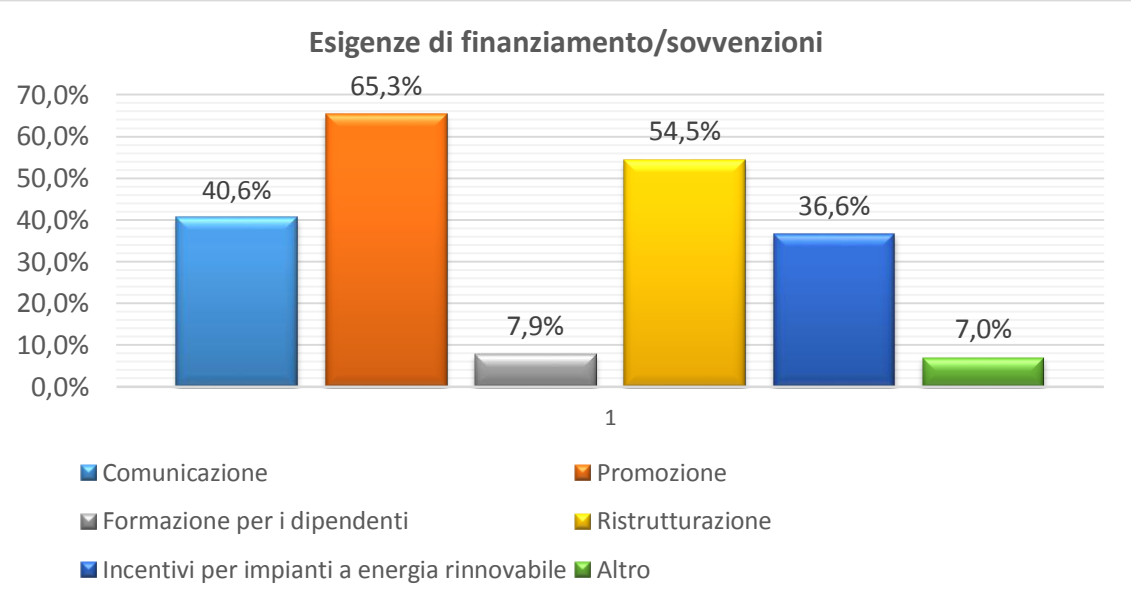


Figura 4.16 - Esigenze di finanziamento/sovvenzioni urgenti. Fonte: elaborazione personale.

Fra le altre esigenze di finanziamento/sovvenzioni troviamo richieste di attrezzature, ampliamento per nuove attività connesse e il miglioramento del supporto offerto dalle associazioni di categoria.

4.2.2 Focus sulla sostenibilità

12. Adottate pratiche per la riduzione dei rifiuti? Se si, quali?

Figura 4.17 - Adozione di pratiche per la riduzione dei rifiuti. Fonte: elaborazione personale.

La maggioranza degli agriturismi intervistati adotta sistemi di riduzione dei rifiuti, questo a sottolineare che rappresenta una pratica ritenuta molto importante per le aziende agrituristiche del Veneto.

Nel grafico sottostante vengono mostrate le principali pratiche utilizzate per la riduzione dei rifiuti da parte degli agriturismi che hanno risposto al questionario:

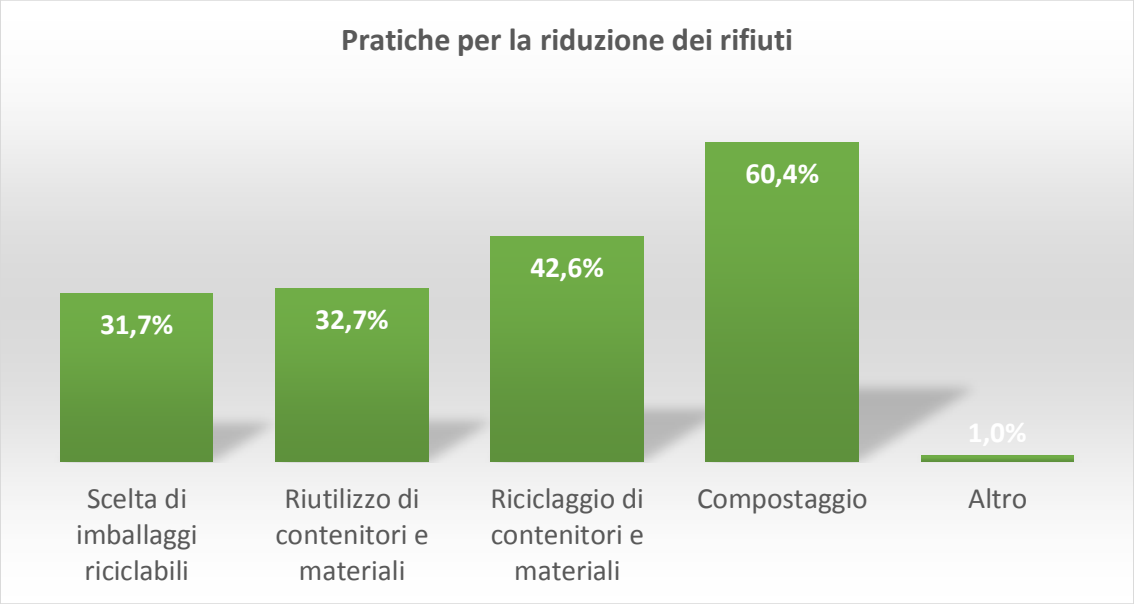


Figura 4.18 - Pratiche per la riduzione dei rifiuti.

La pratica più utilizzata risulta essere il compostaggio domestico, tramite l'utilizzo di appositi composter; subito dopo troviamo l'attenzione alla differenziazione e al riciclaggio dei materiali e contenitori.

Seguono poi il riutilizzo di contenitori e materiali e la scelta di imballaggi riciclabili. Fra le altre pratiche rientra l'uso di prodotti ecologici.

Analizzando infine le aziende più "virtuose", il numero di agriturismi che dimostra maggior impegno nell'impiego di pratiche volte alla riduzione dei rifiuti è abbastanza esiguo; infatti solo il 14,61% degli agriturismi mette in atto tutte le principali pratiche indicate nel questionario.

13. Presso la sua azienda praticate la raccolta differenziata dei rifiuti?

In questo caso la totalità degli agriturismi intervistati dichiara di praticare la raccolta differenziata dei rifiuti.

Questo è dimostrato anche dal fatto che il Veneto risulta attualmente la regione più "riciclona" d'Italia, con una raccolta differenziata pari al 62,5% del totale (Corriere della Sera, 31 luglio 2013).

14. La sua azienda è dotata di strutture e facilitazioni per l'accesso e l'ospitalità di persone con disabilità?

Dalle risposte date al questionario è emerso che ben il 77% delle aziende agrituristiche intervistate è dotata di strutture e facilitazioni per l'accesso e l'ospitalità di persone con disabilità. Un buon dato che comunque dovrebbe essere migliorato fino ad arrivare ad una percentuale totale degli agriturismi in modo da garantire un servizio minimo a quelle persone disabili ovunque esse soggiornino.

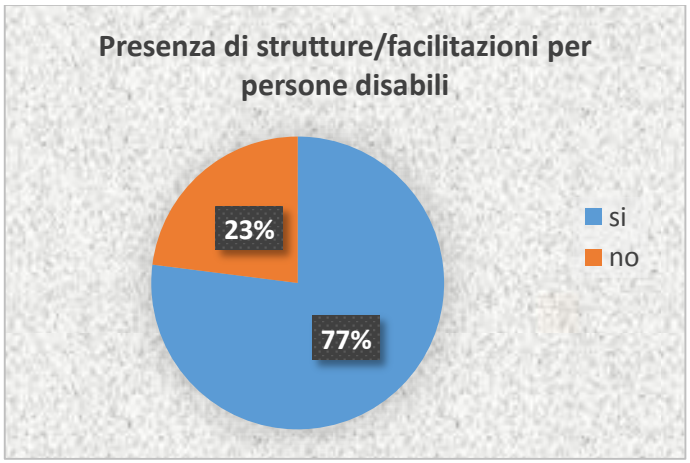


Figura 4.19 - Presenza di strutture/facilitazioni per persone disabili. Fonte: elaborazione personale.

15. Adottate misure per la gestione e la diminuzione del consumo elettrico?

Come si può vedere dal grafico sottostante il 91% degli agriturismi che hanno risposto al questionario adotta misure per la gestione e la diminuzione del consumo elettrico.

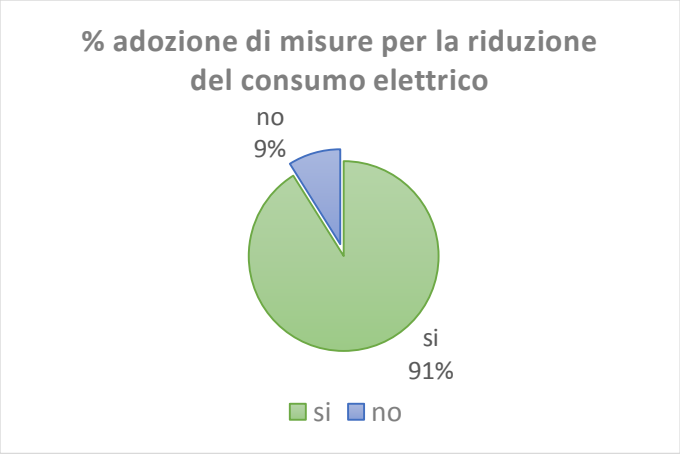


Figura 4.20 - Percentuale di adozione di misure per la riduzione del consumo elettrico. Fonte: elaborazione personale.

Analizzando ora in dettaglio quali sono le misure maggiormente adottate, dalla rappresentazione grafica seguente è agevole osservare che la misura più adottata, anche se non da tutti gli agriturismi, è l'utilizzo di lampadine a basso consumo energetico. Infatti l'uso di lampadine normali, a lungo andare, costituisce una delle principali fonti di spreco e consumo energetico. Basti pensare che normalmente una lampadina classica trasforma solo il 5% dell'energia in luce, mentre il restante 95% non è nient'altro che "spreco energetico"[10].

Seguono poi con l'utilizzo di elettrodomestici a classe di efficienza energetica A (67%) e l'uso di sensori per lo spegnimento automatico delle luci (57%). Infine solo circa il 12 % degli agriturismi ha deciso di utilizzare pannelli fotovoltaici per l'autoproduzione di energia elettrica, probabilmente a causa dell'investimento iniziale che bisogna intraprendere.

[10] www.lifegate.it

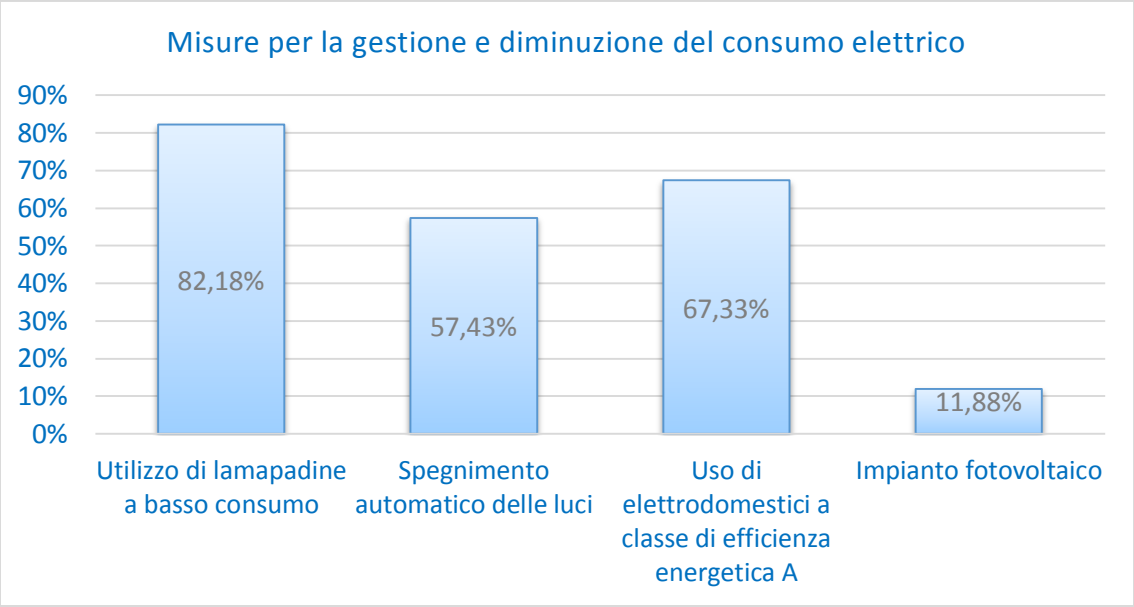


Figura 4.21 - Misure per la gestione e diminuzione del consumo elettrico. Fonte: elaborazione personale.

16. Adottate misure per la gestione e la diminuzione del consumo idrico?

In questo caso degli agriturismi che hanno risposto al questionario la percentuale di quelli che dichiarano di adottare misure per la gestione e la diminuzione del consumo idrico non è molto alta come si può notare dal grafico che segue.

Figura 4.22 - Percentuale di adozione di misure per la gestione e la diminuzione del consumo idrico. Fonte: elaborazione personale.

Le misure maggiormente utilizzate sono le seguenti:

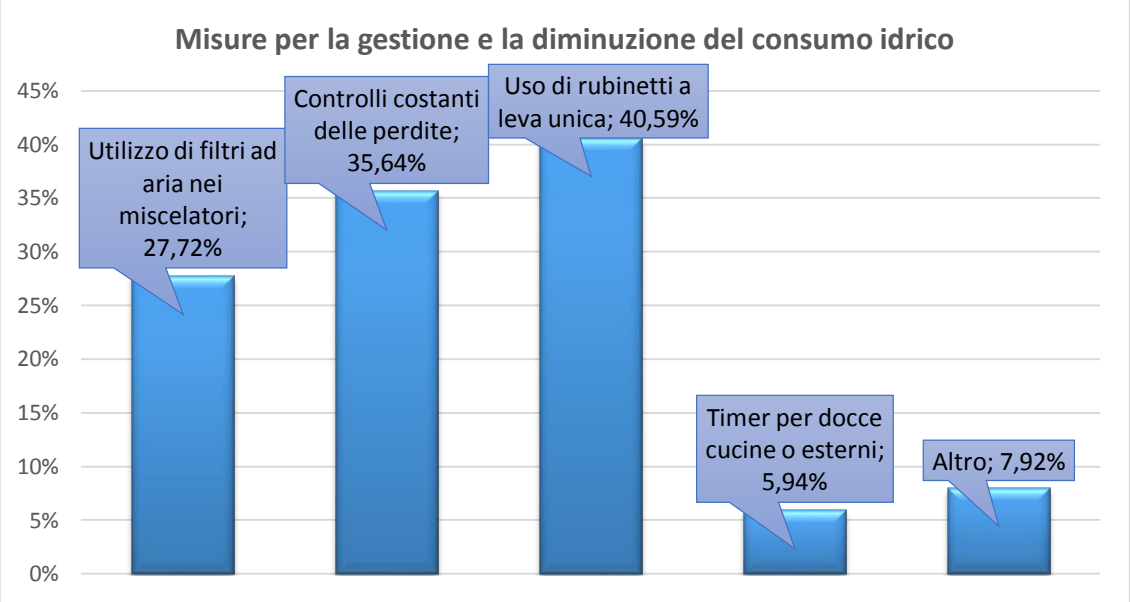


Figura 4.23 - Misure per la gestione e la diminuzione del consumo idrico. Fonte: elaborazione personale.

Altre misure adottate da alcuni agriturismi consistono nell'utilizzo di fotocellule ai rubinetti, pozzi artesiani, cisterne per la raccolta dell'acqua piovana (utilizzabile anche per l'acqua di scarico dei WC), sistemi di irrigazione goccia a goccia per le piante e infine probabilmente la misura più semplice che può e dovrebbe essere adotta da chiunque: l'attenzione personale a non sprecare.

17. Adottate misure per la gestione e la diminuzione del consumo termico?

Molto meglio per quanto riguarda la percentuale di aziende agrituristiche che adottano sistemi di gestione e diminuzione del consumo termico, questa infatti si attesta intorno all'82 %.

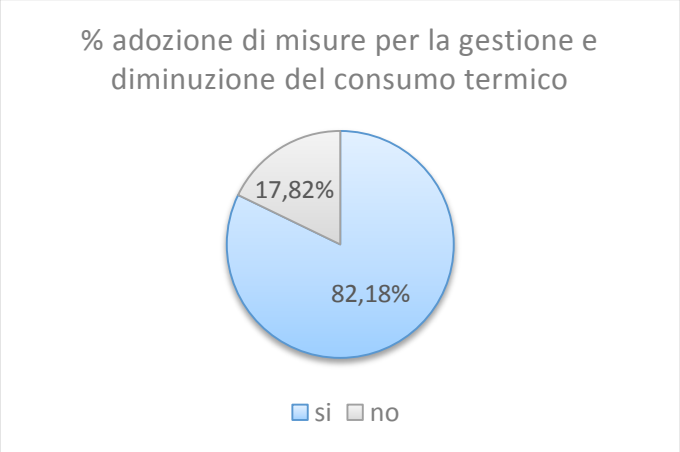


Figura 4.24 - Percentuale di adozione di misure per la gestione e la diminuzione del consumo termico. Fonte: elaborazione personale.

Come si può facilmente notare dal grafico sottostante la percentuale di agriturismi che ha optato per l'isolamento termico dell'edificio è abbastanza elevata: ben il 61% ha infatti adottato questa soluzione che, pur essendo onerosa in termini di costi da supportare inizialmente, consente un significativo risparmio energetico in quanto la dispersione di calore durante l'inverno è ridotta al minimo.

Seguono poi con percentuali ben più basse l'utilizzo di condizionamento a classe di efficienza energetica minima B (29,7%), l'interesse a comunicare agli ospiti la propria politica ambientale in modo da stimolare un comportamento in linea con i principi dell'agriturismo stesso (23,7%) ed infine la predisposizione di suggerimenti per gli ospiti al fine di evitare sprechi (35,6%).

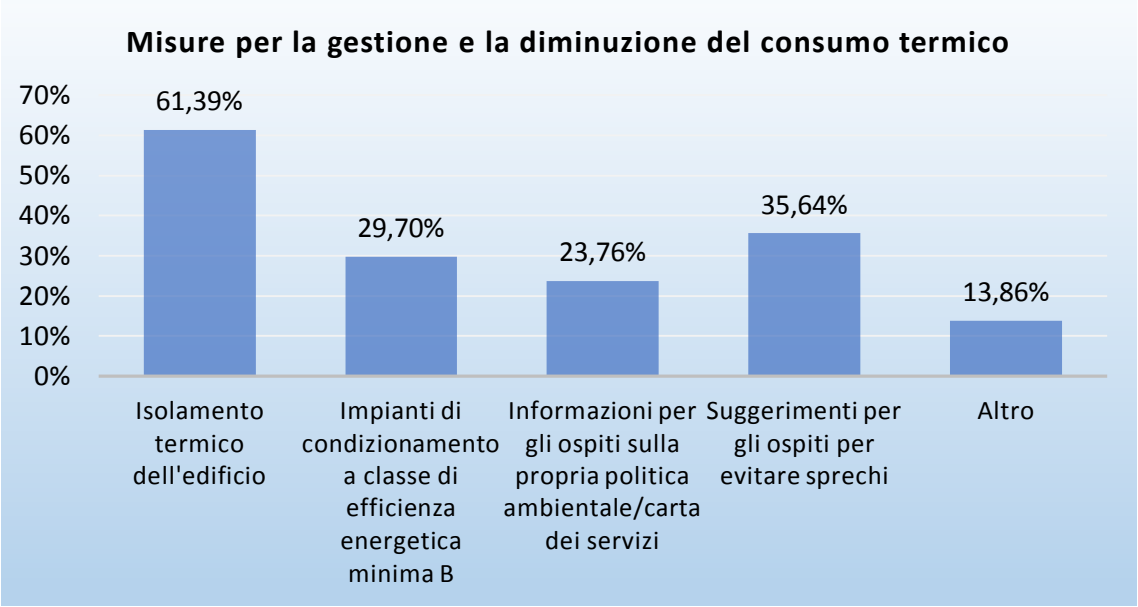


Figura 4.25 - Misure per la gestione e la diminuzione del consumo termico. Fonte: elaborazione personale.

Altre soluzioni adottate da alcuni agriturismi sono:

- Impianti solari termici per l'acqua sanitaria;
- Disattivazione automatica del riscaldamento e del condizionamento se le finestre sono aperte;
- Riscaldamento tramite stufe a legna;
- Impianto a biomasse;
- Utilizzo di caldaie con funzionamento a pallet.

Un'osservazione importante da fare è che gli agriturismi "virtuosi", cioè quelli che adottano tutte le soluzioni elencate nel questionario sono veramente pochi: soltanto 6 su 83.

18. Presso la sua azienda si utilizzano in modo responsabile disinfettanti, detergenti e detersivi?

Dalle risposte date al questionario è emerso che la percentuale di agriturismi che non utilizzano in modo responsabile detersivi, detergenti e disinfettanti è quasi inesistente: appena il 3%.

Ottima invece la percentuale degli agriturismi che costantemente li utilizza in modo responsabile: ben l'85%. Il 12% invece li utilizza soltanto in parte.

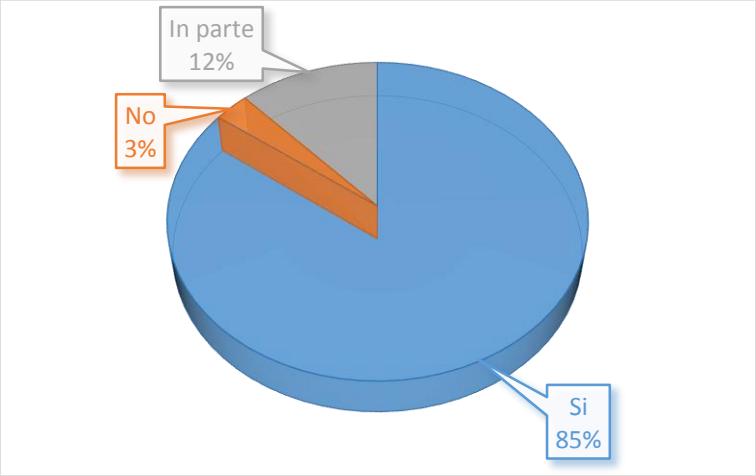


Figura 4.26 - Uso responsabile costante di disinfettanti, detergenti e detersivi. Fonte: elaborazione personale.

Per quanto riguarda le misure maggiormente adottate per prevenire gli sprechi, la situazione è quella rappresentata nel seguente grafico:

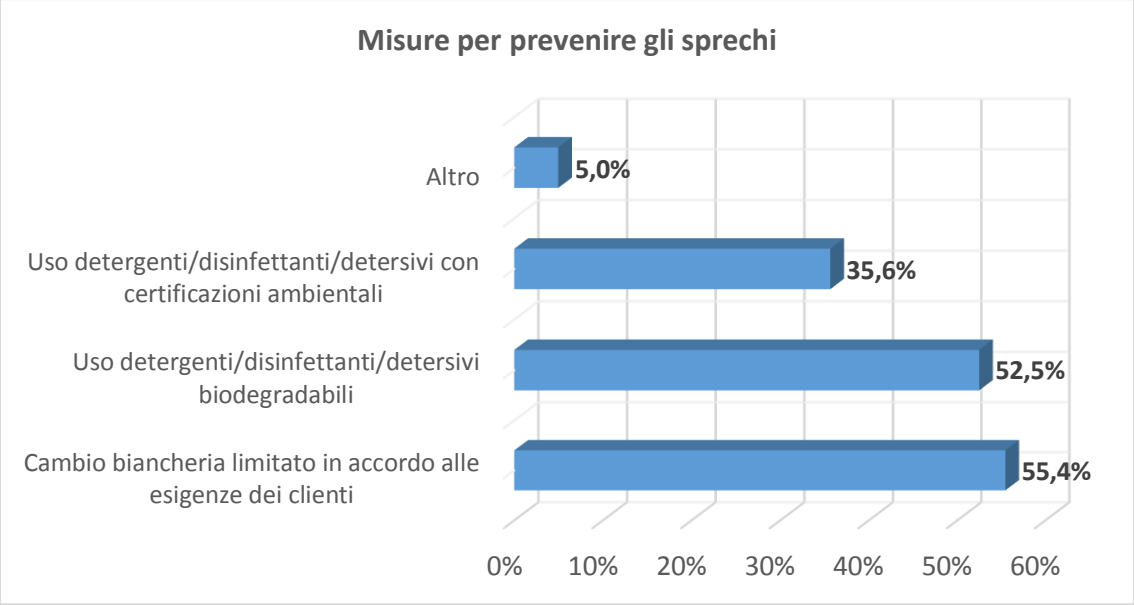


Figura 4.27 - Misure per la prevenzione degli sprechi. Fonte: elaborazione personale.

Come si può facilmente notare, non c'è una misura che prevale sulle altre. Le percentuali di adozione di ogni misura si attesta massimo sul 50% circa degli agriturismi intervistati.

A mio avviso, una situazione che potrebbe essere migliorata facilmente e senza grandi difficoltà attraverso interventi di sensibilizzazione verso le aziende agrituristiche.
Se ogni azienda agrituristica, ma in generale qualsiasi tipo di azienda e privato, dimostrasse maggior impegno nell'adottare questi piccoli accorgimenti, come qualsiasi altro piccolo accorgimento volto a preservare l'ambiente in cui viviamo, il risultato sarebbe sicuramente più che positivo. Alla fine, è sempre la somma delle piccole cose che fa la differenza sul totale.
Fra le altre misure adottate, da circa il 5% degli agriturismi, vi sono: l'uso della pulizia con il vapore, l'uso dei prodotti nella corretta quantità e l'uso di prodotti naturali.

19. Presso la sua struttura adottate misure per la gestione ambientale? Se si, quali?

La percentuale di aziende agrituristiche che adotta una o più misure per la gestione ambientale è abbastanza elevata: un 64% che le adotta pienamente, più un 25% che le adotta solo in parte.
Mentre una percentuale abbastanza bassa di agriturismi (11%) dichiara di non adottare misure per la gestione ambientale.

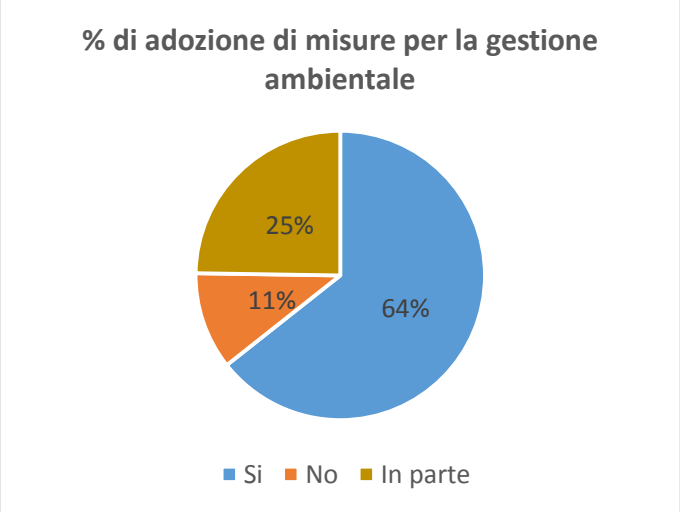


Figura 4.28 - Percentuale di adozione di misure per la gestione ambientale. Fonte: elaborazione personale.

Fra le misure adottate, come si può vedere dal grafico che segue, la più diffusa è sicuramente la manutenzione delle apparecchiature, dovuta molto probabilmente ad un esigenza di funzionamento delle stesse nel tempo. Circa il 50% delle aziende agrituristiche invece svolge controlli periodici dei consumi, una soluzione a mio avviso molto utile in quanto consente di monitorare gli sprechi e attivare eventuali accorgimenti. Poco utilizzata è invece la predisposizione ai clienti di informazioni sulla propria politica ambientale (circa il 28%).

Infine solo due delle aziende intervistate ha adottato delle misure diverse da quelle elencate nel questionario:

- Una ha deciso di creare un bosco per l'abbattimento dell'inquinamento prodotto;
- L'altra azienda agrituristica ha invece deciso di produrre vino biodinamico, cioè vino prodotto attraverso sistemi "sostenibili" che mantengono equilibrato in modo naturale il terreno con tutti i suoi organismi[11].

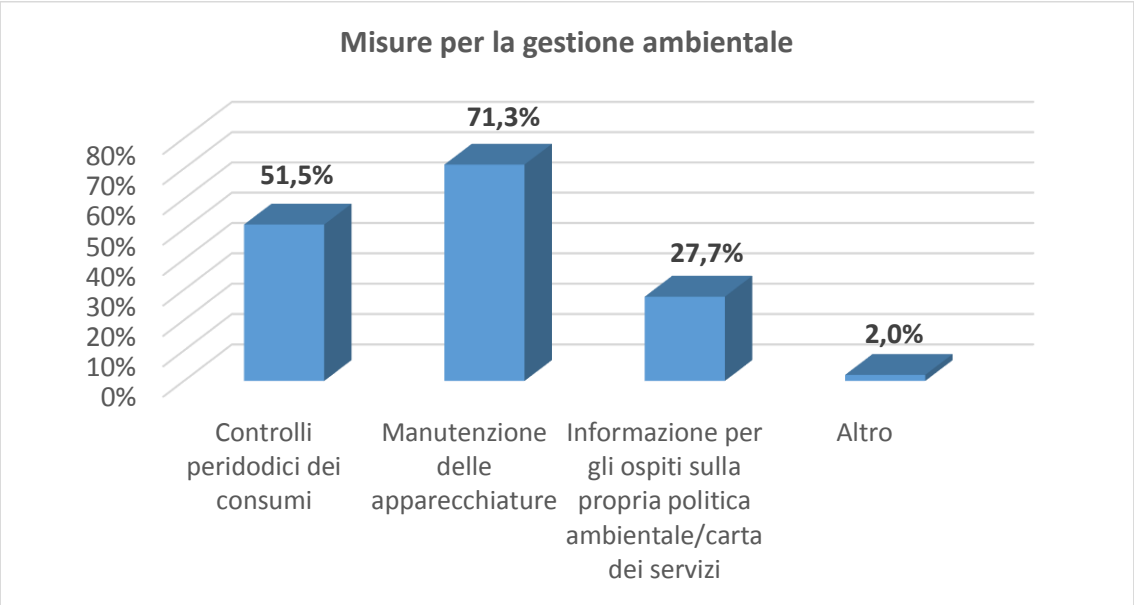


Figura 4.29 - Misure per la gestione ambientale. Fonte: elaborazione personale.

[11] www.wineblog.it

20. Vi è intenzione per il futuro di adottare misure di gestione ambientale?

Come è evidente dal grafico a torta sottostante la maggior parte degli agriturismi intervistati, l'89%, ha preferito non dire niente circa l'intenzione di adottare in futuro una o più misure per la gestione ambientale.
Solo l'11% ha risposto alla domanda in questione, di cui il 9% ha dato risposta positiva e il 2% negativa.

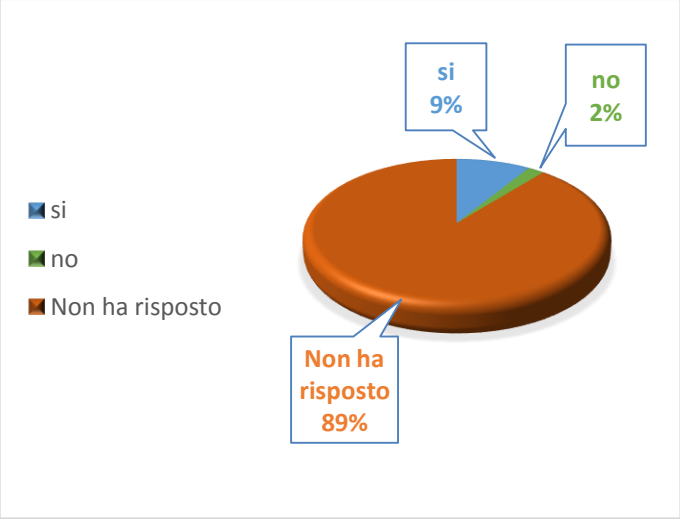


Figura 4.30 - Intenzione di adottare in futuro una qualche misura di gestione ambientale. Fonte: elaborazione personale.

21. È possibile accedere alla sua azienda attraverso i mezzi del servizio pubblico?

Dal questionario è emerso che una parte consistente degli agriturismi non è raggiungibile direttamente attraverso mezzi pubblici come l'autobus. Questo costituisce sicuramente uno svantaggio per gli agriturismi stessi poiché perdono quelle categorie di clienti che preferiscono muoversi con i mezzi del servizio pubblico, oltre agli anziani che hanno difficoltà o non possono utilizzare mezzi propri.

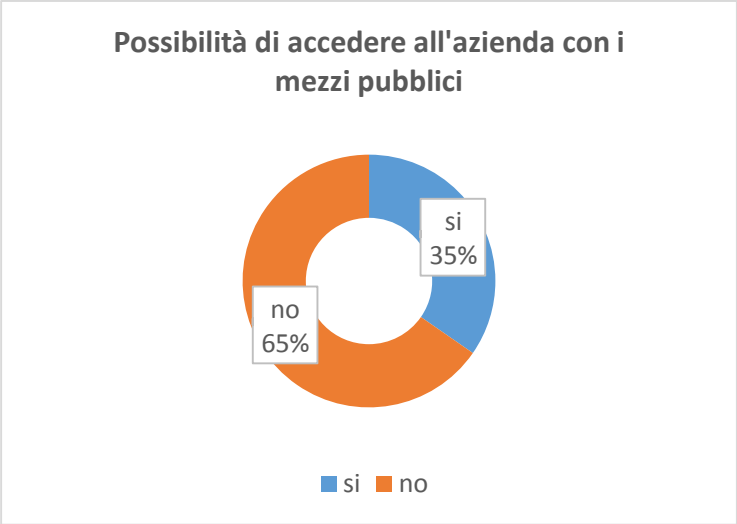


Figura 4.31 - Possibilità di accedere all'azienda con i mezzi pubblici. Fonte: elaborazione personale.

22. Vengono fornite adeguate informazioni agli ospiti su come raggiungere la struttura e altre destinazioni con i mezzi pubblici?

Fra gli agriturismi che nella precedente domanda hanno dichiarato che è possibile accedervi attraverso i mezzi del servizio pubblico, la quasi totalità (83%) fornisce informazioni adeguate su come raggiungere la struttura e le altre destinazioni con i mezzi pubblici.

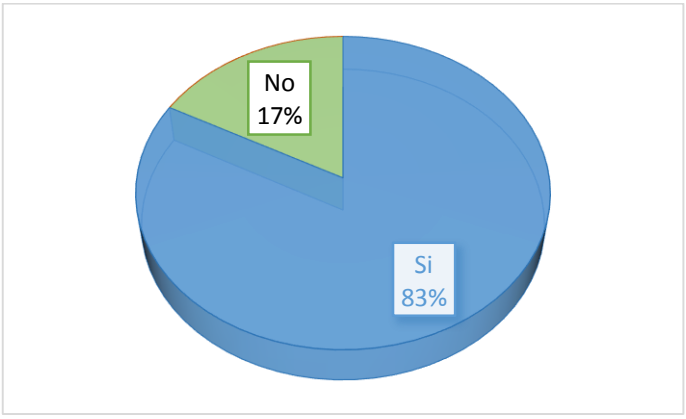


Figura 4.32 - Adeguatezza delle informazioni fornite agli ospiti su come raggiungere la struttura e le altre destinazioni turistiche con i mezzi pubblici. Fonte: elaborazione personale.

23. Presso la sua azienda sono presenti impianti per la produzione di energia rinnovabile?

La percentuale di aziende agrituristiche che ha installato e utilizza impianti per la produzione di energia rinnovabile non è elevatissima, appena il 58%. Una situazione abbastanza positiva ma sicuramente migliorabile, soprattutto se incentivata dallo Stato o dall'Unione Europea.

A mio parere, comunque, la percentuale di aziende agrituristiche, e in generale di qualsiasi tipo di azienda, che deciderà di dotarsi di impianti per la produzione di energia rinnovabile è destinata a crescere. Basti pensare che i governi, a causa dell'aumento dell'inquinamento, stanno investendo sempre di più nel settore delle energie alternative incentivandone il loro uso.

La percentuale di produzione sul totale di energie da fonti rinnovabili è infatti in continuo aumento[12].

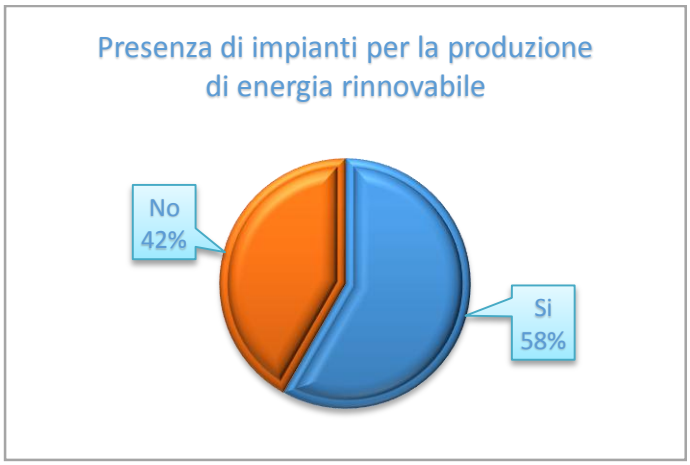


Figura 4.33 - Presenza di impianti per la produzione di energia rinnovabile. Fonte: elaborazione personale.

Come mostrato dal grafico che segue la misura maggiormente, se non unicamente, adottata consiste nella produzione di energia dal solare (pannelli fotovoltaici, geotermia).

[12] www.repubblica.it

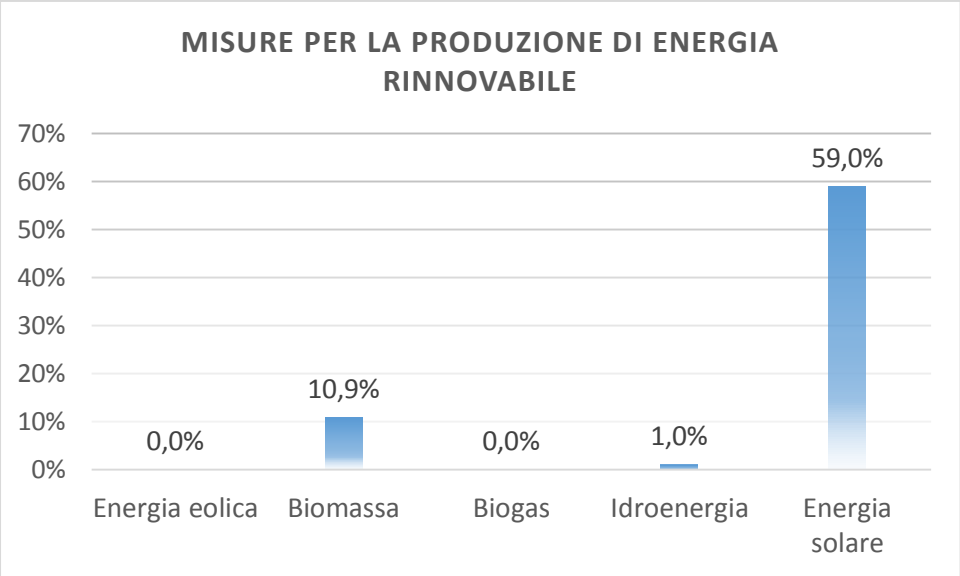


Figura 4.34 - Misure per la produzione di energia rinnovabile. Fonte: elaborazione personale.

Infatti circa il 60% degli agriturismi ha optato per l'installazione di pannelli fotovoltaici o di impianti solari termici. Appena il l'11% invece utilizza impianti a biomasse per la produzione di calore o di energia elettrica.

24. La sua azienda ha beneficiato di una o più misure di sostegno allo sviluppo rurale nell'ultimo triennio?

Come si può facilmente vedere dal grafico che segue la percentuale di agriturismi che ha beneficiato nell'ultimo triennio di misure di sostegno allo sviluppo rurale non sono molte, appena il 36%.

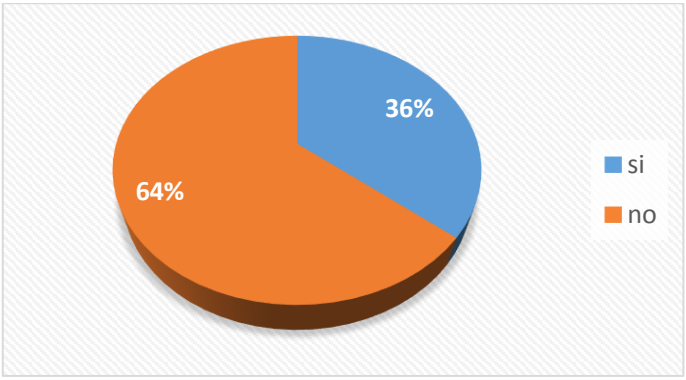


Figura 4.35 - Beneficio di misure di sostegno allo sviluppo rurale, ultimo triennio. Fonte: elaborazione personale.

Fra le misure considerate nel questionario, le due maggiormente beneficiate sono state la misura 311 (14,95) e la misura 121 (15,8%). Le altre misure sono state ottenute in pochissimi casi, soprattutto se consideriamo la misura 216 (appena il 2%) e la misura 213 (solo l'1%).

Un agriturismo ha poi beneficiato di un'altra misura non presente fra le alternative del questionario, cioè la misura 331 ("Formazione e informazione rivolte agli operatori economici nei settori che rientrano nell'Asse 3").

Per un approfondimento sulle misure di sostegno allo sviluppo rurale adottabili dalle aziende agrituristiche è possibile consultare il sito: "www.agricoltura.regione.lombardia.it" e andare nella sezione "Programma di Sviluppo Rurale 2007/2013 – 2014/2020".

Di seguito il grafico a barre contenente le misure di sostegno allo sviluppo rurale e le relative percentuali di agriturismi che ne hanno beneficiato.

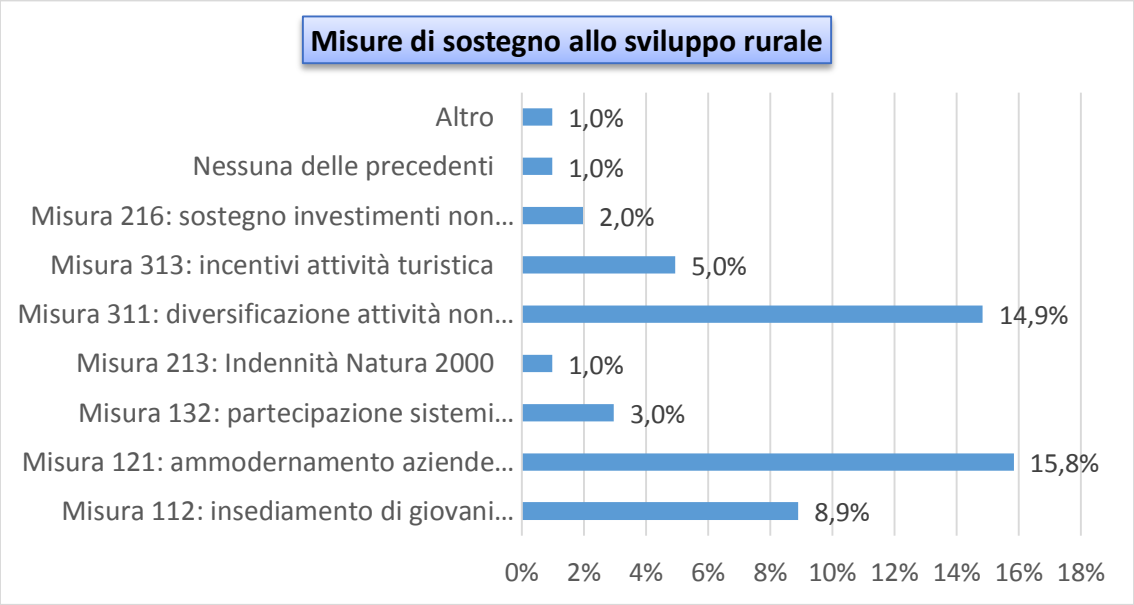


Figura 4.36 - Misure di sostegno allo sviluppo rurale. Fonte: elaborazione personale.

25. La sua struttura è in possesso di certificazioni ambientali?

Dal questionario è emerso che appena il 5% degli agriturismi è in possesso di almeno una certificazione ambientale.

Figura 4.37 - Percentuale di adozione di certificazioni ambientali. Fonte: elaborazione personale.

Di questi pochissimi agriturismi che hanno ottenuto una certificazione ambientale, soltanto l'1% ha ottenuto la certificazione EMAS, mentre non si è registrato nessun agriturismo che abbia ottenuto una certificazione ECOLABEL e ISO14001.
Si sono verificati due casi di azienda agrituristica che ha ottenuto una certificazione di produzione biologica.
Questi dati derivano probabilmente dal fatto che prima di tutto tali certificazioni sono molto difficili da ottenere, in particolar modo l'ECOLABEL. In secondo luogo sono onerose in termini finanziari, poiché sono molti gli aspetti a cui doversi adattare per soddisfare i requisiti richiesti da queste tipologie di certificazione.

26. Vi è l'intenzione per il futuro di adottare certificazioni ambientali?

Come di può facilmente vedere dal grafico che segue la percentuale di agriturismi che ha intenzioni positive nell'ottenere una certificazione ambientale è abbastanza bassa, il 17%. Infatti la maggior parte delle aziende agrituristiche è indecisa se ottenere o meno una

qualche certificazione ambientale (il 59%), mentre il 24% di queste dichiara di non avere nessuna intenzione nel volerne ottenere una.

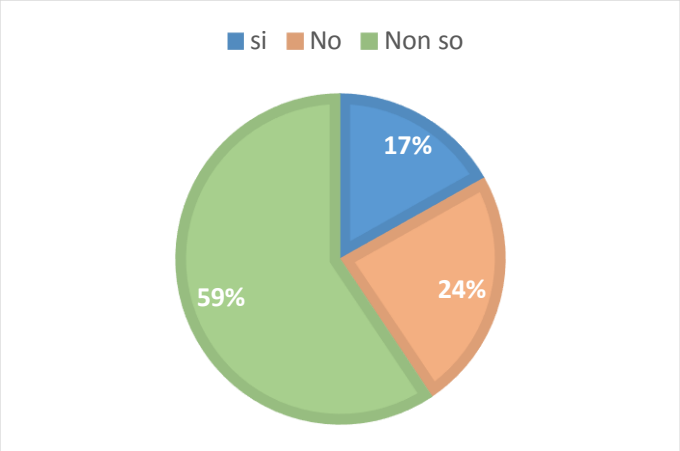


Figura 4.38 - Intenzione futura di adottare una qualche certificazione ambientale. Fonte: elaborazione personale.

Infine, delle poche aziende agrituristiche che hanno risposto "si" o "non so" circa l'intenzione di adottare per il futuro una certificazione ambientale, come si può vedere chiaramente dal grafico che segue la maggior parte degli agriturismi, circa l'86%, è indeciso (o non vuole specificare) su quale misura adottare. Il 9% con molta probabilità adotterà la certificazione EMAS, mentre circa il 5% opterà per l'ISO14001.
Nessun agriturismo ha invece intenzione di adottare l'ECOLABEL, probabilmente per il fatto che si tratta della certificazione più difficile e onerosa da ottenere.

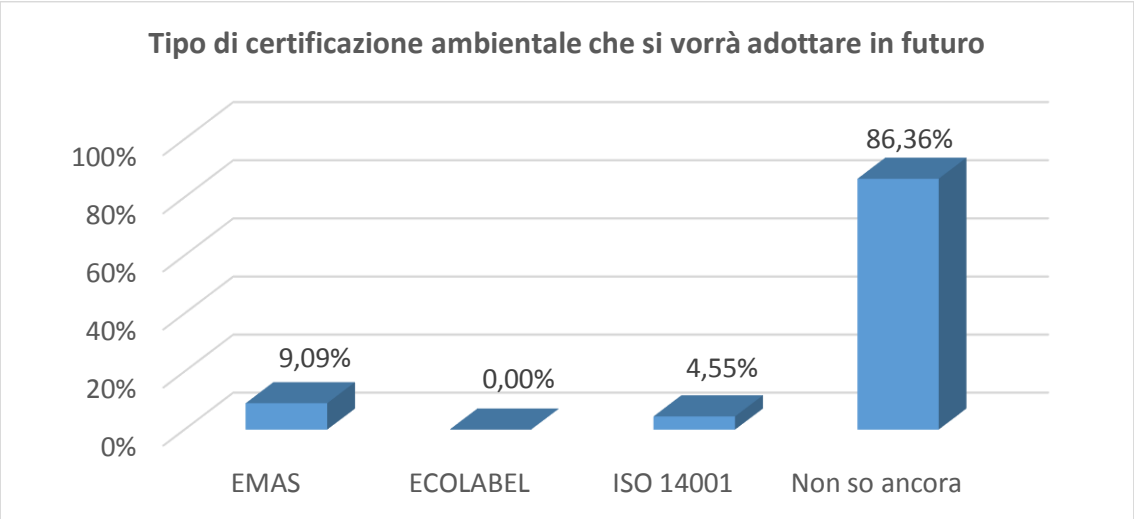
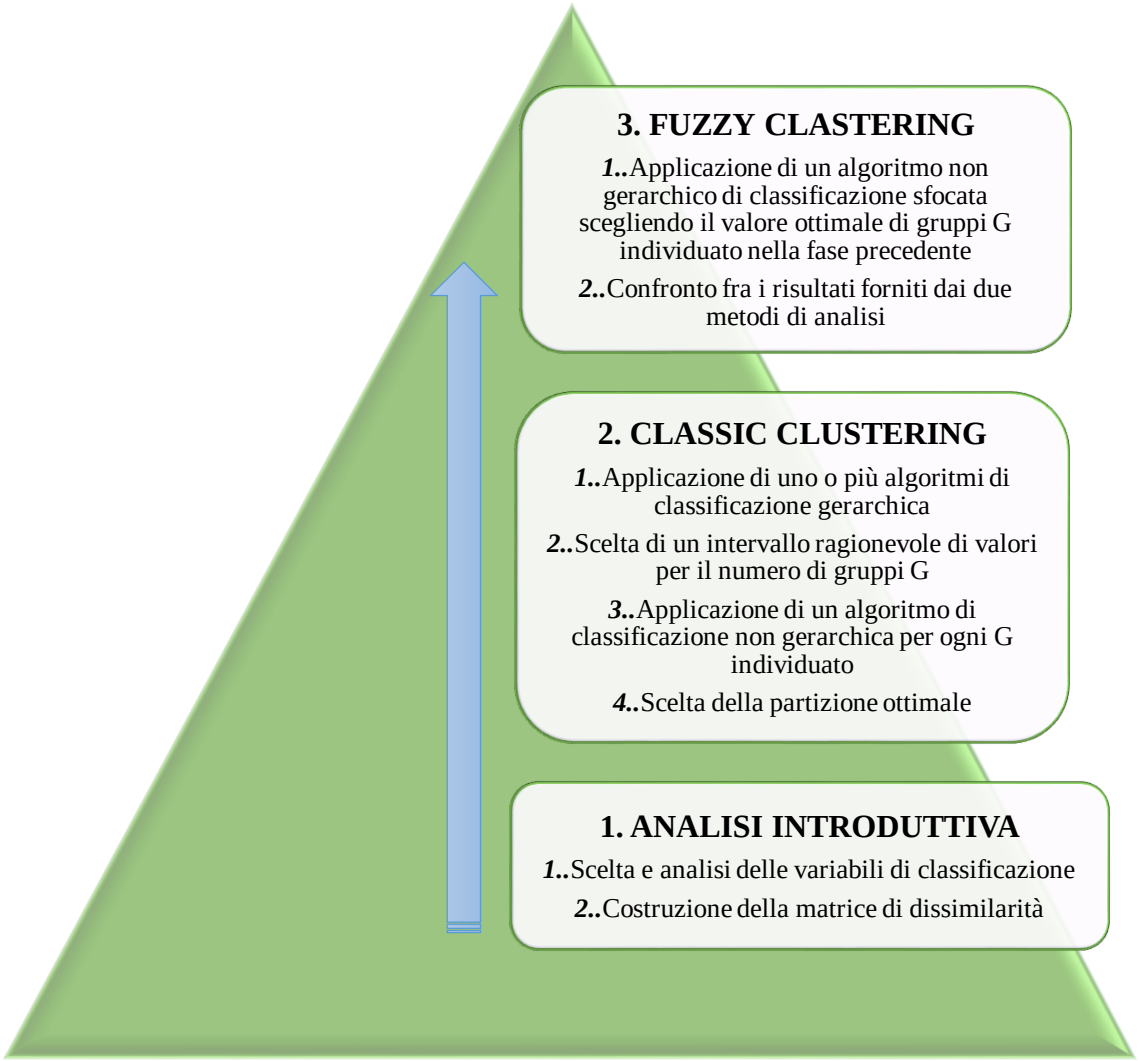


Figura 4.39 - Tipo di certificazione ambientale che si vorrà adottare in futuro. Fonte: elaborazione personale.

4.3 Analisi dei cluster

Questa seconda parte del capitolo si propone di trovare dei gruppi di agriturismi, fra quelli che hanno fornito le risposte al questionario, che presentino un certo grado di omogeneità interna ed eterogeneità esterna sulla base di determinate caratteristiche.
Nel presente lavoro l'obiettivo è quello di identificare dei gruppi di agriturismi che presentino caratteristiche simili, a partire da alcune variabili selezionate inerenti l'adozione di pratiche per la gestione ambientale nell'attività delle aziende agrituristiche.

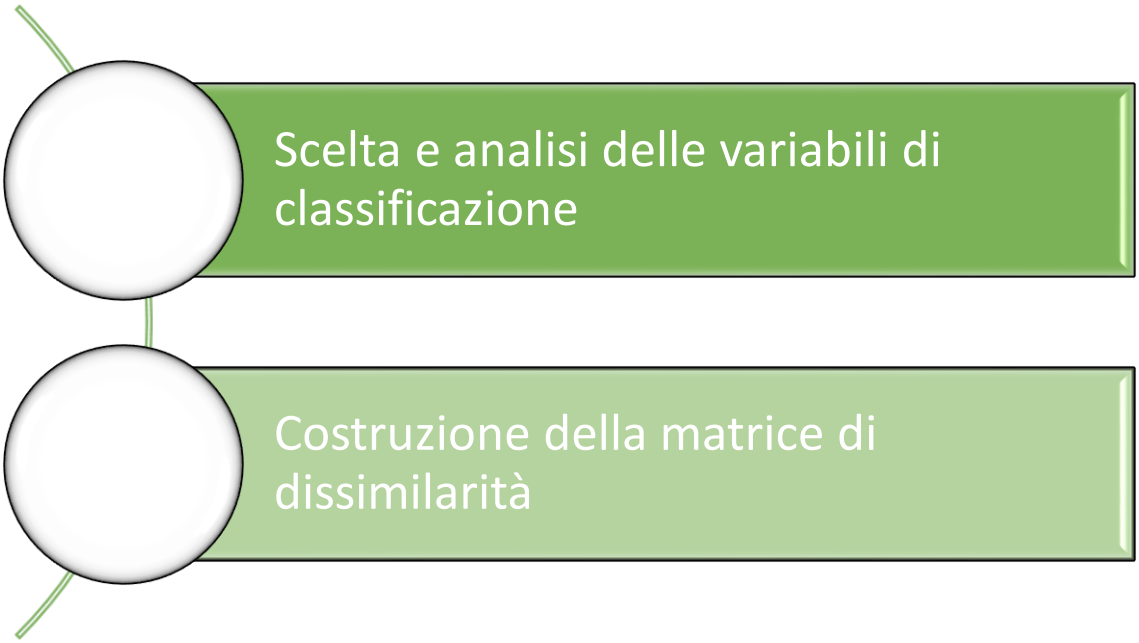
Per perseguire questo obiettivo si è deciso di sviluppare l'analisi come indicato di seguito:



4.3.1 Analisi introduttiva

Questa primo paragrafo si propone di fornire un introduzione alle successive metodologie di analisi dei cluster.

Verranno qui discussi i seguenti due punti:



Scelta delle variabili di classificazione

La prima cosa che bisogna fare nel momento in cui si decide di effettuare un'analisi dei cluster consiste nel scegliere le variabili di classificazione, o caratteri, le quali permettono di raggruppare le unità statistiche sulla base del profilo descritto da tali variabili.

In altre parole, le unità statistiche vengono suddivise in un determinato numero di cluster in base al loro livello di somiglianza valutato a partire dai valori assunti in ciascuna unità dalle variabili prese in considerazione per l'analisi.

La scelta di tali variabili è un'operazione affetta da un forte grado di soggettività poiché rispecchia le convinzioni e le idee di chi sta effettuando la ricerca. Questo può portare infatti a non considerare variabili che sono fortemente selettive e quindi ad ottenere una

partizione sbagliata delle unità; può capitare anche di includere variabili selettive che però non sono utili ai fini dell'analisi che stiamo effettuando.
Di conseguenza si otterrebbero risultati di scarso rilievo.
Questo ci fa capire quanto sia essenziale porre particolare attenzione a questa fase dell'analisi.
Le variabili dovrebbero essere scelte:

- ✓ Innanzitutto in base all'obiettivo che ci siamo prefissati di raggiungere: in questo caso l'obiettivo è quello di individuare gruppi di agriturismi simili sulla base del loro impegno all'adozione di strumenti per prevenire sprechi e salvaguardare l'ambiente;
- ✓ Valutando la loro rilevanza: le variabili devono permettere di distinguere le unità statistiche oggetto di analisi. In altre parole non avrebbe senso selezionare delle variabili che assumono sempre la stessa modalità, non avrebbero valore discriminante.

Le variabili scelte per la presente analisi sono le seguenti:

VARIABILI SCELTE PER LA CLASSIFICAZIONE		
Codifica della variabile	**Descrizione variabile**	**Frequenza data-set (su 101 unità statistiche)**
X49	Misure per la gestione e la diminuzione del consumo idrico	0,61
X72	Possibilità di accedere all'azienda attraverso i mezzi del servizio pubblico	0,35
X74	Presenza di impianti per la produzione di energia rinnovabile	0,58

X81	Beneficio di misure per lo sviluppo rurale	0,36
X91	Ottenimento di certificazioni ambientali	0,05
X96SI	Intenzione "SI" di adottare certificazioni ambientali in futuro	0,24
X96NonSo	Intenzione "NON SO" di adottare certificazioni ambientali in futuro	0,60

Tabella 4.2 - Variabili di classificazione. Fonte: elaborazione personale.

Come si può notare sono tutte variabili inerenti l'aspetto ambientale; fanno infatti parte della parte del questionario incentrata sulla sostenibilità. Osservando le medie, si nota che sono variabili che presentano un certo grado di variabilità.
L'unica eccezione è rappresentata dalla variabile inerente l'ottenimento di certificazioni ambientali: la presenza di tale modalità è infatti molto bassa all'interno del campione di unità statistiche considerato e quindi risulterà ininfluente nel momento in cui si effettuerà la classificazione.

Analisi delle variabili scelte per la classificazione

In questa parte verrà effettuata una breve analisi sulle variabili che verranno utilizzate per classificare gli agriturismi. Si tratta di un'analisi bivariata che ha lo scopo di verificare eventuali relazioni fra di esse.
A tale scopo verranno costruite le tabelle di contingenza (tabelle a doppia entrata) contenenti le frequenze congiunte delle variabili.
Successivamente verrà calcolato per ogni coppia di variabile l'indice di associazione di chi-quadrato e infine l'indice v di Cramér.

Per motivi di spazio verranno di seguito riportate soltanto le tabelle di associazione costruite prima con l'indice di chi-quadrato e poi con l'indice *v* di Cramér, mentre le tabelle di contingenza saranno riportate in appendice.
Le analisi presenti in questa parte del lavoro sono state effettuate utilizzando il software di statistica SPSS.

Partiamo dalla costruzione della matrice di associazione contenente gli indici di chi-quadrato.
L'indice di chi-quadrato è calcolato come segue:

$$\chi^2 = \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{c_{st}^2}{n'_{st}}$$

S e T rappresentano rispettivamente il numero di modalità della prima e della seconda variabile. $n'_{st} = (n_{s.} - n_{.t})/n$ indica le frequenze teoriche di indipendenza nella tabella doppia, in altre parole le frequenze che si sarebbero osservate, dati i valori marginali di riga e di colonna, nel caso di perfetta indipendenza fra le due variabili.
$c_{st} = (n_{st} - n'_{st})$ rappresenta le contingenze, cioè le differenze tra le frequenze osservate e le frequenze attese di indipendenza (vedi tabella distribuzione di chi-quadrato)[13].
Se le due variabili sono perfettamente indipendenti, l'indice assume valore 0 poiché tutti i numeratori dei rapporti sono pari a 0. Tanto più le frequenze osservate di discostano dalle frequenze attese, tanto più l'indice chi-quadrato assume valori elevati.

[13] Approfondimento sul come si calcola l'indice χ^2:
http://segr-did2.fmag.unict.it/Vademecum/Programmi0607/Zira%20Hichy/lezione4.pdf

Di seguito la tabella con gli indici chi-quadrato relativi alle variabili analizzate:

MATRICE DELLE ASSOCIAZIONI CHI-QUADRATO							
	x49	**x72**	**x74**	**x81**	**x91**	**x96si**	**x96nonso**
x49							
x72	0,423						
x74	0,546	1,076					
x81	0,220	1,168	0,167				
x91	0,004	0,499	1,009	0,044			
x96si	0,124	1,296	5,58	0,498	3,813		
x96nonso	1,140	4,320	1,182	1,357	0,915	48,008	

Tabella 4.3 - Matrice associazioni con indice chi-quadrato. Fonte: elaborazione personale.

Come si può notare, l'indice assume valori molto bassi generalmente per tutte le coppie di variabili, per cui possiamo considerarle indipendenti le une dalle altre. Anche in questo caso l'eccezione è rappresentata dalla coppia "x96si, x96nonso" che presenta un indice pari a 48,008, per cui la dipendenza fra queste due variabili risulta evidente. Valori mediamente più alti si verificano per le altre due coppie identificate nell'analisi precedente, e cioè le coppie "x74, x96si" e "x72, x96nonso".

Non esiste un valore massimo definito per l'indice χ^2 e il suo valore, a parità di associazione, dipende dal numero di unità statistiche.

Per questo viene utilizzato un altro indice di associazione che assume valori compresi fra 0 e 1 e che permette quindi di capire meglio quale sia il livello di associazione fra le diverse coppie di variabili.

Stiamo parlando dell'indice *v* di Cramér, o indice medio di contingenza. È dato da:

$$v = \left(\frac{\chi^2/_n}{min[(S-1),(T-1)]} \right)^{1/2}$$

Tale indice assume valore 0 nel caso in cui le variabili sono indipendenti e valori sempre più prossimi ad 1 man mano che l'associazione fra le variabili in esame diventa più forte.

Nel caso assuma valore 1, fra i caratteri vi è massima associazione.
Di seguito viene riportata la matrice di associazione contenente gli indici *v* di Cramér per ciascuna coppia di unità. La matrice è simmetrica con valori unitari sulla diagonale principale.

Matrice delle associazioni V di Cramér							
	x49	x72	x74	x81	x91	x96si	x96nonso
x49	1						
x72	0,065	1					
x74	0,074	0,103	1				
x81	0,047	0,108	0,041	1			
x91	0,006	0,070	0,100	0,021	1		
x96si	0,035	0,113	0,235	0,070	0,194	1	
x96nonso	0,106	0,207	0,108	0,116	0,095	0,689	1

Tabella 4.4 - Matrice delle associazioni con indice V di Cramér. Fonte: elaborazione personale.

Anche in quest'ultima analisi è facile osservare che il valore più alto si registra per la coppia di variabili "x96si, x96nonso". Per tale coppia infatti assume il valore 0,689, per cui si può affermare che l'associazione fra questi due caratteri è abbastanza elevata.
Le altre due coppie che presentano valori più alti rispetto alla media sono ancora la coppia "x72, x96nonso" e la coppia "x74, x96si". Il resto degli indici assume valore prossimo allo 0, segno che le variabili sono indipendenti.

Costruzione della matrice di dissimilarità

Una volta scelte le variabili rispetto alle quali effettuare l'analisi dei cluster, è necessario definire un criterio in base al quale misurare quanto ogni unità è simile rispetto a ciascun'altra unità del campione in riferimento a tutti i caratteri.
Partendo dalla matrice dei dati e una volta calcolate tutte le distanze tra le n unità statistiche, si ottiene una matrice delle distanze D di dimensione *n*x*n*. Si tratta di una matrice simmetrica con valori pari a 0 sulla diagonale principale.

Le misure di distanza variano in relazione alla tipologia dei caratteri presenti nella matrice dei dati. Nel caso di variabili quantitative si parla di misure di distanza, mentre nel caso di variabili qualitative di misure di similarità.
Le variabili che abbiamo scelto per l'analisi dei cluster sono tutte variabili qualitative binarie, per cui si tratterà di trovare una misura adatta a esprimere la similarità fra ciascuna coppia di unità statistica di questa tipologia.
Quando si ha a che fare con variabili binarie, due generiche righe della matrice dei dati possono essere sintetizzate in una tabella di contingenza come quella che segue:

	Unità p		
Unità s		1	0
	1	a	b
	0	c	d

Tabella 4.5 - Tabella di contingenza. Fonte: elaborazione personale.

"a" indica il numero di caratteri presenti in entrambe le unità, "b" il numero di caratteri presenti nell'unità *s* e assenti nell'unità *p*, "c" il numero di caratteri presenti nell'unità *p* ma assenti nell'unità *s* e "d" il numero di caratteri assenti in entrambe le unità statistiche.
A partire da quanto appena detto possono essere calcolate diverse misure di similarità in base all'approccio utilizzato.
È possibile infatti ricorrere ad alcuni approcci principali che qui di seguito vengono elencati:

- Simple Matching (Sokal & Michener, 1958);
- Indice di distanza di Jaccard (Sokal & Sneath,1963);
- Indice di distanza di Czekanowsky (Sokal & Sneath,1963).

Per la presente analisi invece si è deciso di utilizzare un indicatore che consente di calcolare una misura sintetica di similarità tra ciascuna coppia di unità statistica a prescindere dalle tipologie di variabili presenti nei dati.

Si tratta dell'*indice di distanza di Gower* definito come segue:

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

dove:

- "p" indica il numero di caratteri,
- $d_{ij,\,k}$ è la misura della distanza fra le righe *i* e *j* della matrice dei dati in relazione al *k*-esimo carattere,
- $\delta_{ij,\,k}$ è una variabile dicotomica che acquisisce valore 1 se le due unità possono essere confrontate rispetto al carattere *k* e zero nel caso contrario.

La misura della distanza $d_{ij,\,k}$ varia in relazione al tipo di carattere. Nel caso in esame, essendo di fronte a tutti caratteri qualitativi dicotomici, $d_{ij,\,k}$ assume valori 0 e 1 a seconda che le due unità abbiano modalità uguali o diverse. $\delta_{ij,\,k}$ assume valore 0 soltanto nel caso in cui entrambe le due unità hanno valore 0 per il carattere considerato, in altre parole quando non presentano quel dato carattere.
Riassumendo, i valori assunti nei diversi casi da $d_{ij,\,k}$ e da $\delta_{ij,\,k}$ sono indicati nella seguente tabella:

$d_{ij,\,k}$	Unità s	
Unità p	Si	No
Si	0	1
No	1	0
$\delta_{ij,\,k}$	Unità s	
Unità p	Si	No
Si	1	1
No	1	0

Ponendo $\delta_{ij,k} = 0$ soltanto nel caso di assenza del fenomeno in entrambi i caratteri significa adottare per questa tipologia di caratteri la misura di distanza proposta da Jaccard[14], il quale non considera indicatore di similarità tra due unità il fatto di non possedere entrambe un certo carattere.

Per il calcolo della matrice di dissimilarità è stato utilizzato il software di statistica "R". La funzione utilizzata si chiama "Daisy", e permette di calcolare la distanza fra coppie di unità statistiche specificando diverse metriche. Nel caso in esame, la metrica utilizzata è quella di Gower e specificandola nella funzione "Daisy" attraverso l'argomento "metric", viene applicata la formula generica dell'indice di Gower mostrata precedentemente.

4.3.2 Classificazione classica

Nel momento in cui abbiamo calcolato la matrice di dissimilarità, il passo successivo è quello di scegliere un algoritmo di raggruppamento delle unità statistiche.
È da sottolineare che in generale i gruppi che vengono identificati dipendono:

- dal tipo di distanza usata per creare la matrice di dissimilarità;
- dal tipo di metodo che viene scelto, se partitivo o aggregativo;
- dal tipo di algoritmo di aggregazione.

Non esiste il metodo migliore in assoluto, ogni metodo ha i suoi vantaggi e i suoi svantaggi.
L'obiettivo è ottenere una partizione che presenti la più piccola varianza interna possibile e allo stesso tempo la più grande varianza possibile fra i gruppi. In altre parole gruppi omogenei al loro interno ed eterogenei fra di loro.

[14] Vedi: http://www.stat.unipg.it/luca/AdM/AdM_cap4.pdf

Verranno applicati quindi due diversi metodi gerarchici aggregativi con lo scopo di identificare il numero ottimale di gruppi in cui dividere le unità statistiche.
Una volta identificati i valori ragionevoli per G (numero dei gruppi), per ognuno di essi verrà applicato un metodo non gerarchico di classificazione in modo da riallocare le unità in maniera più efficiente, determinare il numero ottimale definitivo di cluster attraverso diversi strumenti di analisi e consentire infine un confronto con la successiva classificazione sfocata.

4.3.2.1 Clustering gerarchico

In questa prima parte verranno applicati i seguenti due algoritmi gerarchici di classificazione:

1. Ward
2. Legame completo

Il risultato dei metodi gerarchici consiste in una serie di partizioni che possono essere rappresentate graficamente attraverso lo strumento del dendrogramma. Tale strumento permette anche di effettuare una prima valutazione circa il numero ragionevole di gruppi in cui il campione di un'unità statistiche deve essere diviso.
Il dendrogramma riporta sull'asse delle ascisse le unità statistiche e sull'asse delle ordinate la distanza a cui avviene la fusione.
Come vedremo successivamente il numero di gruppi viene identificato tagliando il dendrogramma con una linea orizzontale in corrispondenza della massima distanza fra i livelli di fusione. Siccome l'obiettivo è ottenere il minor numero di gruppi con massima omogeneità, bisognerà tagliare alla base dell'insieme dei rami più lunghi (le linee verticali più lunghe). Non sempre però risulta intuitivo osservando semplicemente il grafico: in questi casi perciò è possibile ricorrere ad alcune statistiche relative alle classi individuate in modo da rendere la scelta più sicura.

Metodo di Ward

Applicando il metodo di Ward, il relativo dendrogramma risulta essere il seguente:

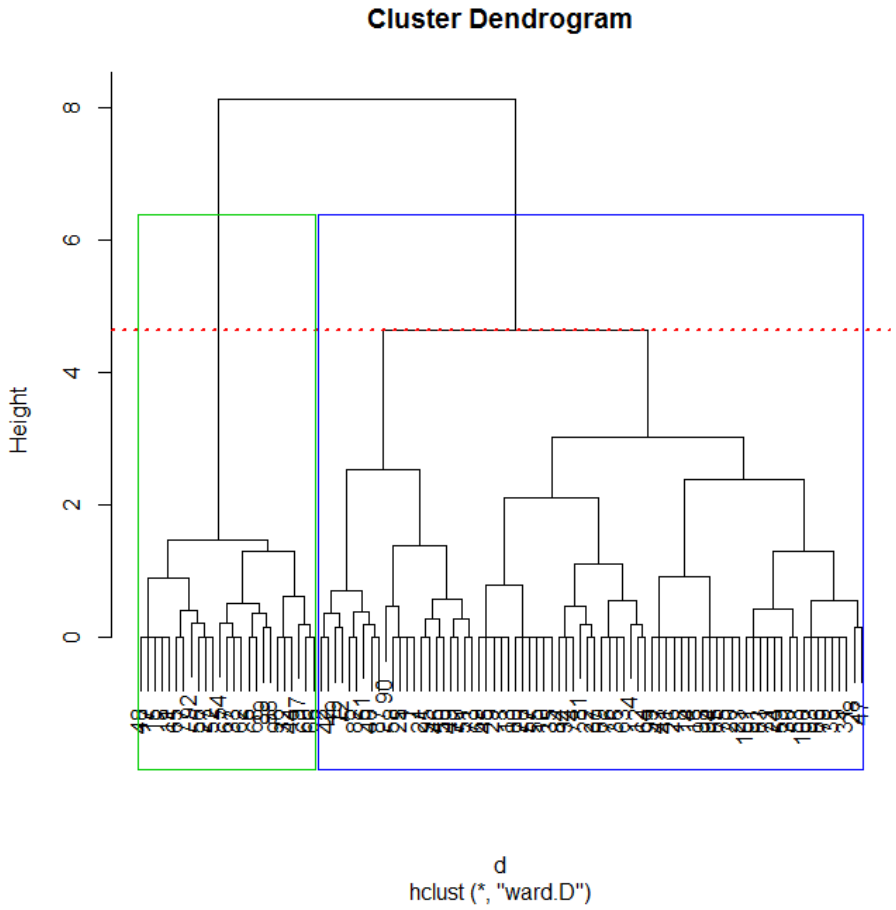


Figura 4.40 - Dendrogramma metodo di Ward. Fonte: elaborazione personale.

Come è evidente dal grafico, il segmento a cui corrisponde la massima differenza fra i livelli di fusione è quello posto al di sopra della linea rossa, per cui possiamo affermare che il grafico dovrebbe essere tagliato proprio in corrispondenza di quel segmento. Viene identificata dunque una partizione composta da due cluster:

- Il primo comprendente le unità all'interno del rettangolo verde (25 unità);
- Il secondo comprendente le unità all'interno del rettangolo blu (76 unità).

Analizziamo ora il processo di aggregazione, qui rappresentato a partire dallo step 81.

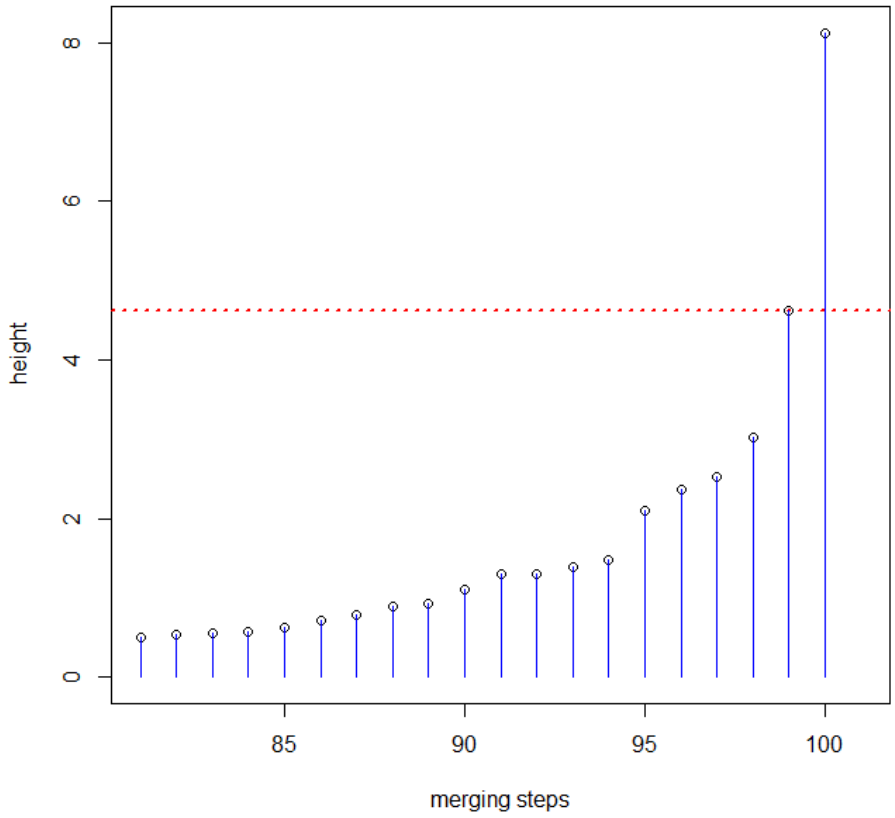


Figura 4.41 - Processo di aggregazione metodo di Ward. Fonte: elaborazione personale.

Come si può facilmente notare il salto maggiore si verifica fra il penultimo e l'ultimo step agglomerativo (la linea rossa identifica il penultimo passaggio), cioè fra la divisione in 2 e 1 cluster. Questo significa che il numero di gruppi che hanno maggiore coesione interna e divisione esterna corrisponde a 2.

Di seguito il processo di aggregazione (sempre a partire dallo step 81) in numeri, dove le prime due colonne indicano la fusione fra singole unità statistiche (identificate dal simbolo "-") e cluster (nessun simbolo), mentre la terza colonna indica il livello a cui avviene la fusione.

[81,]	71	75	0.5000000
[82,]	56	66	0.5442177
[83,]	20	64	0.5555556
[84,]	72	73	0.5714286
[85,]	49	69	0.6190476
[86,]	74	76	0.7063492
[87,]	8	25	0.7792208
[88,]	14	77	0.8874459
[89,]	32	37	0.9230769
[90,]	80	82	1.1041340
[91,]	78	83	1.2931839
[92,]	81	85	1.3027211
[93,]	79	84	1.3890110
[94,]	88	92	1.4755473
[95,]	87	90	2.1081419
[96,]	89	91	2.3710407
[97,]	86	93	2.5319125
[98,]	95	96	3.0251323
[99,]	97	98	4.6254145
[100,]	94	99	8.1272582

Tabella 4.6 - Parte del processo di aggregazione in numeri. Fonte: elaborazione personale.

In giallo è evidenziato il passaggio in cui si registra la maggiore differenza fra i livelli di fusione (3,50), quindi fra lo step 99 e lo step 100. Qui infatti la differenza è superiore a qualsiasi altro passaggio compresi quelli non riportati nella tabella (i livelli di fusione sono ordinati in modo crescente). Dunque come suggerito dal metodo di Ward, il numero più adatto di gruppi in cui suddividere il campione corrisponde a 2.

Nell'analisi non gerarchica di classificazione verrà quindi posto G = 2 come scelta ottimale dei gruppi da costruire, ma come ulteriore prova verrà poi posto G = 3 (a cui corrisponde la seconda più grande differenza fra i livelli di fusione) in modo da evidenziare le differenze e confermare che il primo partizionamento è il più efficiente.

Metodo del legame completo

A solo scopo di conferma e per mostrare che i risultati ottenuti con l'algoritmo di Ward possono essere ottenuti anche con un altro metodo di classificazione gerarchica, viene ora applicato il metodo del legame completo.

Il dendrogramma che si ottiene è il seguente:

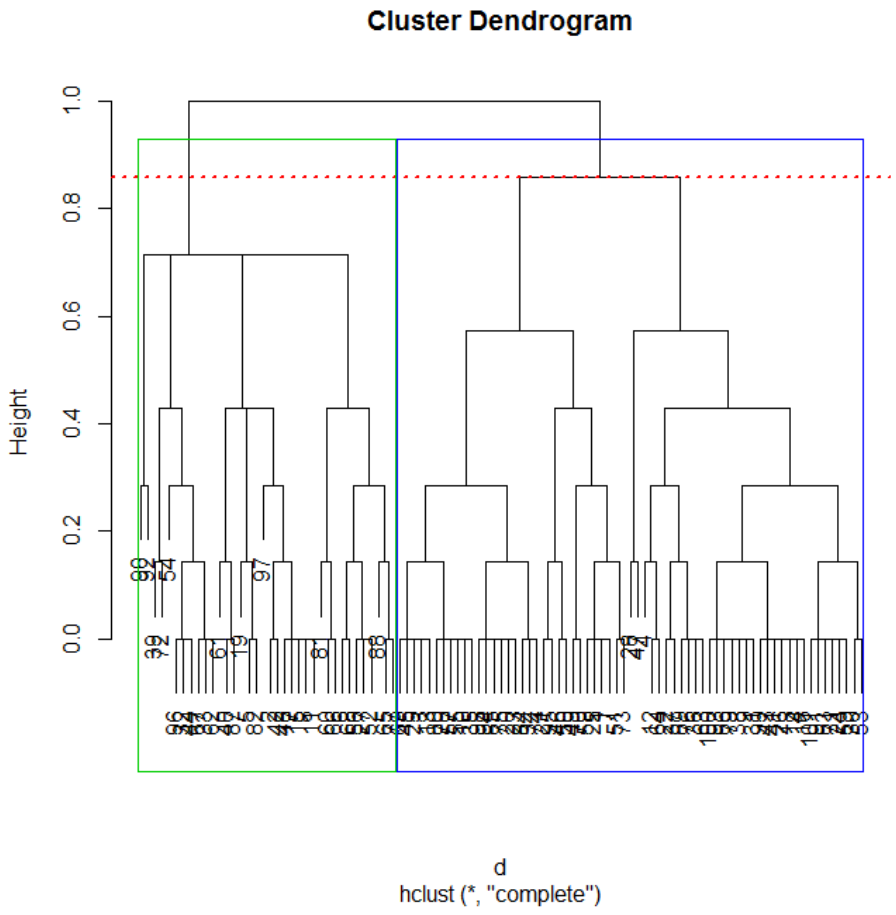


Figura 4.42 - Dendrogramma metodo "complete". Fonte: elaborazione personale.

Come si può facilmente notare, anche questo metodo porta ad una classificazione composta da 2 cluster ben distinti, il primo composto da 35 unità statistiche e il secondo da 66 unità statistiche.

4.3.2.2 Clustering non gerarchico

Per l'analisi non gerarchica di classificazione si è deciso di utilizzare il metodo del partizionamento attorno ai medoidi, noto come "*k-medoids*".
Tale metodo, a differenza del più noto e utilizzato algoritmo delle "*k medie*", si basa sulla ricerca di oggetti rappresentativi all'interno del campione di unità statistiche. Tali punti, noti col nome di "medoidi", sono identificati in modo tale che la dissimilarità totale di tutti i punti dai medoidi più vicini sia la più piccola possibile. In pratica si tratta di ricercare l'insieme dei medoidi $\{m_1, \dots, m_k\} \subset \{1, \dots, n\}$ che minimizza la seguente funzione obiettivo:

$$F(\{m_1, \dots, m_k\}) = \sum_{i=1}^{n} \min_{f=1\dots k} d(i, m_f)$$

La motivazione per cui è stato scelto il metodo dei k-medoids consiste nel fatto che è una versione più robusta del classico algoritmo delle k-medie poiché minimizza una somma di distanze assolute rendendolo così meno sensibile alla presenza di possibili outlier. Tale metodo inoltre funziona molto bene in presenza di piccoli data set, come quello utilizzato per la presente analisi.

La procedura porta alla formazione di due cluster delle seguenti dimensioni:

Partizionamento	
1° cluster	2° cluster
32 unità	69 unità

Tabella 4.7 – Partizionamento. Fonte: elaborazione personale.

Valutiamo ora la qualità dei gruppi che si sono formati. Per perseguire questo obiettivo costruiamo il grafico della Silhouette. Esso indica per ciascun cluster le unità che sono state correttamente assegnate al gruppo, quelle che dovrebbero essere assegnate ad un altro gruppo e quelle che si trovano in una posizione intermedia.

Brevemente e a solo scopo indicativo, l'ampiezza della Silhouette relativa all'*i*-esima unità viene calcolata in questo modo:

$$-1 \leq s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \leq 1$$

a(i) è la dissimilarità media dell'*i*-esima unità rispetto a tutte le unità che appartengono al cluster mentre *b(i)* è la dissimilarità più piccola tra l'unità *i*-esima e le altre unità che non appartengono al cluster.

Si possono verificare 3 casi:

- ***s(i)* = 1 →** significa che *a(i)* = 0, per cui l'unità statistica è stata assegnata correttamente al cluster. Questo significa che mediamente dista 0 rispetto a ciascuna delle altre unità appartenenti al cluster. In altre parole il cluster è omogeneo al suo interno.
- **s(i) = 0 →** significa che a(i) = b(i), per cui l'assegnazione dell'unità al cluster o a quello più vicino è ambigua. Si verifica infatti che la distanza media calcolata all'interno del gruppo è uguale a quella minima valutata rispetto alle unità non appartenenti al cluster.
- **s(i) = -1 →** la dissimilarità rispetto alle unità non appartenenti al cluster è pari a 0. L'unità deve essere riassegnata a quello più vicino.

Una volta calcolate le ampiezze di silhouette per ogni unità, si può procedere calcolando la silhouette media relativa al singolo cluster, quindi sommando le silhouette di tutte le unità del cluster e dividendo il risultato per la numerosità del cluster.
Infine è possibile calcolare l'ampiezza media di silhouette relativa all'intero partizionamento semplicemente sommando le ampiezze medie di silhouette di ogni cluster e dividendo poi il risultato per il numero totale di cluster.

Si può così definire il coefficiente di silhouette:

$$SC = \max \bar{s}(g)$$

Tale coefficiente permette di capire qual è il partizionamento migliore: viene scelto quello per cui la silhouette media è massima.
Inoltre fornisce un'indicazione circa la struttura dei cluster. Per capire se la struttura dei cluster che abbiamo calcolato è forte oppure è debole, o nel caso peggiore manca del tutto, si può fare riferimento alla seguente tabella:

Validità di una struttura di clustering	
SC	**Struttura di Clustering**
(0,70 – 1]	Forte
(0,50 – 0,70]	Ragionevole
(0,25 – 0,50]	Debole
Inferiore a 0,25	Assenza di struttura

Tabella 4.8 - Validità di una struttura di clustering. Fonte: Ricolfi L. (1992).

Il grafico della silhouette relativa al caso in esame risulta essere:

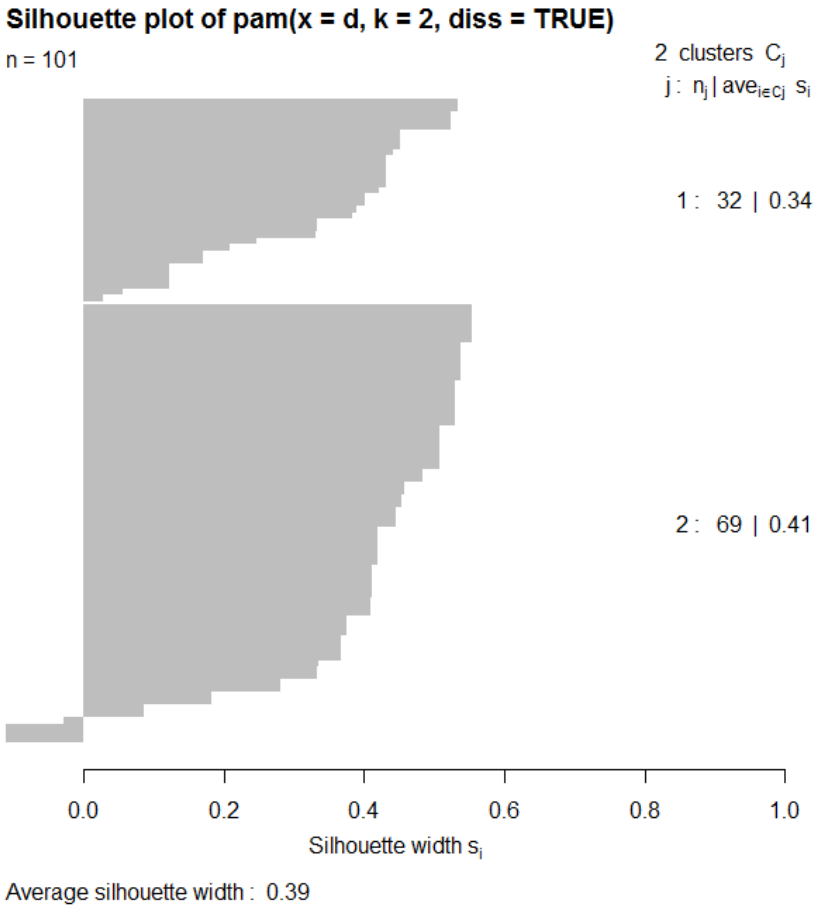


Figura 4.43 - Grafico silhouette procedura PAM con 2 gruppi. Fonte: elaborazione personale.

Il coefficiente di silhouette è pari a 0,39, per cui la struttura di clustering risulta debole.
La silhouette media relativa ai singoli cluster risulta essere abbastanza buona.
Come si può notare ci sono alcune unità (esattamente 4) che presentato silhouette negativa. Questo significa che la loro appartenenza al secondo cluster non è corretta, di conseguenza dovrebbero essere assegnate al primo cluster.

Le unità in questione sono le seguenti:

Unità statistiche allocate in modo errato secondo la Silhouette		
Unità statistica	**Cluster di assegnazione**	**Silhouette**
32	2	-0.110
44	2	-0.110
96	2	-0.110
97	2	-0.027

Tabella 4.9 - Unità statistiche allocate in modo errato. Fonte: elaborazione personale.

Tali unità presentano valori negativi molto vicini allo 0, per cui si tratta di unità la cui assegnazione al 2° cluster o al 1° risulta abbastanza ambigua. Di conseguenza non è totalmente sbagliata l'assegnazione da parte del metodo PAM al 2° cluster.

Tali unità verranno utilizzate come confronto nell'analisi sfocata dei cluster del prossimo paragrafo.

Riapplicando la procedura con G = 3, il grafico della silhouette risulta essere il seguente:

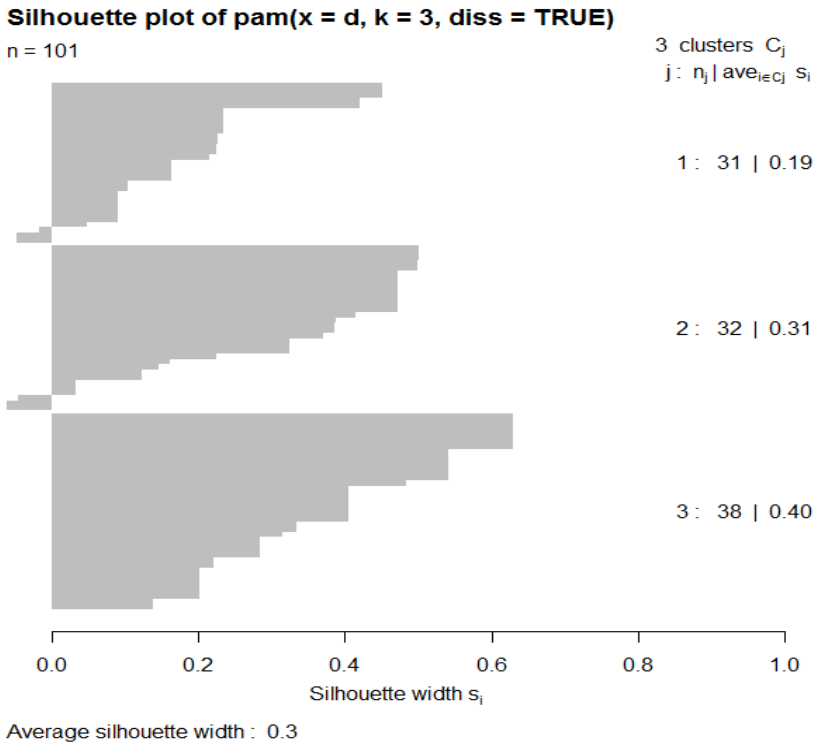


Figura 4.44 - Grafico silhouette procedura PAM con 3 gruppi. Fonte: elaborazione personale.

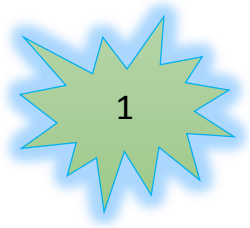
Come si voleva dimostrare, il partizionamento migliore risulta essere quello con 2 gruppi. Con 3 gruppi infatti il coefficiente di silhouette è pari a 0,30, quindi inferiore rispetto alla classificazione in 2 cluster.
Inoltre le unità con silhouette negativa diventano 6, questo significa che rispetto al caso precedente ci sono 2 unità in più che non stanno bene all'interno del loro cluster di assegnazione.
In definitiva, quindi, il partizionamento migliore risulta essere quello ottenuto ponendo il numero di gruppi uguale a 2.

4.3.3 Fuzzy clustering

Questo paragrafo si propone di effettuare un'analisi sfocata dei cluster sul data set oggetto di studio. In particolar modo l'obiettivo è quello di operare un confronto fra la partizione ottenuta precedentemente con l'applicazione del metodo non gerarchico classico dei *k-medoidi* e la partizione che si otterrà applicando il corrispondente metodo non gerarchico di classificazione sfocata.
Il coefficiente di sfocatura verrà posto uguale a 2 poiché è con tale coefficiente che si ottengono i risultati migliori. Verrà comunque effettuata una prova ponendo m=3 per vedere se si verificano differenze sostanziali nella classificazione ottenuta.

Specificatamente, si è deciso di sviluppare l'analisi in questo modo:

1) Applicazione dell'algoritmo ponendo il numero di gruppi G pari a 2 e l'indice di sfocatura *m* uguale a 2, e successivo confronto con i risultati ottenuti con il metodo dei *k-medoidi*;
2) Applicazione dell'algoritmo ponendo il numero di gruppi G pari a 3 e l'indice di sfocatura *m* uguale a 2, allo scopo di verificare se la funzione *fanny* può portare ad una partizione qualitativamente superiore rispetto a quanto ottenuto ponendo il numero di gruppi pari a 3 con il metodo classico di classificazione;
3) Applicazione dell'algoritmo ponendo il numero di gruppi G pari a 2 e il coefficiente di sfocatura *m* uguale a 3.



G = 2, m = 2

Lavorando su un piccolo data set è possibile riportare in una tabella i gradi di appartenenza di ogni unità a ciascun cluster. Inoltre verrà indicato il cluster crisp di assegnazione e la silhouette per ognuna delle unità statistiche. Tale tabella è consultabile in appendice.

L'applicazione dell'algoritmo FANNY ponendo il numero di gruppi e il coefficiente di sfocatura uguali a 2 porta alla seguente partizione:

Partizionamento	
1° cluster	2° cluster
40 unità	61 unità

Tabella 4.10 – Partizionamento. Fonte: elaborazione personale.

L'indice di Dunn standardizzato[15] (valori compresi fra 0 e 1: massimo grado di sfocatura nel primo caso, assenza di sfocatura nel secondo) assume valore 0,0449. Essendo un valore molto vicino allo 0 significa che nei gruppi si ha un alto grado di sfocatura.

In altre parole significa che mediamente ciascuna unità statistica assume nei due gruppi un grado di appartenenza simile, quindi vicino allo 0,5. In altre parole, non si verificano molti casi di unità che appartengono in modo preponderante ad un gruppo piuttosto che all'altro.

I casi di unità che hanno un "rilevante" grado di appartenenza ad uno dei due cluster sono infatti soltanto 15. Nella presente analisi si è stabilito che si ha un'appartenenza rilevante ad uno dei due cluster quando si verifica, per l'unità statistica in esame, una differenza assoluta fra i gradi di appartenenza ai due cluster di almeno 0,3.

[15] Indice di Dunn: $F_k = \sum_{i=1}^{n}\sum_{v=1}^{k}\frac{u_{i,v}^2}{n}$. È compreso fra [1/k, 1]. *n* indica il numero di unità statistiche, mentre *k* il numero di cluster.

Indice di Dunn standardizzato: $F_k' = \frac{F_k - 1/k}{1 - 1/k} = \frac{kF_k - 1}{k-1}$. È compreso fra [0,1].

Le unità in questione presentano i seguenti gradi di appartenenza:

Unità statistiche con rilevante grado di appartenenza ad un cluster			
Unità statistica	Appartenenza al 1° cluster	Appartenenza al 2° cluster	Cluster classico (crisp) di assegnazione
[6,]	0,6559686	0,3440314	1
[10,]	0,3470705	0,6529295	2
[20,]	0,3237037	0,6762963	2
[23,]	0,3237037	0,6762963	2
[33,]	0,6503975	0,3496025	1
[35,]	0,3470705	0,6529295	2
[55,]	0,3470705	0,6529295	2
[62,]	0,6503975	0,3496025	1
[67,]	0,6503975	0,3496025	1
[68,]	0,6559687	0,3440313	1
[75,]	0,3237037	0,6762963	2
[80,]	0,3470705	0,6529295	2
[85,]	0,3237036	0,6762964	2
[94,]	0,3237036	0,6762964	2
[98,]	0,3237036	0,6762964	2

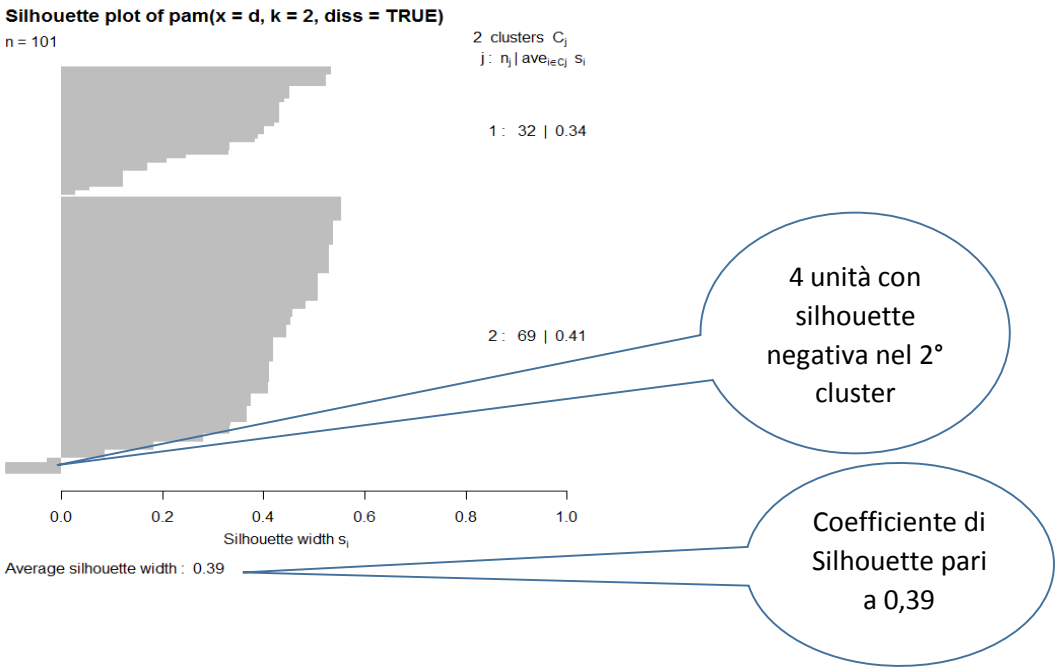
Tabella 4.11 - Unità statistiche con rilevante grado di appartenenza ad un cluster. Fonte: elaborazione personale.

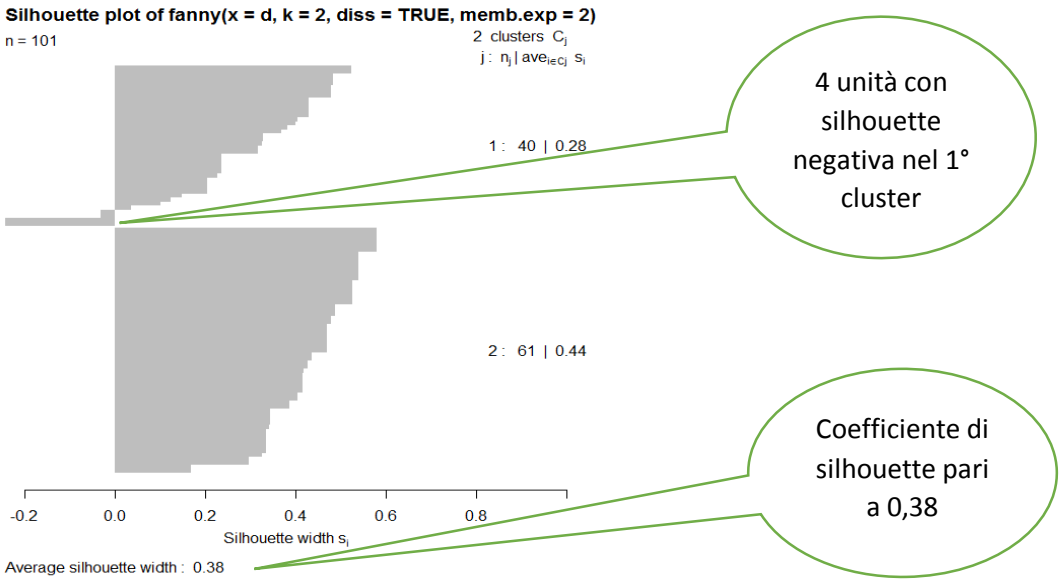
Tutte le altre unità presentano mediamente valori di appartenenza equamente distribuiti fra i due cluster, questo significa che non hanno caratteristiche tali da essere attribuite in modo preponderante ad uno o all'altro cluster.

Al di là di quanto appena detto, quello che risulta interessante per il presente lavoro è effettuare un confronto diretto fra i principali risultati ottenuti precedentemente con l'algoritmo "PAM" e quelli ottenuti adesso con l'algoritmo "FANNY".

Partiamo dal confronto fra i grafici della silhouette.

Il primo grafico rappresenta quanto ottenuto con la classificazione classifica effettuata con l'algoritmo "PAM", mentre il secondo rappresenta quanto ottenuto con la classificazione sfocata effettuata con l'algoritmo "FANNY".





A prima vista si nota subito che essenzialmente la struttura dei cluster è molto simile. Nella classificazione ottenuta con l'algoritmo "FANNY" le unità con silhouette negativa compaiono nel primo cluster anziché nel secondo.

Anche la qualità delle partizioni ottenute è pressoché identica. Il coefficiente di silhouette infatti assume valore 0,39 nel primo caso e 0,38 nel secondo. La qualità del secondo cluster risulta migliore, mentre quella del primo cluster diminuisce. Questo dipende da un valore negativo abbastanza elevato dell'ampiezza di silhouette assunto dalle unità statistiche 12 e 64. Però, come verrà dimostrato successivamente dai coefficienti di appartenenza, tali valori non indicano una "reale" mal allocazione di codeste unità.
Con l'algoritmo "FANNY" le unità statistiche che presentano silhouette negativa sono le seguenti:

Unità statistiche allocate in modo errato		
Unità statistica	**Cluster di assegnazione**	**Silhouette**
12	1	-0,243
40	1	-0,032
64	1	-0,243
87	1	-0,032

Tabella 4.12 - Unità statistiche allocate in modo errato. Fonte: elaborazione personale.

Le unità statistiche 40 e 87 presentano un valore di silhouette negativo ma quasi pari a 0, questo significa che la loro attribuzione al 1° o al 2° cluster, in termini "classici", risulta ambigua. Diverso invece per le unità 12 e 64, che presentano valori negativi di silhouette più elevati anche se comunque non rilevanti, per cui si potrebbe desumere con maggior certezza che la loro attribuzione al 1° cluster, sempre in termini classici, risulta sbagliata. In questa analisi però risulta più utile osservare i gradi di appartenenza di tali unità ai cluster, piuttosto che soltanto le ampiezze di silhouette.
Il punto di forza della classificazione sfocata sta proprio nel venirci incontro nel caso in cui si presentano unità la cui attribuzione all'uno o all'altro cluster risulta difficile in termini classici, cioè per quelle unità che risulterebbe difficile attribuire ad un cluster piuttosto che ad un altro.
La classificazione sfocata ci fornisce un'importante informazione aggiuntiva rispetto alla classificazione classica: il grado di appartenenza di ciascuna unità a ciascun cluster. In questo modo possiamo dire quanto un'unità appartiene ad un dato cluster.

Nel caso in esame, l'algoritmo "FANNY" porta ai seguenti cambiamenti rispetto alla precedente classificazione ottenuta con l'algoritmo "PAM":

1) 8 unità che nella procedura "PAM" erano state attribuite al secondo cluster vengono ora spostate nel 1° cluster;
2) 4 di queste unità sono proprio quelle unità che, in base ai valori di silhouette, erano state attribuite in modo sbagliato al 2° cluster nella procedura "PAM" (cioè le unità 32, 44, 96, 97). Si ottiene per cui una migliore attribuzione di tali unità;
3) Le altre 4 unità sono unità che con l'algoritmo "FANNY" assumono silhouette negativa e che presentano un coefficiente di appartenenza ai due cluster vicino allo 0,5. Questo significa che si trovano in una posizione intermedia fra i due cluster e di conseguenza la loro totale attribuzione all'uno o all'altro gruppo risulterebbe sbagliata in un'analisi di classificazione classica. A mio avviso sono proprio queste le unità di maggior interesse per la presente analisi;
4) Tutte le altre unità vengono attribuite agli stessi cluster da entrambi gli algoritmi.

Di seguito viene riportata una tabella riassuntiva delle unità classificate in maniera differente dai due metodi, dove le prime quattro unità (in arancione) sono quelle che presentano ampiezza di silhouette negativa nella procedura "FANNY" mentre le seconde 4 unità (in giallo) sono quelle che presentavano silhouette negativa nella precedente procedura "PAM".

Unità considerate	Cluster di assegnazione FANNY	Silhouette procedura PAM	Silhouette procedura FANNY	Cluster di assegnazione PAM	Appartenenza al 1° cluster	Appartenenza al 2° cluster
12	1	0,28104	-0,24301	2	0,5211311	0,4788689
40	1	0,08556	-0,03167	2	0,5093808	0,4906192
64	1	0,28104	-0,24301	2	0,5211311	0,4788689
87	1	0,08556	-0,03167	2	0,5093808	0,4906192
32	1	-0,10994	0,23513	2	0,5211311	0,4788689
44	1	-0,10994	0,23513	2	0,6147515	0,3852485
96	1	-0,10994	0,23513	2	0,6147516	0,3852484
97	1	-0,02710	0,12208	2	0,5815297	0,4184703

Tabella 4.13 - Unità statistiche allocate in modo differente dai due metodi di classificazione analizzati. Fonte: elaborazione personale.

Analizzando tali unità e osservando prima di tutto il coefficiente di appartenenza, si può concludere che:

- ✓ Delle quattro unità che nella procedura "PAM" presentavano silhouette negativa soltanto l'unità 32 risulta di difficile assegnazione. Essa presenta infatti dei coefficienti di appartenenza molto vicini allo 0,5 a differenza delle altre 3 unità. Da notare che la sola silhouette non ci è di molto aiuto nell'analisi "FANNY". Infatti guardando la silhouette dell'unità 97 e quella dell'unità 32 ci verrebbe da dire che l'attribuzione della prima al 1° cluster risulta più ambigua rispetto all'attribuzione della seconda al 1° cluster. I coefficienti di appartenenza invece ci suggeriscono l'incontrario: l'unità 97 appartiene sicuramente di più al 1° cluster mentre l'unità 32 risulta particolarmente sfumata fra i due gruppi e quindi di difficile attribuzione. La procedura "FANNY", quindi, porta sicuramente ad una migliore attribuzione delle unità [44, 96, 97], mentre l'unità 32 può essere attribuita indistintamente ad uno dei due cluster.

- ✓ Per le quattro unità che presentano silhouette negativa nella procedura "FANNY" si può invece osservare che sono tutte unità che si trovano in posizione intermedia fra i due cluster. I relativi coefficienti di appartenenza sono infatti vicini allo 0,5 in particolar modo per le unità statistiche 40 e 87. La classificazione sfocata risulta molto utile per classificare queste unità: essa infatti ci mostra che la percentuale di appartenenza ai due cluster è pressoché identica, un'informazione che invece non avevamo con la classificazione classica. Come si può desumere, la procedura "PAM" sbagliava in pieno attribuendo tali unità completamente al 2° cluster. Esse, sulla base delle caratteristiche che posseggono (date dalle variabili di clusterizzazione) appartengono per metà ad un cluster e per l'altra metà all'altro cluster, si trovano cioè in posizione intermedia. La silhouette relativa alla singola unità statistica, come già detto in precedenza, non risulta molto utile in un'analisi di classificazione sfocata: essa, se negativa, ci dice soltanto che in termini classici l'attribuzione di quell'unità al cluster sarebbe sbagliata, ma in termini di analisi sfocata quello che bisogna osservare sono i coefficienti di appartenenza, sono loro

che ci forniscono l'informazione principale e ci permettono di capire il vero punto di forza della classificazione sfocata.

L'ultima cosa che è necessario specificare è che ci sono stati alcuni casi di unità statistiche che, nonostante si trovassero in una posizione intermedia (vedi l'elevata sfumatura che le caratterizza), sono state assegnate dalla procedura "PAM" allo stesso cluster di assegnazione della procedura "FANNY". Essendo la classificazione sfocata più adatta a classificare unità di questo tipo, possiamo concludere che per questi casi la procedura di classificazione PAM è riuscita a fornire il risultato corretto.
Le unità statistiche in questione sono le seguenti:

Unità statistica	Appartenenza al 1° cluster	Appartenenza al 2° cluster	Cluster classico (crisp) di assegnazione: PAM e FANNY
[4,]	0,4701927	0,5298073	2
[19,]	0,5269588	0,4730412	1
[24,]	0,4850267	0,5149733	2
[51,]	0,4714343	0,5285657	2
[71,]	0,4850267	0,5149733	2
[73,]	0,4714343	0,5285657	2
[81,]	0,5325863	0,4674137	1
[83,]	0,4850267	0,5149733	2
[91,]	0,4850267	0,5149733	2
[101,]	0,4850267	0,5149733	2

Tabella 4.14 - Unità statistiche assegnate agli stessi clusters dai due metodi di classificazione analizzati. Fonte: elaborazione personale.

Dalla tabella è possibile avere conferma che tali unità sono caratterizzate da un alto grado di sfocatura, cioè si trovano in una posizione pressoché centrale fra i due cluster individuati.

2

G = 3, m = 2

L'obiettivo di questo secondo punto è quello di verificare se l'algoritmo "FANNY" può portare ad un miglioramento qualitativo della classificazione rispetto alla precedente classificazione classica in 3 gruppi.

Applicando la procedura "FANNY" ponendo il numero di gruppi uguale a 3 e mantenendo un coefficiente di sfocatura pari a 2, si ottiene una partizione così caratterizzata:

- L'indice di Dunn è uguale a 0,0445. Tale valore risulta essere leggermente inferiore a quello ottenuto con la classificazione in due gruppi. Questo significa che il grado di sfocatura aumenta, anche se in maniera impercettibile.
- Una particolarità della classificazione ottenuta ponendo il numero di gruppi uguale a 3 è che ciascuna unità statistica presenta lo stesso coefficiente di appartenenza al secondo e al terzo cluster. Questo significa che quei due gruppi non hanno caratteristiche tali da essere considerati differenti, sono cioè perfettamente identici.
- Le unità con silhouette negativa diventano 7, mentre nel caso precedente erano 6.
- Di seguito il confronto fra il grafico della silhouette generato dalla funzione "PAM" e quello ottenuto con la procedura "FANNY":

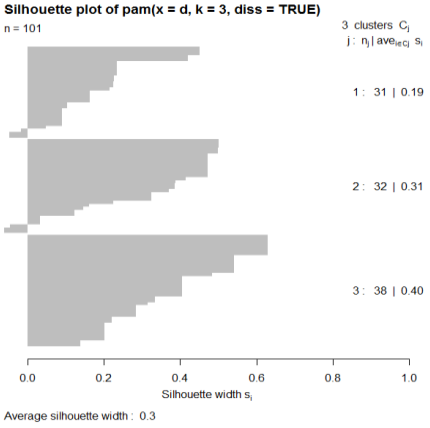


Figura 4.45 - Silhouette procedura PAM

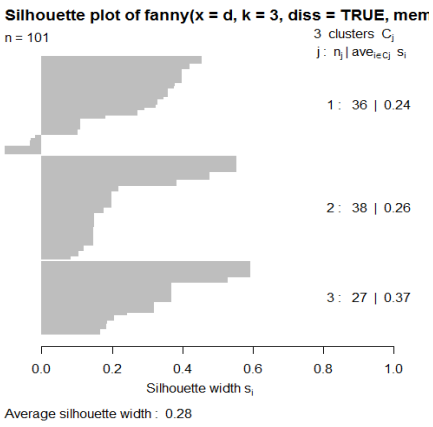


Figura 4.46 - Silhouette procedura FANNY

- Come si può notare la qualità complessiva della partizione diminuisce leggermente: il coefficiente di silhouette passa da 0,3 a 0,28. Le unità statistiche con silhouette negativa compaiono tutte all'interno del primo cluster.

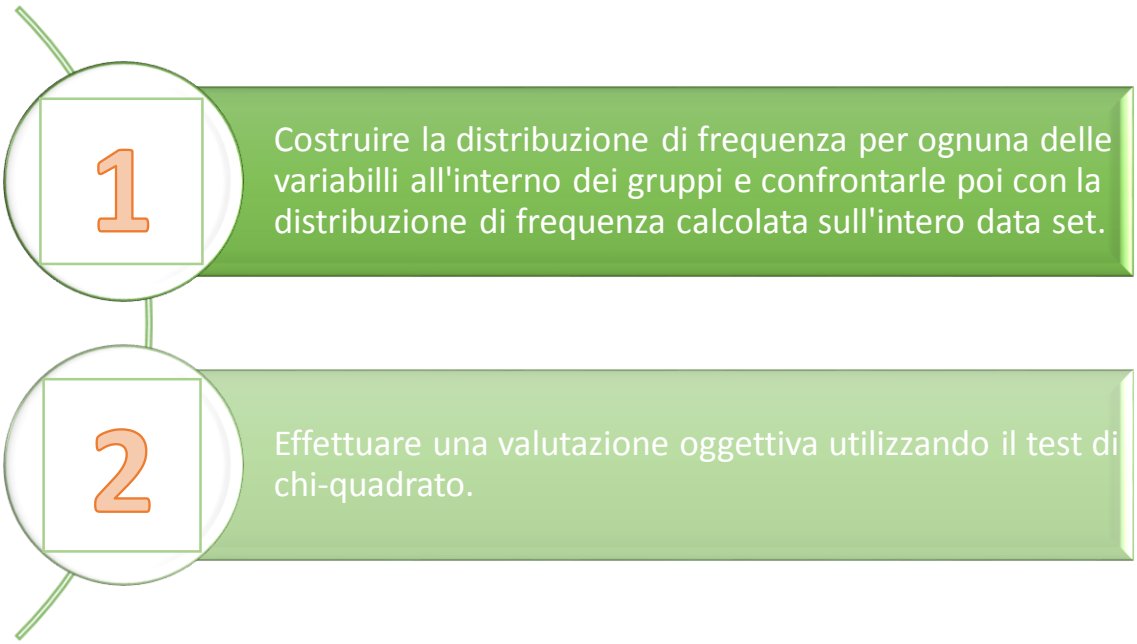
Quello che si può concludere sulla partizione ottenuta, è che ponendo il numero di gruppi pari a 3 non si ottiene un partizionamento che ha senso. Il fatto che il secondo e il terzo cluster siano composti da unità statistiche che presentano un grado di appartenenza ripartito in modo perfettamente identico fra i due cluster in questione, non ci fornisce un'informazione supplementare. In altre parole ha più senso considerare questi due gruppi un unico gruppo formato da 65 unità statistiche.
In base a quanto detto possiamo concludere che il partizionamento in 2 gruppi risulta essere quello qualitativamente migliore.

G = 2, m = 3

Questo ultimo breve punto vuole dimostrare che in questo caso, trattandosi di una partizionamento già estremamente sfocato, aumentare il grado di sfocatura *m* non è possibile: quello che si ottiene è infatti una partizionamento composto da unità statistiche aventi tutte una grado di appartenenza pari a 0,5. In altre parole le unità vengono considerate con caratteristiche tali da trovarsi in una posizione perfettamente intermedia ai due cluster. I due gruppi che si ottengono sono per cui gruppi identici.

4.3.4 Valutazione della composizione dei cluster

Il seguente paragrafo si propone di effettuare un'analisi volta a fornire una valutazione dell'importanza di ciascuna variabile di clusterizzazione nella formazione dei gruppi.
Sarà poi possibile definire un nome per i cluster utilizzando quelle variabili che sono risultate particolarmente importanti nella formazione dei cluster stessi.
Per capire se e in che misura una variabile che abbiamo utilizzato per effettuare l'analisi dei cluster è importante nella formazione dei gruppi stessi possiamo procedere in due modi, tenendo presente che nel caso in esame abbiamo a che fare con tutte variabili qualitative:



In entrambi i due punti sopra elencati l'obiettivo è quello di verificare se la composizione dei gruppi è simile a quella dell'intera popolazione in esame. Nel primo caso si vanno a confrontare i grafici delle distribuzioni di frequenza di ciascuna variabile utilizzata per l'analisi dei cluster con le distribuzioni di frequenza delle stesse variabili ma riferite all'intera popolazione.
Nel secondo caso invece si effettua una valutazione oggettiva utilizzando il test chi-quadrato.

Confronto fra le rappresentazioni grafiche delle distribuzioni di frequenza

La seguente tabella contiene un prospetto completo delle frequenze relative di ciascuna variabile utilizzata per l'analisi dei cluster: la parte evidenziata in arancione contiene le frequenze riferite ai clusters, mentre la parte evidenziata in verde contiene le frequenze riferite all'intero data set.

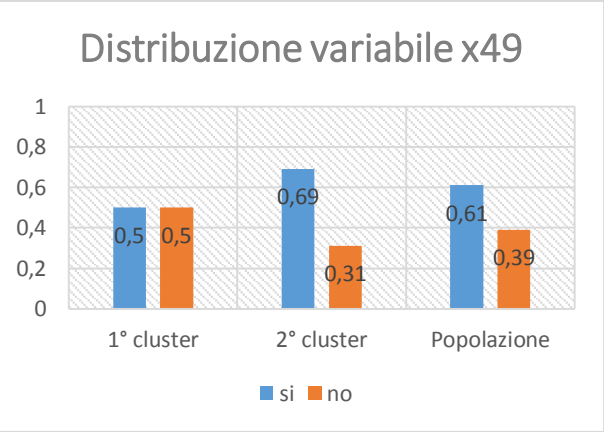
Variabile	**Frequenze Cluster 1**		**Frequenze Cluster 2**		**Frequenze popolazione**	
Modalità	**Si**	**No**	**Si**	**No**	**Si**	**No**
x49	0,50	0,50	0,69	0,31	0,61	0,39
x72	0,18	0,83	0,46	0,54	0,35	0,65
x74	0,70	0,30	0,51	0,49	0,58	0,42
x81	0,48	0,53	0,28	0,72	0,36	0,64
x91	0,08	0,93	0,03	0,97	0,05	0,95
x96sì	0,60	0,40	0	1	0,24	0,76
x96nonso	0,05	0,95	0,97	0,03	0,60	0,40

Tabella 4.15 - Frequenze variabili di clusterizzazione. Fonte: elaborazione personale.

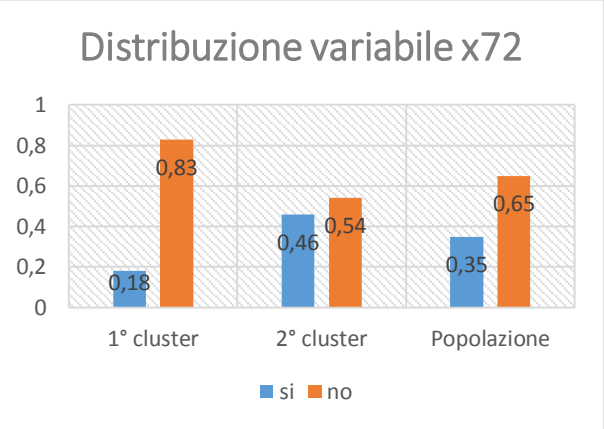
Attraverso questa tabella è possibile procedere con la rappresentazione grafica delle distribuzioni la quale ci permette di capire con maggiore facilità l'importanza di ciascuna variabile nella formazione dei gruppi.

Ogni grafico che segue contiene la distribuzione della variabile in esame nel 1° cluster, nel 2° cluster ed infine nella popolazione. In questo modo è possibile fornire una prima valutazione circa l'uguaglianza delle distribuzioni all'interno dei cluster con quella riferita all'intero data set.

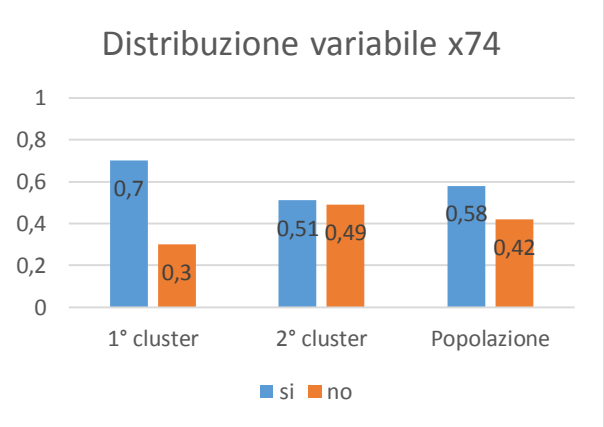
Se si verifica che la distribuzione della variabile in esame, riferita per esempio al 1° cluster, è simile a quella della popolazione allora tale variabile non è importante nella formazione di quel gruppo.



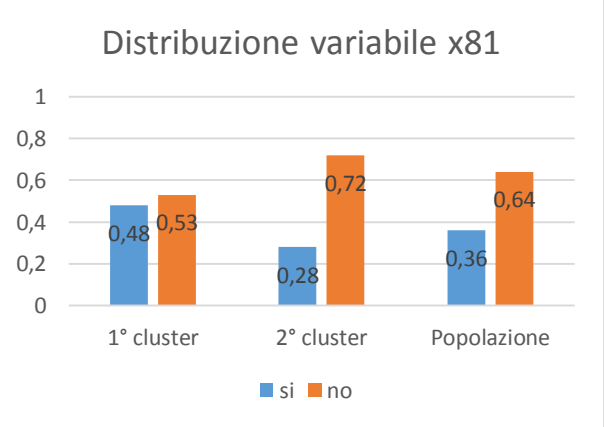
La distribuzione della variabile all'interno dei due gruppi non presenta differenze significative rispetto a quella nella popolazione, per cui osservando semplicemente i grafici risulta difficile affermare se sia importante o meno nella formazione dei due gruppi.



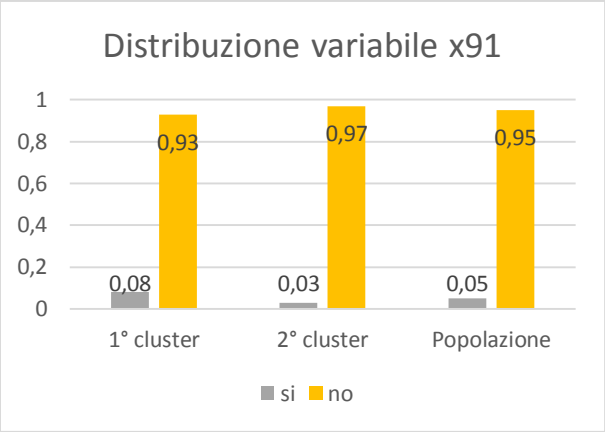
In questo caso invece la distribuzione all'interno del 1° cluster è significativamente diversa da quella nella popolazione, mentre risulta più complicato trarre delle conclusioni sul 2° cluster.



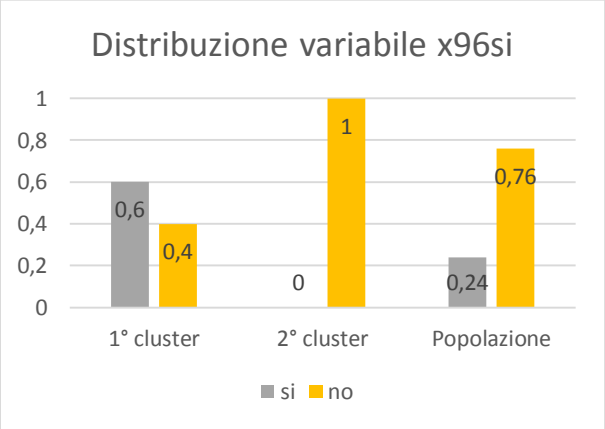
La distribuzione della variabile in esame all'interno del 1° cluster presenta una certa differenza con quella riferita all'intera popolazione, mentre quella del 2° cluster risulta molto simile. Il test chi-quadrato chiarirà le idee.



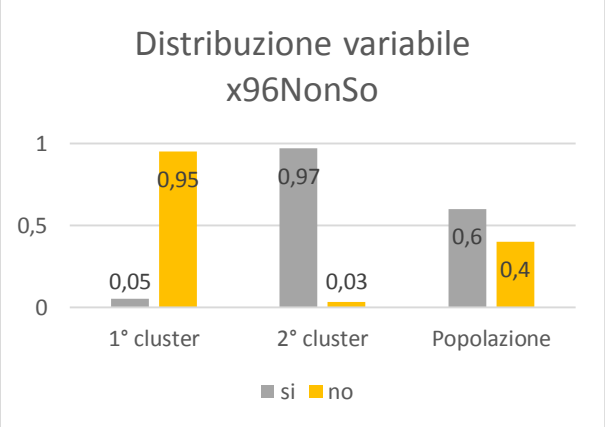
Anche per questa variabile, come per la X49 risulta difficile trarre delle conclusioni poiché i valori delle frequenze sono abbastanza simili. Ovviamente il test chi-quadrato potrà fornire una valutazione migliore.



La variabile x91 come si può facilmente osservare non è importante nella formazione di entrambi i gruppi. Le distribuzioni nei due gruppi sono pressoché identiche a quella nella popolazione.



Per la variabile x96si invece possiamo affermare con sicurezza che essa è importante per la formazione di entrambi i cluster.



Anche per quest'ultima variabile possiamo affermare con certezza che essa risulta importante nella formazione dei due gruppi.

Come si può capire una valutazione basata sul confronto delle rappresentazioni grafiche delle distribuzioni può altro che non essere una valutazione puramente soggettiva, nel senso che si tratta di una valutazione influenzata dal soggetto che esegue l'analisi.
Tale confronto risulta comunque molto importante perché ci fornisce una prima idea circa l'importanza delle variabili di clusterizzazione nella formazione dei gruppi.
Per avere una valutazione oggettiva si può ricorrere, nel caso di variabili qualitative, al test non parametrico chi-quadrato.

Test Chi-quadrato

Il test chi-quadrato è l'unico test che si può utilizzare con variabili qualitative. Esso è utilizzato per valutare il grado di somiglianza di due distribuzioni: quella empirica e quella teorica. Nel nostro caso la distribuzione empirica è la distribuzione riferita ai clusters mentre quella teorica è quella riferita all'intero data set.
Se il valore della statistica chi-quadrato è pari a zero significa che le due distribuzioni sono identiche, mentre tanto più è grande tanto più le distribuzioni sono diverse.
Di conseguenza se confrontiamo la distribuzione di una variabile all'interno di un singolo cluster con la distribuzione su tutti i dati, tale variabile risulta tanto più importante nella formazione di quel cluster quanto più elevato è il valore della statistica test. In altre parole, tanto maggiore è il valore della statistica test tanto più la variabile è probabilmente importante nel distinguere quel cluster dagli altri.

Esiste un valore critico che ci permette di dire se la variabile è importante o meno. Tale valore è determinato dal *p-value* con cui effettuiamo il test e determina l'ampiezza della regione di rifiuto dell'ipotesi nulla. I valori passibili per il p-value sono: 0,10 – 0,05 – 0,01.
Il sistema di ipotesi per il caso in esame è il seguente:

$$\begin{cases} H_0: \ Distribuzione\ cluster\ = distribuzione\ popolazione \\ H_1: \ Distribuzione\ cluster \neq distribuzione\ popolazione \end{cases}$$

L'obiettivo è rifiutare l'ipotesi nulla per poter affermare che la variabile presa in esame è utile nella formazione del cluster. Il *p-value* inizialmente utilizzato è 0,05. Nel caso si

dovesse accettare l'ipotesi nulla con tale livello del test, verrà effettuata un'ulteriore prova con un *p-value* di 0,10.

Verrà effettuato un test per ciascuna variabile utilizzando un intervallo di confidenza pari a 0,95 a cui corrisponde un *p-value* di 0,05. Se ciò portasse ad accettare l'ipotesi nulla verrà effettuato un ulteriore test con intervallo di confidenza uguale a 0,90 e quindi *p-value* di 0,10.

Variabile	Valore di chi-quadrato	p-value test	Accetto H_1 con p-value 0,05?	Accetto H_1 con p-value 0,10?
X49	2.87	0.09021	No	Si
X72	7.40	0.006533	Si	Si
X74	2.91	0.08794	No	Si
X81	3.24	0.07151	No	Si
X91	0.24	0.6259	No	No
X96SI	44.75	2.235e-11	Si	Si
X96NONSO	81.18	2.2e-16	Si	Si

Tabella 4.16 - Valori test chi-quadrato. Fonte: elaborazione personale.

Osservando la tabella sopra riportata si nota che:

- Con un p-value di 0,05 soltanto 3 variabili delle 7 utilizzate risultano importanti nella formazione dei gruppi. Si tratta delle variabili "x72, x96si ed x96nonso". Considerando il p-value ottenuto con il test abbiamo la certezza che tali variabili, in particolar modo la "x96si" e la "x96nonso", sono importanti nel differenziare i due gruppi. Infatti anche effettuando il test con un p-value di 0,01 rifiuteremmo comunque l'ipotesi nulla. Inoltre, come ulteriore conferma, abbiamo un valore molto elevato della statistica test chi-quadrato.

- Con un p-value di 0,10 invece le variabili importanti nel differenziare i clusters diventano 6 su 7. In questo caso abbiamo la certezza che la variabile "x91" non è risultata importante nel processo di clusterizzazione alla luce dell'elevato valore del p-value ottenuto con il test e del valore della statistica test chi-quadrato prossima allo zero. È da precisare però che le variabili "x49, x74 ed x81" cadono appena dopo il valore critico che separa la regione di accettazione dell'ipotesi nulla da quella di rifiuto. Esse presentano infatti un valore della statistica test chi-quadrato abbastanza basso.

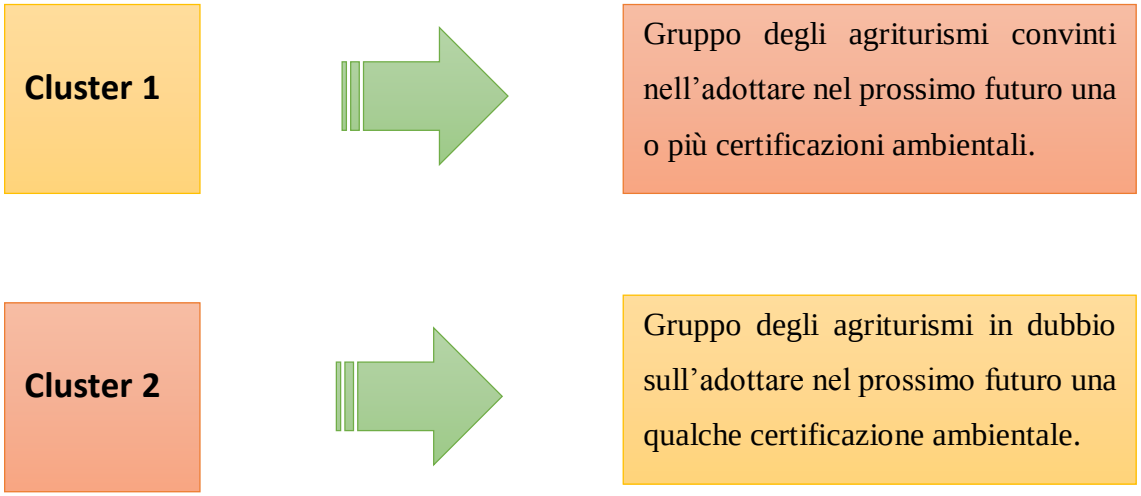
A conclusione del paragrafo si cercherà ora di dare un nome ai cluster ottenuti, utilizzando le tre variabili che sono risultate particolarmente importanti nel processo di clusterizzazione.
Trattandosi di variabili qualitative confrontiamo le frequenze relative assunte da tali variabili all'interno dei cluster con quelle riferite all'intero data-set.

Variabile	Frequenza modalità "Si" nel 1° cluster	Frequenza modalità "Si" nel 2° cluster	Frequenza modalità "Si" nel data-set
X72	0,18	0,46	0,35
X96SI	0,60	0	0,24
X96NONSO	0,05	0,97	0,60

Tabella 4.17 - Frequenze delle variabili ritenute importanti nel processo di formazione dei gruppi. Fonte: elaborazione personale.

Nel caso in esame il nome dei cluster può essere dato utilizzando soltanto le ultime due variabili della tabella poiché sono quelle che presentano le maggiori differenze nelle frequenze assunte all'interno dei cluster rispetto a quelle assunte nell'intero data-set.

Ricordando che la variabile “X96si” indica l’intenzione di adottare in futuro una qualche certificazione ambientale e la variabile “X96NonSo” indica il dubbio se adottare o meno in futuro una qualche certificazione ambientale, i due cluster possono essere così nominati:



4.3.5 Profilazione dei gruppi

All'interno di quest'ultima parte verrà affrontata la profilazione dei gruppi, cioè la descrizione del contenuto dei clusters ottenuti in seguito al processo di clusterizzazione. Per perseguire questo obiettivo vengono utilizzate delle variabili dette "variabili di profilazione": si tratta di variabili che non sono state utilizzate per classificare le unità statistiche e che servono appunto per valutare quello che c'è dentro ai gruppi.

Per poter profilare i gruppi è necessario seguire alcuni passi:

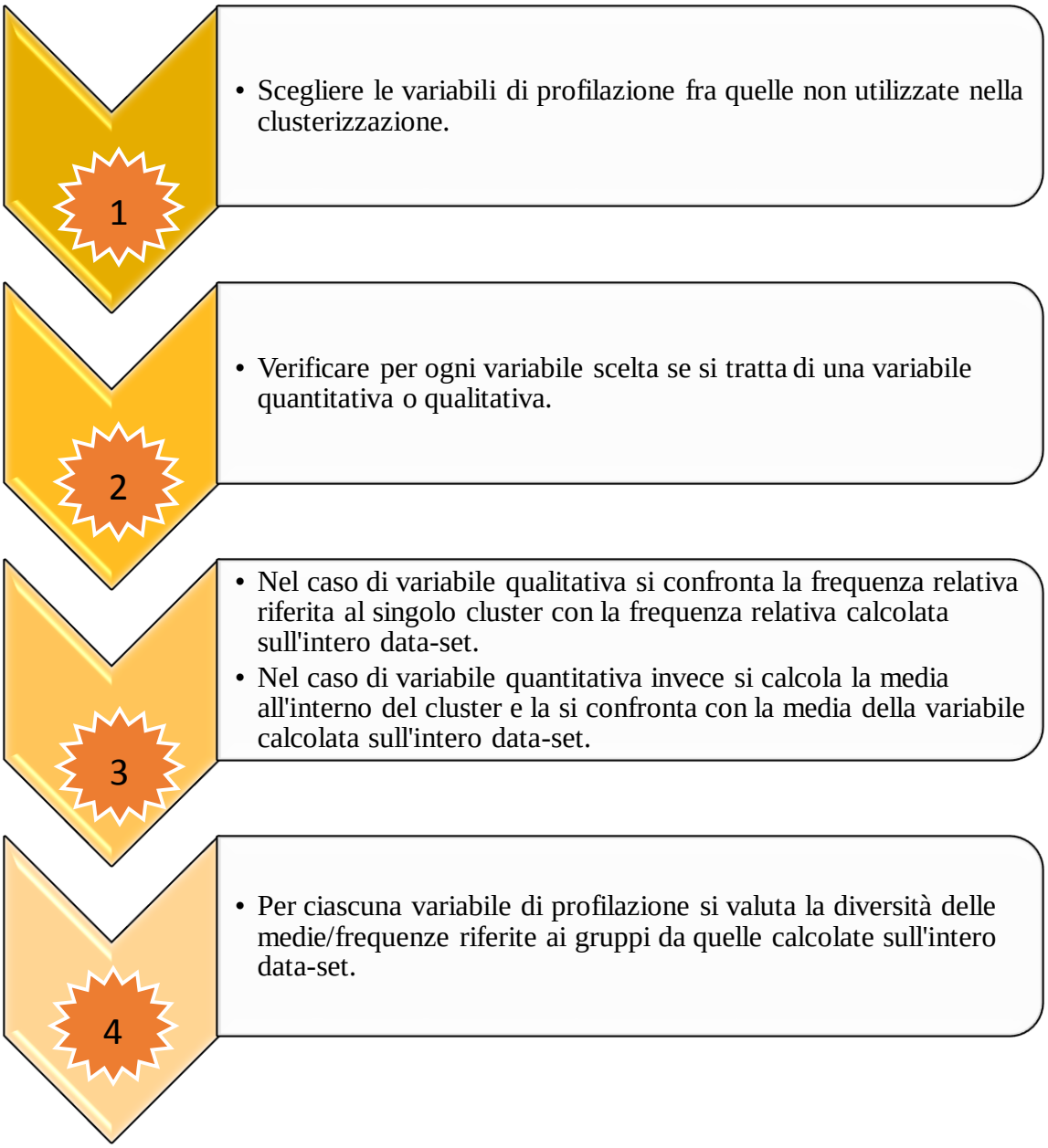


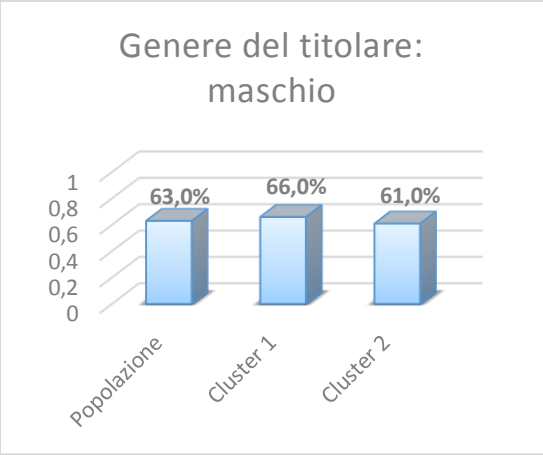
Figura 4.47 - Processo di profilazione adottato nel presente lavoro. Fonte: elaborazione personale.

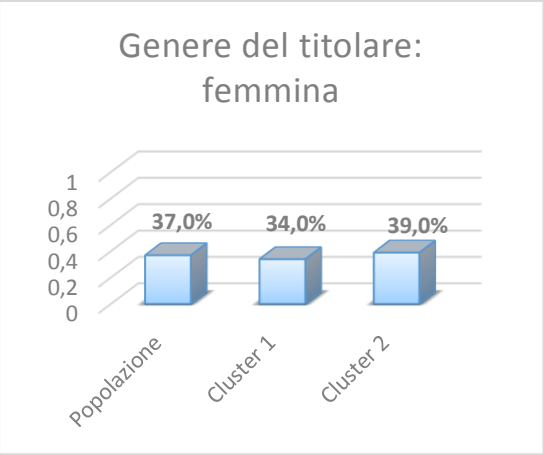
Per il caso in esame si è deciso di utilizzare le seguenti variabili riferite alle diverse domande contenute nel questionario:

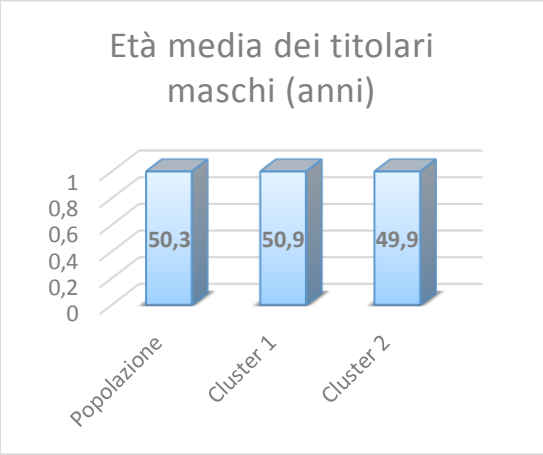
Variabile	Descrizione	Tipo variabile	Frequenze/ medie pop.	Cluster 1	Cluster 2
a1	Genere titolare: maschio	Qualitativa	0,63	0,8	0,79
a3	Genere titolare: femmina	Qualitativa	0,37	0,5	0,46
a5-a9	Età titolari maschi (media)	Quantitativa	50,3	50,9	49,9
a11-a15	Età titolari femmina (media)	Quantitativa	47,3	48,3	46,7
a17-19	Titolo di studio licenza elementare media maschio	Qualitativa	0,28	0,35	0,1
a17-19	Titolo di studio licenza superiore maschio	Qualitativa	0,47	0,68	0,13
a17-19	Titolo di studio laure/post laurea maschio	Qualitativa	0,25	0,2	0,24
a20-a22	Titolo di studio licenza elementare media femmina	Qualitativa	0,21	0,1	0,05
a20-a22	Titolo di studio licenza superiore femmina	Qualitativa	0,46	0,1	0,13
a20-a22	Titolo di studio laure/post laurea femmina	Qualitativa	0,33	0,2	0,06
a23-a27	Dimensione economica complessiva: meno di 20.000 euro	Qualitativa	0,13	0,12	0,13
a23-a27	Dimensione economica complessiva: tra 20.000 e 40.000 euro	Qualitativa	0,21	0,25	0,18
a23-a27	Dimensione economica complessiva: tra 40.000 e 100.000 euro	Qualitativa	0,26	0,23	0,28
a23-a27	Dimensione economica complessiva: tra 100.000 e 500.000 euro	Qualitativa	0,33	0,3	0,34
a23-a27	Dimensione economica complessiva: oltre 500.000 euro	Qualitativa	0,05	0,05	0,05
a70	Servizio di alloggio	Qualitativa	0,75	0,8	0,72
a76	Servizio di ristorazione	Qualitativa	0,41	0,53	0,33

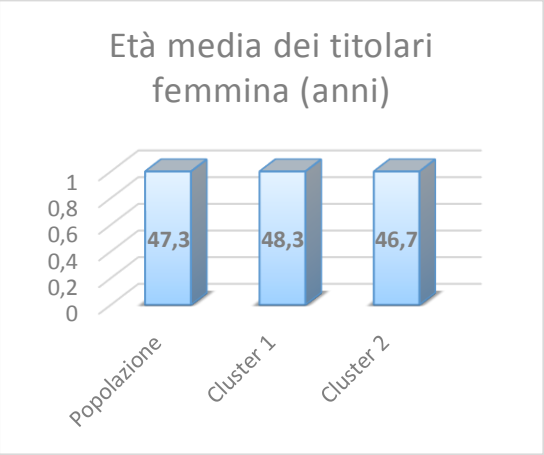
Tabella 4.18 - Variabili considerate per la profilazione. Fonte: elaborazione personale.

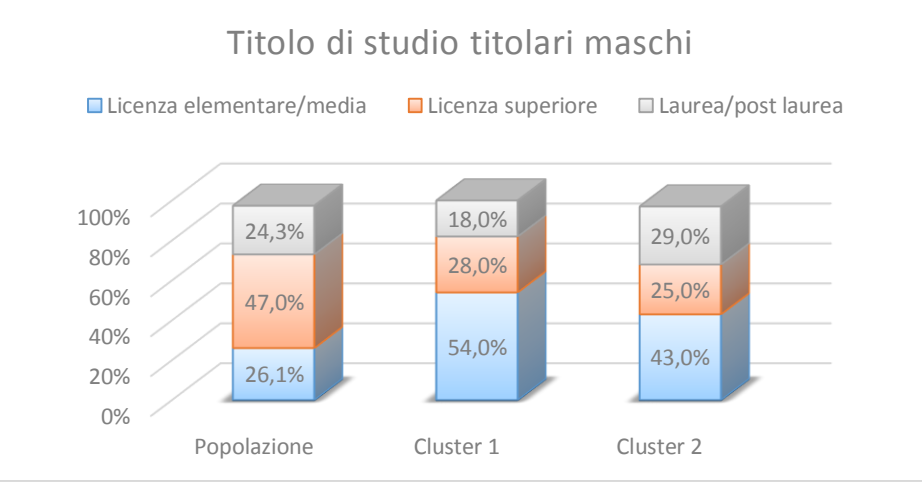
Per capire meglio quali sono le caratteristiche predominanti degli agriturismi nei due segmenti che abbiamo ottenuto si può procedere alla rappresentazione grafica delle variabili appena indicate. Sarà poi così possibile definire il profilo dei due gruppi.

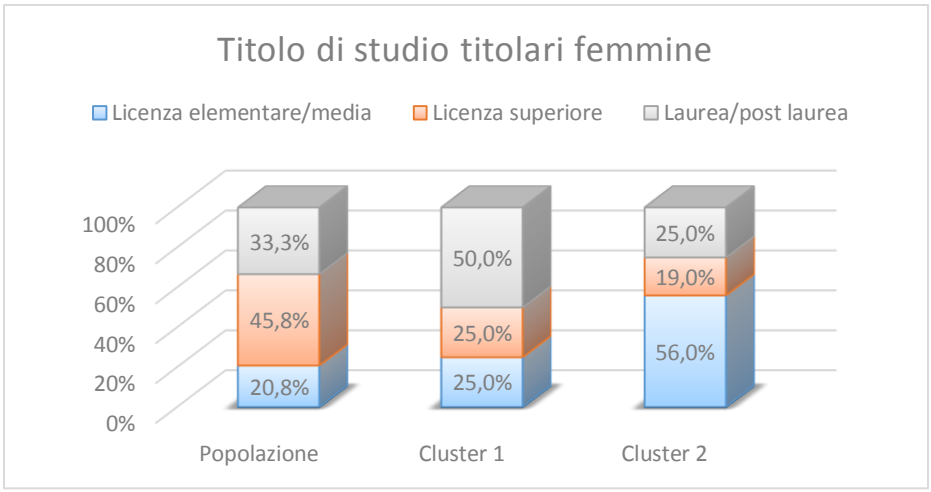

Genere del titolare: maschio
1
0,8
0,6
0,4
0,2
0
63,0%
66,0%
61,0%
Popolazione
Cluster 1
Cluster 2

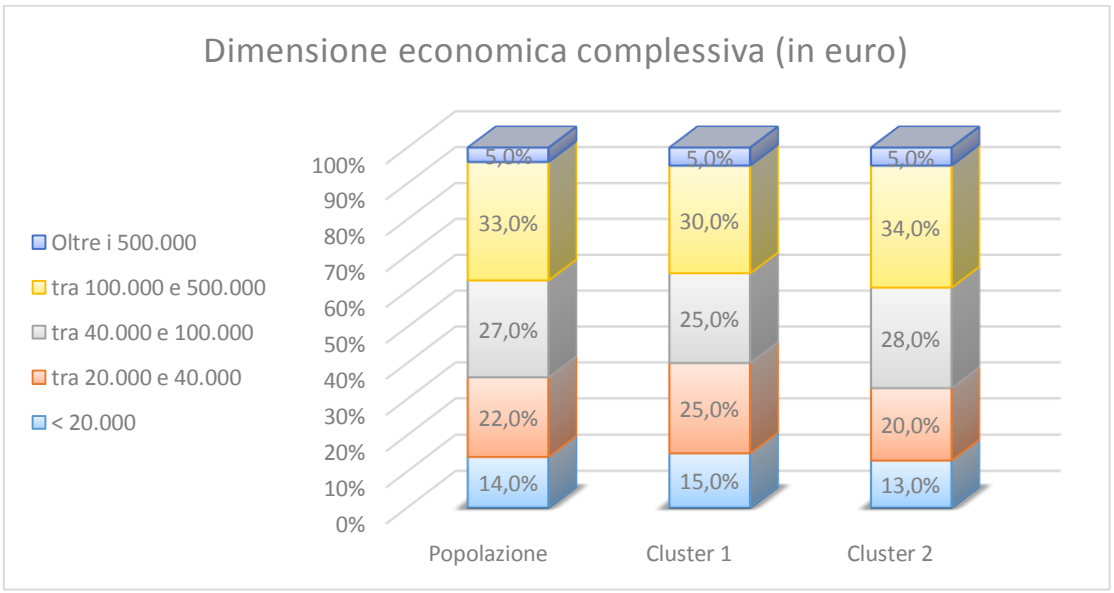

Genere del titolare: femmina
1
0,8
0,6
0,4
0,2
0
37,0%
34,0%
39,0%
Popolazione
Cluster 1
Cluster 2

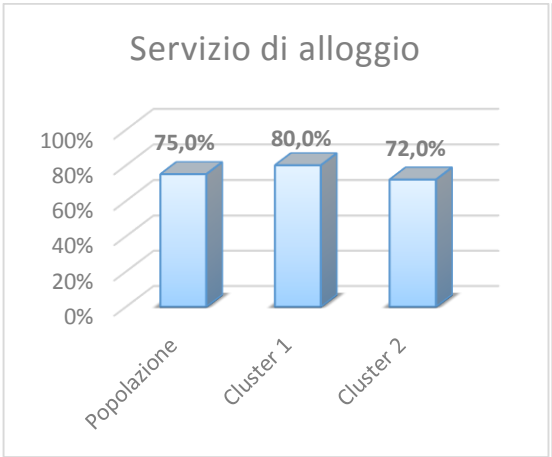

Età media dei titolari maschi (anni)
1
0,8
0,6
0,4
0,2
0
50,3
50,9
49,9
Popolazione
Cluster 1
Cluster 2


Età media dei titolari femmina (anni)
1
0,8
0,6
0,4
0,2
0
47,3
48,3
46,7
Popolazione
Cluster 1
Cluster 2


Titolo di studio titolari maschi
Licenza elementare/media
Licenza superiore
Laurea/post laurea
100%
80%
60%
40%
20%
0%
24,3%
47,0%
26,1%
18,0%
28,0%
54,0%
29,0%
25,0%
43,0%
Popolazione
Cluster 1
Cluster 2


Titolo di studio titolari femmine
Licenza elementare/media
Licenza superiore
Laurea/post laurea
100%
80%
60%
40%
20%
0%
33,3%
45,8%
20,8%
50,0%
25,0%
25,0%
25,0%
19,0%
56,0%
Popolazione
Cluster 1
Cluster 2


Dimensione economica complessiva (in euro)
Oltre i 500.000
tra 100.000 e 500.000
tra 40.000 e 100.000
tra 20.000 e 40.000
< 20.000
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
5,0%
33,0%
27,0%
22,0%
14,0%
5,0%
30,0%
25,0%
25,0%
15,0%
5,0%
34,0%
28,0%
20,0%
13,0%
Popolazione
Cluster 1
Cluster 2


Servizio di alloggio
100%
80%
60%
40%
20%
0%
75,0%
80,0%
72,0%
Popolazione
Cluster 1
Cluster 2

Servizio di ristorazione
100%
80%
60%
40%
20%
0%
41,0%
53,0%
33,0%
Popolazione
Cluster 1
Cluster 2

Sulla base delle rappresentazioni grafiche sopra riportate, i due gruppi sono mediamente composti da agriturismi aventi le seguenti caratteristiche:

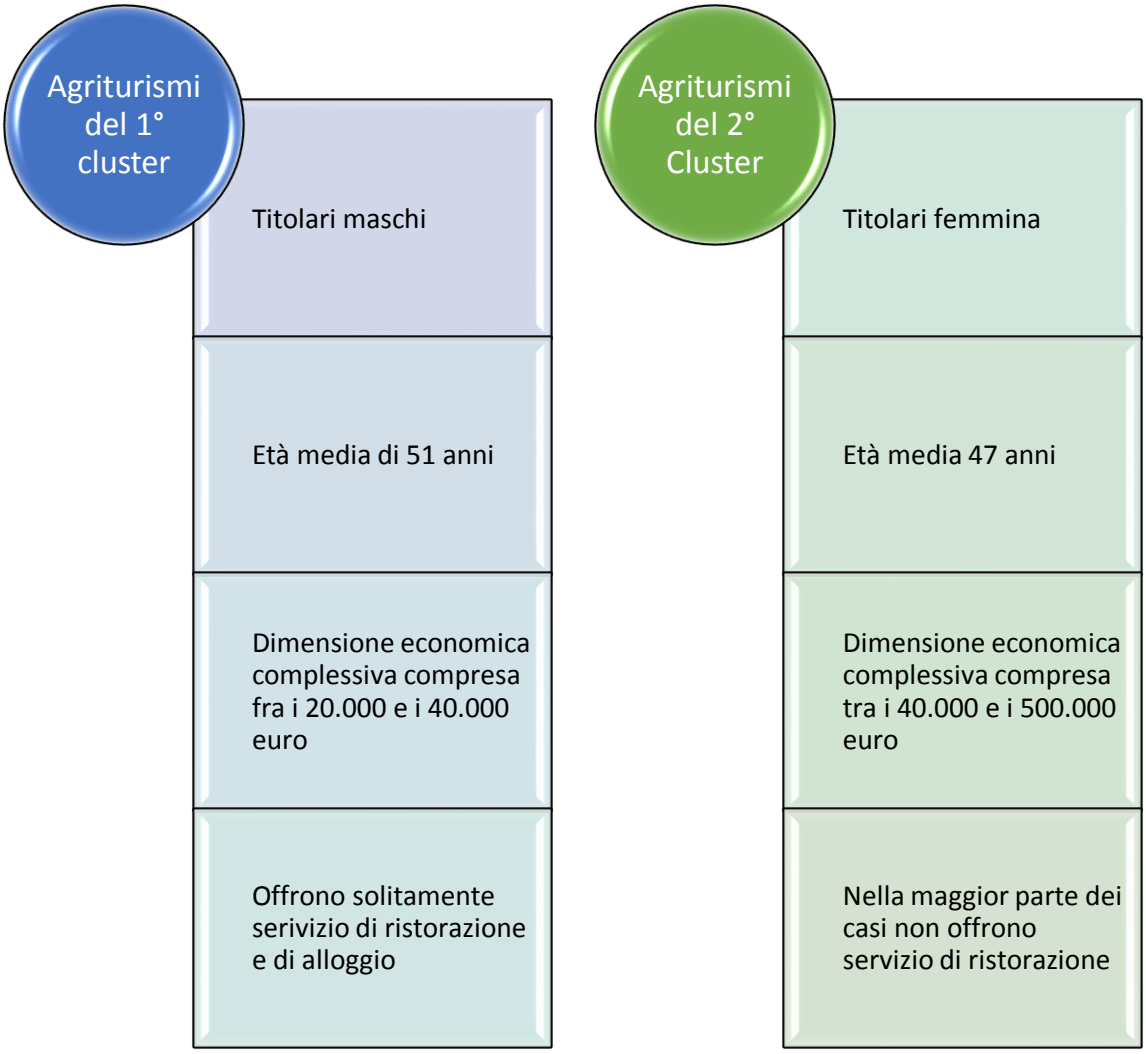


Figura 4.48 - Profili dei clusters. Fonte: elaborazione personale.

Conclusioni

Come già detto all'inizio dell'ultimo capitolo di questo lavoro, la prima intenzione era quella di definire la situazione attuale in termini di sostenibilità ambientale dell'attività degli agriturismi nella regione Veneto. Questo attraverso un'analisi dei cluster volta ad identificare gruppi di agriturismi che presentassero caratteristiche simili sulla base del profilo descritto da un insieme di variabili inerenti la sostenibilità.

Si è dovuto però cambiare obiettivo a causa del riscontro molto deludente da parte degli agriturismi intervistati: infatti solo l'8,6 %, cioè 101 agriturismi su 1178, ha fornito risposta al questionario nonostante le ripetute sollecitazioni.

Dunque un'analisi del genere diventava priva di senso. Si è quindi deciso di effettuare un'analisi con l'obiettivo di mostrare in modo completo l'iter di svolgimento di un'analisi dei cluster.

In particolar modo l'obiettivo vero e proprio del lavoro è stato il confronto fra due metodologie di analisi dei cluster: quella classica e quella sfumata.

A partire dal data set contenente le risposte al questionario fornite dagli agriturismi del Veneto sono state scelte alcune variabili ritenute interessanti, sulla base di determinate valutazioni, per poter effettuare l'analisi di classificazione. Si è trattato di variabili riferite a domande riguardanti gli strumenti adottabili dagli agriturismi per la gestione sostenibile della propria attività.

Riguardo alle variabili scelte per la classificazione si è notato che, dopo avere effettuato le analisi, sono risultate quasi tutte indipendenti. Solo la coppia "Intenzione 'Si' di adottare certificazioni ambientali in futuro, Intenzione 'Non So' di adottare certificazioni ambientali in futuro" ha mostrato una associazione elevata. Tali variabili infatti fanno riferimento alla stessa domanda del questionario, di conseguenza è naturale che siano dipendenti l'una dall'altra.

Effettuate tutte le analisi sulle variabili scelte per la classificazione si è proceduto con la costruzione della matrice delle dissimilarità, elemento essenziale per poter effettuare qualsiasi processo di classificazione. Nel caso affrontato si è utilizzato l'indice di Gower poiché si adatta bene alla tipologia di variabili trattate.

Si è effettuata a questo punto l'analisi classica dei cluster, applicando inizialmente due algoritmi gerarchici, cioè il metodo del legame completo e il metodo di Ward, allo scopo di capire quale fosse il numero ragionevole di gruppi in cui suddividere le unità statistiche, il quale è risultato essere due. Tale suddivisione era prevedibile alla luce del fatto che il data set in esame è di piccole dimensioni.

La conferma si è avuta dapprima attraverso lo strumento del dendrogramma, il quale ha mostrato dove si verificava la massima differenza fra i livelli di fusione dei gruppi, e in seguito attraverso il grafico del processo di aggregazione delle unità. Un ulteriore conferma si è avuta poi osservando i valori numerici dei livelli di fusione.

Con il partizionamento in due gruppi si verifica quindi la massima coesione interna e divisione esterna tra i gruppi.

A questo punto si è effettuata un'analisi classica di tipo non gerarchico, utilizzando l'algoritmo dei K medoidi. Ponendo il numero di gruppi pari a due si è valutata la qualità della partizione attraverso lo strumento della Silhouette e si è confrontata poi con i risultati ottenuti ponendo il numero di gruppi pari a tre. L'analisi ha portato a confermare che, qualitativamente, il partizionamento migliore è quello ottenuto suddividendo le unità in due soli gruppi.

Si è osservata poi la presenza di quattro unità statistiche con valore di silhouette negativo ma prossimo allo zero: significa che la loro assegnazione all'uno o all'altro cluster risulta essere ambigua, cioè possono essere indistintamente assegnate all'uno o all'altro gruppo. Queste unità sono state poi prese in considerazione per essere confrontate con i risultati ottenuti con la classificazione di tipo sfocato.

Si è dunque passati all'analisi vera e propria del lavoro: è stata infatti effettuata una classificazione sfocata delle unità statistiche utilizzando un algoritmo di tipo non gerarchico, ponendo il numero di gruppi e l'indice di sfocatura pari a due.

Dal punto di vista della struttura dei cluster i due approcci, quello classico e quello sfumato, hanno fornito risultati molto simili. Nel primo caso si è ottenuto un partizionamento con due cluster, il primo composto da 32 unità statistiche e il secondo da 69 unità statistiche. Nel caso sfumato si è invece ottenuto un partizionamento composto da due cluster, rispettivamente di 40 e 61 unità statistiche.

Anche dal punto di vista qualitativo la classificazione finale è risultata pressoché identica, ottenendo un'ampiezza media di silhouette di 0,39 nel caso classico e di 0,38 nel caso sfumato.

A questo punto quello che viene naturale chiedersi è: che utilità ha avuto, per il presente lavoro, aver effettuato un'analisi di classificazione sfocata se i due approcci hanno fornito più o meno gli stessi risultati dal punto di vista della classificazione finale?

La risposta è semplice e risiede nel grado di appartenenza ai cluster attribuito dai metodi di classificazione sfocata a tutte le unità statistiche.

Con l'approccio classico si è ottenuto un partizionamento di cui si è potuto valutare soltanto la qualità utilizzando il grafico della silhouette.

Con tale approccio si è potuto infatti dedurre che se una determinata unità statistica è stata assegnata ad un certo cluster, vuol dire che possiede per la maggior parte caratteristiche corrispondenti a quel gruppo, senza sapere in che misura.

La classificazione sfocata, assegnando ad ogni unità il suo grado di appartenenza a ciascun gruppo, permette di superare questo inconveniente ed ottenere così un risultato che fornisce più informazioni circa la classificazione ottenuta.

Grazie a questa informazione si è potuto osservare che nel caso in esame le unità statistiche presentano mediamente un alto grado di sfocatura nei gruppi. Questo significa che hanno delle caratteristiche tali da non poter essere attribuite in modo preponderante ad uno dei due cluster. In altre parole, non avendo avuto conoscenze pregresse sulla struttura dei dati, si è potuto scoprire che la maggior parte di tali unità si trovano in una posizione intermedia fra i due gruppi.

Il metodo classico forzava le unità statistiche ad appartenere totalmente ad un cluster piuttosto che ad un altro, pur avendo caratteristiche intermedie ai due gruppi. Il risultato era quindi una classificazione non perfettamente corrispondente alla realtà analizzata.

Si è potuto infine osservare la presenza di 15 unità statistiche aventi caratteristiche tali da essere attribuite con sicurezza ad un cluster piuttosto che all'altro così da essere certi che, grazie all'informazione fornita dai gradi di appartenenza, tali unità appartengono in modo preponderante ad uno dei due cluster.

APPENDICE A

Tabella corrispondenze domande questionario <--> variabili

a1	**[Maschio (numero)] Genere del/i titolare/i**
a2	[Commento] Genere del/i titolare/i
a3	[Femmina (numero)] Genere del/i titolare/i
a4	[Commento] Genere del/i titolare/i
a5	[Titolare Maschio 1] Età del/i titolare/i
a6	[Commento] Età del/i titolare/i
a7	[Titolare Maschio 2] Età del/i titolare/i
a8	[Commento] Età del/i titolare/i
a9	[Titolare Maschio 3] Età del/i titolare/i
a10	[Commento] Età del/i titolare/i
a11	[Titolare Femmina 1] Età del/i titolare/i
a12	[Commento] Età del/i titolare/i
a13	[Titolare Femmina 2] Età del/i titolare/i
a14	[Commento] Età del/i titolare/i
a15	[Titolare Femmina 3] Età del/i titolare/i
a16	[Commento] Età del/i titolare/i
a17	[Titolare Maschio 1] Titolo di studio del/i titolare/i
a18	[Titolare Maschio 2] Titolo di studio del/i titolare/i
a19	[Titolare Maschio 3] Titolo di studio del/i titolare/i
a20	[Titolare Femmina 1] Titolo di studio del/i titolare/i
a21	[Titolare Femmina 2] Titolo di studio del/i titolare/i
a22	[Titolare Femmina 3] Titolo di studio del/i titolare/i
a23	[meno di 20.000 euro] Dimensione economica complessiva
a24	[tra 20.000 e 40.000 euro] Dimensione economica complessiva
a25	[tra 40.000 e 100.000 euro] Dimensione economica complessiva
a26	[tra 100.000 e 500.000 euro] Dimensione economica complessiva
a27	[oltre 500.000 euro] Dimensione economica complessiva
a28	[inferiore al 5%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?
a29	[dal 5% al 10%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?

a30	[dal 10% al 20%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?
a31	[dal 20% al 30%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?
a32	[dal 30% al 40%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?
a33	[oltre il 40%] Quale percentuale dei ricavi aziendali realizza attraverso l'attività di Agriturismo?
a34	[Vendita diretta (%)] Quali altre attività integrativa del reddito aziendale effettua?
a35	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a36	[Fattoria didattica (%)] Quali altre attività integrativa del reddito aziendale effettua?
a37	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a38	[Produzione prodotti biologici (%)] Quali altre attività integrativa del reddito aziendale effettua?
a39	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a40	[Produzione energia da fonti rinnovabili (%)] Quali altre attività integrativa del reddito aziendale effettua?
a41	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a42	[Produzione prodotti tipici/tradizionali (Dop/Igp, Doc/Igt) (%)] Quali altre attività integrativa del reddito aziendale effettua?
a43	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a44	[Nessuna delle precedenti] Quali altre attività integrativa del reddito aziendale effettua?
a45	[Commento] Quali altre attività integrativa del reddito aziendale effettua?
a46	[Altro] Quali altre attività integrativa del reddito aziendale effettua?
a47	[Altro commento] Quali altre attività integrativa del reddito aziendale effettua?
a48	Numero complessivo di addetti (familiari/dipendenti) impiegati nell'azienda agricola
a49	[Addetto 1] Numero medio di giornate lavorate per addetto all'anno
a50	[Addetto 2] Numero medio di giornate lavorate per addetto all'anno
a51	[Addetto 3] Numero medio di giornate lavorate per addetto all'anno
a52	[Addetto 4] Numero medio di giornate lavorate per addetto all'anno
a53	[Addetto 5] Numero medio di giornate lavorate per addetto all'anno
a54	[Addetto 6] Numero medio di giornate lavorate per addetto all'anno
a55	[Addetto 7] Numero medio di giornate lavorate per addetto all'anno
a56	[Addetto 8] Numero medio di giornate lavorate per addetto all'anno
a57	[Addetto 9] Numero medio di giornate lavorate per addetto all'anno
a58	[Addetto 10] Numero medio di giornate lavorate per addetto all'anno
a59	Numero complessivo di addetti (familiari/dipendenti) impiegati in attività agrituristica
a60	[Addetto 1] Numero medio di giornate lavorate per addetto nell'agriturismo

a61	[Addetto 2] Numero medio di giornate lavorate per addetto nell'agriturismo
a62	[Addetto 3] Numero medio di giornate lavorate per addetto nell'agriturismo
a63	[Addetto 4] Numero medio di giornate lavorate per addetto nell'agriturismo
a64	[Addetto 5] Numero medio di giornate lavorate per addetto nell'agriturismo
a65	[Addetto 6] Numero medio di giornate lavorate per addetto nell'agriturismo
a66	[Addetto 7] Numero medio di giornate lavorate per addetto nell'agriturismo
a67	[Addetto 8] Numero medio di giornate lavorate per addetto nell'agriturismo
a68	[Addetto 9] Numero medio di giornate lavorate per addetto nell'agriturismo
a69	[Addetto 10] Numero medio di giornate lavorate per addetto nell'agriturismo
a70	La struttura offre servizio di alloggio?
a71	Quante persone vengono mediamente ospitate in un anno?
a72	Stima delle presenze registrate in un anno
a73	[In crescita] Rispetto agli ultimi tre anni, i clienti che alloggiano sono:
a74	[Stabili] Rispetto agli ultimi tre anni, i clienti che alloggiano sono:
a75	[In diminuzione] Rispetto agli ultimi tre anni, i clienti che alloggiano sono:
a76	La struttura offre servizio di ristorazione?
a77	Generalmente, quanti clienti/coperti registrate all'anno?
a78	In media, qual è la spesa per persona?
a79	[In crescita] Rispetto agli ultimi tre anni i clienti/coperti sono:
a80	[Stabili] Rispetto agli ultimi tre anni i clienti/coperti sono:
a81	[In diminuzione] Rispetto agli ultimi tre anni i clienti/coperti sono:
x1	Turista italiano [SQ001]
x2	Commento [SQ001comment]
x3	Turista straniero [SQ002]
x4	Commento [SQ002comment]
x5	Cliente locale [SQ003]
x6	Commento [SQ003comment]
x7	Passaparola [SQ001]
x8	Depliant e volantini [SQ002]
x9	Aziende di promozione turistica APT [SQ003]
x10	Agenzie di viaggio [SQ005]
x11	Manifestazioni locali [SQ006]
x12	Sito Internet dell'Agriturismo [SQ007]
x13	Altri siti Inernet [SQ008]
x14	Guide turistiche [SQ009]
x15	Altro Quali .. [other]
x16	La sua azienda propone eventi, attività, serate a tema

x17	Cicloturismo .. [SQ001]
x18	Escursionismo/visite naturalistiche [SQ002]
x19	Attività sportive [SQ003]
x20	Turismo fluviale [SQ004]
x21	Laboratori didattici [SQ005]
x22	Mostre/esposizioni d'arte, fotografia [SQ006]
x23	Corsi (cucina, meditazione) [SQ007]
x24	Altro [other]
x25	Importantissimi [SQ001]
x26	Molto importanti [SQ002]
x27	Abbastanza importanti [SQ003]
x28	Poco importanti [SQ004]
x29	Per niente importanti [SQ005]
x30	Comunicazione [SQ001]
x31	Promozione [SQ002]
x32	Formazione per i dipendenti [SQ003]
x33	Ristrutturazione [SQ004]
x34	Incentivi per impianti a energia rinnovabile [SQ005]
x35	Altro [other]
x36	Adottate pratiche per la riduzione dei rifiuti?
x37	Scelta di imballaggi riciclabili [SQ001]
x38	Riutilizzo di materiali e contenitori [SQ002]
x39	Riciclaggio di materiali e contenitori [SQ003]
x40	Compostaggio [SQ004]
x41	Altro [other]
x42	Presso la sua azienda praticate la raccolta differenziata dei rifiuti?
x43	La sua azienda è dotata di strutture e facilitazioni per l'accesso e l'ospitalità di persone con disabilità?
x44	Adottate misure per la gestione e la diminuzione del consumo elettrico?
x45	Utilizzo di lampadine a basso consumo [SQ001]
x46	Spegnimento automatico luci [SQ002]
x47	Uso elettrodomestici a classe di efficienza energetica A [SQ003]
x48	Altro [other]
x49	Adottate misureper la gestione e la diminuzione del consumo idrico?
x50	Utilizzo di filtri ad aria nei miscelatori [SQ001]
x51	Controlli costanti perdite [SQ002]
x52	Uso rubinetti a leva unica [SQ003]

x53	Timer per docce cucine e/o esterni [SQ004]
x54	Altro [other]
x55	Adottate misure per la gestione e la diminuzione del consumo termico?
x56	Isolamento termico edificio [SQ001]
x57	Impianti di condizionamento a classe di efficienza energetica minima B[SQ002]
x58	Informazioni per gli ospiti sulla propria politica ambientale/carta dei servizi [SQ003]
x59	Suggerimenti per evitare sprechi [SQ004]
x60	Altro [other]
x61	Presso la sua azienda si utilizzano in modo responsabile disinfettanti, detergenti e detersivi?
x62	Cambio biancheria limitato in accordo con le esigenze dei clienti [SQ001]
x63	Uso detergenti detergenti/disinfettanti/detersivi biodegradabili [SQ002]
x64	Uso detergenti detergenti/disinfettanti/detersivi con certificazioni ambientali [SQ003]
x65	Altro [other]
x66	Presso la sa struttura adottate misure per la gestione ambientale?
x67	Controlli periodici dei consumi [SQ001]
x68	Manutenzione delle apparecchiature [SQ002]
x69	Informazione per gli ospiti sulla propria politica ambientale/carta dei servizi [SQ003]
x70	Altro [other]
x71	Vi è l'intenzione per il futuro di adottare misure di gestione ambientale?
x72	È possibile accedere alla sua azienda attraverso i mezzi del servizio pubblico?
x73	Vengono fornite adeguate informazioni agli ospiti su come raggiungere la struttura e altre destinazioni con i mezzi pubblici?
x74	Presso la sua azienda sono presenti impianti per la produzione di energia rinnovabile?
x75	Energia eolica [SQ001]
x76	Biomassa [SQ002]
x77	Biogas [SQ003]
x78	Idroenergia [SQ004]
x79	Energia solare [SQ005]
x80	Altro [other]
x81	La sua azienda ha beneficiato di una o più misure di sostegno allo sviluppo rurale nell'ultimo triennio?
x82	Misura 112: insediamento di giovani agricoltori [SQ001]
x83	Misura 121: ammodernamento aziende agricole [SQ002]
x84	Misura 132: partecipazione sistemi qualità alimentare [SQ003]
x85	Misura 213: Indennità Natura 2000 [SQ004]
x86	Misura 311: diversificazione attività non agricole [SQ005]
x87	Misura 313: incentivi attività turistica [SQ006]

x88	Misura 216: sostegno investimenti non produttivi [SQ007]
x89	Nessuna delle precedenti [SQ008]
x90	Altro [other]
x91	La sua struttura è in possesso di certificazioni ambientali
x92	EMAS [SQ001]
x93	ECOLABEL [SQ002]
x94	ISO 14001 [SQ003]
x95	Altro [other]
x96sì	Vi è l'intenzione per il futuro di adottare certificazioni ambientali? SI
x96no	Vi è l'intenzione per il futuro di adottare certificazioni ambientali? NO
x96no nso	Vi è l'intenzione per il futuro di adottare certificazioni ambientali? NON SO
x97	EMAS [SQ001]
x98	ECOLABEL [SQ002]
x99	ISO 14001 [SQ003]
x100	Non so ancora [SQ004]
x101	Altro [other]

APPENDICE B

Tabelle di contingenza

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x49	0	27	12	39
	1	39	23	62
Totale		66	35	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x72	0	27	39	66
	1	12	23	35
Totale		39	62	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x49	0	18	21	39
	1	24	38	62
Totale		42	59	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x72	0	25	41	66
	1	17	18	35
Totale		42	59	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x49	0	24	15	39
	1	41	21	62
Totale		65	36	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x72	0	40	26	66
	1	25	10	35
Totale		65	36	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x49	0	37	2	39
	1	59	3	62
Totale		96	5	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x72	0	62	4	66
	1	34	1	35
Totale		96	5	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x49	0	29	10	39
	1	48	14	62
Totale		77	24	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x72	0	48	18	66
	1	29	6	35
Totale		77	24	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x49	0	18	21	39
	1	22	40	62
Totale		40	61	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x72	0	31	35	66
	1	9	26	35
Totale		40	61	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x74	0	18	24	42
	1	21	38	59
Totale		39	62	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x81	0	24	41	65
	1	15	21	36
Totale		39	62	101

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x74	0	25	17	42
	1	41	18	59
Totale		66	35	101

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x81	0	40	25	65
	1	26	10	36
Totale		66	35	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x74	0	28	14	42
	1	37	22	59
Totale		65	36	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x81	0	28	37	65
	1	14	22	36
Totale		42	59	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x74	0	41	1	42
	1	55	4	59
Totale		96	5	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x81	0	62	3	65
	1	34	2	36
Totale		96	5	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x74	0	37	5	42
	1	40	19	59
Totale		77	24	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x81	0	51	14	65
	1	26	10	36
Totale		77	24	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x74	0	14	28	42
	1	26	33	59
Totale		40	61	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x81	0	23	42	65
	1	17	19	36
Totale		40	61	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x91	0	37	59	96
	1	2	3	5
Totale		39	62	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x96sì	0	29	48	77
	1	10	14	24
Totale		39	62	101

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x91	0	62	34	96
	1	4	1	5
Totale		66	35	101

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x96sì	0	48	29	77
	1	18	6	24
Totale		66	35	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x91	0	41	55	96
	1	1	4	5
Totale		42	59	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x96sì	0	37	40	77
	1	5	19	24
Totale		42	59	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x91	0	62	34	96
	1	3	2	5
Totale		65	36	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x96sì	0	51	26	77
	1	14	10	24
Totale		65	36	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x91	0	75	21	96
	1	2	3	5
Totale		77	24	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x96sì	0	75	2	77
	1	21	3	24
Totale		96	5	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x91	0	37	59	96
	1	3	2	5
Totale		40	61	101

Tabella di contingenza

Conteggio

		x96nonso		
		0	1	Totale
x96sì	0	16	61	77
	1	24	0	24
Totale		40	61	101

Tabella di contingenza

Conteggio

		x49		
		0	1	Totale
x96nonso	0	18	22	40
	1	21	40	61
Totale		39	62	101

Tabella di contingenza

Conteggio

		x72		
		0	1	Totale
x96nonso	0	31	9	40
	1	35	26	61
Totale		66	35	101

Tabella di contingenza

Conteggio

		x74		
		0	1	Totale
x96nonso	0	14	26	40
	1	28	33	61
Totale		42	59	101

Tabella di contingenza

Conteggio

		x81		
		0	1	Totale
x96nonso	0	23	17	40
	1	42	19	61
Totale		65	36	101

Tabella di contingenza

Conteggio

		x91		
		0	1	Totale
x96nonso	0	37	3	40
	1	59	2	61
Totale		96	5	101

Tabella di contingenza

Conteggio

		x96sì		
		0	1	Totale
x96nonso	0	16	24	40
	1	61	0	61
Totale		77	24	101

APPENDICE C

Gradi di appartenenza ai cluster

Unità statistica	Fanny - Cluster crisp di assegnazione	Appartenenza al 1° cluster	Appartenenza al 2° cluster
[1,]	1	0.5721085	0.4278915
[2,]	1	0.5780508	0.4219492
[3,]	2	0.4115057	0.5884943
[4,]	2	0.4701927	0.5298073
[5,]	2	0.4556064	0.5443936
[6,]	1	0.6559686	0.3440314
[7,]	1	0.5721085	0.4278915
[8,]	1	0.6346301	0.3653699
[9,]	2	0.4011171	0.5988829
[10,]	2	0.3470705	0.6529295
[11,]	1	0.6346301	0.3653699
[12,]	1	0.5211311	0.4788689
[13,]	2	0.4115057	0.5884943
[14,]	2	0.3597389	0.6402611
[15,]	1	0.6346301	0.3653699
[16,]	2	0.3597389	0.6402611
[17,]	1	0.6346301	0.3653699
[18,]	2	0.3597389	0.6402611
[19,]	1	0.5269588	0.4730412
[20,]	2	0.3237037	0.6762963
[21,]	2	0.4336794	0.5663206
[22,]	2	0.4158290	0.5841710
[23,]	2	0.3237037	0.6762963
[24,]	2	0.4850267	0.5149733
[25,]	1	0.5721085	0.4278915
[26,]	2	0.3597389	0.6402611
[27,]	1	0.6431584	0.3568416
[28,]	2	0.4256574	0.5743426
[29,]	2	0.4115057	0.5884943
[30,]	2	0.4337812	0.5662188
[31,]	2	0.4011171	0.5988829
[32,]	1	0.6147515	0.3852485
[33,]	1	0.6503975	0.3496025
[34,]	2	0.3787652	0.6212348
[35,]	2	0.3470705	0.6529295
[36,]	2	0.4645819	0.5354181
[37,]	2	0.4158290	0.5841710
[38,]	2	0.4011171	0.5988829
[39,]	1	0.5913620	0.4086380
[40,]	1	0.5093808	0.4906192
[41,]	2	0.3597389	0.6402611
[42,]	1	0.5780509	0.4219491
[43,]	2	0.3597389	0.6402611
[44,]	1	0.6147515	0.3852485
[45,]	2	0.4115057	0.5884943
[46,]	2	0.3960600	0.6039400
[47,]	2	0.4575543	0.5424457

[48,]	1	0.6346301	0.3653699
[49,]	2	0.4314918	0.5685082
[50,]	2	0.3960600	0.6039400
[51,]	2	0.4714343	0.5285657
[52,]	1	0.6431584	0.3568416
[53,]	2	0.4337812	0.5662188
[54,]	1	0.6044150	0.3955850
[55,]	2	0.3470705	0.6529295
[56,]	1	0.6431584	0.3568416
[57,]	2	0.3470705	0.6529295
[58,]	1	0.5721085	0.4278915
[59,]	2	0.4850267	0.5149733
[60,]	1	0.6030631	0.3969369
[61,]	1	0.5397767	0.4602233
[62,]	1	0.6503975	0.3496025
[63,]	2	0.4645819	0.5354181
[64,]	1	0.5211311	0.4788689
[65,]	1	0.5836012	0.4163988
[66,]	1	0.6030631	0.3969369
[67,]	1	0.6503975	0.3496025
[68,]	1	0.6559687	0.3440313
[69,]	2	0.4115057	0.5884943
[70,]	2	0.4314918	0.5685082
[71,]	2	0.4850267	0.5149733
[72,]	1	0.5527433	0.4472567
[73,]	2	0.4714343	0.5285657
[74,]	2	0.3787651	0.6212349
[75,]	2	0.3237037	0.6762963
[76,]	2	0.4645819	0.5354181
[77,]	1	0.5836012	0.4163988
[78,]	2	0.3470705	0.6529295
[79,]	2	0.4011170	0.5988830
[80,]	2	0.3470705	0.6529295
[81,]	1	0.5325863	0.4674137
[82,]	2	0.4556064	0.5443936
[83,]	2	0.4850267	0.5149733
[84,]	2	0.3787651	0.6212349
[85,]	2	0.3237036	0.6762964
[86,]	2	0.4011170	0.5988830
[87,]	1	0.5093808	0.4906192
[88,]	1	0.6019388	0.3980612
[89,]	2	0.4645819	0.5354181
[90,]	1	0.5821775	0.4178225
[91,]	2	0.4850267	0.5149733
[92,]	1	0.5986400	0.4013600
[93,]	2	0.4011170	0.5988830
[94,]	2	0.3237036	0.6762964
[95,]	2	0.4336794	0.5663206
[96,]	1	0.6147516	0.3852484
[97,]	1	0.5815297	0.4184703
[98,]	2	0.3237036	0.6762964
[99,]	2	0.3597389	0.6402611
[100,]	2	0.4011170	0.5988830
[101,]	2	0.4850267	0.5149733

Bibliografia e sitografia

Testi

Angelo Bruscino (2011), *Il turismo sostenibile*, libreriauniversitaria.it, pp.23-26

ARPAV Staff Emas (2006), "*A proposito di….Ecolabel Europeo*"

ARPAV (2005), "*Annali italiani del Turismo internazionale*", volume n.1, pp. 31-52

Bezdek J. C. (1981), "*Pattern Recognition with Fuzzy Objective Function Algorithm*", Plenum Press, New York

Bezdek J. C. (1991), "*Numerical comparison of the RFCM and AP algorithms for clustering relational data*", in "*Pattern Recognition*", pp, 786-791

Bezdek J. C. e Hathaway R. J. (1988), "*Recent convergence results for the fuzzy c-means clustering algorithms*", pp. 237-247

Cerbara L. (1997), *Hierachical fuzzy clustering: an example of spatio-temporal analysis*", in "*Book of Short Papers: Classification and Data Analysis*", Università degli Studi G. d'Annunzio di Pescara

Dunn J. C. (1974), "*Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems*", in "*J. Cybernetic*", 4, pp. 1-15

Dunn J. C. (1974), "Well separated clusters and optimal fuzzy partitions", in "*J. Cybernetic*", 4, pp. 95-104

E. H. Ruspini (1970), "*Numerical methods for fuzzy clustering*", Information Sciences, 2

Giles R. (1988), "*The concept of grade of membership*", in "*Fuzzy Sets and Systems*", 25, pp. 297-323

Gordon A.D e Vichi M. (1997), "*Partitions of Partitions*"

Hartigan J. (1975), "*Clustering algorithms*", Wiley, New York

Iacovacci G. (1995), "*Sull'utilizzo del metodo delle c-medie sfocato per la classficazione dei comuni italiani a seconda del grado di urbanità e ruralità*", in "*Statistica Applicata*", 7, Rocco Curto Editore

Iacovacci G. (1997), "*A semi-fuzzy c-means algorithm for soft clustering*" in "*Book of Short Papers: Classification and Data Analysis*", Università degli Studi G. d'Annunzio di Pescara

Jari Oksanen (26/01/2014), "*Cluster Analysis: Tutorial Whit R*"

Kamel M. S. e Selim S. Z. (1991), "*A thresholded fuzzy c-means algorithm for semi-fuzzy clustering*", in Pattern Recognition, 24, pp. 825-883

Kaufmann A. (1975), "*Introduction to the theory of fuzzy subsets*", 1, Academic Press, New York

MacQueen J. (1967), "*Some methods for classification and analysis of multivariate observation*", pp. 281-297

Mioli M. A. (1994), "*Confronto fra partizioni sfocate nell'analisi di dati territoriali*", in "*Atti della XXXVII Riunione Scientifica (SIS) San Remo 6-8 aprile, 2*"

Ricolfi L. (1992), "*Helga, nuovi principi di analisi dei gruppi*", FrancoAngeli, Milano

Rizzi A. (1978), "*Analisi dei gruppi*", La Goliardica

Rizzi A. (1985), "*Analisi dei dati. Applicazioni dell'informatica alla statistica*", Studi Superiori NIS

Selim S. Z. e Ismail M. A. (1984), "*Soft clustering of multidimensional data: a semi-fuzzy approach*", in "*Pattern Recognition*", 10, pp. 294-301

Zadeh L. A. (1965), "*Fuzzy sets*", pp. 338-353

Zaddeh L. A. (1977), "*Fuzzy set and their application to pattern classification and clustering*", in "*Classification and Clustering*", J. Van Ryzin, Accademic Press, New York, pp. 251-299

Zadeh L. A. (1965), "*Fuzzy sets*", pp. 338-353

Zani S. (1989), "*Classificazioni 'sfocate' di unità territoriali: un'applicazione alle regioni italiane*", in Atti delle giornate di studio "*Analisi Statistica di Dati Territoriali*", Cacucci, Bari, pp. 495-506

Zani S. (1988), "*Un metodo di classificazione 'sfocata', in G. Diana, C. Provasi e R. Vevaldi, "Metodi statistici per la tecnologia e l'analisi dei dati multidimensionali*", Università degli Studi di padova, pp. 281-288

Zani S. (1993), "*Classificazioni di unità territoriali e spaziali", in S. Zani, Metodi statistici per le analisi territoriali*", FrancoAngeli, Milano, pp. 93-121

Zani S. e Cerioli A. (2007), "*Analisi dei dati e data mining per le decisioni aziendali*", Giuffrè Editore

Zimmermann (1976), "*Description and optimization of fuzzy systems*", pp. 209-215

Siti, documenti e articoli internet consultati

"*Agenda per un turismo europeo sostenibile e competitivo*", 15/02/2008, http://europa.eu/legislation_summaries/enterprise/industry/l10132_it.htm

Alessandro La Rocca, Istat - Direzione centrale per le indagini su condizioni e qualità della vita, "*Fuzzy Clustering: la logica, i metodi*" www3.istat.it/dati/pubbsci/contributi/Contributi/contr_2006/2006_5.pdf

Alessandro Maurilli, "Agriturismo: *La certificazione ambientale fa la differenza*" 22/09/2012, www.novagricoltura.com/agriturismo-la-certificazione-ambientale-fa-la-differenza/0,1254,105_ART_6695,00.html

Arpav (luglio 2010), "*ISO 14001 e regolamento EMAS*", www.arpa.veneto.it/servizi-ambientali/certificazioni-ambientali/gli-strumenti-per-la-certificazione/iso-14001-e-regolamento-emas

Arpav (21/01/2014), "*I numeri Ecolabel*", www.arpa.veneto.it/servizi-ambientali/certificazioni-ambientali/gli-strumenti-per-la-certificazione/ecolabel/i-numeri-ecolabel

Beatrice Lazzarini, Dipartimento dell'ingegneria dell'informazione di Pisa, "*Introduzione agli Insiemi Fuzzy e alla Logica Fuzzy*", http://info.iet.unipi.it/~lazzerini/icse/DispensaFuzzySpec.pdf

"Carta europea del turismo durevole", www.tures.it/uploaddocumenti/2012/2_42_IHCMGDIOFFAS.pdf

"Città sostenibili: conferenza europea a Ginevra", 2013 www.lapiazzadelleidee.comune.mirano.ve.it/?p=85

"Cluster Analysis in R: Hierarchical Clustering" (30 giugno 2010), http://statisticaconr.blogspot.it/2010/06/cluster-analysis-in-r-1-hierarchical.html

"*Conferenza Rio+20: una sfida importante*, 26/09/2013,
www.minambiente.it/pagina/conferenza-rio20-una-sfida-importante#rio

Davide Luciani (2005-2006), tesi di laurea, "*Logica fuzzy: storia e sue applicazioni*",
www.apav.it/mat/logica_fuzzy/logica_fuzzy_Davide_Luciani_tesi_completa.pdf

Elisa Poggiali, Gruppo di Coordinamento Agenda 21 "Terre di Siena", "*Turismo Sostenibile: principi di base e possibili strategie per le Pubbliche Amministrazioni*",
www.agenda21.provincia.siena.it/upload_settori/turismo%20sost_azioni%20PA.pdf

"Fuzzy Clustering",
http://homes.di.unimi.it/~valenti/SlideCorsi/Bioinformatica05/Fuzzy-Clustering-lecture-Babuska.pdf

"Fuzzy Clustering",
www.irpps.cnr.it/it/system/files/wp2_98_0.pdf

Gruppo per la Sostenibilità del Turismo, "*Azione per un turismo europeo più sostenibile*", 2007,
www.sinanet.isprambiente.it/gelso/files/azione-per-un-turismo-europeo.pdf

International Conference of Environment Ministers on Biodiversity and Tourism 6-8 March, 1997, Berlin, "*The Berlin Declaration On Biological Diversity And Sustainable Tourism*",
www.gdrc.org/uem/eco-tour/berlin.html

ISPRA, "*Principali promotori del turismo sostenibile*",
www.sinanet.isprambiente.it/ gelso/tematiche/buone-pratiche-per-il-turismo/principali-promotori-del-turismo-sostenibile

James Bezdek, Robert Ehrlich,William Full, "*FCM: The fuzzy c-Means Clsutering Algorithm*", 1984,
http://web-ext.u-aizu.ac.jp/course/bmclass/documents/FCM%20-%20The%20Fuzzy%20c-Means%20Clustering%20Algorithm.pdf

"L'agriturismo in Italia",
www.agriturismo-on-line.com/agriturismo_in_italia.cfm

"L'agriturismo è donna", 11 ottobre 2011,
http://blog.agriturismo.it/?p=752

La Repubblica, *"L'inarrestabile crescita di fonti rinnovabili e gas"*,
www.repubblica.it/economia/finanza-personale/2013/11/17/news/milano_-_hanno_avuto_una_battuta_d_arresto_in_particolare_in_europa_dove_alcuni_paesi_cone_l_italia_e_la_spagna_hanno_camb-71223068/

L'informatore agrario, 32/2009, "*Le certificazioni ambientali qualificano l'attività agrituristica*",
www.informatoreagrario.it

"Le tappe cronologiche dello sviluppo sostenibile",
www.padovanet.it/dettaglio.jsp?id=7965#.VHcIkDGG_j-

Mario Adua (1 aprile 2014), "*Originalità dell'azienda agricola al femminile: i prodotti di qualità e gli agriturismi*",
www.istat.it/it/files/2014/03/Relazione-_Adua.pdf

Ministero dell'ambiente e della tutela del territorio e del mare, "*Vertice mondiale sullo Sviluppo Sostenibile 2002*";
www.minambiente.it/pagina/vertice-mondiale-sullo-sviluppo-sostenibile-2002

Ministero dell'ambiente e della tutela del territorio e del mare, "*Conferenza Rio+20: una sfida importante*";
www.minambiente.it/pagina/conferenza-rio20-una-sfida-importante#rio

"Package Cluster" (04/09/2014),
http://cran.r-project.org/web/packages/cluster/cluster.pdf

Pap Team & University of Alexandria, "P*riority Actions Programme Regional Activity Centre Split*", June 1999,
www.pap-thecoastcentre.org/pdfs/CCA%20for%20Tourism%20 Development.pdf

"Procedure di Cluster Analysis con S-Plus: nuove tecniche per la costruzione di gruppi",
www.rescoop.com/seminarioCluster/semCluster.htm

Provincia di Rimini, "*Il manifesto della mobilità sostenibile*", 2001, http://www.a21italy.it/medias/C2F34F2DAE037F58.pdf

Provincia di Rimini (2000), "*Guida alle buone pratiche del turismo sostenibile*", www.a21italy.it/medias/9A52DFBBB778E3F0.pdf

Provincia di Varese, "*Turismo sostenibile: concetti, azioni, esperienze*", www.provincia.va.it/ProxyVFS.axd/null/r26847/Turismo-sostenibile-concetti-azioni-esperienze-pdf?ext=.pdf

R. Giacomozzi, "Gestione delle problematiche ambientali all'interno dell'impresa", http://archivio.ambiente.it/impresa/monografie/problematiche/gestione.htm

Regione del Veneto, Università degli studi di Padova, "*Il settore turistico e la sostenibilità: le certificazioni ambientali a carattere volontario*" Luglio 2012, www.regione.veneto.it/c/document_library/get_file?uuid=dd578ac8-e986-4328-9236-875ca433c9cb&groupId=10813

S. Terzi, "*La cluster analysis*", http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/185_903.pdf

Serenella Stasi (2011), "*La logica fuzzy nella ricerca sociale con particolare riguardo alle scale di atteggiamenti*", http://padis.uniroma1.it/bitstream/10805/1246/2/Tesi%204.pdf;jsessionid=0706D3B7E4098B82B66690D42CBA5D3D

T. Soni Madhulatha (2012), "*An overview on clustering methods*", IOSR Journal of Engineering, http://arxiv.org/ftp/arxiv/papers/1205/1205.1117.pdf

"*Vertice mondiale sullo sviluppo sostenibile 2002*", 26/09/2013, www.minambiente.it/pagina/vertice-mondiale-sullo-sviluppo-sostenibile-2002

"Vino biodinamico…Ma cos'è esattamente la "biodinamica"?, www.wineblog.it/index.php/vino-biodinamico%E2%80%A6%E2%80%A6-ma-che-cos%E2%80%99-e-esattamente-%E2%80%9Cbiodinamica%E2%80%9D/

www.cittasostenibili.minori.it/guida/agenda21.htm

www.eea.europa.eu/it/about-us/who

www.wwfnature.it/informati/dossier-wwf/item/337-carte-e-codici-di-condotta.html

www.iso.org/iso/home/standards/certification/iso-survey.htm?certificate=ISO%2014001&countrycode=IT#countrypick

http://www2.stat.unibo.it/montanari/Didattica/Multivariate/CA_lab.pdf

http://www.statmethods.net/advstats/cluster.html