UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA GIORGIO FUÀ

---

Corso di Laurea Magistrale in Data Science per l'Economia e le Imprese

# Clustering con Media Geometrica penalizzata: dati ISPRA sul consumo del suolo

# Clustering via Penalized Geometric Mean: ISPRA land use data

Relatore: Chiar.ma
Prof.ssa Maria Cristina Recchioni

Tesi di Laurea di:
Anastasia Zinovyeva

A.A. 2022/ 2023

# Indice

# Elenco delle tabelle

# Elenco delle figure

# Capitolo 1 Introduzione

## 1.1 Descrizione del problema del suolo

Il cambiamento climatico rappresenta una delle sfide più urgenti del nostro tempo. Quando ci confrontiamo con questa problematica, è essenziale tenere sempre presente un obiettivo fondamentale: preservare il nostro pianeta. Tutte le forme di vita, dagli organismi più piccoli come i vermi fino all'uomo, dipendono dai primi metri di suolo, dove avvengono i processi vitali più importanti. Sebbene i microbi che lo abitano siano di vitale importanza, il danno vero e proprio si manifesta quando questo strato di suolo, che sostiene l'intera biosfera, viene compromesso (l'Agenzia delle Nazioni Unite per l'Agricoltura, 2019) .

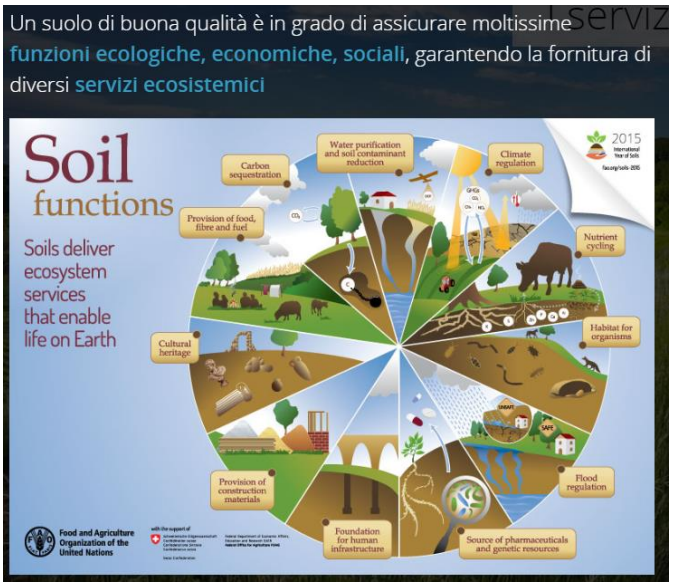


Figura 1.1 Le funzioni del suolo
Fonte: https://www.fao.org/

Tra gli altri problemi ambientali, come l'inquinamento dell'aria, dell'acqua e acustico, la conservazione del suolo costituisce un elemento cruciale. Perché, per quanto riguarda gli altri aspetti, il pianeta è in grado di autocorreggersi. Anche

l'inquinamento dell'aria, sebbene sia di grande rilevanza, non può sostituire l'importanza di fermare la perdita del suolo (Shefali, Pankaj , & Kanchan, 2020).

La perdita di suolo sta procedendo molto velocemente. Il fattore critico riguarda l'aumento dell'urbanizzazione nelle valli, nelle pianure e nelle zone costiere, il quale ha determinato una significativa diminuzione della copertura forestale, un consumo insostenibile del suolo e una compromissione delle reti ecologiche (Barbera, Gallerano, Nicoletti, & Raimond, 2020). Per consumo di suolo si intende l'occupazione di una porzione originariamente agricola, naturale o seminaturale, con una copertura artificiale.

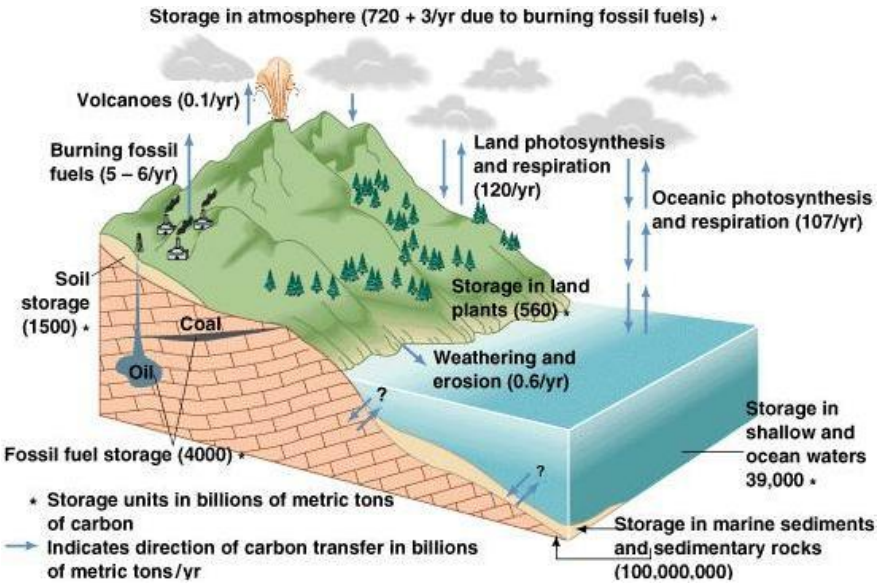


Figura 1.2 Il diagramma del ciclo del carbonio.

Fonte: https://www.fao.org/soils-portal/data-hub/soil-properties/

Il diagramma del ciclo del carbonio illustra lo scambio di carbonio tra la biosfera, pedosfera, geosfera, idrosfera e atmosfera della Terra. Questo processo è fondamentale per il riciclaggio e il riutilizzo dell'elemento più abbondante del pianeta. I microrganismi presenti nel suolo svolgono un ruolo significativo nel riciclaggio dei nutrienti, consumando il materiale organico e producendo CO2,

H2O ed humus. L'humus costituisce una riserva di carbonio a lenta decomposizione che può durare diverse centinaia o addirittura migliaia di anni. Il tempo di permanenza del carbonio nella maggior parte dei suoli è di circa 20-30 anni. I microrganismi del suolo sono altamente sensibili alla quantità di carbonio organico, alla temperatura e alla disponibilità di acqua, con una maggiore velocità di respirazione in presenza di concentrazioni e temperature più elevate (Bot & Benites, 2005).

Il suolo rappresenta l'habitat di milioni di specie viventi ed è la base fondamentale per la crescita delle piante, l'alimentazione degli animali e degli esseri umani. È pertanto di vitale importanza prestare particolare attenzione alla tutela del suolo, senza trascurare gli altri aspetti dell'inquinamento. Il lockdown imposto dalla pandemia di Coronavirus ha dimostrato che aria e acqua possono depurarsi rapidamente (Shefali, Pankaj , & Kanchan, 2020), mentre il suolo necessita di un lungo periodo di recupero, addirittura duecento anni (Dellasala & Goldstein, 2018). Non dobbiamo sottovalutare l'impatto della degradazione del suolo e pensare che la sola cura dell'aria pulita possa garantire la nostra qualità di vita. La perdita di suolo rappresenta una minaccia concreta per la sopravvivenza del pianeta e della nostra specie. È di cruciale importanza intervenire prima che diventi irrimediabile (l'Agenzia delle Nazioni Unite per l'Agricoltura, 2019).

# Capitolo 2 Case Study: Province Italiane

## 2.1 Fonte del dataset

La seguente analisi si basa su dati provenienti dalla banca dati online dell'Agenzia Europea dell'Ambiente (Fonte: https://groupware.sinanet.isprambiente.it/uso-copertura-e-consumo-di-suolo), prodotta da L'ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale) è un Ente Pubblico di Ricerca (EPR) dotato di personalità giuridica di diritto pubblico e dotato di autonomia statutaria e regolamentare nonché tecnica, scientifica, organizzativa, finanziaria, gestionale, amministrativa, patrimoniale e contabile; l'Istituto è sottoposto alla vigilanza del Ministro della Transizione Ecologica che impartisce direttive annuali o pluriennali, declinate dagli organi dell'ente in priorità strategiche e attività da perseguire.

I dati, le informazioni, i pareri e le valutazioni fornite da ISPRA sono il riferimento per l'assunzione di decisioni pubbliche in materia ambientale, incluse normative e atti amministrativi di autorizzazione e di controllo, svolgendo un ruolo essenziale e con un impatto diretto sull'operato di innumerevoli aziende e organizzazioni.


Figura 2.1 ISPRA Logo

Fonte: https://www.isprambiente.gov.it/it/istituto

ISPRA raccoglie dati ed evidenze sull'ambiente italiano attraverso il monitoraggio e la valutazione scientificamente basati, integrati con l'attività di ricerca dell'Istituto. Questo è fondamentale per verificare il raggiungimento degli obiettivi stabiliti a livello internazionale. ISPRA utilizza diverse fonti di dati, tra cui **il monitoraggio in loco, l'analisi di campioni raccolti dalle Agenzie** che compongono il SNPA (Sistema Nazionale Protezione Ambiente), i **dati raccolti da satelliti e la collaborazione con istituzioni europee**. ISPRA dispone di serie storiche di dati che talvolta risalgono a oltre cento anni fa. Inoltre, i cittadini collaborano volontariamente con l'Istituto per la raccolta di dati di monitoraggio, come nel caso del tracciamento dei viaggi degli uccelli migratori e del monitoraggio delle specie non indigene nel Mediterraneo (Fonte: https://www.isprambiente.gov.it/it/sistema-nazionale-protezione-ambiente).

Il Programma Copernicus è un progetto di osservazione della Terra dell'Unione Europea, che utilizza una costellazione di satelliti per monitorare il nostro pianeta a beneficio di tutti i cittadini europei. I dati forniti ogni giorno sono utilizzabili sia dai tecnici che dai cittadini e sono utili per comprendere "lo stato di salute" del nostro pianeta su una vasta gamma di temi, tra cui l'acqua, il suolo, l'atmosfera, le emergenze, la sicurezza e il cambiamento climatico.

Per garantire l'accuratezza dei dati, Copernicus utilizza alcune reti di controllo come le stazioni meteorologiche terrestri, le boe oceaniche e le reti di monitoraggio della qualità dell'aria. Grazie a Copernicus, possiamo avere una visione più completa e dettagliata del nostro pianeta, contribuendo così a proteggerlo e preservarlo per le future generazioni.


Figura 2.2 Programma Copernicus
Fonte: https://www.copernicus.eu/it/

ISPRA ha anche il ruolo fondamentale di **definire e armonizzare i metodi per le attività di monitoraggio ambientale**, sia all'interno del SNPA (Sistema Nazionale Protezione Ambiente) che in contesti internazionali più ampi. Questo è essenziale poiché senza metodi solidi e uniformi non è possibile aggregare e confrontare i dati in modo significativo (L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), 2021).

Il dataset sul suolo del sito Ispra (Fonte: http://www.consumosuolo.isprambiente.it) contiene 120 variabili. Sono stati considerati i seguenti anni: 2012, 2015, 2016, 2017, 2018, 2019, 2020. Nella fase di pulizia dei dati, è stata effettuata un'operazione volta a rimuovere le colonne che presentavano valori mancanti o contenevano solo zeri. Successivamente, sono state selezionate solo le colonne di interesse, che includono:

- "NOME_Provincia": Rappresenta il nome della provincia.
- "CSUOLO9: Densita_consumo_[m2/ha]": Rappresenta la densità di consumo di suolo in metri quadrati per ettaro.
- "DISECO3 Superficie_alterata_dal_consumo_di_suolo_60m_[%]": Indica la superficie alterata dal consumo di suolo entro una distanza di 60 metri.
- "FORMET5 Rapporto tra aree ad alta densità di urbanizzazione e aree ad alta e bassa densità [%]": Rappresenta il rapporto percentuale tra le aree ad alta densità di urbanizzazione e le aree ad alta e bassa densità.

Queste colonne sono state selezionate per la successiva analisi e sono state considerate rilevanti per l'ambito della tesi.

Per condurre l'analisi di dipendenza del numero delle aziende manifatturiere e agricole al consumo, è stato utilizzato il dataset fornito da InfoCamere attraverso il loro servizio Movimprese è un'analisi statistica trimestrale sulla nati-mortalità delle imprese, condotta da InfoCamere per conto dell'Unioncamere, basata sugli archivi di tutte le Camere di Commercio italiane (Fonte: https://www.infocamere.it/).

L'utilizzo del dataset di Movimprese fornito da InfoCamere offre una solida base di dati provenienti da diverse Camere di Commercio italiane. Ciò consente di condurre un'analisi accurata e rappresentativa del numero delle aziende manifatturiere e agricole, fornendo una panoramica significativa sullo stato e le dinamiche di questi settori nell'economia italiana.

Dai dati del dataset, sono state selezionate le seguenti categorie per la variabile "Numero delle Aziende agricole" nel settore A Agricoltura: A 01 Coltivazioni

agricole e produzione di prodotti animali, A 02 Silvicoltura ed utilizzo di aree forestali, ed escluso A 03 Pesca e acquacoltura.

"Numero delle Aziende Manifatturiere" include dati relativi al settore C Attività manifatturiere.

Le variabili scelte sono tutte variabili quantitative (numeriche).

## 2.2 Descrizione delle variabili

La copertura artificiale del suolo aumenta quando nuovi edifici, infrastrutture, strade, cantieri, e così via, vengono costruiti su una superficie originariamente naturale o semi-naturale. Questo fenomeno è noto come **consumo di suolo** e viene rilevato in un periodo di tempo specifico, ad esempio due anni consecutivi. Il Sistema Nazionale per la Protezione Ambientale (SNPA) tiene traccia del consumo di suolo e aggiorna annualmente la "Carta Nazionale" relativa (Munafò, 2022)**.**

Figura 2.3 Il consumo del suolo ha un impatto significativo sul potenziale agricolo a causa dell'occupazione di suoli fertili.

Fonte:https://www.isprambiente.gov.it/it/istituto

Con **consumo di suolo** si intende **l'incremento della copertura artificiale del suolo**, generalmente su base annuale, mentre con **suolo consumato** si intende la

**quantità complessiva di suolo a copertura artificiale in un dato**. Ad esempio, il consumo di suolo netto 2020-2021 è uguale alla **differenza tra il suolo consumato 2021 e il suolo consumato 2020**, ovvero alla crescita delle superfici artificiali in un anno tra il 2020 e il 2021 (Munafò, 2022).

Il territorio nazionale continua a subire una rapida trasformazione a causa del consumo di suolo. Nel corso dell'ultimo anno, sono state ricoperte altre 69,1 km2 di terreno con nuove costruzioni artificiali, pari a circa 19 ettari al giorno in media. Questo incremento rappresenta un'accelerazione evidente rispetto ai dati degli anni precedenti, invertendo il trend di riduzione registrato in passato e portando a una perdita di 2,2 metri quadrati di suolo ogni secondo in Italia. Queste sono le principali conclusioni emerse dal "Rapporto sul consumo di suolo, dinamiche territoriali e servizi ecosistemici" (Munafò, 2022).

Tabella 2.1 Velocità del consumo di suolo giornaliero netto degli ultimi 15 anni. Fonte: Elaborazioni ISPRA su cartografia SNPA

| | Consumo di suolo netto (ha/giorno) | Consumo di suolo netto revisionato[33] (ha/giorno) |
|---|---|---|
| 2006-2012 | 27,4 | 28,7 |
| 2012-2015 | 15,1 | 15,2 |
| 2015-2016 | 14,4 | 14,7 |
| 2016-2017 | 15,4 | 15,6 |
| 2017-2018 | 16,7 | 17,1 |
| 2018-2019 | 16,1 | 17,2 |
| 2019-2020 | 14,2 | 15,9 |
| 2020-2021 | 17,3 | - |

### 2.2.1 Variabile CSUOLO9

La variabile CSUOLO9 rappresenta densità di consumo di suolo [m2] rispetto all'area totale [ha]. Incremento del consumo di suolo annuale tra due anni conseguenti in metri quadrati per ogni ettaro di territorio.

La densità di consumo di suolo è una misura che esprime la quantità di suolo consumato in metri quadrati rispetto all'area totale in ettari. Indica la quantità di suolo che viene convertita o occupata da attività umane, come l'edificazione, l'urbanizzazione o l'infrastrutturazione, in relazione alla dimensione complessiva dell'area considerata. L'incremento del consumo di suolo annuale tra due anni successivi si riferisce alla variazione nella quantità di suolo consumato durante un determinato periodo di tempo.

Tabella 2.2 Tabella dei valori riassuntivi per la variabile CSUOLO9

| X2012 | X2015 | X2016 | X2017 | X2018 | X2019 | X2020 |
|---|---|---|---|---|---|---|
| Min. : 1.605 | Min. :-0.7011 | Min. :-0.09629 | Min. :-0.2391 | Min. :0.1002 | Min. :-0.01248 | Min. :0.1923 |
| 1st Qu.:10.021 | 1st Qu.: 2.6830 | 1st Qu.: 0.84455 | 1st Qu.: 0.7093 | 1st Qu.:1.0380 | 1st Qu.: 0.91877 | 1st Qu.:0.8269 |
| Median :19.262 | Median : 5.3071 | Median : 1.65011 | Median : 1.4016 | Median :1.9485 | Median : 1.66101 | Median :1.2290 |
| Mean :22.715 | Mean : 5.9946 | Mean : 1.90141 | Mean : 2.0190 | Mean :2.2577 | Mean : 2.23662 | Mean :1.9561 |
| 3rd Qu.:32.342 | 3rd Qu.: 7.6175 | 3rd Qu.: 2.45849 | 3rd Qu.: 2.5304 | 3rd Qu.:3.0280 | 3rd Qu.: 3.20194 | 3rd Qu.:2.3535 |
| Max. :98.357 | Max. :33.5534 | Max. : 7.99563 | Max. :10.2264 | Max. :8.3642 | Max. : 8.07609 | Max. :8.3695 |

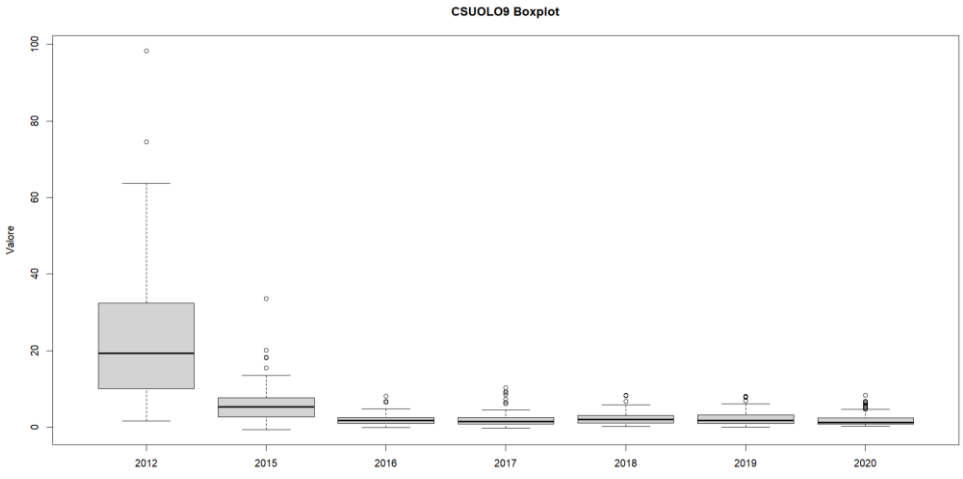


Figura 2.4 Box plot della variabile CSUOLO9

La variazione di CSUOLO9 nel 2012 si attesta tra 1.605 e 98.357, con una distribuzione ampia evidenziata dai quartili. La mediana è di 19.262 e la media è di 22.715, indicando una tendenza verso valori più elevati.

Nel 2015, la distribuzione dei valori di CSUOLO9 è meno ampia rispetto al 2012, con un range da -0.7011 a 33.5534. I quartili sono più bassi, con una mediana di 5.3071 e una media di 5.9946.

Nel 2016, la distribuzione dei valori di CSUOLO9 ha un range più piccolo rispetto al 2015, con un range da -0.09629 a 7.99563. I quartili sono più bassi rispetto all'anno precedente, con una mediana di 1.65011 e una media di 1.90141.

Nel 2017, la distribuzione dei valori di CSUOLO9 è simile a quella del 2016, con un range da -0.2391 a 10.2264. I quartili sono simili, con una mediana di 1.4016 e una media di 2.0190.

Nel 2018, la distribuzione dei valori di CSUOLO9 è ancora meno variabile rispetto al 2017, con un range da 0.1002 a 8.3642. I quartili sono ancora più alti, con una mediana di 1.9485 e una media di 2.2577.

Nel 2019, la distribuzione dei valori di CSUOLO9 è simile a quella del 2018, con un range da -0.01248 a 8.07609. I quartili sono simili, con una mediana di 1.66101 e una media di 2.23662.

Nel 2020, la distribuzione dei valori di CSUOLO9 è ancora più ristretta rispetto al 2019, con un range da 0.1923 a 8.3695. I quartili sono ancora più bassi, con una mediana di 1.2290 e una media di 1.9561.

I valori minimi nel 2012 si verificano nelle province di Valle d'Aosta/Vallée d'Aoste e Sardegna del Sud. A partire dal 2015, alcune province iniziano ad avere

valori negativi. Ad esempio, nel 2015 la provincia di Massa-Carrara presenta un valore negativo, nel 2016 è la provincia di Biella, nel 2017 è la provincia di Como, e nel 2019 è la provincia di Vercelli. Mentre i valori massimi tradizionalmente si verificano nelle province di Milano e Monza e della Brianza.

### 2.2.2 Variabile DISECO3

La variabile DISECO3 rappresenta la percentuale di superficie impattata dalla presenza di coperture artificiali entro una distanza di 60 metri. Per valutare correttamente l'impatto del consumo del suolo, è essenziale esaminare gli effetti sia diretti che indiretti della superficie coperta artificialmente sull'ambiente circostante. Questi effetti possono interessare importanti servizi ecosistemici di regolazione climatica ed idrogeologica, oltre alla biodiversità. Per quantificare l'estensione dell'area potenzialmente influenzata dalla presenza di coperture artificiali, si è scelto di utilizzare un approccio basato su buffer di 60, 100 e 200 metri, che consentono di generalizzare gli impatti senza assegnare pesi specifici ai diversi comparti ambientali coinvolti. Tuttavia, è importante tenere presente che questa analisi si limita alla dimensione orizzontale della superficie terrestre e che gli effetti indiretti e di disturbo potrebbero avere conseguenze significative per l'ecosistema e per la società nel suo insieme (Munafò, 2022).

La valutazione dell'impatto del consumo di suolo ha rivelato che una porzione significativa del territorio nazionale è interessata dalla copertura artificiale. In particolare, considerando buffer di diverse distanze (60, 100 e 200 metri), la superficie coinvolta è stata del 42,2%, 56,0% e 75,5% rispettivamente. Questi risultati rappresentano indicatori preoccupanti della portata del disturbo del consumo di suolo, poiché oltre la metà del territorio nazionale ha una copertura artificiale entro 100 metri di distanza e i tre quarti entro 200 metri. Inoltre,

aumentando la distanza di impatto a 1.000 metri, la quasi totalità del territorio nazionale sarebbe coperta (98%, con picchi del 99,9% in Liguria e Toscana). (L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) , 2016)

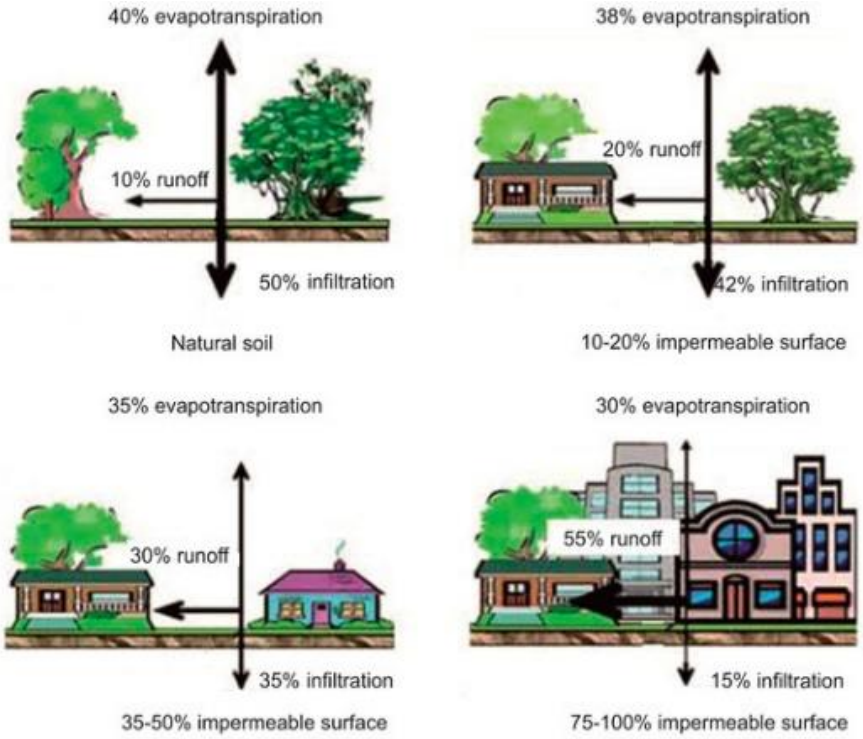


Figura 2.5 L'impermeabilizzazione del suolo riduce l'assorbimento dell'acqua.
Fonte: https://extension.unr.edu/publications.aspx

Inoltre, l'impermeabilizzazione delle aree antropizzate (come edifici e infrastrutture di trasporto) ha un impatto negativo sull'evapotraspirazione. Questa riduzione nell'assorbimento di calore dall'aria può portare alla creazione di climi urbani nuovi e severi, specialmente se associata a cambiamenti nell'uso del suolo e un aumento della densità edilizia (Moretti & Loprencipe, 2018).

Tabella 2.3 Tabella dei valori riassuntivi per la variabile DISECO3

| X2012 | X2015 | X2016 | X2017 | X2018 | X2019 | X2020 |
|---|---|---|---|---|---|---|
| Min. :12.04 | Min. :12.09 | Min. :12.10 | Min. :12.12 | Min. :12.13 | Min. :12.13 | Min. :12.13 |
| 1st Qu.:30.33 | 1st Qu.:30.43 | 1st Qu.:30.45 | 1st Qu.:30.48 | 1st Qu.:30.50 | 1st Qu.:30.51 | 1st Qu.:30.52 |
| Median :36.18 | Median :36.23 | Median :36.25 | Median :36.26 | Median :36.28 | Median :36.31 | Median :36.35 |
| Mean :36.92 | Mean :37.00 | Mean :37.03 | Mean :37.05 | Mean :37.08 | Mean :37.11 | Mean :37.13 |
| 3rd Qu.:41.72 | 3rd Qu.:41.79 | 3rd Qu.:41.81 | 3rd Qu.:41.83 | 3rd Qu.:41.88 | 3rd Qu.:41.90 | 3rd Qu.:41.94 |
| Max. :82.21 | Max. :82.30 | Max. :82.38 | Max. :82.41 | Max. :82.41 | Max. :82.42 | Max. :82.45 |

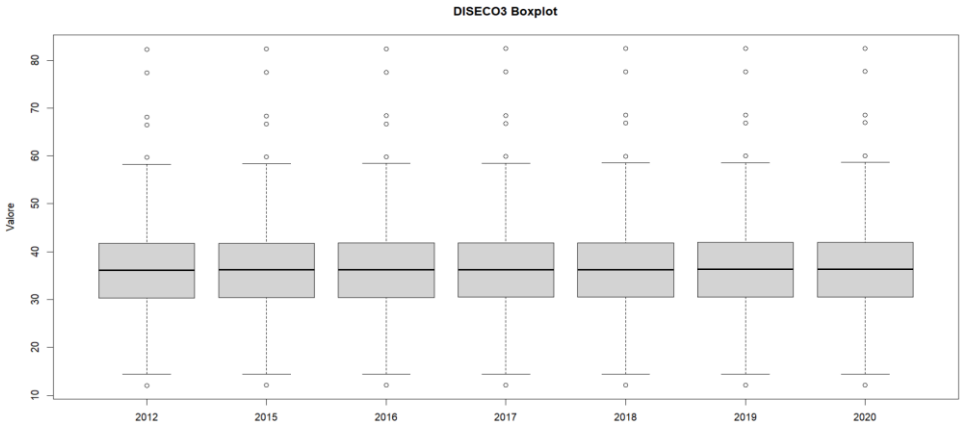


Figura 2.6 Box plot della variabile DISECO3

Si può notare che i valori di DISECO3 non variano considerevolmente da un anno all'altro. Ad esempio, nel 2012 il valore minimo è 12.04 e il valore massimo è 82.21, mentre nel 2018 il valore minimo è 12.13 e il valore massimo è 82.41.

Le medie annuali di DISECO3 mostrano una certa stabilità nel tempo, con valori che oscillano intorno a una media di circa 37.

I valori minimi si verificano nelle province di Valle d'Aosta/Vallée d'Aoste, Verbano-Cusio-Ossola e Bolzano/Bozen, mentre i valori massimi si verificano nelle province di Monza e della Brianza, Napoli e Milano.

### 2.2.3 Variabile FORMET5

La variabile FORMET5 rappresenta l'Indice di Dispersione Urbana in percentuale. Negli ultimi anni, ISPRA ha iniziato ad analizzare le forme di urbanizzazione e le tipologie di insediamenti, identificando alcuni indicatori efficaci per rappresentare i fenomeni di trasformazione territoriale. L'IDU (Indice di Dispersione Urbana)

FORMET5 rappresenta la dispersione territoriale attraverso il rapporto tra aree ad alta densità e aree ad alta e bassa densità e descrive la variazione di densità di urbanizzazione (EEA, 2006; ESPON, 2011).

Secondo il libro "Sprawl: A Compact History" (Bruegmann, 2005), **la dispersione urbana, detta anche città diffusa o invasione urbana (in inglese: urban sprawl e urban encroachment),** è un fenomeno urbanistico che si verifica quando una città si espande rapidamente e in modo disordinato, senza una adeguata e sostenibile pianificazione urbanistica. Questo fenomeno si manifesta soprattutto nelle zone periferiche, che sono soggette a continui mutamenti e caratterizzate da una polarizzazione tra il centro e la periferia urbana.

La bassa densità abitativa è il segno distintivo della dispersione urbana in città di medie e grandi dimensioni (oltre i 100.000 abitanti). Tra gli effetti di questo fenomeno ci sono la **riduzione degli spazi verdi, il consumo del suolo, la dipendenza dalle autovetture a causa della maggiore distanza dai servizi, dal posto di lavoro, dai mezzi di trasporto pubblico locale e la mancanza di infrastrutture per la mobilità sostenibile**, come piste ciclabili, marciapiedi o attraversamenti pedonali adeguatamente connessi (Bruegmann, 2005). Consumo di suolo e urban sprawl sono dunque fenomeni strettamente correlati e sono sempre più spesso considerati il risultato di una pianificazione territoriale poco efficiente.

Figura 2.7 Un esempio di espansione urbana è l'Area Metropolitana di Sydney

Fornte*:* http://danielkimsgeographyblog.blogspot.com/

Tabella 2.4 Tabella dei valori riassuntivi per la variabile FORMET5

| X2012 | X2015 | X2016 | X2017 | X2018 | X2019 | X2020 |
|---|---|---|---|---|---|---|
| Min. :53.54 | Min. :53.26 | Min. :53.27 | Min. :53.11 | Min. :53.04 | Min. :52.90 | Min. :52.83 |
| 1st Qu.:82.24 | 1st Qu.:82.14 | 1st Qu.:82.12 | 1st Qu.:82.07 | 1st Qu.:82.05 | 1st Qu.:81.97 | 1st Qu.:81.89 |
| Median :86.64 | Median :86.57 | Median :86.62 | Median :86.60 | Median :86.52 | Median :86.46 | Median :86.42 |
| Mean :85.31 | Mean :85.27 | Mean :85.26 | Mean :85.23 | Mean :85.18 | Mean :85.14 | Mean :85.10 |
| 3rd Qu.:90.22 | 3rd Qu.:90.11 | 3rd Qu.:90.12 | 3rd Qu.:90.12 | 3rd Qu.:90.09 | 3rd Qu.:90.05 | 3rd Qu.:89.99 |
| Max. :96.36 | Max. :96.37 | Max. :96.38 | Max. :96.39 | Max. :96.40 | Max. :96.42 | Max. :96.42 |

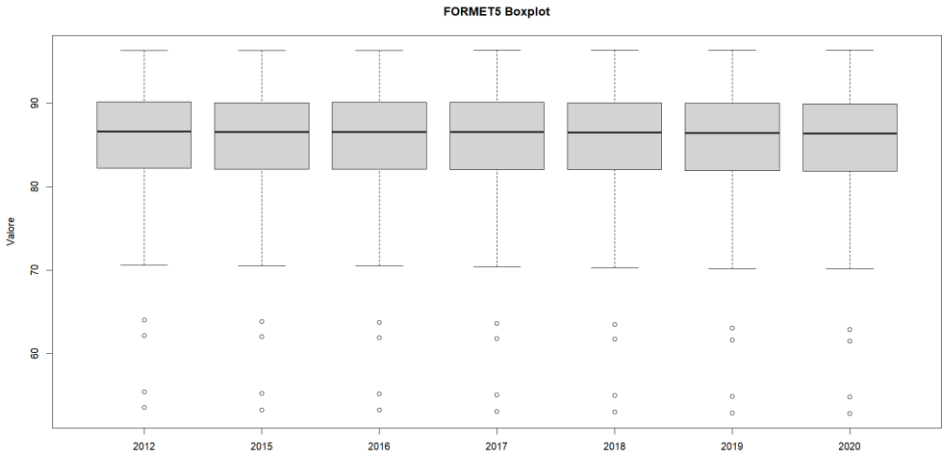


Figura 2.8 Box plot della variabile FORMET5

Possiamo osservare i seguenti punti di confronto tra gli anni:

Minimo e massimo: I valori minimi e massimi delle variabili non mostrano una certa variabilità tra gli anni. Ad esempio, la variabile nel 2012 ha un valore minimo di 53.54 e un valore massimo di 96.36, mentre le altre variabili hanno valori minimi e massimi simili.

Quartili: i valori dei quartili (1° quartile, mediana, 3° quartile) sono generalmente molto vicini tra gli anni. Ciò indica che la distribuzione dei dati è abbastanza simile tra gli anni considerati.

Anche la media delle variabili mostra una certa coerenza tra gli anni. Si aggira intorno a 85, con piccole variazioni tra le diverse variabili.

Le città con i valori minori sono Monza, Torino, Napoli e Milano, le quali sono caratterizzate da centri urbani compatti all'interno del limite comunale. Al contrario, i valori più alti si riscontrano per Perugia, Benevento, Latina e Catanzaro, che sono le città dove i processi di espansione della superficie urbanizzata a bassa densità hanno avuto un impatto maggiore sul territorio comunale.

Il range dei valori varia da 0,18 (Monza, città con fenomeno della diffusione soprattutto distribuito nella relativa conurbazione, al di fuori dei limiti amministrativi comunali) a 0,85 (Catanzaro).

## 2.2.4 Variabile Numero delle Aziende Agricole

Tabella 2.5 Tabella dei valori riassuntivi per la variabile Numero delle Aziende Agricole

| X2012 | X2015 | X2016 | X2017 | X2018 | X2019 | X2020 |
|---|---|---|---|---|---|---|
| Min. : 3.00 | Min. : 3.0 | Min. : 3.0 | Min. : 3.0 | Min. : 3.0 | Min. : 3.0 | Min. : 3.0 |
| 1st Qu.: 31.00 | 1st Qu.: 31.0 | 1st Qu.: 30.0 | 1st Qu.: 31.5 | 1st Qu.: 32.0 | 1st Qu.: 34.0 | 1st Qu.: 34.0 |
| Median : 63.00 | Median : 65.0 | Median : 66.0 | Median : 69.0 | Median : 65.0 | Median : 64.0 | Median : 69.0 |
| Mean : 97.77 | Mean :101.1 | Mean :101.4 | Mean :102.3 | Mean :102.6 | Mean :102.6 | Mean :103.7 |
| 3rd Qu.:138.00 | 3rd Qu.:137.0 | 3rd Qu.:135.5 | 3rd Qu.:135.0 | 3rd Qu.:138.5 | 3rd Qu.:138.0 | 3rd Qu.:142.5 |
| Max. :528.00 | Max. :548.0 | Max. :553.0 | Max. :569.0 | Max. :586.0 | Max. :602.0 | Max. :604.0 |

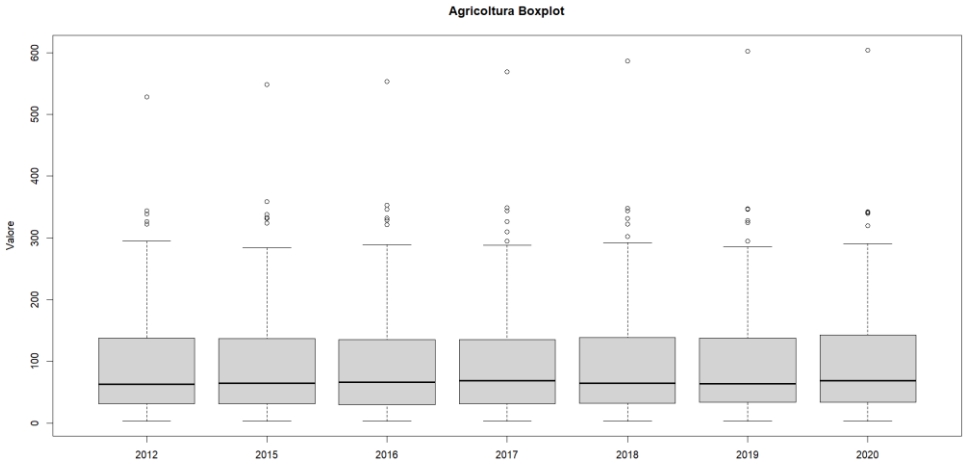


Figura 2.9 Box plot della variabile Numero delle Aziende Agricole

Analizzando i dati della variabile "Numero delle Aziende Agricole" nei diversi anni, possiamo osservare quanto segue:

Il numero minimo di aziende agricole si registra nel 2012 con un valore minimo di 3. Questo valore aumenta nel corso degli anni successivi.

La mediana, che rappresenta il valore centrale della distribuzione, aumenta gradualmente da un valore di 63 nel 2012 a 69 nel 2020. Ciò suggerisce un incremento generale nel numero medio di aziende agricole nel periodo considerato.

La media, invece, mostra una tendenza simile, aumentando da 97.77 nel 2012 a 103.7 nel 2020.

I quartili (1° e 3° quartile) forniscono informazioni sulla distribuzione dei dati. Possiamo notare che nel corso degli anni, il 25% inferiore dei valori si riduce leggermente, mentre il 25% superiore dei valori rimane relativamente stabile.

Il valore massimo delle aziende agricole aumenta nel corso degli anni, passando da 528 nel 2012 a 604 nel 2020.

Questi dati indicano un leggero aumento del numero medio di aziende agricole nel periodo considerato, con una distribuzione che tende ad allargarsi. L'analisi dei dati relativi alla variabile "Numero delle Aziende Agricole" mostra che le province con il numero minimo di aziende agricole sono Trieste, Barletta-Andria-Trani, Siracusa e Monza e della Brianza. D'altra parte, le province con il numero massimo di aziende agricole sono Cuneo, Perugia e Bolzano

### 2.2.5 Variabile Numero delle Aziende Manifatturiere

Tabella 2.6 Tabella dei valori riassuntivi per la variabile Numero delle Aziende Manifatturiere

| X2012 | X2015 | X2016 | X2017 | X2018 | X2019 | X2020 |
|---|---|---|---|---|---|---|
| Min. : 644 | Min. : 585 | Min. : 579 | Min. : 578 | Min. : 580 | Min. : 581 | Min. : 583 |
| 1st Qu.: 1942 | 1st Qu.: 1805 | 1st Qu.: 1760 | 1st Qu.: 1728 | 1st Qu.: 1695 | 1st Qu.: 1664 | 1st Qu.: 1644 |
| Median : 3468 | Median : 3283 | Median : 3215 | Median : 3167 | Median : 3124 | Median : 3062 | Median : 3041 |
| Mean : 4646 | Mean : 4376 | Mean : 4313 | Mean : 4261 | Mean : 4205 | Mean : 4134 | Mean : 4065 |
| 3rd Qu.: 5478 | 3rd Qu.: 5134 | 3rd Qu.: 5024 | 3rd Qu.: 4963 | 3rd Qu.: 4877 | 3rd Qu.: 4818 | 3rd Qu.: 4724 |
| Max. :28847 | Max. :27740 | Max. :27457 | Max. :27347 | Max. :27080 | Max. :26791 | Max. :25328 |

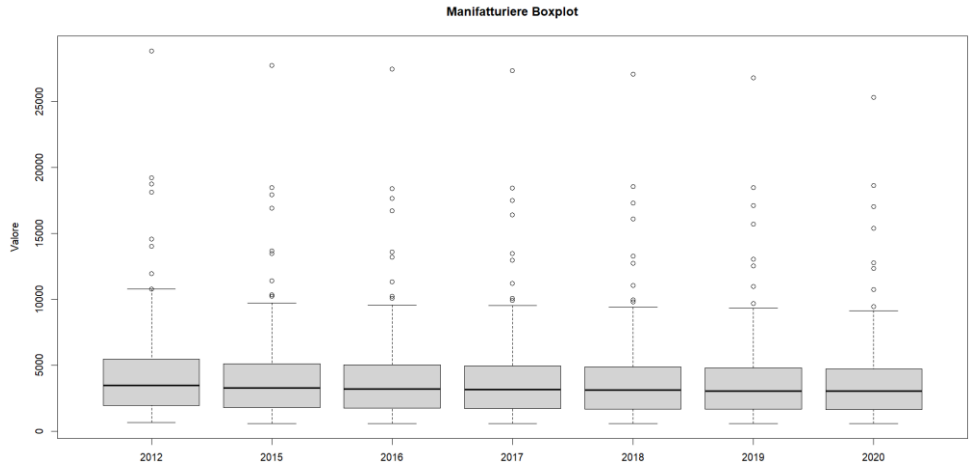


Figura 2.10 Box plot della variabile Numero delle Aziende Manifatturiere

Analizzando ulteriormente i dati della variabile "Numero delle Aziende Manifatturiere" nei diversi anni, possiamo notare quanto segue:

Il valore massimo delle aziende manifatturiere si registra nel 2012 con 28847, indicando una forte presenza di aziende manifatturiere in quell'anno. Tuttavia, nel corso degli anni successivi, si osserva una tendenza alla diminuzione del numero di aziende manifatturiere, con un valore massimo di 25328 nel 2020.

Il valore minimo delle aziende manifatturiere è di 578 nel 2017, rappresentando una bassa presenza di aziende manifatturiere in quell'anno. Questo suggerisce una potenziale contrazione del settore manifatturiero in determinate province in quel periodo.

La mediana dei dati, che rappresenta il valore al centro della distribuzione, si mantiene relativamente stabile nel periodo considerato, oscillando tra 3041 nel 2020 e 3215 nel 2016.

La media mostra una diminuzione nel corso degli anni, passando da 4646 nel 2012 a 4065 nel 2020. Questo indica una tendenza generale verso una riduzione del numero medio di aziende manifatturiere nel periodo considerato.

I quartili (1° e 3° quartile) indicano che la maggior parte dei dati si concentra nella parte inferiore della scala, con una diminuzione nel corso degli anni. Questo suggerisce una riduzione della dispersione dei dati e un'ulteriore concentrazione delle aziende manifatturiere verso valori inferiori.

Complessivamente, l'analisi dei dati evidenzia una tendenza alla diminuzione del numero di aziende manifatturiere nel periodo considerato, con una maggiore concentrazione di aziende verso valori inferiori. Ciò potrebbe riflettere sfide o cambiamenti strutturali nel settore manifatturiero delle province considerate.

# Capitolo 3 Indicatori Compositi e Medie Penalizzate

Secondo il libro “Open issues in composite indicators. A starting point and a reference on some state-of-the-art issues”, il termine "indicatore" ha origine dal latino indicator, che significa "colui che indica" o "elemento che indica o segnala qualcosa". Un indicatore può essere qualsiasi variabile normalizzata, ma di solito si riferisce ad un rapporto composto da un numeratore (la variabile che fornisce il significato) e un denominatore (la variabile che consente la comparabilità nello spazio e/o nel tempo). Gli "indicatori elementari" sono componenti di un fenomeno complesso da misurare, mentre gli "indicatori compositi" rappresentano la misura del fenomeno stesso. Gli indicatori elementari sono utilizzati per monitorare o valutare il successo o l'adeguatezza delle attività implementate, mentre gli indicatori compositi sono espressioni quantitative composte da diverse variabili in grado di riassumere la tendenza del fenomeno a cui si riferiscono.

Il fenomeno complesso viene definito con un approccio multidimensionale e, quindi, è un fattore latente che richiede un processo per definire il concetto e decomporlo in indicatori individuali. Quando si calcola un indicatore composito, si parte dalla matrice originale costruita con le dimensioni del fenomeno in studio e si cercano di "trattare" statisticamente gli indicatori individuali per ridurre le dimensioni nello spazio e perdere il minor numero possibile di informazioni, per

arrivare a un numero singolo. (Otoiu, Pareto, Grimaccia, Mazziotta, & Terzi, 2021)

La power mean, o media potenziata, è una misura di aggregazione che combina valori di un insieme di indicatori in un unico valore rappresentativo. La power mean di ordine p di un insieme di valori viene calcolata elevando ciascun valore alla potena p, calcolando la media aritmetica di queste potenze e infine elevando il risultato all'inverso di p.

Formalmente, la power mean di ordine p associata a un insieme di valori $I = [x_1, x_2, \dots, x_n]$ è definita come:

$$M_p(I) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}} \quad (1)$$

La media aritmetica, la media geometrica e la media armonica sono casi particolari della power mean (media di potenza) per p = 1, p = 0 e p = −1, rispettivamente.

La media aritmetica è il caso particolare della power mean con ordine p = 1. Viene calcolata come la somma di tutti i valori divisa per il numero totale di valori.

$$M_1(I) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2)$$

La media geometrica è il caso particolare della power mean con ordine p = 0. Viene calcolata come la radice n-esima del prodotto di tutti i valori.

$$M_0(\mathrm{I}) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \tag{3}$$

La media armonica è il caso particolare della power mean con ordine p = −1. Viene calcolata come il reciproco della media delle reciprocità dei valori.

$$M_{-1}(\mathrm{I}) = \left(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{x_i}\right)^{-1} \tag{4}$$

# 3.1 Nuova classe di indicatori compositi

L'articolo "A New Class of Composite Indicators: The Penalized Power Mean" (Recchioni, Mariani, & Ciommi, 2023 (sottomesso)), introduce una nuova classe di indicatori compositi basata su una penalizzazione della media potenziata. L'idea alla base di questo approccio consiste nel moltiplicare la media potenziata per un fattore che tiene conto dell'eterogeneità orizzontale tra gli indicatori, penalizzando le unità con una maggiore eterogeneità.

Il processo di penalizzazione coinvolge una serie di passaggi per calcolare un fattore che modifica la media e cattura l'eterogeneità orizzontale tra gli indicatori. Ecco una lista dettagliata dei passaggi coinvolti:

1. **<u>Calcolo della media potenza:</u>**

   Per ogni unità, calcolare la media potenza degli indicatori normalizzati associati a quella unità. La media potenza rappresenta il valore medio, dando maggiore peso a determinati indicatori in base alla loro importanza.

$$M_{p,i} = \begin{cases} \left(\frac{1}{m}\sum_{j=1}^{m} I_{i,j}^{p}\right)^{\frac{1}{p}}, p \neq 0, \\ \left(\prod_{j=1}^{m} I_{i,j}\right)^{\frac{1}{m}}, p = 0. \end{cases} \tag{5}$$

dove:

- $M_{p,i}$ rappresenta la media potenza di ordine p associata all'unità i. È il risultato della media ponderata degli indicatori normalizzati relativi all'unità i, dove il peso di ciascun indicatore è determinato dalla sua importanza (attraverso l'esponente p)
- m rappresenta il numero totale di indicatori
- p rappresenta l'ordine della media potenza. Quando p è diverso da zero, la formula calcola la media potenza considerando l'esponente p. Quando p è uguale a zero, la formula calcola la media semplice degli indicatori normalizzati.
- $I_{i,j}$ rappresenta il vettore dei valori normalizzati degli indicatori relativi all'unità i. È un vettore colonna di dimensione m

2. **<u>Ridimensionamento degli indicatori:</u>**

Ridimensionare gli indicatori dividendo ciascun indicatore per la sua media potenza corrispondente. Questo passaggio aiuta a standardizzare gli indicatori rispetto alle loro medie potenza.

$$\tilde{I}_{i,j} = \frac{I_{i,j}}{M_{p,i}}, j = 1,2, \dots, m, i = 1,2, \dots, n. \quad (6)$$

dove:

- $\tilde{I}_{i,j}$ rappresenta il vettore degli indicatori normalizzati scalati

3. **<u>Applicazione della trasformazione di Box-Cox:</u>**

Applicare la trasformazione di Box-Cox agli indicatori ridimensionati.

La trasformazione di Box-Cox è una tecnica statistica generalmente utilizzata per stabilizzare la varianza di una distribuzione e ottenere una distribuzione approssimativamente normale. Prende il nome dai suoi sviluppatori, George Box e David Cox (Box & Cox, 1964). La trasformazione di Box-Cox è definita da una famiglia di trasformazioni parametriche che includono l'esponenziale come caso particolare. La forma generale della trasformazione di Box-Cox è data dall'equazione:

$$y_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & se\ \lambda\ \neq 0, \\ log(x) & se\ \lambda = 0. \end{cases} \quad (7)$$

dove:

- x è la variabile di partenza
- $y(\lambda)$ è la variabile trasformata
- Il parametro λ controlla la forma della trasformazione e può assumere qualsiasi valore reale. La scelta di λ dipende dall'obiettivo della trasformazione. Se λ=1, la trasformazione di Box-Cox diventa una trasformazione lineare, mentre per λ=0 diventa una trasformazione logaritmica. Valori diversi di λ consentono di esplorare trasformazioni non lineari.

Nel contesto descritto, si utilizza la trasformazione di Box-Cox con il parametro p per ottenere il valore trasformato.

$$\mathrm{h_p}(x) = \begin{cases} \frac{x^\mathrm{p} - 1}{p} & se\ \mathrm{p} \neq 0, \\ ln(x) & \mathrm{se\ p} = 0. \end{cases} \quad (8)$$

Il valore trasformato ottenuto viene impiegato per risolvere problemi di ottimizzazione.

4. **Calcolo della varianza degli indicatori trasformati:**

Calcolare la varianza degli indicatori trasformati ottenuti dal passaggio precedente. La varianza misura la dispersione dei punti dati intorno alla media. Fornisce un'indicazione della variabilità o eterogeneità tra gli indicatori trasformati. La varianza è data dalla formula (9) in quanto $\mathrm{E}[\mathrm{h_p}(I) = 0]$

$$\tilde{\mathrm{S}}^2_{\mathrm{p,i}} = \frac{1}{\mathrm{m}}\sum_{\mathrm{j=1}}^{\mathrm{m}} (\mathrm{h_p}(\tilde{\mathrm{I}}_{\mathrm{i,j}}))^2 , \mathrm{i} = 1, 2, \ldots \mathrm{n}. \quad (9)$$

Questa formula calcola l'errore o la perdita di informazione causata dalla sostituzione del vettore degli indicatori $\mathrm{h_p}(\tilde{\mathrm{I}}_{\mathrm{i,j}})$ con $\mathrm{h_p}(1) = 0$. Si dimostra facilmente che $\mathrm{h_p}(1) = 0$ e $M_p(\tilde{\mathrm{I}}_{\mathrm{i}})$=1 per i=1, 2, …,n, basandosi sulla proprietà di omogeneità dei mezzi di potenza di ordine p.

La formula può essere scomposta nel seguente modo:

- $\tilde{S}^2_{p,i}$ rappresenta l'errore o la perdita di informazione per l'indicatore *i* nella popolazione p.
- m è il numero di sottopopolazioni o strati.
- j rappresenta l'indice per ogni sottopopolazione.
- $\mathrm{h_p}(\tilde{\mathrm{I}}_{\mathrm{i,j}})$ rappresenta il valore dell'indicatore *i* nella sottopopolazione j dopo aver applicato la trasformazione di Box-Cox con il parametro p. Questo valore trasformato è utilizzato per risolvere problemi di ottimizzazione.

5. **Calcolo dell'immagine inversa della varianza:**

Utilizzare la funzione di Box-Cox per ottenere l'immagine inversa della varianza calcolata nel passaggio precedente. Questa immagine inversa serve come fattore di penalizzazione.

$$g_{p,i}^{\pm} = h_p^{-1}\left(\pm\tilde{S}_{p,i}^2\right) = (1 \pm p\tilde{S}_{p,i}^2)^{\frac{1}{p}} = \begin{cases} (1 \pm p\tilde{S}_{p,i}^2)^{\frac{1}{p}}, p \neq 0 \\ \exp\left(\pm\tilde{S}_{0,i}^2\right), p = 0 \end{cases} \quad (10)$$

**Interpretazione del fattore di penalizzazione:** Il fattore di penalizzazione ottenuto rappresenta una "varianza" nello spazio trasformato. Rappresenta l'errore relativo o la perdita di informazione associata alla sostituzione della media potenza con il vettore di indicatori trasformati. Considera l'eterogeneità tra gli indicatori e riflette l'impatto dell'utilizzo della media potenza come valore rappresentativo.

Seguendo questi passaggi, il fattore di penalizzazione viene calcolato per ogni unità, considerando le medie potenza, il ridimensionamento, la trasformazione di Box-Cox e i calcoli della varianza nello spazio trasformato. Questo fattore contribuisce a considerare l'eterogeneità orizzontale tra gli indicatori e quantifica la perdita di informazione quando si sostituisce la media potenza con il vettore di indicatori trasformati.

Nella descrizione fornita, viene esaminato il comportamento del termine di penalizzazione $g_{p,i}$ con polarità positiva in funzione dell'ordine p. È importante notare che poiché $\tilde{S}_{p,i}^2$ (la varianza trasformata) è sempre non negativa, il fattore $(1 - p\tilde{S}_{p,i}^2)^{\frac{1}{p}}$ risulta essere positivo solo se vale l'ineguaglianza $p\tilde{S}_{p,i}^2 \leq 1$. Questa disuguaglianza può essere soddisfatta o meno a seconda del segno di p. Per valori di p maggiori di zero, si può dimostrare che $\tilde{S}_{p,i}^2$ rappresenta una funzione positiva non decrescente di p. Ciò significa che all'aumentare di p, $\tilde{S}_{p,i}^2$ diminuisce. Inoltre, si osserva che il limite di p che tende all'infinito di $p\tilde{S}_{p,i}^2$è zero ($lim_{p\to+\infty}\, p\tilde{S}_{p,i}^2 = 0$) e il limite di p che tende a zero positivo di $p\tilde{S}_{p,i}^2$ è zero ($lim_{p\to 0}\ + p\tilde{S}_{p,i}^2 = 0$). Queste proprietà indicano che per valori sufficientemente grandi e piccoli di p, l'ineguaglianza $p\tilde{S}_{p,i}^2 \leq 1$ è soddisfatta.

D'altra parte, quando p è zero o negativo, il termine di penalizzazione è sempre nonnegativo. Inoltre, quando p < 0, il termine $(1 - p\tilde{S}_{p,i}^2)^{\frac{1}{p}}$ è inferiore a uno perché l'esponente $\frac{1}{p}$ è negativo e la base della potenza $1 - p\tilde{S}_{p,i}^2$ è maggiore di uno. Ciò significa che non ci sono restrizioni sulla grandezza $p\tilde{S}_{p,i}^2$ quando p < 0. Nel caso in cui p = 0, la penalizzazione è inferiore a uno per la polarità positiva, rappresentata da $exp(-\tilde{S}_{0,i}^2)$.

Per la polarità negativa dei sottoindicatori, il termine di penalizzazione $(1 + p\tilde{S}_{p,i}^2)^{\frac{1}{p}}$, dove $p \neq 0$, e $exp(-\tilde{S}_{0,i}^2)$, quando p = 0 è superiore a uno. Questo perché valori più elevati dell'indicatore composito indicano posizioni di classifica più basse a causa della polarità negativa.

**La media potenziata penalizzata** ($PM_{p,i}^{\pm}$) è definita come la media ponderata ($M_{p,i}$) moltiplicata per un fattore di penalizzazione ($g_{(p,i)}^{\pm}$) che dipende dalla perdita di informazioni relativa ($\tilde{S}_{p,i}^2$). Il fattore ($g_{(p,i)}^{\pm}$) penalizza le unità in base all'equilibrio tra gli indicatori, favorendo le unità con indicatori più bilanciati. Il segno dipende dalla polarità (positiva o negativa) del fenomeno misurato.

$$PM_{p,i}^{\pm} = M_{p,i} g_{p,i}^{\pm} \tag{11}$$

dove:

- $M_{p,i}$ è la media ponderata penalizzata di ordine p
- Il segno ± dipende dalla polarità (positiva o negativa) del fenomeno misurato.
- $g_{(p,i)}^{\pm}$ è il fattore di penalizzazione che dipende dalla perdita relativa di informazione $\tilde{S}_{p,i}^2$.

La media geometrica (non-penalizzata/penalizzata) con ordine p = 0 dovrebbe essere preferita rispetto ad altre medie di potenza, perché è non-compensativa.

Approccio non-compensativo penalizza lo squilibrio tra gli indicatori, incoraggiando il miglioramento degli indicatori deboli e penalizzando le prestazioni degli indicatori elevati.

## 3.2 Creazione di un nuovo indicatore composito per la misurazione del consumo di suolo

Per ottenere l'indicatore composito per la misurazione del consumo di suolo, vengono utilizzate tre variabili relative alle Province Italiane (n = 107) per gli anni 2012, 2015, 2016, 2017, 2018, 2019, 2020: CSUOLO9, DISECO3 e FORMET5.

Prima di procedere con il calcolo, i dati di queste variabili vengono normalizzati utilizzando il valore massimo, al fine di ottenere valori compresi tra 0 e 1.

Tabella 3.1:Summary Valori Normalizzate 2012

| CSUOLO9 | DISECO3 | FORMET5 |
|---|---|---|
| Min. :0.01632 | Min. :0.1464 | Min. :0.5557 |
| 1st Qu.:0.10189 | 1st Qu.:0.3690 | 1st Qu.:0.8534 |
| Median :0.19584 | Median :0.4401 | Median :0.8991 |
| Mean :0.23094 | Mean :0.4491 | Mean :0.8854 |
| 3rd Qu.:0.32882 | 3rd Qu.:0.5075 | 3rd Qu.:0.9363 |
| Max. :1.00000 | Max. :1.0000 | Max. :1.0000 |

Successivamente, viene calcolata la composizione degli indicatori sui dati utilizzando sia il media classica che la media penalizzata per diversi valori dell'ordine p.

- Quando "p" è uguale a 1, si ottiene la media aritmetica.
- Quando "p" è uguale a 0, si ottiene la media geometrica.
- Quando "p" è uguale a −1, si ottiene la media armonica.

Dopo il calcolo delle medie classiche e penalizzate, si procede al calcolo del ranking utilizzando le medie ottenute.

In questa sezione, viene investigato se la classifica indotta dalle medie classiche di ordine p e le corrispondenti medie penalizzate differiscano

Tabella 3.2 Confronto il Rank Top 20 Province con Consumo di suolo basso 2012

| NOME_Provincia | RankMC_ aritmetic | RankMC_ geometric | RankMC_ harmonic | RankMP_ aritmetic | RankMP_ geometric | RankMP_ harmonic |
|---|---|---|---|---|---|---|
| Valle d'Aosta/Vallée d'Aoste | 1 | 2 | 23 | 1 | 1 | 1 |
| Sud Sardegna | 2 | 3 | 5 | 2 | 6 | 7 |
| Verbano-Cusio-Ossola | 3 | 2 | 8 | 3 | 2 | 2 |
| Bolzano/Bozen | 4 | 4 | 20 | 4 | 3 | 3 |
| Belluno | 5 | 8 | 36 | 6 | 4 | 5 |
| Sondrio | 6 | 7 | 25 | 10 | 5 | 4 |
| Grosseto | 7 | 10 | 20 | 7 | 10 | 10 |
| Nuoro | 8 | 5 | 6 | 14 | 7 | 6 |
| Rieti | 10 | 17 | 62 | 8 | 9 | 11 |
| Trento | 11 | 9 | 15 | 15 | 8 | 9 |
| Oristano | 12 | 14 | 20 | 12 | 13 | 16 |
| Cagliari | 13 | 6 | 1 | 19 | 20 | 23 |
| Isernia | 14 | 22 | 66 | 17 | 12 | 14 |
| Siena | 15 | 23 | 43 | 14 | 14 | 23 |
| Enna | 17 | 21 | 39 | 18 | 16 | 18 |
| L'Aquila | 19 | 11 | 23 | 23 | 12 | 8 |
| Crotone | 23 | 13 | 15 | 26 | 16 | 12 |
| Sassari | 23 | 15 | 17 | 25 | 17 | 15 |
| Potenza | 27 | 29 | 51 | 27 | 19 | 17 |
| Matera | 29 | 20 | 25 | 35 | 19 | 13 |

La tabella fornisce i cambiamenti nel ranking delle province in base a diverse medie calcolate, includendo le medie aritmetiche, geometriche e armoniche per l'Indicatore Media Consumo di Suolo (RankMC) e l'Indicatore Media Penalizzata (RankMP). Ogni colonna rappresenta un tipo di media, mentre le righe corrispondono alle diverse province. Dalla tabella fornita, possiamo osservare che la media armonica (RankMC_harmonic e RankMP_harmonic) si comporta in modo diverso rispetto alle medie aritmetiche (RankMC_aritmetic e RankMP_aritmetic) e geometriche (RankMC_geometric e RankMP_geometric).

Per esempio, prendendo come esempio la provincia di Valle d'Aosta/Vallée d'Aoste, la quale occupa il primo posto sia per RankMC_aritmetic che per RankMP_aritmetic, ma ottiene un piazzamento inferiore (ventitreesimo posto) sia

per RankMC_harmonic che per RankMP_harmonic. Questo suggerisce che la media armonica tende a penalizzare le province che hanno valori estremamente alti o bassi rispetto alle altre province. Allo stesso modo, la provincia di Sud Sardegna si posiziona in modo diverso nelle diverse medie. Ad esempio, ottiene il secondo posto sia per RankMC_aritmetic che per RankMP_aritmetic, mentre si posiziona al sesto posto per RankMC_geometric e al settimo posto per RankMP_harmonic. Ciò implica che le medie geometriche e armoniche possono rispondere in modo diverso alle variazioni dei valori delle province, rispetto alle medie aritmetiche.

Tabella 3.3 Confronto il Rank Bottom 20 Province con Consumo di suolo alto 2012

| NOME_Provincia | RankMC_ aritmetic | RankMC_ geometric | RankMC_ harmonic | RankMP_ aritmetic | RankMP_ geometric | RankMP_ harmonic |
|---|---|---|---|---|---|---|
| Monza e della Brianza | 107 | 107 | 107 | 107 | 107 | 107 |
| Brindisi | 106 | 106 | 106 | 106 | 106 | 106 |
| Milano | 105 | 103 | 103 | 105 | 105 | 105 |
| Padova | 105 | 105 | 104 | 104 | 104 | 104 |
| Napoli | 103 | 105 | 105 | 103 | 103 | 103 |
| Treviso | 102 | 101 | 101 | 101 | 101 | 101 |
| Lecce | 101 | 102 | 102 | 97 | 100 | 102 |
| Roma | 101 | 99 | 98 | 102 | 102 | 100 |
| Ragusa | 99 | 100 | 100 | 99 | 99 | 99 |
| Ravenna | 98 | 98 | 97 | 100 | 99 | 98 |
| Varese | 97 | 96 | 85 | 98 | 97 | 97 |
| Latina | 96 | 97 | 99 | 95 | 95 | 95 |
| Venezia | 95 | 92 | 76 | 96 | 96 | 94 |
| Vicenza | 94 | 94 | 89 | 92 | 93 | 93 |
| Bari | 93 | 95 | 93 | 88 | 94 | 96 |
| Cremona | 92 | 91 | 87 | 93 | 91 | 89 |
| Rovigo | 91 | 93 | 91 | 90 | 88 | 87 |
| Novara | 90 | 88 | 81 | 91 | 90 | 88 |
| Gorizia | 89 | 83 | 71 | 94 | 92 | 91 |
| Rimini | 88 | 86 | 79 | 89 | 89 | 92 |

Quando si analizzano i rank relativi ai valori alti del consumo, si osserva che le medie aritmetiche, geometriche e armoniche generano valori identici per ogni provincia. Di conseguenza, non si riscontrano discrepanze o variazioni nel posizionamento delle province nel ranking.

Questo significa che, nel contesto delle province con valori di consumo elevati, le diverse medie utilizzate per calcolare i rank producono risultati coerenti e concordanti. Le tre tipologie di media mostrano un accordo completo nel determinare la posizione relativa delle province all'interno del ranking.

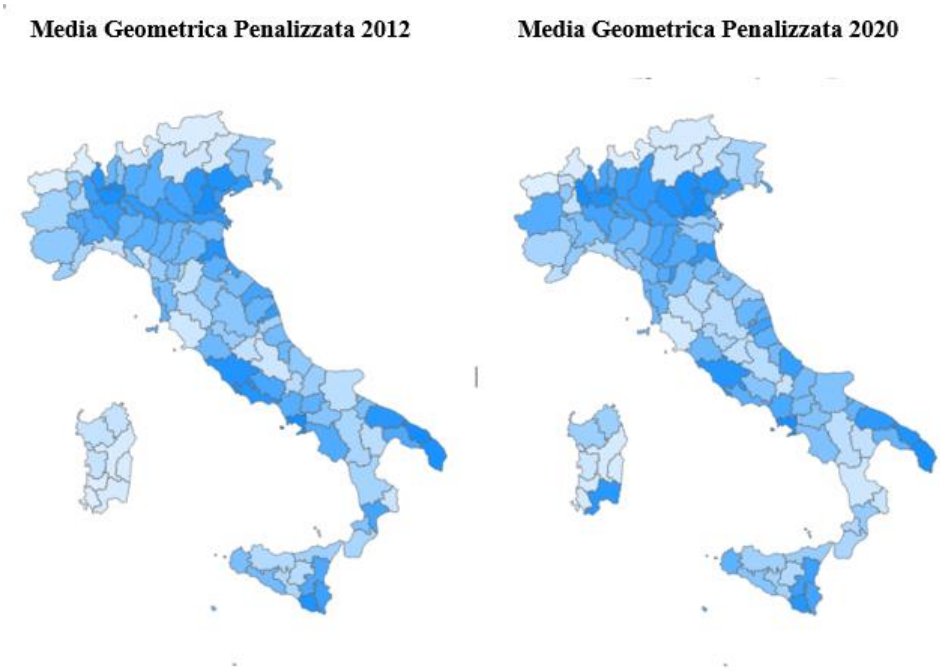


Figura 3.1 Saturazione del colore in base al rank delle province Media Geometrica Penalizzata 2012, Media Geometrica Penalizzata 2020

Tabella 3.4 Confronto il Rank Media Geometrica Penalizzata Bottom 20 Province con Consumo di suolo alto anni 2012-2020:

| NOME_Provincia | 2012_Rank MP_geometric | 2015_Rank MP_geometric | 2016_Rank MP_geometric | 2017_Rank MP_geometric | 2018_Rank MP_geometric | 2019_Rank MP_geometric | 2020_Rank MP_geometric |
|---|---|---|---|---|---|---|---|
| Padova | 104 | 97 | 102 | 103 | 107 | 103 | 107 |
| Monza e della Brianza | 107 | 104 | 101 | 104 | 100 | 102 | 106 |
| Novara | 90 | 64 | 76 | 91 | 70 | 75 | 105 |
| Napoli | 103 | 105 | 107 | 107 | 105 | 106 | 104 |
| Lecce | 100 | 102 | 104 | 99 | 104 | 105 | 103 |
| Vicenza | 93 | 92 | 100 | 101 | 97 | 90 | 102 |
| Milano | 105 | 107 | 105 | 100 | 102 | 94 | 101 |
| Brindisi | 106 | 98 | 93 | 94 | 101 | 101 | 99 |
| Treviso | 101 | 103 | 106 | 106 | 106 | 107 | 98 |
| Roma | 102 | 101 | 85 | 81 | 94 | 90 | 97 |
| Ragusa | 99 | 106 | 94 | 98 | 93 | 91 | 96 |
| Ravenna | 99 | 90 | 59 | 58 | 80 | 83 | 95 |
| Bari | 94 | 100 | 96 | 93 | 91 | 99 | 93 |
| Varese | 97 | 88 | 60 | 87 | 98 | 81 | 90 |
| Cremona | 91 | 75 | 61 | 64 | 61 | 88 | 86 |
| Venezia | 96 | 92 | 70 | 102 | 96 | 95 | 78 |
| Latina | 95 | 83 | 74 | 76 | 37 | 49 | 69 |
| Rimini | 89 | 80 | 84 | 68 | 83 | 58 | 58 |
| Rovigo | 88 | 25 | 68 | 28 | 56 | 41 | 39 |
| Gorizia | 92 | 89 | 92 | 89 | 95 | 39 | 35 |

Tabella 3.5 Confronto il Rank Media Geometrica Penalizzata Top 20 Province con Consumo di suolo basso anni 2012-2020:

| NOME_Provincia | 2012_Rank MP_geometric | 2015_Rank MP_geometric | 2016_Rank MP_geometric | 2017_Rank MP_geometric | 2018_Rank MP_geometric | 2019_Rank MP_geometric | 2020_Rank MP_geometric |
|---|---|---|---|---|---|---|---|
| Valle d'Aosta/Vallée d'Aoste | 1 | 2 | 2 | 1 | 1 | 1 | 1 |
| Verbano-Cusio-Ossola | 2 | 1 | 5 | 2 | 8 | 8 | 3 |
| Bolzano/Bozen | 3 | 9 | 4 | 37 | 6 | 5 | 5 |
| Belluno | 4 | 3 | 1 | 14 | 4 | 3 | 9 |
| Sondrio | 5 | 4 | 9 | 3 | 3 | 2 | 2 |
| Sud Sardegna | 6 | 7 | 7 | 6 | 5 | 10 | 6 |
| Nuoro | 7 | 5 | 10 | 5 | 2 | 4 | 5 |
| Trento | 8 | 14 | 11 | 4 | 9 | 13 | 8 |
| Rieti | 9 | 8 | 17 | 16 | 11 | 22 | 17 |
| Grosseto | 10 | 16 | 15 | 12 | 12 | 9 | 7 |
| Isernia | 12 | 11 | 12 | 8 | 19 | 14 | 13 |
| L'Aquila | 12 | 6 | 3 | 7 | 28 | 7 | 21 |
| Oristano | 13 | 12 | 22 | 15 | 10 | 12 | 11 |
| Siena | 14 | 17 | 25 | 13 | 13 | 16 | 14 |
| Crotone | 16 | 44 | 8 | 9 | 8 | 11 | 12 |
| Enna | 16 | 28 | 16 | 17 | 20 | 15 | 24 |
| Sassari | 17 | 20 | 41 | 29 | 25 | 25 | 35 |
| Matera | 19 | 18 | 6 | 10 | 14 | 20 | 16 |
| Potenza | 19 | 23 | 19 | 22 | 42 | 18 | 18 |
| Cagliari | 20 | 80 | 50 | 37 | 92 | 96 | 94 |

L'Indicatore Media Classica e Media Penalizzata è un metodo utilizzato per valutare le province in base al loro livello di consumo di suolo. Tuttavia, al fine di ottenere una visione più approfondita, viene eseguita un'analisi dei cluster.

# Capitolo 4 Cluster

Il presente capitolo tratta diverse tecniche di clusterizzazione degli indici di consumo di suolo all'interno del dataset. L'analisi dei cluster è una tecnica di Data Mining non supervisionata che ha come obiettivo principale la suddivisione dei dati in gruppi il più coesi e omogenei possibile, con istanze che presentano caratteristiche simili. Poiché si tratta di una tecnica non supervisionata, il cluster cerca di catturare la struttura naturale dei dati originali senza l'utilizzo di alcun dato esterno.

L'obiettivo è ottenere cluster in modo da minimizzare la distanza all'interno dei gruppi e massimizzare, invece, la distanza tra i gruppi stessi, in modo che gli oggetti appartenenti a cluster diversi siano il più possibile differenziati.

## 4.1 Cluster Gerarchico

I metodi di clustering gerarchico sono finalizzati alla creazione di una struttura gerarchica che riflette la similarità o la dissimilarità dei dati, consentendo di esplorare il dataset a diversi livelli di dettaglio e comprendere le relazioni tra i punti. I metodi gerarchici utilizzano una procedura per passi che consente il cambiamento dei raggruppamenti di un solo individuo o gruppo alla volta. Questo approccio produce una sequenza completa di raggruppamenti e preserva l'inclusione dei gruppi nei passi successivi. Inoltre, l'assegnazione delle unità ai cluster è stabile e non viene modificata nelle fasi successive del processo di clustering.

Durante l'esecuzione delle tecniche di clustering gerarchico, è possibile applicare diverse misure di distanza o similarità. Tra le misure di distanza comuni utilizzate nell'analisi dei cluster, si includono la distanza euclidea, la distanza di Manhattan, la distanza di Ward, la correlazione di Pearson e la distanza di coseno. La scelta della misura di distanza dipende dal tipo di dati e dalle caratteristiche specifiche del problema di clustering. Nel presente lavoro, sono state utilizzate le tecniche del metodo Ward, applicando sia la distanza di Canberra che la distanza euclidea. Dopo aver ottenuto i risultati, è stata presa la decisione di utilizzare il metodo di Ward in combinazione con la distanza di Canberra.

### 4.1.1 Metodo Ward

Il metodo di Ward, noto anche come metodo di legame di Ward (Ward's linkage), prende il nome da Joe H. Ward Jr., uno statistico americano. Il metodo di Ward è uno dei metodi di aggregazione gerarchica utilizzati nell'analisi cluster, che si basa sull'approccio ANOVA, cerca di minimizzare l'incremento nella variabilità, cioè, cerca di trovare una combinazione di cluster che riduca al minimo la differenza o la dissimilarità rispetto al gruppo totale. L'obiettivo è creare cluster che siano il più omogenei possibile all'interno e che presentino una bassa variabilità interna.

Per spiegare il funzionamento del metodo di Ward si utilizzano le seguenti notazioni:

si assume di aver suddiviso la popolazione in C cluster, indicati con $r$ =1,...,C, con $n_r$ unità all'interno di ciascun cluster.

La somma dei quadrati interni per il cluster r, che contiene $n_r$ unità, è calcolata come la somma dei quadrati delle deviazioni degli individui (unità) all'interno del cluster dalla media del cluster stesso (centroide). È espressa dalla formula:

$$SSW_r = \sum_{i=1}^{n_r} \sum_{j=1}^{m} (x_{ijr} - \overline{x}_{jr})^2 \quad (12)$$

dove:

- $SSW_r$ rappresenta la somma dei quadrati interni per il cluster r
- $x_{ijr}$ la realizzazione della variabile j per l'individuo i che appartiene al cluster r
- $\bar{x}_{jr}$ indica la media della variabile j calcolata su tutte le unità appartenenti al cluster r
- $\bar{x}_j$ indica la media della variabile j calcolata su tutte le unità statistiche (gli individui)

Sia *t* il cluster ottenuto unendo i cluster *r* e *s*. La dissimilarità tra i due cluster è misurata come l'incremento nella variabilità rispetto al gruppo viene calcolata come:

$$d(C_r, C_s) = SSW_t - SSW_r - SSW_s \quad (13)$$

dove:

- $SSW_t$ rappresenta la somma dei quadrati interni per il cluster t, ottenuto unendo i cluster r e s.
- $SSW_r$ rappresenta la somma dei quadrati interni per il cluster r.
- $SSW_s$ rappresenta la somma dei quadrati interni per il cluster s.

In sostanza, la dissimilarità tra due cluster misura quanto l'aggregazione dei due cluster aumenta la variabilità all'interno del gruppo combinato rispetto alla variabilità dei singoli cluster separatamente.

### 4.1.2 Distanza di Canberra

La distanza di Canberra è una misura di dissimilarità utilizzata per calcolare la distanza tra due punti in uno spazio multidimensionale. La formula della distanza di Canberra tiene conto delle differenze relative tra le coordinate dei punti e considera la somma delle differenze normalizzate tra le coordinate dei punti. (Lance & Williams, 1967) . La formula per calcolare la distanza di Canberra tra due unità con osservazioni, A=[$x_1, x_2, \ldots x_m$] e B=[$y_1, y_2, \ldots y_m$], è la seguente:

$$\mathrm{d(Canberra)(A,B)} = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i| + |y_i|} \qquad (14)$$

dove $|x_i - y_i|$rappresenta la differenza assoluta tra le coordinate dei punti.

### 4.1.3 Analisi del Dendrogramma per la Determinazione del Numero Ottimale di Gruppi nel Clustering Gerarchico

La decisione di suddividere il cluster in gruppi è una scelta soggettiva basata sul dendrogramma e sul livello desiderato di dettaglio nella soluzione di clustering. Il dendrogramma viene creato utilizzando il metodo di collegamento di Ward con la metrica di distanza di Canberra.

Metodo del gomito (l'Elbow Method) è una tecnica utilizzata per determinare il numero ottimale di cluster da utilizzare in un'analisi di clustering. Questo metodo prende il nome dalla forma del grafico ottenuto, che assomiglia a un gomito o un angolo. L'obiettivo del metodo del gomito è identificare il punto in cui si verifica tale curvatura significativa, indicando il numero ottimale di cluster. Questo punto corrisponde al "gomito" nel grafico.

Questo grafico metodo del gomito visualizza le distanze (altezze) alle quali i cluster vengono uniti durante il processo di clustering gerarchico. Nel complesso,

questo consente di visualizzare la fusione dei cluster a diverse distanze (altezze) e evidenzia una specifica altezza (5.5 in questo caso) con una linea tratteggiata rossa, il che può aiutare nella determinazione del numero di cluster o del livello desiderato di dettaglio nel clustering. Nel presente caso, il numero ottimale di gruppi si situa tra 3 e 4, ed è stato scelto di effettuare il clustering con 4 gruppi.

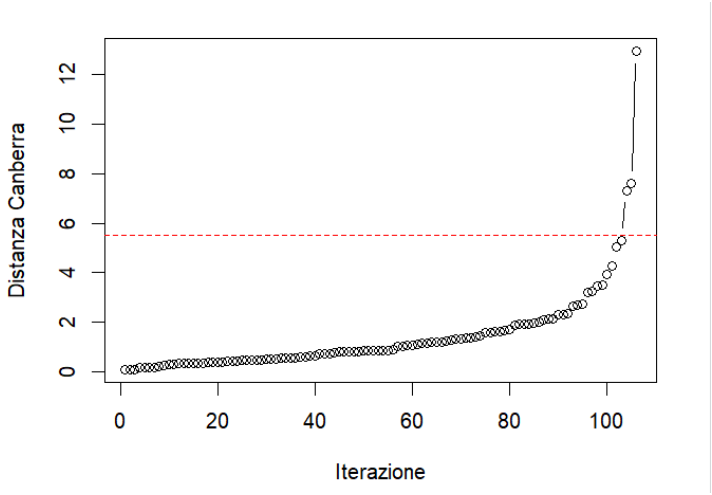


Figura 4.1 Visualizzazione delle Altezze di Fusione dei Cluster nel Processo di Clustering Gerarchico

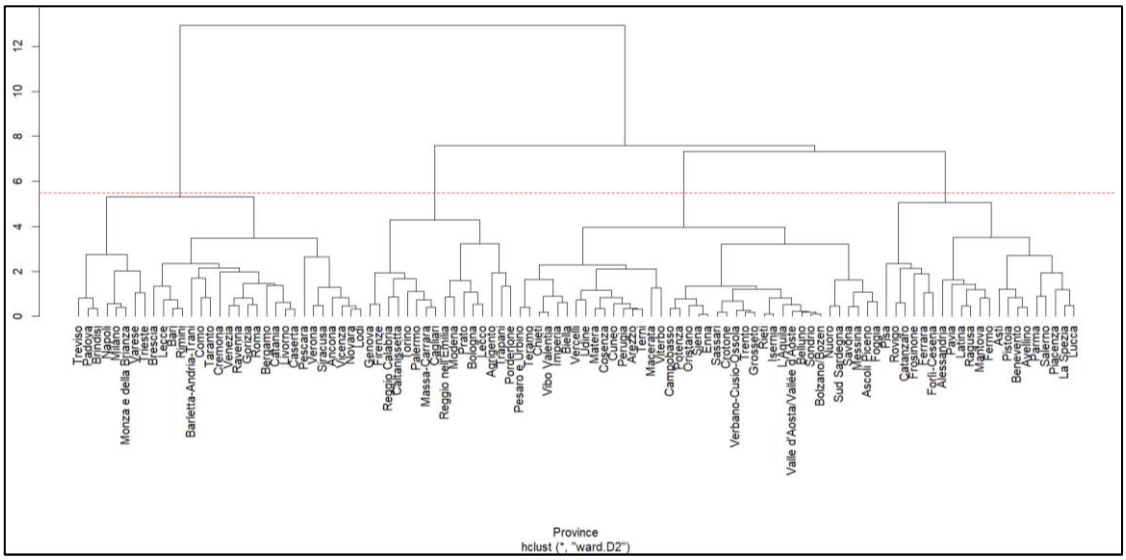


Figura 4.2 Dendrogramma con Taglio all'Altezza 5.5 per la Visualizzazione del Numero Ottimale di Gruppi

## 4.2 Applicazione delle tecniche di clustering e confronto dei risultati

La clusterizzazione dei dati è stata eseguita utilizzando tre diverse approcci su tre diversi dataset relativi agli anni 2012, 2016 e 2020, ottenendo sempre quattro gruppi di cluster. Nel primo approccio, sono state considerate tre variabili originali: CSUOLO9, DISECO3 e FORMET5. Nel secondo approccio, sono state utilizzate le medie geometriche classiche delle variabili. Nel terzo approccio, sono state utilizzate le medie geometriche penalizzate delle tre variabili originali. L'obiettivo di questa analisi è valutare quale dei tre approcci sia in grado di fornire la migliore clusterizzazione dei dati.

**Cluster fatto con Tre Variabili Originali anno 2012**

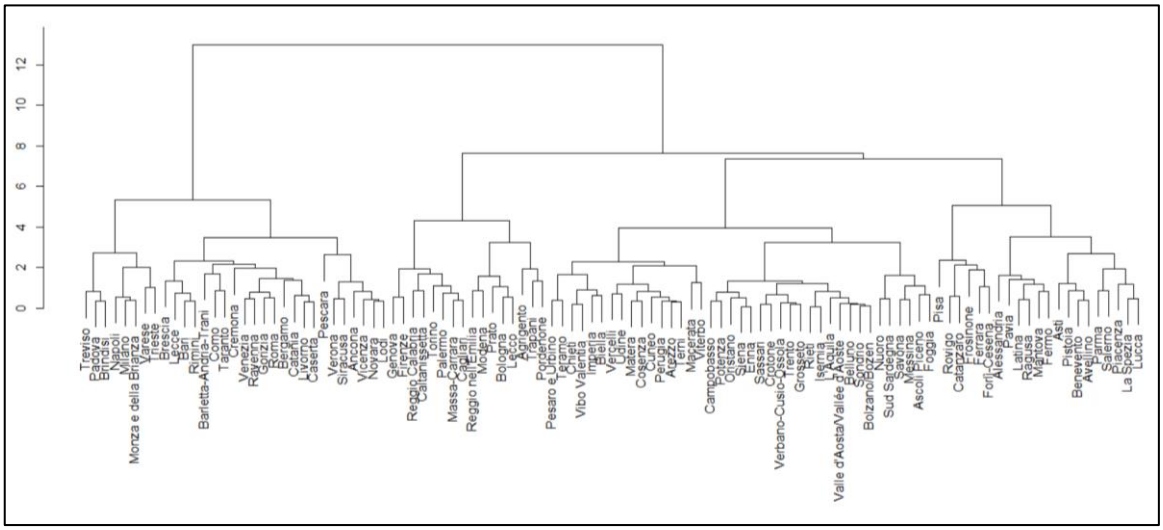


Figura 4.3 Dendrogramma Cluster Tre Variabili Originali 2012

**Cluster fatto con una Variabile Media Geometrica Classica 2012**

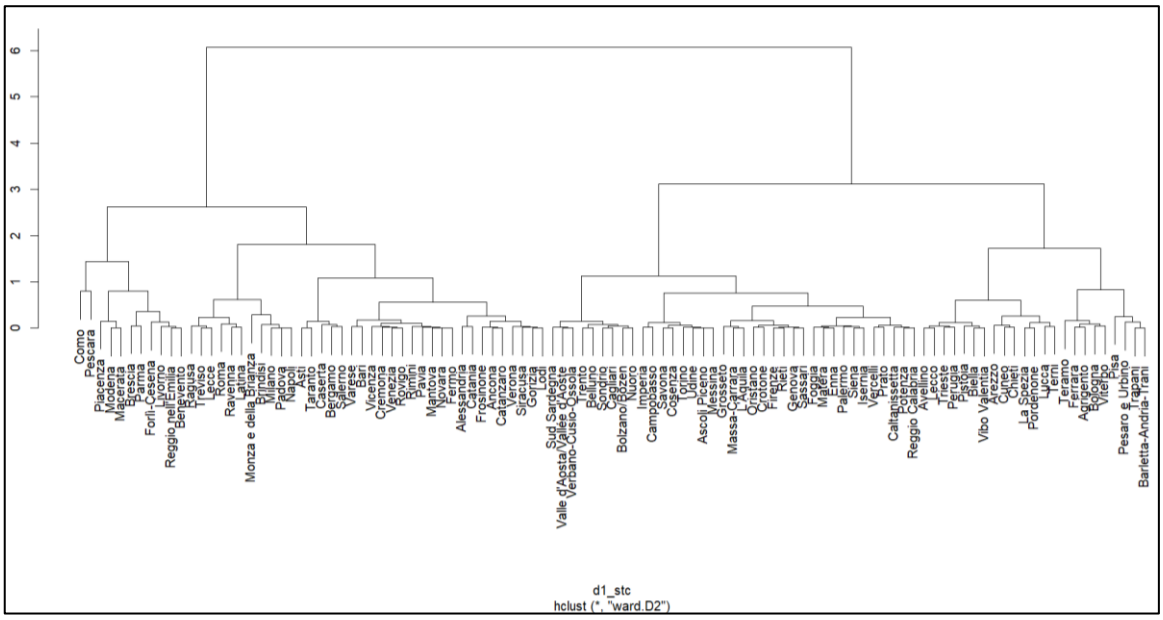


Figura 4.4 Dendrogramma Cluster con una Variabile Media Geometrica Classica 2012

**Cluster fatto con una Variabile Media Geometrica Penalizzata 2012**

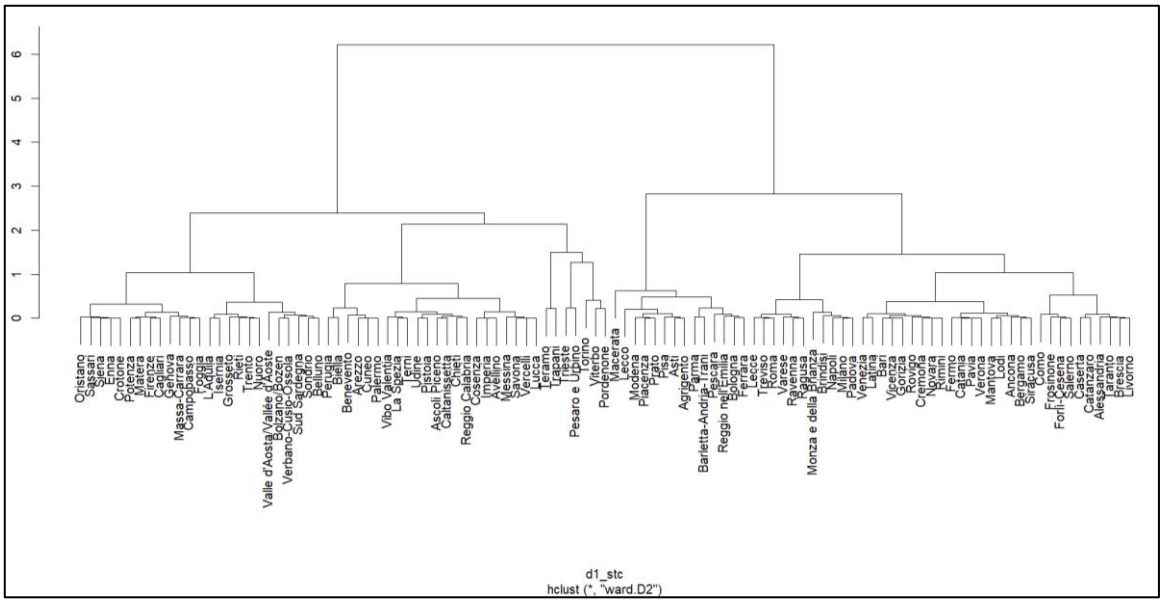


Figura 4.5 Dendrogramma Cluster con una Variabile Media Geometrica Penalizzata 2012

Tabella 4.1 Tabella dei Principali Indicatori dei cluster del 2012 utilizzando 3 approcci distinti

| Gruppi_3VR_12 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 3 | 31 | 40.46 | 48.94 | 78.74 |
| 4 | 21 | 24.46 | 40.06 | 89.25 |
| 1 | 16 | 14.76 | 35.91 | 80.90 |
| 2 | 39 | 10.94 | 26.10 | 90.23 |
| **Total** | **107** | **22.71** | **36.92** | **85.31** |

| Gruppi_MC_2012 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 2 | 36 | 40.78 | 47.87 | 82.21 |
| 4 | 11 | 22.65 | 39.25 | 85.17 |
| 3 | 23 | 16.65 | 36.36 | 86.97 |
| 1 | 37 | 8.93 | 25.93 | 87.35 |
| **Total** | **107** | **22.71** | **36.92** | **85.31** |

| Gruppi_MP_12 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 2 | 39 | 40.10 | 47.03 | 81.81 |
| 3 | 14 | 21.07 | 38.87 | 83.86 |
| 1 | 29 | 14.27 | 34.27 | 87.50 |
| 4 | 25 | 6.31 | 23.14 | 89.05 |
| **Total** | **107** | **22.71** | **36.92** | **85.31** |

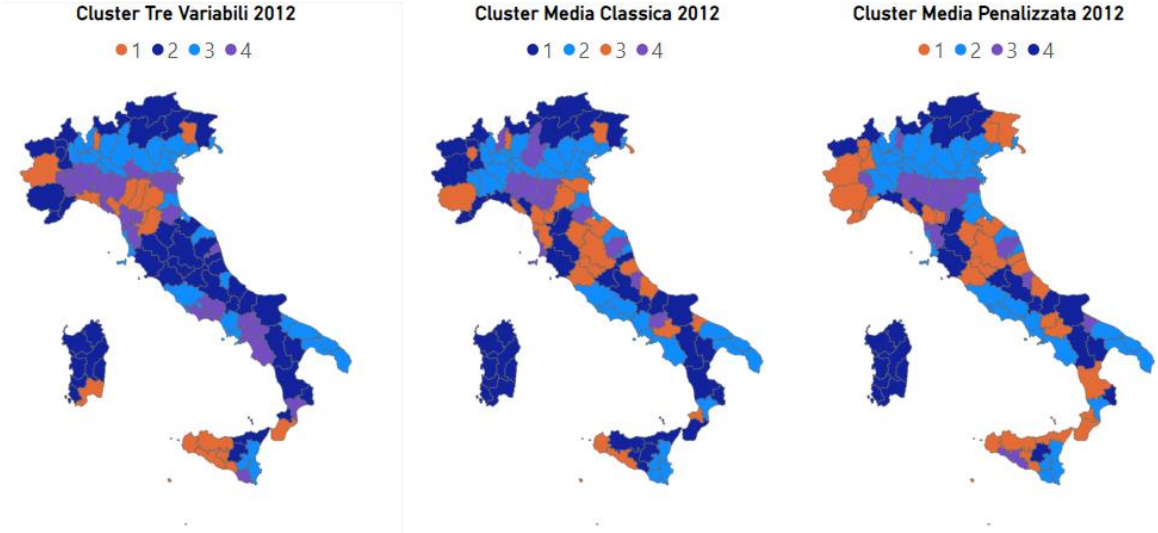


Figura 4.6 Mappe dei cluster del 2012 utilizzando 3 approcci distinti

I gruppi possono essere suddivisi in base alle caratteristiche del consumo di suolo:

1. Gruppo “Sostenibile” di colore Blu Scuro (corrispondente al gruppo numero 2 nel caso di cluster Tre Variabili, gruppo numero 1 nel caso di cluster Media classica e numero 4 nel caso di media penalizzata): rappresenta un utilizzo del suolo sostenibile, che tiene conto dell'equilibrio tra le esigenze umane e la conservazione dell'ambiente.
2. Gruppo “Soddisfacente” di colore Arancione (corrispondente al gruppo numero 1 nel caso di cluster Tre Variabili, gruppo numero 3 nel caso di cluster Media classica e numero 1 nel caso di media penalizzata): indica un livello di consumo di suolo adeguato.
3. Gruppo “Insoddisfacente” di colore Viola (corrispondente al gruppo numero 4 nel caso di cluster Tre Variabili, gruppo numero 4 nel caso di cluster Media classica e numero 3 nel caso di media penalizzata) si riferisce a un consumo di suolo insufficiente o inadeguato.
4. Gruppo “Eccessivo” di colore Blu Chiaro (corrispondente al gruppo numero 3 nel caso di cluster Tre Variabili, gruppo numero 2 nel caso di

cluster Media classica e numero 2 nel caso di media penalizzata) denota un consumo di suolo che supera significativamente i livelli sostenibili e può causare una serie di conseguenze negative come la perdita di habitat naturali, la frammentazione del territorio, l'impermeabilizzazione del suolo e la distruzione delle risorse ambientali.

Dall'analisi condotta utilizzando la media penalizzata come approccio di clustering, emerge che il gruppo con consumo di suolo "Eccessivo" presenta un maggior numero di elementi rispetto agli altri approcci di clustering. Al contrario, il gruppo con consumo di suolo "Sostenibile" mostra un minor numero di elementi, caratterizzato da un valore medio più basso del parametro CSUOLO9. Questi risultati indicano che l'utilizzo della media penalizzata come metodo di clustering tende a raggruppare un numero più elevato di osservazioni nel gruppo con consumo di suolo "Eccessivo", evidenziando una concentrazione significativa di aree con un consumo di suolo superiore ai livelli sostenibili.

**Cluster fatto con Tre Variabili Originali 2016**

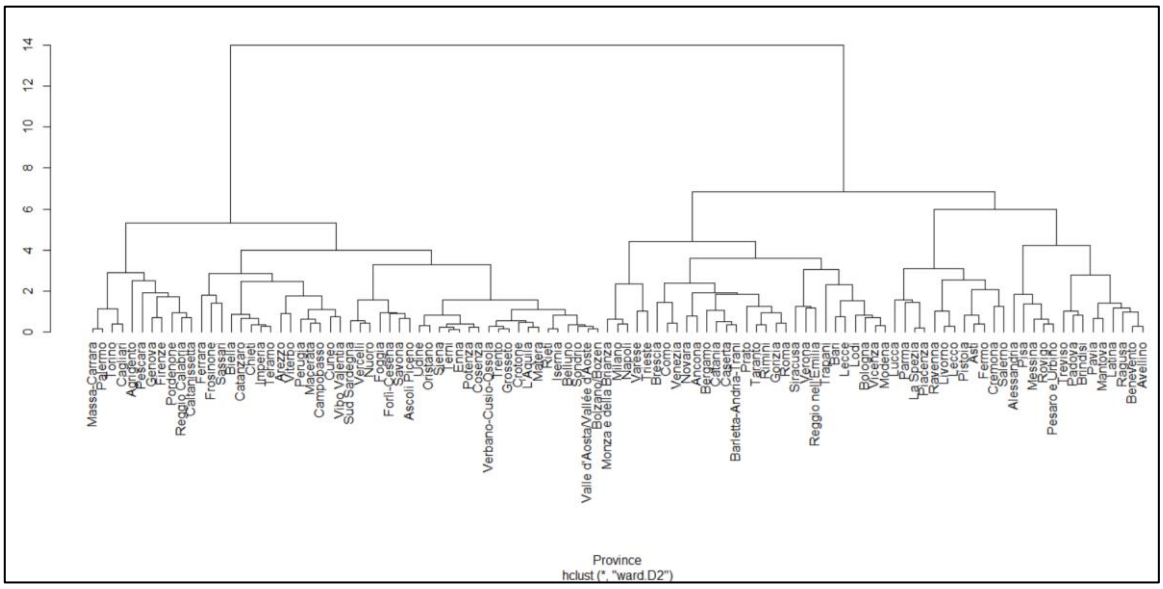


Figura 4.7 Dendrogramma Cluster con Tre Variabili Originali 2016

**Cluster fatto con una Variabile Media Geometrica Classica 2016**

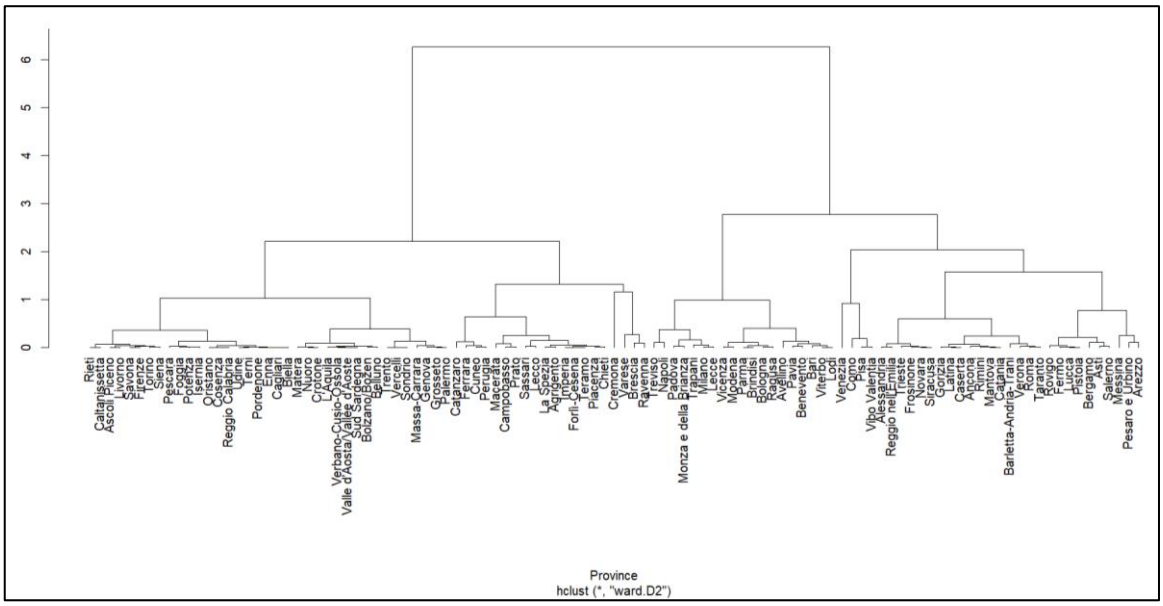


Figura 4.8 Dendrogramma Cluster con una Variabile Media Geometrica Classica 2016

**Cluster fatto con una Variabile Media Geometrica Penalizzata 2016**

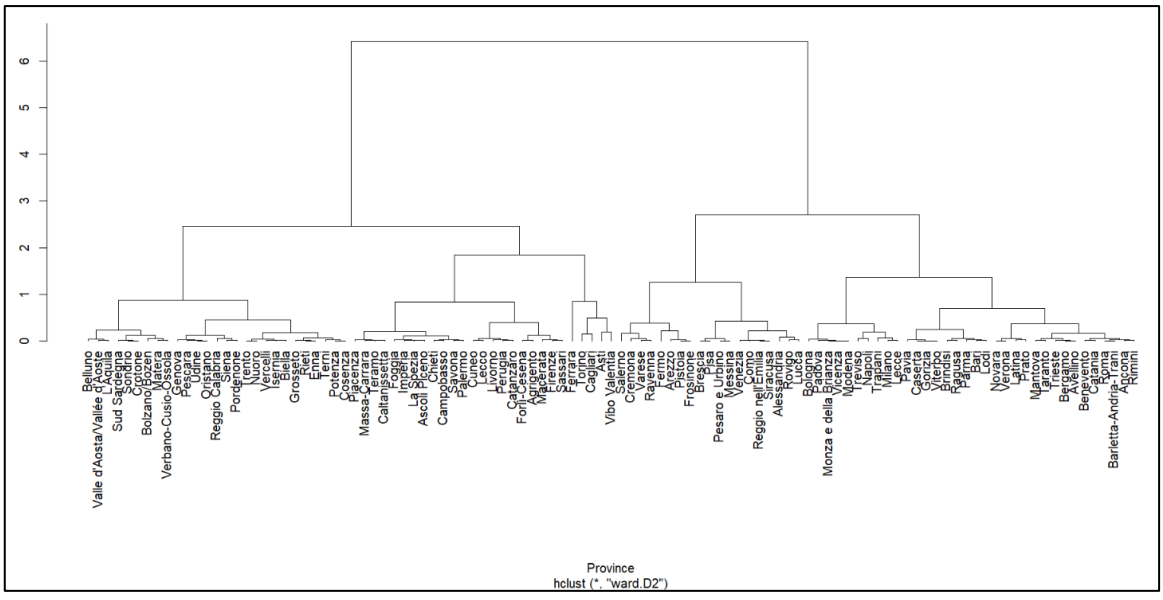


Figura 4.9 Dendrogramma Cluster con una Variabile Media Geometrica Penalizzata 2016

Tabella 4.2 Tabella dei Principali Indicatori dei cluster del 2016 utilizzando 3 approcci distinti

| Gruppi_3VR_2016 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMETS |
|---|---|---|---|---|
| 2 | 29 | 3.31 | 48.29 | 77.55 |
| 4 | 14 | 2.78 | 44.63 | 88.35 |
| 3 | 12 | 1.53 | 40.31 | 86.37 |
| 1 | 52 | 0.97 | 27.94 | 88.48 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

| Gruppi_MC_2016 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMETS |
|---|---|---|---|---|
| 4 | 19 | 4.35 | 52.38 | 81.73 |
| 2 | 31 | 2.21 | 41.10 | 84.69 |
| 3 | 20 | 1.30 | 36.09 | 85.41 |
| 1 | 37 | 0.71 | 26.24 | 87.47 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

| Gruppi_MP_2016 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMETS |
|---|---|---|---|---|
| 3 | 34 | 3.57 | 48.63 | 80.87 |
| 4 | 19 | 1.87 | 39.83 | 85.53 |
| 1 | 27 | 1.15 | 33.55 | 86.15 |
| 2 | 27 | 0.57 | 23.93 | 89.71 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

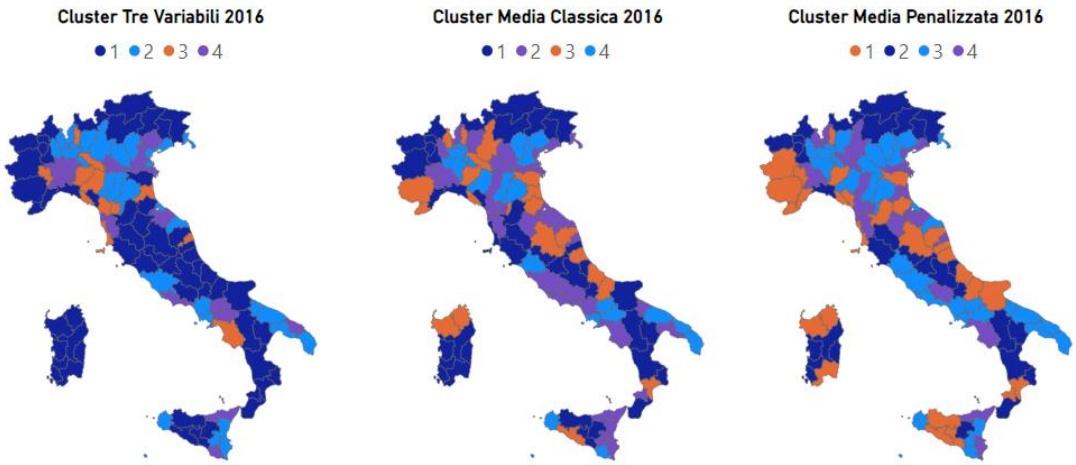


Figura 4.10 Mappe dei cluster del 2016 utilizzando 3 approcci distinti

Nel clustering del 2016, si osserva che l'utilizzo della media penalizzata come approccio di clustering porta a un gruppo con consumo di suolo "Eccessivo" (rappresentato dal colore Blu Chiaro) con un maggior numero di elementi rispetto agli altri approcci di clustering. Allo stesso tempo, sia nel clustering basato su Tre Variabili che nella media classica, il gruppo con consumo di suolo "Sostenibile" (rappresentato dal colore Blu Scuro) mostra un maggior numero di elementi. L'approccio basato sulla media penalizzata tende a raggruppare un numero inferiore di osservazioni nel gruppo con consumo di suolo sostenibile, caratterizzato da un valore medio più basso del parametro CSUOLO9. Inoltre, il clustering ottenuto con la media penalizzata distribuisce in modo più uniforme il numero di elementi tra i gruppi con consumo di suolo soddisfacente (colore arancione) e consumo insoddisfacente (gruppo di colore viola).

**Cluster fatto con Tre Variabili Originali 2020**

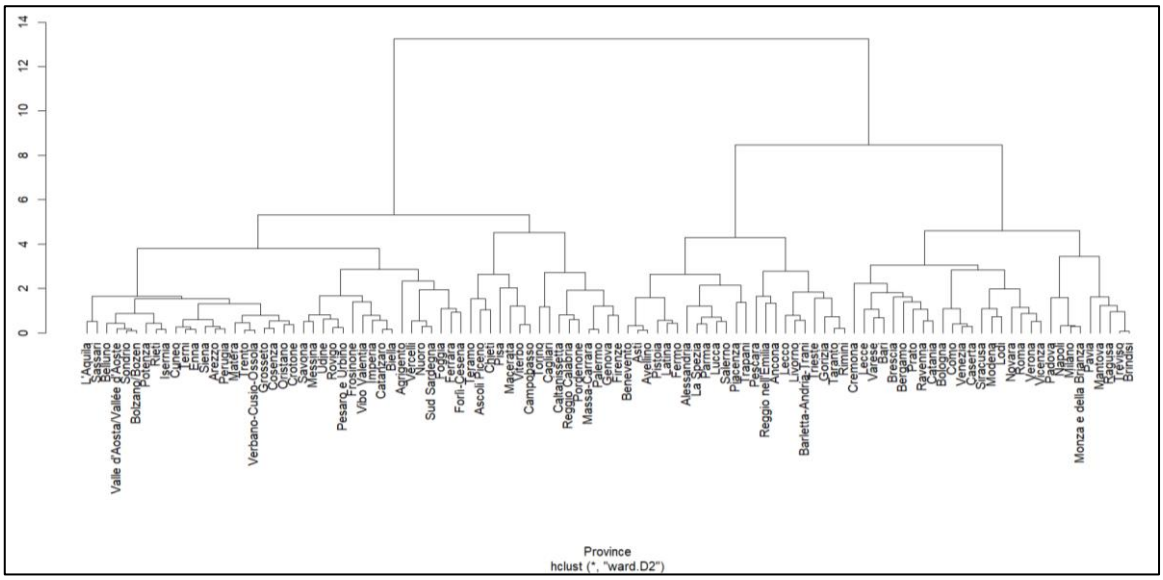


Figura 4.11 Dendrogramma Cluster con Tre Variabili Originali 2020

**Cluster con una Variabile Media Geometrica Classica 2020**

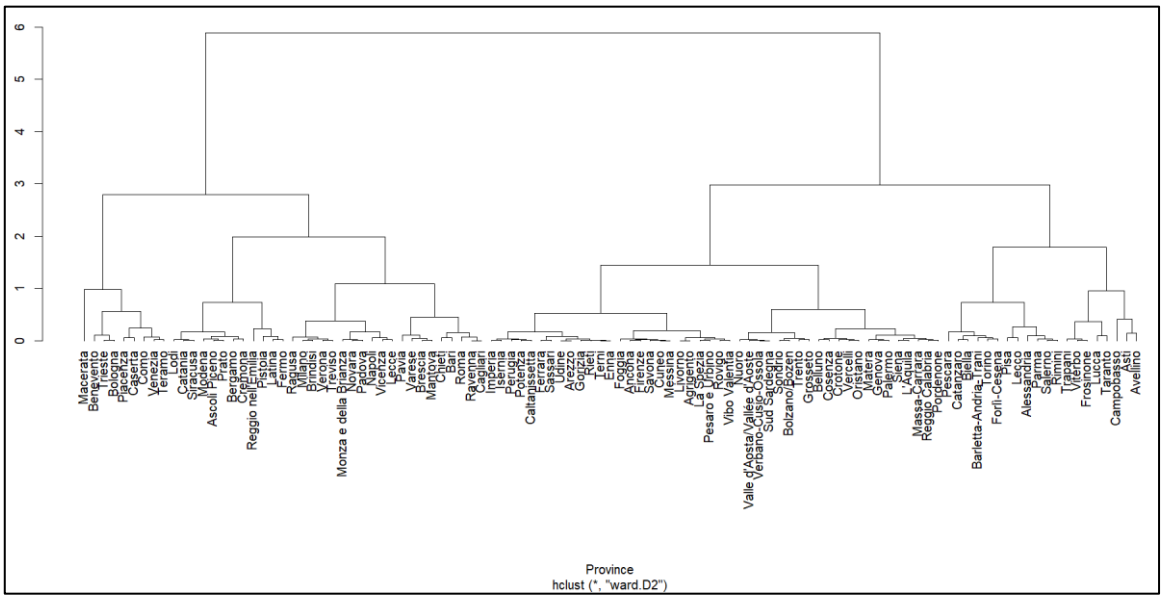


Figura 4.12 Dendrogramma Cluster con una Variabile Media Geometrica Classica 2020

**Cluster con una Variabile Media Geometrica Penalizzata 2020**

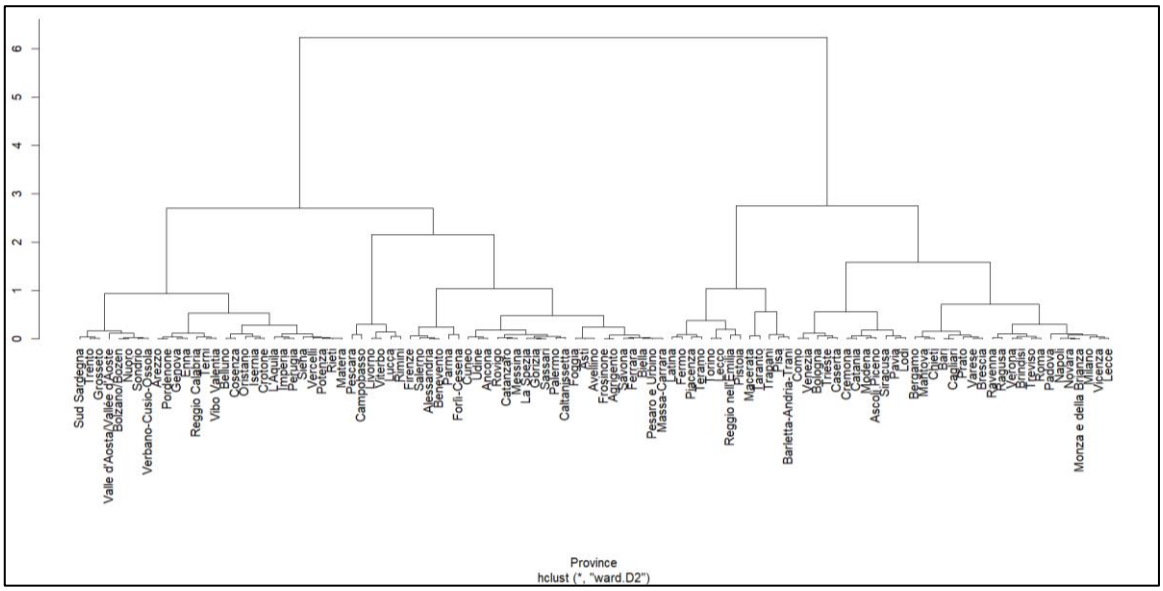


Figura 4.13 Dendrogramma Cluster con una Variabile Media Geometrica Penalizzata 2020

Tabella 4.3 Tabella dei Principali Indicatori dei cluster del 2020 utilizzando 3 approcci distinti

| Gruppi_3VR_20 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 3 | 29 | 3.30 | 49.74 | 79.14 |
| 4 | 23 | 2.13 | 41.96 | 85.37 |
| 1 | 16 | 1.26 | 32.31 | 84.82 |
| 2 | 39 | 0.99 | 26.60 | 89.92 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

| MC-20 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 3 | 32 | 2.92 | 48.52 | 80.67 |
| 4 | 9 | 2.22 | 40.77 | 83.46 |
| 1 | 20 | 2.13 | 38.00 | 87.00 |
| 2 | 46 | 1.03 | 27.88 | 88.05 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

| MP-20 | Count of Provincia | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|
| 3 | 34 | 3.57 | 48.63 | 80.87 |
| 4 | 19 | 1.87 | 39.83 | 85.53 |
| 1 | 27 | 1.15 | 33.55 | 86.15 |
| 2 | 27 | 0.57 | 23.93 | 89.71 |
| **Total** | **107** | **1.90** | **37.03** | **85.26** |

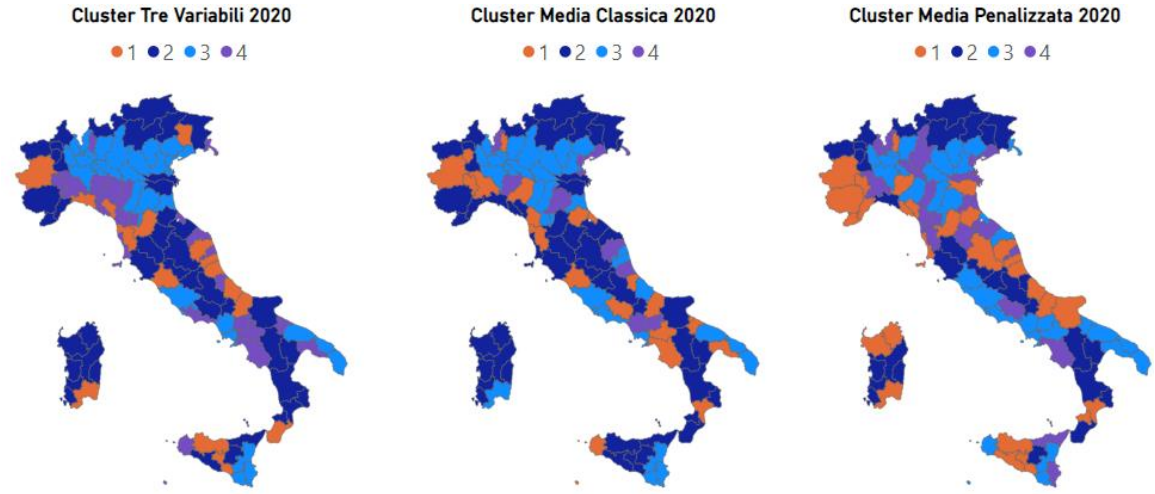


Figura 4.14 Mappe dei cluster del 2020 utilizzando 3 approcci distinti

Nel clustering del 2020, si osserva che l'utilizzo della media penalizzata come approccio di clustering porta a un gruppo con consumo di suolo "Eccessivo"

(rappresentato dal colore Blu Chiaro) con un maggior numero di elementi, mentre si tende a raggruppare un numero inferiore di osservazioni nel gruppo con consumo di suolo "Sostenibile", rispetto agli altri approcci di clustering. Allo stesso modo, sia nel clustering basato su Tre Variabili che nella media classica, il gruppo con consumo di suolo "Sostenibile" rappresentato dal colore Blu Scuro mostra un maggior numero di elementi.

### 4.2.1 Rappresentazioni tridimensionali dei dati per tre diverse tecniche di clustering

I seguenti grafici mostrano una rappresentazione tridimensionale dei dati utilizzando tre variabili: FORMET5, CSUOLO9 e DISECO3. I punti sono colorati in base ai gruppi I tre grafici rappresentano i risultati di tre diversi cluster:

1. Grafico Cluster Tre Variabili Originali
2. Grafico Cluster Media Geometrica Classica
3. Grafico Cluster Media Geometrica Penalizzata

Il dataset delle Province viene utilizzato per tracciare i grafici, selezionando le colonne FORMET5, CSUOLO9 e DISECO3 come coordinate x, y e z rispettivamente.

Nel grafico, ogni punto rappresenta una provincia, e il colore dei punti dipende dalla variabile "Group" presente nel dataset.

**Grafici tridimensionali dei cluster utilizzando tre approcci distinti per l'anno 2012**

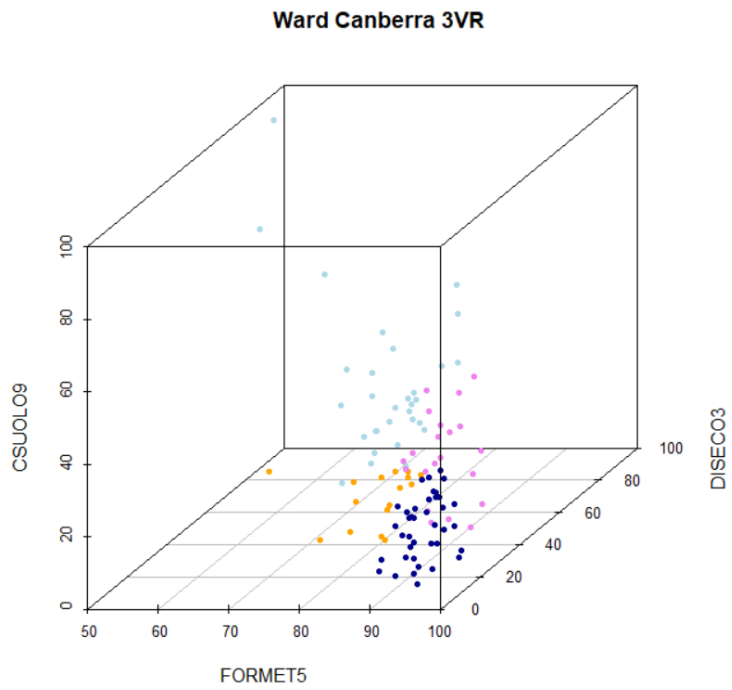


Figura 4.15 Grafico Cluster Tre Variabili Originali 2012

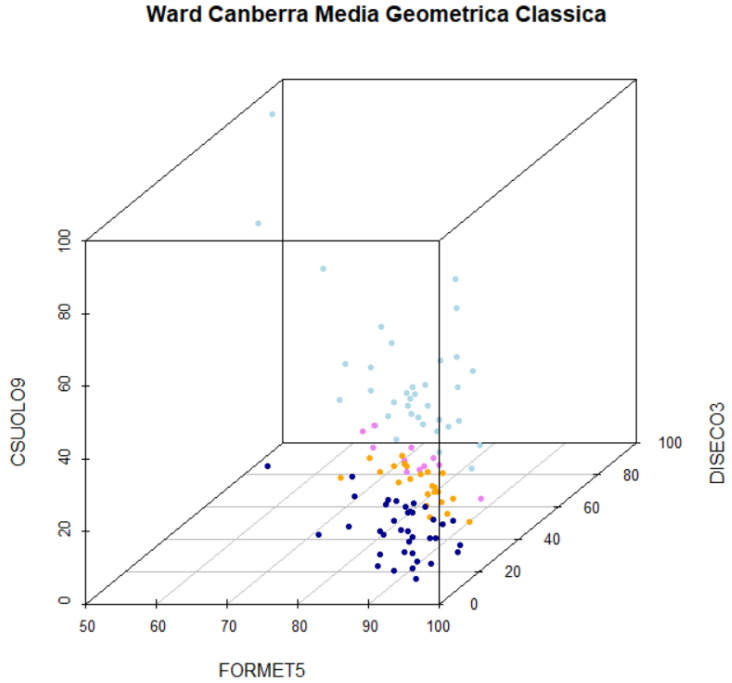


Figura 4.16 Grafico Cluster Media Geometrica Classica 2012

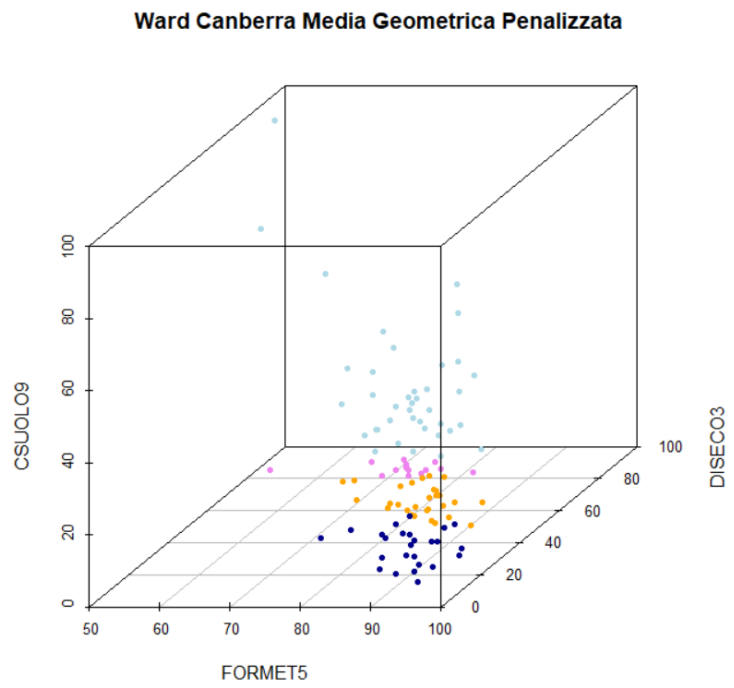


Figura 4.17 Grafico Cluster Media Geometrica Penalizzata 2012

**Grafici tridimensionali dei cluster utilizzando tre approcci distinti per l'anno 2016**

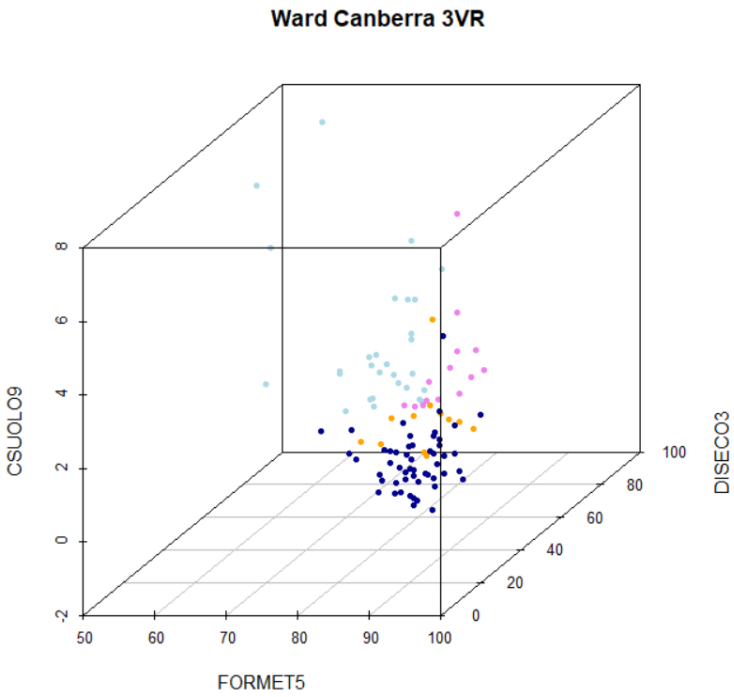


Figura 4.18 Grafico Cluster Tre Variabili Originali 2016

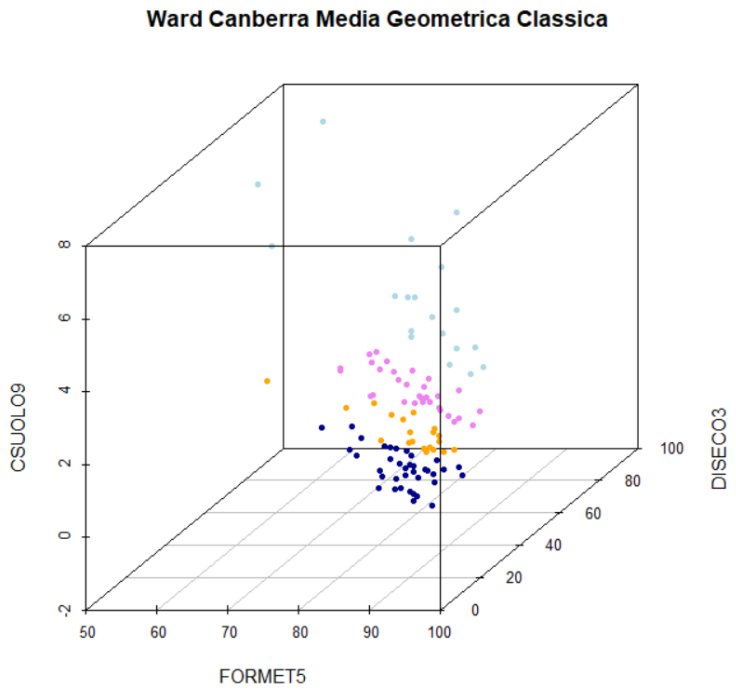


Figura 4.19 Grafico Cluster Media Geometrica Classica 2016

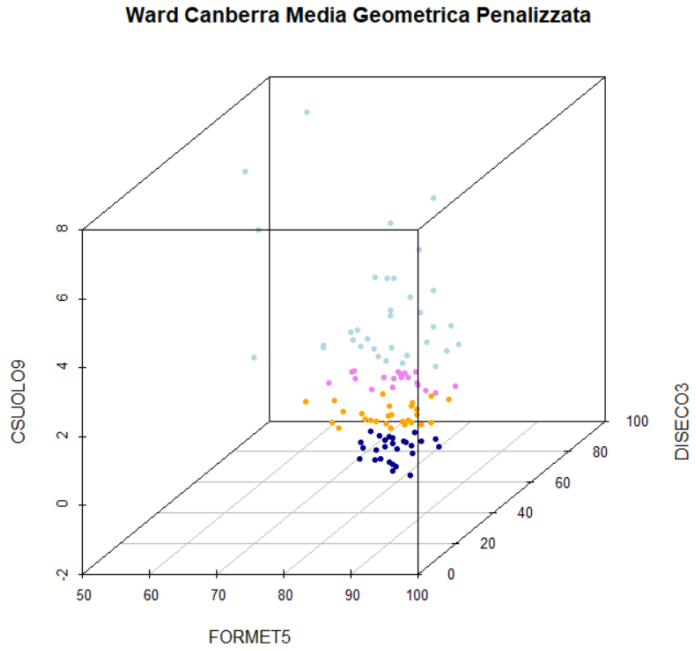


Figura 4.20 Grafico Cluster Media Geometrica Penalizzata 2016

**Grafici tridimensionali dei cluster utilizzando tre approcci distinti per l'anno 2016**

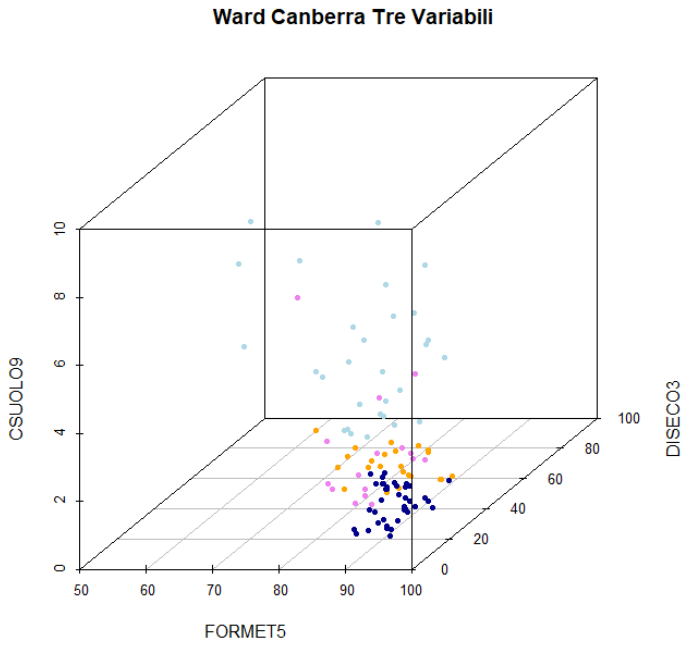


Figura 4.21 Grafico Cluster Tre Variabili Originali 2020

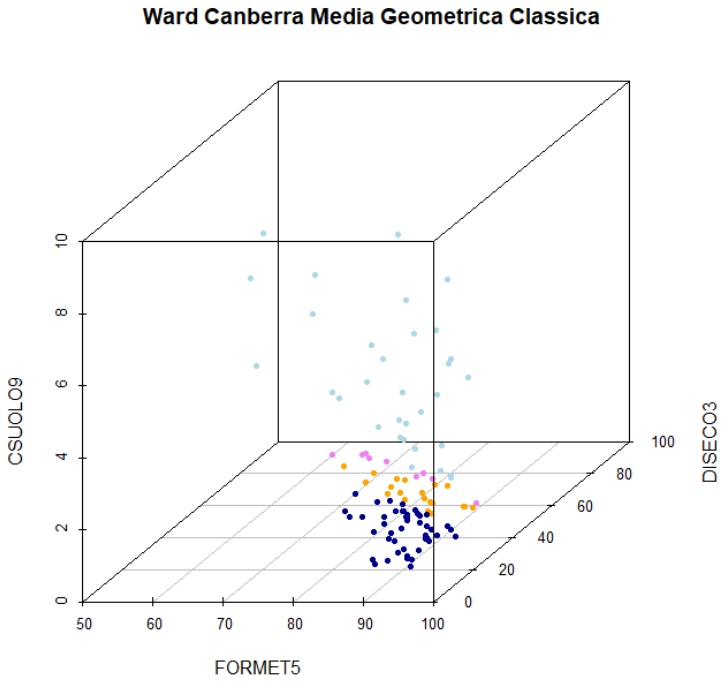


Figura 4.22 Grafico Cluster Media Geometrica Classica 2020

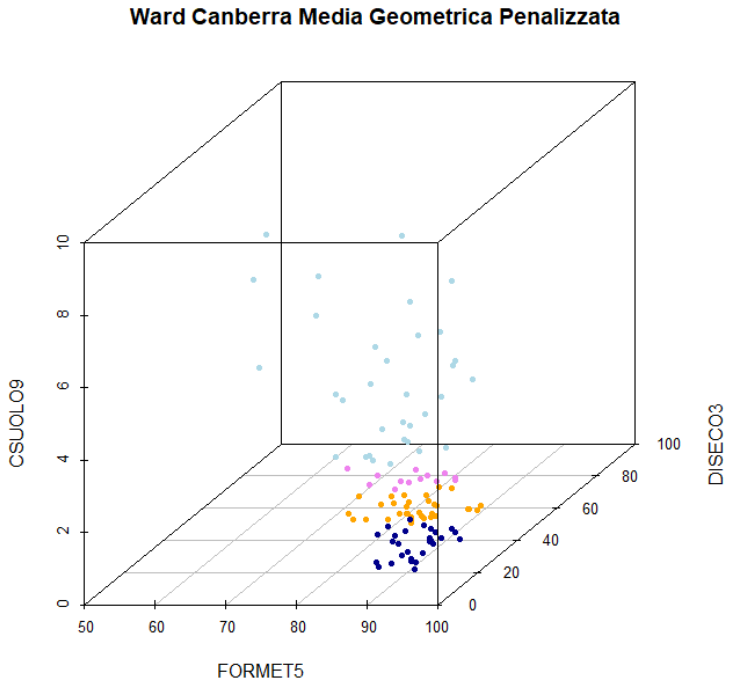


Figura 4.23 Grafico Cluster Media Geometrica Penalizzata 2020

Nel Grafico Cluster Tre Variabili Originali (Figure 4.24, 4.25, 4.26) si può osservare come gli elementi dei diversi gruppi si mescolino nel cluster. Alcuni punti si trovano all'interno della zona di un gruppo diverso, indicando errori di classificazione.

Il cluster basato sulla media geometrica classica sembra avere una clusterizzazione leggermente migliore (Figure 4.27, 4.28, 4.29); tuttavia, i punti risultano non completamente all'interno dell'area del loro gruppo, alcuni escono al di fuori e si fondono con un altro gruppo.

I cluster ottenuto utilizzando la media geometrica penalizzata (Figure 4.30, 4.31, 4.32) permette di distinguere in modo più accurato i diversi gruppi di dati, con una separazione molto omogenea e precisa. Questo evidenzia che l'indicatore

composito della media penalizzata è in grado di ottenere informazioni in modo più efficace.

Complessivamente, l'analisi dei risultati del clustering mostra una differenziazione delle province in base alle variabili considerate, rivelando diverse tipologie di consumo del suolo. L'approccio basato sulle medie geometriche delle variabili sembra offrire una migliore separazione dei cluster e una maggiore coerenza all'interno dei gruppi rispetto agli altri approcci considerati.

Alla luce di tali evidenze, si procede con l'analisi empirica utilizzando esclusivamente il cluster ottenuto con la media geometrica penalizzata.

## 4.3 Analisi dei cluster basata sulla Media Geometrica Penalizzata. Esplorazione delle dinamiche nel corso del tempo

Nel presente capitolo, è stata condotta un'analisi dei cluster utilizzando la media penalizzata su dati relativi a diversi anni. Nel contesto dell'analisi condotta, è stato applicato il metodo di clustering Ward utilizzando la distanza di Canberra. Questo approccio consente di identificare modelli e relazioni tra le osservazioni nel corso del tempo, rivelando possibili cambiamenti o tendenze significative. L'utilizzo della media penalizzata nell'analisi dei cluster contribuisce a ottenere risultati più precisi e coerenti, fornendo una migliore comprensione delle dinamiche sottostanti presenti nei dati durante il periodo considerato.

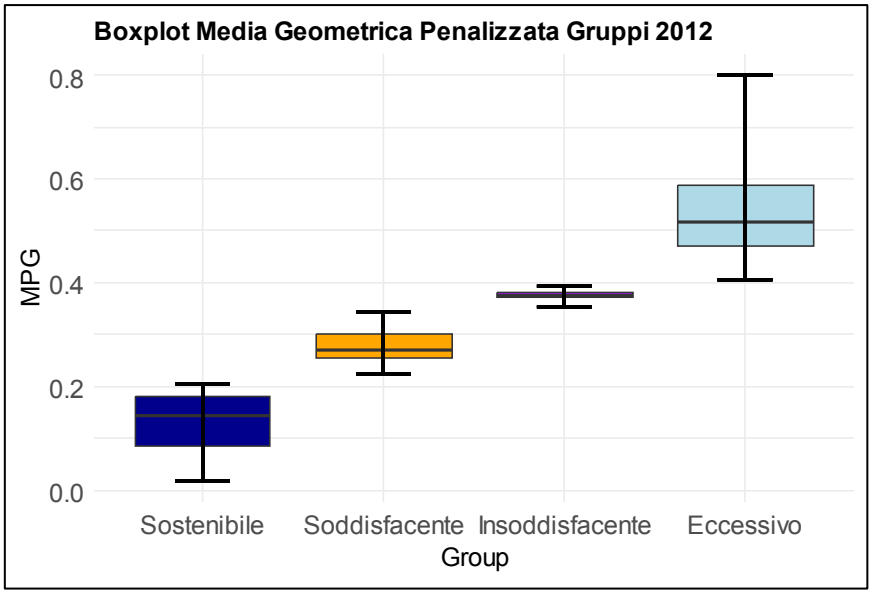


Figura 4.24 Box plot Gruppi Cluster Media Geometrica Penalizzata 2012

Tabella 4.4 Tabella dei valori minimi e massimi del consumo di suolo per ciascun gruppo 2012

| Consumo 2012 | Gruppi | MP_geometric_min | MP_geometric_max | MP_geometric_median | MP_geometric_mean |
|---|---|---|---|---|---|
| Sostenibile | 4 | 0.02 | 0.20 | 0.14 | 0.13 |
| Soddisfacente | 1 | 0.22 | 0.34 | 0.27 | 0.28 |
| Insoddisfacente | 3 | 0.35 | 0.39 | 0.37 | 0.38 |
| Eccessivo | 2 | 0.41 | 0.80 | 0.52 | 0.53 |

I valori minimi e massimi rappresentano delle soglie che possono essere utilizzate per distinguere le province in base al consumo di suolo.


Figura 4.25 Mappa Cluster Media Geometrica Penalizzata 2012

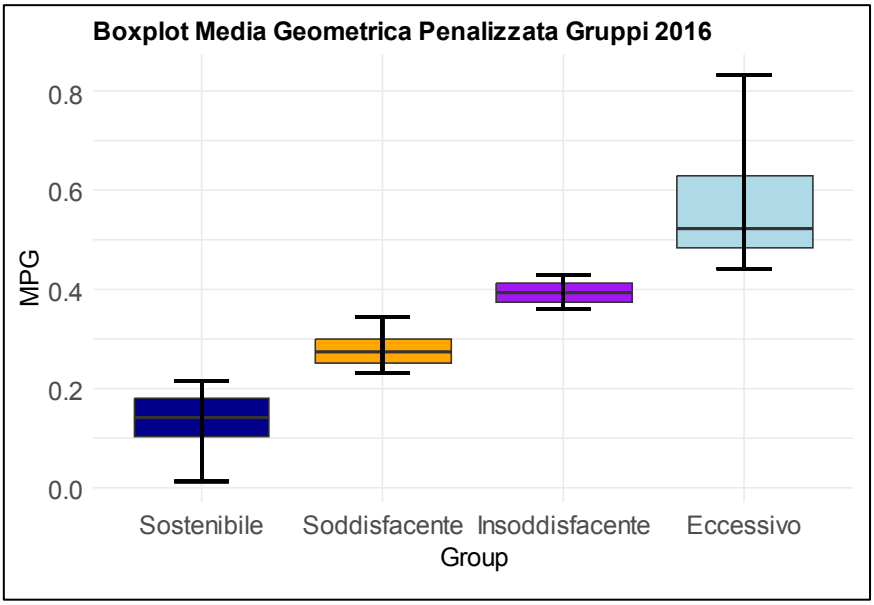


Figura 4.26 Box plot Gruppi Cluster Media Geometrica Penalizzata 2016

I valori minimi e massimi per ciascun gruppo rappresentano delle soglie che possono essere utilizzate per distinguere le province in base al consumo di suolo.

Tabella 4.5 Tabella dei valori minimi e massimi del consumo di suolo per ciascun gruppo 2016

| Consumo 2016 | Gruppi | MP_geometric_min | MP_geometric_max | MP_geometric_median | MP_geometric_mean |
|---|---|---|---|---|---|
| Sostenibile | 2 | 0.01 | 0.21 | 0.14 | 0.13 |
| Soddisfacente | 1 | 0.23 | 0.34 | 0.27 | 0.28 |
| Insoddisfacente | 4 | 0.36 | 0.43 | 0.39 | 0.39 |
| Eccessivo | 3 | 0.44 | 0.83 | 0.52 | 0.56 |


Figura 4.27 Mappa Cluster Media Geometrica Penalizzata 2016

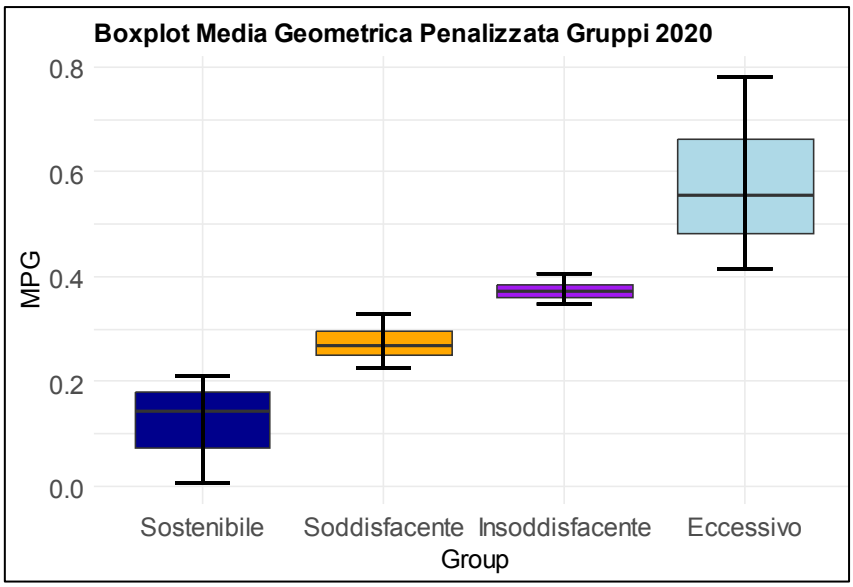


Figura 4.28 Box plot Gruppi Cluster Media Geometrica Penalizzata 2020

Tabella 4.6 Tabella dei valori minimi e massimi del consumo di suolo per ciascun gruppo 2020

| Consumo 2020 | Gruppi | MP_geometric_min | MP_geometric_max | MP_geometric_median | MP_geometric_mean |
|---|---|---|---|---|---|
| Sostenibile | 2 | 0.01 | 0.21 | 0.14 | 0.12 |
| Soddisfacente | 4 | 0.23 | 0.33 | 0.27 | 0.28 |
| Insoddisfacente | 1 | 0.35 | 0.40 | 0.37 | 0.37 |
| Eccessivo | 3 | 0.42 | 0.78 | 0.56 | 0.57 |

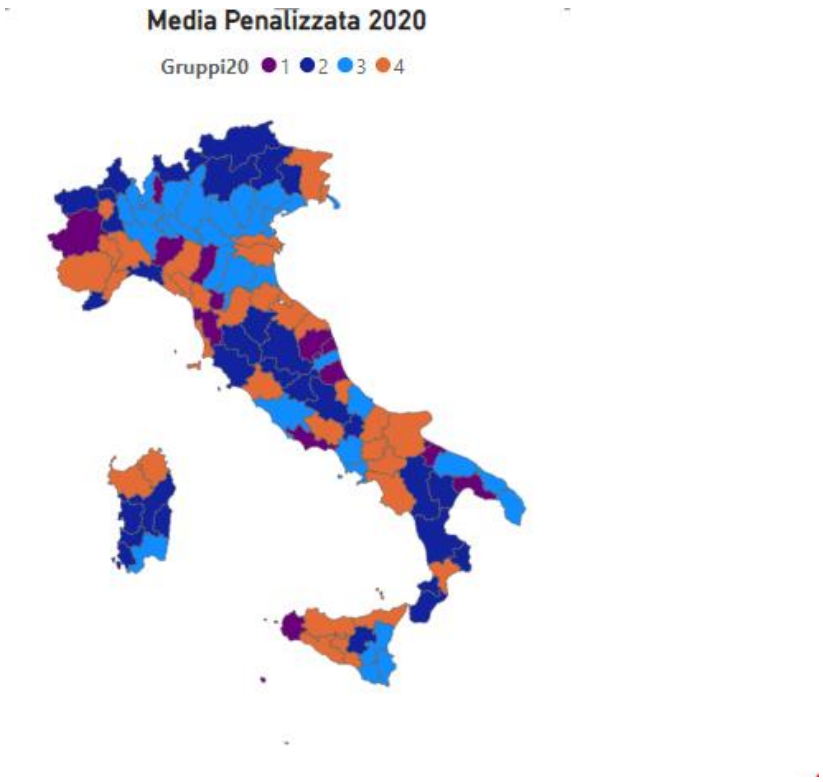


Figura 4.29 Mappa Cluster Media Geometrica Penalizzata 2020

### 4.3.1 Analisi della distribuzione delle aziende agricole e manifatturiere nei gruppi con consumo di suolo nel corso degli anni 2012, 2016 e 2020

Nella presente tabella vengono riportati i risultati per i tre anni 2012, 2016 e 2020 suddivisi in base al Gruppo con Consumo di Suolo. Le colonne "Ratio_Agricole" e "Ratio_Manifatturiere" rappresentano rispettivamente la proporzione delle aziende agricole e delle aziende manifatturiere per ciascun gruppo. I rapporti indicano la proporzione di "Aziende_Agricole" e "Aziende_Manifatturiere" all'interno di ogni gruppo.

Il rapporto è un modo per esprimere il contributo relativo o la percentuale di una quantità rispetto al totale di più quantità. In questo caso, i rapporti indicano la proporzione di "Aziende_Agricole" e "Aziende_Manifatturiere" all'interno di ciascun gruppo, rispetto al loro totale combinato.

Tabella 4.7 Tabella dei rapporti di Aziende Agricole e Aziende Manifatturiere nei diversi gruppi

| Gruppo con Consumo di Suolo | | 2012 | | 2016 | | 2020 | |
|---|---|---|---|---|---|---|---|
| | | Ratio Agricole | Ratio Manifatturiere | Ratio Agricole | Ratio Manifatturiere | Ratio Agricole | Ratio Manifatturiere |
| Sostenibile | | 0.052 | 0.948 | 0.063 | 0.937 | 0.071 | 0.929 |
| Sodisfacente | | 0.040 | 0.960 | 0.034 | 0.966 | 0.037 | 0.963 |
| Insodisfacente | | 0.015 | 0.985 | 0.024 | 0.976 | 0.019 | 0.981 |
| Eccesivo | | 0.013 | 0.987 | 0.014 | 0.986 | 0.012 | 0.988 |

Nell'analisi condotta, emerge chiaramente che la presenza delle aziende agricole è inferiore rispetto alle aziende manifatturiere. Questa disparità è particolarmente evidente nel confronto dei rapporti delle aziende agricole tra i diversi gruppi, rispetto ai rapporti delle aziende manifatturiere. Si osserva una variazione relativamente minima nel rapporto delle aziende manifatturiere tra i diversi gruppi, indicando una presenza più significativa di attività manifatturiere. Questo dato suggerisce che l'attività manifatturiera è più diffusa rispetto all'attività agricola all'interno dei gruppi presi in considerazione.

Si può affermare che le province con una maggiore presenza di attività manifatturiere mostrano un consumo di suolo più elevato. Al contrario, una maggiore presenza di aziende agricole è associata a un minor consumo di suolo.

## 4.4 Analisi delle dinamiche di consumo di suolo e sviluppo urbano nei gruppi intersecanti durante tre anni distinti

Nel contesto specifico dell'analisi dei dati di diversi anni, è possibile osservare le intersezioni tra i cluster. Ad esempio, le province possono trovarsi nello stesso gruppo sia nel cluster basato sulla media geometrica penalizzata del 2012 che nel cluster del 2016. Inoltre, vengono confrontati con i cluster del 2016 con quelli del 2020.

Durante l'analisi delle intersezioni, viene calcolata la lunghezza dell'intersezione tra i gruppi dei cluster di diversi anni. In questo modo, viene individuato il gruppo con la massima intersezione. Questo calcolo permette di identificare il gruppo con la massima intersezione, evidenziando le province che mantengono una similarità nel raggruppamento tra i diversi anni.

L'analisi delle intersezioni dei cluster permette quindi di comprendere come le province si associano tra loro nel corso del tempo.

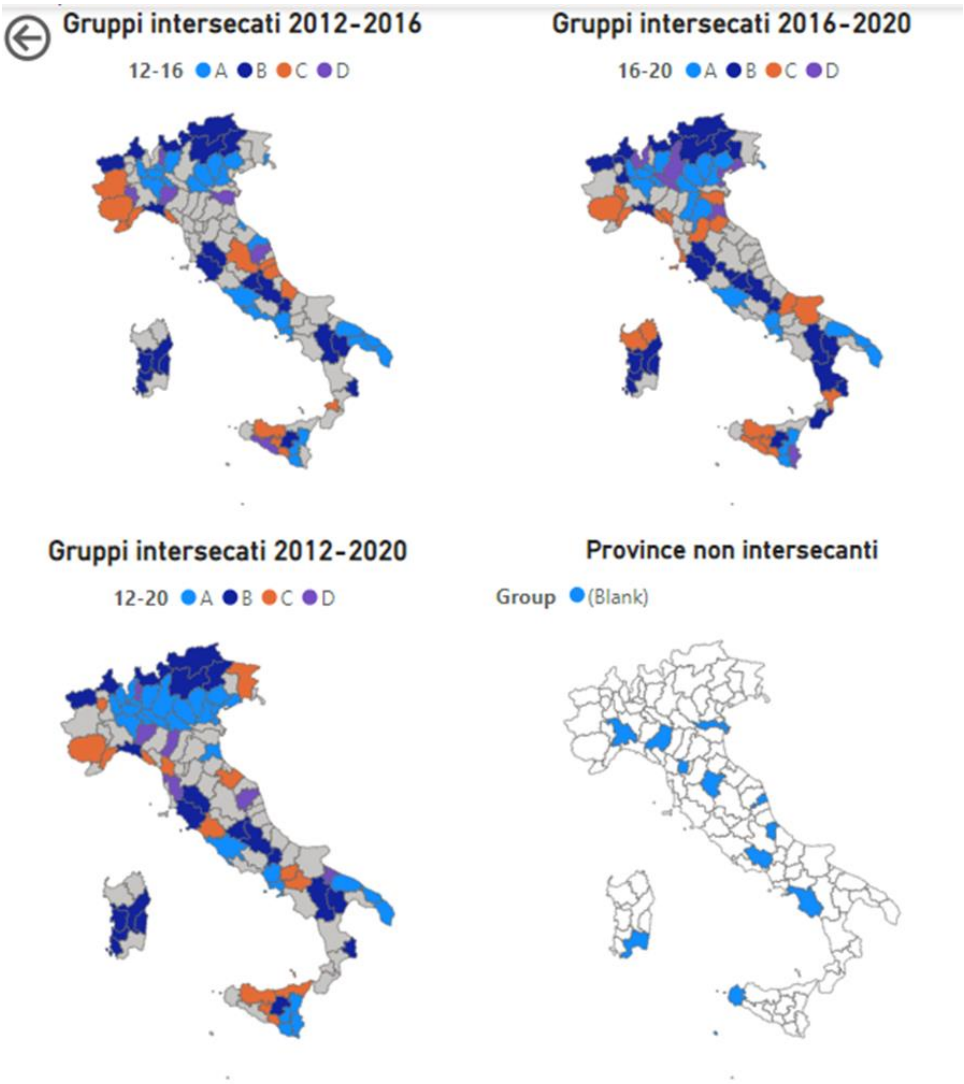


Figura 4.30 Intersezioni dei cluster nel corso degli anni

Tabella 4.8 Medie delle variabili originali nei gruppi intersecanti durante tre anni 2012, 2016, 2020

| Gruppo | Province | Densità di consumo di suolo rispetto all'area totale | Indice di Dispersione Urbana | Superficie impattata dalla presenza di coperture artificiali considerando una distanza (buffer) di 60 metri |
|---|---|---|---|---|
| A | 19 | 4.58 | 78.99 | 53.40 |
| B | 19 | 0.63 | 90.59 | 21.21 |
| C | 5 | 0.96 | 85.72 | 32.34 |
| D | 2 | 1.81 | 84.71 | 39.68 |

Analizzando la media delle variabili per i tre anni considerati e i gruppi che intersecano tra di essi, possiamo distinguere i seguenti indicatori principali:

Densità di consumo di suolo rispetto all'area totale: La media più elevata si osserva nel Gruppo A, seguita dal Gruppo D, Gruppo C e infine dal Gruppo B. Ciò indica che il Gruppo A ha il più alto tasso di urbanizzazione e utilizzo del suolo, mentre il Gruppo B ha il tasso più basso.

Indice di Dispersione Urbana: La media più alta si registra nel Gruppo B, seguita dal Gruppo C, Gruppo D e infine dal Gruppo A. Questo indica che il Gruppo B ha una distribuzione urbana più concentrata, mentre il Gruppo A ha una distribuzione urbana relativamente più equilibrata.

Superficie impattata dalla presenza di coperture artificiali considerando una distanza di 60 metri: La media più elevata si osserva nel Gruppo A, seguita dal Gruppo D, Gruppo C e infine dal Gruppo B. Ciò indica che il Gruppo A ha la maggiore superficie coperta da infrastrutture artificiali, mentre il Gruppo B ha la minore superficie impattata.

Questi indicatori evidenziano le differenze nei modelli di consumo di suolo e sviluppo urbano tra i gruppi nel corso dei tre anni considerati. Il Gruppo A mostra una tendenza verso una maggiore urbanizzazione e impatto sul territorio, mentre il Gruppo B ha una minore urbanizzazione e impatto. I Gruppi C e D presentano valori intermedi rispetto ai primi due gruppi.

### 4.4.1 Gruppo A

Nell'analisi delle intersezioni emergono 19 province con una stabilità di appartenenza al gruppo.

Tabella 4.9 Gruppo A. Principali Indicatori

| Gruppo 2012 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| A | 19 | 0.58 | 45.67 | 53.07 | 79.35 |
| Total | 19 | 0.58 | 45.67 | 53.07 | 79.35 |

| Gruppo 2016 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| A | 19 | 0.58 | 3.77 | 53.23 | 79.26 |
| Total | 19 | 0.58 | 3.77 | 53.23 | 79.26 |

| Gruppo 2020 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| A | 19 | 0.62 | 4.58 | 53.40 | 78.99 |
| Total | 19 | 0.62 | 4.58 | 53.40 | 78.99 |

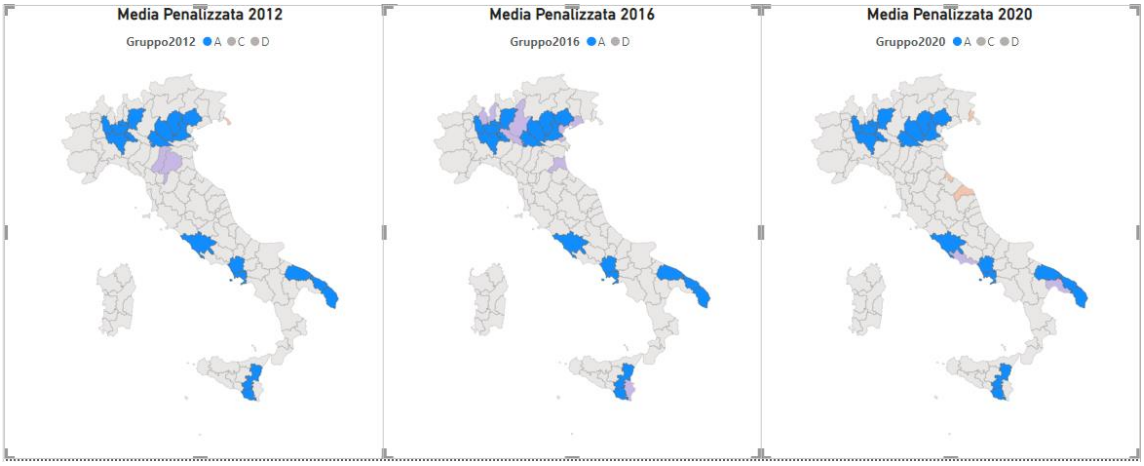


Figura 4.31 Intersezioni tra Province del Gruppo A

Le province menzionate non sono tutte vicine geograficamente tra loro. Molte di queste province hanno una popolazione numerosa. Questo indica una concentrazione significativa di persone che contribuiscono alla vitalità e alla dinamicità di queste aree. Inoltre, queste molte di queste province si distinguono per la diversificazione delle attività economiche.

Tabella 4.10 Gruppo A - Coerenza nel tempo e classifica delle province

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Monza e della Brianza | A | 107 | A | 101 | A | 106 |
| 2 | Brindisi | A | 106 | A | 93 | A | 99 |
| 3 | Milano | A | 105 | A | 105 | A | 101 |
| 4 | Padova | A | 104 | A | 102 | A | 107 |
| 5 | Napoli | A | 103 | A | 107 | A | 104 |
| 6 | Roma | A | 102 | A | 85 | A | 97 |
| 7 | Treviso | A | 101 | A | 106 | A | 98 |
| 8 | Lecce | A | 100 | A | 104 | A | 103 |
| 9 | Ragusa | A | 99 | A | 94 | A | 96 |
| 10 | Bari | A | 94 | A | 96 | A | 93 |
| 11 | Vicenza | A | 93 | A | 100 | A | 102 |
| 12 | Novara | A | 90 | A | 76 | A | 105 |
| 13 | Lodi | A | 87 | A | 95 | A | 82 |
| 14 | Mantova | A | 86 | A | 78 | A | 87 |
| 15 | Bergamo | A | 85 | A | 81 | A | 89 |
| 16 | Catania | A | 82 | A | 88 | A | 85 |
| 17 | Pavia | A | 81 | A | 89 | A | 81 |
| 18 | Verona | A | 81 | A | 77 | A | 100 |
| 19 | Caserta | A | 78 | A | 90 | A | 77 |

Le province appartenenti al gruppo A si distinguono per il rank più elevato rispetto alle province appartenenti agli altri gruppi. Questo indica che tali province presentano un consumo di suolo particolarmente elevato. Monza e della Brianza, Brindisi, Milano, Padova, Napoli, Roma, Treviso e Lecce occupano costantemente le ultime posizioni nel ranking nel corso degli anni considerati. Questa situazione evidenzia una significativa pressione sul suolo e una perdita costante di aree non urbanizzate in queste province.

Tabella 4.11 Le dinamiche di cambiamento all'interno del Gruppo A

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Ravenna | A | 99 | D | 59 | A | 95 |
| 2 | Varese | A | 97 | D | 60 | A | 90 |
| 3 | Venezia | A | 96 | D | 70 | A | 78 |
| 4 | Latina | A | 95 | A | 74 | D | 69 |
| 5 | Gorizia | A | 92 | A | 92 | C | 35 |
| 6 | Cremona | A | 91 | D | 61 | A | 86 |
| 7 | Rimini | A | 89 | A | 84 | C | 58 |
| 8 | Rovigo | A | 88 | D | 68 | C | 39 |
| 9 | Siracusa | A | 84 | D | 72 | A | 80 |
| 10 | Ancona | A | 83 | A | 83 | C | 31 |
| 11 | Fermo | A | 79 | D | 55 | D | 70 |
| 12 | Catanzaro | A | 77 | C | 48 | C | 38 |
| 13 | Brescia | A | 76 | D | 66 | A | 91 |
| 14 | Livorno | A | 75 | C | 49 | C | 60 |
| 15 | Taranto | A | 74 | A | 79 | D | 63 |
| 16 | Alessandria | A | 73 | D | 69 | C | 52 |
| 17 | Como | A | 72 | D | 71 | A | 79 |
| 18 | Frosinone | A | 71 | D | 58 | C | 46 |
| 19 | Forlì-Cesena | A | 70 | C | 44 | C | 54 |
| 20 | Salerno | A | 70 | D | 62 | C | 50 |
| 21 | Barletta-Andria-Trani | D | 68 | A | 86 | D | 64 |
| 22 | Parma | D | 67 | A | 97 | C | 55 |
| 23 | Bologna | D | 65 | A | 98 | A | 75 |
| 24 | Prato | D | 62 | A | 75 | A | 92 |
| 25 | Modena | D | 60 | A | 99 | A | 83 |
| 26 | Trapani | C | 54 | A | 103 | D | 66 |
| 27 | Trieste | C | 52 | A | 80 | A | 76 |
| 28 | Viterbo | C | 49 | A | 92 | C | 61 |
| 29 | Benevento | C | 45 | A | 87 | C | 52 |
| 30 | Chieti | C | 38 | C | 31 | A | 88 |
| 31 | Ascoli Piceno | C | 35 | C | 34 | A | 84 |
| 32 | Avellino | C | 27 | A | 82 | C | 48 |
| 33 | Cagliari | B | 20 | C | 50 | A | 94 |

Analizzando la tabella, si può notare che le province presentano cambiamenti significativi nel corso del tempo riguardo all'appartenenza ai diversi gruppi. Alcune province mostrano una maggiore variabilità nel loro gruppo di appartenenza. Ad esempio, Barletta-Andria-Trani ha alternato il suo status dal Gruppo D al Gruppo A e successivamente è ritornata al Gruppo D, mentre Trapani è passata dal Gruppo C al Gruppo A e infine al Gruppo D. Altre province che hanno sperimentato cambiamenti significativi includono Venezia, Latina, Gorizia, Cremona, Rimini, Rovigo, Siracusa e Ancona, le quali hanno cambiato gruppo almeno una volta nel corso degli anni considerati.

In generale, si può osservare che le province del gruppo A hanno mostrato una certa variabilità nel loro status di gruppo nel corso del tempo, spostandosi occasionalmente verso i gruppi D e C. Queste dinamiche evidenziano processi di cambiamento e adattamento, in cui alcune province sono state in grado di ridurre temporaneamente la loro pressione sul suolo e di transitare nei gruppi D o C. Tale variabilità suggerisce la possibilità di adottare interventi e politiche volte a ridurre il consumo di suolo e promuovere un utilizzo sostenibile delle risorse territoriali.

**Box plot delle Variabili Originali. Gruppo A**

**CSUOLO9 (Densità di consumo di suolo rispetto all'area totale)**

Dal confronto dei valori, si osserva una variazione significativa nel parametro CSUOLO9 nel gruppo A nel corso degli anni. Nel 2012, si registrano valori più elevati con una media di 40.10 e un valore massimo di 98.36. Nel 2016 e nel 2020, i valori diminuiscono notevolmente con medie di 3.57 e 4.08 rispettivamente. Si osserva anche una riduzione nei valori massimi e una minore dispersione dei dati, come evidenziato dai quartili inferiori e superiori più bassi. Dal confronto dei valori, possiamo notare un aumento del consumo di suolo nel gruppo A tra il 2016 e il 2020. La media del consumo di suolo è aumentata da 3.57 nel 2016 a 4.08 nel 2020. La mediana, che rappresenta il valore di mezzo, è leggermente aumentata da 3.00 a 4.05. Inoltre, sia il valore minimo che il valore massimo sono aumentati nel 2020 rispetto al 2016.

I quartili mostrano che c'è stata una maggiore distribuzione dei dati nel 2020 rispetto al 2016. Il primo quartile è passato da 2.53 nel 2016 a 2.57 nel 2020, mentre il terzo quartile è aumentato da 4.38 a 5.21. Sono stati rilevati outliers nel

2016 Monza e della Brianza, Milano, Brindisi e Napoli nel 2016, rappresentando valori anomali nella distribuzione dei dati.

In sintesi, i dati indicano un aumento complessivo del consumo di suolo nel gruppo A tra il 2016 e il 2020, con una maggiore variabilità nei valori nel 2020.

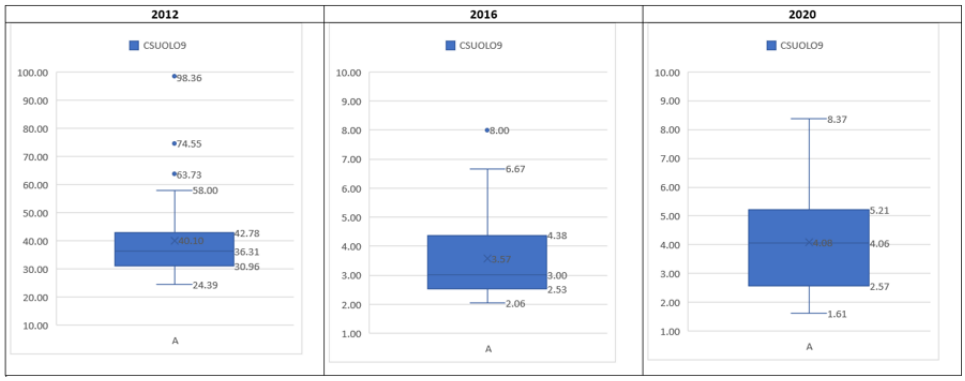


Figura 4.32 Box Plot della variabile CSUOLO9 nel Gruppo A

**DISECO3 (Superficie impattata dalla presenza di coperture artificiali)**

Anno 2012: Nel gruppo A, il parametro DISECO3 presenta una media di 47.03 e una mediana di 42.21. I valori variano da 35.12 a 82.21, con una distribuzione compatta intorno ai quartili 39.88 e 52.94. Sono stati rilevati due outliers con valori 82.21 Monza e della Brianza e 77.32 Napoli.

Anno 2016: Nel gruppo A, il parametro DISECO3 registra una media di 48.63 e una mediana di 45.24. I valori oscillano tra 30.89 e 82.38, con una distribuzione compatta intorno ai quartili 41.16 e 54.12. Sono stati rilevati due outliers con valori 82.38 Monza e della Brianza e 77.51 Napoli.

Anno 2020: Nel gruppo A, il parametro DISECO3 mostra una leggera diminuzione, con una media di 48.50 e una mediana di 46.02. I valori si estendono da 30.09 a 82.45, mantenendo una distribuzione compatta intorno ai quartili 41.09

e 55.74. Sono stati rilevati due outliers con valori 82.45 Monza e della Brianza e 77.64 Napoli.

In conclusione, nel gruppo A si osserva una stabilità nel parametro DISECO3 nel corso degli anni. Sono stati rilevati outliers in tutti e tre gli anni, rappresentando valori anomali nella distribuzione dei dati.

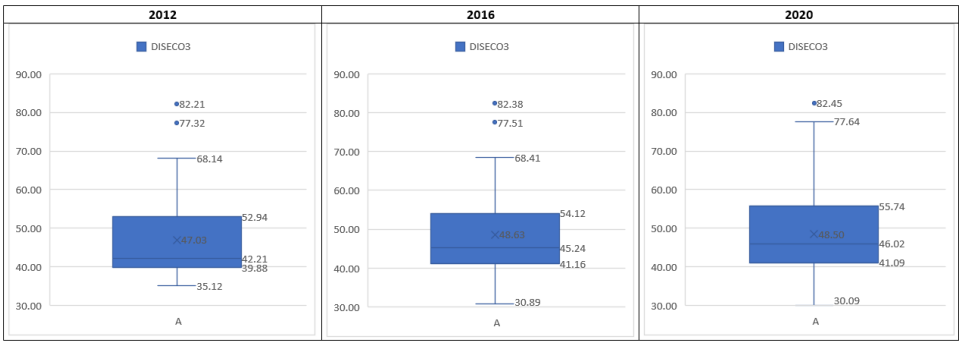


Figura 4.33 Box Plot della variabile DISECO3 nel Gruppo A

**FORMET5 (Indice di Dispersione Urbana)**

Anno 2012: Nel gruppo A, il parametro FORMET5 presenta una media di 81.81 e una mediana di 83.83. I valori variano da 53.54 a 95.48, con una distribuzione compatta intorno ai quartili 79.54 e 86.82. Sono stati rilevati tre outliers con valori 55.42 Milano, 53.54 Monza e della Brianza e 62.15 Napoli.

Anno 2016: Nel gruppo A, il parametro FORMET5 registra una media di 80.87 e una mediana di 83.12. I valori oscillano tra 53.27 e 93.96, con una distribuzione compatta intorno ai quartili 79.18 e 86.55. Sono stati rilevati quattro outliers con valori 55.17 Milano, 53.26 Monza e della Brianza, 61.92 Napoli e 63.73 Prato.

Anno 2020: Nel gruppo A, il parametro FORMET5 mostra una diminuzione, con una media di 79.03 e una mediana di 81.70. I valori si estendono da 52.83 a 90.93, mantenendo una distribuzione compatta intorno ai quartili 76.55 e 85.20. Sono

stati rilevati quattro outliers con valori 54.80 Milano, 52.83 Monza e della Brianza, 61.53 Napoli e 62.91 Prato.

In conclusione, nel gruppo A si osserva una leggera diminuzione nel parametro FORMET5 nel corso degli anni, passando da una media di 81.81 nel 2012 a 79.03 nel 2020. La distribuzione dei dati rimane compatta intorno ai quartili, indicando una certa stabilità nelle caratteristiche del parametro. Sono stati rilevati outliers in tutti e tre gli anni, rappresentando valori anomali nella distribuzione dei dati.

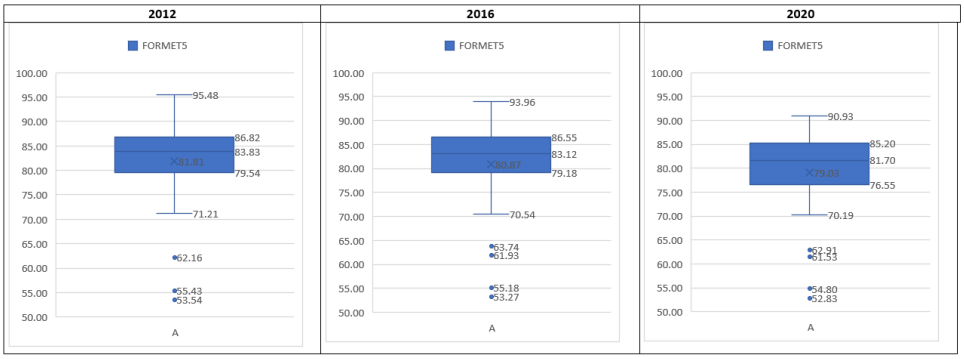


Figura 4.34 Box Plot della variabile FORMET5 nel Gruppo A

## 4.4.2 Gruppo B

Tabella 4.12 Cluster Media Penalizzata 2012, 2016, 2020. Tabella dei Principali Indicatori

| Gruppo 2012 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| B | 19 | 0.07 | 5.90 | 21.08 | 90.65 |
| Total | 19 | 0.07 | 5.90 | 21.08 | 90.65 |

| Gruppo 2016 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| B | 19 | 0.09 | 0.60 | 21.15 | 90.66 |
| Total | 19 | 0.09 | 0.60 | 21.15 | 90.66 |

| Gruppo 2020 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| B | 19 | 0.09 | 0.63 | 21.21 | 90.59 |
| Total | 19 | 0.09 | 0.63 | 21.21 | 90.59 |


Figura 4.35 Intersezioni tra Province del Gruppo B

Le province del gruppo B sono caratterizzate da una presenza geografica comune nelle regioni montuose. Questo paesaggio montagnoso può influenzare l'agricoltura, l'economia locale e la densità della popolazione. Ad esempio, l'agricoltura intensiva potrebbe essere limitata a causa delle caratteristiche montuose del territorio. La presenza di zone montuose può influire sulla densità della popolazione, con una popolazione potenzialmente più ridotta rispetto ad aree pianeggianti.

Tabella 4.13 Gruppo B con coerenza nel tempo e classifica delle province

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Genova | B | 22 | B | 21 | B | 25 |
| 2 | Matera | B | 19 | B | 6 | B | 16 |
| 3 | Potenza | B | 19 | B | 19 | B | 18 |
| 4 | Crotone | B | 16 | B | 8 | B | 12 |
| 5 | Enna | B | 16 | B | 16 | B | 24 |
| 6 | Siena | B | 14 | B | 25 | B | 14 |
| 7 | Oristano | B | 13 | B | 22 | B | 11 |
| 8 | Isernia | B | 12 | B | 12 | B | 13 |
| 9 | L'Aquila | B | 12 | B | 3 | B | 21 |
| 10 | Grosseto | B | 10 | B | 15 | B | 7 |
| 11 | Rieti | B | 9 | B | 17 | B | 17 |
| 12 | Trento | B | 8 | B | 11 | B | 8 |
| 13 | Nuoro | B | 7 | B | 10 | B | 5 |
| 14 | Sud Sardegna | B | 6 | B | 7 | B | 6 |
| 15 | Sondrio | B | 5 | B | 9 | B | 2 |
| 16 | Belluno | B | 4 | B | 1 | B | 9 |
| 17 | Bolzano/Bozen | B | 3 | B | 4 | B | 5 |
| 18 | Verbano-Cusio-Ossola | B | 2 | B | 5 | B | 3 |
| 19 | Valle d'Aosta/Vallée d'Aoste | B | 1 | B | 2 | B | 1 |

Le province del gruppo B hanno mostrato indicatori migliori rispetto ad altri gruppi, indicando una gestione più sostenibile del suolo.

Le province di Trento, Nuoro, Sud Sardegna, Sondrio, Belluno, Bolzano/Bozen, Verbano-Cusio-Ossola e Valle d'Aosta/Vallée d'Aoste mantengono una posizione stabile nel gruppo B nel corso degli anni e risultano costantemente ai primi posti del rank. Queste province mostrano una buona gestione del consumo del suolo, mantenendo valori relativamente bassi della Media Penalizzata e ottenendo risultati positivi in termini di classifica. Siena mostra una variazione significativa

ma successivamente si stabilizza nel 2020. Nonostante ciò, la provincia mantiene una posizione all'interno del gruppo B nel ranking.

Altre come Sondrio e Belluno, nonostante mostrino variazioni nel corso del tempo, mantengono una posizione all'interno del gruppo B nel ranking, confermando una certa coerenza nel consumo di suolo.

Tabella 4.14 Le dinamiche di cambiamento all'interno del Gruppo B

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Pescara | D | 66 | B | 24 | C | 57 |
| 2 | Pordenone | C | 48 | B | 26 | B | 23 |
| 3 | Biella | C | 47 | B | 13 | C | 42 |
| 4 | Perugia | C | 46 | C | 47 | B | 20 |
| 5 | Arezzo | C | 44 | D | 56 | B | 23 |
| 6 | Vibo Valentia | C | 41 | C | 52 | B | 26 |
| 7 | Terni | C | 39 | B | 20 | B | 27 |
| 8 | Reggio Calabria | C | 37 | B | 27 | B | 28 |
| 9 | Udine | C | 34 | B | 23 | C | 30 |
| 10 | Vercelli | C | 30 | B | 14 | B | 15 |
| 11 | Cosenza | C | 28 | B | 19 | B | 10 |
| 12 | Imperia | C | 27 | C | 35 | B | 19 |
| 13 | Massa-Carrara | B | 25 | C | 36 | C | 44 |
| 14 | Foggia | B | 24 | C | 32 | C | 49 |
| 15 | Campobasso | B | 23 | C | 30 | C | 56 |
| 16 | Firenze | B | 21 | C | 40 | C | 53 |
| 17 | Cagliari | B | 20 | C | 50 | A | 94 |
| 18 | Sassari | B | 17 | C | 41 | C | 35 |

Analizzando le dinamiche di cambiamento dei gruppi nel corso del tempo, è possibile osservare diverse tendenze. Alcune province hanno mostrato un miglioramento nel consumo del suolo, passando a gruppi con rank più alti. Ad esempio, Pordenone è passata dal gruppo C nel 2012 con un rank di 48 al gruppo B nel 2016, mantenendo tale posizione nel 2020 con un rank di 23. Biella ha seguito un percorso diverso, passando dal gruppo C nel 2012 con un rank di 47 al gruppo B nel 2016, ma successivamente hanno peggiorato la sua posizione, raggiungendo il gruppo C nel 2020 con un rank di 42

Alcune province, come Perugia, sono rimaste nel gruppo C nel corso del tempo, ma hanno registrato un miglioramento nel rank, passando da 47 nel 2016 a 20 nel

2020. Altre province, come Arezzo, sono passate dal gruppo C nel 2012 con un rank di 44 al gruppo D nel 2016, ma successivamente hanno migliorato la loro posizione, raggiungendo il gruppo B nel 2020 con un rank di 23.

D'altra parte, alcune province hanno mostrato un peggioramento nel consumo del suolo. Ad esempio, Massa-Carrara è passata dal gruppo B nel 2012 con un rank di 25 al gruppo C nel 2016, mantenendo tale posizione nel 2020 con un rank di 44. Foggia e Campobasso hanno seguito una tendenza simile, passando dal gruppo B nel 2012 a gruppi C nel 2016 e mantenendo tali posizioni nel 2020 con rank più bassi.

**Box plot delle Variabili Originali. Gruppo B**

**CSUOLO9 (Densità di consumo di suolo rispetto all'area totale)**

Anno 2012: Nel gruppo B, il parametro CSUOLO9 presenta una media di 6.31 e una mediana di 5.52. I valori variano da 1.61 a 16.64, con una distribuzione compatta intorno ai quartili 3.82 e 9.35.

Anno 2016: Nel gruppo B, il parametro CSUOLO9 registra una media di 0.57 e una mediana di 0.58. I valori oscillano tra -0.10 e 0.94, con una distribuzione compatta intorno ai quartili 0.40 e 0.78.

Anno 2020: Nel gruppo B, il parametro CSUOLO9 mostra un aumento, con una media di 0.61 e una mediana di 0.55. I valori si estendono da 0.19 a 1.61, mantenendo una distribuzione intorno ai quartili 0.42 e 0.81. È stato rilevato un outlier con valore 1.61 L’Aquila.

In conclusione, nel gruppo B si osserva una diminuzione nel parametro CSUOLO9 nel corso degli anni, passando da una media di 6.31 nel 2012 a 0.61 nel 2020. È stato rilevato un outlier nel 2020 con valore 1.61 L'Aquila.

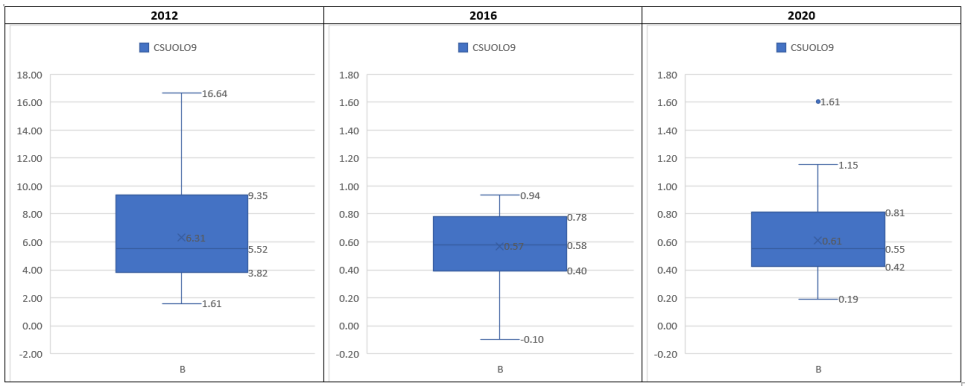


Figura 4.36 Box Plot della variabile CSUOLO9 nel Gruppo B

**DISECO3 (Superficie impattata dalla presenza di coperture artificiali)**

Anno 2012: Nel gruppo B, il parametro DISECO3 presenta una media di 23.14 e una mediana di 22.85. I valori oscillano tra 12.04 e 35.69, con una distribuzione compatta intorno ai quartili 17.59 e 28.46.

Anno 2016: Nel gruppo B, il parametro DISECO3 registra una media leggermente superiore a 23.93 e una mediana di 23.82. I valori variano da 12.10 a 37.44, evidenziando una distribuzione compatta intorno ai quartili 18.36 e 28.67.

Anno 2020: Nel gruppo B, il parametro DISECO3 mostra un'ulteriore crescita, con una media di 24.12 e una mediana di 24.32. I valori si estendono tra 12.13 e 34.87, mantenendo una distribuzione compatta intorno ai quartili 18.56 e 29.13.

Nel gruppo B si osserva un leggero aumento nel parametro DISECO3 nel corso degli anni, passando da una media di 23.14 nel 2012 a 24.12 nel 2020. La distribuzione dei dati rimane compatta intorno ai quartili, indicando una certa

stabilità nelle caratteristiche del parametro. Non sono stati rilevati outliers significativi.

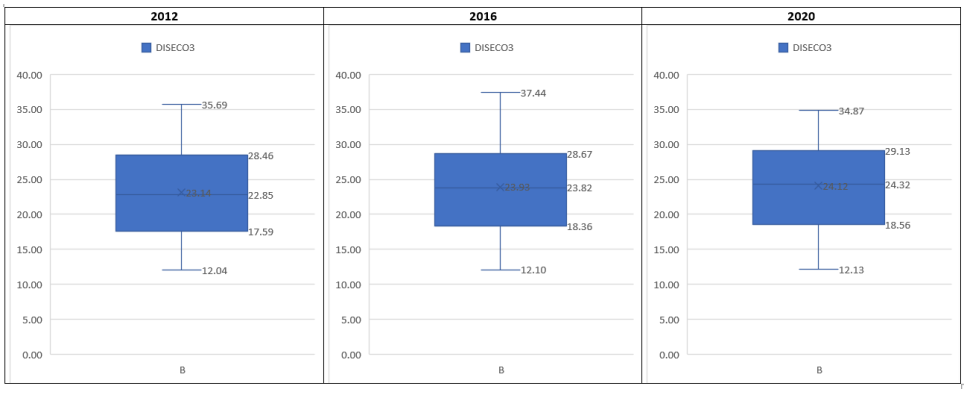


Figura 4.37 Box Plot della variabile DISECO3 nel Gruppo B

**FORMET5 (Indice di Dispersione Urbana)**

Anno 2012: Nel gruppo B, il parametro FORMET5 presenta una media di 89.05 e una mediana di 89.79. I valori variano da 74.66 a 96.36, con una distribuzione compatta intorno ai quartili 86.21 e 92.87. È stato rilevato un outlier con valore 74.66 Cagliari.

Anno 2016: Nel gruppo B, il parametro FORMET5 registra una media di 89.71 e una mediana di 89.83. I valori oscillano tra 82.13 e 96.38, con una distribuzione compatta intorno ai quartili 86.90 e 92.11.

Anno 2020: Nel gruppo B, il parametro FORMET5 mostra un aumento con una media di 90.04 e una mediana di 90.13. I valori si estendono da 82.10 a 96.42, mantenendo una distribuzione compatta intorno ai quartili 88.16 e 92.36.

In conclusione, nel gruppo B si osserva un aumento leggero ma costante nel parametro FORMET5 nel corso degli anni, passando da una media di 89.05 nel 2012 a 90.04 nel 2020.

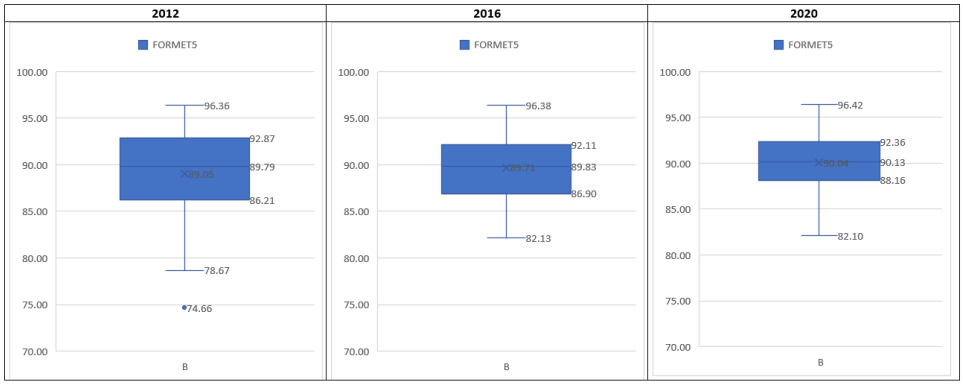


Figura 4.38 Box Plot della variabile FORMET5 nel Gruppo B

Complessivamente, l'analisi dei dati evidenzia che le province del gruppo B presentano indicatori migliori rispetto ad altri gruppi, il che suggerisce una gestione più consapevole del suolo e dell'ambiente.

### 4.4.3 Gruppo C

Tabella 4.15 Gruppo C. Principali Indicatori

| Gruppo 2012 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| C | 5 | 0.27 | 13.99 | 32.20 | 85.80 |
| Total | 5 | 0.27 | 13.99 | 32.20 | 85.80 |

| Gruppo 2016 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| C | 5 | 0.25 | 1.02 | 32.27 | 85.84 |
| Total | 5 | 0.25 | 1.02 | 32.27 | 85.84 |

| Gruppo 2020 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| C | 5 | 0.24 | 0.96 | 32.34 | 85.72 |
| Total | 5 | 0.24 | 0.96 | 32.34 | 85.72 |

Le province del gruppo C che mantengono lo stesso gruppo nel corso degli anni sono poche rispetto ad altri gruppi. Molte province di questo gruppo si trovano lungo le coste e hanno un'economia influenzata dall'agricoltura e dall'industria agroalimentare.

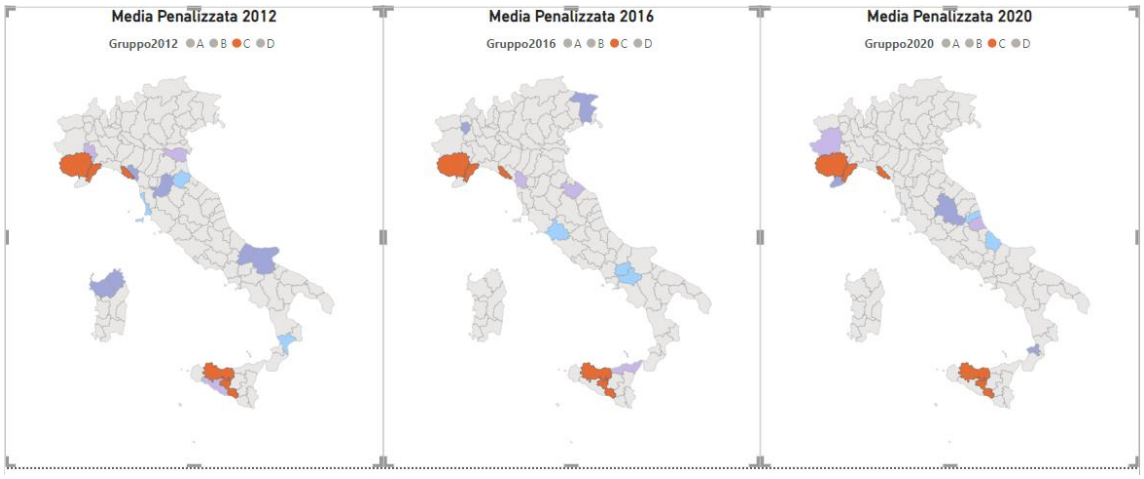


Figura 4.39 Intersezioni tra Province del Gruppo C

Nel gruppo C, le province presentano valori medi del ranking che si situano tra quelli del gruppo B e del gruppo D. Ciò indica che, in generale, le province del gruppo C hanno una posizione intermedia in termini di consumo di suolo rispetto agli altri gruppi.

Le province del gruppo C hanno mostrato cambiamenti nel loro ranking nel corso del tempo. Alcune province hanno registrato miglioramenti, mentre altre hanno subito diminuzioni. Ad esempio, Savona ha mostrato un miglioramento nel ranking nel periodo considerato, passando dal 31º posto nel 2012 al 28º posto nel 2016. Tuttavia, ha subito una diminuzione nel 2020, scendendo al 40º posto. Cuneo, Palermo e La Spezia hanno registrato miglioramenti nel ranking. Caltanissetta ha mantenuto una posizione relativamente stabile nel ranking nel corso degli anni.

Tabella 4.16 Gruppo C con coerenza nel tempo e classifica delle province

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Cuneo | C | 44 | C | 46 | C | 29 |
| 2 | Palermo | C | 44 | C | 29 | C | 33 |
| 3 | La Spezia | C | 40 | C | 33 | C | 36 |
| 4 | Caltanissetta | C | 36 | C | 38 | C | 32 |
| 5 | Savona | C | 31 | C | 28 | C | 40 |

Tabella 4.17 Le dinamiche di cambiamento all'interno del Gruppo C

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Gorizia | A | 92 | A | 92 | C | 35 |
| 2 | Rimini | A | 89 | A | 84 | C | 58 |
| 3 | Rovigo | A | 88 | D | 68 | C | 39 |
| 4 | Ancona | A | 83 | A | 83 | C | 31 |
| 5 | Catanzaro | A | 77 | C | 48 | C | 38 |
| 6 | Livorno | A | 75 | C | 49 | C | 60 |
| 7 | Alessandria | A | 73 | D | 69 | C | 52 |
| 8 | Frosinone | A | 71 | D | 58 | C | 46 |
| 9 | Forlì-Cesena | A | 70 | C | 44 | C | 54 |
| 10 | Salerno | A | 70 | D | 62 | C | 50 |
| 11 | Parma | D | 67 | A | 97 | C | 55 |
| 12 | Pescara | D | 66 | B | 24 | C | 57 |
| 13 | Ferrara | D | 64 | C | 54 | C | 41 |
| 14 | Piacenza | D | 61 | C | 39 | D | 68 |
| 15 | Agrigento | D | 59 | C | 43 | C | 46 |
| 16 | Asti | D | 59 | C | 53 | C | 48 |
| 17 | Lecco | D | 56 | C | 45 | D | 71 |
| 18 | Macerata | D | 55 | C | 42 | D | 62 |
| 19 | Trapani | C | 54 | A | 103 | D | 66 |
| 20 | Teramo | C | 53 | C | 37 | D | 67 |
| 21 | Pesaro e Urbino | C | 52 | D | 63 | C | 44 |
| 22 | Trieste | C | 52 | A | 80 | A | 76 |
| 23 | Torino | C | 50 | C | 51 | D | 72 |
| 24 | Viterbo | C | 49 | A | 92 | C | 61 |
| 25 | Pordenone | C | 48 | B | 26 | B | 23 |
| 26 | Biella | C | 47 | B | 13 | C | 42 |
| 27 | Perugia | C | 46 | C | 47 | B | 20 |
| 28 | Benevento | C | 45 | A | 87 | C | 52 |
| 29 | Arezzo | C | 44 | D | 56 | B | 23 |
| 30 | Vibo Valentia | C | 41 | C | 52 | B | 26 |
| 31 | Terni | C | 39 | B | 20 | B | 27 |
| 32 | Chieti | C | 38 | C | 31 | A | 88 |
| 33 | Reggio Calabria | C | 37 | B | 27 | B | 28 |
| 34 | Ascoli Piceno | C | 35 | C | 34 | A | 84 |
| 35 | Pistoia | C | 34 | D | 58 | D | 73 |
| 36 | Udine | C | 34 | B | 23 | C | 30 |
| 37 | Messina | C | 32 | D | 64 | C | 37 |
| 38 | Vercelli | C | 30 | B | 14 | B | 15 |
| 39 | Lucca | C | 29 | D | 67 | C | 59 |
| 40 | Cosenza | C | 28 | B | 19 | B | 10 |
| 41 | Avellino | C | 27 | A | 82 | C | 48 |
| 42 | Imperia | C | 27 | C | 35 | B | 19 |
| 43 | Massa-Carrara | B | 25 | C | 36 | C | 44 |
| 44 | Foggia | B | 24 | C | 32 | C | 49 |
| 45 | Campobasso | B | 23 | C | 30 | C | 56 |
| 46 | Firenze | B | 21 | C | 40 | C | 53 |
| 47 | Cagliari | B | 20 | C | 50 | A | 94 |
| 48 | Sassari | B | 17 | C | 41 | C | 35 |

È interessante notare che il gruppo C ha mostrato un numero maggiore di province che hanno registrato variazioni significative nel loro ranking rispetto agli altri gruppi. Questo suggerisce che all'interno del gruppo C vi siano dinamiche

particolari che hanno portato a un maggior spostamento delle posizioni nel corso degli anni considerati.

Le province di Gorizia, Rimini, Rovigo, Ancona, Catanzaro, Livorno, Alessandria, Frosinone, Forlì-Cesena e Salerno hanno mostrato un miglioramento nel ranking nel corso del tempo. Sono riuscite a passare dal gruppo A, caratterizzato da un eccessivo consumo di suolo, al gruppo C, mentre le province di Cagliari, Chieti, Ascoli Piceno, Trieste, Pistoia, Torino, Lecco, Piacenza, Teramo, Trapani e Macerata hanno registrato una significativa diminuzione nel ranking nel corso del tempo. Sono passate dal gruppo C, caratterizzato da una minore pressione sul suolo, al gruppo A con peggiori posizioni nel ranking.

**Box plot delle Variabili Originali. Gruppo C**

**CSUOLO9 (Densità di consumo di suolo [m2] rispetto all'area totale [ha])**

Nel 2012, nel gruppo C, il parametro CSUOLO presenta una media di 14.27 e una mediana di 14.95. I valori variano da 4.19 a 22.40, con una distribuzione compatta intorno ai quartili 10.02 e 18.13. Nel 2016, nel gruppo C, il parametro CSUOLO registra una media di 1.15 e una mediana di 1.03. I valori oscillano tra 0.62 e 1.82, con una distribuzione compatta intorno ai quartili 0.87 e 1.52. Nel 2020, nel gruppo C, il parametro CSUOLO mostra una media di 1.06 e una mediana di 1.00. I valori si estendono da 0.35 a 1.88, mantenendo una distribuzione compatta intorno ai quartili 0.87 e 1.20. Sono stati rilevati alcuni outliers con valori di 0.35 Gorizia, 1.88 Viterbo e 1.83 Campobasso.

In conclusione, nel gruppo C si osserva una diminuzione nel parametro CSUOLO nel corso degli anni, passando da una media di 14.27 nel 2012 a 1.06 nel 2020.

Tuttavia, nel 2020 sono stati rilevati alcuni outliers che si discostano significativamente dalla distribuzione principale.

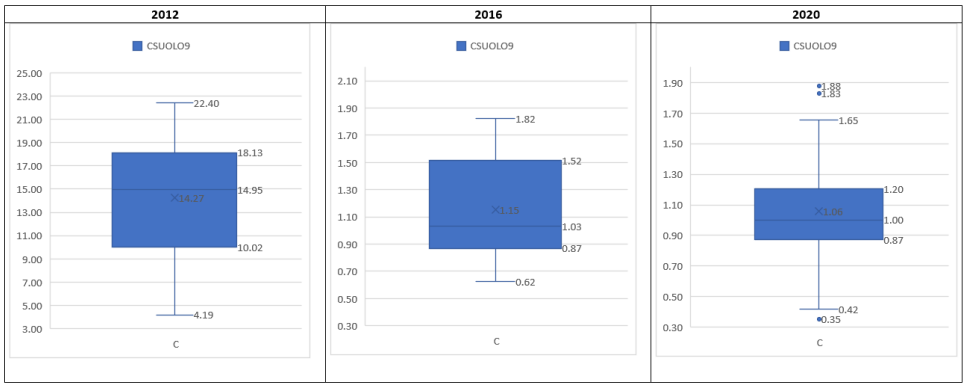


Figura 4.40 Box Plot della variabile CSUOLO9 nel Gruppo C

**DISECO3 (Superficie impattata dalla presenza di coperture artificiali)**

Nel 2012, nel gruppo C, il parametro DISECO3 presenta una media di 34.27 e una mediana di 33.78. I valori variano da 24.71 a 55.34, con una distribuzione compatta intorno ai quartili 30.65 e 37.05. È stato rilevato un outlier con valore 55.34 Trieste.

Nel 2016, nel gruppo C, il parametro DISECO3 registra una media di 33.55 e una mediana di 33.89. I valori oscillano tra 23.57 e 40.93, con una distribuzione compatta intorno ai quartili 30.95 e 35.75. È stato rilevato un outlier con valore 23.57 Sassari.

Nel 2020, nel gruppo C, il parametro DISECO3 mostra una media di 35.85 e una mediana di 35.79. I valori si estendono da 23.65 a 48.01, mantenendo una distribuzione compatta intorno ai quartili 31.46 e 39.59.

In conclusione, nel gruppo C si osserva una leggera variazione nel parametro DISECO3 nel corso degli anni, passando da una media di 34.27 nel 2012 a 35.85 nel 2020.

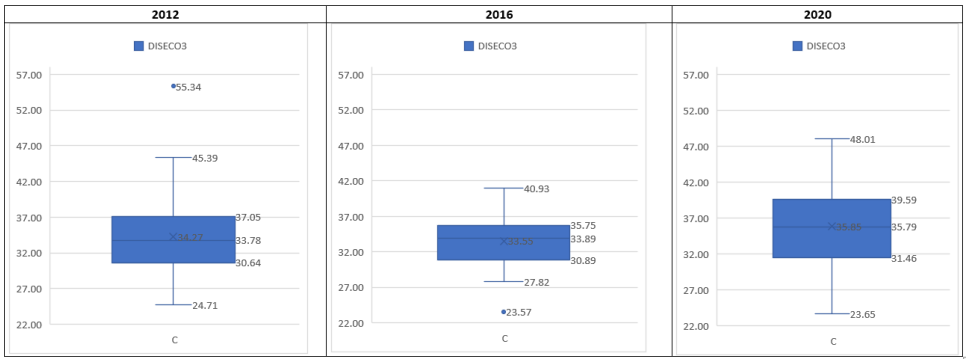


Figura 4.41 Box Plot della variabile DISECO3 nel Gruppo C

**FORMET5 (Indice di Dispersione Urbana)**

Nel 2012, nel gruppo C, il parametro FORMET5 presenta una media di 87.50 e una mediana di 88. I valori variano da 70.60 a 93.96, con una distribuzione compatta intorno ai quartili 85.47 e 90.92. È stato rilevato un outlier con valore 70.60 Trieste.

Nel 2016, nel gruppo C, il parametro FORMET5 registra una media di 86.15 e una mediana di 86.77. I valori oscillano tra 74.88 e 93.38, con una distribuzione compatta intorno ai quartili 82.11 e 90.37. Non sono stati rilevati outliers.

Nel 2020, nel gruppo C, il parametro FORMET5 mostra un aumento, con una media di 86.88 e una mediana di 87.28. I valori si estendono da 77.45 a 95.33, mantenendo una distribuzione compatta intorno ai quartili 84.77 e 89.50. È stato rilevato un outlier con valore 77.45 Livorno e 78.67 Massa Carrara.

In conclusione, nel gruppo C si osserva una tendenza generale di aumento nel parametro FORMET5 nel corso degli anni, passando da una media di 87.50 nel

2012 a 86.88 nel 2020. La distribuzione dei dati rimane compatta intorno ai quartili, indicando una certa stabilità nelle caratteristiche del parametro. Sono stati rilevati alcuni outliers nel 2012 e nel 2020, che rappresentano valori anomali rispetto alla distribuzione dei dati.

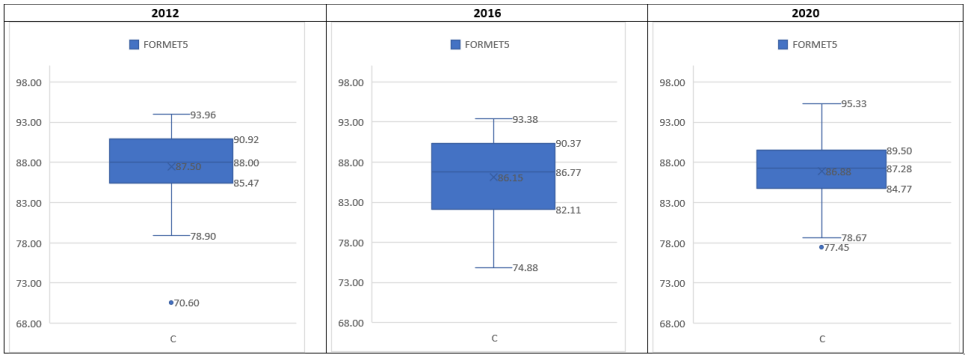


Figura 4.42 Box Plot della variabile FORMET5 nel Gruppo C

## 4.4.4 Gruppo D

Tabella 4.18 Gruppo D. Principali Indicatori

| Gruppo 2012 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| D | 2 | 0.37 | 20.26 | 39.55 | 85.07 |
| Total | 2 | 0.37 | 20.26 | 39.55 | 85.07 |

| Gruppo 2016 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| D | 2 | 0.41 | 2.04 | 39.61 | 84.96 |
| Total | 2 | 0.41 | 2.04 | 39.61 | 84.96 |

| Gruppo 2020 | Count of Province | Average of MP_geometric | Average of CSUOLO9 | Average of DISECO3 | Average of FORMET5 |
|---|---|---|---|---|---|
| D | 2 | 0.38 | 1.81 | 39.68 | 84.71 |
| Total | 2 | 0.38 | 1.81 | 39.68 | 84.71 |

Le province del gruppo D, Reggio nell'Emilia e Pisa, si caratterizzano per la loro posizione costante all'interno di tale gruppo nel corso degli anni considerati.

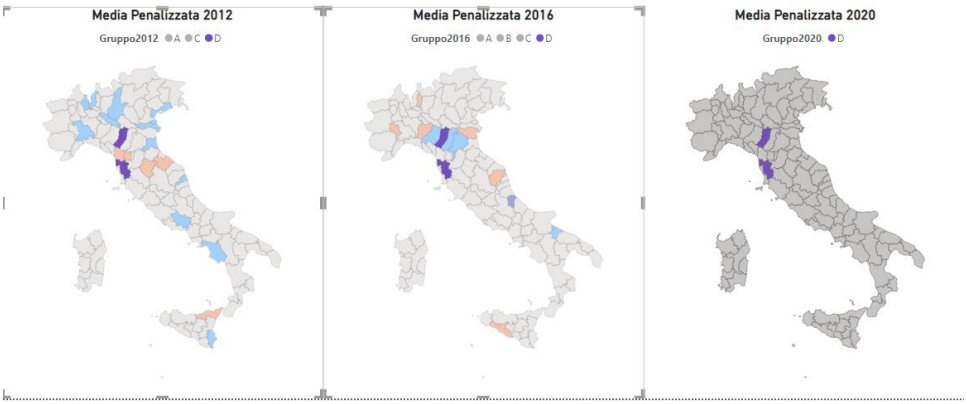


Figura 4.43: Intersezioni tra Province del Gruppo D

Tabella 4.19 Gruppo D con coerenza nel tempo e classifica delle province

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Reggio nell'Emilia | D | 63 | D | 73 | D | 74 |
| 2 | Pisa | D | 57 | D | 65 | D | 65 |

Le province del gruppo D, Reggio nell'Emilia e Pisa, che non hanno mai cambiato gruppo, presentano un comportamento diverso in termini di rango e consumo del suolo nel corso degli anni considerati. Entrambe le province hanno mostrato un aumento del consumo del suolo nel corso degli anni, sebbene abbiano registrato un cambiamento nel loro rango. Reggio nell'Emilia ha avuto un leggero aumento del rango nel 2020, mentre Pisa ha mantenuto lo stesso rango del 2016.

Tabella 4.20 Le dinamiche di cambiamento all'interno del Gruppo D

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Ravenna | A | 99 | D | 59 | A | 95 |
| 2 | Varese | A | 97 | D | 60 | A | 90 |
| 3 | Venezia | A | 96 | D | 70 | A | 78 |
| 4 | Latina | A | 95 | A | 74 | D | 69 |
| 5 | Cremona | A | 91 | D | 61 | A | 86 |
| 6 | Rovigo | A | 88 | D | 68 | C | 39 |
| 7 | Siracusa | A | 84 | D | 72 | A | 80 |
| 8 | Fermo | A | 79 | D | 55 | D | 70 |
| 9 | Brescia | A | 76 | D | 66 | A | 91 |
| 10 | Taranto | A | 74 | A | 79 | D | 63 |
| 11 | Alessandria | A | 73 | D | 69 | C | 52 |
| 12 | Como | A | 72 | D | 71 | A | 79 |
| 13 | Frosinone | A | 71 | D | 58 | C | 46 |
| 14 | Salerno | A | 70 | D | 62 | C | 50 |
| 15 | Barletta-Andria-Trani | D | 68 | A | 86 | D | 64 |
| 16 | Parma | D | 67 | A | 97 | C | 55 |
| 17 | Pescara | D | 66 | B | 24 | C | 57 |
| 18 | Bologna | D | 65 | A | 98 | A | 75 |
| 19 | Ferrara | D | 64 | C | 54 | C | 41 |
| 20 | Prato | D | 62 | A | 75 | A | 92 |
| 21 | Piacenza | D | 61 | C | 39 | D | 68 |
| 22 | Modena | D | 60 | A | 99 | A | 83 |
| 23 | Agrigento | D | 59 | C | 43 | C | 46 |
| 24 | Asti | D | 59 | C | 53 | C | 48 |
| 25 | Lecco | D | 56 | C | 45 | D | 71 |
| 26 | Macerata | D | 55 | C | 42 | D | 62 |
| 27 | Trapani | C | 54 | A | 103 | D | 66 |
| 28 | Teramo | C | 53 | C | 37 | D | 67 |
| 29 | Pesaro e Urbino | C | 52 | D | 63 | C | 44 |
| 30 | Torino | C | 50 | C | 51 | D | 72 |
| 31 | Arezzo | C | 44 | D | 56 | B | 23 |
| 32 | Pistoia | C | 34 | D | 58 | D | 73 |
| 33 | Messina | C | 32 | D | 64 | C | 37 |
| 34 | Lucca | C | 29 | D | 67 | C | 59 |

Analizzando i dati forniti, si può osservare che le province del gruppo D tendono a cambiare di solito verso i gruppi A o C.

Durante il periodo considerato, alcune province hanno registrato un notevole aumento di posizioni nella classifica. Tra queste province, Trapani si distingue per aver sperimentato il maggior incremento di rango, salendo di ben 49 posizioni rispetto al 2012. Tuttavia, nel 2020 ritorna alla posizione 66. Alcune province come Parma e Pescara hanno seguito dinamiche simili, peggiorando nel 2016 e poi migliorando nelle posizioni successive. Invece, Ravenna, Varese, Venezia, Latina, Cremona, Rovigo, Siracusa, Fermo, Brescia, Taranto, Alessandria, Como, Frosinone e Salerno hanno avuto l'andamento opposto, con un miglioramento significativo nel 2016, ma poi una successiva diminuzione nelle posizioni di classifica.

**Box plot delle Variabili Originali. Gruppo D**

**CSUOLO9 (Densità di consumo di suolo rispetto all'area totale)**

Nel 2012, nel gruppo D, il parametro CSUOLO9 presenta una media di 21.07 e una mediana di 21.27. I valori variano da 15.99 a 26.60, con una distribuzione compatta intorno ai quartili 18.97 e 23.23.

Nel 2016, nel gruppo D, il parametro CSUOLO9 registra una media di 1.87 e una mediana di 1.93. I valori oscillano tra 1.07 e 2.28, con una distribuzione compatta intorno ai quartili 1.67 e 2.16.

Nel 2020, nel gruppo D, il parametro CSUOLO9 mostra una diminuzione, con una media di 1.74 e una mediana di 1.74. I valori si estendono da 1.09 a 2.37, mantenendo una distribuzione compatta intorno ai quartili 1.58 e 1.91.

In conclusione, nel gruppo D si osserva una diminuzione nel parametro CSUOLO9 nel corso degli anni, passando da una media di 21.07 nel 2012 a 1.74 nel 2020. La distribuzione dei dati rimane compatta intorno ai quartili, indicando una certa stabilità nelle caratteristiche del parametro. Si nota una significativa riduzione dei valori nel 2016 e un lieve calo nel 2020.

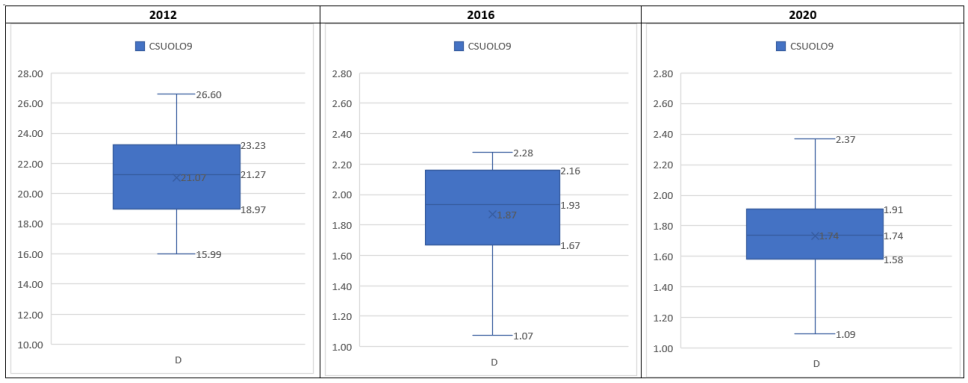


Figura 4.44 Box Plot della variabile CSUOLO9 nel Gruppo D

**DISECO3 (Superficie impattata dalla presenza di coperture artificiali)**

Nel 2012, nel gruppo D, il parametro DISECO3 presenta una media di 38.87 e una mediana di 38.78. I valori variano da 32.32 a 46.03, con una distribuzione compatta intorno ai quartili 36.27 e 41.34.

Nel 2016, nel gruppo D, il parametro DISECO3 registra una media di 39.83 e una mediana di 39.35. I valori oscillano tra 31.42 e 55.93, con una distribuzione compatta intorno ai quartili 35.14 e 41.85. È stato rilevato un outlier con valore 55.93 Varese.

Nel 2020, nel gruppo D, il parametro DISECO3 mostra una stabilità, con una media di 39.51 e una mediana di 39.31. I valori si estendono da 31.57 a 47.44, mantenendo una distribuzione compatta intorno ai quartili 34.44 e 44.38.

In conclusione, nel gruppo D si osserva una certa stabilità nel parametro DISECO3 nel corso degli anni, con valori medi e mediani simili tra il 2012 e il 2020. La

distribuzione dei dati rimane compatta intorno ai quartili, indicando una relativa consistenza nelle caratteristiche del parametro.

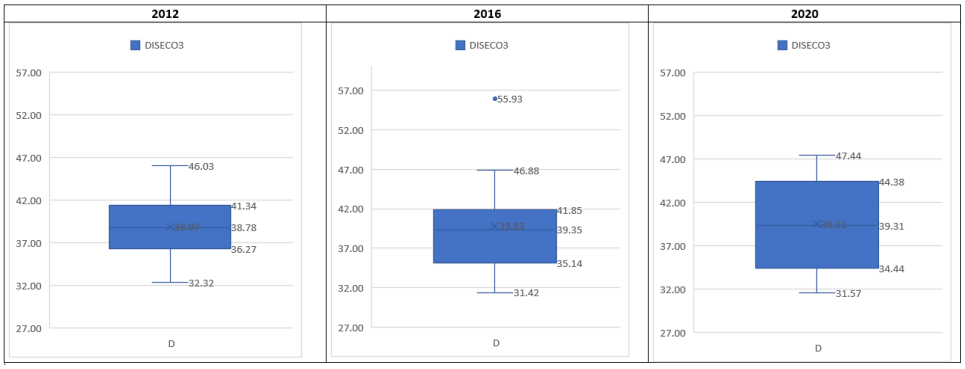


Figura 4.45 Box Plot della variabile DISECO3 nel Gruppo D

**FORMET5 (Indice di Dispersione Urbana)**

Nel 2012, nel gruppo D, il parametro FORMET5 presenta una media di 83.86 e una mediana di 84.97. I valori variano da 64.05 a 93.17, con una distribuzione compatta intorno ai quartili 81.81 e 87.45. È stato rilevato un outlier con valore 64.05 Prato.

Nel 2016, nel gruppo D, il parametro FORMET5 registra una media di 85.53 e una mediana di 86.62. I valori oscillano tra 71.08 e 95.40, con una distribuzione compatta intorno ai quartili 81.47 e 88.99.

Nel 2020, nel gruppo D, il parametro FORMET5 mostra una leggera diminuzione, con una media di 85.31 e una mediana di 85.06. I valori si estendono da 78.43 a 90.86, mantenendo una distribuzione compatta intorno ai quartili 80.50 e 89.62.

In conclusione, nel gruppo D si osserva una leggera diminuzione nel parametro FORMET5 nel corso degli anni, passando da una media di 83.86 nel 2012 a 85.31 nel 2020. La distribuzione dei dati rimane compatta intorno ai quartili, indicando una certa stabilità nelle caratteristiche del parametro. È stato rilevato un outlier nel 2012 con un valore anomalo.

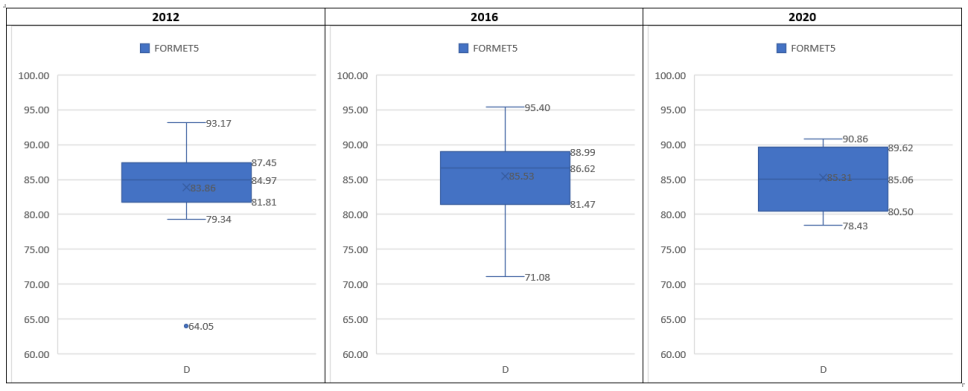


Figura 4.46 Box Plot della variabile FORMET5 nel Gruppo D

### 4.4.5 Province non intersecanti nei gruppi del cluster: Analisi delle province che non condividono gruppi tra i cluster degli anni 2012, 2016 e 2020

Le province che non intersecano tra i cluster degli anni 2012, 2016 e 2020 indicano una situazione in cui la classificazione delle province è cambiata significativamente nel corso del tempo. Questo significa che le province hanno mostrato modelli di consumo di suolo o altre caratteristiche territoriali che le hanno spinte a spostarsi da un cluster all'altro nel corso degli anni.

La mancanza di intersezione tra i cluster suggerisce che le province hanno avuto dinamiche e tendenze diverse rispetto alle altre nel corso del periodo considerato.

Figura 4.47 Mappa Province Non Intersecanti

Nella tabella fornita, sono elencate le province con le relative informazioni di classificazione e rango per gli anni 2012, 2016 e 2020. Le province che non intersecano tra i cluster degli anni considerati sono:

Tabella 4.21 Le dinamiche di cambiamento dei gruppi e il valore della media penalizzata delle province non intersecanti

| | Province | Gruppo2012 | Rank_2012 | Gruppo2016 | Rank_2016 | Gruppo2020 | Rank_2020 |
|---|---|---|---|---|---|---|---|
| 1 | Rovigo | A | 88 | D | 68 | C | 39 |
| 2 | Alessandria | A | 73 | D | 69 | C | 52 |
| 3 | Frosinone | A | 71 | D | 58 | C | 46 |
| 4 | Salerno | A | 70 | D | 62 | C | 50 |
| 5 | Parma | D | 67 | A | 97 | C | 55 |
| 6 | Pescara | D | 66 | B | 24 | C | 57 |
| 7 | Trapani | C | 54 | A | 103 | D | 66 |
| 8 | Arezzo | C | 44 | D | 56 | B | 23 |
| 9 | Cagliari | B | 20 | C | 50 | A | 94 |

Queste province hanno mostrato un'instabilità nel loro posizionamento nella classifica nel corso degli anni, perdendo o guadagnando un numero considerevole di posizioni. Queste dinamiche incerte e imprevedibili sono emerse in province come Rovigo, Alessandria, Frosinone, Salerno, Parma, Pescara, Trapani, Arezzo e Cagliari. Queste province hanno mostrato un comportamento irregolare, con fluttuazioni significative nel loro rank, rendendo difficile prevedere il loro posizionamento e indicando una situazione di cambiamento caotico nel tempo.

Queste differenze di rank indicano cambiamenti nella posizione relativa delle province rispetto ad altre province italiane nel contesto del consumo di suolo.

## 4.5 Conclusione dell'Analisi del Clustering

Durante questo capitolo, è stata condotta un'analisi dei cluster utilizzando dati relativi a tre anni differenti e tre approcci distinti al fine di determinare il metodo migliore. Sulla base di questa analisi, il clustering ottenuto utilizzando la media

geometrica penalizzata si è dimostrato superiore, in quanto è stato in grado di separare gli elementi in modo più accurato rispetto alle variabili originali.

Successivamente, sono stati individuati i valori massimi e minimi dei quattro gruppi, definendo così delle soglie che possono essere utilizzate per classificare le province in base al consumo di suolo. Questi risultati forniscono una base solida per l'analisi empirica successiva, concentrandosi sul cluster ottenuto con la media geometrica penalizzata per ulteriori indagini e interpretazioni.

L'analisi del cluster basato sulla media geometrica penalizzata ha permesso di distinguere quattro gruppi distinti in base al consumo di suolo. I dati raccolti durante l'analisi delle intersezioni hanno evidenziato che due di questi quattro gruppi si distinguono in modo significativo, mostrando una ripetizione delle province all'interno di essi nel corso degli anni.

Il primo gruppo, denominato A, è caratterizzato da un consumo di suolo eccessivo e comprende province più popolate con un elevato sviluppo economico. Il secondo gruppo, denominato B, si trova in zone montane in cui le attività manifatturiere e l'agricoltura intensiva non sono possibili, e presenta un consumo di suolo sostenibile.

Gli altri due gruppi identificati sono il gruppo C, che comprende cinque province che rimangono all'interno del gruppo nel corso degli anni e presentano un consumo di suolo soddisfacente, e il gruppo D, meno numeroso, con solo due province che rimangono nel gruppo nel corso del tempo, ma altre 34 province che sono state nel gruppo D nel corso degli anni. Il gruppo D è caratterizzato da un consumo di suolo insoddisfacente, tanto che le province appartenenti a questo gruppo tendono a passare nel gruppo A, caratterizzato da un consumo di suolo eccessivo.

Inoltre, è stata condotta un'analisi sulla distribuzione delle aziende agricole e manifatturiere nei gruppi con consumo di suolo nel periodo 2012, 2016, 2020, evidenziando che le province con una maggiore concentrazione di attività manifatturiere mostrano un consumo di suolo più elevato, mentre una maggiore presenza di aziende agricole è associata a un minor consumo di suolo. Al fine di approfondire la relazione tra consumo di suolo e la presenza delle attività agricole e manifatturiere, si ritiene necessario condurre un'analisi di regressione. Questo approccio statistico permetterà di valutare se esiste una relazione significativa tra le variabili in esame e di determinarne l'entità.

# Capitolo 5 Regressione Lineare

Nel presente capitolo, si è condotto un'analisi della regressione al fine di confrontare gli indici ottenuti dalle tre variabili originali: CSUOLO9, DISECO3 e FORMET5. In particolare, sono state considerate due misure della Media Geometrica: la Media Geometrica Classica e la Media Geometrica Penalizzata.

L'obiettivo di questo confronto è valutare se l'indice Media Geometrica Penalizzata risulta più significativo e in grado di catturare in modo più efficace le informazioni desiderate rispetto alla Media Geometrica Classica.

Si analizzano le due varianti dell'indice come variabili dipendenti. Si utilizza un'analisi di regressione lineare univariata per esaminare la relazione tra queste variabili e il numero delle aziende agricole, nonché separatamente il numero delle aziende manifatturiere.

Successivamente, si conduce un'analisi di regressione lineare multivariata, includendo sia il numero delle aziende agricole che il numero delle aziende manifatturiere come due variabili indipendenti.

Si cerca di valutare se esiste un legame significativo tra il consumo di suolo e il numero delle aziende agricole e manifatturiere, al fine di valutare l'impatto dell'attività economica rappresentata da queste aziende sull'indice del consumo di suolo nelle province considerate nello studio.

Nel contesto della regressione lineare univariata, quando si considera come variabile indipendente il numero di aziende agricole, l'equazione del modello è la seguente:

$$Media\_Geometrica = \beta_0 + \beta_1 * Numero\_Aziende\_Agricole + \varepsilon \quad (15)$$

Quando si considera come variabile indipendente il numero di aziende manifatturiere, l'equazione del modello è la seguente:

$$Media\_Geometrica = \beta_0 + \beta_1 * Numero\_Aziende\_Manifatturiere + \varepsilon \quad (16)$$

dove:

- $Media\_Geometrica$ rappresenta la variabile dipendente, che può essere sia la Media Geometrica Classica che la Media Geometrica Penalizzata.
- $Numero\_Aziende\_Agricole$ è la variabile indipendente, che rappresenta il numero di aziende agricole considerate.
- $Numero\_Aziende\_Manifatturiere$ è la variabile indipendente, che rappresenta il numero di aziende agricole considerate.

- $\beta_0$ e $\beta_1$ sono i coefficienti di regressione, che misurano rispettivamente l'intercetta e l'effetto della variabile indipendente sulla $Media\_Geometrica$.

- ε rappresenta l'errore residuo, che rappresenta la parte della variabilità non spiegata dal modello.

Successivamente, nella regressione lineare multivariata, si include anche il numero delle aziende manifatturiere come seconda variabile indipendente. L'equazione del modello diventa:

$$Media\_geometrica = \beta_0 + \beta_1 * Numero\_Aziende\_Agricole + \beta_2 * Numero\_Aziende\_Manifatturiere + \varepsilon \quad (17)$$

dove:

- $Media\_Geometrica$, $Numero\_Aziende\_Agricole$ e $Numero\_Aziende\_Manifatturiere$ hanno gli stessi significati della regressione univariata.

- $\beta_0$, $\beta_1$ e $\beta_2$ rappresentano i coefficienti di regressione, che misurano rispettivamente l'intercetta, l'effetto del $Numero\_Aziende\_Agricole$ e l'effetto del $Numero\_Aziende\_Manifatturiere$ sulla $Media\_Geometrica$.

- ε rappresenta l'errore residuo, che rappresenta la parte della variabilità non spiegata dal modello.

Nell'analisi di regressione lineare, si cerca di stimare i coefficienti di regressione ottimali ($\beta_0$, $\beta_1$ , $\beta_2$ ecc.) per determinare la relazione tra le variabili indipendenti e la variabile dipendente e misurare l'entità dell'influenza delle variabili indipendenti sul risultato desiderato.

## 5.1 Valutazione dell'adeguatezza del modello di regressione

Per valutare la bontà del modello di regressione, sono disponibili diverse misure, tra cui:

1. Il coefficiente di determinazione $R^2$ è un indice di bontà di adattamento del modello di regressione lineare. $R^2$ è una misura della proporzione di varianza della variabile dipendente che può essere spiegata dalle variabili indipendenti nel modello. $R^2$ varia da 0 a 1, dove 0 indica che le variabili indipendenti non spiegano la variazione nella variabile dipendente e 1 indica una spiegazione completa della variazione.

   La formula per calcolare il $R^2$ è:

$$R^2 = 1 - (SSR / SST) \tag{18}$$

   dove:

- $SSR$ (Sum of Squares Residuals) rappresenta la somma dei quadrati dei residui, che sono le differenze tra i valori osservati e i valori predetti dal modello.
- $SST$ (Sum of Squares Total) rappresenta la somma dei quadrati totali, che rappresenta la variabilità totale dei dati rispetto alla media dei valori osservati.

2. $R^2\ corretto$ (adjusted) è una versione del $R^2$ che tiene conto del numero di variabili indipendenti nel modello e del numero di osservazioni nel campione. L'obiettivo del $R^2$ corretto o adjusted è penalizzare l'inclusione di variabili indipendenti che non apportano un significativo miglioramento alla capacità predittiva del modello.

$$R^2\ corretto = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-k-1}\right)\right] \tag{19}$$

dove:

- $n$ è il numero totale di osservazioni nel campione
- $k$ è il numero di variabili indipendenti nel modello

3. P-value è una misura di significatività statistica che indica la probabilità che il coefficiente di regressione sia significativamente diverso da zero. Un p-value basso (di solito inferiore a 0,05) suggerisce una significativa relazione tra la variabile indipendente e la variabile dipendente nel modello. Il p-value viene calcolato utilizzando un test di ipotesi sul coefficiente di regressione, in cui l'ipotesi nulla è che il coefficiente sia uguale a zero. Se il p-value è inferiore a una soglia di significatività predefinita (ad esempio, 0,05), l'ipotesi nulla viene rifiutata e il coefficiente è considerato statisticamente significativo.
4. Il test F per la regressione lineare viene utilizzato per valutare complessivamente la significatività del modello di regressione. Test F confronta il modello di regressione lineare con un modello null in cui non sono presenti le variabili indipendenti.

   La formula per calcolare il test F per la regressione lineare è la seguente:

$$F = \frac{SSR/k}{SSE/(n-k-1)} \tag{20}$$

   dove:

- $SSR$ (Sum of Squares Regression) rappresenta la somma dei quadrati della regressione, ovvero la variazione spiegata dal modello di regressione
- k è il numero di variabili indipendenti nel modello di regressione (escluso l'intercetta)

- $SSE$ (Sum of Squares Error) rappresenta la somma dei quadrati degli errori, ovvero la variazione residua non spiegata dal modello di regressione
- n è il numero totale di osservazioni nel campione
- k + 1 è il numero totale di coefficienti stimati nel modello di regressione (compreso l'intercetta)

La statistica del test F segue una distribuzione di probabilità F con k e n - k - 1 gradi di libertà. Il valore p associato al test F indica la probabilità di ottenere una statistica di test uguale o più estrema sotto l'ipotesi nulla che non ci sia alcuna relazione significativa tra le variabili indipendenti e dipendenti.

Ipotesi Nulla ($H_0$): Non vi è alcuna relazione significativa tra le variabili indipendenti e la variabile dipendente nel modello di regressione.

Ipotesi Alternativa ($H_A$): Vi è una relazione significativa tra almeno una delle variabili indipendenti e la variabile dipendente nel modello di regressione.

Il valore p associato al test F indica la probabilità di ottenere una statistica di test uguale o più estrema sotto l'ipotesi nulla. Se il valore p è inferiore a un certo livello di significatività prefissato (solitamente 0.05), si può concludere che vi è una relazione significativa tra le variabili indipendenti e dipendente nel modello di regressione.

## 5.2 Valutazione della validità del modello di regressione lineare

Per valutare la validità del modello di regressione lineare, vengono eseguiti test statistici significativi. Uno di questi test è il test di normalità OLS (Ordinary Least Squares), una procedura statistica utilizzata per verificare l'assunzione di normalità degli errori nel modello di regressione lineare. La normalità degli errori è

un'assunzione essenziale per garantire la validità e l'affidabilità dei risultati statistici ottenuti.

Il test di normalità OLS si basa sull'ipotesi nulla che gli errori residuali seguano una distribuzione normale. Per testare questa ipotesi, vengono impiegati diversi metodi statistici, come il test di Shapiro-Wilk o il test di Kolmogorov-Smirnov. Questi test confrontano la distribuzione degli errori residuali con una distribuzione normale teorica e forniscono una statistica di test e un valore p associato.

Il p-value associato alla statistica del test viene confrontato con una soglia predefinita (ad esempio, $\alpha = 0,05$) per determinare se rifiutare o accettare l'ipotesi nulla.

L'output del test di normalità OLS fornisce la statistica del test e il valore p, che indica la probabilità di ottenere i risultati osservati se l'ipotesi nulla di normalità degli errori fosse verificata. Sulla base del valore p, è possibile valutare se i dati supportano o meno l'ipotesi di normalità degli errori.

### 5.2.1 Test Shapiro-Wilk

Il test di Shapiro-Wilk è uno dei test più potenti per la verifica della normalità dei dati, soprattutto quando si lavora con piccoli campioni. È stato introdotto nel 1965 da Samuel Shapiro e Martin Wilk.

Il test si basa sulla comparazione tra uno stimatore non parametrico, che utilizza una combinazione lineare ottimale delle statistiche d'ordine di una variabile aleatoria normale, e lo stimatore parametrico tradizionale, che corrisponde alla varianza campionaria. Il rapporto tra questi due stimatori costituisce la statistica del test.

Se le osservazioni seguono una distribuzione normale, il rapporto tenderà ad avvicinarsi a uno. In altre parole, se i dati sono normalmente distribuiti, il test di Shapiro-Wilk produrrà un risultato che indica una buona aderenza alla distribuzione normale.

Il test di Shapiro-Wilk è particolarmente utile per piccoli campioni, in quanto è in grado di rilevare anche deviazioni dalla normalità che potrebbero non essere evidenti con altri test.

- Ipotesi nulla ($H_0$): i valori hanno una distribuzione normale.
- Si mette in ordine crescente le osservazioni;
- per n osservazioni, poniamo m=n/2 se n è pari e m=(n−)/2 se n è dispari;
- si calcola la statistica del test di Shapiro e Wilk

$$W = \frac{(\sum_{i=1}^{m} k_i (X_{n+1-i} - X_i))^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{21}$$

più il valore di W nel test di Shapiro-Wilk si avvicina a 1, maggiore è l'evidenza che i dati si distribuiscano in modo normale.

### 5.2.2 Test Kolmogorov Smirnov

È un test che confronta la distribuzione empirica con la teorica.

Ipotesi nulla ($H_0$): i valori hanno una distribuzione normale. Se il p-value ottenuto dal test è maggiore di 0,05 (o altra soglia di significatività predefinita), allora l'ipotesi nulla viene accettata

$$\delta = max|F(x) - \Phi(\mathrm{x})| \tag{22}$$

dove:

- $\delta$ è la massima differenza assoluta tra le funzioni di distribuzione empirica e ipotizzata.
- $F(x)$ è la funzione di distribuzione empirica, che assegna una probabilità cumulativa ai dati osservati.
- $\Phi(x)$ è la funzione di distribuzione teorica ipotizzata, che rappresenta la distribuzione normale cumulativa.
- $\delta_0$ è una quantità prefissata positiva, che rappresenta un valore critico. Viene utilizzata per determinare la regione di rifiuto per il test.

Si può dimostrare che, se l'ipotesi da verificare fosse vera, la probabilità di ottenere casualmente un valore di $\delta$ non inferiore ad una prefissata quantità (positiva) $\delta_0$ sarebbe data da

$$Pr(\delta \geq \delta_0) = F_{KS}(\delta'_0) \tag{23}$$

dove $F_{KS}$ è la serie

$$F_{KS}(x) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} \tag{24}$$

## 5.3 Normalizzazione dei dati

Per quanto riguarda la normalizzazione dei dati, si utilizza la tecnica della normalizzazione per il massimo. Questo significa che ogni valore nella matrice o nel vettore viene diviso per il valore massimo presente nella colonna corrispondente.

La formula per la normalizzazione dei dati utilizzando la tecnica della normalizzazione per il massimo è la seguente:

$$x\_normalized = x / max(x) \quad (25)$$

dove:

- $x$ è il valore originale del dato
- $max(x)$ è il massimo valore presente nella colonna corrispondente

In questo modo, ogni valore viene scalato in modo che il valore massimo della colonna diventi 1 e gli altri valori vengano ridimensionati proporzionalmente. Questo processo consente di eliminare le differenze di scala tra le variabili.

Valutazione della validità del modello di regressione lineare attraverso test statistici

## 5.4 Analisi comparativa di modelli di regressione lineare univariata con differenti lag nelle variabili

Nel contesto di questa tesi, è stato effettuato un ampio lavoro di modellazione attraverso diverse regressioni lineari al fine di determinare quale di esse fosse più significativa. Sono state considerate sia analisi univariate che multivariate, valutando l'effetto delle diverse variabili sul fenomeno oggetto di studio.

Nella fase di modellazione univariata, sono stati esaminati gli effetti individuali delle variabili indipendenti sulla variabile dipendente di interesse. Tuttavia, è importante notare che i modelli univariati con la variabile dipendente "Media geometrica classica" spesso non soddisfano l'assunzione di normalità degli errori nel contesto della regressione lineare. I risultati dei tre test di normalità nei modelli univariati hanno evidenziato valori di p inferiori a 0.05, indicando che gli errori residuali non seguono una distribuzione normale. Ciò suggerisce che i modelli univariati potrebbero non essere adeguati a rappresentare correttamente le

relazioni tra le variabili nel contesto della regressione lineare. Nella tabella sono riportati i risultati dei modelli con la variabile dipendente Media Classica (MC) e Media Penalizzata (MP) considerando diversi lag. I modelli sono stati testati per la normalità degli errori e sono inclusi solo quelli che hanno superato il test di normalità.

Tabella 5.1 Tabella dei modelli univariati solo con test di normalità superati

| Lag | Modello | R^2 | P-value | Coefficienti Numero delle Aziende Agricole | Coefficienti Numero delle Aziende Manifatturiere | Residui normali |
|---|---|---|---|---|---|---|
| 0 | Univariata MP 2012 con numero delle Aziende Agricole 2012 | 0.116 | 0.001334 | -0.0006038 | | sì |
| 0 | Univariata MP 2012 con numero delle Aziende Manifatturiere 2012 | 0.2363 | 2.084e-06 | | 1.795e-05 | sì |
| 0 | Univariata MP 2016 con numero delle Aziende Agricole 2016 | 0.08517 | 0.006403 | -0.0005242 | | sì |
| 4 | Univariata MP 2016 con numero delle Aziende Manifatturiere 2012 | 0.2698 | 2.98E-07 | | 2.010e-05 | sì |
| 0 | Univariata MP 2016 con numero delle Aziende Manifatturiere 2016 | **0.2676** | 3.39e-07 | | 2.122e-05 | sì |
| 0 | Univariata MP 2020 con numero delle Aziende Manifatturiere 2020 | 0.315 | 1.894e-08 | 2.61e-15 | | sì |

Inoltre, è importante notare che i modelli univariati hanno mostrato un coefficiente di determinazione $R^2$ molto basso. Il coefficiente di determinazione misura la proporzione di varianza della variabile dipendente che può essere spiegata dalle variabili indipendenti nel modello. Un valore basso di $R^2$ indica che il modello univariato ha una capacità limitata nel spiegare la variazione nella variabile dipendente.

Pertanto, alla luce di questi risultati, è necessario esplorare ulteriormente modelli più complessi, come i modelli multivariati, e considerare l'inclusione di altre variabili indipendenti o l'uso di diverse trasformazioni delle variabili per ottenere risultati più significativi e una migliore rappresentazione delle relazioni tra le variabili nel contesto della regressione lineare.

## 5.5 Analisi comparativa di modelli di regressione lineare multivariata con differenti lag nelle variabili.

Successivamente, è stata effettuata un'analisi multivariata, che considera contemporaneamente l'effetto di più variabili indipendenti sulla variabile dipendente. Questo tipo di modellazione consente di esaminare l'interazione e l'influenza congiunta delle diverse variabili sul fenomeno in esame. Sono stati testati diversi modelli multivariati, includendo differenti combinazioni di variabili indipendenti e valutando la loro significatività attraverso metodi statistici appropriati.

Nella tabella sono riportati i risultati dei modelli multivariati con la variabile dipendente "Media Classica" e "Media Penalizzata" considerando diversi lag.

Tabella 5.2 Confronto dei modelli multivariati con variabile dipendente "Media Classica" e "Media Penalizzata"

| Lag | Modello | R^2 | R^2 corretto | P-value | Coefficienti Numero delle Aziende Agricole | Coefficienti Numero delle Aziende Manifatturiere | Residui normali |
|---|---|---|---|---|---|---|---|
| 0 | Multivariata MC 2012 con numero delle Aziende Agricole e Manifatturiere 2012 | 0.3724 | 0.3573 | 4.012e-09 | -0.18886 | 0.29798 | sì |
| 0 | Multivariata MP 2012 con numero delle Aziende Agricole e Manifatturiere 2012 | 0.4052 | 0.3908 | 4.336e-10 | -0.48512 | 0.72259 | sì |
| 0 | Multivariata MC 2016 con numero delle Aziende Agricole e Manifatturiere 2016 | 0.3749 | 0.3598 | 3.407e-09 | -0.20265 | 0.36199 | sì |
| 0 | Multivariata MP 2016 con numero delle Aziende Agricole e Manifatturiere 2016 | 0.4108 | 0.3966 | 2.92e-10 | -0.45715 | 0.78177 | sì |
| 0 | Multivariata MC 2020 con numero delle Aziende Agricole e Manifatturiere 2020 | 0.4208 | 0.4069 | 1.44E-10 | -0.25348 | 0.40153 | no |
| 0 | Multivariata MP 2020 con numero delle Aziende Agricole e Manifatturiere 2020 | **0.4963** | 0.4841 | 4.379e-13 | -0.54722 | 0.88767 | sì |
| 4 | Multivariata MC 2016 con numero delle Aziende Agricole e Manifatturiere 2012 | 0.3576 | 0.3421 | 1.057e-08 | -0.18495 | 0.35140 | sì |
| 4 | Multivariata MP 2016 con numero delle Aziende Agricole e Manifatturiere 2012 | 0.3971 | 0.3826 | 7.6e-10 | -0.42415 | 0.76229 | sì |

I modelli multivariati con la variabile dipendente "Media Penalizzata" mostrano valori di $R^2$ più elevati, indicando una maggiore varianza spiegata dalla variabile indipendente. Inoltre, i coefficienti delle variabili indipendenti e i valori p associati indicano che le variabili considerate sono significative nel modello.

È importante notare che la presenza di residui normali, come indicato nella colonna "Residui normali", è stata verificata nei modelli che hanno superato il test di normalità.

Attraverso l'analisi dei modelli di regressione lineare, sia univariati che multivariati, considerando diverse specificità come i lag, è possibile valutare l'importanza e la significatività delle variabili indipendenti nel fornire una spiegazione al fenomeno di interesse. Dai risultati ottenuti da questi modelli, emerge che l'utilizzo della Media Penalizzata come variabile dipendente risulta essere il più performante, mostrando una maggiore capacità di spiegare la variazione dei dati. Inoltre, si osserva che il lag 0 presenta un $R^2$ più elevato rispetto agli altri valori di lag, indicando una migliore aderenza dei dati al modello quando non viene introdotta una temporizzazione tra le variabili.

Successivamente, verranno analizzati e interpretati i risultati ottenuti dai modelli basati sulla media penalizzata. Questa analisi consentirà di comprendere in modo più approfondito le implicazioni e i significati dei risultati derivati da tali modelli.

### 5.4.1 Modello di regressione lineare con media penalizzata e interpretazione dei risultati per l'anno 2012:

La formula della regressione lineare utilizzata nel modello è:

$$Media\_Geometrica\_Penalizzata\_2012 = \beta_0 + \beta_1 * Aziende\_Agricole\_2012 + \beta_2 * Aziende\_Manifatturiere\_2012 \quad (26)$$

dove:

- $Media\_Geometrica\_Penalizzata\_2012$ è la variabile dipendente (variabile risposta) che si cerca di predire.
- $Aziende\_Agricole\_2012$ è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende nel settore dell'agricoltura nel 2012
- $Aziende\_Manifatturiere\_2012$ è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende manifatturiere nel 2012.
- I coefficienti $\beta_0$, $\beta_1$, $\beta_2$ rappresentano rispettivamente l'intercetta del modello, il coefficiente di regressione associato a $Aziende\_Agricole\_2012$ . e il coefficiente di regressione associato a $Aziende\_Manifatturiere\_2012$ .

Tabella 5.3 Tabella delle statistiche delle variabili originali 2012

| | **MP_geometric_2012** | **Aziende_Agricole_2012** | **Aziende_Manifatturiere_2012** |
|---|---|---|---|
| Primo Quartile | 0.22365 | 31 | 1942 |
| Mediana | 0.34484 | 63 | 3468 |
| Media | 0.34294 | 97.77 | 4646 |
| Terzo Quartile | 0.4809 | 138 | 5478 |

Tabella 5.4 Tabella delle statistiche delle variabili scalati 2012

| | **MP_geometric_2012** | **Aziende_Agricole_2012** | **Aziende_Manifatturiere_2012** |
|---|---|---|---|
| Primo Quartile | 0.3002 | 0.072443 | 0.07324 |
| Mediana | 0.4371 | 0.135417 | 0.12577 |
| Media | 0.4366 | 0.211747 | 0.175 |
| Terzo Quartile | 0.5952 | 0.303504 | 0.20643 |

La formula con i coefficienti sarebbe:

$$Media\_Geometrica\_Penalizzata\_2012 = 0.41288 - 0.48512 * Aziende\_Agricole\_2012 + 0.72259 * Aziende\_Manifatturiere\_2012 \quad (27)$$

dove:

- $Media\_Geometrica\_Penalizzata\_2012$ è la variabile dipendente
- $Aziende\_Agricole\_2012$ è la variabile indipendente normalizzata
- $Aziende\_Manifatturiere\_2012$ è la variabile indipendente normalizzata
- 0.41288, -0.48512, 0.72259 sono i coefficienti stimati per l'intercetta e le variabili indipendenti rispettivamente.

Tabella 5.5 Risultati della Regressione Lineare 2012

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.52772 -0.12013  0.01434  0.11688  0.45717

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0.41288    0.03278  12.594  < 2e-16 ***
W$df.Aziende_Agricole_2012.1.86.    -0.48512    0.09994  -4.854 5.61e-06 ***
W$df.Aziende_Manifatturiere_2012.1.86.  0.72259    0.11375   6.352 1.08e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1724 on 83 degrees of freedom
Multiple R-squared:  0.4052,    Adjusted R-squared:  0.3908
F-statistic: 28.27 on 2 and 83 DF,  p-value: 4.336e-10
```

Il coefficiente $R^2$ per il modello di regressione lineare è 0.4052, il che indica che il 40.52% della variazione nella variabile dipendente (Media Geometrica) può essere spiegato dalle variabili indipendenti (Numero delle Aziende Agricole e Aziende Manifatturiere) nel modello. Il $R^2\ corretto$ è 0.3908, il che indica che il 39.08% della variazione nella variabile dipendente può essere spiegato dalle variabili indipendenti nel modello, tenendo conto del numero di variabili indipendenti e delle osservazioni.

I p-value per entrambe le variabili indipendenti sono molto piccoli (p-value < 0.001), il che suggerisce che entrambe le variabili hanno un effetto significativo sull'indicatore di consumo di suolo.

Il residual standard error (errore standard residuo) è 0.1724, che rappresenta la stima della deviazione standard dei residui del modello. Indica la misura di quanto i punti dati del modello si discostano dai valori stimati.

Il test F ha un valore di 28.27 con 2 e 83 gradi di libertà, e un p-value molto piccolo (p-value < 0.001), il che suggerisce che il modello di regressione è statisticamente significativo e che almeno una delle variabili indipendenti ha un effetto significativo sull'indicatore di consumo di suolo.

Dal grafico dei residui e dalla tabella dei residui, possiamo osservare che i residui seguono una distribuzione normale

Tabella 5.6 Test Statistici 2012

| Test | Statistic | pvalue |
|---|---|---|
| Shapiro-Wilk | 0.988 | 0.6185 |
| Kolmogorov-Smirnov | 0.0448 | 0.9922 |
| Cramer-von Mises | 19.9882 | 0.0000 |
| Anderson-Darling | 0.2516 | 0.7318 |


Figura 5.1 Grafico OLS Residui-QQ - Distribuzione dei residui confrontata con la distribuzione normale teorica

**Interpretazione dei risultati**

I coefficienti stimati nel modello di regressione rappresentano la relazione stimata tra le variabili indipendenti (numero delle Aziende agricole e Manifatturiere) e la variabile dipendente (Media Geometrica).

Il coefficiente associato alla variabile Aziende_Agricole_2012 ( −0.48512) indica che, mantenendo costante la variabile Aziende_Manifatturiere_2012, un aumento di una unità nella variabile Aziende_Agricole_2012 corrisponde a una diminuzione di 0.48512 nella Media_Geometrica_Penalizzata_2012. Questo coefficiente indica la relazione lineare inversa tra la variabile Aziende_Agricole_2012 e la Media_Geometrica_Penalizzata_2012.

Il coefficiente associato alla variabile Aziende_Manifatturiere_2012 (0.72259) indica che, mantenendo costante la variabile Aziende_Agricole_2012, un aumento di una unità nella variabile Aziende_Manifatturiere_2012 corrisponde a un aumento di 0.72259 nella Media_Geometrica_Penalizzata_2012. Questo coefficiente indica la relazione lineare diretta tra la variabile Aziende_Manifatturiere_2012 e la Media_Geometrica_Penalizzata_2012.

### 5.4.2 Modello di regressione lineare con media penalizzata e interpretazione dei risultati per l'anno 2016

La formula della regressione lineare utilizzata nel modello è:

$$Media\_Geometrica\_Penalizzata\_2016 = \beta_0 + \beta_1 * Aziende\_Agricole\_2016 + \beta_2 * Aziende\_Manifatturiere\_2016 \quad (28)$$

dove:

- $Media\_Geometrica\_Penalizzata\_2016$ è la variabile dipendente (variabile risposta) che si cerca di predire.

- $Aziende\_Agricole\_2016$ è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende nel settore dell'agricoltura nel 2016.
- $Aziende\_Manifatturiere\_2016$ è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende manifatturiere nel 2016.
- I coefficienti $\beta_0$, $\beta_1$, $\beta_2$ rappresentano rispettivamente l'intercetta del modello, il coefficiente di regressione associato a $Aziende\_Agricole\_2016$ e il coefficiente di regressione associato a $Aziende\_Manifatturiere\_2016$.

Tabella 5.7 Tabella delle statistiche delle variabili originali 2016

| | **MP_geometric_2016** | **Aziende_Agricole_2016** | **Aziende_Manifatturiere_2016** |
|---|---|---|---|
| Primo Quartile | 0.22322 | 30 | 1760 |
| Mediana | 0.34374 | 66 | 3215 |
| Media | 0.34622 | 101.4 | 4313 |
| Terzo Quartile | 0.47874 | 135.5 | 5024 |

Tabella 5.8 Tabella delle statistiche delle variabili scalati 2016

| | **MP_geometric_2016** | **Aziende_Agricole_2016** | **Aziende Manifatturiere_2016** |
|---|---|---|---|
| Primo Quartile | 0.28547 | 0.074141 | 0.07215 |
| Mediana | 0.44258 | 0.141953 | 0.12434 |
| Media | 0.42474 | 0.209386 | 0.1706 |
| Terzo Quartile | 0.57172 | 0.30425 | 0.19103 |

La formula con i coefficienti sarebbe:

$$Media\_Geometrica\_Penalizzata\_2016 = 0.38708 - 0.45715 * Aziende\_Agricole\_2016 + 0.78177 * Aziende\_Manifatturiere\_2016 \quad (29)$$

dove:

- $Media\_Geometrica\_Penalizzata\_2016$ è la variabile dipendente
- $Aziende\_Agricole\_2016$ è la variabile indipendente normalizzata
- $Aziende\_Manifatturiere\_2016$ è la variabile indipendente normalizzata

- 0.38708, −0.45715, 0.78177 sono i coefficienti stimati per l'intercetta e le variabili indipendenti rispettivamente.

Tabella 5.9 Risultati della Regressione Lineare 2016

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.41214 -0.10813 -0.01318  0.12468  0.36794

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0.38708    0.03251  11.908  < 2e-16 ***
W$df.Aziende_Agricole_2016.1.86.    -0.45715    0.10179  -4.491 2.27e-05 ***
W$df.Aziende_Manifatturiere_2016.1.86.  0.78177    0.11542   6.773 1.68e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1728 on 83 degrees of freedom
Multiple R-squared:  0.4108,    Adjusted R-squared:  0.3966
F-statistic: 28.94 on 2 and 83 DF,  p-value: 2.92e-10
```

Il coefficiente $R^2$ è pari a 0.4108, il che significa che circa il 41.08% della variazione della variabile Media Geometrica Penalizzata può essere spiegata dalle variabili Aziende_Agricole_2016 e Aziende_Manifatturiere_2016. Il $R^2$ $corretto$ è 0.3966, il che indica che il 39.66% della variazione nella variabile dipendente può essere spiegato dalle variabili indipendenti nel modello, tenendo conto del numero di variabili indipendenti e delle osservazioni.

I p-value sono tutti molto piccoli (indicati come 2e-16, 2.27e-05, 1.68e-09), il che suggerisce che esiste una relazione significativa tra le variabili indipendenti e la variabile dipendente.

L'errore standard residuo è pari a 0.1728, il che significa che in media le previsioni del modello si discostano di circa 0.1728 unità dai valori osservati.

L'F-statistic è pari a 28.94, con un p-value molto piccolo (2.92e-10). Questo indica che il modello di regressione nel suo complesso è statisticamente significativo e che almeno una delle variabili indipendenti contribuisce in modo significativo alla spiegazione della variazione della variabile dipendente.

Dal grafico dei residui e dalla tabella dei residui, possiamo osservare che i residui seguono approssimativamente una distribuzione normale

Tabella 5.10 Test Statistici 2016

| Test | Statistic | pvalue |
|---|---|---|
| Shapiro-Wilk | 0.9903 | 0.7804 |
| Kolmogorov-Smirnov | 0.0509 | 0.9710 |
| Cramer-von Mises | 19.7069 | 0.0000 |
| Anderson-Darling | 0.2456 | 0.7518 |

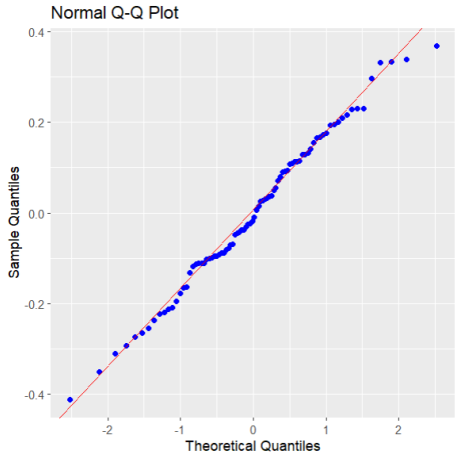


Figura 5.2 Grafico OLS Residui-QQ - Distribuzione dei residui confrontata con la distribuzione normale teorica

**Interpretazione dei risultati**

Il coefficiente per la variabile "Aziende_Agricole_2016." è stimato a −0.45715. Questo indica che, tenendo costante il valore della variabile "Aziende_Manifatturiere_2016", ogni unità di aumento nel numero delle Aziende agricole è associata a una diminuzione media di 0.45715 nell'indicatore di consumo di suolo.

Il coefficiente per la variabile "Aziende_Manifatturiere_2016" è stimato a 0.78177. Questo indica che, tenendo costante il valore della variabile "Aziende_Agricole_2016.", ogni unità di aumento nel numero delle Aziende

manifatturiere è associata a un aumento medio di 0.78177 nell'indicatore di consumo di suolo.

### 5.4.3 Applicazione di modelli di regressione lineare con media penalizzata e interpretazione dei risultati per l'anno 2020

La formula della regressione lineare utilizzata nel modello è:

$$Media\_Geometrica\_Penalizzata\_2020 = \beta_0 + \beta_1 * Aziende\_Agricole\_2020 + \beta_2 * Aziende\_Manifatturiere\_2020 \quad (30)$$

dove:

- Media_Geometrica_Penalizzata_2020 è la variabile dipendente (variabile risposta) che si cerca di predire.
- Aziende_Agricole_2020 è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende nel settore dell'agricoltura nel 2020.
- Aziende_Manifatturiere_2020 è la variabile indipendente (variabile esplicativa) che rappresenta il numero di aziende manifatturiere nel 2020
- I coefficienti $\beta_0$ , $\beta_1$ , $\beta_2$ rappresentano rispettivamente l'intercetta del modello, il coefficiente di regressione associato a $Aziende\_Agricole\_2020$ e il coefficiente di regressione associato a $Aziende\_Manifatturiere\_2020$.

Tabella 5.11 Tabella delle statistiche delle variabili originali 2020

| | MP_geometric_2020 | Aziende_Agricole_2020 | Aziende_Manifatturiere_2020 |
|---|---|---|---|
| Primo Quartile | 0.20707 | 34 | 1644 |
| Mediana | 0.30243 | 69 | 3041 |
| Media | 0.33754 | 103.7 | 4065 |
| Terzo Quartile | 0.46246 | 142.5 | 4724 |

Tabella 5.12 Tabella delle statistiche delle variabili scalati

| | MP_geometric_2020 | Aziende_Agricole_2020 | Aziende_Manifatturiere_2020 |
|---|---|---|---|
| Primo Quartile | 0.29749 | 0.071192 | 0.07297 |
| Mediana | 0.38345 | 0.139073 | 0.12836 |
| Media | 0.43769 | 0.195634 | 0.1742 |
| Terzo Quartile | 0.55622 | 0.274834 | 0.20272 |

La formula con i coefficienti sarebbe:

$$\text{Media\_Geometrica\_Penalizzata\_2020} = 0.39805 - 0.58780 * \text{Aziende\_Agricole\_2020} + 0.88767 * \text{Aziende\_Manifatturiere\_2020} \quad (31)$$

dove:

- $Media\_Geometrica\_Penalizzata\_2020$ è la variabile dipendente
- $Aziende\_Agricole\_2020$ è la variabile indipendente normalizzata
- $Aziende\_Manifatturiere\_2020$ è la variabile indipendente normalizzata
- 0.39805, −0.58780, 0.88767 sono i coefficienti stimati per l'intercetta e le variabili indipendenti rispettivamente.

Tabella 5.13 Risultati della Regressione Lineare 2020

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.46082 -0.09384 -0.02005  0.06959  0.52376

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0.39805    0.03224  12.347  < 2e-16 ***
W$df.Aziende_Agricole_2020.1.86.    -0.58780    0.10756  -5.465 4.76e-07 ***
W$df.Aziende_Manifatturiere_2020.1.86.  0.88767    0.11171   7.946 8.33e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1713 on 83 degrees of freedom
Multiple R-squared:  0.4963,    Adjusted R-squared:  0.4841
F-statistic: 40.88 on 2 and 83 DF,  p-value: 4.379e-13
```

Il coefficiente $R^2$ del modello è 0.4963, il che indica che il 49.63% della variazione nella variabile dipendente può essere spiegato dalle variabili indipendenti nel modello. Il $R^2$ $corretto$ è 0.4841, il che indica che il 48.41% della variazione nella variabile dipendente può essere spiegato dalle variabili indipendenti nel modello, tenendo conto del numero di variabili indipendenti e delle osservazioni.

I p-value associati ai coefficienti delle variabili indipendenti sono tutti molto piccoli (<0.05), indicando una forte evidenza statistica che le variabili indipendenti hanno un impatto significativo sulla variabile dipendente.

La deviazione standard dei residui (residual standard error) è 0.1713, che rappresenta la misura della discrepanza tra i valori osservati e i valori predetti dal modello.

Il test F con un valore di 40.88 e un p-value di 4.379e-13 conferma che il modello nel complesso è statisticamente significativo e fornisce un miglior adattamento rispetto a un modello con solo l'intercetta.

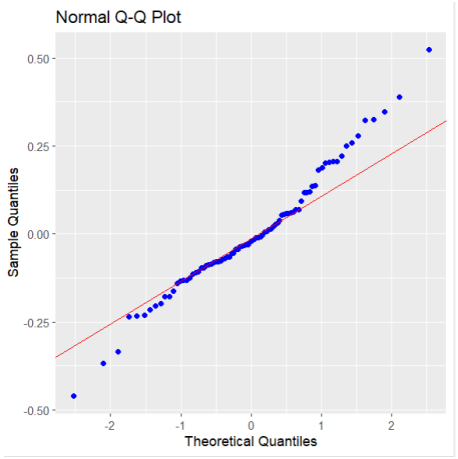


Figura 5.3 Grafico OLS Residui-QQ - Distribuzione dei residui confrontata con la distribuzione normale teorica

Tabella 5.14 Test Statistici 2020

| Test | Statistic | pvalue |
|---|---|---|
| Shapiro-Wilk | 0.9773 | 0.1337 |
| Kolmogorov-Smirnov | 0.0962 | 0.3796 |
| Cramer-von Mises | 20.3767 | 0.0000 |
| Anderson-Darling | 0.8481 | 0.0279 |

Dal grafico dei residui e dalla tabella dei residui, possiamo osservare che i residui seguono approssimativamente una distribuzione normale. Nel grafico, i punti dei residui sono distribuiti intorno alla linea di riferimento, senza evidenti pattern o deviazioni sistematiche. Nella tabella dei residui, i valori dei residui non mostrano deviazioni significative dalla media e non presentano evidenti pattern o strutture.

**Interpretazione dei risultati**

L'intercetta (0.39805) rappresenta il valore previsto della Media_Geometrica_Penalizzata_2020 quando entrambe le variabili indipendenti, Aziende_Agricole_2020 e Aziende_Manifatturiere_2020, sono pari a zero.

Il coefficiente associato alla variabile Aziende_Agricole_2020 ( −0.58780) indica che, mantenendo costante la variabile Aziende_Manifatturiere_2020, un aumento

di una unità nella variabile Aziende_Agricole_2020 corrisponde a una diminuzione di 0.58780 nella Media_Geometrica_Penalizzata_2020. Questo coefficiente indica la relazione lineare inversa tra la variabile Aziende_Agricole_2020 e la media geometrica penalizzata.

Il coefficiente associato alla variabile Aziende_Manifatturiere_2020 (0.88767) indica che, mantenendo costante la variabile Aziende_Agricole_2020, un aumento di una unità nella variabile Aziende_Manifatturiere_2020 corrisponde a un aumento di 0.88767 nella Media_Geometrica_Penalizzata_2020. Questo coefficiente indica la relazione lineare diretta tra la variabile Aziende_Manifatturiere_2020 e la Media_Geometrica_Penalizzata_2020.

## 5.5 Conclusione dell'Analisi della Regressione Lineare

Nei modelli di regressione multivariata, i valori di $R^2$ sono significativamente elevati, indicando una buona capacità dei modelli di spiegare la variazione dei dati. I p-value associati ai coefficienti sono molto bassi, confermando la loro significatività statistica. Rispetto ai modelli con la variabile "Media Classica" (MC), i modelli con la variabile "Media Penalizzata" (MP) presentano un $R^2$ leggermente superiore. Ad esempio, il modello Multivariata MP 2012 contro Aziende 2012 ha un $R^2$ di 0.4052, mentre il modello Multivariata MC 2012 contro Aziende 2012 ha un $R^2$ di 0.3724. Inoltre, i p-value dei coefficienti nei modelli MP sono generalmente più bassi rispetto ai modelli MC, indicando una maggiore significatività statistica dei coefficienti.

I modelli multivariati presentano un $R^2$ più elevato e una maggiore varietà di coefficienti rispetto ai modelli univariati. Ciò indica che i modelli multivariati

offrono una migliore spiegazione della variazione dell'Indice del consumo del suolo utilizzando più variabili indipendenti.

I risultati dell'analisi indicano un significativo impatto del numero di aziende agricole e industriali sulla variabile di destinazione. Un aumento del numero di aziende agricole si traduce in una diminuzione del valore del consumo di suolo, mentre un aumento del numero di aziende industriali porta ad un aumento del valore del consumo di suolo

# Capitolo 6 Conclusioni

La tesi si è focalizzata su un aspetto metodologico: l'introduzione di un clustering basato su una nuova distanza ottenuta combinando la media geometrica penalizzata dei dati di una unità, la distanza di Canberra e il metodo di Ward. I risultati hanno mostrato che l'utilizzo della media geometrica penalizzata permette una clusterizzazione più efficace, consentendo di distinguere con maggiore precisione i gruppi di osservazioni.

Nel contesto del clustering dei gruppi di province con diversi livelli di consumo di suolo, l'approccio della media geometrica penalizzata si rivela particolarmente rilevante poiché penalizza le osservazioni caratterizzate da un consumo di suolo elevato e con alta variabilità. Ciò permette di ottenere una distanza che considera in modo adeguato la particolare sensibilità e importanza del consumo di suolo nella formazione dei gruppi.

L'introduzione di una distanza basata sulla media geometrica penalizzata ha arricchito il panorama metodologico, offrendo una prospettiva innovativa per la

valutazione delle relazioni tra le osservazioni nel contesto del consumo di suolo. Questo approccio ha il potenziale per migliorare la comprensione e la capacità di analizzare la distribuzione dei gruppi.

L'introduzione di soglie basate sui valori massimi e minimi dei gruppi identificati ha fornito un metodo per classificare le province in base al consumo di suolo, consentendo di identificare e monitorare le province con consumi elevati o sostenibili.

I risultati dell'analisi empirica hanno evidenziato due gruppi significativamente differenti nel consumo di suolo, con province che si ripetono all'interno di essi nel corso degli anni. Il gruppo A è caratterizzato da un consumo di suolo eccessivo e comprende province popolate con un alto sviluppo economico, mentre il gruppo B è localizzato in zone montane con un consumo di suolo sostenibile. Gli altri due gruppi, C e D, presentano rispettivamente un consumo soddisfacente e insoddisfacente, con alcune province che mostrano il rischio di sviluppare un consumo di suolo eccessivo.

L'analisi della regressione ha confermato l'importante impatto del numero di aziende agricole e industriali sul consumo di suolo. Un aumento delle aziende agricole è associato a una diminuzione del consumo di suolo, mentre un aumento delle aziende industriali comporta un aumento del consumo di suolo. Queste relazioni sottolineano l'importanza di considerare il ruolo delle attività agricole e industriali nella gestione sostenibile del suolo e suggeriscono l'adozione di politiche che promuovano una maggiore sostenibilità ambientale.

In definitiva, l'utilizzo della media geometrica penalizzata rappresenta un importante progresso metodologico che apre nuove prospettive per la ricerca e

l'applicazione pratica nel campo dell'analisi dei dati territoriali e della pianificazione ambientale.

# Ringraziamenti

Desidero esprimere i miei sinceri ringraziamenti alla Prof.ssa Maria Cristina Recchioni, il mio relatore di tesi, per la sua preziosa guida e supporto durante tutto il percorso di ricerca. La sua esperienza e competenza sono state fondamentali per la realizzazione di questo lavoro e mi ha fornito una prospettiva illuminante sulla tematica trattata.

In particolare, desidero ringraziarla per avermi introdotto al nuovo indicatore Media Penalizzata e per avermi fornito le competenze necessarie per utilizzarlo nel contesto della mia ricerca. Grazie alla sua chiarezza nell'esplicare i concetti e nella sua fiducia in me, sono stata in grado di applicare con successo questo strumento innovativo nella mia analisi.

Vorrei inoltre esprimere il mio profondo apprezzamento a tutte le organizzazioni che si occupano del problema del suolo, in particolare ai volontari del movimento “SaveSoil”, per il loro impegno e la loro dedizione a preservare e proteggere la terra. Il loro lavoro instancabile e la loro passione per l'ambiente sono stati una fonte di ispirazione per me e hanno rafforzato la mia convinzione nell'importanza di affrontare le sfide legate alla salute del suolo.

Ringrazio tutti coloro che hanno contribuito in vario modo alla realizzazione di questa tesi, compresi i miei familiari, per il loro costante sostegno e incoraggiamento lungo tutto il percorso. Senza il loro supporto, non avrei potuto portare a termine questo lavoro con successo.

# Bibliografia


Barbera, F., Gallerano, L., Nicoletti, A., & Raimond, S. (2020, Maggio). Biodiversità a rischio.

Bot, A., & Benites, J. (2005). The importance of soil organic matter. *FAO SOILS BULLETIN 80*.

Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*.

Bruegmann, R. (2005). *Sprawl. A Compact History.* The University of Chicago Press.

Dellasala, D., & Goldstein, M. (2018). *Encyclopedia of the Anthropocene.*

L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) . (2016, Luglio). Consumo di suolo, dinamiche territoriali e servizi ecosistemici.

L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). (2021, Ottobre). Bilancio Di Sostenibilità.

l'Agenzia delle Nazioni Unite per l'Agricoltura. (2019). Soil erosion must be stopped 'to save our future'. *UN News*.

Lance, G., & Williams. (1967). The Canberra distance metric.

Moretti, L., & Loprencipe, G. (2018, Novembre). Climate Change and Transport Infrastructures: State of the Art.

*Movimprese*. (s.d.). Tratto da https://www.infocamere.it/

Munafò, M. (2022, Luglio). Consumo di suolo, dinamiche territoriali e servizi ecosistemici. Edizione 2022. Report SNPA 32/22.

Otoiu, A., Pareto, A., Grimaccia, E., Mazziotta, M., & Terzi, S. (2021). *Open issues in composite indicators. A starting point and a reference on some state-of-the-art issues.* ROMA TRE-PRESS.

Recchioni, M., Mariani, F., & Ciommi, M. (2023 (sottomesso)). A New Class of Composite Indicators: The Penalized Power Mean.

Shefali, A., Pankaj , K., & Kanchan, D. (2020). Coronavirus lockdown helped the environment to bounce back. *The Science of the Total Environment*.