



UNIVERSITÀ
DI SIENA
1240

DIPARTIMENTO DI ECONOMIA POLITICA E STATISTICA
SCUOLA DI ECONOMIA E MANAGEMENT

CORSO DI LAUREA IN SCIENZE ECONOMICHE E BANCARIE

AFFARI E LAVORO: ANALISI STATISTICA MULTIVARIATA DELLE PROVINCE ITALIANE

Relatrice
Prof.ssa Marzia Marcheselli

Correlatrice
Prof.ssa Caterina Pisani

Candidato
Dario Macii

Anno accademico 2023/2024

Indice

Introduzione.....	1
1 Il Dataset	2
1.1 Analisi descrittiva degli indicatori.....	2
2 Analisi delle componenti principali	26
2.1 Matrice di correlazione e componenti principali	26
2.2 Biplot	28
2.3 Indice sintetico delle province basato sulla prima componente principale	30
3 Clustering.....	34
3.1 Matrice delle distanze	34
3.2 Il metodo gerarchico agglomerativo e costruzione del dendrogramma.....	35
3.3 Descrizione dei cluster	37
Conclusioni	42
Bibliografia	43
Siti internet consultati.....	43

Indice delle figure

Figura 1.1.1 - Distribuzione delle startup innovative ogni 1000 società di capitale nelle province italiane.....	4
Figura 1.1.2 - Distribuzione delle imprese che fanno ecommerce ogni 100 imprese registrate nelle province italiane.....	5
Figura 1.1.3 - Distribuzione di imprese con titolare under 35 ogni 100 imprese registrate nelle province italiane.....	7
Figura 1.1.4 – Distribuzione delle nuove iscrizioni ogni 100 imprese registrate nelle province italiane.....	8
Figura 1.1.5 – Distribuzione delle cessazioni ogni 100 imprese registrate nelle province italiane ...	10
Figura 1.1.6 – Distribuzione delle imprese in fallimento ogni 100 imprese registrate nelle province italiane.....	11
Figura 1.1.7 – Distribuzione delle imprese straniere ogni 100 imprese registrate nelle province italiane.....	13
Figura 1.1.8 – Distribuzione del tasso di occupazione nelle province italiane	14
Figura 1.1.9 – Distribuzione della percentuale di Neet nelle province italiane	16
Figura 1.1.10 – Distribuzione del gender pay gap nelle province italiane.....	17
Figura 1.1.11 – Distribuzione dei lavoratori domestici ogni 1000 abitanti nelle province italiane ...	19
Figura 1.1.12 – Distribuzione della quota di export sul Pil in percentuale nelle province italiane ...	20
Figura 1.1.13 – Distribuzione della partecipazione alla formazione continua in percentuale nelle province italiane.....	22
Figura 1.1.14 – Distribuzione degli infortuni sul lavoro ogni 10'000 occupati nelle province italiane	23
Figura 1.1.15 – Distribuzione delle pensioni di vecchiaia ogni 1000 abitanti.....	25
Figura 2.1.1 - Matrice di correlazione.....	27
Figura 2.2.1 - Biplot.....	29
Figura 3.1.1 - Matrice delle distanze.....	35
Figura 3.2.1 - Dendrogramma.....	37
Figura 3.3.1 - Mappa delle province italiane colorate per cluster.....	38
Figura 3.3.2 – Grafici a stella dei cluster	39

Indice delle tabelle

Tabella 1.1.1 – Principali statistiche descrittive della variabile Startup innovative	3
Tabella 1.1.2 - Principali statistiche descrittive della variabile Imprese che fanno ecommerce.....	3
Tabella 1.1.3 - Principali statistiche descrittive della variabile Imprenditorialità giovanile.....	6
Tabella 1.1.4 - Principali statistiche descrittive della variabile Nuove iscrizioni	6
Tabella 1.1.5 - Principali statistiche descrittive della variabile Cessazioni	9
Tabella 1.1.6 - Principali statistiche descrittive della variabile Imprese in fallimento	9
Tabella 1.1.7 - Principali statistiche descrittive della variabile Imprese straniere	12
Tabella 1.1.8 - Principali statistiche descrittive della variabile Tasso di occupazione.....	12
Tabella 1.1.9 - Principali statistiche descrittive della variabile Giovani che non studiano e non lavorano (Neet)	15
Tabella 1.1.10 - Principali statistiche descrittive della variabile Gender pay gap	15
Tabella 1.1.11 - Principali statistiche descrittive della variabile Lavoratori domestici	18
Tabella 1.1.12 - Principali statistiche descrittive della variabile Quota di export sul Pil	18
Tabella 1.1.13 - Principali statistiche descrittive della variabile Partecipazione alla formazione continua.....	21
Tabella 1.1.14 - Principali statistiche descrittive della variabile Infortuni sul lavoro	21
Tabella 1.1.15 - Principali statistiche descrittive della variabile Numero pensioni di vecchiaia.....	24
Tabella 2.3.1 - Pesi dell'autovettore associato alla prima componente principale	31
Tabella 2.3.2 - Punteggi della prima componente principale delle province italiane	31
Tabella 3.3.1 – Media aritmetica delle variabili dei cluster	41

Introduzione

Descrivere la situazione lavorativa di un Paese è un’operazione che richiede una grande quantità di informazioni per catturare tutte le sfaccettature del mondo del lavoro. L’Italia in particolare, è caratterizzata da una forte eterogeneità territoriale, con differenze anche nel contesto lavorativo. L’obiettivo di questo elaborato è quindi quello di analizzare la situazione lavorativa e delle imprese delle province italiane utilizzando i 15 indicatori della categoria “Affari e lavoro” creata dal Sole 24 Ore all’interno dell’indagine sulla qualità della vita del 2023 delle province italiane.

Le tecniche di analisi multivariata come l’analisi delle componenti principali e il clustering hanno un ruolo centrale nella comprensione del contesto economico e lavorativo delle province italiane: la prima consente di ridurre la complessità dei dati riassumendo le informazioni in un insieme di poche variabili, mentre il clustering permette di individuare gruppi di unità con caratteristiche simili, facilitando così il confronto tra le province.

La tesi si articola quindi in tre capitoli: nel primo viene effettuata una descrizione degli indicatori utilizzati nell’analisi, fornendo sia alcune statistiche che ne sintetizzano la distribuzione che informazioni sulla loro distribuzione geografica. Nel secondo capitolo viene svolta l’analisi delle componenti principali, con l’obiettivo di ridurre la dimensionalità del dataset originale e di facilitare l’interpretazione dei dati, anche attraverso l’utilizzo del biplot, che permette di ridurre la complessità del dataset in sole due dimensioni in modo da utilizzare un piano cartesiano per visualizzare i dati. Verrà inoltre creato un indice sintetico di ciascuna provincia così da poter creare un punteggio per confrontarne le performance. Infine, nel terzo capitolo viene svolta l’analisi dei cluster con l’obiettivo di identificare gruppi di province con caratteristiche omogenee e per osservare le differenze che ci sono tra le varie aree del paese. Sarà inoltre effettuata una descrizione dei cluster per confrontare le caratteristiche dei gruppi ottenuti.

Capitolo 1

1 Il Dataset

Il Sole 24 Ore conduce annualmente l'indagine sulla qualità della vita, che valuta le condizioni socioeconomiche delle province italiane attraverso 90 indicatori. Gli indicatori sono suddivisi in sei macrocategorie, ognuna delle quali descrive un diverso aspetto del benessere: “Ricchezza e consumi”, “Affari e lavoro”, “Demografia e società”, “Ambiente e servizi”, “Giustizia e sicurezza” e “Cultura e tempo libero”. I dati utilizzati per effettuare l'indagine sulla qualità della vita 2023 sono stati acquisiti dal Sole 24 Ore da fonti ufficiali ed istituti di ricerca e sono stati pubblicati nella pagina GitHub del Sole 24 Ore¹. Dei 90 indicatori presenti, sono stati selezionati i 15 appartenenti alla macrocategoria “Affari e lavoro”.

È stato quindi ottenuto un dataset di 15 colonne, pari al numero di indicatori utilizzati, e 107 righe corrispondenti al numero delle province italiane. Il dataset contiene solo variabili quantitative e non presenta valori mancanti.

1.1 Analisi descrittiva degli indicatori

Per effettuare l'analisi descrittiva degli indicatori sono stati utilizzati i principali indici di posizione: media aritmetica, minimo, primo quartile, mediana, terzo quartile e massimo, con l'obiettivo di fornire una panoramica della distribuzione dei valori, e il coefficiente di variazione, un indice di dispersione adimensionale che permette di confrontare variabili con unità di misura diverse. È stata inoltre utilizzata una mappa coropletica per visualizzare la distribuzione geografica degli indicatori relativi alle province italiane.

Per effettuare l'analisi è stato utilizzato il software statistico R (R Core Team, 2023): gli indici di posizione sono stati calcolati attraverso la funzione *summary()*, mentre la mappa coropletica è stata creata utilizzando il file geografico aggiornato al 2023 relativo alle province italiane pubblicato dall'Istat² ed è stato letto utilizzando il pacchetto *sf*. La rappresentazione grafica della mappa coropletica è stata poi elaborata utilizzando il pacchetto *ggplot2*.

¹ Disponibile al link: <https://github.com/IlSole24ORE/QDV2023>

² Disponibile al link: <https://www.istat.it/notizia/confini-delle-unita-amministrative-a-fini-statistici-al-1-gennaio-2018-2/>

- **Startup innovative**

L'indicatore rappresenta il numero di startup innovative per ogni mille società di capitale. La fonte originale dei dati è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.1 – Principali statistiche descrittive della variabile Startup innovative

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
1.1	4.21	6.08	6.12	7.71	13.7	0.42

Il valore minimo registrato è di 1.1 nella provincia di Vercelli, mentre il valore massimo raggiunge 13.7 nella provincia di Milano. La media delle province è di 6.12, con una mediana leggermente inferiore pari a 6.08. Il 25% delle province presenta un valore inferiore a 4.21, mentre il 75% si colloca al di sotto di 7.71. Il coefficiente di variazione è pari a 0.42. Osservando la mappa coropletica in Figura 1.1.1, si rilevano i valori più elevati nelle province di Milano, Trieste e Terni, senza però evidenziare differenze territoriali significative.

- **Imprese che fanno ecommerce**

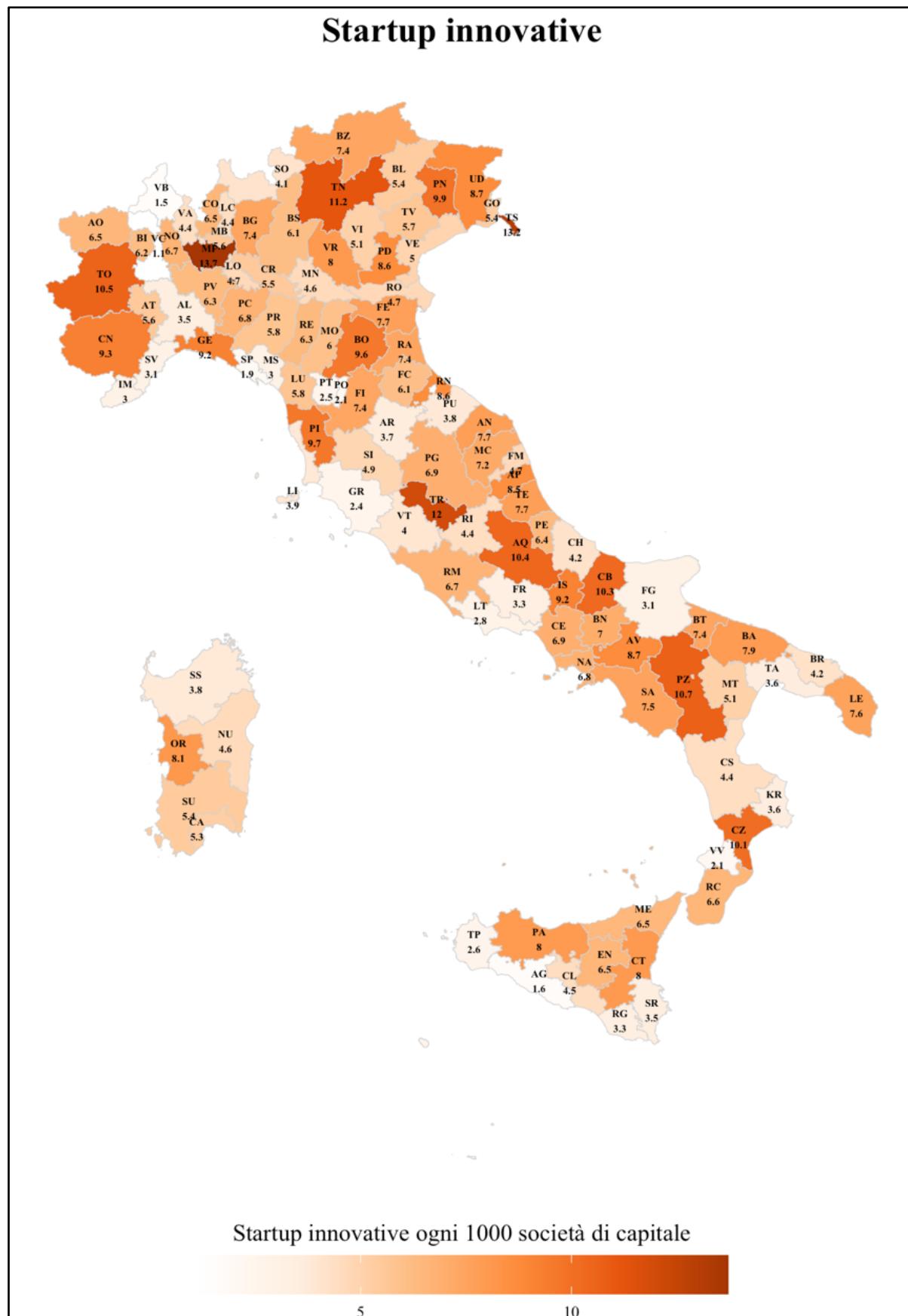
L'indicatore rappresenta il numero di imprese che fanno e-commerce ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.2 - Principali statistiche descrittive della variabile Imprese che fanno ecommerce

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
0	3.38	4.62	4.92	6.24	10.7	0.41

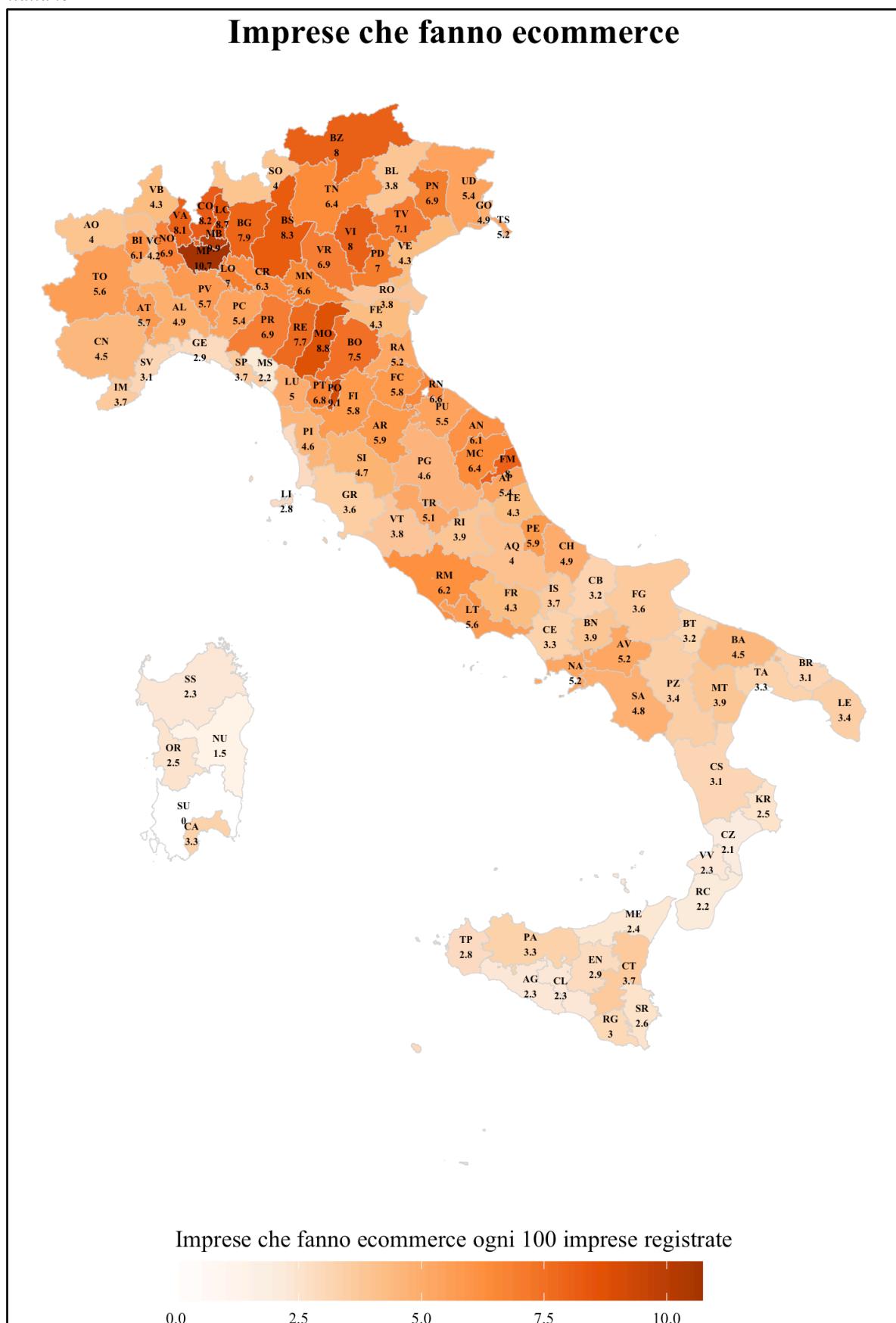
Il valore minimo registrato è di 0 nella provincia del Sud Sardegna, mentre il valore massimo raggiunge 10.7 nella provincia di Milano. La media delle province è di 4.92, con una mediana leggermente inferiore pari a 4.62. Il 25% delle province presenta un valore inferiore a 3.38, mentre il 75% si colloca al di sotto di 6.24. Il coefficiente di variazione è pari a 0.41. Osservando la mappa coropletica in Figura 1.1.2 si nota una maggiore concentrazione di imprese che fanno e-commerce nelle province del Centro-Nord con Milano seguito da Monza e Brianza e Prato.

Figura 1.1.1 - Distribuzione delle startup innovative ogni 1000 società di capitale nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

Figura 1.1.2 - Distribuzione delle imprese che fanno ecommerce ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

- **Imprenditorialità giovanile**

L'indicatore rappresenta il numero di imprese con titolare under 35, ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.3 - Principali statistiche descrittive della variabile Imprenditorialità giovanile

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
6.17	7.02	7.94	8.19	9.03	12.0	0.17

Il valore minimo registrato è di 6.17 nella provincia di Pesaro e Urbino, mentre il valore massimo raggiunge 12 nella provincia di Vibo Valentia. La media delle province è di 8.19, con una mediana leggermente inferiore pari a 7.94. Il 25% delle province presenta un valore inferiore a 7.02, mentre il 75% si colloca al di sotto di 9.03. Il coefficiente di variazione è pari a 0.17. Facendo riferimento alla mappa coropletica in Figura 1.1.3 si osserva un'impreditorialità giovanile più elevata nelle province del Sud.

- **Nuove iscrizioni**

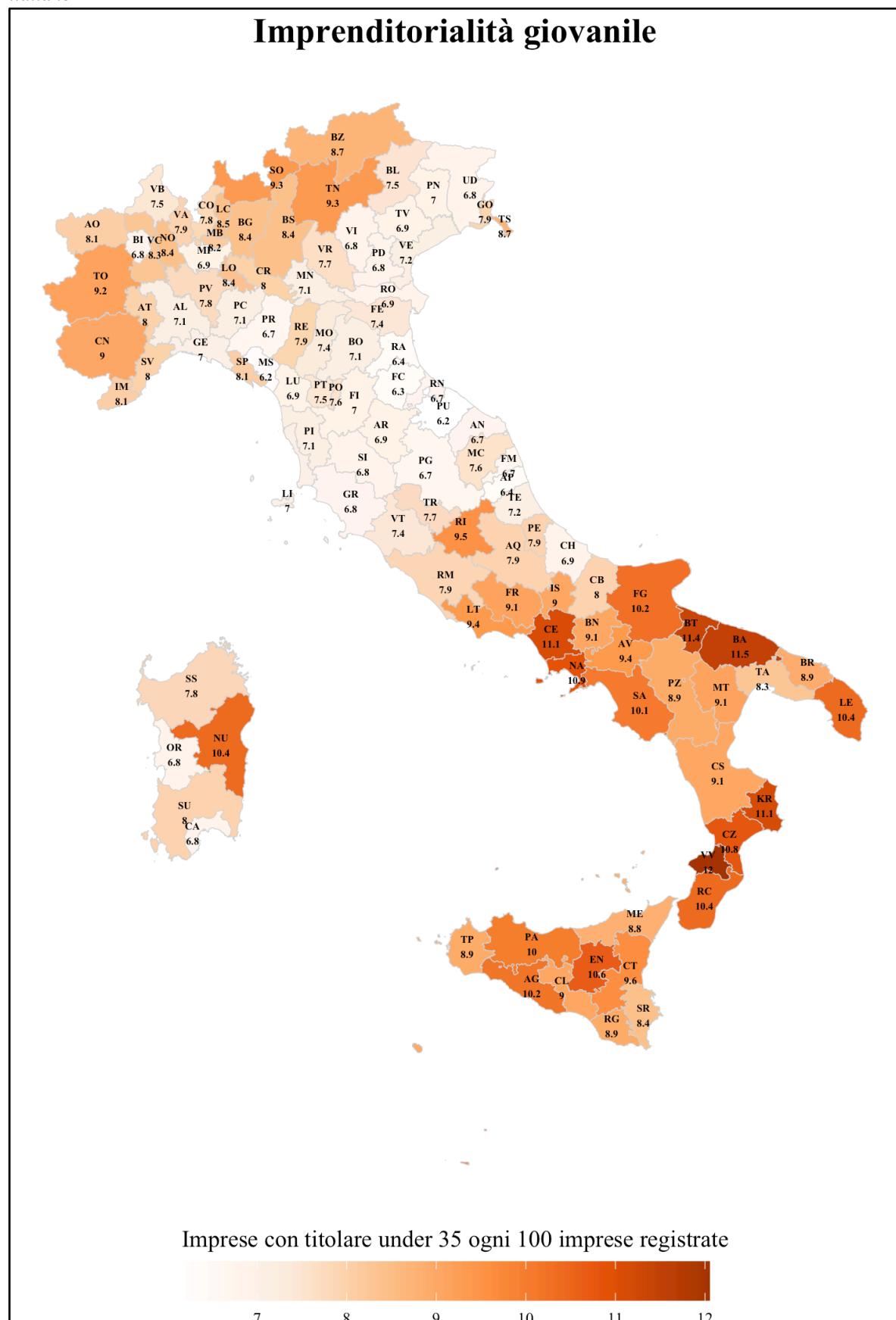
L'indicatore rappresenta il numero di nuove iscrizioni ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.4 - Principali statistiche descrittive della variabile Nuove iscrizioni

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
3.40	4.54	4.87	4.92	5.26	7.21	0.13

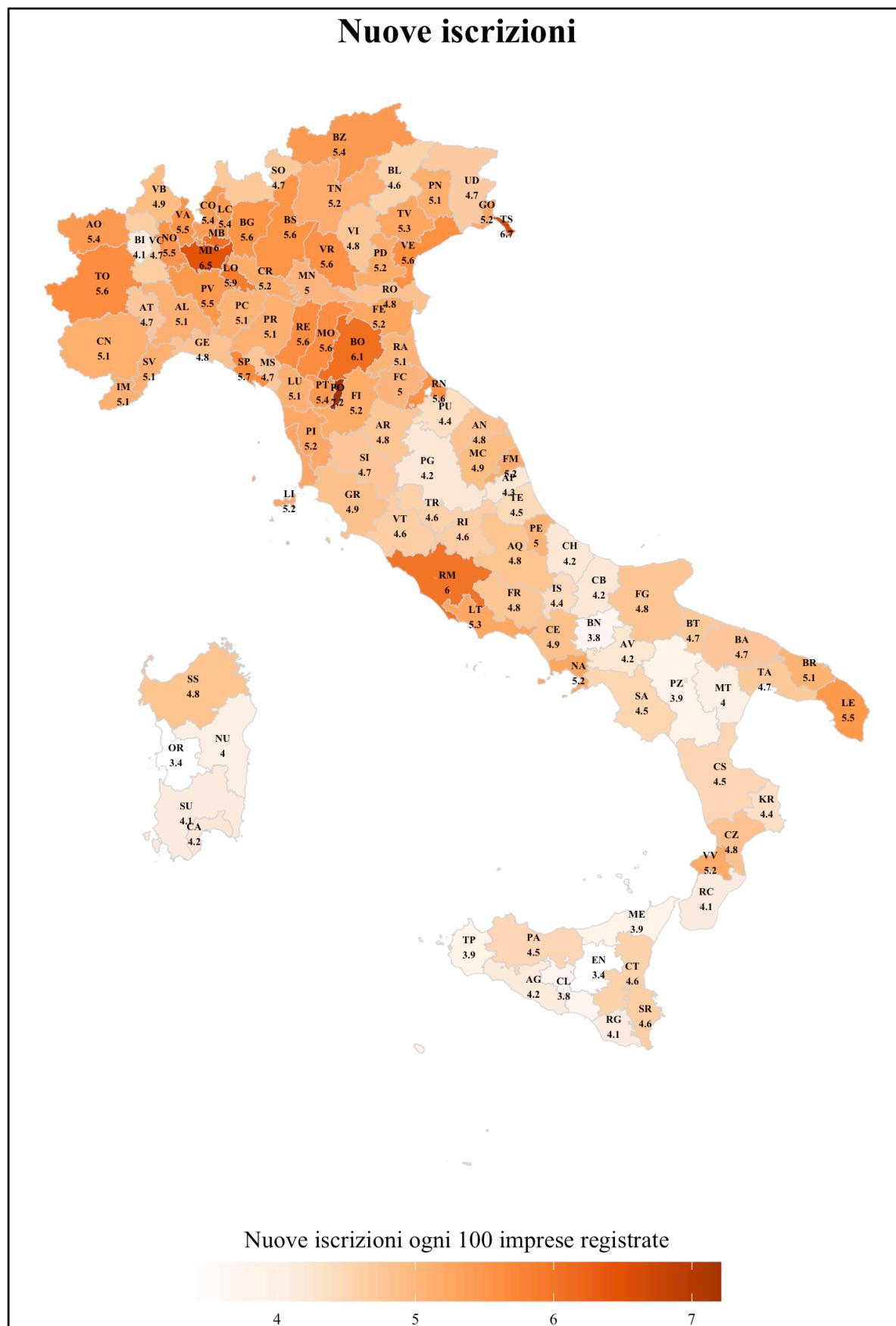
Il valore minimo registrato è di 3.4 nella provincia di Oristano, mentre il valore massimo raggiunge 7.21 nella provincia di Prato. La media delle province è di 4.92, con una mediana leggermente inferiore pari a 4.87. Il 25% delle province presenta un valore inferiore a 4.54, mentre il 75% si colloca al di sotto di 5.26. Il coefficiente di variazione è pari a 0.13. Osservando la mappa coropletica in Figura 1.1.4 si nota che le province con i valori più elevati di nuove iscrizioni si concentrano nel Centro-Nord, con Prato al primo posto, seguita da Trieste e Milano.

Figura 1.1.3 - Distribuzione di imprese con titolare under 35 ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

Figura 1.1.4 – Distribuzione delle nuove iscrizioni ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

- **Cessazioni**

L'indicatore rappresenta il numero di cessazioni ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.5 - Principali statistiche descrittive della variabile Cessazioni

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
3.12	4.12	4.59	4.58	5.11	6.44	0.14

Il valore minimo registrato è di 3.12 nella provincia di Oristano, mentre il valore massimo raggiunge 6.44 nella provincia di Prato. La media delle province è di 4.58, con una mediana leggermente superiore pari a 4.59. Il 25% delle province presenta un valore inferiore a 4.12, mentre il 75% si colloca al di sotto di 5.11. Il coefficiente di variazione è pari a 0.14. Osservando la Figura 1.1.5, la mappa coropletica rivela valori sensibilmente più alti nel Centro-Nord.

- **Imprese in fallimento**

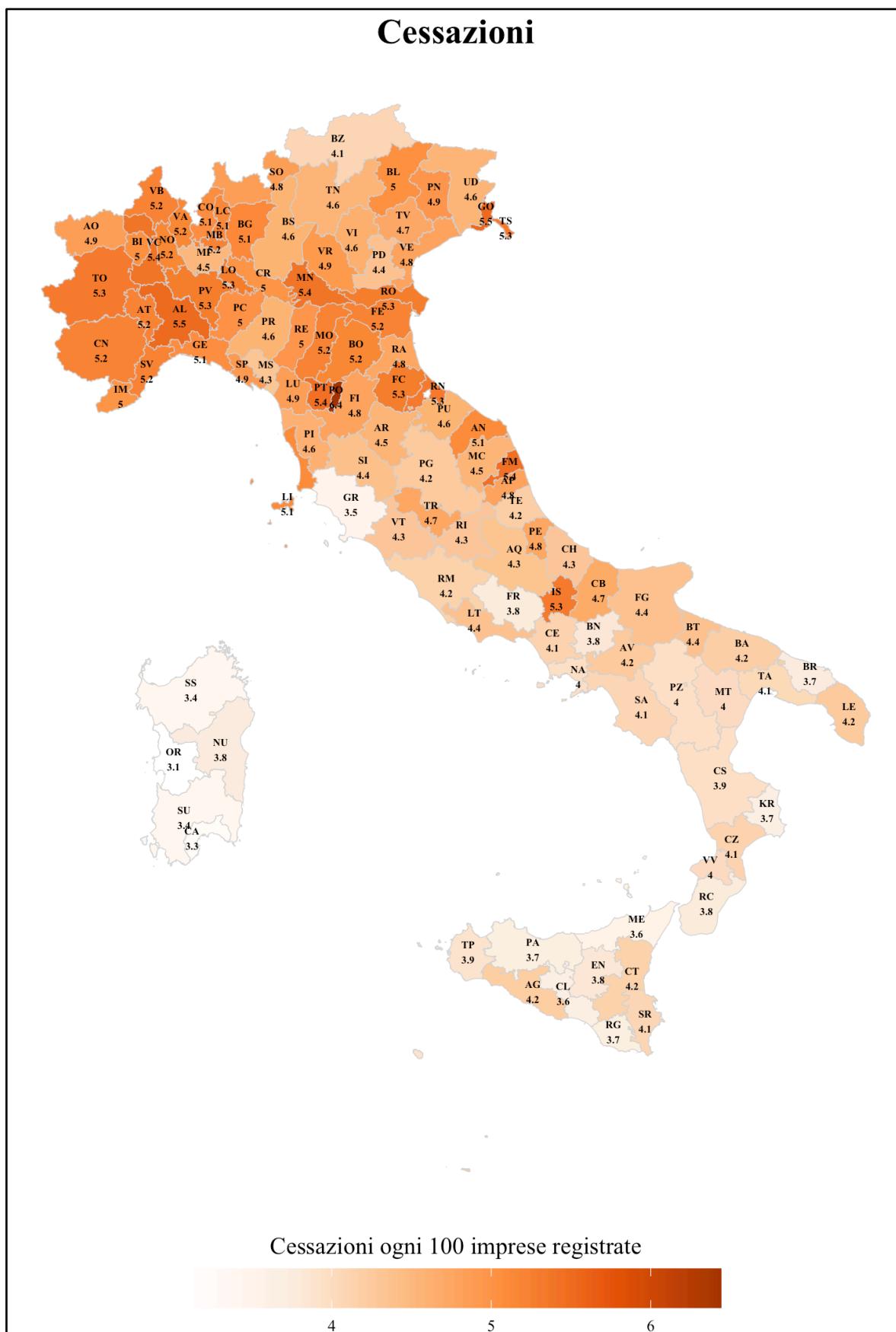
L'indicatore rappresenta il numero di imprese in fallimento ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.6 - Principali statistiche descrittive della variabile Imprese in fallimento

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
0.30	1.20	1.47	1.58	1.92	3.42	0.35

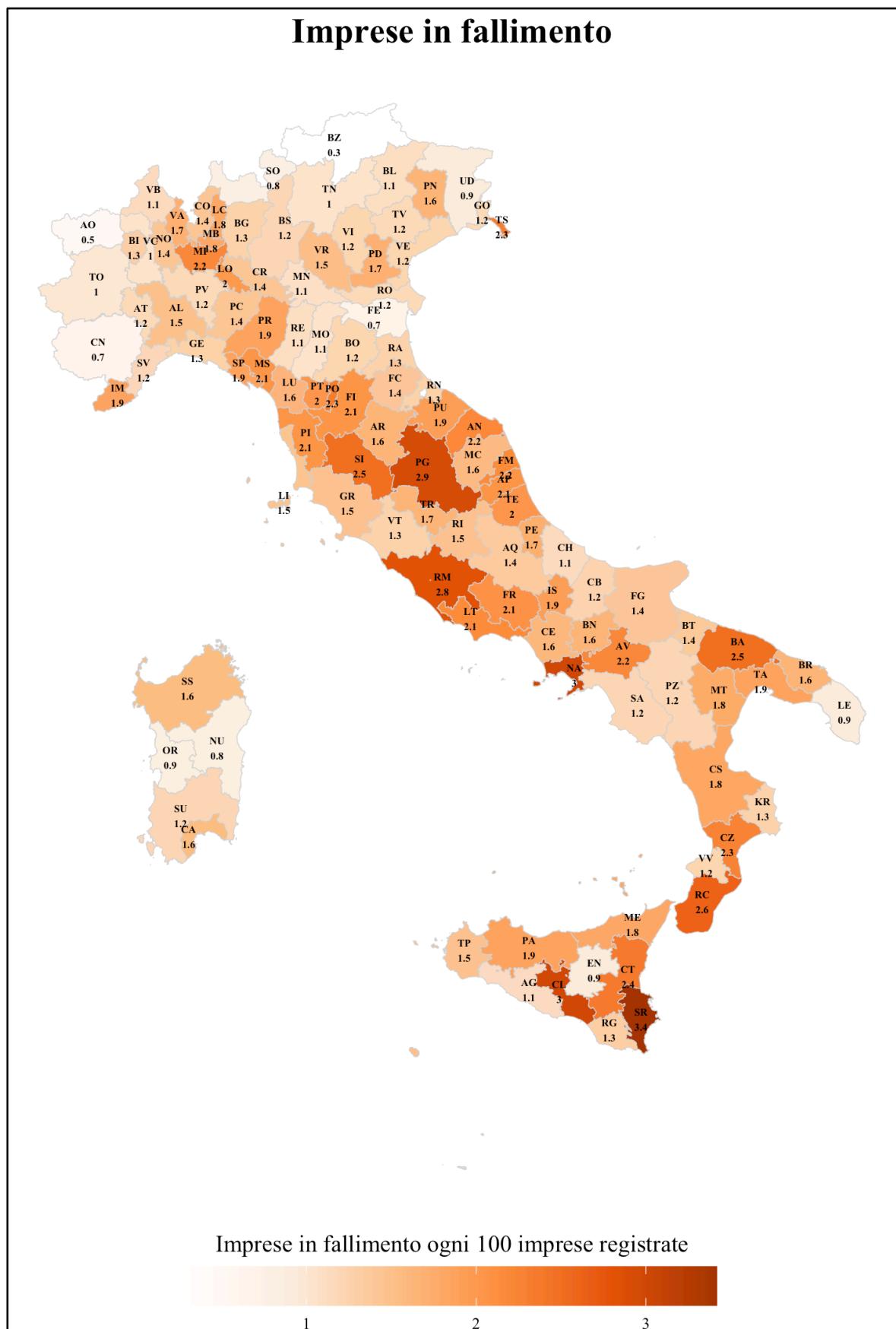
Il valore minimo registrato è di 0.3 nella provincia di Bolzano, mentre il valore massimo raggiunge 3.42 nella provincia di Siracusa. La media delle province è di 1.58, con una mediana leggermente inferiore pari a 1.47. Il 25% delle province presenta un valore inferiore a 1.2, mentre il 75% si colloca al di sotto di 1.92. Il coefficiente di variazione è pari a 0.35. Analizzando la mappa coropletica in Figura 1.1.6, si possono notare valori più alti nell'area del Centro-Sud.

Figura 1.1.5 – Distribuzione delle cessazioni ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

Figura 1.1.6 – Distribuzione delle imprese in fallimento ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

- **Imprese straniere**

L'indicatore rappresenta il numero di imprese straniere ogni 100 imprese registrate. La fonte originale è Infocamere e la rilevazione dell'indicatore risale al 30 settembre 2023.

Tabella 1.1.7 - Principali statistiche descrittive della variabile Imprese straniere

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
3.81	6.97	10.0	10.2	12.7	32.9	0.41

Il valore minimo registrato è di 3.81 nella provincia di Potenza, mentre il valore massimo raggiunge 32.9 nella provincia di Prato. La media delle province è di 10.2, con una mediana leggermente inferiore pari a 10. Il 25% delle province presenta un valore inferiore a 6.97, mentre il 75% si colloca al di sotto di 12.7. Il coefficiente di variazione è pari a 0.41. Osservando la mappa coroletica in Figura 1.1.7 si rilevano valori più elevati nel Centro-Nord, con Prato che presenta un valore significativamente superiore rispetto a Trieste, la seconda provincia con valore più alto, con 20.34.

- **Tasso di occupazione**

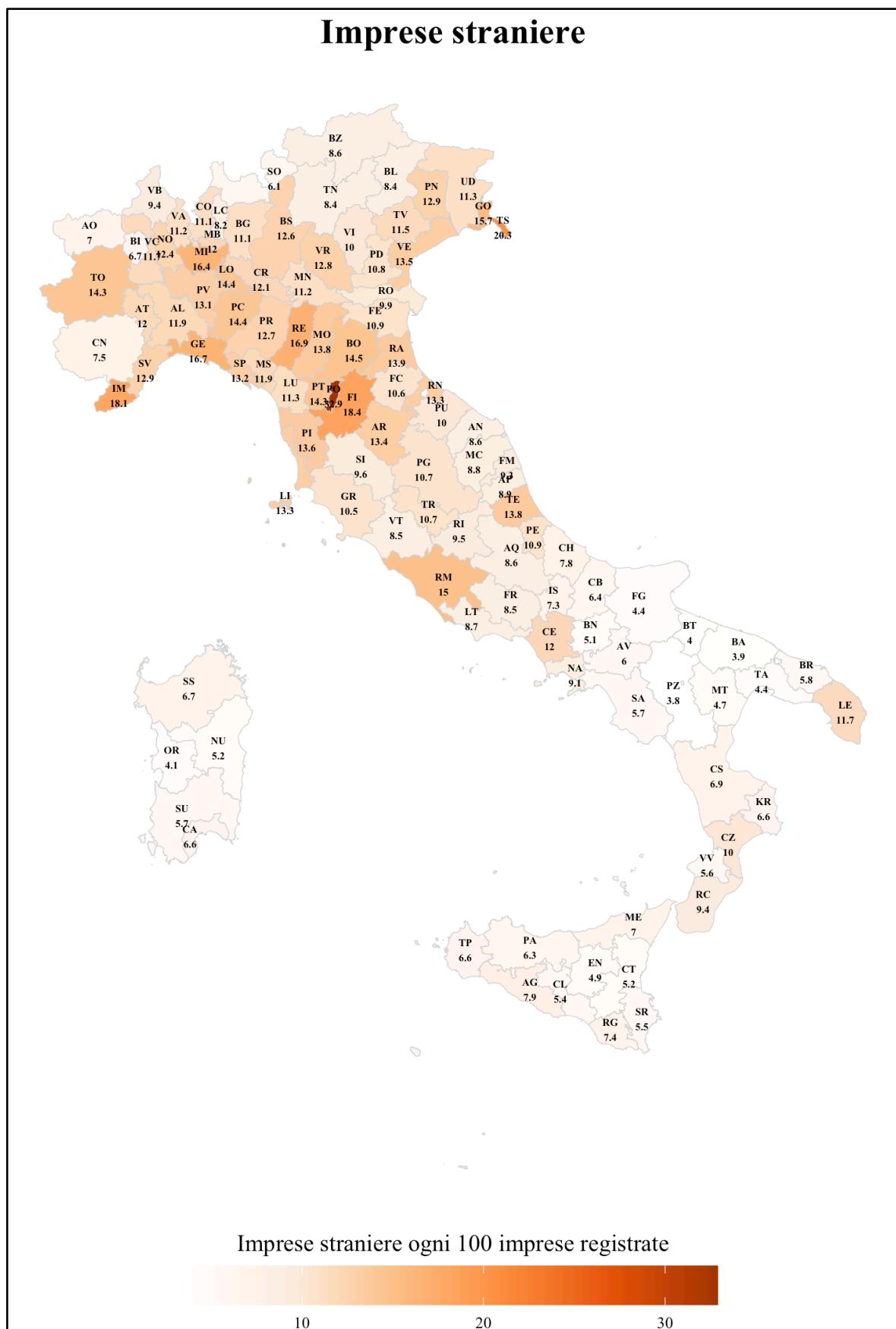
L'indicatore rappresenta il tasso di occupazione calcolato tra persone con età compresa tra 20 e 64 anni, espresso in percentuale. Il tasso di occupazione è il rapporto tra gli occupati e la popolazione di riferimento. In questo caso è il rapporto tra gli occupati con età compresa tra 20 e 64 anni e la popolazione con età compresa tra 20 e 64 anni. La fonte originale è l'Istat la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.8 - Principali statistiche descrittive della variabile Tasso di occupazione

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
38.8	57.5	70.3	65.1	73.2	79.1	0.16

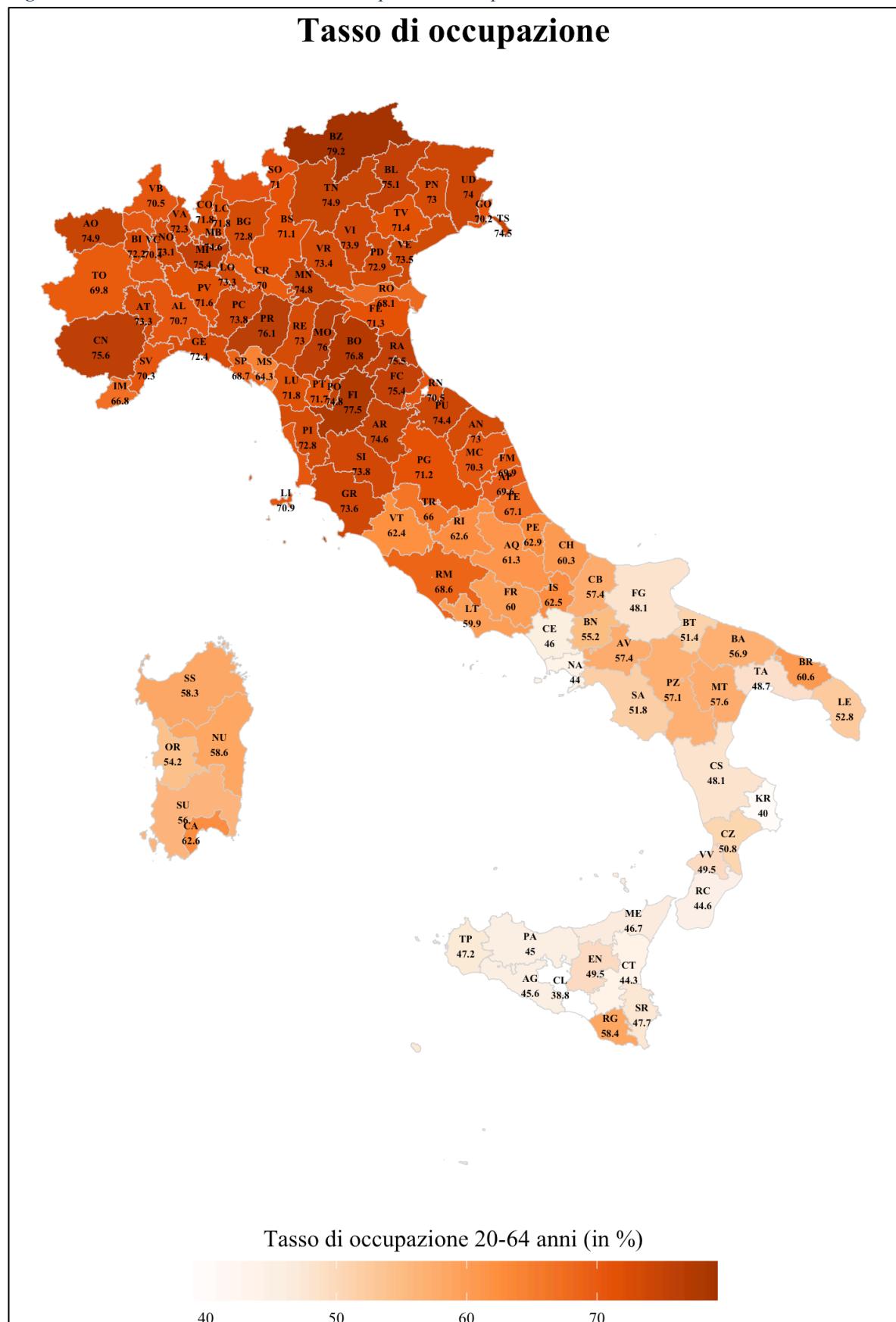
Il valore minimo registrato è 38.8% nella provincia di Caltanissetta, mentre il valore massimo raggiunge il 79.1% nella provincia di Bolzano. La media delle province è del 65.1%, con una mediana pari al 70.3%. Il 25% delle province presenta un valore inferiore a 57.5%, mentre il 75% si colloca al di sotto del 73.2%. Il coefficiente di variazione è pari a 0.16. La mappa coroletica in Figura 1.1.8 mostra chiaramente una concentrazione di valori più alti di tasso di occupazione nel Centro-Nord.

Figura 1.1.7 – Distribuzione delle imprese straniere ogni 100 imprese registrate nelle province italiane



Fonte dei dati: Infocamere, 30 settembre 2023

Figura 1.1.8 – Distribuzione del tasso di occupazione nelle province italiane



Fonte dei dati: Istat, 2022

- **Giovani che non lavorano e non studiano (Neet)**

L'indicatore rappresenta il tasso di neet, ovvero la popolazione di età compresa tra i 15 e i 29 anni che non è né occupata né inserita in un percorso di istruzione o di formazione, espresso in percentuale. La fonte originale è l'Istat la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.9 - Principali statistiche descrittive della variabile Giovani che non studiano e non lavorano (Neet)

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
7.6	13.1	15.9	18.5	22.5	41.6	0.39

Il valore minimo registrato è 7.6% nella provincia di Forlì-Cesena, mentre il valore massimo raggiunge il 41.6% nella provincia di Caltanissetta. La media delle province è del 18.5%, con una mediana pari al 15.9%. Il 25% delle province presenta un valore inferiore al 13.1%, mentre il 75% si colloca al di sotto del 22.5%. Il coefficiente di variazione è pari a 0.39. La distribuzione rappresentata nella mappa coropletica in Figura 1.1.9 evidenzia un tasso di neet decisamente superiore nelle province del Sud.

- **Gender pay gap**

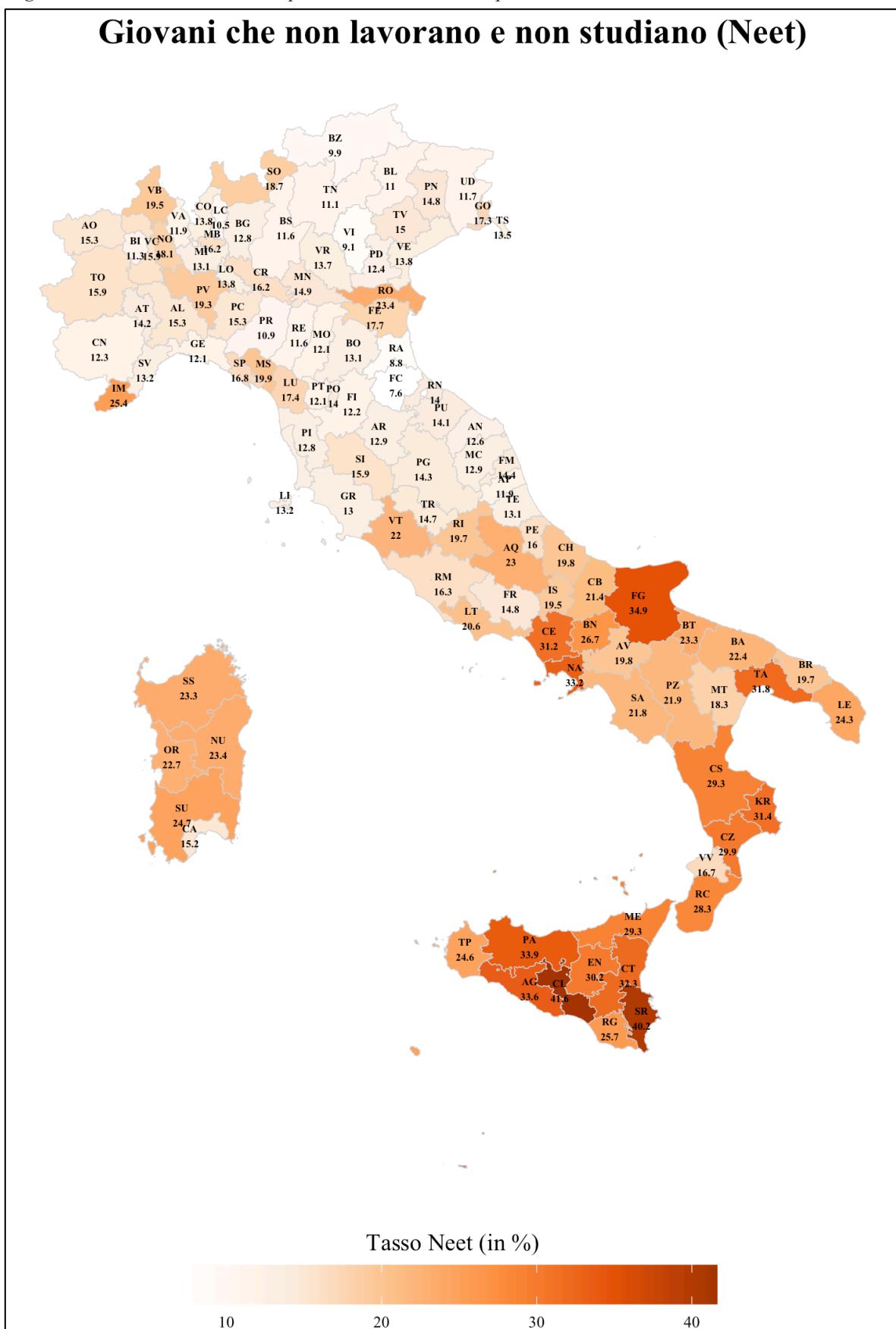
L'indicatore rappresenta il gender pay gap, ovvero la differenza percentuale di retribuzione media annua femminile rispetto a quella maschile, calcolata facendo riferimento solo al settore privato. La fonte originale è l'Istat la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.10 - Principali statistiche descrittive della variabile Gender pay gap

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
20.9	29.0	32.1	31.6	34.2	42.3	0.12

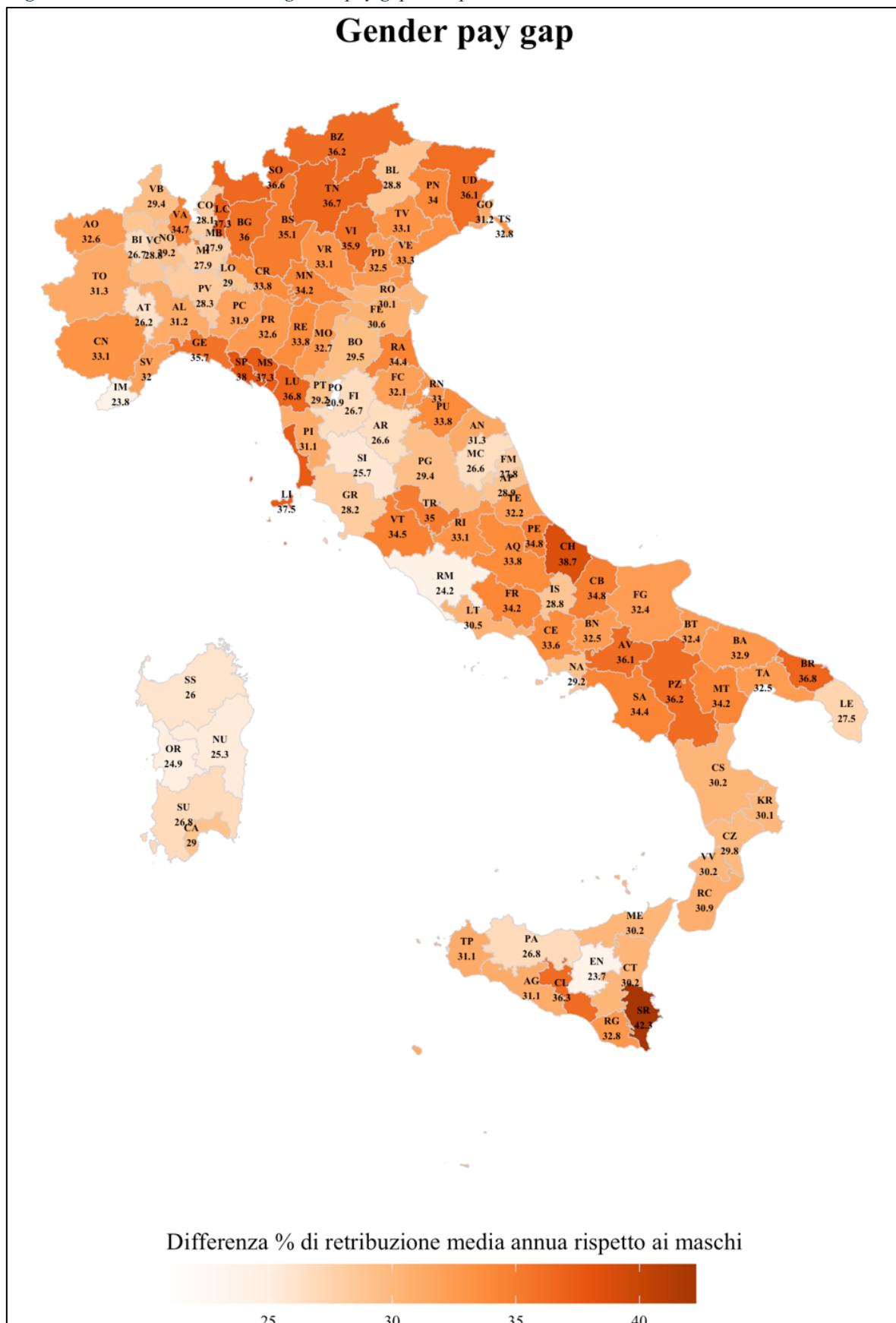
Il valore minimo registrato è 20.9% nella provincia di Prato, mentre il valore massimo raggiunge il 42.3% nella provincia di Siracusa. La media delle province è di 31.6, con una mediana pari a 32.1. Il 25% delle province presenta un valore inferiore al 29%, mentre il 75% si colloca al di sotto del 34.2%. Il coefficiente di variazione è pari a 0.12. La mappa coropletica in Figura 1.1.10 evidenzia una distribuzione del gender pay gap abbastanza uniforme in tutte le province, ad eccezione delle province della Sardegna e di parte della Toscana che presentano valori leggermente inferiori.

Figura 1.1.9 – Distribuzione della percentuale di Neet nelle province italiane



Fonte dei dati: Istat, 2022

Figura 1.1.10 – Distribuzione del gender pay gap nelle province italiane



Fonte dei dati: Inps, 2022

- **Lavoratori domestici**

L'indicatore rappresenta il numero di lavoratori domestici ogni mille abitanti. Le fonti sono Inps e Istat e la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.11 - Principali statistiche descrittive della variabile Lavoratori domestici

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
3.77	9.79	13.6	14.4	17.3	34.5	0.49

Il valore minimo registrato è di 3.77 nella provincia di Siracusa, mentre il valore massimo raggiunge 34.5 nella provincia di Oristano. La media delle province è di 14.4, con una mediana leggermente inferiore pari a 13.6. Il 25% delle province presenta un valore inferiore a 9.79, mentre il 75% si colloca al di sotto di 17.3. Il coefficiente di variazione è pari a 0.49. Osservando la mappa coropletica in Figura 1.1.11 si notano valori più elevati nel Centro-Nord, ma con valori molto elevati anche nelle province della Sardegna.

- **Quota di export sul Pil**

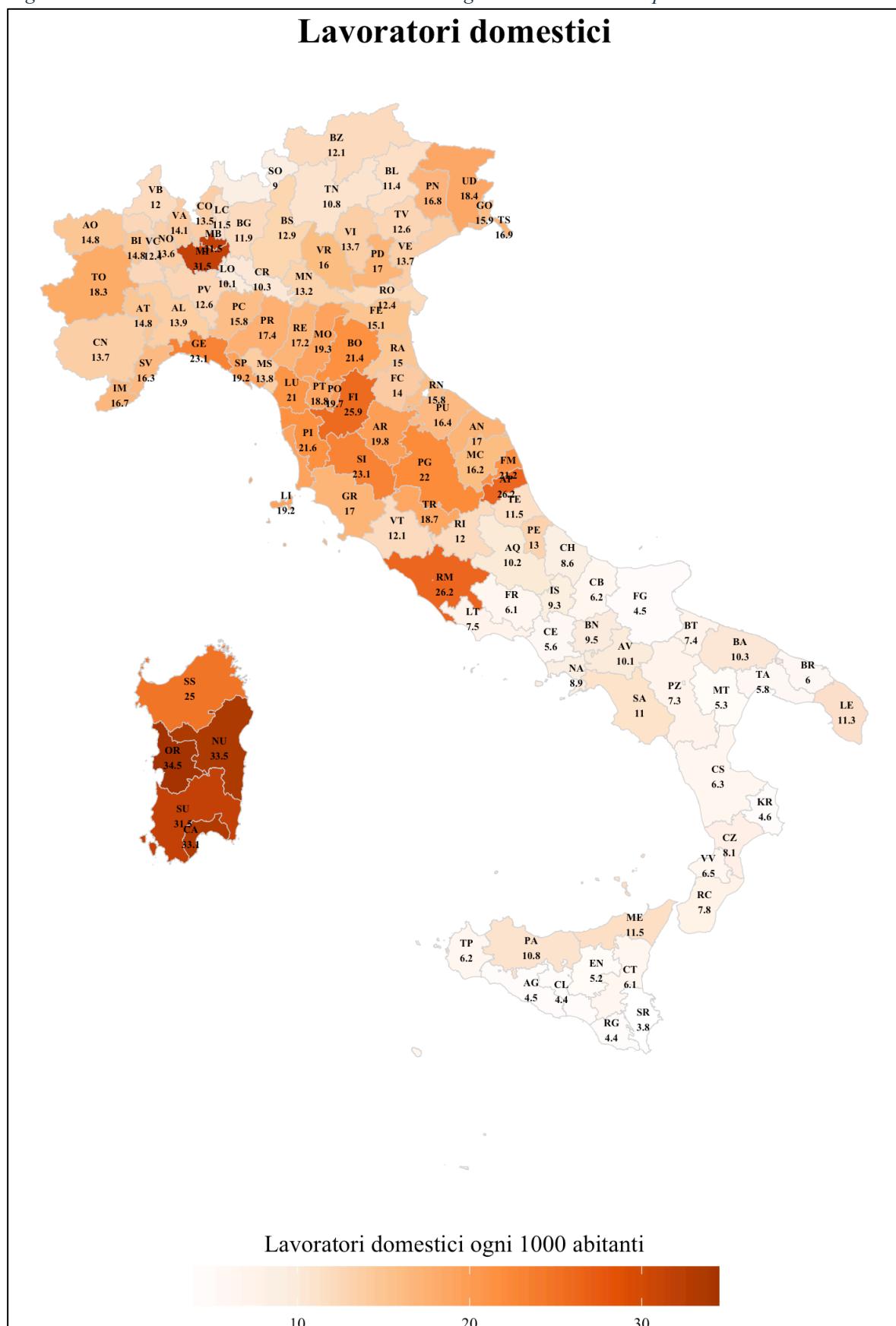
L'indicatore rappresenta il rapporto espresso in percentuale tra esportazioni di beni verso l'estero e valore aggiunto. La fonte originale è Prometeia e la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.12 - Principali statistiche descrittive della variabile Quota di export sul Pil

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
0.73	11.9	26.8	33.1	48.2	174	0.86

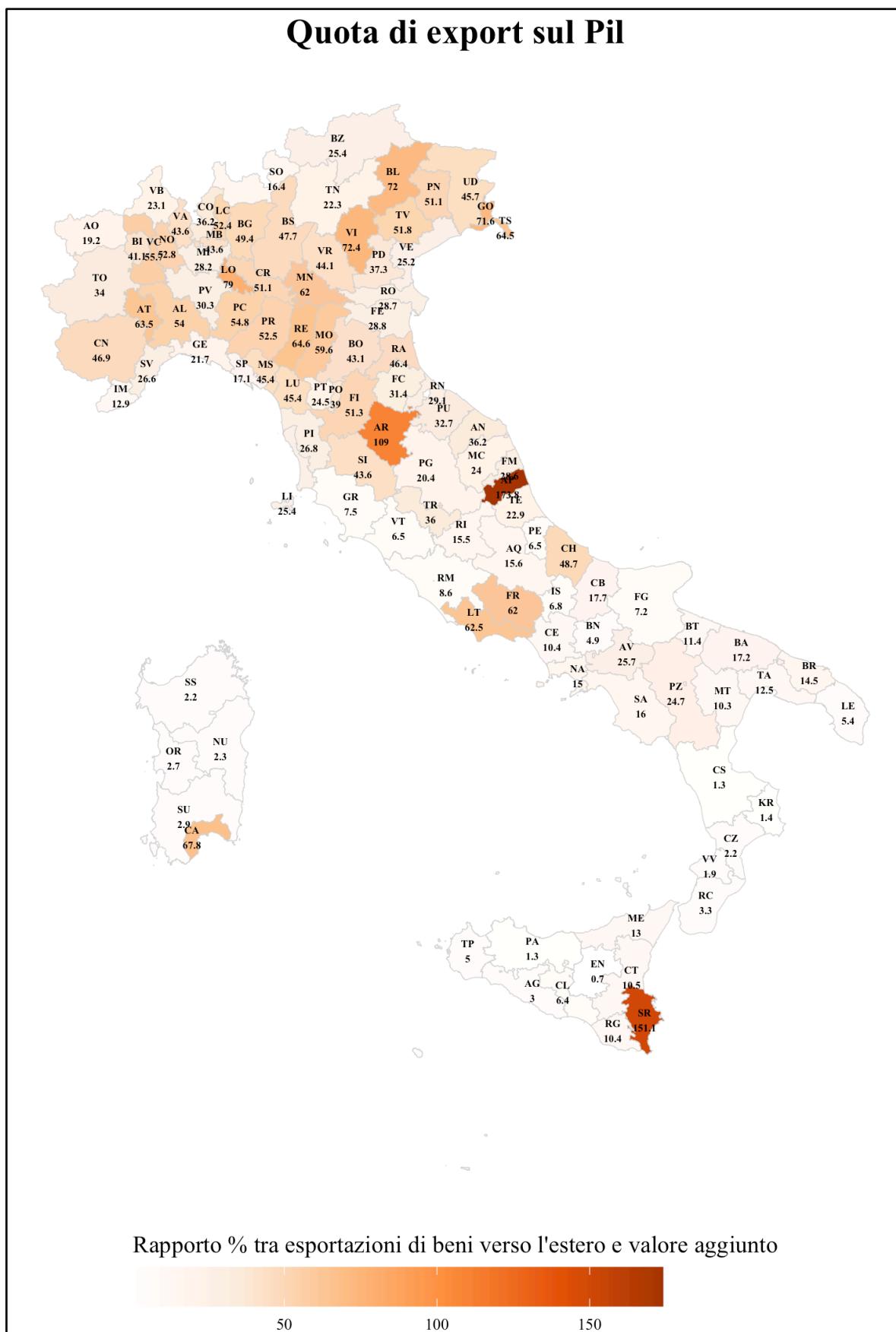
Il valore minimo registrato è 0.73% nella provincia di Enna, mentre il valore massimo raggiunge il 174% nella provincia di Ascoli Piceno. La media delle province è del 33.1%, con una mediana pari al 26.8%. Il 25% delle province presenta un valore inferiore al 11.9%, mentre il 75% si colloca al di sotto del 48.2%. Il coefficiente di variazione è pari a 0.86. Osservando la mappa coropletica in Figura 1.1.12, si nota una quota di export sul PIL più elevata nelle province del Nord; tuttavia, si riscontrano valori significativamente più alti rispetto al resto del territorio nelle province di Ascoli Piceno, Siracusa e Arezzo.

Figura 1.1.11 – Distribuzione dei lavoratori domestici ogni 1000 abitanti nelle province italiane



Fonte dei dati: Inps/Istat, 2022

Figura 1.1.12 – Distribuzione della quota di export sul Pil in percentuale nelle province italiane



Fonte dei dati: Prometeia, 2022

- **Partecipazione alla formazione continua**

L'indicatore rappresenta il tasso di partecipazione alla formazione continua di persone con età compresa fra 25 e 64, espresso in percentuale. La fonte originale è l'Istat e la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.13 - Principali statistiche descrittive della variabile Partecipazione alla formazione continua

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
3.5	7.35	8.8	9.36	10.9	22	0.33

Il valore minimo registrato è 3.5% nella provincia di Imperia, mentre il valore massimo raggiunge il 22% nella provincia di Cagliari. La media delle province è del 9.36%, con una mediana leggermente inferiore pari al 8.8%. Il 25% delle province presenta un valore inferiore al 7.35%, mentre il 75% si colloca al di sotto del 10.9%. Il coefficiente di variazione è pari a 0.33. Analizzando la mappa coroletica in Figura 1.1.13, si nota una leggera prevalenza di province del Centro-Nord impegnate nella formazione continua, anche se i valori più elevati si registrano nelle province di Cagliari e Vibo Valentia.

- **Infortuni sul lavoro**

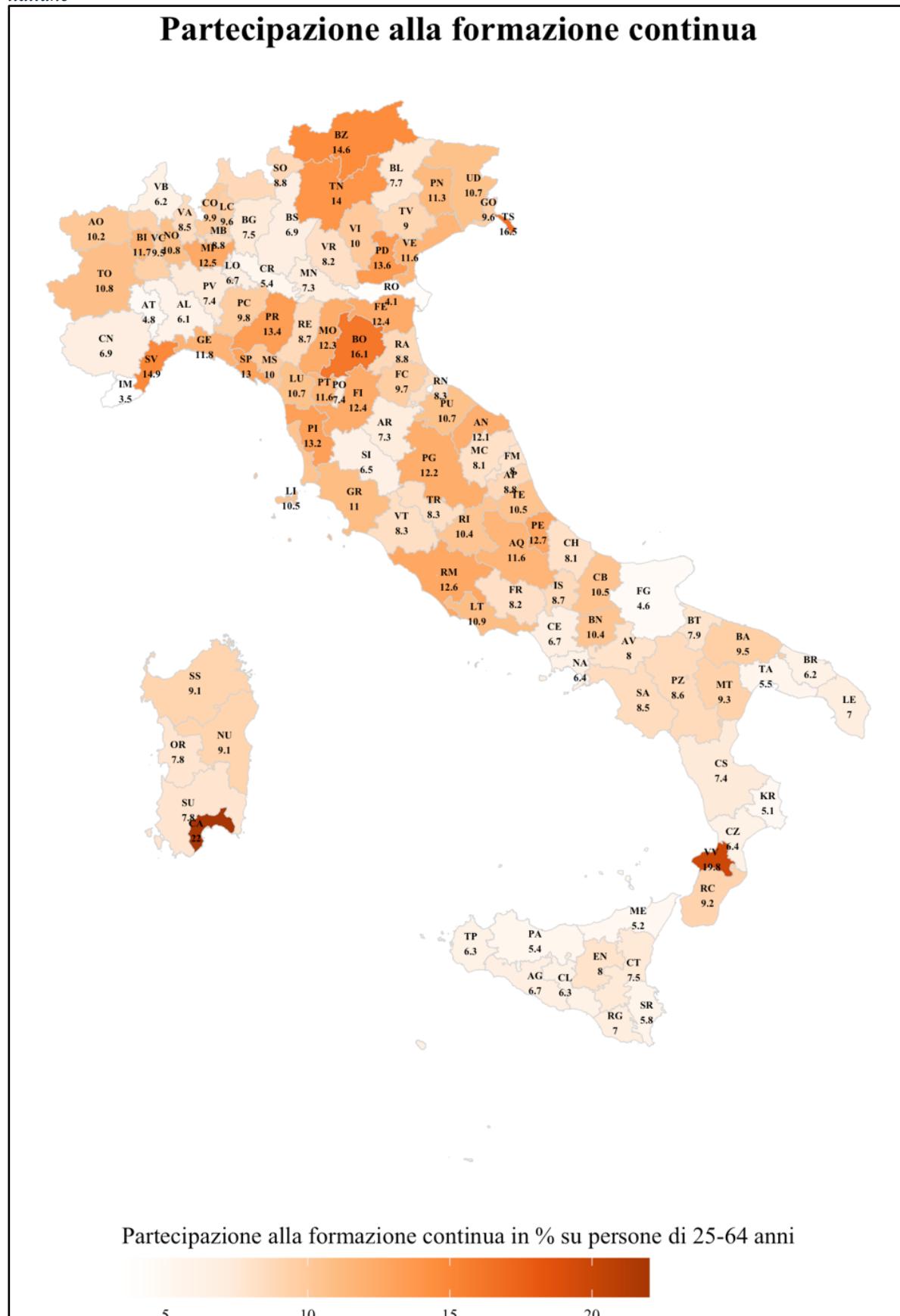
L'indicatore rappresenta il numero di infortuni mortali e inabilità permanente ogni 10000 occupati. La fonte originale è l'Inail e la rilevazione dell'indicatore risale al 2021

Tabella 1.1.14 - Principali statistiche descrittive della variabile Infortuni sul lavoro

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
4.3	9.15	11.2	11.4	14	23.4	0.31

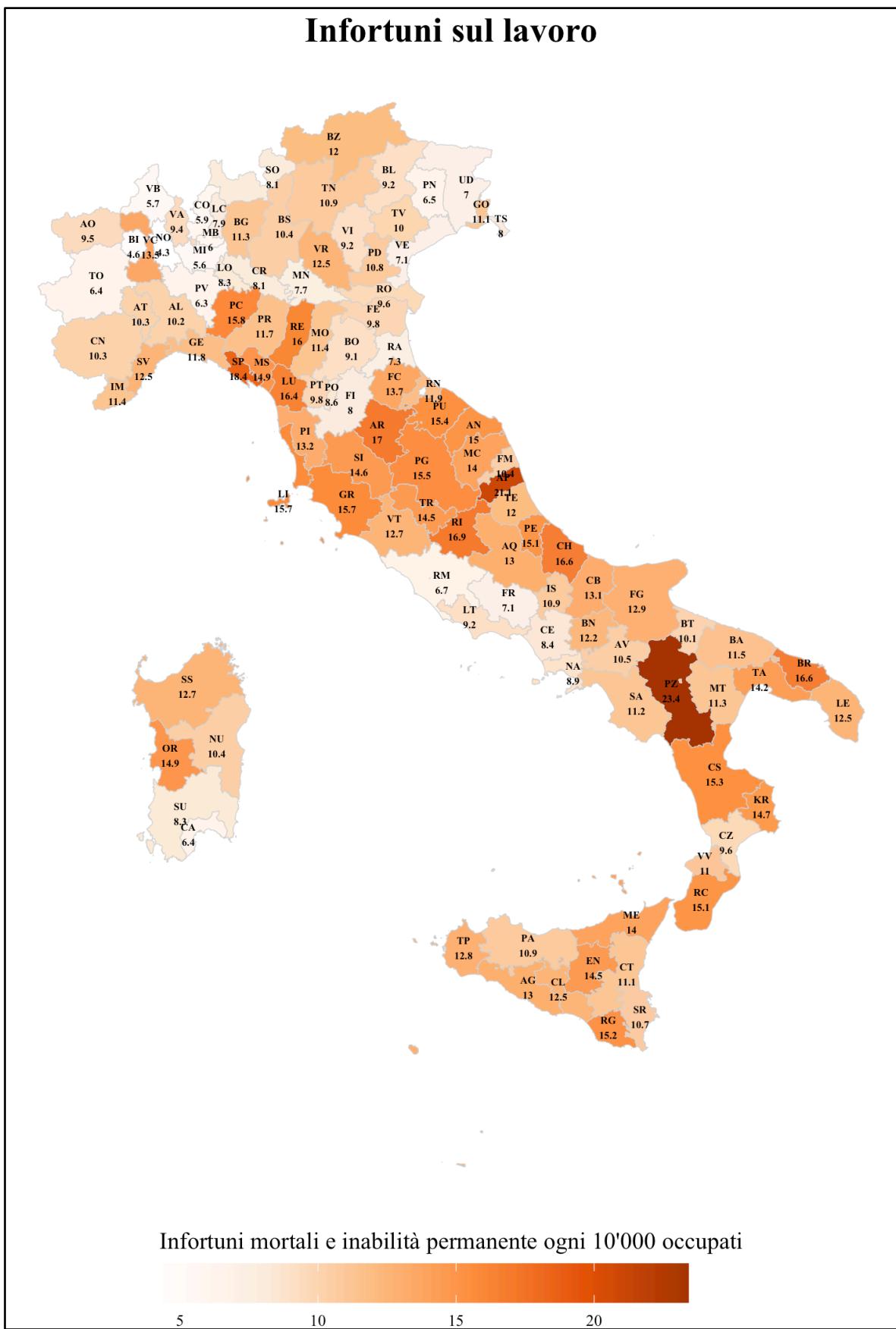
Il valore minimo registrato è di 4.3 nella provincia di Novara, mentre il valore massimo raggiunge 23.4 nella provincia di Potenza. La media delle province è di 11.4, con una mediana leggermente inferiore pari a 11.2. Il 25% delle province presenta un valore inferiore a 9.15, mentre il 75% si colloca al di sotto di 14. Il coefficiente di variazione è pari a 0.31. Osservando la mappa coroletica in Figura 1.1.14, gli infortuni sul lavoro risultano più frequenti nelle province del Centro-Sud.

Figura 1.1.13 – Distribuzione della partecipazione alla formazione continua in percentuale nelle province italiane



Fonte dei dati: Istat, 2022

Figura 1.1.14 – Distribuzione degli infortuni sul lavoro ogni 10'000 occupati nelle province italiane



Fonte dei dati: Inail, 2021

- **Numero pensioni di vecchiaia**

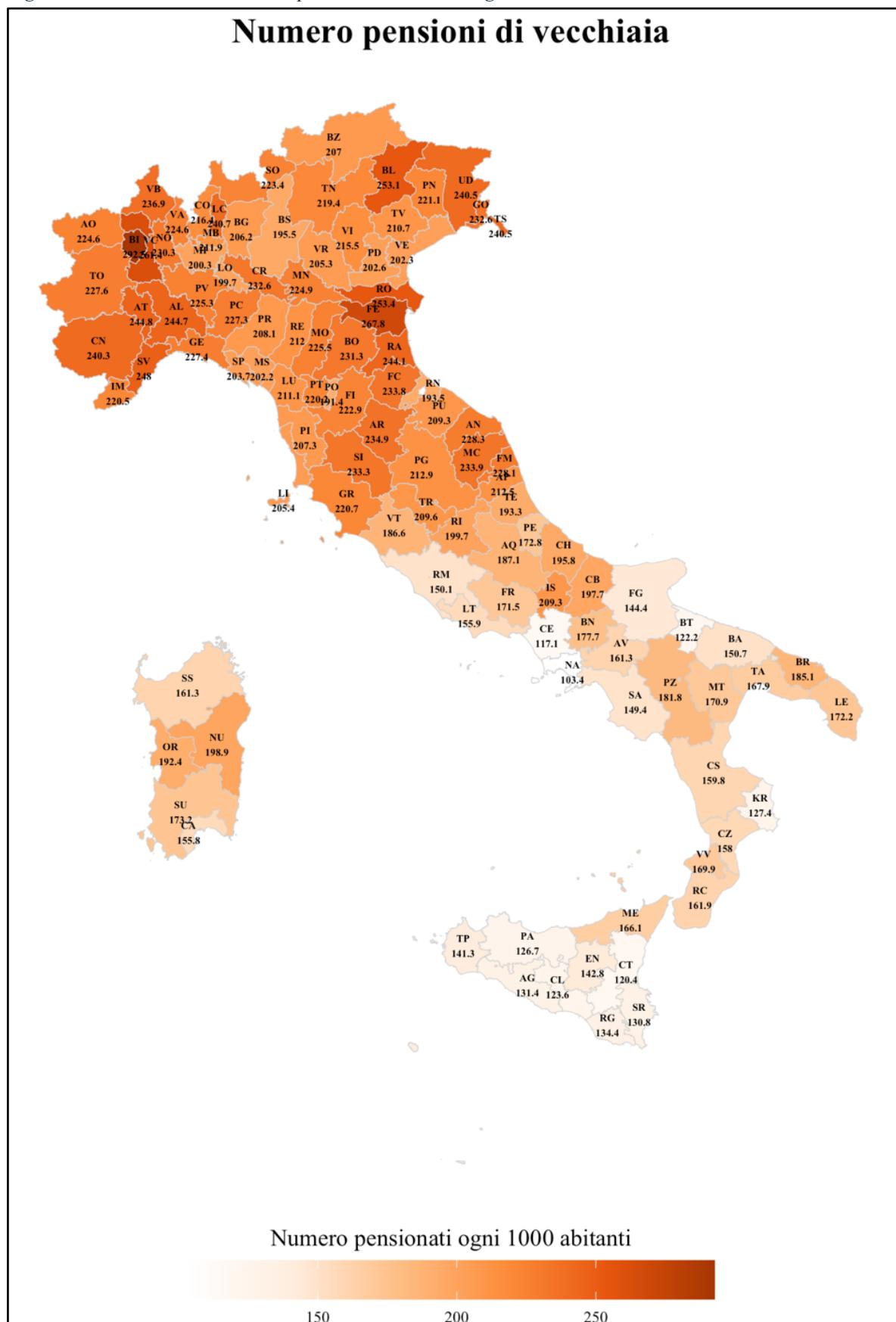
L'indicatore rappresenta il numero di pensionati ogni 1000 abitanti. La fonte originale è l'Inps e la rilevazione dell'indicatore risale al 2022.

Tabella 1.1.15 - Principali statistiche descrittive della variabile Numero pensioni di vecchiaia

Minimo	Primo Quartile	Mediana	Media	Terzo Quartile	Massimo	CV
103	171	206	198	226	292	0.19

Il valore minimo registrato è di 103 nella provincia di Napoli, mentre il valore massimo raggiunge 292 nella provincia di Biella. La media delle province è di 198, con una mediana pari a 206. Il 25% delle province presenta un valore inferiore a 171, mentre il 75% si colloca al di sotto di 226. Il coefficiente di variazione è pari a 0.19. Osservando la mappa coropletica in Figura 1.1.15 emerge chiaramente che le province del Centro-Nord presentano un numero significativamente più elevato di pensioni di vecchiaia.

Figura 1.1.15 – Distribuzione delle pensioni di vecchiaia ogni 1000 abitanti



Fonte dei dati: Inps/Istat, 2022

Capitolo 2

2 Analisi delle componenti principali

L'analisi delle componenti principali è una tecnica di analisi multivariata di riduzione della dimensionalità. L'obiettivo è la creazione di nuove variabili artificiali, chiamate componenti principali, che siano incorrelate tra loro e siano in ordine decrescente rispetto alla loro varianza. Le componenti principali sono combinazioni lineari delle variabili quantitative originali, tra loro correlate. È necessario mantenere la maggiore quantità possibile di informazione delle variabili originali e la misura che viene utilizzata per misurare la quantità di informazione conservata è la varianza³.

2.1 Matrice di correlazione e componenti principali

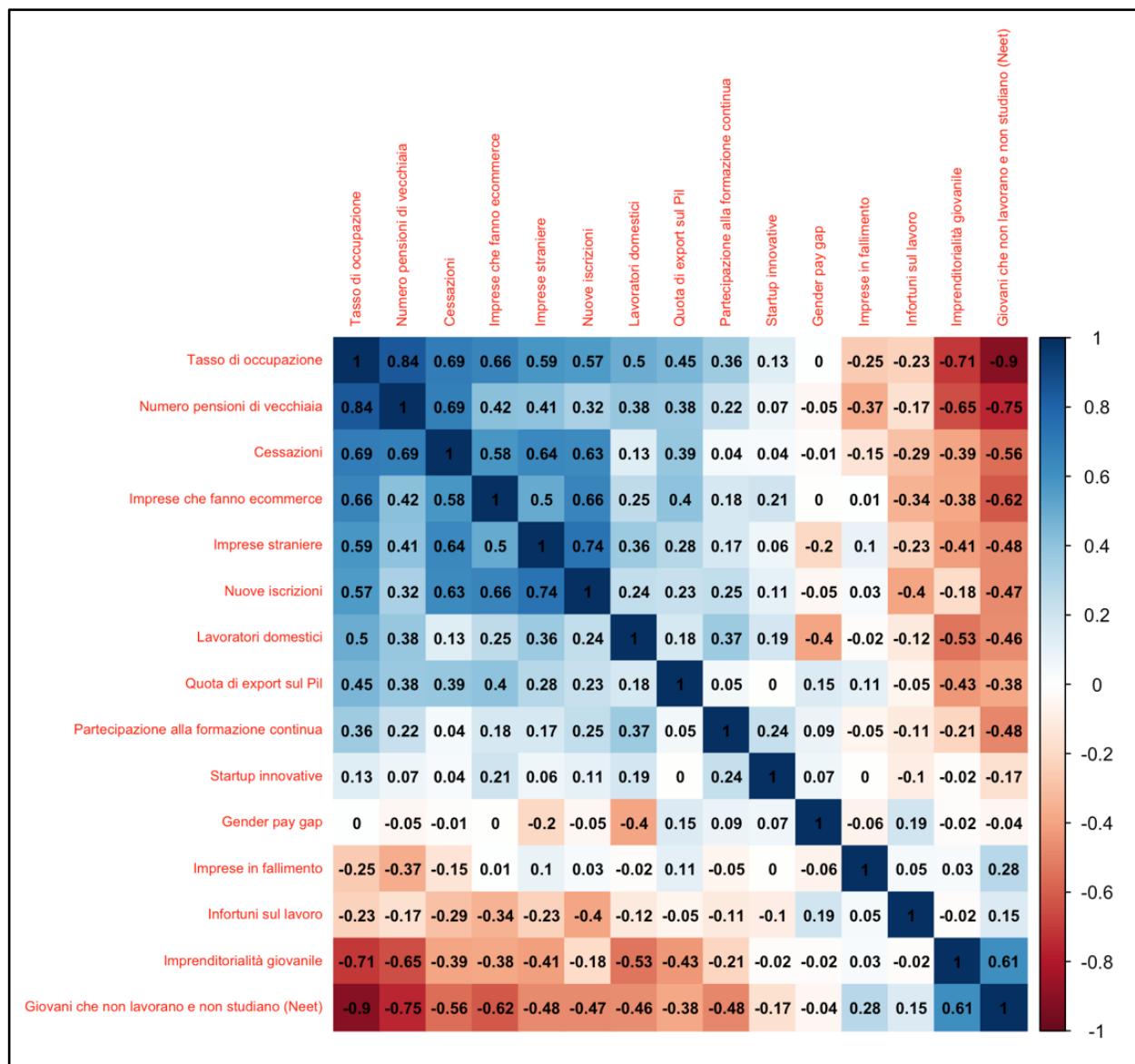
Quando siamo in presenza di variabili quantitative con unità di misura differenti o con ordini di grandezza molto diversi è opportuno procedere alla standardizzazione dei dati prima di effettuare l'analisi delle componenti principali. Si ottiene quindi la matrice dei dati standardizzati $Z = \tilde{X}diag(\frac{1}{s_1}, \dots, \frac{1}{s_p})$, dove \tilde{X} è la matrice degli scarti e s_1, \dots, s_p rappresentano le deviazioni standard di ciascuna variabile. Si utilizza la matrice Z perché le osservazioni relative alle variabili standardizzate, che sono contenute nelle colonne di Z , hanno media zero e varianza unitaria. Questo serve ad evitare che le variabili con varianza maggiore prevalgano sulle componenti principali. Ciò equivale ad assumere come punto di partenza la matrice di correlazione: i punteggi della v -esima componente principale si definiscono allora come $y_v = Z\gamma_v^*$, dove γ_v^* è l'autovettore associato al v -esimo autovalore λ_v^* della matrice di correlazione di Z . λ_v^* rappresenta la varianza della v -esima componente principale⁴.

I punteggi della prima componente principale, ovvero $y_1 = Z\gamma_1^*$, sono stati utilizzati per creare un indice di qualità generale basato sui 15 indicatori della categoria “Affari e lavoro” delle province italiane, accompagnato dall’interpretazione semantica effettuata osservando i pesi dell’autovettore γ_1^* associato al primo autovalore λ_1^* della matrice di correlazione. I punteggi delle componenti principali e gli autovettori sono stati calcolati utilizzando la funzione *prcomp* di R.

³ Jolliffe I.T., (2002), *Principal component analysis*, Springer, pp. 1-2.

⁴ Jolliffe I.T., (2002), *Principal component analysis*, Springer, pp. 21-26.

Figura 2.1.1 - Matrice di correlazione



È importante che le variabili abbiano una buona correlazione tra loro, in questo modo saranno sufficienti poche componenti principali per rappresentare in modo adeguato le variabili originali. Dalla matrice di correlazione in Figura 2.1.1 si può notare, ad esempio, una correlazione molto forte tra il tasso di occupazione e il numero di pensioni di vecchiaia, pari a 0.84, e una correlazione negativa tra il tasso di occupazione e il tasso di Neet pari a -0.9. Si nota anche una forte correlazione negativa fra l'imprenditorialità giovanile e il tasso di occupazione, pari a -0.71. È anche possibile notare una forte correlazione positiva tra tasso di occupazione, cessazioni, imprese che fanno e-commerce, imprese straniere e nuove iscrizioni.

2.2 Biplot

Il biplot è una rappresentazione grafica nella quale la prima e la seconda componente principale sono utilizzate come assi cartesiani, in cui vengono rappresentate contemporaneamente le variabili e le unità statistiche. Dato che si utilizzano congiuntamente due componenti principali è necessario standardizzarle per fare in modo che la differenza nella varianza non influenzi il grafico: la generica unità statistica i sarà allora rappresentata tramite i punteggi standardizzati della prima e della seconda componente principale $\left(\frac{y_{i1}}{\sqrt{\lambda_1^*}}, \frac{y_{i2}}{\sqrt{\lambda_2^*}}\right)$, mentre la generica variabile j è rappresentata dal vettore $(\sqrt{\lambda_1^*}\gamma_{j1}^*, \sqrt{\lambda_2^*}\gamma_{j2}^*)^T$. Il biplot è stato visualizzato tramite la funzione `fviz_pca_biplot` del pacchetto `factoextra`.

Per quanto riguarda le variabili rappresentate nel biplot si osservi che:

- La correlazione tra variabili è tanto più forte quanto più è piccolo l'angolo tra i vettori, dato che il coefficiente di correlazione lineare tra due variabili può essere espresso come il coseno dell'angolo compreso tra i vettori ad esse associati.
- L'angolo compreso tra il vettore che rappresenta una variabile e gli assi cartesiani, che rappresentano le componenti principali, esprime il coefficiente di correlazione. Quindi un angolo piccolo esprime una forte correlazione tra la variabile e la componente principale.
- Il modulo di un vettore relativo ad una variabile, rappresentato dalla lunghezza del vettore, esprime la proporzione di varianza della variabile spiegata dalle due componenti principali⁵.

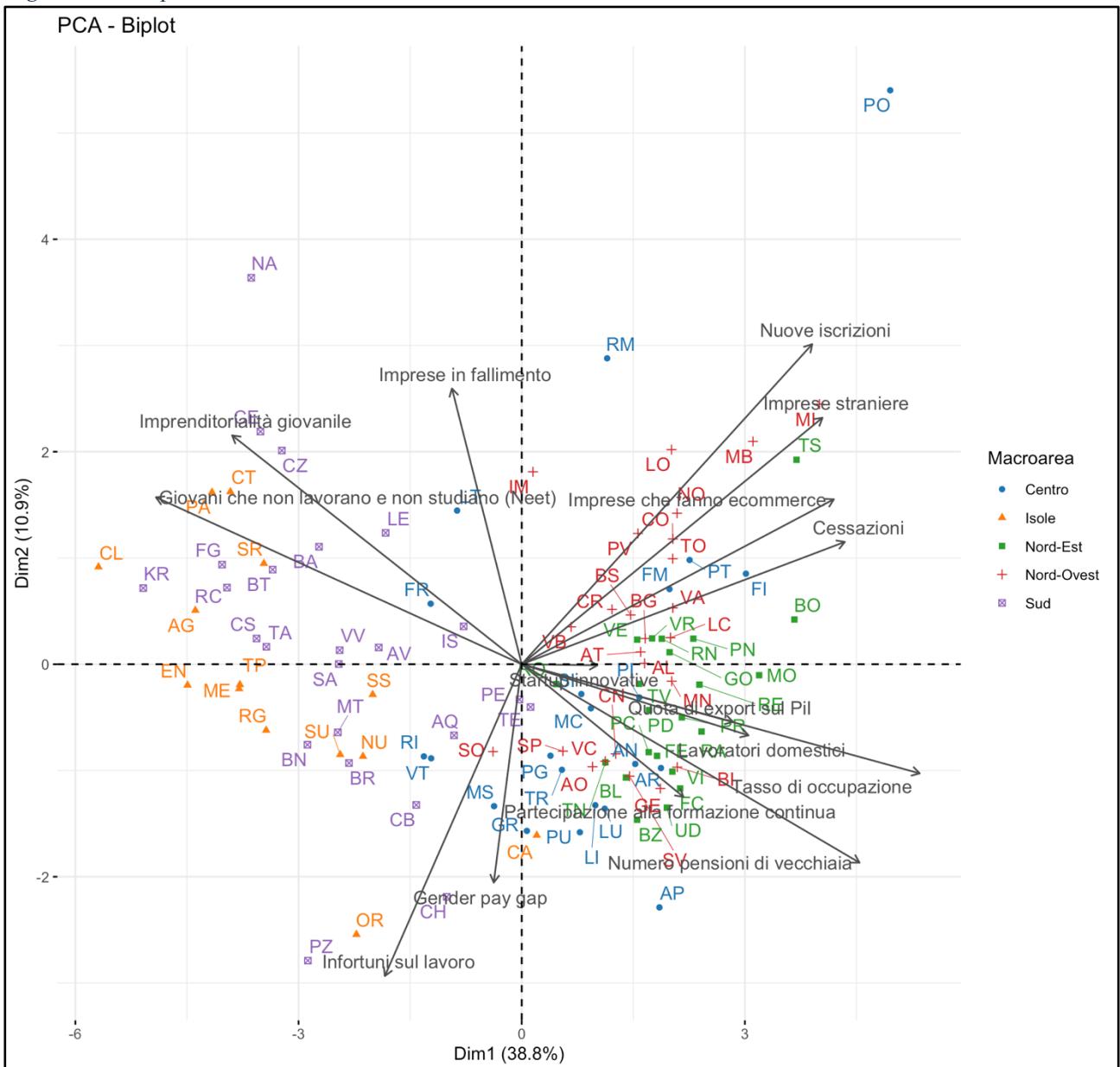
Per quanto riguarda le unità rappresentate nel biplot si osservi che:

- Punti vicini all'origine degli assi segnalano che le corrispondenti unità presentano valori delle variabili prossimi alle rispettive medie.
- Punti lontani dall'origine degli assi e nella direzione di uno degli assi indicano che le relative unità sono caratterizzate da un punteggio di quella componente particolarmente elevato in modulo.
- Un punto lontano dall'origine degli assi nella direzione del vettore corrispondente a una variabile segnala che l'unità mostra un valore di questa variabile notevolmente maggiore della media⁶.

⁵ Jolliffe I.T., (2002), *Principal component analysis*, Springer, pp. 96-99.

⁶ Jolliffe I.T., (2002), *Principal component analysis*, Springer, pp. 100-101.

Figura 2.2.1 - Biplot



Dall'analisi svolta le prime due componenti principali preservano rispettivamente il 38.8% e il 10.9% della varianza originale, per un totale del 49.7%.

Dal biplot in Figura 2.2.1 si nota una chiara divisione delle province del Sud e Isole da quelle del Centro-Nord: le prime, infatti, ad eccezione della provincia di Cagliari, assumono valori negativi nella prima componente principale.

Osservando i vettori del biplot è possibile osservare una correlazione positiva molto forte tra la quota di export sul PIL, lavoratori domestici e tasso di occupazione, oltre che una alta correlazione positiva tra la partecipazione alla formazione continua e il numero di pensioni di vecchiaia. Queste sono le variabili che caratterizzano particolarmente le province del Nord-Est. Le province del Nord-Ovest sembrano invece essere caratterizzate principalmente dalle variabili legate all'impresa, quindi

nuove iscrizioni, imprese straniere, imprese che fanno e-commerce e cessazioni, variabili la cui correlazione positiva viene confermata dal biplot. Si ha anche conferma della correlazione positiva fra Imprenditorialità giovanile e il tasso di Neet, che a loro volta hanno una forte correlazione negativa con il tasso di occupazione. Imprenditorialità giovanile e il tasso di Neet sono le variabili che caratterizzano maggiormente le province del Sud. Si nota anche una correlazione positiva tra gender pay gap e infortuni sul lavoro.

Le variabili maggiormente correlate con la prima componente principale sono il tasso di occupazione e il numero di startup innovative, anche se quest'ultima non viene spiegata particolarmente bene dalle due componenti principali. La seconda componente principale ha invece una forte correlazione positiva con il numero di imprese in fallimento e una forte correlazione negativa con il gender pay gap.

2.3 Indice sintetico delle province basato sulla prima componente principale

La creazione di uno un indice sintetico basato sui 15 indicatori della categoria Affari e lavoro delle province italiane è stata eseguita utilizzando i punteggi della prima componente principale $y_1 = Z\gamma_1^*$, che preserva il 38.8% della varianza totale. Osservando i pesi dell'autovettore γ_1^* associato alla prima componente principale nella Tabella 2.3.1 è possibile dare un'interpretazione della componente principale.

Le variabili che contribuiscono maggiormente in termini positivi sono il tasso di occupazione, il numero di pensioni di vecchiaia e le cessazioni. Quelle che contribuiscono negativamente sono invece il tasso di Neet, l'imprenditorialità giovanile e gli infortuni sul lavoro.

Il contributo negativo dato dall'imprenditorialità giovanile potrebbe essere dovuto al fatto che rappresenta un sintomo di un sistema economico fragile, dove avviare un'impresa rappresenta una risposta alla mancanza di opportunità lavorative. Osservando la matrice di correlazione in Figura 2.1.1 si vede infatti che c'è una forte correlazione negativa tra l'imprenditoria giovanile e il tasso di occupazione, pari a -0.71, oltre che una forte correlazione positiva pari a 0.61 con il tasso di Neet. Questa situazione viene confermata anche nel biplot in Figura 2.2.1. Dal report Istat 2023, inoltre, si può osservare che l'incidenza dell'imprenditoria giovanile si riduce al crescere della dimensione di impresa: nel 98.1% dei casi le imprese under 35 sono microimprese, rispetto al 95.1% di quelle gestite

da over 35. Inoltre, il 75% delle imprese gestite da imprenditori under 35 sono ditte individuali, rispetto al 63% del totale delle imprese⁷.

Tabella 2.3.1 - Pesi dell'autovettore associato alla prima componente principale

Pesi	
Startup innovative	0.075
Imprese che fanno ecommerce	0.309
Imprenditorialità giovanile	-0.287
Nuove iscrizioni	0.288
Cessazioni	0.320
Imprese in fallimento	-0.069
Imprese straniere	0.298
Tasso di occupazione	0.394
Giovani che non lavorano e non studiano (Neet)	-0.362
Gender pay gap	-0.028
Lavoratori domestici	0.225
Quota di export sul Pil	0.209
Partecipazione alla formazione continua	0.160
Infortuni sul lavoro	-0.135
Numero pensioni di vecchiaia	0.335

Dai punteggi della prima componente principale, è stata quindi ottenuta la Tabella 2.3.2, dove sono riportati i punteggi di ciascuna provincia. I valori più alti sono stati ottenuti dalle province di Prato, Milano e Trieste. Le province di Enna, Crotone e Caltanissetta, invece, si trovano in fondo alla classifica.

Tabella 2.3.2 - Punteggi della prima componente principale delle province italiane

Ranking	Provincia	Punteggio
1	Prato	4.9537
2	Milano	4.0032
3	Trieste	3.6942
4	Bologna	3.6659
5	Modena	3.1921
6	Monza e della Brianza	3.1092
7	Firenze	3.0140
8	Ravenna	2.4172

⁷ Istituto Nazionale di Statistica (ISTAT), *Rapporto annuale 2023*, ISTAT, 2023, <https://www.istat.it/storage/rapporto-annuale/2023/Rapporto-Annuale-2023.pdf>.

9	Reggio nell'Emilia	2.3887
10	Pordenone	2.3075
11	Pistoia	2.2567
12	Parma	2.1526
13	Forlì-Cesena	2.1298
14	Biella	2.0907
15	Novara	2.0887
16	Como	2.0326
17	Varese	2.0323
18	Torino	2.0304
19	Vicenza	2.0252
20	Mantova	2.0166
21	Lodi	2.0138
22	Lecco	2.0004
23	Gorizia	1.9883
24	Fermo	1.9867
25	Udine	1.9516
26	Rimini	1.8830
27	Arezzo	1.8719
28	Genova	1.8635
29	Ascoli Piceno	1.8513
30	Ferrara	1.8196
31	Verona	1.7531
32	Padova	1.7122
33	Piacenza	1.7072
34	Bergamo	1.6606
35	Alessandria	1.6517
36	Asti	1.5988
37	Pisa	1.5818
38	Treviso	1.5793
39	Pavia	1.5627
40	Venezia	1.5527
41	Bolzano	1.5515
42	Ancona	1.5263
43	Brescia	1.4646
44	Savona	1.4492
45	Belluno	1.4010
46	Cuneo	1.2643
47	Cremona	1.2138
48	Roma	1.1490
49	Trento	1.1253
50	Vercelli	1.1231
51	Lucca	1.1158
52	Livorno	0.9897
53	Valle d'Aosta	0.9517
54	Macerata	0.9302
55	Siena	0.8019

56	Pesaro e Urbino	0.7821
57	Verbano-Cusio-Ossola	0.6647
58	La Spezia	0.5536
59	Terni	0.5409
60	Rovigo	0.4661
61	Perugia	0.3866
62	Cagliari	0.2026
63	Imperia	0.1510
64	Teramo	0.1188
65	Grosseto	0.0695
66	Pescara	-0.0303
67	Massa-Carrara	-0.3719
68	Sondrio	-0.3858
69	Isernia	-0.7805
70	Latina	-0.8685
71	L'Aquila	-0.9094
72	Chieti	-1.0068
73	Viterbo	-1.2176
74	Frosinone	-1.2240
75	Rieti	-1.3151
76	Campobasso	-1.4177
77	Lecce	-1.8292
78	Avellino	-1.9215
79	Sassari	-2.0030
80	Nuoro	-2.1335
81	Oristano	-2.2215
82	Brindisi	-2.3217
83	Sud Sardegna	-2.4402
84	Vibo Valentia	-2.4471
85	Salerno	-2.4592
86	Matera	-2.4722
87	Bari	-2.7246
88	Potenza	-2.8739
89	Benevento	-2.8797
90	Catanzaro	-3.2240
91	Barletta-Andria-Trani	-3.3479
92	Taranto	-3.4323
93	Ragusa	-3.4355
94	Siracusa	-3.4665
95	Caserta	-3.5122
96	Cosenza	-3.5639
97	Napoli	-3.6355
98	Trapani	-3.7885
99	Messina	-3.7984
100	Catania	-3.9158
101	Reggio Calabria	-3.9618
102	Foggia	-4.0281

103	Palermo	-4.1630
104	Agrigento	-4.3873
105	Enna	-4.4921
106	Crotone	-5.0878
107	Caltanissetta	-5.6896

Capitolo 3

3 Clustering

Il clustering è una tecnica di analisi multivariata che permette di creare dei gruppi di unità statistiche che siano omogenei al loro interno ed eterogenei tra loro.

È stato perciò effettuata un'analisi cluster delle province italiane al fine di identificare gruppi di province che presentano caratteristiche simili tra di loro in base ai 15 indicatori della categoria “Affari e lavoro”. L'analisi è stata eseguita attraverso il clustering gerarchico agglomerativo, usando il metodo di Ward come linkage e la distanza euclidea come misura della distanza, dato che siamo in presenza di sole variabili quantitative. Successivamente è stata calcolata la media aritmetica dei valori assunti dalle variabili nelle province che si trovano nello stesso cluster, con l'obiettivo di fornire una descrizione delle caratteristiche di ciascun gruppo.

Il dendrogramma utilizzato per l'analisi dei cluster è stato ottenuto attraverso la funzione *hclust()* di R, ed è stato visualizzato attraverso la funzione *fviz_dend()* del pacchetto *factoextra*. Per calcolare la matrice delle distanze è stata utilizzata la distanza euclidea, calcolata con l'uso della funzione *dist()* di R.

3.1 Matrice delle distanze

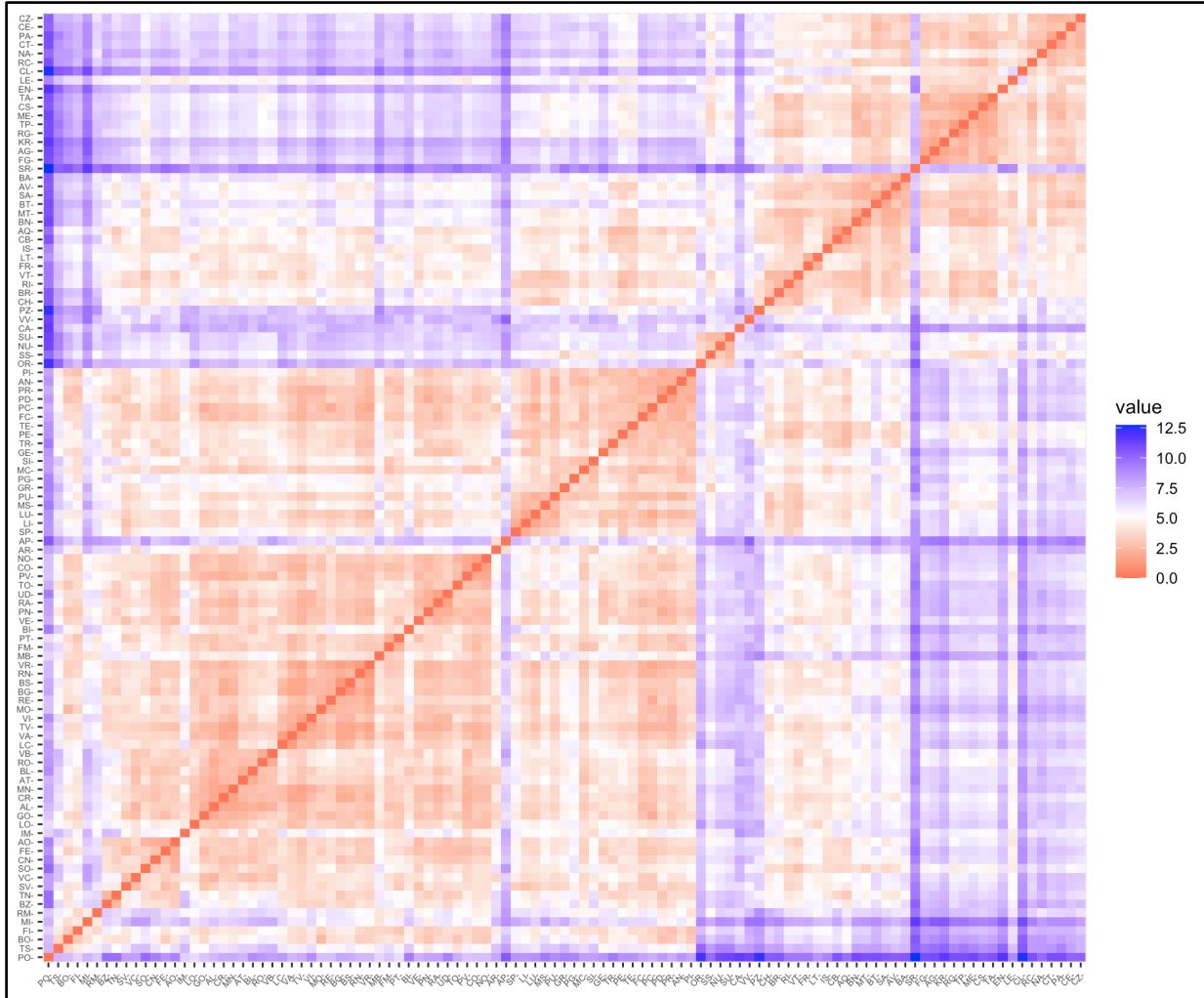
La costruzione della matrice delle distanze deve essere effettuata a partire dalla matrice Z degli scarti standardizzati, per evitare che unità di misura differenti o grandi differenze di scala tra le variabili influenzino il calcolo delle distanze.

La distanza euclidea calcolata fra la generica coppia di unità i e l della matrice degli scarti standardizzati è indicata nella formula 3.1.1 e rappresenta la radice quadrata della somma dei quadrati delle differenze dei valori assunti dalle unità statistiche i e l per ciascuna delle 15 variabili utilizzate⁸. La matrice delle distanze può essere visualizzata in Figura 3.1.1.

⁸ Xu R., Wunsch D., (2009), *Clustering*, Wiley, p. 22.

$$d_e(i, l) = \sqrt{\sum_{j=1}^{15} (z_{ij} - z_{lj})^2} \quad (3.1.1)$$

Figura 3.1.1 - Matrice delle distanze



3.2 Il metodo gerarchico agglomerativo e costruzione del dendrogramma

Il punto di partenza dei metodi gerarchici agglomerativi è la matrice delle distanze. Le unità statistiche vengono considerate inizialmente come cluster distinti, che progressivamente vengono uniti fino a formare un unico cluster. Ad ogni iterazione, la distanza tra due cluster o tra un cluster ed una unità viene calcolata in base al metodo utilizzato. Dopo ogni fusione viene ricalcolata la matrice delle distanze considerando l'unione dei cluster⁹.

⁹ Xu R., Wunsch D., (2009), *Clustering*, Wiley, pp. 32-33.

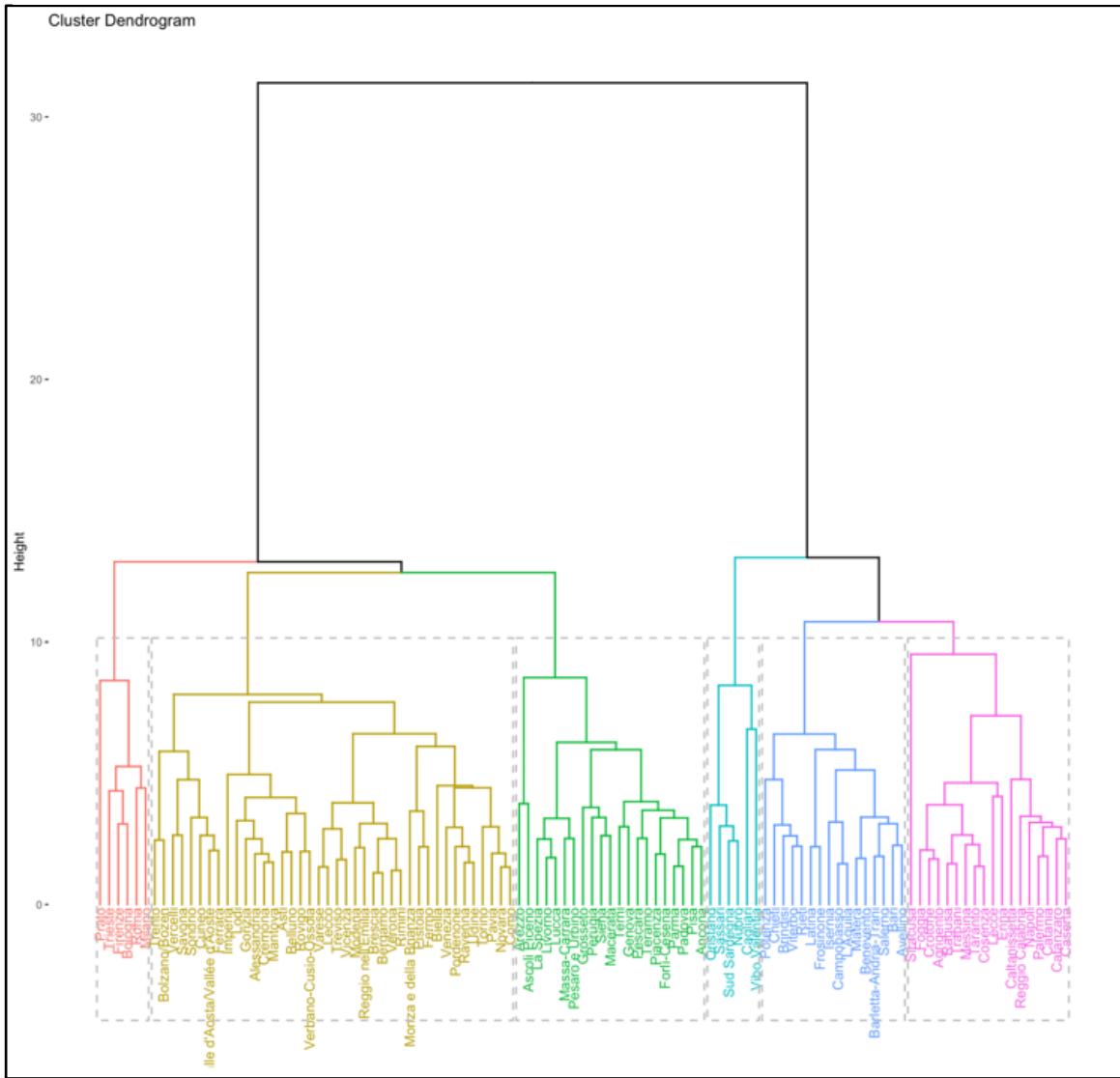
Il metodo scelto nell’analisi, il metodo di Ward, calcola ad ogni iterazione dell’algoritmo la distanza tra ciascuna coppia di cluster in termini di incremento della devianza within risultante dalla loro fusione. La funzione utilizzata per il calcolo delle distanze è la formula 3.2.1, dove n_i e n_j rappresentano la numerosità dei cluster C_i e C_j , mentre \bar{x}_i e \bar{x}_j sono i rispettivi centroidi dei cluster. Viene unita la coppia di cluster che produce il minore incremento della devianza within. Si può dimostrare che la minimizzazione dell’incremento della devianza nei gruppi equivale all’impiego della seguente distanza:

$$d(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\bar{x}_i - \bar{x}_j\|^2 \quad (3.2.1)$$

Il risultato del clustering è rappresentato nel dendrogramma in Figura 3.2.1 ed è stato tagliato in modo da ottenere 6 cluster. Al loro interno è possibile vedere alcune province isolate dal resto del cluster di appartenenza fenomeno che si è verificato utilizzando anche metodi diversi, come il metodo del legame medio, del legame singolo e del legame completo.

La provincia di Prato, ad esempio, è in un ramo isolato del dendrogramma. Questo può essere dovuto al fatto che presenta caratteristiche significativamente diverse dal resto del cluster. Osservando il biplot in Figura 2.2.1, infatti, si nota che la provincia di Prato si trova lontano dal resto delle province e inoltre assume un valore particolarmente elevato nel numero di imprese straniere, come osservabile nella mappa coropletica in Figura 1.1.7. Anche le province di Siracusa, Ascoli Piceno e Arezzo si trovano separate dal resto del proprio cluster. Queste province sono infatti accomunate da un valore di quota di export sul PIL significativamente più grande rispetto al resto delle province, come si può vedere in Figura 1.1.12. Infine, anche le province di Cagliari e Vibo Valentia possono essere considerate delle eccezioni, infatti entrambe sono accomunate dal fatto di avere valori di partecipazione alla formazione continua molto più alti del resto del paese, come si vede in Figura 1.1.13.

Figura 3.2.1 - Dendrogramma



3.3 Descrizione dei cluster

Per descrivere e confrontare i cluster, è stata calcolata la media aritmetica dei valori assunti delle variabili all'interno di ogni cluster. I risultati sono mostrati nella Tabella 3.3.1. Inoltre, è stata realizzata una mappa che rappresenta le province colorate in base al loro cluster di appartenenza (Figura 3.3.1), accompagnata da grafici a stella per facilitare una comparazione qualitativa tra i cluster (Figura 3.3.2), nei quali i valori delle medie aritmetiche degli indicatori sono stati normalizzati su un intervallo da 0 a 1.

Figura 3.3.1 - Mappa delle province italiane colorate per cluster



Figura 3.3.2 – Grafici a stella dei cluster



• Cluster 1

Il cluster 1 è composto da 21 province, principalmente situate nel Centro. Presenta il valore più alto di infortuni sul lavoro e il più basso di imprenditorialità giovanile. Ha valori alti anche nel numero di imprese straniere e di quota di export sul PIL.

• Cluster 2

Il cluster 2 comprende tutte le province della Sardegna, oltre alla provincia di Vibo Valentia, per un totale di 6 province. Questo cluster si distingue per i valori più bassi in Startup innovative, imprese che fanno e-commerce, nuove iscrizioni, cessazioni, imprese in fallimento, imprese straniere e quota di export sul PIL. Registra il valore più alto tra i cluster per la presenza di lavoratori domestici e un livello relativamente elevato di partecipazione alla formazione continua.

- **Cluster 3**

Il cluster 3 è composto da 40 province che fanno parte principalmente del Nord. Presenta i valori più elevati tra i cluster nel numero di cessazioni, nella quota di export sul PIL e nel numero di pensioni di vecchiaia. Mostra inoltre valori alti nel numero di imprese straniere e startup innovative. Registra invece valori bassi nel tasso di Neet, nel numero di imprese in fallimento e negli infortuni sul lavoro.

- **Cluster 4**

Il cluster 4 comprende province distribuite in diverse aree geografiche del paese, ma accomunate da performance molto elevate. Queste province sono tra quelle che hanno il punteggio più elevato della prima componente principale, come riportato nella Tabella 2.3.2. Il cluster si distingue per i valori più alti in startup innovative, imprese che fanno e-commerce, nuove iscrizioni, cessazioni, imprese in fallimento, imprese straniere, tasso di occupazione e partecipazione alla formazione continua. Al contrario, presenta i valori più bassi nel tasso di Neet, gender pay gap e infortuni sul lavoro.

- **Cluster 5**

Il cluster 5 è composto da 16 province, situate nel Centro-Sud. È il cluster che presenta il valore più alto nel gender pay gap e valori elevati negli infortuni sul lavoro e nell'imprenditorialità giovanile. Al contrario, registra valori bassi nel numero di imprese straniere, di lavoratori domestici e nuove iscrizioni.

- **Cluster 6**

Il cluster 6 è composto da 18 province del Sud. Si caratterizza per i valori più alti nell'imprenditorialità giovanile, oltre a registrare valori elevati nel tasso di Neet e negli infortuni sul lavoro. Al contrario, mostra i valori più bassi nel tasso di occupazione, nella presenza di lavoratori domestici, nella partecipazione alla formazione continua e nel numero di pensioni di vecchiaia.

Tabella 3.3.1 – Media aritmetica delle variabili dei cluster

Cluster	Startup innovative	Imprese che fanno ecommerce	Imprenditorialità giovanile	Nuove iscrizioni	Cessazioni	Imprese in fallimento	Imprese straniere
1	6.29	4.94	6.96	4.87	4.64	1.83	11.63
2	4.90	1.98	8.64	4.30	3.50	1.21	5.64
3	5.91	6.16	7.77	5.23	5.07	1.26	11.56
4	8.76	7.42	7.51	6.28	5.07	2.16	19.60
5	6.70	4.10	9.09	4.50	4.23	1.64	6.52
6	5.40	3.01	9.83	4.44	3.95	1.89	7.26
Cluster	Tasso di occupazione	Giovani che non lavorano e non studiano (Neet)	Gender pay gap	Lavoratori domestici	Quota di export sul Pil	Partecipazione alla formazione continua	Infortuni sul lavoro
1	71.22	13.80	32.28	18.24	41.44	10.66	14.87
2	56.53	21.00	27.05	27.33	13.28	12.60	10.62
3	72.43	14.44	31.88	14.70	43.33	9.08	9.24
4	74.60	13.70	27.02	23.59	39.11	12.92	7.67
5	58.40	20.92	34.01	8.69	22.49	9.07	12.89
6	47.02	31.43	31.15	6.65	14.44	6.47	12.57
Cluster	Numero pensioni di vecchiaia						
1	213.84						
2	175.25						
3	228.12						
4	206.08						
5	175.17						
6	140.53						

Conclusioni

Questo studio ha cercato di fotografare la situazione imprenditoriale e lavorativa nelle province italiane attraverso l'analisi degli indicatori della categoria “Affari e lavoro” utilizzati nell'indagine sulla qualità della vita del 2023 svolta dal Sole 24 Ore. L'analisi ha evidenziato una marcata differenza tra le province del Sud Italia e delle Isole rispetto alle province del resto del paese, evidenziata già inizialmente dall'osservazione delle mappe coropletiche e poi confermata dai punteggi della prima componente principale, utilizzati come indice sintetico delle performance delle province, dove le province del Sud e delle Isole hanno ottenuto punteggi negativi ad eccezione della provincia di Cagliari.

Dall'osservazione dell'autovettore associato alla prima componente principale è stato possibile osservare il contributo di ciascuna variabile ai punteggi. Il risultato più interessante è stato forse il contributo negativo dato dall'imprenditorialità giovanile, che è stata perciò identificata come un indicatore di un'economia che non offre sufficienti opportunità lavorative.

Tramite l'analisi dei cluster sono stati poi identificati 6 cluster delle province italiane. È evidente una suddivisione geografica, si può infatti osservare un cluster composto principalmente da province del Nord, uno da quelle del Centro, un altro include province del Centro-Sud, uno da quelle del Sud, e un cluster che comprende principalmente le province della Sardegna. Anche in questo caso si è visto che i cluster composti da province del Sud e del Centro-Sud sono caratterizzati principalmente da variabili con contributi negativi all'indice sintetico creato attraverso l'analisi delle componenti principali. È presente un ultimo cluster che raccoglie le province con performance particolarmente alte, le cui province si possono identificare nei principali centri del lavoro e delle imprese.

Bibliografia

Istituto Nazionale di Statistica (ISTAT), (2023), *Rapporto annuale 2023*, ISTAT.

Jolliffe I.T., (2002), *Principal component analysis*, Springer, pp. 1-26.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>

Xu R., Wunsch D., (2009), *Clustering*, Wiley.

Siti internet consultati

<https://www.istat.it/>

<https://github.com/>