POLITECNICO DI TORINO
Corso di Laurea Magistrale in Ingegneria Matematica

Elaborato finale per il corso di Data Spaces

# Indicatori socio-economici delle regioni italiane


Enrico Armando
s231568

Febbraio 2017

# Indice

**1 Introduzione** **2**

**2 Il dataset** **2**
2.1 Origine dei dati . . . . . 2
2.2 Struttura dei dati . . . . . 2
2.3 Esplorazione dei dati . . . . . 4
2.4 Correlazione tra le variabili . . . . . 10

**3 Clustering** **14**
3.1 Richiami teorici . . . . . 14
3.2 Applicazione al dataset . . . . . 16

**4 Analisi delle componenti principali** **18**
4.1 Richiami teorici . . . . . 18
4.2 Applicazione al dataset . . . . . 19

**5 Linear Discriminant Analysis** **22**
5.1 Richiami teorici . . . . . 22
5.2 Applicazione al dataset . . . . . 23

**6 Conclusioni** **26**

# 1 Introduzione

Lo scopo della seguente trattazione è approfondire l'applicazione di alcuni dei metodi studiati durante i corsi di Data Spaces e Modelli Statistici ad un insieme di dati reale, per poter apprezzare come queste tecniche riescano ad analizzare in modo diverso il comportamento delle variabili, e a spiegare fenomeni che non sempre sono alla portata di colui che si limita a leggere il contenuto di un database riga per riga.
La scelta è ricaduta su un dataset piccolo di indicatori socio-demografici molto eterogenei riguardanti le regioni italiane sia per cercare un legame tra queste variabili e vedere come questo si riflettesse sul raggruppamento e la classificazione dei record, sia per poter verificare con il senso comune il significato dei risultati ottenuti, essendo le realtà regionali oggetto di discorsi nella quotidianità.

# 2 Il dataset

## 2.1 Origine dei dati

Tra i vari siti web su cui è possibile reperire informazioni socio-demografiche riguardo all'Italia, è stato scelto *dati.italiaitalie*, contenente una grande quantità di dati statistici annuali riguardanti le provincie italiane con le aggregazioni in regioni e macro-regioni. A loro volta i dati provengono dal Dipartimento per lo sviluppo delle economie territoriali del Governo italiano. I dati utilizzati sono relativi al 2014.

## 2.2 Struttura dei dati

Il database online offre la possibilità di scaricare un foglio Excel per ogni indicatore socio-economico, contenente i valori per ogni provincia e aggregato, oppure un foglio excel per ogni provincia contenente i valori di più indicatori di una stessa tematica. Partendo da tabelle del primo tipo, sono stati utilizzati alcuni indicatori di vario tipo per costruire il dataframe, affiancando le colonne per mezzo del software Excel. Il livello di aggregazione considerato è stata la regione, dal momento che alcune provincie (soprattutto quelle di più recente costituizione) presentano dati mancanti. Ad una maggiore quantità di osservazioni si è preferito lavorare con dei dati aggregati

meno soggetti ad outlier, più facilmente confrontabili tra loro (ad esempio per dimensione geografica e popolazione).

Le regioni sono state arbitrariamente suddivise in tre gruppi a seconda della loro posizione geografica, secondo il senso comune, ottenendo un attributo di tipo categorico (**Zona**), con i seguenti valori: Nord (Piemonte, Valle d'Aosta, Lombardia, Trentino-Alto Adige, Veneto, Friuli-Venezia-Giulia, Liguria, Emilia-Romagna), Centro (Toscana, Umbria, Marche, Lazio, Abruzzo, Molise), Sud (Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna).

Gli indicatori ritenuti più interessanti per l'analisi sono i seguenti:

- **Pop**: numero di abitanti,
- **Dens**: densità di popolazione in $^{ab.}/_{km^2}$,
- **Age**: età media della popolazione residente,
- **For**: numero di stranieri ogni 1000 abitanti,
- **Unemp**: numero di disoccupati ogni 1000 abitanti,
- **Income**: reddito pro-capite medio,
- **Bank**: numero di sportelli bancari ogni 100.000 abitanti,
- **Diff**: raccolta differenziata in $^{kg}/_{ab.}$,
- **Exp**: esportazioni in $^{€}/_{ab.}$,
- **Patent**: numero di brevetti registrati ogni 100.000 abitanti,
- **Hotel**: numero di strutture ricettive ogni 100.000 abitanti.

Queste variabili sono state appositamente scelte in ambiti molto diversi, per rendere il più possibile osservabile la differenza tra le varie regioni sotto più aspetti. La figura 1 mostra il dataset così ottenuto, visualizzato tramite l'interfaccia di RStudio.

| | Region | Zona | Pop | Dens | Age | For | Unemp | Income | Bank | Diff | Exp | Patent | Hotel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Piemonte | Nord | 4424467 | 174.28033 | 45.9 | 96.16 | 51.02 | 19.30 | 56.37 | 247.13 | 9.663 | 125.39 | 134.62 |
| 2 | Valle d'Aosta | Nord | 128298 | 39.34438 | 44.8 | 70.73 | 42.07 | 20.94 | 74.05 | 253.34 | 4.738 | 55.34 | 869.85 |
| 3 | Lombardia | Nord | 10002615 | 419.15692 | 44.2 | 115.20 | 37.79 | 20.09 | 60.02 | 244.73 | 10.954 | 162.26 | 78.63 |
| 4 | Trentino-Alto Adige | Nord | 1055934 | 77.61082 | 42.6 | 91.06 | 27.08 | 19.92 | 86.75 | 302.96 | 6.891 | 72.92 | 1224.98 |
| 5 | Veneto | Nord | 4927596 | 267.69623 | 44.3 | 103.81 | 33.90 | 18.99 | 66.71 | 290.29 | 10.982 | 102.32 | 1070.83 |
| 6 | Friuli-Venezia Giulia | Nord | 1227122 | 156.07668 | 46.4 | 87.65 | 35.05 | 19.91 | 71.55 | 262.97 | 9.789 | 87.11 | 486.91 |
| 7 | Liguria | Nord | 1583263 | 292.31915 | 48.1 | 87.60 | 46.05 | 19.26 | 54.13 | 177.15 | 4.469 | 60.44 | 269.76 |
| 8 | Emilia-Romagna | Nord | 4450508 | 198.21637 | 45.4 | 120.60 | 38.93 | 20.41 | 72.35 | 330.99 | 11.901 | 139.31 | 199.91 |
| 9 | Toscana | Centro | 3752654 | 163.25082 | 46.1 | 105.41 | 45.97 | 18.25 | 61.21 | 250.06 | 8.520 | 102.94 | 342.88 |
| 10 | Umbria | Centro | 894762 | 105.70975 | 45.8 | 110.22 | 49.69 | 17.51 | 57.78 | 241.09 | 3.842 | 74.21 | 438.78 |
| 11 | Marche | Centro | 1550796 | 164.95416 | 45.4 | 93.58 | 45.25 | 17.99 | 70.61 | 273.35 | 8.050 | 116.84 | 280.63 |
| 12 | Lazio | Centro | 5892425 | 341.94089 | 43.9 | 108.02 | 55.84 | 17.93 | 43.84 | 140.23 | 3.106 | 125.48 | 157.20 |
| 13 | Abruzzo | Centro | 1331574 | 122.93148 | 44.9 | 64.77 | 51.31 | 14.83 | 47.54 | 193.26 | 5.205 | 50.47 | 189.93 |
| 14 | Molise | Centro | 313348 | 70.24720 | 45.5 | 34.47 | 57.49 | 14.62 | 43.40 | 78.63 | 1.185 | 34.15 | 147.12 |
| 15 | Campania | Sud | 5861529 | 428.75806 | 41.1 | 37.11 | 73.98 | 11.74 | 25.57 | 191.27 | 1.611 | 37.79 | 97.84 |
| 16 | Puglia | Sud | 4090105 | 209.30992 | 43.0 | 28.78 | 76.42 | 12.94 | 31.78 | 103.89 | 1.982 | 45.40 | 129.48 |
| 17 | Basilicata | Sud | 576619 | 57.24218 | 44.3 | 31.58 | 54.63 | 13.39 | 39.71 | 92.98 | 1.965 | 25.32 | 140.13 |
| 18 | Calabria | Sud | 1976631 | 129.85440 | 43.1 | 46.22 | 80.83 | 12.50 | 23.52 | 62.15 | 0.164 | 22.66 | 148.28 |
| 19 | Sicilia | Sud | 5092080 | 197.12000 | 42.6 | 34.19 | 73.95 | 12.04 | 31.05 | 63.13 | 1.895 | 27.30 | 116.47 |
| 20 | Sardegna | Sud | 1663286 | 69.01595 | 44.9 | 27.10 | 75.44 | 14.15 | 39.32 | 227.37 | 2.790 | 29.76 | 272.47 |

Figura 1: Il dataset

## 2.3 Esplorazione dei dati

Per comprendere se le differenze nei valori degli attributi scelti siano significative tra le varie regioni, e soprattutto tra le tre macro-aree, è possibile osservare le quantità attraverso uno strumento più tradizionale come il *box-plot*, oppure tramite una mappa dell'Italia. Quest'ultima visualizzazione è resa possibile in R grazie al pacchetto `mapIT` sviluppato da *Quantide*.
Si riportano i risultati più interessanti.

**Popolazione.** Come si può osservare in figura 2, la Lombardia appare come la regione più popolosa, ma non ci sono molte differenze tra Nord, Centro e Sud, siccome le regioni più piccole e meno popolose sono equamente distribuite. Proprio per la Lombardia si ha al Nord la maggiore variabilità.

**Età media.** L'età media dei cittadini italiani non è uguale in tutte le regioni, ma varia considerevolmente tra i 41 e i 48 anni. In particolare si vede in figura 3 che il Sud può definirsi più giovane, in particolare la Campania, mentre il Nord lo è di meno, anzi è emblematico il dato della Liguria, forse perchè meta di pensionati che scelgono di trascorrervi gli ultimi anni di vita.

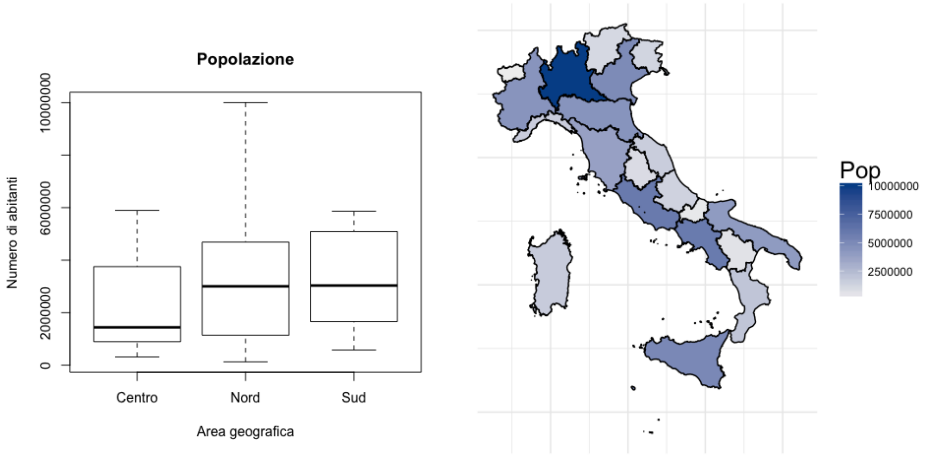


Figura 2: Popolazione

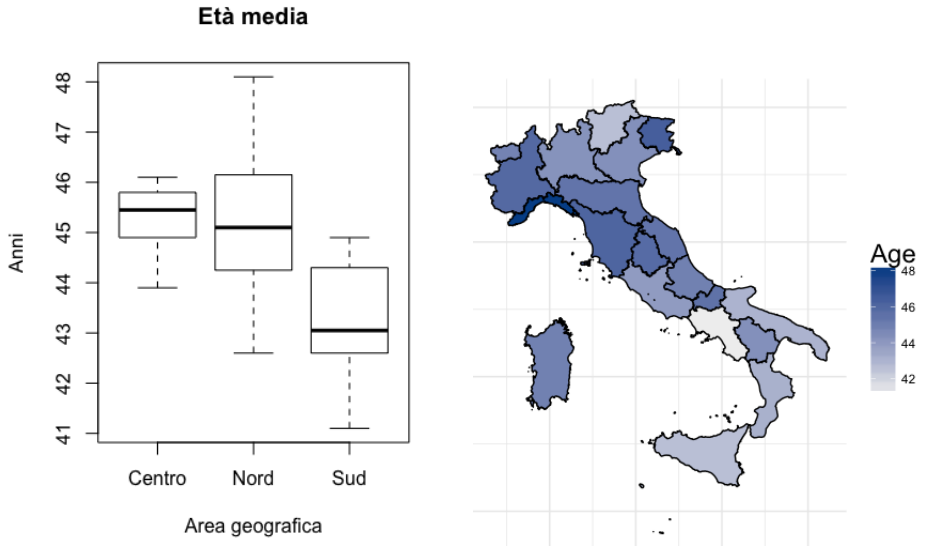


Figura 3: Età media

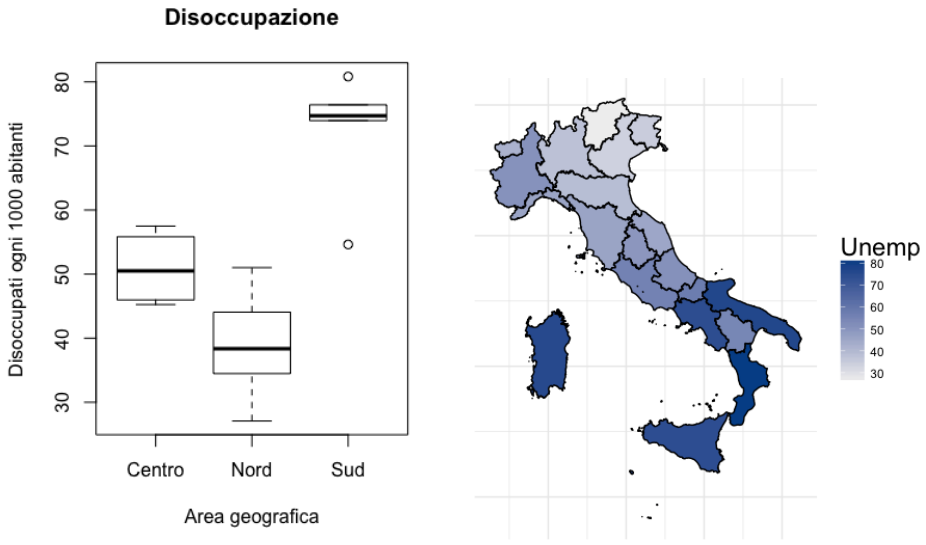


Figura 4: Disoccupazione

**Disoccupazione.** Il tasso di disoccupazione (qui espresso in numero di disoccupati ogni 1000 abitanti) è il primo attributo che marca una netta differenza tra le aree geografiche. Si può osservare che la media per le regioni del Sud è addirittura doppia di quella del Nord, e soltanto la Basilicata sembra farvi eccezione.

**Stranieri.** Contrariamente a quanto appena visto per la disoccupazione, il numero di stranieri residenti (ogni 1000 abitanti) è più alto al Nord rispetto al Sud: non stupisce questo forte legame con la variabile precedente, dal momento che è preferibile insediarsi dove è più facile trovare lavoro.

**Reddito pro-capite.** La distribuzione del reddito medio pro-capite in Italia è analoga a quella della popolazione straniera (figura 6). Questo accostamento potrebbe apparire inconsueto, ma entrambi questi dati sono indicatori del benessere di un'area. I valori più alti al Nord si incontrano nelle regioni a statuto speciale, e la variabilità all'interno delle tre aree è ridotta.

**Banche.** La presenza di sportelli bancari sul territorio (qui in occorrenze ogni 100.000 abitanti) è un altro indicatore di ricchezza. Infatti non stupisce la distribuzione simile alle due precedenti. Il valore alto può dipendere

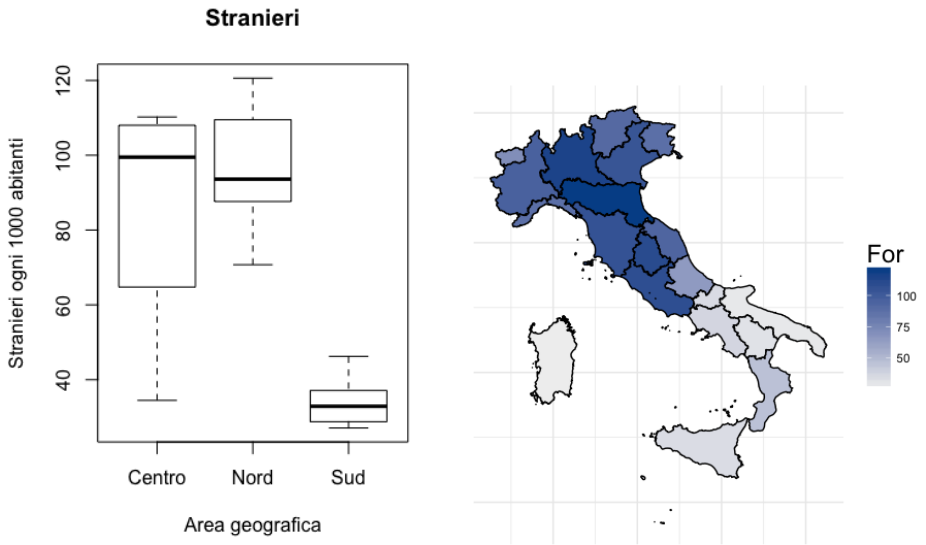


Figura 5: Presenza di stranieri

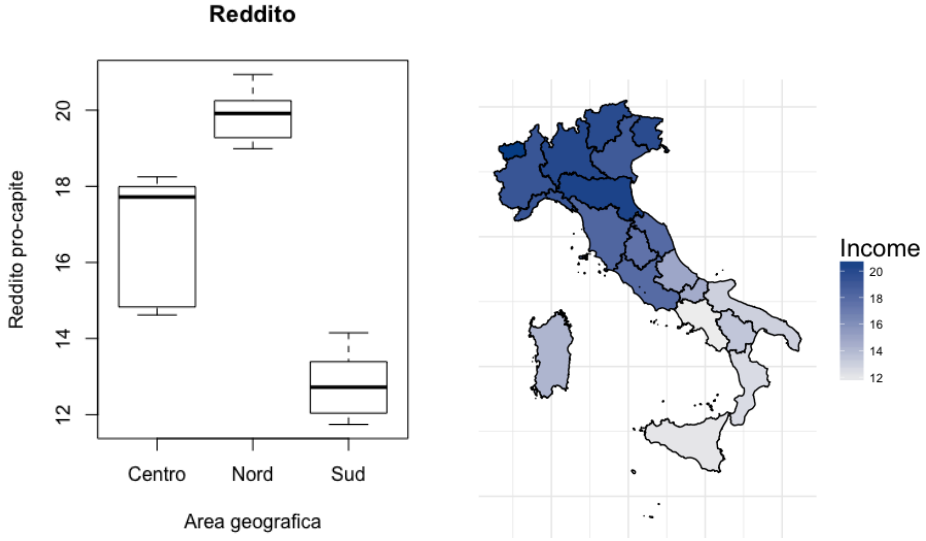


Figura 6: Reddito medio pro-capite

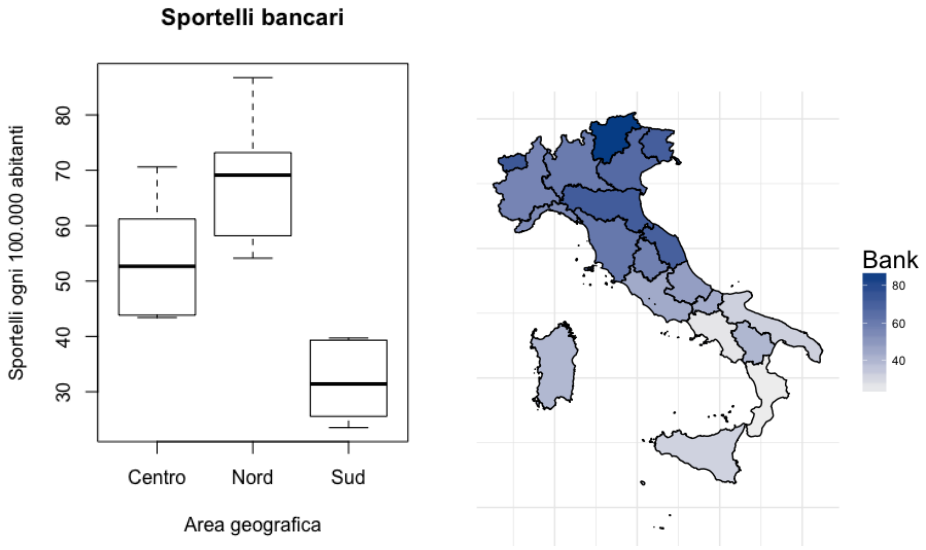


Figura 7: Presenza di sportelli bancari

anche da fattori come la distanza geografica tra i centri abitati, in casi come Trentino e Valle d'Aosta.

**Raccolta differenziata.** Ancora una volta, le regioni più ricche riciclano di più. Si può osservare però che la Campania, forse in seguito all'attenzione mediatica, costituisce un esempio virtuoso per le regioni vicine.

**Esportazioni.** Altro indice di ricchezza, con valori che decrescono scendendo lungo la penisola. Come si vede in figura 9, i valori più alti si riscontrano nelle regioni della Pianura Padana, mentre al Sud i valori sono bassi e simili tra di loro.

**Brevetti.** Un indicatore curioso è il numero di brevetti depositati in ogni regione, qui normalizzato rispetto alla popolazione. Anche in questo caso il Nord si dimostra più attivo, ma i valori più interessanti riguardano le due principali città, Roma e Milano, in quanto sede di aziende e università.

**Ricettività.** Infine, ecco il dato relativo alla presenza di strutture ricettive ed alberghiere, che vuole sintetizzare lo sviluppo del turismo nelle regioni.

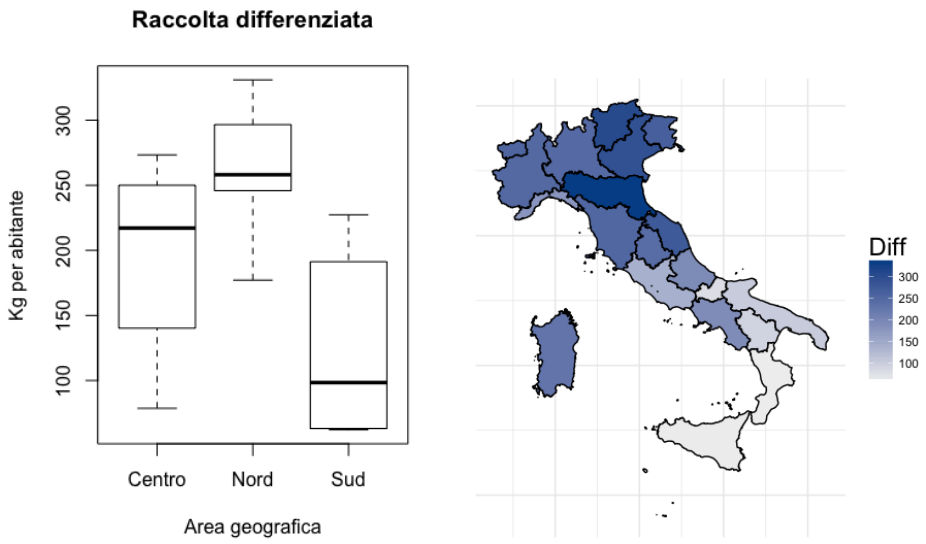


Figura 8: Raccolta differenziata

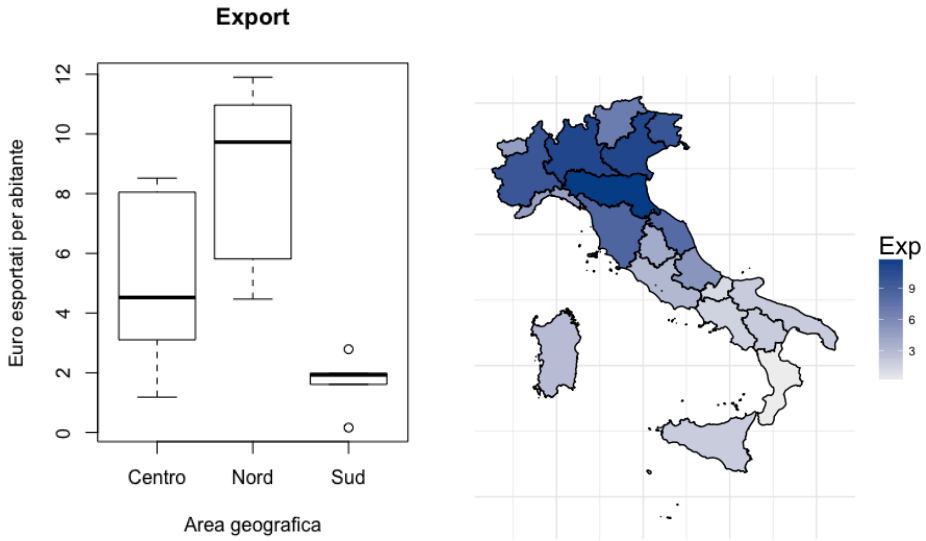


Figura 9: Esportazioni

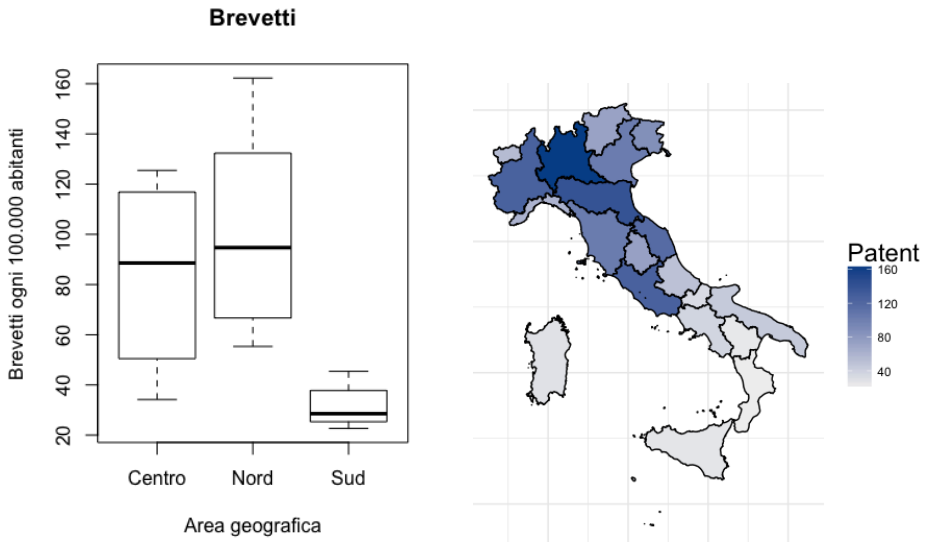


Figura 10: Brevetti depositati

L'attributo, sempre normalizzato rispetto alla popolazione, presenta dei valori singolarmente alti in Trentino-Alto Adige, Valle d'Aosta e Veneto: questo potrebbe essere dovuto per le prime due regioni al calcolo dei rifugi alpini insieme agli hotel tradizionali, e nel caso del Veneto alla presenza di Venezia e della frammentazione delle strutture ricettive tra le isole che la costituiscono.

## 2.4 Correlazione tra le variabili

A questo punto risulta naturale chiedersi se ci siano dei legami tra le variabili, avendo osservato come i grafici siano molto simili tra loro per alcuni degli indicatori di interesse.

Per evidenziare le correlazioni è naturale costruire un correlogramma. Per iniziare, è utile uno schema contenente tutte le variabili, come quello in figura 12, costruito per mezzo del pacchetto `corrgram`. L'intensità del colore rappresenta la forza del legame (valori in modulo prossimi a uno), mentre il blu indica correlazione positiva, il rosso negativa.

Appaiono evidenti alcuni risultati, poichè molte variabili sono fortemente correlate tra di loro: di tratta della coppia Popolazione-Densità e di tutto

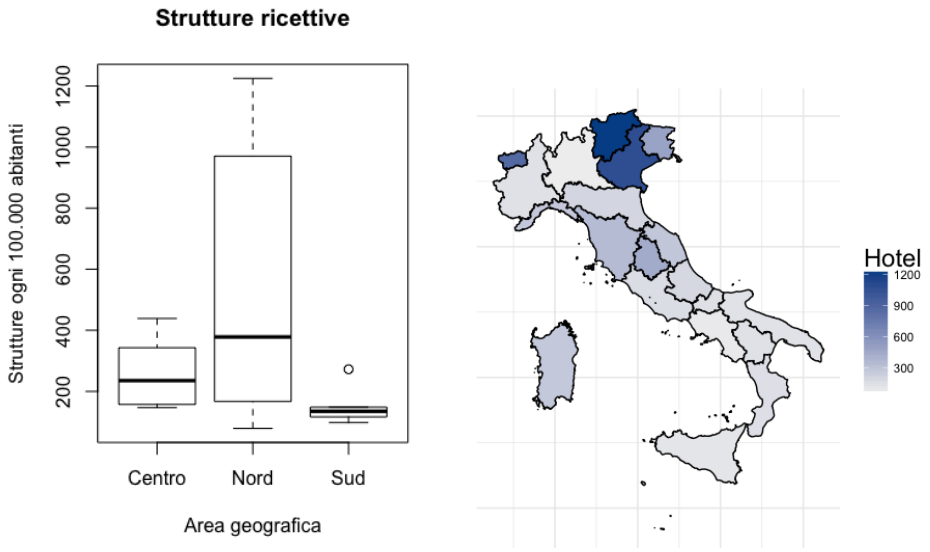


Figura 11: Strutture ricettive

il gruppo Stranieri-Reddito-Banche-Differenziata-Esportazioni-Brevetti. A quest'ultimo gruppo, che si potrebbe definire *Benessere*, va aggiunta anche la Disoccupazione, semplicemente correlata negativamente. Questo legame non stupisce, avendo osservato grafici molto simili per la distrbuzione di questi indicatori tra le regioni.

Si può vedere la correlazione nel gruppo *Benessere* ancora più nel dettaglio, per mezzo della funzione `pairs`, che produce la tabella di figura 13: i valori (in valore assoluto) variano da 0.6 a 0.9, mentre gli scatterplot sotto la diagonale mostrano la linearità o quasi della correlazione.

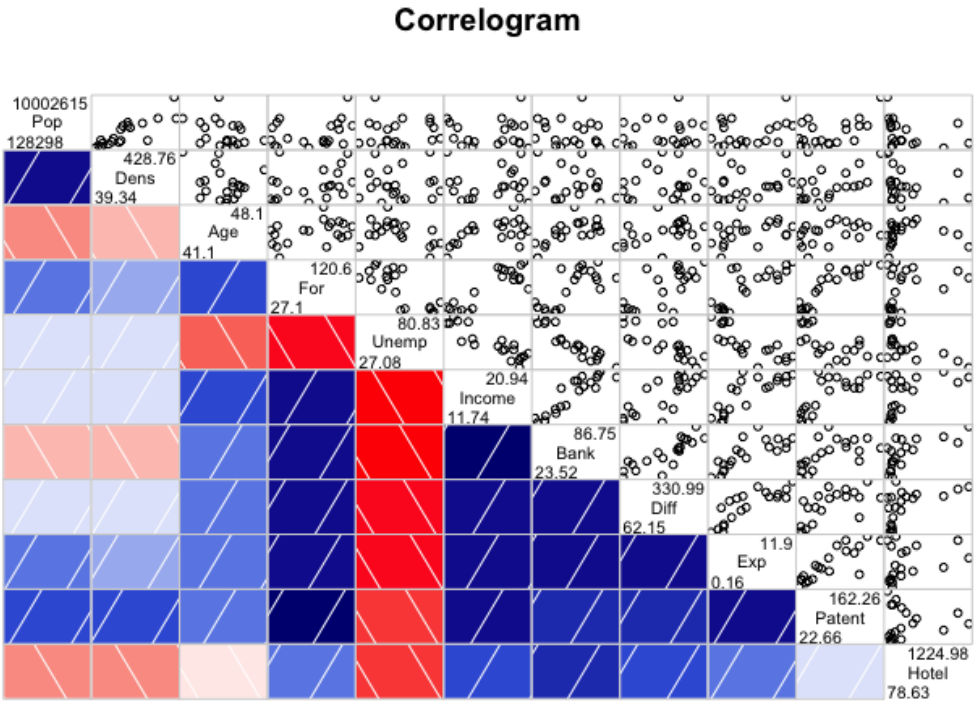


Figura 12: Correlogramma con tutte le variabili

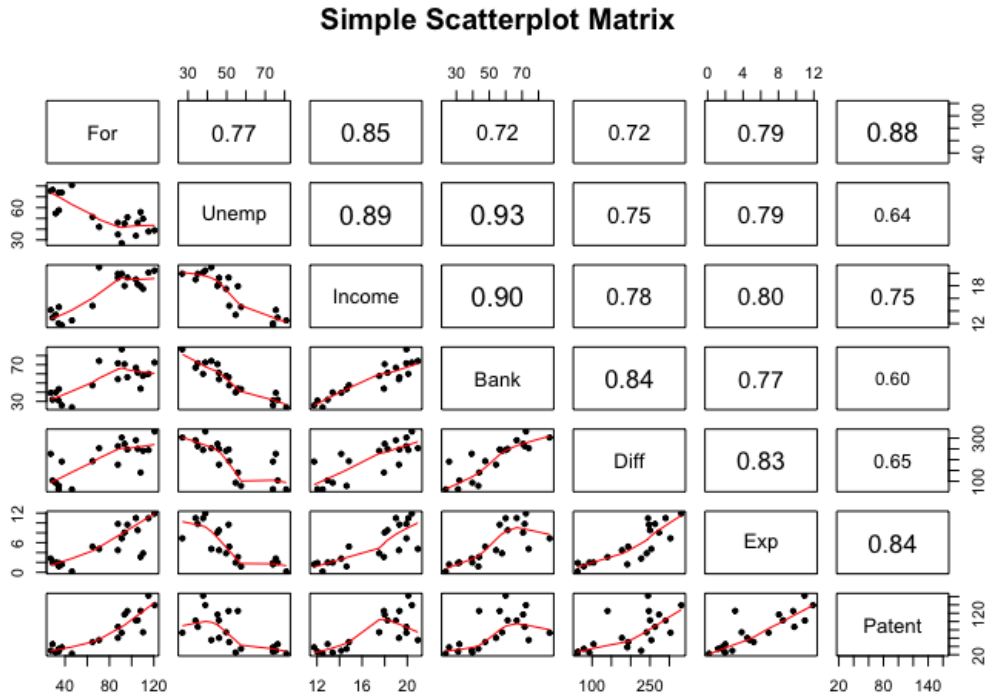


Figura 13: Correlogramma per il gruppo *Benessere*

# 3 Clustering

## 3.1 Richiami teorici

Il clustering, o analisi dei gruppi, è una tecnica che mira al raggruppamento dei dati in insiemi omogenei di osservazioni. Anzitutto è necessario definire il concetto di similarità - o dissimilarità - tra i dati, che è concepita come distanza in uno spazio multi-dimensionale. Le metriche più comunemente utilizzate negli algoritmi di clustering, per definire la lontananza tra un oggetto $p$ e un oggetto $q$, sono le seguenti:

- la distanza di Manhattan, derivata dalla norma $l_1$

$$d_1(p,q) = \sum_{i=1}^{n} |p_i - q_i|,$$

- la distanza euclidea, derivata dall'omonima norma

$$d_2(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2},$$

- la distanza di Minkowski, che generalizza le precedenti

$$d_m(p,q) = \sqrt[m]{\sum_{i=1}^{n} |p_i - q_i|^m},$$

- la distanza di Canberra, versione pesata di quella di Manhattan

$$d_c(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|},$$

- la distanza di Mahalanobis, basata sulle correlazioni tra le variabili.

Le tecniche di clustering si dividono in due sottogruppi, gerarchiche e non gerarchiche. Il secondo caso è rappresentato dal *k-means clustering*, dove il numero di gruppi è assegnato in partenza. Il clustering gerarchico, invece, si suddivide ulteriormente in due tipologie:

- agglomerativo, in cui si parte da una configurazione in cui ogni elemento rappresenta un gruppo, e poi attraverso di *similarità* si raggruppano uno per volta gli elementi, fino ad arrivare ad un unico cluster con tutte le osservazioni,
- divisivo, che contrariamente al precedente parte dall'insieme di tutti gli oggetti nello stesso cluster e procede separando gli elementi individuando la *dissimilarità*.

Un'ultima scelta per chi disegna il modello è il criterio di similarità appunto, per misurare la distanza tra insiemi di punti, in particolare i cluster ad ogni passaggio. Anche in questo caso sono possibili diversi criteri:

- *single linkage*, che tiene conto della distanza tra i due punti più vicini
- *complete linkage*, che tiene conto della distanza tra i due punti più lontani,
- *average linkage*, una media delle distanze tra tutte le possibili coppie di punti.

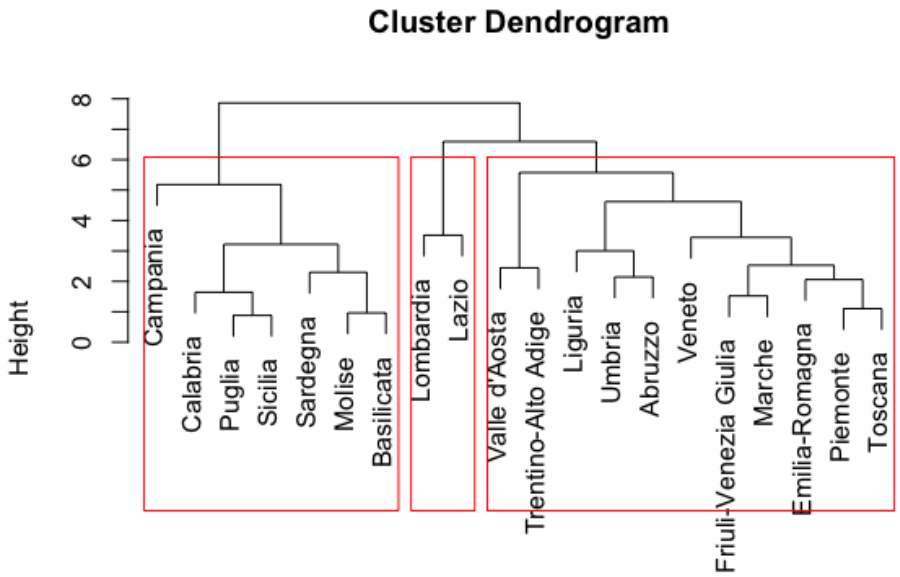


Figura 14: 3 cluster, distanza di Minkowski, average linkage

## 3.2 Applicazione al dataset

Per il dataset in esame si è scelto di operare con un algoritmo gerarchico agglomerativo. È interessante notare come scelte diverse della distanza o della similarità portino a diversi raggruppamenti delle regioni.

Il primo esempio, riportato in figura 14, è stato ottenuto scegliendo la distanza di Minkowski e il criterio di *complete linkage*; i rettangoli rossi evidenziano la composizione dei cluster scelti in numero di 3, ottenuti andando a tagliare l'albero (detto *dendrogramma*) due passi prima del raggruppamento totale. Si è scelto arbitrariamente il numero 3 per evidenziare alcuni risultati: il gruppo di sinistra comprende tutte le regioni del Sud più Abruzzo e Molise, considerate dunque vicine anche da questo algoritmo, mentre il gruppo centrale comprende le due regioni con le principali città, Roma e Milano, affini tra loro e con caratteristiche diverse dagli altri territori, e nel terzo gruppo tutte le altre regioni.

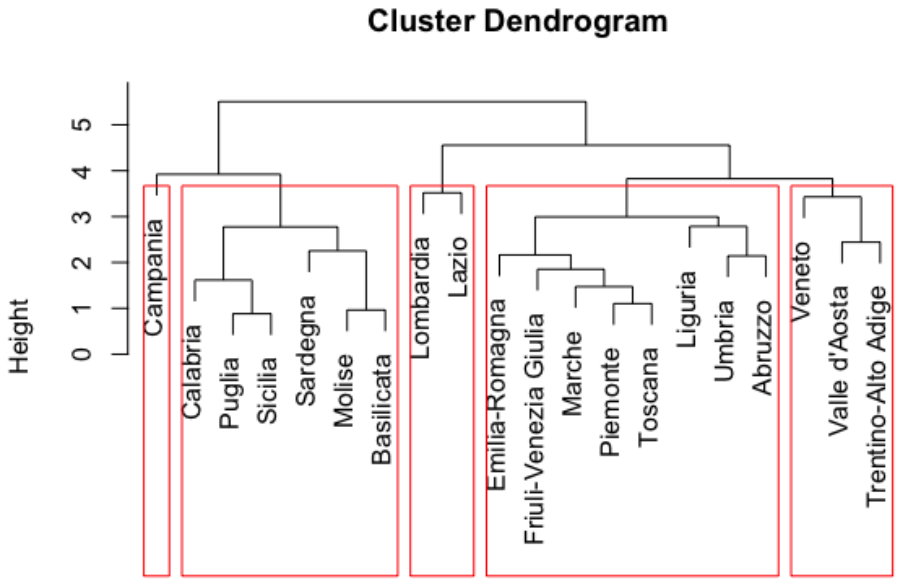


Figura 15: 5 cluster, distanza euclidea, complete linkage

Può essere interessante considerare un dendrogramma tagliato ad una diversa altezza: quello in figura 15 ad esempio è stato troncato alla quinta

ramificazione. In primo luogo l'albero è molto simile a quello ottenuto nel caso precedente, tuttavia la divisione in 5 gruppi mette in evidenza altre similarità tra i dati: rimangono insieme Lazio e Lombardia, mentre la Campania viene separata dalle altre regioni del Sud (si ricorda infatti un'età media più bassa) mentre Valle d'Aosta, Veneto e Trentino-Alto Adige si separano dal gruppo Centro-Nord, principalmente per l'influenza dell'attributo Hotel.

In entrambi i casi il numero di cluster è stato scelto arbitrariamente per poter mostrare alcune proprietà che rispecchiassero il significato delle variabili. Tuttavia, non è stato rispettato alcun criterio di ottimalità per determinare il miglior numero di cluster. Se ne possono considerare due:

- il WSS, ossia *within sum of squares*, che prevede di calcolare per ogni gruppo la varianza tra gli elementi che ne fanno parte. Graficamente, con il *metodo del gomito* si sceglie il numero di cluster relativo al punto di massima decrescenza della curva del WSS;
- il metodo della *silhouette*, in cui per ogni cluster viene definita il valore di silhouette come rapporto tra coesione tra i suoi membri e la separazione degli altri gruppi. I valori assumibili sono tra -1 e 1, e 1 indica un clustering perfetto. Si sceglie il numero di cluster per cui il valore medio di silhouette sia più vicino a 1.

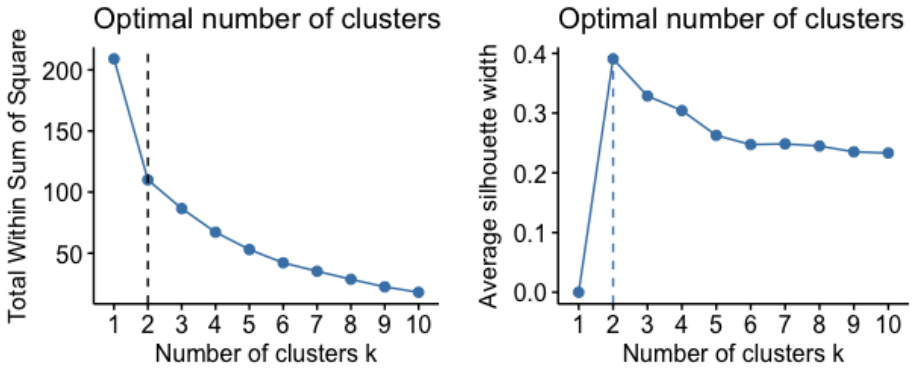


Figura 16: Scelta del numero ottimale di cluster

Come si osserva i figura 16 entrambe i criteri, applicati ai dati in esame, hanno indicato che il numero ottimale di cluster è 2, indicando dunque che

le regioni del Sud + Abruzzo + Molise sono molto simili tra loro e lontane da tutte le altre.

# 4 Analisi delle componenti principali

## 4.1 Richiami teorici

La PCA, o Analisi delle componenti principali, è una tecnica di apprendimento non supervisionato che permette di trovare una combinazione lineare delle variabili in esame tale da sintetizzare i dati nel miglior modo possibile, permettendo in secondo luogo di attuare una riduzione della dimensionalità. Consideriamo un dataset $X$ di $n \times p$ osservazioni, costruito in modo che ciascuna riga sia generata da una normale p-variata:

$$X^i_{(\cdot)} = N_p(\mu, \Sigma), \;\; i : 1...n,$$

con $\mu$ media delle osservazioni e $\Sigma$ matrice di varianze e covarianze, simmetrica e definita non negativa. Il procedimento è il seguente:

1. Si ruota $X$ per ottenere delle variabili $Y^i$ scorrelate, a partire da una nuova matrice di varianze e covarianze $\Lambda$ diagonale. Infatti esiste $\Gamma$ matrice ortogonale tale che

$$\Lambda = \Gamma^t \Sigma \Gamma \qquad e \qquad Y = \Gamma X.$$

2. Si ordinano in modo decrescente gli autovalori della matrice diagonale $\Gamma$:

$$\lambda_1 \geq \lambda_2 \geq ... \geq 0.$$

   In questo modo le combinazioni lineari $Y^1, Y^2, ..., Y^n$ si dicono rispettivamente prima, seconda, ..., ultima componente principale.

La proprietà più interessante di questo ordinamento è che le $Y^i$ sono disposte per varianza decrescente, e che non esiste alcuna altra combinazione delle $X^i$ con varianza maggiore di $Y^1$. Sarà sufficiente selezionare solo alcune delle componenti principali ottenute per descrivere il nostro dataset con un livello di variabilità fissato.

## 4.2 Applicazione al dataset

Al dataset vengono rimosse le colonne corrispondenti agli attributi categorici, ossia il nome e l'area geografica, e i dati vengono poi scalati in modo che ogni colonna abbia media 0 e varianza unitaria.

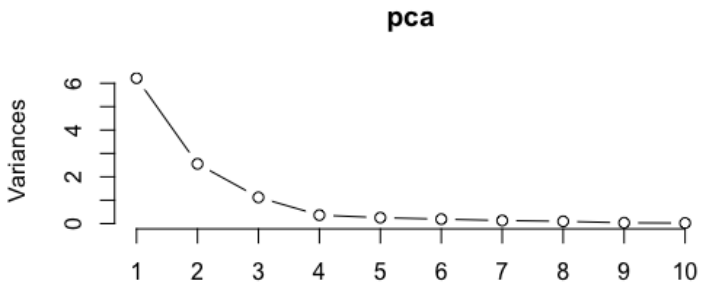


Figura 17: Varianza delle componenti principali

Come mostra la figura 17, solo le prime componenti principali calcolate portano con sè una parte consistente di varianza. Già la prima (56.5%) e la seconda (23.2%) riassumono circa l'80% della variabilità dei dati. Si riportano in figura 18 i valori delle combinazioni per le prima quattro componenti, la cui somma delle varianze ammonta quasi alla totalità.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Pop | -0.05232661 | 0.59599141 | -0.14842013 | 0.06802870 |
| Dens | -0.03849361 | 0.56624521 | -0.08144708 | -0.48629216 |
| Age | -0.19294989 | -0.19574018 | 0.73443394 | -0.20714179 |
| For | -0.36084859 | 0.14424945 | 0.11080079 | -0.21509495 |
| Unemp | 0.36970779 | 0.11417203 | 0.06636649 | 0.25584986 |
| Income | -0.38298867 | -0.05048526 | 0.07402591 | -0.23596114 |
| Bank | -0.37047479 | -0.19766196 | -0.11622250 | 0.03581305 |
| Diff | -0.35213785 | -0.03572523 | -0.16927663 | 0.46559100 |
| Exp | -0.36411823 | 0.12215763 | 0.02226973 | 0.43851592 |
| Patent | -0.32825652 | 0.30828178 | 0.13146723 | 0.15576511 |
| Hotel | -0.20829684 | -0.31487743 | -0.59147413 | -0.33758721 |

Figura 18: Le prime quattro componenti principali

Un tale risultato permette una visualizzazione bidimensionale delle osservazioni, in cui gli assi cartesiani sono costituiti proprio dalle due componenti principali. Si ottiene dunque il *bi-plot* di figura 19, sul quale si possono fare delle interessanti osservazioni.

Innanzitutto si è scelto di mostrare anche l'attributo categorico Zona, escluso dall'analisi PCA: si vede come già la prima componente principale abbia permesso una separazione netta tra Nord e Sud, mentre gli elementi del Centro possono essere incontrati in entrambi gli altri gruppi, essendo idealmente disposti su un ellisse trasversale. Osservando i valori numerici di figura 18, si vede come la prima componente principale sia formata dagli attributi del gruppo *Benessere* in pressochè uguale misura, mentre la seconda sia individuata essenzialmente da Densità e Popolazione, due variabili che, come osservato al punto precedente, sono fortemente correlate tra loro ma indipendenti dalle altre.

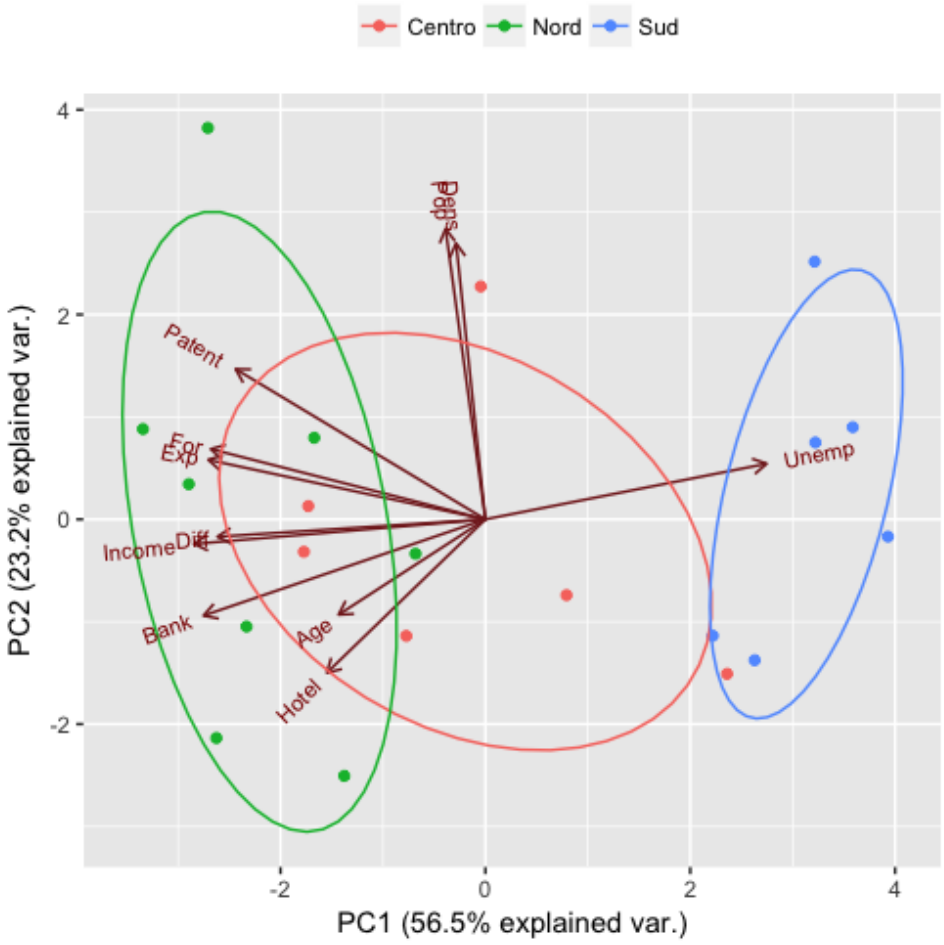


Figura 19: Le prime quattro componenti principali

# 5 Linear Discriminant Analysis

## 5.1 Richiami teorici

La LDA è una tecnica di *supervised learning*, che utilizza la riduzione della dimensionalità di un dataset per creare raggruppamenti in classi. Il risultato è simile a quello dell'Analisi delle componenti principali, ma il punto di partenza è diverso, dal momento che la PCA non tiene conto a priori delle differenze di classe, mentre la LDA le considera, essendo fornite al modello (apprendimento supervisionato) e anzi cerca di spiegarle.

Detti $\Pi_1$, $\Pi_2$, ..., $\Pi_k$ i gruppi in cui sono suddivisi i dati, e $\pi_i$ le rispettive probabilità *a priori* di appartenenza, e supponendo di conoscere queste grandezze, ci si interroga sulla probabilità *a posteriori* di appartenenza ad una classe. A seconda della conoscenza della funzione di densità di $P(X = x|X \in \Pi_i)\ \forall i$, che indica la ripartizione degli elementi in un gruppo, la LDA si differenzia in tre tipi:

- **LDA semplice**: è nota tale distribuzione, perciò per classificare un elemento è sufficiente applicare il Teorema di Bayes per calcolare

  $$P(X \in \Pi_i|X) = \frac{P(X|X \in \Pi_i)\pi_i}{P(X)}$$

  per ciascuna delle classi, ed infine assegnare all'elemento $X$ la classe $\Pi_j$ tale da massimizzare l'appartenenza a quella classe, ossia l'indice corrispondente a $\max_i(P(X \in \Pi_i|X))$;

- **LDA parametrico**: è un caso analogo al precedente, ma la distribuzione di $P(X = x|X \in \Pi_i)$ è nota a meno di parametri da determinare. Ad esempio si potrebbe sapere che tale distribuzione è normale, ma di media e varianza sconosciuti; il procedimento prevede di stimare in primo luogo i parametri e poi ricondursi al caso precedente;

- **LDA non parametrico**: la distribuzione non è nota, perciò si deve procedere diversamente, con un approccio simile a quello della PCA. Detta $S$ la matrice di varianze e covarianze del dataset $X$, si cercano delle combinazioni lineari di $X$, $Y = a^tX$, la cui varianza sia il più possibile distribuita tra un gruppo e l'altro piuttosto che all'interno dei

singoli gruppi. Ossia,

$$\mathbf{VarCov}(Y) = a^t S a = a^t B a + a^t W a,$$

dove $B$ rappresenta la varianza *between groups* e $W$ quella *within groups*, si vuole massimizzare il rapporto $\frac{a^t B a}{a^t W a}$.

## 5.2 Applicazione al dataset

Nel caso in oggetto, le 20 regioni sono state suddivise (per convenzione geografica) in 3 zone: Nord, Centro e Sud. L'analisi sarà di tipo non parametrico, ovviamente, in quanto non sono note a priori le peculiarità che determinano l'appartenenza di un elemento ad un gruppo. Il programma R ha fornito come output del calcolo del LDA soltanto due nuove variabili, in grado di spiegare rispettivamente l'88% e il 12% della varianza totale. I vettori di trasformazione sono mostrati in figura 20, ed evidenziano una grande rilevanza della variabile Income, con piccoli contributi di Exp e Age, trascurando quasi le altre variabili.

| | LD1 | LD2 |
|---|---|---|
| Pop | 0.000 | 0.000 |
| Dens | 0.008 | 0.003 |
| Age | -0.737 | -1.001 |
| For | -0.014 | -0.027 |
| Unemp | -0.045 | 0.081 |
| Income | 2.122 | 1.326 |
| Bank | -0.119 | -0.079 |
| Diff | -0.012 | -0.005 |
| Exp | 0.841 | 0.826 |
| Patent | -0.063 | -0.078 |
| Hotel | -0.001 | -0.002 |

Figura 20: La trasformazione nelle nuove variabili

Il grafico in figura 21 rappresenta la visualizzazione bi-dimensionale (*biplot*) della proiezione del dataset lungo le nuove variabili. Sono stati messi in risalto i gruppi di origine in modo da rendere evidente che la classificazione porta dei buoni risultati: i due ellissi relativi a Nord e Sud non presentano intersezioni (ossia dubbi di classificazione), mentre l'ellisse rosso relativo al

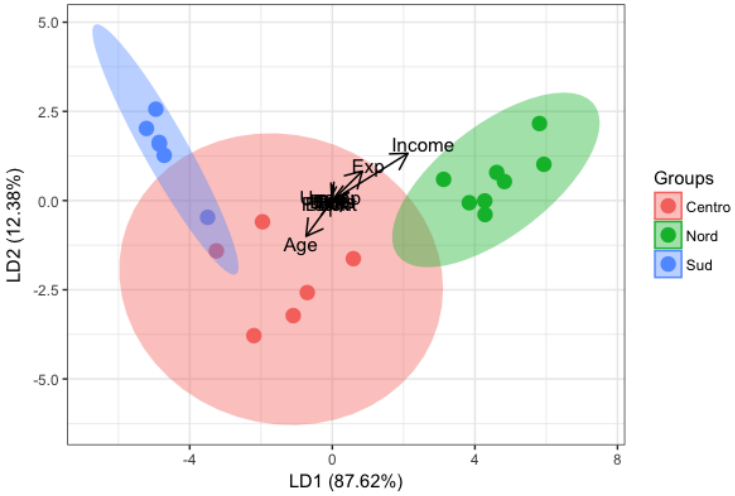


Figura 21: *bi-plot* con le nuove variabili

gruppo Centro li interseca entrambi per piccole porzioni.

Viene naturale a questo punto interrogarsi sulla bontà della classificazione. Uno dei parametri che è possibile specificare nella funzione `lda` di R è `CV=true`, che indica la richiesta dell'utente di effettuare la cross-validation. L'algoritmo insomma mette alla prova la propria bontà, cercando di classificare i record che vengono dati in input. L'approccio naturalmente è quello di dividere l'insieme fornito in *training set* e *test set* in modo da allenare e testare il modello; in questo caso la strategia è del tipo *leave-one-out*, in cui ciascun record, a turno, viene usato come *test set* mentre tutti gli altri costituiscono il *train set*.

In figura 22 si mostra il risultato di questa operazione, e per ogni regione si vede sulle colonne la probabilità di assegnazione in una determinata classe. Si può osservare che, mentre per molte regioni la probabilità è del 100% e l'assegnazione è corretta, per altre non c'è certezza e anzi, il risultato della classificazione porta ad errori. È il caso della Liguria e della Basilicata, assegnate alla classe Centro anzichè a Nord e Sud rispettivamente, oppure di Abruzzo e Molise, classificate come Sud anzichè Centro. Come caso intermedio, ad esempio, il Trentino-Alto Adige ha una probabilità del 61% di essere assegnato al Nord, e del 39% di essere ssegnato al Centro.

| | Centro | Nord | Sud |
|---|---|---|---|
| Piemonte | 0.001 | 0.999 | 0.000 |
| Valle d'Aosta | 0.000 | 1.000 | 0.000 |
| Lombardia | 0.000 | 1.000 | 0.000 |
| Trentino-Alto Adige | 0.391 | 0.609 | 0.000 |
| Veneto | 0.000 | 1.000 | 0.000 |
| Friuli-Venezia Giulia | 0.000 | 1.000 | 0.000 |
| Liguria | 1.000 | 0.000 | 0.000 |
| Emilia-Romagna | 0.000 | 1.000 | 0.000 |
| Toscana | 0.759 | 0.241 | 0.000 |
| Umbria | 1.000 | 0.000 | 0.000 |
| Marche | 1.000 | 0.000 | 0.000 |
| Lazio | 0.990 | 0.010 | 0.000 |
| Abruzzo | 0.042 | 0.000 | 0.958 |
| Molise | 0.000 | 0.000 | 1.000 |
| Campania | 0.000 | 0.000 | 1.000 |
| Puglia | 0.000 | 0.000 | 1.000 |
| Basilicata | 1.000 | 0.000 | 0.000 |
| Calabria | 0.000 | 0.000 | 1.000 |
| Sicilia | 0.000 | 0.000 | 1.000 |
| Sardegna | 0.290 | 0.000 | 0.710 |

Figura 22: Probabilità di assegnazione alle classi

Questo ragionamento permette di evidenziare quali regioni, pur appartenendo geograficamente ad un'area, abbiano proprietà (secondo gli indicatori scelti) affini a quelle di altre aree. Volendo riassumere il risultato della classificazione, è possibile costruire la *matrice di confusione*: questa tabella presenta sulle righe le classi reali a cui appartengono gli oggetti, e sulle colonne le classi predette. In figura 23, la matrice per il dataset in esame, riassume che 2 regioni del Centro sono erroneamente assegnate al Sud, e il Centro ottiene una regione di troppo dal Nord e una dal Sud. Come indice di *accuratezza* globale, si può affermare che l'80% delle regioni ($\frac{4+7+5}{20}$) sono state classificate correttamente, mentre all'interno di ciascun gruppo 4/6 regioni del Centro, 7/8 regioni del Nord e 5/6 regioni del Sud.

| | Centro | Nord | Sud |
|---|---|---|---|
| Centro | 4 | 0 | 2 |
| Nord | 1 | 7 | 0 |
| Sud | 1 | 0 | 5 |

Figura 23: Matrice di confusione

# 6 Conclusioni

Nel corso dell'elaborato sono state utilizzate tre delle principali tecniche di modellazione statistica, due di apprendimento non supervisionato, una di apprendimento supervisionato, e i risultati sono stati interessanti.

Già l'esplorazione delle variabili ha permesso di scoprire correlazioni tra indicatori apparentemente indipendenti, ed il machine learning ha individuato i pesi con cui essi intervengono nel caratterizzare le diverse marco-aree.

Così il clustering ha evidenziato dei nuovi gruppi con il massimo di caratteristiche condivise, l'analisi delle componenti principali ha permesso di ridurre la dimensionalità del problema verso un nuovo insieme di variabili più esplicative, ed infine la linear discriminant analysis ha trovato la combinazione di variabili migliore per descrivere la composizione dei tre gruppi già definiti.

Sono soddisfacenti i risultati di tutte e tre le tecniche, perchè non hanno portato ad affermazioni controintuitive, anzi hanno messo in luce vicinanze-lontananze tra i dati in modo più completo rispetto alle analisi sulle singole variabili. Inoltre la classificazione tramite apprendimento supervisionato ha mostrato un'accuratezza piuttosto alta.

In generale si deve tenere conto della scarsa dimensionalità del campione: i dati, le regioni, sono soltanto 20, presentano caratteristiche simili e c'è molta omogeneità tra gli attributi, senza variazioni (outlier) evidenti. Gli analoghi risultati su un campione più grande sarebbero sicuramente più apprezzabili.