Progetto di

Data Technology and Machine Learning

Pietro Mattia Campi, 794156
Marta Giltri, 795267

Parte I

Introduzione

L'obiettivo di questo progetto era cercare di clusterizzare i paesi del mondo tenendo conto del consumo di sostanze inquinanti quali combustibili fossili e carbone, della produzione di gas nocivi quali CO_2 e polveri sottili PM2.5, della produzione di energia tramite fonti rinnovabili e del prodotto interno lordo (PIL) del paese stesso.
Per questo motivo abbiamo selezionato molteplici dataset provenienti dalle banche dati più accessibili e adatte alla ricerca di elementi di nostro interesse, quali UNdata e WorldBank, da cui abbiamo potuto scaricare gratuitamente i seguenti dataset:

- UNdata:
    - Dataset sul consumo di carbone complessivo.
    - Dataset sull'utilizzo e sul consumo di combustibili fossili quali benzina e diesel.
- WorldBank:
    - Dataset sulla percentuale di energia rinnovabile utilizzata rispetto all'intera quantità di energia utilizzata.
    - Dataset sulla produzione complessiva della CO_2.
    - Dataset sulla popolazione registrata.
    - Dataset sulla presenza di polveri sottili PM2.5.
    - Dataset sul prodotto interno lordo (PIL).

Una volta selezionati questi dataset, si è proceduto ad integrare i dati provenienti da queste fonti così diverse al fine di ottenere un dataset finale contenente paesi a cui fossero assegnati valori validi per ognuna delle feature selezionate. Dato che, però, i dataset provenienti da UNdata non contenevano una chiave univoca per l'identificazione del paese, trovata nel codice ISO a tre cifre presente invece nei dataset di WorldBank, è stato necessario uno step ulteriore per poter consentire la migliore integrazione possibile dei dati raccolti, integrando inizialmente un nuovo dataset proveniente da un progetto su GitHub e contenente una lista a cui ogni paese era associato il relativo codice ISO con i due provenienti da UNdata.

Parte II

Data Technologies

Obiettivi

Come descritto brevemente nell'introduzione, dato l'obiettivo posto per il nostro progetto abbiamo deciso di ricercare i dataset relativi alle feature da esplorare all'interno delle banche dati pubbliche UNdata e WorldBank.

Selezione ed acquisizione delle sorgenti dati

L'acquisizione vera e propria dei dataset in questione è stata relativamente semplice, perché entrambe le banche dati consultate per la ricerca disponevano della possibilità di scaricare i propri dataset direttamente in formato .csv o .xlsx, e di conseguenza di poter accedere ai dati di nostro interesse senza ricorrere ad algoritmi di scraping e, soprattutto, in formati di facile trattazione per l'integrazione dei dati stessi.
Per questo motivo e per la compatibilità con Pentaho, software di data integration utilizzato in tutti i processi di pulizia e fusione dei dataset che poi andremo ad illustrare, abbiamo deciso quindi di lavorare esclusivamente con file .csv o .xlsx, senza andare a complicare ulteriormente il lavoro di integrazione richiesto da questo progetto utilizzando database completi o file di formati meno comuni e di più difficile trattamento.

Dimensioni di qualità su singoli dataset

Una volta acquisiti i dataset a noi più congeniali, ciò che abbiamo fatto inizialmente è stato analizzare tre diverse dimensioni di qualità per ognuno dei documenti scaricati, in modo da verificarne lo stato effettivo.

1. Prima dimensione: Completeness.

La prima dimensione di qualità analizzata è stata quella di completeness, ossia completezza, con cui abbiamo verificato con quale copertura il fenomeno osservato

in ogni dataset fosse rappresentato tramite le istanze presenti. Abbiamo deciso di verificare questa misura di qualità tenendo conto di tre diverse misure: completezza di tabella, completezza di tupla e completezza di attributi.

Completezza di tabella:

Dataset	Percentuale Completezza di Tabella
CO_2	76.87%
Carbone	92.58%
Carburanti	85.64%
Energia Rinnovabile	96.05%
Popolazione	94.58%
PIL	90.17%
PM2.5	48.92%
Elenco codici paesi	96.59%

Completezza di attributi:

Dataset CO_2

Nome attributo	Completezza dell'attributo
Series Name	1
Series Code	1
Country Name	1
Country Code	1
1990 [YR1990]	0.797235
2000 [YR2000]	0.917051
2008 [YR2008]	0.935484
2009 [YR2009]	0.935484
2010 [YR2010]	0.935484
2011 [YR2011]	0.935484

2012 [YR2012]	0.949309
2013 [YR2013]	0.949309
2014 [YR2014]	0.944700
2015 [YR2015]	0
2016 [YR2016]	0
2017 [YR2017]	0

Dataset Carbone

Nome attributo	Completezza dell'attributo
Country or Area	1
Commodity - Transaction	1
Year	1
Unit	1
Quantity	1
Quantity Footnotes	0.554808

Dataset Carburanti

Nome attributo	Completezza dell'attributo
Country or Area	1
Commodity - Transaction	1
Year	1
Unit	1
Quantity	1
Quantity Footnotes	0.13831

Dataset Energia Rinnovabile

Nome attributo	Completezza dell'attributo
Country Name	1

Country Code	1
Indicator Name	1
Indicator Code	1
2000	0.931818
2001	0.935606
2002	0.943182
2003	0.943182
2004	0.943182
2005	0.946970
2006	0.946970
2007	0.954545
2008	0.954545
2009	0.954545
2010	0.954545
2011	0.954545
2012	0.962121
2013	0.962121
2014	0.962121

Dataset Popolazione

Nome attributo	Completezza dell'attributo
Country Name	1
Country Code	1
2000	0.996212
2001	0.996212
2002	0.996212
2003	0.996212
2004	0.996212

2005	0.996212
2006	0.996212
2007	0.996212
2008	0.996212
2009	0.996212
2010	0.996212
2011	0.996212
2012	0.992424
2013	0.992424
2014	0.992424
2015	0.992424
2016	0.992424

Dataset PIL

Nome attributo	Completezza dell'attributo
Country Name	1
Country Code	1
2000	0.928030
2001	0.928030
2002	0.943182
2003	0.943182
2004	0.946970
2005	0.946970
2006	0.950758
2007	0.950758
2008	0.946970
2009	0.943182
2010	0.946970

2011	0.946970
2012	0.931818
2013	0.935606
2014	0.928030
2015	0.924242
2016	0.893939

Dataset PM2.5

Nome attributo	Completezza dell'attributo
Country Name	1
Country Code	1
2000	0.909091
2001	0
2002	0
2003	0
2004	0
2005	0.909091
2006	0
2007	0
2008	0
2009	0
2010	0.909091
2011	0.909091
2012	0.909091
2013	0.909091
2014	0.909091

2015	0.909091
2016	0

Dataset Elenco codici paesi

Nome attributo	Completezza dell'attributo
official_name_en	1
ISO3166-1-Alpha-3	1
CLDR display name	1
FIFA	0.951807
IOC	0.907631
ITU	0.935743

Completezza di tupla:

Dataset	Numero di tuple complete
CO_2	0 su 217
Carbone	2308 su 4160
Carburanti	3194 su 23093
Energia Rinnovabile	250 su 264
Popolazione	262 su 264
PIL	236 su 264
PM2.5	0 su 264
Elenco codici paesi	226 su 249

2. Seconda dimensione: Currency.

La seconda dimensione che abbiamo deciso di analizzare è stata quella di currency, una dimensione di qualità che permette di definire se i dati presenti all'interno dei dataset siano effettivamente aggiornati oppure obsoleti. Verificando le misure di

aggiornamento applicate sia da UNdata che da WorldBank, abbiamo potuto verificare che i loro dataset vengono aggiornati annualmente, per la maggiore, e per tutti i file che abbiamo considerato l'ultimo update risale a gennaio 2018. Questo sembra contraddire il fatto che i dataset non presentino dati riguardanti gli anni passati più recenti, ma questo potrebbe essere dovuto al fatto che questi dati non siano stati ancora resi pubblici e che quindi le due banche dati non abbiano potuto inserirli nei propri dataset, riuscendo quindi ad aggiornare solamente i valori più vecchi. Per questo motivo possiamo ritenere soddisfacente la currency dei dati da noi utilizzati.

3. Terza dimensione: Pertinence.

La terza ed ultima dimensione di qualità che abbiamo considerato è la pertinence, o pertinenza, dimensione che definisce quanti attributi inutili al nostro scopo contengono i dataset che abbiamo deciso di considerare. Volendo utilizzare per ognuno dei file le colonne contenenti il nome dei paesi, il codice ISO associato e i valori complessivi di consumo o produzione relativi ad un anno preciso (2014), le colonne che ci sono state effettivamente utili e che sono state utilizzate nel processo di creazione del dataset finale destinato al machine learning sono le seguenti:

Nome Dataset	Colonne utili
CO_2	3 su 16
Carbone	4 su 6
Carburanti	4 su 6
Energia Rinnovabile	3 su 19
Popolazione	3 su 19
PIL	3 su 19
PM2.5	3 su 19
Elenco codici paesi	3 su 6

Come possiamo vedere, molte delle colonne dei vari dataset sono state scartate, ma questo non ha influito in alcun modo sulla completezza del nostro lavoro in quanto tutti gli attributi non considerati risultavano inutili ai fini della nostra analisi.

Processo di Data Integration

Terminata l'analisi riguardante le dimensioni di qualità sui singoli dataset presi in considerazione, siamo passati alla fase di integrazione dei dati al fine di ottenere, per la parte di Machine Learning, un dataset contenente per un buon numero di paesi del mondo i dati relativi al consumo di combustibili fossili e carbone, alla produzione di anidride carbonica e polveri sottili, alla produzione di energia rinnovabile, alla popolazione e al prodotto interno lordo.

Un primo passaggio fondamentale, prima di passare all'integrazione vera e propria, è stato quello di fornire ai dati provenienti da UNdata il codice a tre cifre ISO relativo ad ogni paese: non potendo basare l'integrazione dei dataset sui nomi dei paesi, scritti secondo una diversa convenzione a seconda della provenienza dei dati stessi, ed essendo i dataset provenienti da WorldBank provvisti di codice ISO di identificazione universalmente riconosciuto, abbiamo deciso di fornire questo identificativo specifico e non ambiguo anche ai dataset provenienti da UNdata utilizzando un elenco di paesi e relativo codice ISO prelevato da un progetto pubblico su GitHub.
In ambiente Python si è quindi proceduto inizialmente ad espandere eventuali sigle e abbreviazioni e ad eliminare restanti caratteri non alfanumerici presenti all'interno dei dataset UNdata in modo che questi non potessero interferire con i passaggi successivi, normalizzando di fatto i dati, per poi passare alla tokenizzazione dei nomi dei paesi sia di questi dataset che del dataset ISO-3, in modo da poter associare ad ogni paese un codice a seconda del risultato ottenuto all'applicazione sui token ricavati dell'indice di Jaccard.

L'indice di Jaccard, anche noto come "coefficiente di similarità di Jaccard", è una misura statistica utilizzata per confrontare la similarità e la diversità di insiemi campionari, ed è definito come la dimensione dell'intersezione fratto la dimensione dell'unione calcolate sugli insiemi campionari.

Dato che i nomi dei paesi presentano perlopiù elementi riordinati a seconda della convenzione specifica applicata dalla banca dati di provenienza, e quindi privi di semplici errori di battitura o altre eterogeneità di questo genere, applicare una funzione di edit distance per verificare la distanza dei nomi dei dataset UNdata da quelli dell'elenco ISO-3 e di conseguenza confrontare carattere per carattere i nomi stessi non ci avrebbe consentito di associare i codici come invece avevamo in mente.
La tokenizzazione dei nomi dei paesi ci ha fornito invece, per ogni entry dei dataset, un insieme contenente le singole parole che componevano il nome del paese, e confrontare questi insiemi di parole fra di loro senza tenere conto del loro ordine di

apparizione ci ha consentito di collegare con successo i paesi dei dataset a prescindere dalle convenzioni poste dalle banche dati, e quindi di poter associare ad ogni entry nei dataset UNdata il rispettivo codice ISO.

Una volta terminato questo passaggio di pulizia e aggiunta di codice ISO, tutti i dataset di nostro interesse per la finale analisi di machine learning risultavano idonei al processo vero e proprio di integrazione, che abbiamo deciso di svolgere utilizzando il software di data integration Pentaho.

Quello che inizialmente abbiamo fatto è stato importare tutti i .csv di nostro interesse all'interno di un nuovo workflow, utilizzando anche quello relativo all'elenco di codici ISO-3: questo perché, inizialmente, avevamo in mente di raccogliere dati su più nazioni possibile, includendo quindi nella tabella finale non ancora definitiva tutti i paesi in cui non fosse null almeno una delle feature da noi considerate.
Data la natura dell'obiettivo finale del progetto, abbiamo poi pensato di prendere in esame i dati relativi ad un anno solo di quelli contenuti all'interno dei vari dataset, e la nostra scelta è ricaduta sul 2014 nel momento in cui abbiamo constatato che la maggior parte dei dataset caricati su UNdata e WorldBank registrano istanze datate solamente fino a quell'anno, o al più fino a due anni più tardi ma con istanze sempre meno complete.
Si è quindi proceduto a selezionare, tramite delle funzioni di Select Values e di Select Rown, le colonne e le righe dei dataset relative all'anno 2014 e che, in generale, risultavano riportare il consumo o la produzione complessivi dell'oggetto in analisi durante l'intero arco di quell'anno.

Una volta terminato il lavoro di selezione, abbiamo preparato tutti i vari dataset per le successive procedure di join assicurandoci che fossero tutti ordinati in ordine lessicografico sulla base della colonna contenente il codice ISO-3, che avevamo scelto come chiave di riferimento per poter integrare i dataset nella maniera più completa possibile.

Concluso anche il riordinamento delle istanze, si è proceduto ad effettuare per i dataset contenenti i dati su consumo di carbone, consumo di carburante e produzione di CO_2 un left-outer join con, come elemento posto a sinistra nella selezione, l'elenco di paesi con codice ISO-3: questo passaggio ci ha consentito di ottenere, per tutta la lista di paesi inclusa nell'elenco sopracitato, il valore relativo alla feature del secondo dataset nel 2014, che fosse null oppure no. Da questa prima procedura di left-outer join sono stati esclusi i dataset relativi a energia rinnovabile, popolazione, PIL e polveri sottili, in quanto più estesi dei dataset di UNdata e quindi più completi rispetto alla presenza di dati su più paesi del mondo.

Successivamente a questo passaggio di join siamo quindi passati, al fine di integrare effettivamente tutti i dati in un'unica tabella utile in una maniera sufficientemente pulita, a selezionare dai tre dataset sopra considerati le colonne di nostro interesse eliminando quelle duplicate dall'associazione con l'elenco ISO-3, quali nome del paese e relativo codice, sempre tramite funzioni Select Values.
In questo modo abbiamo quindi estratto tre tabelle più leggere che, tramite la funzione Multiway Merge Join, si è poi proceduto ad integrare con i rimanenti quattro dataset, utilizzando ancora come chiave di riferimento il codice paese. Questo ci ha consentito di ottenere un insieme di dati molto ampio contenente, per ogni paese dell'elenco ISO-3, i dati relativi ad ogni feature da noi considerata in partenza, rifinita poi sempre da Select Values per eliminare ogni colonna duplicata e rinominare adeguatamente tutte le rimanenti in modo da meglio specificare il contenuto presentato.
Una volta terminato anche questo passaggio di integrazione quello che abbiamo fatto è stato, nell'ottica di dover poi eseguire apprendimento non supervisionato tramite algoritmo K-Means, estrarre dalla tabella ottenuta in precedenza tutte le righe in cui tutte le feature prese in considerazione fossero not null: questa decisione è stata presa poiché l'algoritmo di machine learning da utilizzare suddivide i dati passati in ingresso tenendo conto di tutti i valori dei loro attributi per calcolare le distanze, e incontrando un null si limita a ignorare l'istanza stessa in quanto, per questo dato in particolare, non può calcolare la distanza necessaria a causa della mancanza dei valori di uno o più attributi. Nonostante l'algoritmo ignori questi casi in autonomia, abbiamo deciso di bypassare il problema alla base escludendo le istanze non idonee dal dataset definitivo ben prima di passare all'esecuzione dell'apprendimento non supervisionato.

Al termine quindi dell'integrazione dei dataset di partenza ciò che abbiamo ottenuto è stato un nuovo dataset, ordinato per codice paese, contenente una lista di paesi con relativo consumo di carburanti fossili, consumo di carbone, produzione di CO_2, percentuale di energia rinnovabile prodotta rispetto all'intera produzione energetica, quantità di polveri sottili, PIL e popolazione, il tutto relativo allo specifico anno 2014.

Dimensioni di qualità sul dataset finale

Una volta finalizzato il dataset da utilizzare nella parte di apprendimento non supervisionato, si è quindi proceduto a misurare le dimensioni di qualità verificate per i singoli dataset iniziali sul prodotto finale dell'integrazione appena effettuata.

1. Prima dimensione: Completeness.

Per prima cosa verifichiamo la completezza di questo nuovo set di dati.

- *Completezza di tabella:* 100.00%, con 1377 elementi non nulli su 1377 totali.
- *Completezza di attributi:*

Nome attributo	Completezza dell'attributo
Country code	1
Country name	1
Population	1
CO_2 production (kilotonnes)	1
Charcoal consumption (kilotonnes)	1
Fuel oil consumption (kilotonnes)	1
Renewable energy consumption (percentage)	1
PM2.5 (micrograms)	1
GDP (current US dollars)	1

- *Completezza di tupla:* 153 su 153.

Come era prevedibile, considerando il modo in cui abbiamo costruito la nostra tabella finale, il dataset risultante dalla data integration che abbiamo applicato risulta perfettamente completo.

2. Seconda dimensione: Currency.

Per i motivi illustrati alla descrizione della currency per i singoli dataset, e considerando in particolare i dati relativi al 2014, possiamo dire che il nostro dataset risulta aggiornato.

3. Terza dimensione: Pertinency.

Sempre seguendo il nostro processo di integrazione, abbiamo avuto la premura di eliminare qualsiasi colonna non dovesse essere di nostro interesse prima di

finalizzare l'aspetto finale della nostra tabella. Di conseguenza, ai fini della nostra analisi di machine learning, tutte le 9 colonne del dataset risultano essere utili ed utilizzabili.

Analisi descrittive dei dati integrati

Da ultimo, abbiamo voluto effettuare la seguente analisi descrittiva sul dataset finale, per poter avere una panoramica su quali tipi di dati saremmo andati a considerare.

	Population	CO_2 prod. (kton)	Charcoal cons. (kton)	Fuel oil cons. (kton)	Renewable energy cons. (%)	PM2.5 (µg)	GDP (US $)
$\bar{x}$	4.39e+7	1.991e+5	437.158	6735.986	35.687	30.244	4.32e+11
σ	1.56e+8	9.599e+5	1056.316	22262.31	29.877	21.075	1.74e+12
min	7.28e+4	6.234e+1	0.029	0.985	0.0	3.398	1.78e+9
max	1.36e+9	1.029e+7	6411.0	190050.0	93.859	115.872	1.73e+13

Parte III

Machine Learning

Obiettivi

L'obiettivo della nostra analisi a cluster è quello di raggruppare i paesi mondiali secondo il loro impatto ecologico rispetto alla produzione dei gas serra, dell'inquinamento dell'aria, del tipo di materie prime utilizzate utilizzate per la produzione di energia elettrica (carbone, benzine-gasoli, energia rinnovabile), del loro prodotto interno lordo (GDP) e popolazione.
In particolare, abbiamo scelto come unità di normalizzazione la popolazione, in quanto riteniamo che sia più rappresentativo confrontare le emissioni e i consumi pro-capite piuttosto che assoluti.
Utilizzando le quantità assolute, infatti, si avrebbe una netta prevalenza dei paesi più avanzati rispetto a quelli in via di sviluppo o a quelli poveri, che non permetterebbe di cogliere le differenze significative nelle politiche energetiche e ambientali. Inoltre, paesi più popolosi inquinano di più, ma ciò non significa che inquinino maggiormente pro-capite.

La domanda a cui abbiamo cercato di rispondere con la nostra analisi è stata "Una persona che vive nel paese X quanto inquina rispetto a una che vive nel paese Y?"

Training set: creazione ed analisi esplorativa

Creazione

Il training set è stato creato utilizzando tutte le istanze complete ricavate dal processo di integrazione dei dati descritto nella parte relativa al Data Management. I dati sono privi di etichetta, cioè non vi è una classificazione a priori.
Prima di utilizzarli per l'induzione di un modello k-means, abbiamo riportato le unità di misura originali (kilotonnellate) in unità del Sistema Internazionale (chilogrammi), per poter successivamente dividere per la popolazione (in alcuni casi un numero

molto grande) ottenendo così valori distanti dallo zero, che avrebbero invece potuto originare instabilità numerica.

Il dataset, salvato in un foglio excel, contiene 153 istanze di paesi con i seguenti attributi:

- Country Code: String di 3 lettere contenente il codice ISO del paese
- Country Name: String
- Population: Integer
- CO2 production (kg): Integer
- Charcoal consumption (kg): Integer
- Fuel oil consumption (kg): Integer
- Renewable energy consumption (% del totale): Float
- PM2.5 (micrograms): Float
- GDP (current US $): Integer

Le distribuzioni dei dati originali è mostrata nelle Figura 1 alla pagina seguente, mentre la Figura 2 mostra la distribuzione della popolazione.

Le features sono molto sbilanciate verso il basso con l'eccezione del consumo di energie rinnovabili, che presenta una distribuzione più uniforme. In particolare il consumo di carburanti fossili presenta un solo data point che si discosta significativamente dallo zero, situato a circa 4.5e+8 (kg).

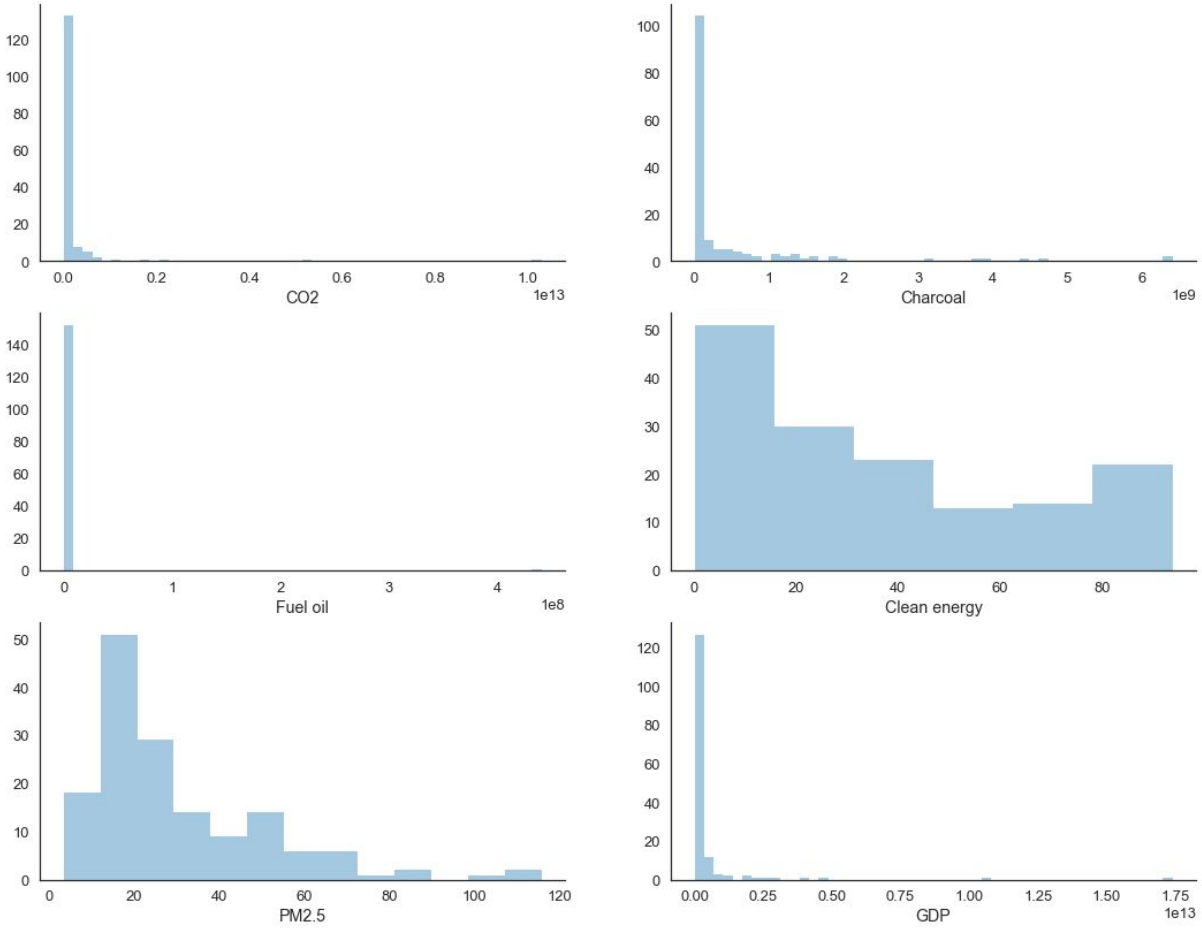


Figura 1: distribuzione dei dati originali

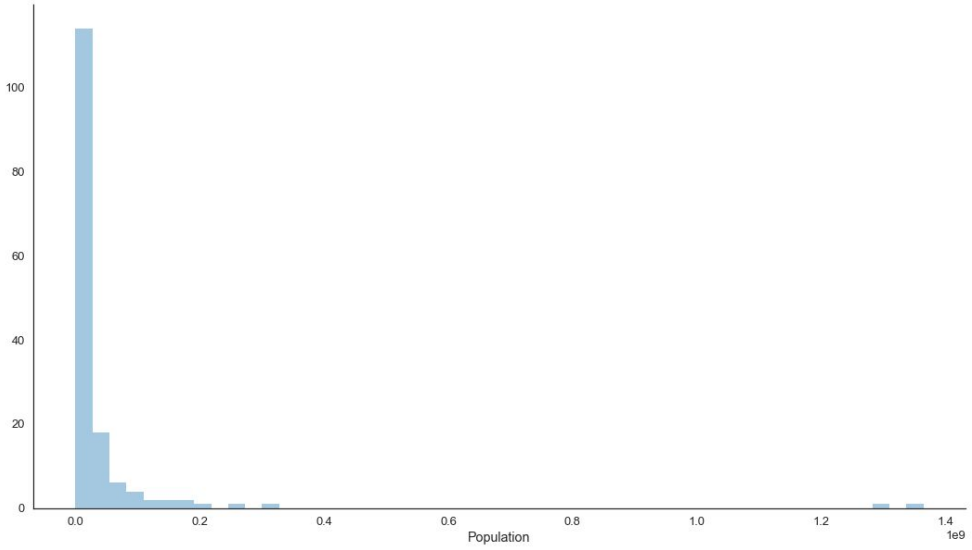
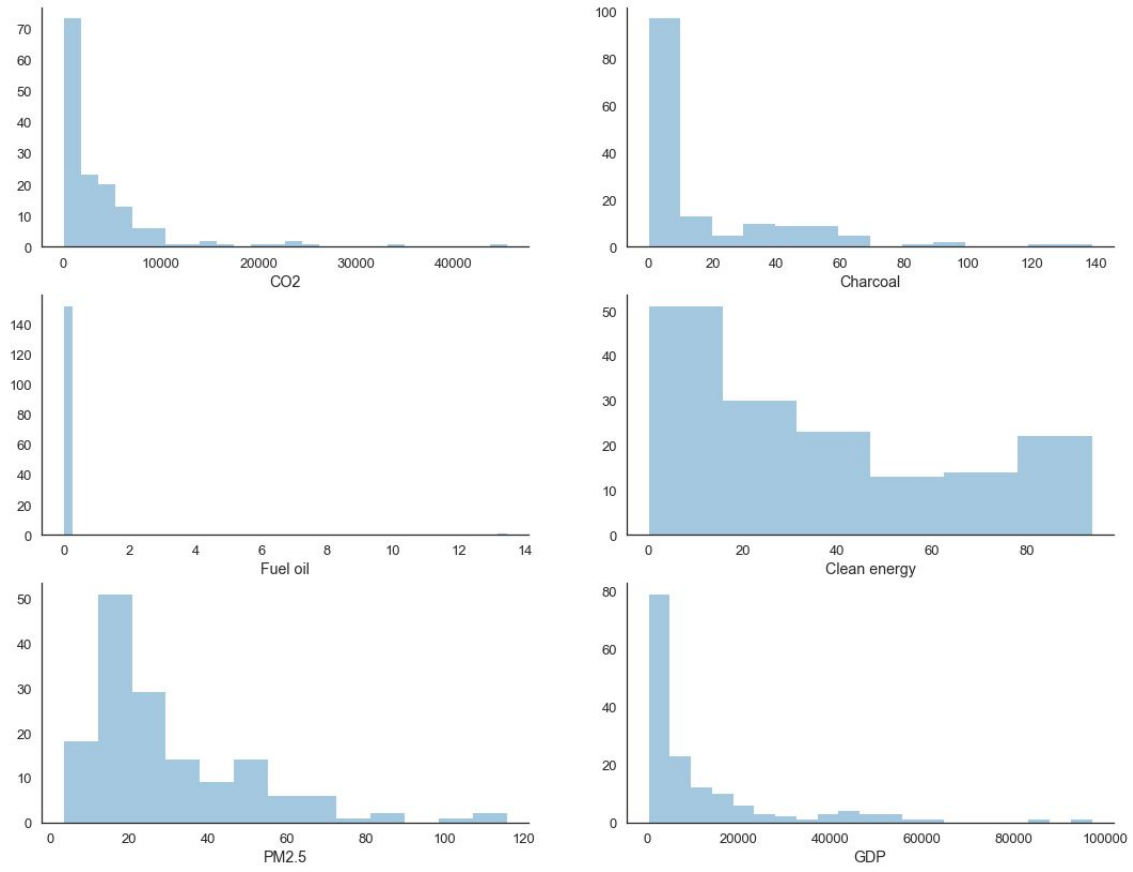


Figura 2: istogramma della popolazione

Per l'analisi si è scelto di normalizzare i consumi e le emissioni rispetto alla popolazione, per ottenere dati più rappresentativi dei reale pattern di inquinamento. Non sono state normalizzate le features relative al PM2.5, poiché erano già pro-capite, e alla percentuale di energia rinnovabile utilizzata.

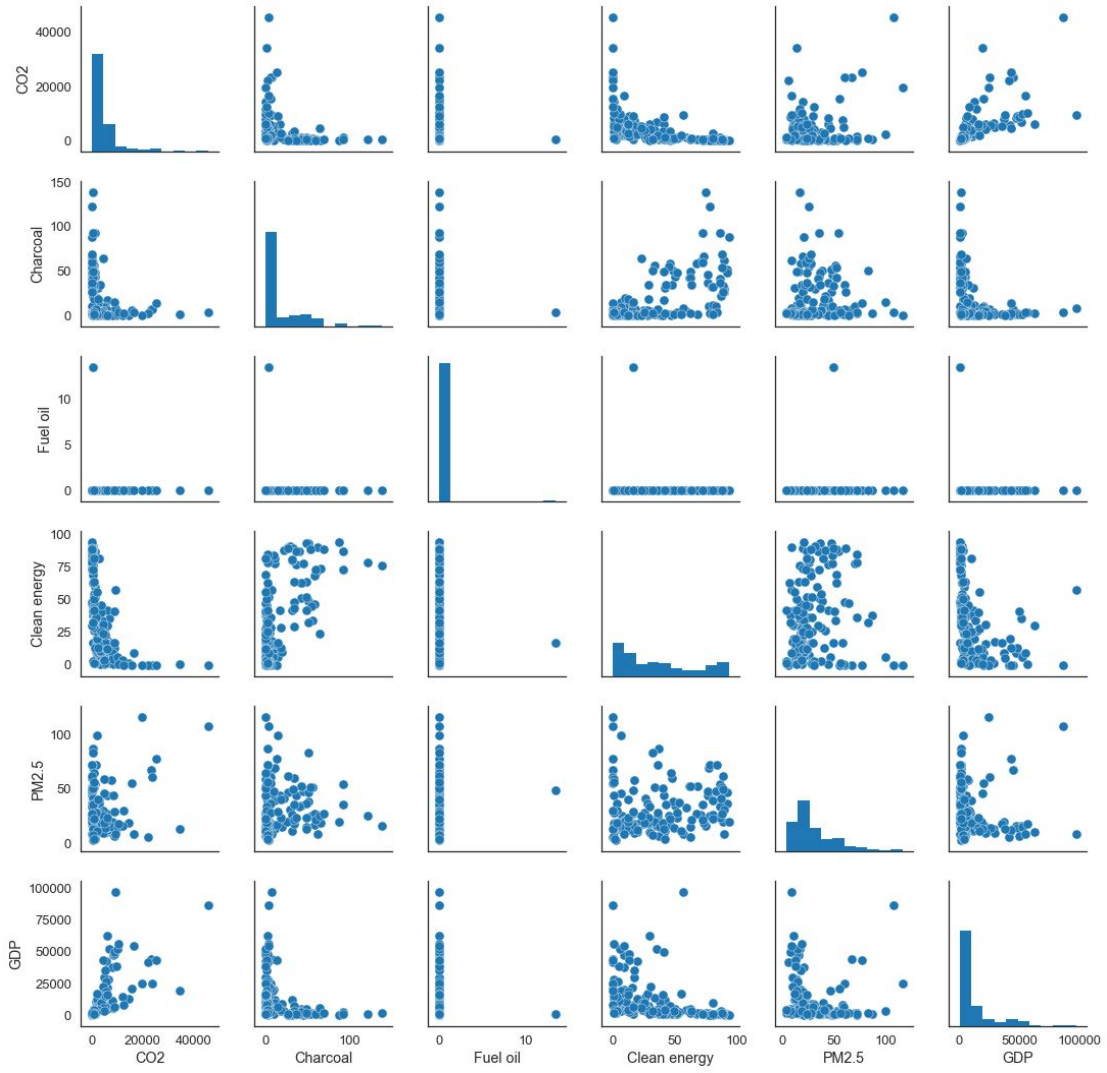
Di seguito è mostrata la distribuzione del dataset normalizzato rispetto alla popolazione.



Come si può vedere, vi è un apprezzabile cambiamento nelle distribuzioni del GDP, consumo di carbone ed emissioni di CO2. La feature "Fuel oil", invece, non mostra un significativo cambiamento. Per questo abbiamo cercato quale fosse il punto isolato più a destra visibile nel grafico dei carburanti fossili: è l'Afghanistan, dove ogni cittadino in un anno consuma circa 13.5kg di carburanti fossili.
Tutti gli altri paesi sono compresi in un intervallo [0 - 0.4].

Analisi esplorativa del training set

Prima di indurre un modello K-Means, abbiamo proceduto ad analizzare le distribuzioni accoppiate degli attributi: poiché ogni istanza appartiene ad R^6, una visualizzazione 3-D non era possibile.
La figura seguente mostra il grafico a coppie degli attributi. Avendo utilizzato la popolazione come attributo per normalizzare gli altri, abbiamo provveduto a toglierla dal dataset.

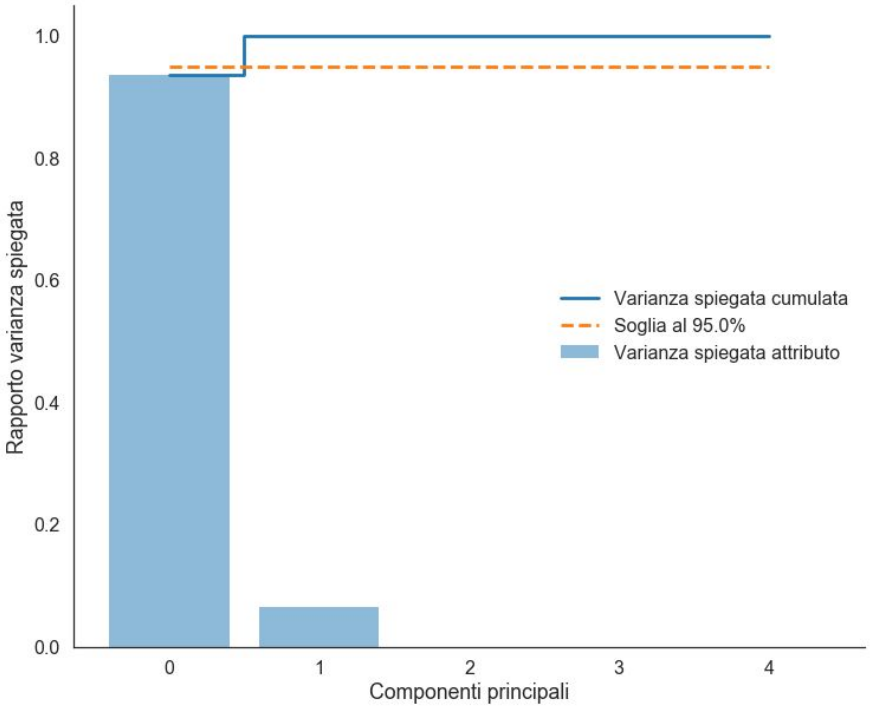


Di seguito, invece, le statistiche descrittive del dataset.

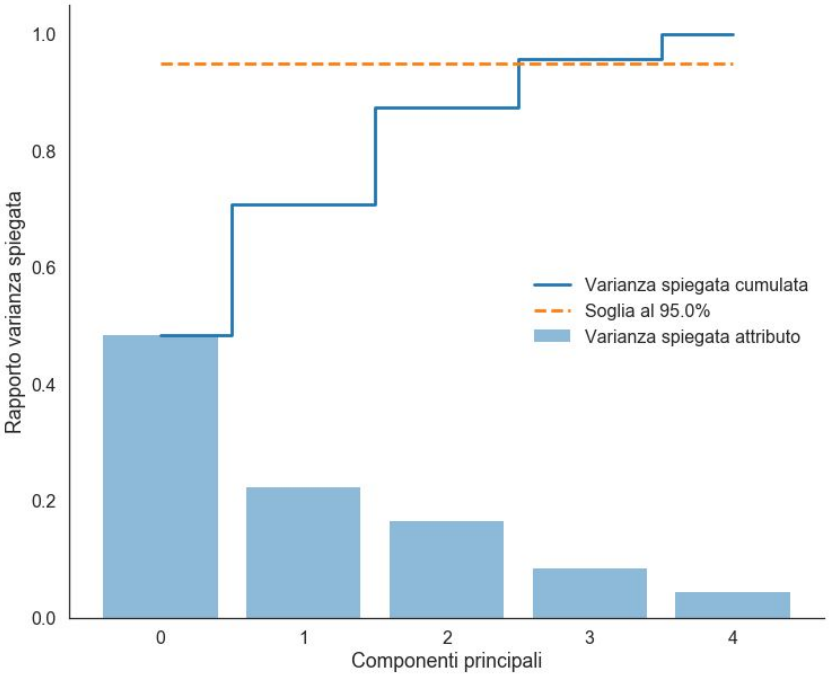
	CO2 (kg)	Charcoal (kg)	Fuel oil (kg)	Clean energy (%)	PM 2.5 micrograms	GDP $
$\bar{x}$	4190.1	17.17	0.088	35.7	30.24	11757
σ	6372.36	25.6	1.09	29.9	21.1	16910
min	44.48	0.009	5.18e-7	0.0	3.40	312.75
max	45423.24	138.78	13.45	93.86	115.87	97200

Gli attributi sono espressi in diverse unità di misura, su diverse scale e con statistiche molto diverse. Per verificare se ciò crea problemi per il clustering, abbiamo svolto una analisi dei componenti principali (PCA), la quale indica che una sola feature contribuisce per oltre il 93% della varianza spiegata. In questo caso il

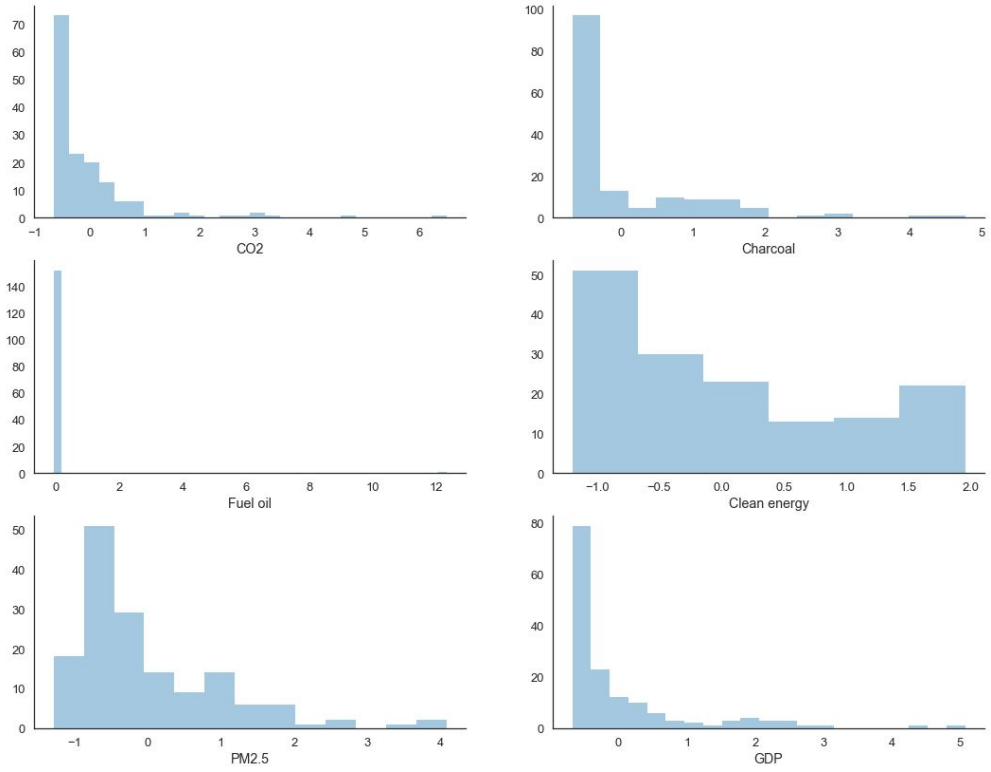
modello K-Means tenderà a separare i cluster nella direzione della feature con più varianza. Il grafico cumulativo della varianza spiegata ottenuto con l'analisi PCA sui dati non standardizzati mostra proprio questo:



Per ridimensionare l'impatto negativo di questo sbilanciamento, abbiamo standardizzato il dataset. Dopo la standardizzazione, l'analisi PCA indica una distribuzione della varianza più equilibrata.

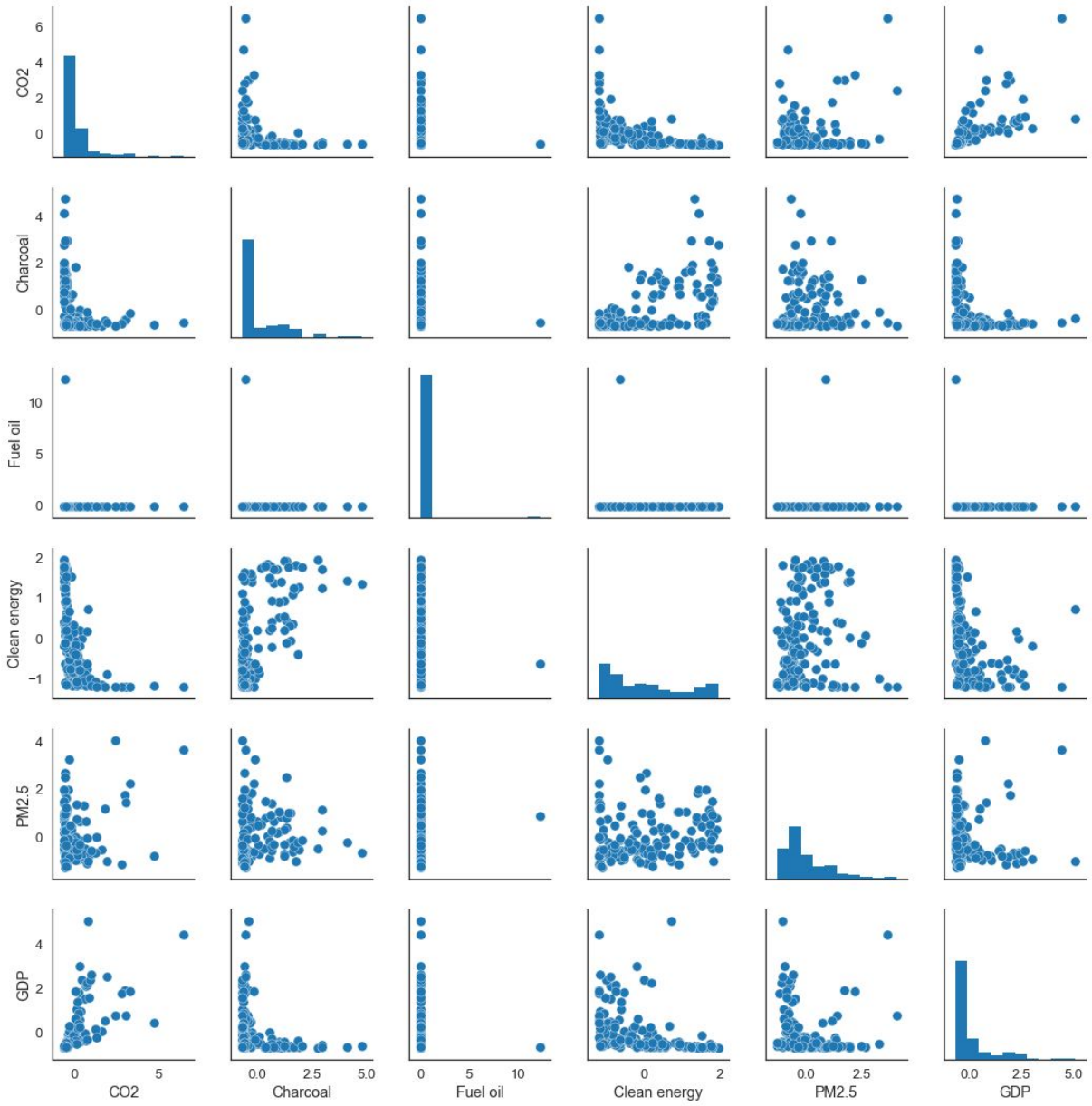


Di seguito sono riportati gli istogrammi relativi agli attributi standardizzati.

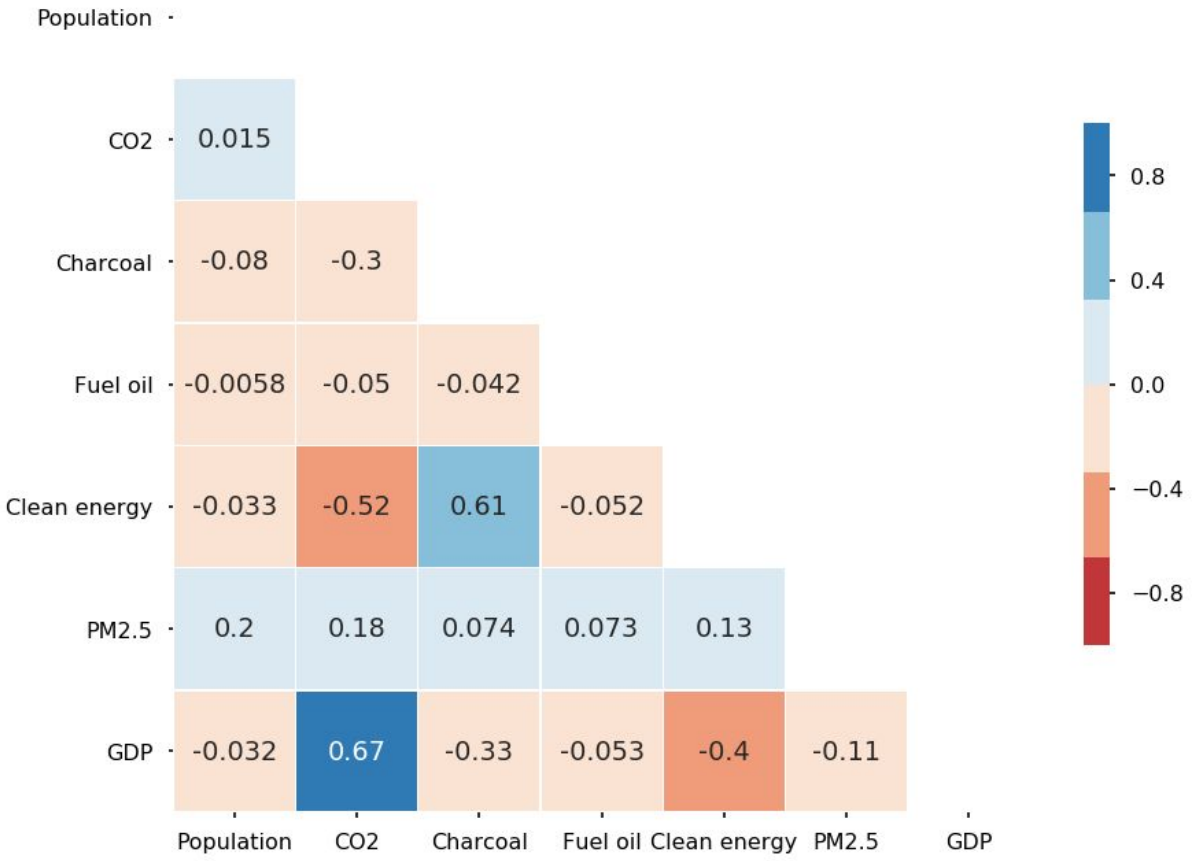


Ovviamente, la distribuzione non è cambiata, solo la scala.

Come si vede dal grafico, l'attributo relativo ai carburanti fossili continua a presentare il singolo punto spostato verso destra. Per verificare che questa feature sia effettivamente significativa, abbiamo utilizzato un plot accoppiato di tutte le variabili di interesse del dataset standardizzato e una analisi di correlazione riportata nella heatmap alle pagine seguenti.



Si vede chiaramente come l'attributo "Fuel oil" vari pochissimo rispetto agli altri, con l'eccezione del punto che rappresenta l'Afghanistan.



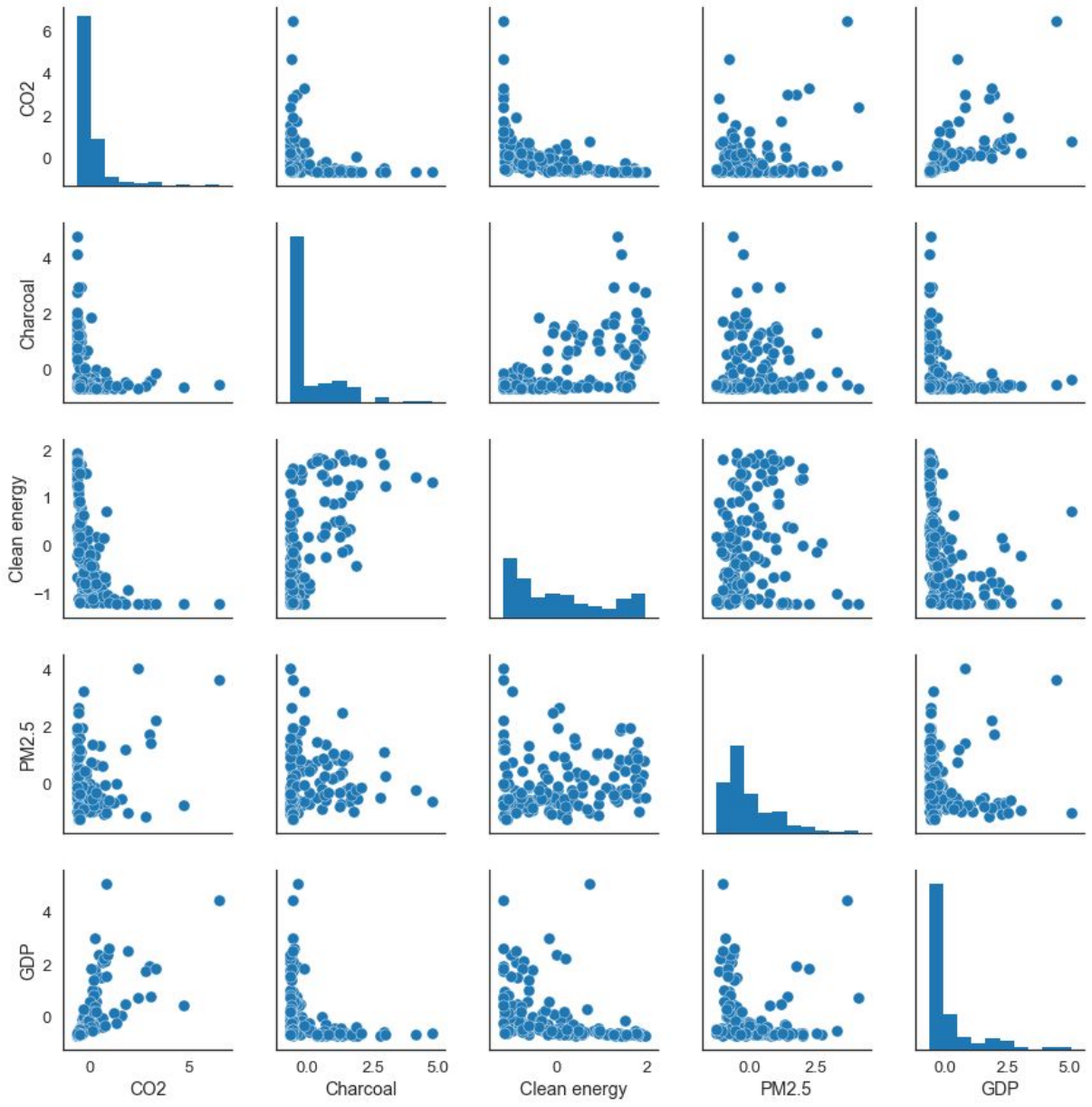
Come già suggerito dai grafici accoppiati, la heatmap mostra che non c'è correlazione tra l'attributo "Fuel oil" e gli altri. Tutti i suoi coefficienti di correlazione sono infatti molto bassi (≤ 0.05).

Si evidenzia invece una correlazione di 0.67 tra il prodotto interno lordo e le emissioni di CO_2 di un paese, così come una correlazione inversa di -0.52 tra il consumo di energia rinnovabile e la produzione di CO_2.
Questi due insight sono coerenti con quanto ci si potrebbe aspettare: un paese sviluppato consuma di più, quindi emette più gas serra, ed inoltre per le attuali politiche energetiche sono pochissimi i paesi sviluppati che utilizzano prevalentemente energia rinnovabile.

Sorprende, invece, che ci sia una correlazione di 0.61 tra il consumo di carbone e quello di energia pulita.

Avendo compreso che l'attributo "Fuel oil" non incide in modo significativo sulle altre variabili, si è deciso di scartarlo e di ritenere soltanto "CO_2", "Charcoal", "Clean energy", "PM 2.5" e "GDP".

La seguente figura riporta il grafico accoppiato degli attributi finali del dataset, normalizzati rispetto alla popolazione e standardizzati.



Clustering con modello K-Means

Il modello K-Means è un algoritmo di clustering partitivo ed esclusivo che permette di dividere N osservazioni in K gruppi, minimizzando la varianza totale intra-cluster. Formalmente, date N osservazioni $X = (x_1, x_2, ..., x_N)$ dove ogni $x_i \in R^n$, l'algoritmo partiziona X in k insiemi $S = \{S_1, S_2, ..., S_k\}$ al fine di minimizzare la quantità

$$\sum_{i=1}^{k} \sum_{x \in S_i} \left| x - \mu_i \right|^2$$

Poiché si tratta di un modello di apprendimento non supervisionato, utilizziamo l'indice di *Silhouette* per valutare la performance:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}$$

dove

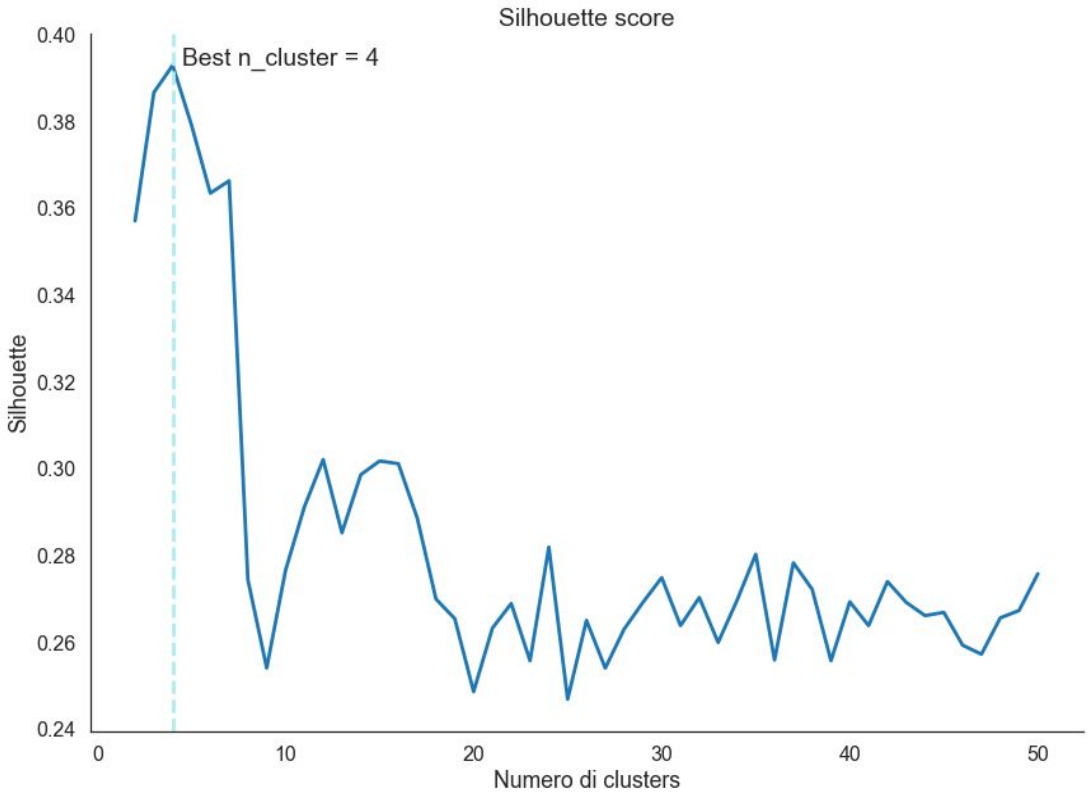
- x_i è l'i-esima osservazione
- $a(x_i)$ è la media delle distanze tra x_i e tutti gli altri x del suo cluster
- $b(i)$ è la più bassa distanza media di x_i dai punti degli altri cluster (la distanza, intuitivamente, dal cluster più vicino a cui non appartiene)

Stima del numero ottimale di cluster

Selezioneremo come numero ottimale di clusters k quello che massimizza l'indice di silhouette medio dei cluster, ovvero

$$k_{opt} = max_k \{ s_{avg} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} s(x_j) \}$$

Per trovare k_{opt} abbiamo quindi iterato tra 2 e 50 clusters, ottenendo il risultato riportato in Figura YYTT

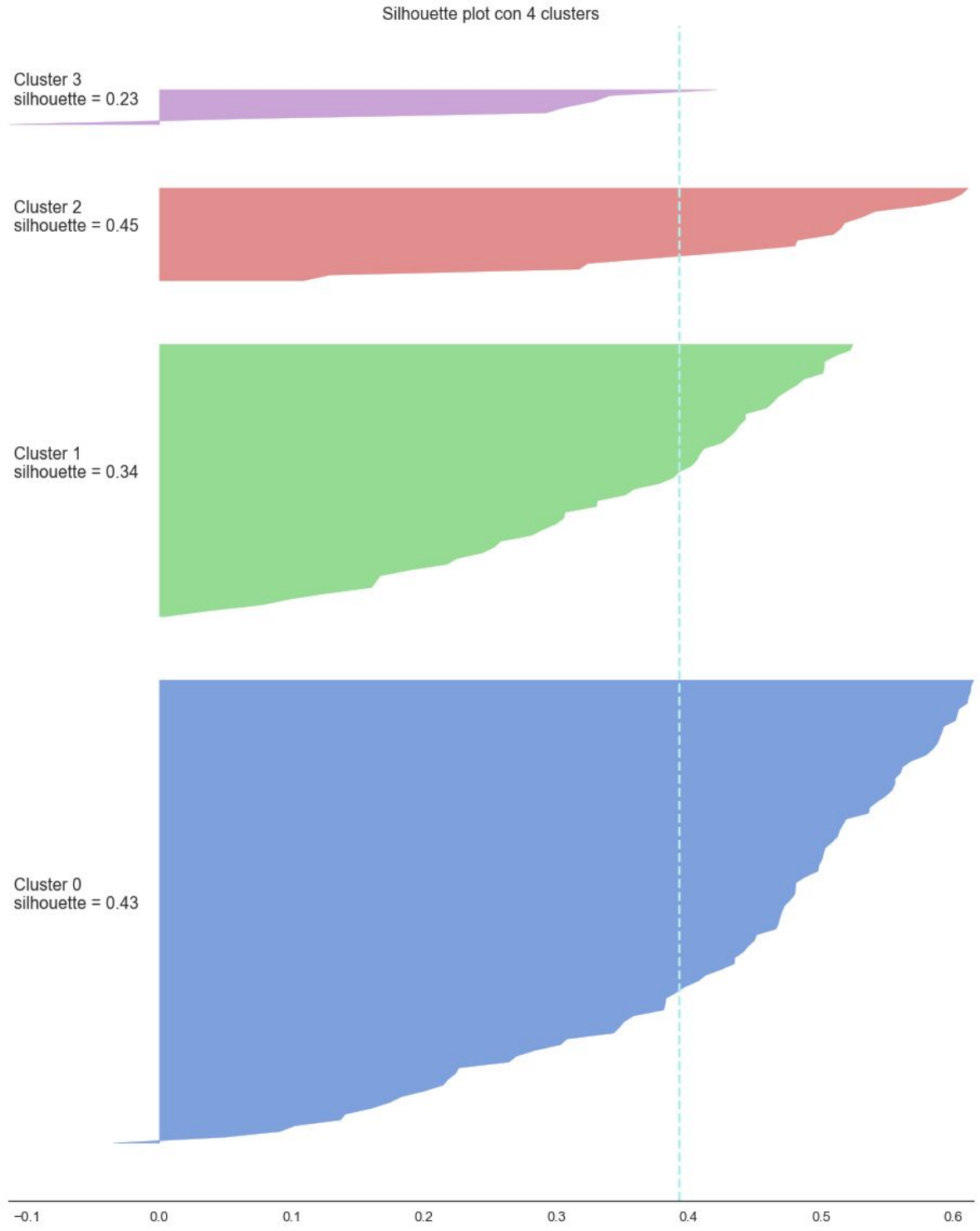


da cui si evince che il numero migliore di cluster sia $k_{opt} = 4$, per cui l'indice di silhouette media vale $s_{avg} = 0.3929$.
Tuttavia, la performance non è molto buona.

I risultati sono stati ottenuti utilizzando la funzione KMeans del package Python *sklearn*, utilizzando il metodo di inizializzazione dei centroidi *k-means++* (per aumentare la velocità di convergenza) e selezionando il risultato migliore di 10 trial randomizzati per ogni k.

Silhouette di ogni cluster con $k = k_{opt}$

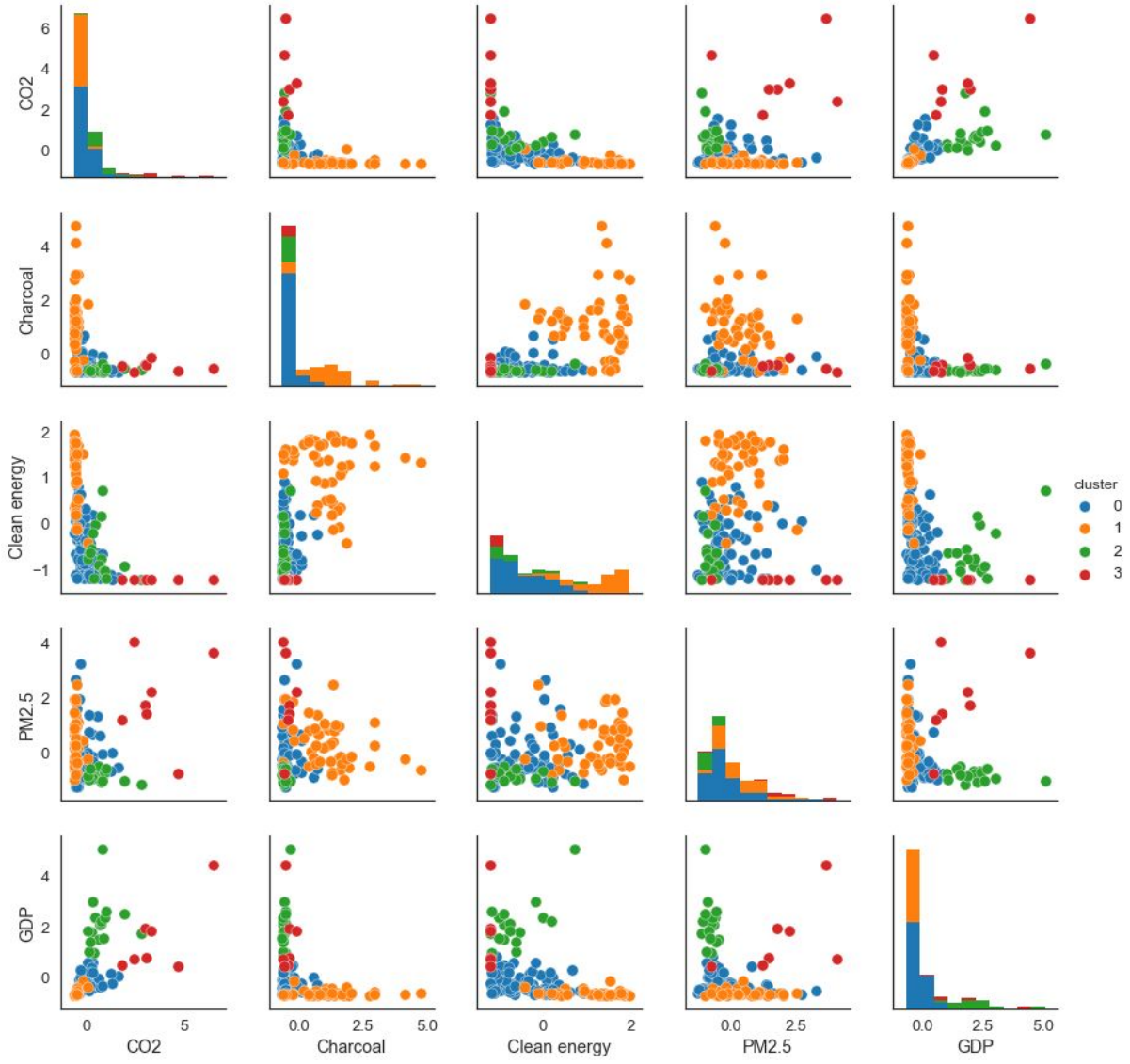
Rispetto alla soluzione di clustering ottimale, riportiamo nella figura alla pagina seguente il grafico degli indici di silhouette per ciascun cluster.



La seguente tabella mostra i valori di silhouette media per ogni cluster:

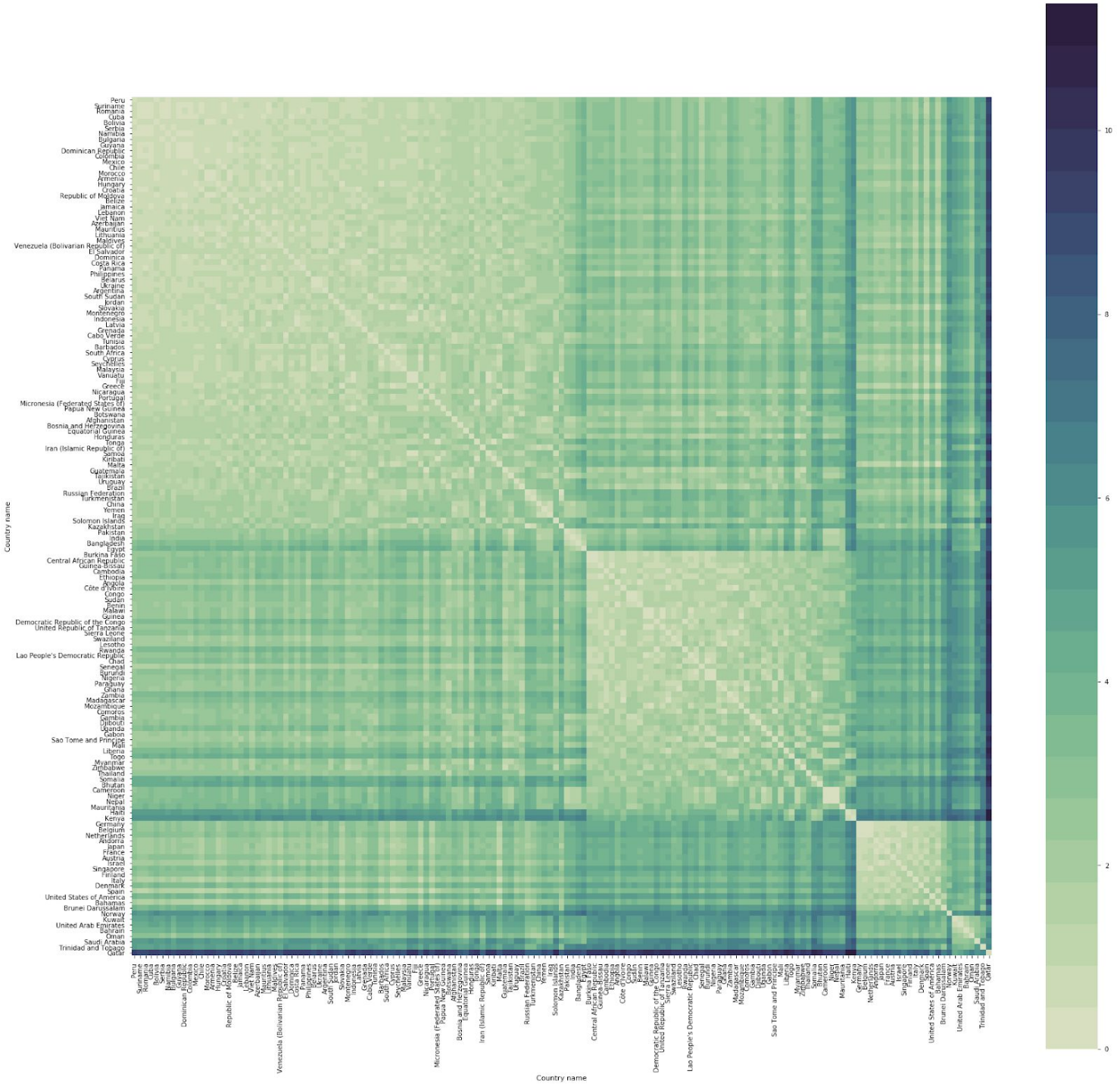
Cluster	0	1	2	3
Silhouette	0.43	0.34	0.45	0.23

La clusterizzazione non è molto buona, come suggerisce anche la presenza di elementi, al bordo inferiore di ogni cluster, che hanno indice di silhouette negativa.
Il plot accoppiato delle feature assieme alla clusterizzazione (i cluster sono divisi per colore) è mostrato di seguito.



Stima della matrice di dissimilarità con $k = k_{opt}$

Da ultimo, vediamo la stima della matrice di dissimilarità per la clusterizzazione ottimale con l'ausilio di una visualizzazione a heatmap.



Si può vedere dalla heatmap che i paesi sono divisi in 4 cluster rappresentati dalle zone quadrate in verde più chiaro: ciò è coerente con il fatto che la distanza tra elementi dello stesso cluster sia minore rispetto ad elementi appartenenti a cluster diversi.

Analisi dei risultati e conclusioni

La nostra analisi suggerisce la divisione dei paesi del mondo, raggruppati secondo le loro emissioni di CO_2, l'utilizzo di carburanti fossili e di energia pulita, nonché dei livelli di inquinamento medio, in quattro diversi gruppi.

Il primo è quello più ampio, costituito da 81 paesi la cui maggior parte sono di piccola-media grandezza e poco popolosi, come Cipro, le Barbados e l'Armenia. Tuttavia, esso include notevoli eccezioni come l'India, il Brasile, l'Argentina e la Cina. Nel primo gruppo figurano anche diversi paesi dell'Est Europa (Romania, Montenegro, Lituania) e diversi paesi africani (Namibia, Libano, Sudan).
Si potrebbe pensare che i paesi contenuti nel primo gruppo siano quelli in cui ogni abitante consuma relativamente di più rispetto agli altri.

Nel secondo gruppo, di 48 elementi, figurano principalmente i paesi poveri dell'Africa (Congo, Nigeria, Rwanda, Kenya) assieme ad alcune isole (Madagascar, Haiti). Si può presumere che questo gruppo contenga i paesi più poveri e popolosi, e per questo con un tasso di inquinamento pro-capite minore.

Nel terzo gruppo, composto di 17 paesi tra cui i più ricchi dell'Eurozona (Italia, Francia, Spagna, Olanda), sono condensati i paesi più sviluppati a livello mondiale. In particolare, troviamo i maggiori produttori ed utilizzatori di energia nucleare (Francia, USA e Giappone).

Nell'ultimo gruppo di 7 elementi si trova invece la maggior parte dei paesi produttori di petrolio (Emirati Arabi Uniti, Bahrain, Kuwait, Qatar, Arabia Saudita) e accreditati del minor consumo assoluto di energia pulita. Si nota che è incluso anche Trinidad e Tobago, che infatti risulta avere un coefficiente di silhouette negativo.

A conclusione di questo lavoro, possiamo affermare che visto il basso indice di silhouette per la soluzione ottimale (0.39), potrebbe essere utile inserire attributi aggiuntivi e più rappresentativi delle reali differenze nelle politiche energetiche dei vari paesi e del loro contributo all'inquinamento globale.
Inoltre, a causa dell'incompletezza di alcuni dataset sono stati scartati circa 70 paesi, i quali non figurano nell'analisi. Potrebbe essere interessante aggiungere anche i paesi mancanti, utilizzando sorgenti dati diverse ma con uguale autorevolezza (World Bank, UN).