Progetto Data Mining

Historical public debt data

Ivan Diliso
Matricola: 761053

Laurea magistrale in Data Science
Anno di corso 2021/2022

Indice

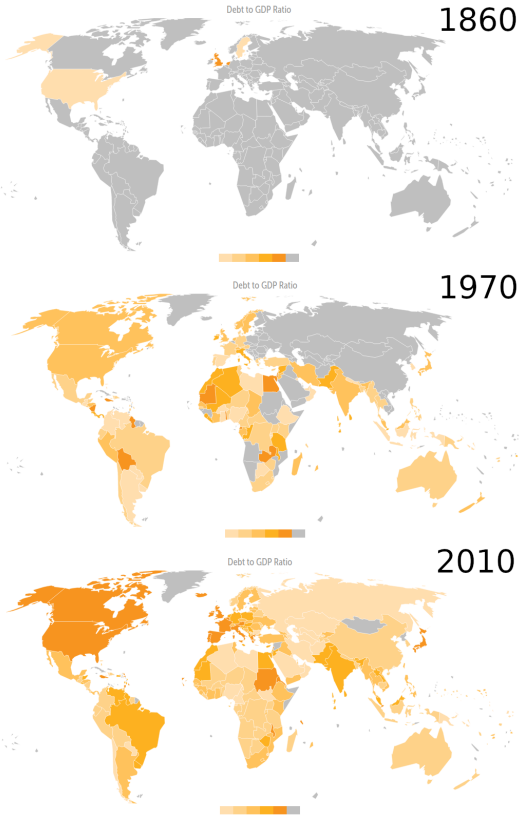
1 Data understanding

1.1 Descrizione del dataset

Il dataset utilizzato contiene i dati non bilanciati del prodotto interno lordo, debito lordo e rapporto lordo tra prodotto e debito per 187 paesi. La serie di dati per ogni paese va dal 1800 al 2020 tuttavia i dati di ciascun paese dipendono dalla loro data di indipendenza dalla reperibilità dei dati.

1.2 Analisi del dominio

Molti dei paesi presenti non esistevano fino a molto dopo il 1800, portando ad una grande quantità di dati mancanti, basti pensare alla serbia, montenegro, croazia, slovenia, bosnia, macedonia del nord (ex jugoslavia, 2003), tutti i paesi formatosi dopo la scissione dell'urss (1991) e i paesi africani nati dal colonialismo europeo. Queste informazioni hanno portato a pensare che il dataset fosse altamente sparso. Analizzando infatti il sito web del dataset è dispobile una visualizzazione del cambiamento del rateo del debito dei diversi paesi al variare del tempo e possiamo subito notare come nei primi anni solo pochissimi paesi sono avvalorati, e questo numero cresce solo all'aumentare degli anni.



1.3 Descrizione dei dati

Il dataset contiene i seguenti attributi:

1. Country name: Nome del paese (esempio "Italy")
2. Country code: Codice univoco del paese (esempio "103")
3. Indicator name: Nome dell'indicatore
    - Gross Domestic Product: Prodotto interno lordo espresso in dollari USD.
    - Gross Government Debt: Debito totale del paese espresso in dollari USD.
    - Debt to GDP ratio: Rapporto tra debito e prodotto interno lordo.
4. Indicator code: Codice dell'indicatore
    - NGPD: Gross Domestic Product
    - GGXWDG: Gross Government Debt
    - GGXWDG_GDP: Debt to GDP ratio
5. Attribute: Ha valore value per tutte le colonne
6. 1800 - 2020: Rappresentano l'informazione temporale della serire

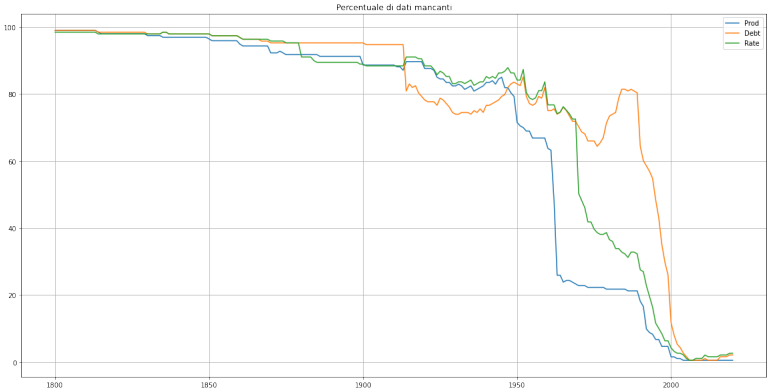
La serie temporale analizza l'andamento dei tre indicatori dal 1800 al 2020 per un totale di 226 colonne (5 per la descrizione degli attributi e 221 per le date) e 570 righe. Alcune righe presentano lo stesso paese ripetuto, le ripetizioni sono massimo 3 (una per ogni diverso indicatore del debito). Molti paesi non presentano tutti gli indicatori, il totale dei paesi osservati è 193, di questi 193 hanno informazioni sul loro prodotto interno lordo, 188 sul debito e 189 sul rateo.

1.4 Analisi statistica

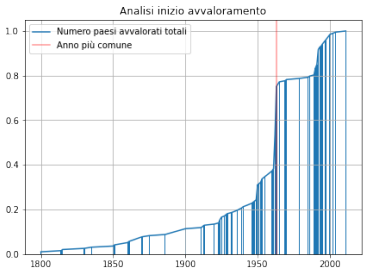
Viste le informazioni ritrovate nelle fasi precendenti la prima fase di analisi pone l'obiettivo di analizzare la sparsità del dataset, per poi procedere con una analisi della distribuzione dei valori per la ricerca di dati errati o outlier. Per la prima fase verrà ricercata informazione sull'anno ottimale per ridurre la lunghezza della serire temporale andando solo a modificare l'anno di inizio della serie, in modo da mantenere la continuità dei dati.

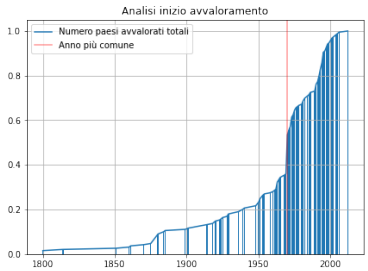
1.4.1 Analisi data primo avvaloramento

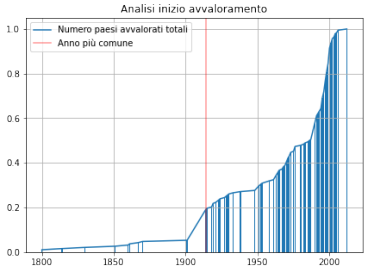
In questa prima fase di analisi si va ad analizzare la data di primo avvaloramento per ogni paese, questa analisi è motivata dal fatto che molti paesi potrebbero non essere nati prima dell'inizio del 1900 portando ad una grande quantità di dati non avvalorati. Le immagini mostrano la percentuale di dati avvalorati all'aumentare del tempo nell'ordine: prodotto interno lordo, rateo, debito.



Si nota infatti che fino al 1900 solo il 5% dei paesi ha almeno un valore nella serie temporale, questa percentuale sale al 30% attorno al 1950 e al 70% solo nel 1975. Dai grafici creati si scopre che l'anno più comune per il primo avvaloramento per il pil è il 1963, per il rateo il 1970 e il 1914 per il debito pubblico.







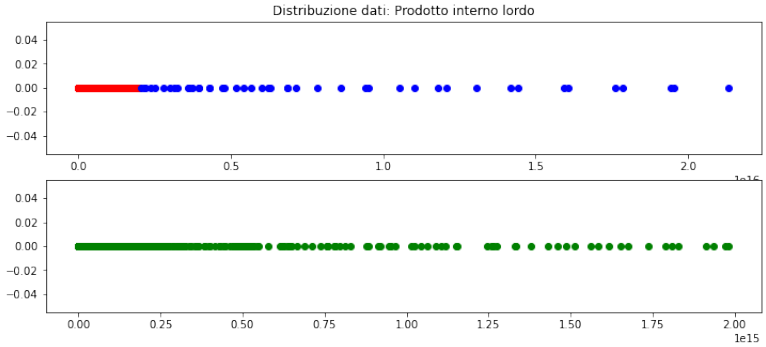
Prendendo il 1970 come anno di esempio andiamo ad analizzare come cambia la percentuale dei dati mancanti per ognuno degli indicatori.

	Percentuale del periodo temporale					
	Prodotto		Rateo		Debito	
Percentuale di dati mancanti	1800 - 2020	1970 - 2020	1800 - 2020	1963 - 2020	1800 - 2020	1970 - 2020
Minore del 25%	25.34%	**100%**	13.12%	**56.86%**	9.50%	**41.18%**
Tra il 25% e il 50%	1.36%	**0.00%**	9.50%	**41.18%**	2.26%	**9.80%**
Tra il 50% e il 75%	5.43%	**0.00%**	2.71%	**1.96%**	14.48%	**35.29%**
Superiore del 75%	67.87%	**0.00%**	74.66%	**0.00%**	73.76%	**13.73%**

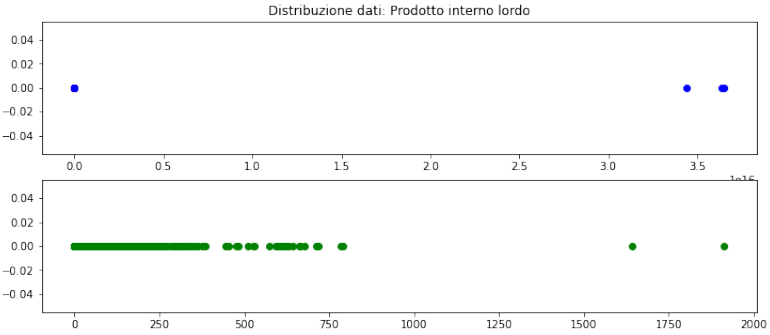
	Prodotto	Rateo	Debito
Media dei dati mancanti dal 1800 al 2020	69.64%	74.43%	78.56%
Media dei dati mancanti dal 1800 al 1970	86.83%	90.62%	89.92%
Media dei dati mancanti dal 1970 al 2020	10.64%	18.66%	38.92%

1.4.2 Analisi della distribuzione dei dati

L'analisi della distribuzione dei dati mira a trovare informazioni sui valori stessi dei diversi indicatori in modo da scovare la presenza di eventuali outlier o dati errati, questa fase viene inoltre utilizzata per la stima dei parametri di distanza minima per algoritmi come DBSCAN (parametro eps).



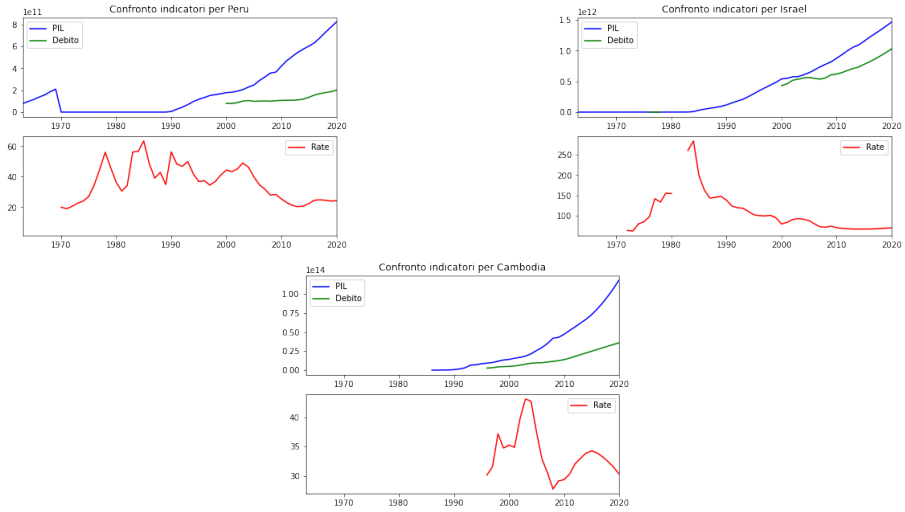
Per questo indicatore notiamo che il 99.65% dei dati si trova tra 0 e 2×10^{15}. La stessa analisi eseguita sul debito pubblico mostra risultati identici, è stata quindi omessa.



Il rateo è un indicatore percentuale, ci si aspetta dei valori nel range che va dal 0% fino ad un massimo del 2000% (che significherebbe per un paese un debito grande circa 2000 volte il prodotto interno lordo, valore possibile in periodi di forte crisi o default dello stato). Notiamo invece dei valori attorno al 3.5×10^{16} decisamente fuori scala per il rateo, questi vengono identificati come dati errati e successivamente andranno rimossi dal dataset. Come mostrato nel grafico sottostante a parte alcuni outlier il 99.944% dei dati si torva nel range $0\% - 150\%$.

1.5 Scelta degli indicatori da analizzare

Analizzando e confrontando le serie temporali del prodotto interno lordo e del debito si nota non solo una maggiore quantità di dati mancanti (come già dimostrato nelle sezioni precedenti) ma anche una similarità dell'andamento delle serie. Lo stesso non si rispecchia nell'analisi dell'andamento del rateo. Si è scelto quindi di procedere con l'applicazione di clustering e pattern mining solo sugli indicatori di rateo e prodotto interno lordo, escludendo il debito. Nelle immagini seguenti alcuni esempi delle serie temporali.



2 Data preparation

2.1 Selezione degli attributi

La colonna attribute ha valore value per tutte le tuple, viene quindi eliminata. Vengono prese in considerazione le colonne relative al nome di ogni paese, al codice identificativo del'indicatore del paese e dei dati degli anni dal 1800 al 2020, vengono quindi eliminate le colonne relative al codice del paese, al nome

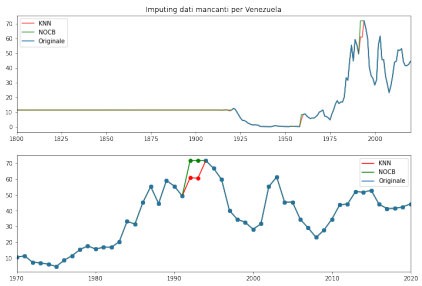
completo dell'indicatore e l'ultima colonna che risulta non avvalorata per tutte le tuple.

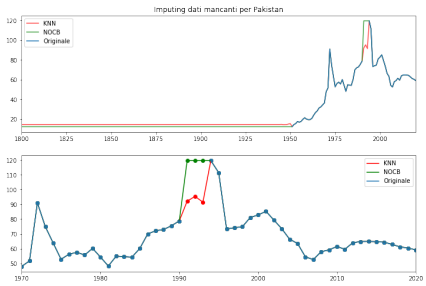
2.2 Rimozione dati mancanti

Sulla base delle infomazioni ritrovate nella fase di data understanding vengono effettuate due strategie per la rimozione dei dati mancanti (nelle successive fasi di clustering e pattern mining, si andranno ad applicare gli algoritmi non solo al dataset ridotto tramite la rimozione dei dati ma anche al dataset completo):

1. Viene utilizzato come anno di inizio della serie temporale l'anno di inizio avvaloramento più comune per l'indicatore analizzato, le serire saranno quindi:
    - 1963-2020: Prodotto interno lordo
    - 1970-2020: Rateo tra debito e prodotto
    - 1914-2020: Debito
2. Per avvalorare i restanti dati mancanti vengono utilizzate due tecniche:
    - K Nearest Neighbor (KNN): Il valore mancante viene sostituito con la media dei 5 valori più vicini nella serie temporale.
    - Next Observation Carried Backward (NOCB): Il valore mancante viene sostituito con il primo elemento avvalorato nella serie temporale.

Sono state scelte queste metodologie in modo da mantenere l'informazione sull'andamento della serie temporale, evitando di utilizzare un singolo valore fisso (come ad esempio la media) per ognuno dei valori null, a seguire alcuni esempi di applicazione dell'imputing tramite le due metodologie





2.3 Calcolo incremento annuale per pattern mining

La fase di pattern mining verrà effettuata sul dataset con righe e colonne invertite (anni come righe e i paesi come colonne) trasformando i dati da valori annuali a incremento annuale degli indicatori. Una volta calcolato l'incremento

annuale questo viene discretizzato in diversi range di valori, per poi applicare gli algoritmi di patter mining ai diversi indicatori binari. Per quanto riguarda i dati mancanti questi possono essere ignorati in quanto un valore nan in una riga può essere interpretato come un valore false, quindi come un paese che non è presente nel set di item della riga.

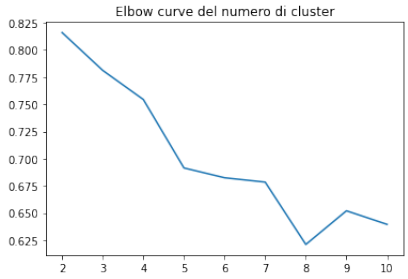
name	Guyana	Nicaragua	Liberia
2016	0.07	2.94	12.04
2017	0.66	1.66	5.25
2018	-0.37	1.90	0.60
2019	-1.46	1.43	-1.94
2020	-2.52	1.99	-3.99

name	Guyana	Nicaragua	Liberia
2016	True	True	True
2017	True	True	True
2018	False	True	True
2019	False	True	False
2020	False	True	False

3 Clustering

3.1 Scelta numero di cluster

Nelle sezioni a seguire per i modelli che richiedono di definire a priopri il numero di cluster è stata analizzata la curva formata dal variare della misura di valutazione al variare del parametro per la scelta del numero di cluster ottimale. Non è stato inserito questo grafico per ogni modello per non appensantire la presentazione. Nella foto di esempio è mostrata la elbow curve del modello KMeans applicato sul prodotto interno lordo nel periodo di tempo 1963-2020.



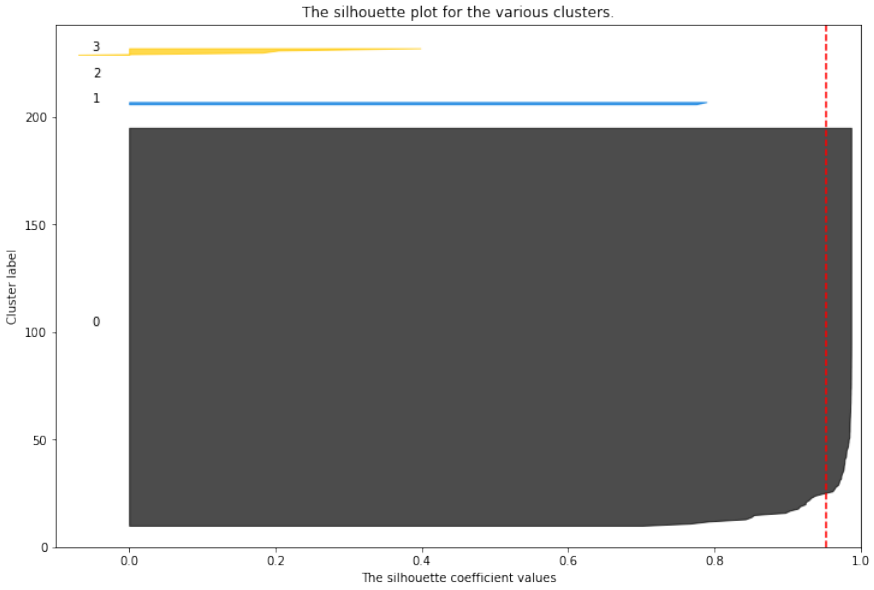
3.2 Approccio multidimensionale

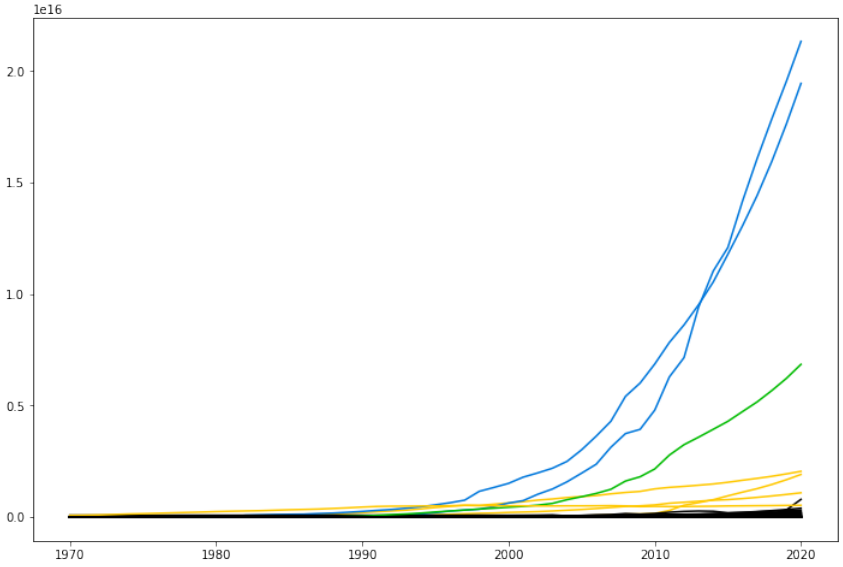
Il dataset viene trattato come un dataset multidimensionale, utilizzando ogni anno come feature a se stante. I modelli vengono addestrati sia sul dataset completo sia sul dataset con la rimozione dei dati null. Viene utilizzato il coefficente di Silhouette per la valutazione dei modelli.

3.2.1 KMeans

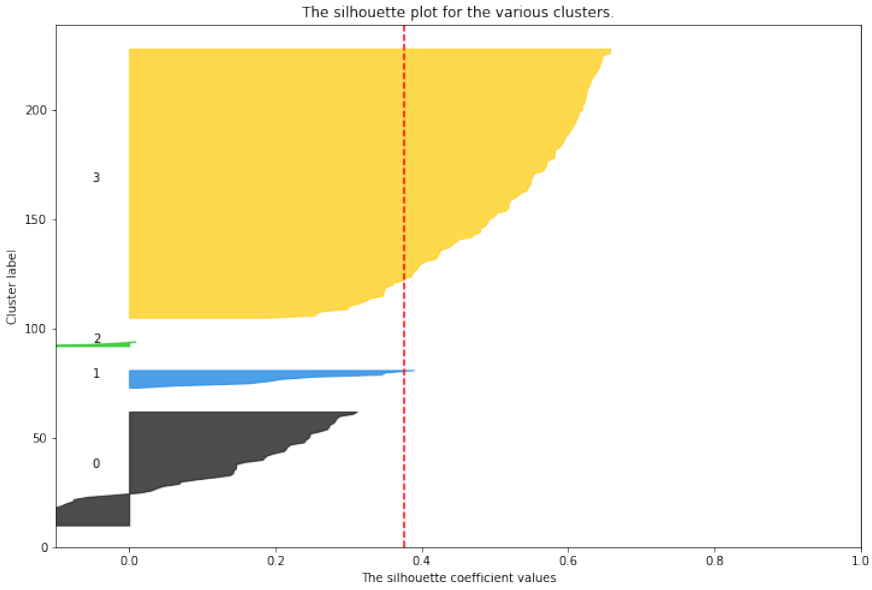
Nel primo caso viene applicato l'algoritmo KMeans alle serire dell'indicatore del PIL. In questo caso sia l'utilizzo della tecnica NOCB sia KNN non cambiano

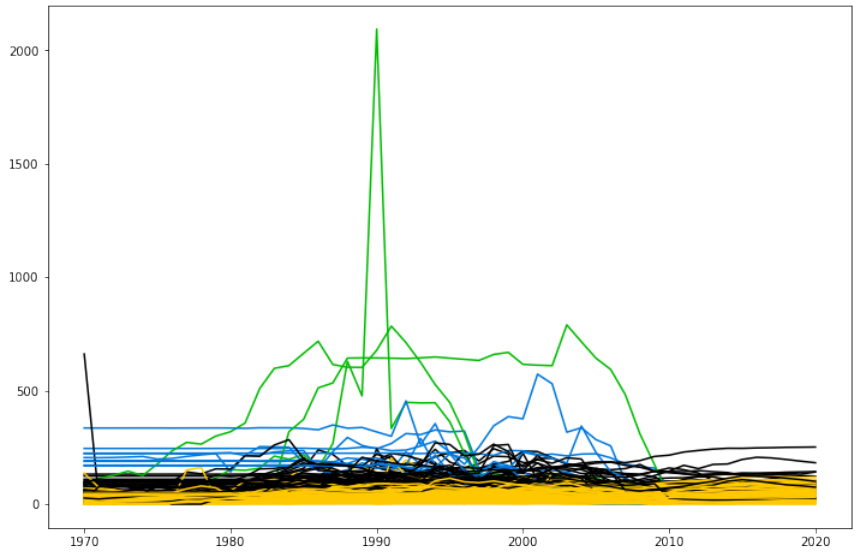
l'ouput del modello, così come la rimozione dei dati mancanti. Nelle immagini è mostrato il modello applicato al dataset con i dati null rimossi, k=4 e KNN come tecnica di imputing.





L'algoritmo scopre un unico cluster contenente circa il 98% dei dati, i risulati risultatno quindi non soddisfacenti. L'algoritmo presenta dei risultati più interessanti quando applicato all'indicatore del rateo.





In questo caso nonostante l'abbassamento del coefficente di silhouette medio si hanno dei cluster più interessanti che sembrano rispecchiare le forme comuni delle serie temporali. L'applicazione di questo stesso metodo al dataset completo e con l'utilizzo della tecnica NOCB ha portato ad un innalzamento del coefficente silhouette ma alla scoperta di un solo cluster unico anche per l'indicatore del rateo. Gli ulteriori test sono stati effettuati utilizzando diverse combinazioni dei metodi di imputing, numero di cluster, indicatore scelto per l'analisi e periodo di tempo considerato, per un totale di 16 test effettuati, molti dei risultati

ottenuti risultano simili tra loro, i risultati più interesasnti (mostrati nelle figure precedenti) sono i seguenti:

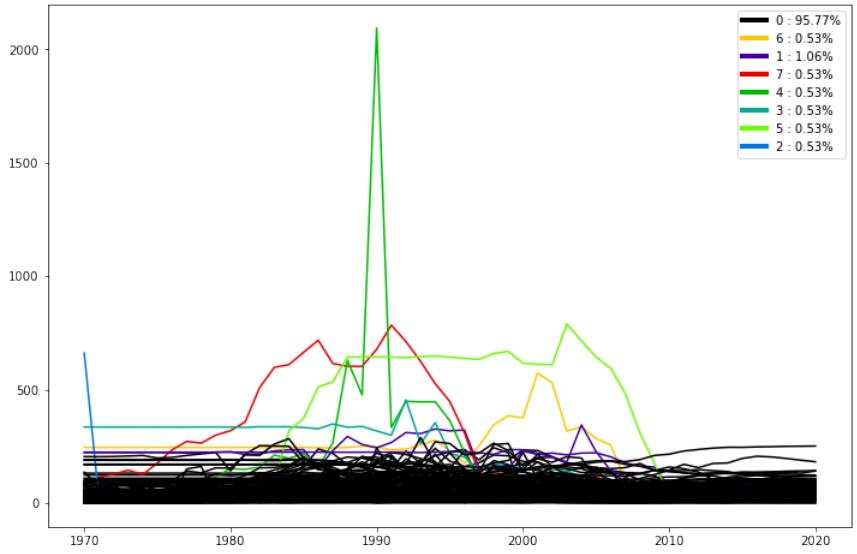
Modello e parametri	**Silhouette**
KMEANS, 1970-2020, k=4, KNN, PIL	*0.9518*
KMEANS, 1970-2020, k=4, KNN, RATEO	*0.3742*

3.2.2 Agglomerative

Nell'applicazione dell'algoritmo di clustering agglomerativo vengono considerati i seguenti parametri:

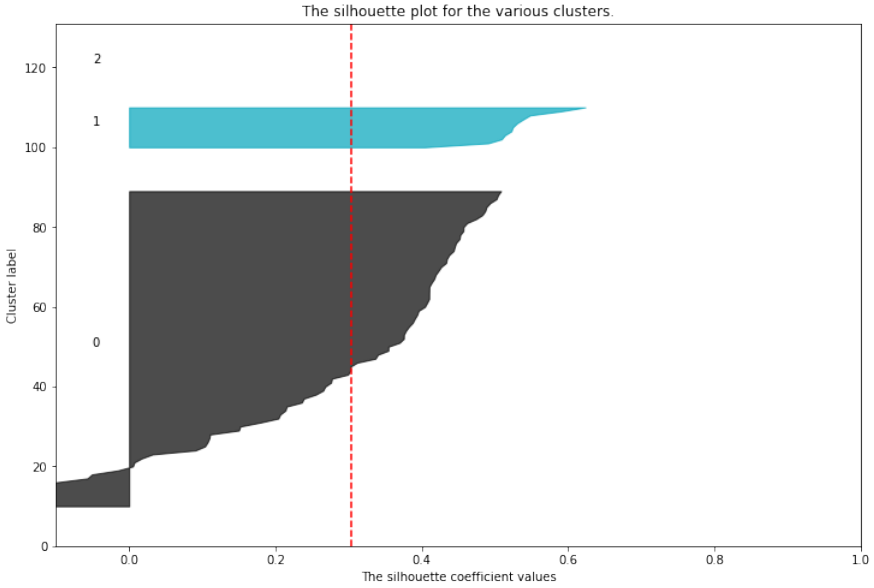
- Misura di distanza: euclidea, manhattan, coseno
- Linkage: single, complete, average, ward

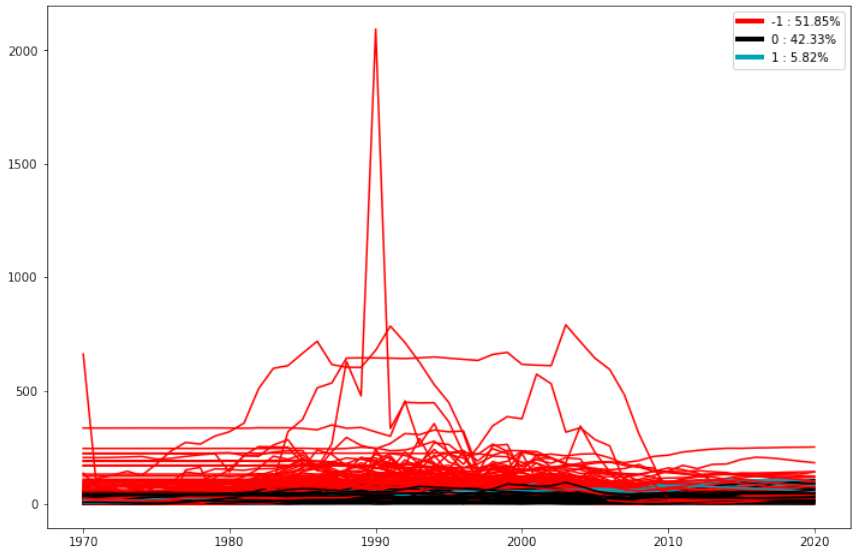
Nonostante il modello agglomerativo sia stato provato con tutte le possibili configurazioni non ha prodotto risultati interessanti, con cluster in linea con i risultati ottenuti con l'applicazione dell'algoritmo KMEANS. Quando applicato senza specificare il numero di cluster ma utilizzando una soglia di distanza per agglomerare i cluster questo ha portato alla creazione di molti cluster ognuno con il 1-2% dei dati al suo interno. Nella foto vengono mostrati i cluster creati dall'algoritmo agglomerativo applicato sull'indicatore del rateo con rimozione dei dati mancanti, con misura di distanza L2 e single linkage. Vengono trovati 8 cluster per un coefficente silhouette del 48%.



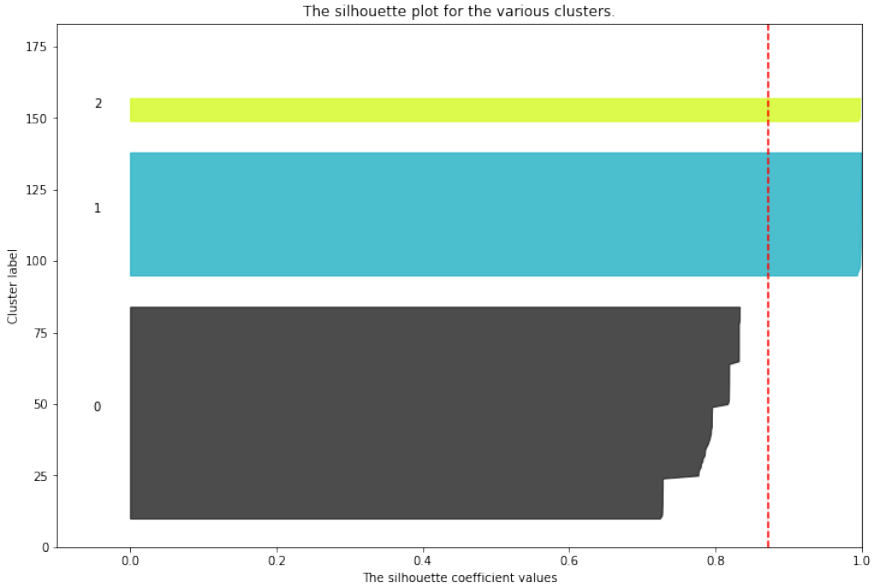
3.2.3 DBSCAN

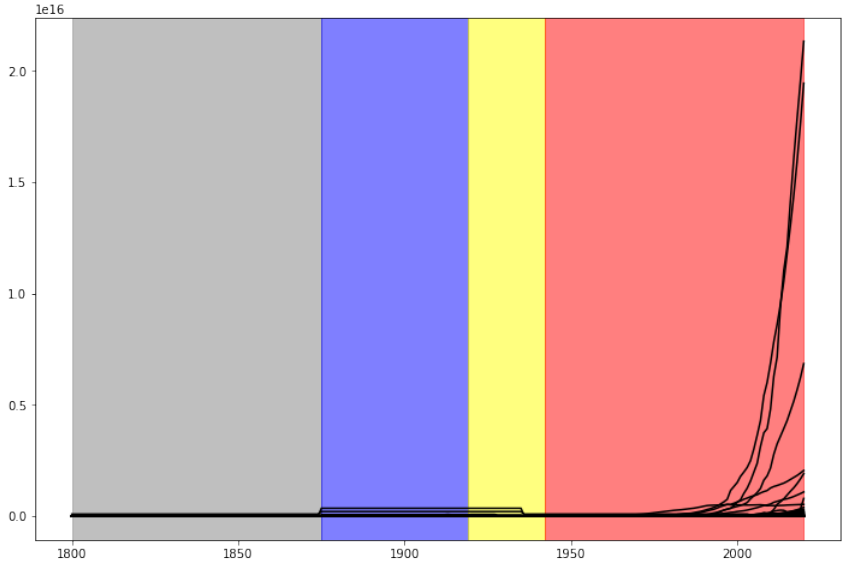
Quando applicato ai dati completi questo algoritmo non produce alcun cluster, tutti i dati vengono taggati come rumorosi a prescindere dalla scelta dei parametri eps e campione minimo. L'unica applicazione che ha prodotto risultati è stata applicata sul periodo di tempo 1970-2020, sull'indicatore RATEO, utilizzando eps=100, campione minimo 5 e imputing tramite KNN.





Possiamo notare che il 51% dei dati viene taggato come rumoroso e vengono trovati solo due cluster rispettivamente con il 42% dei dati e il 5% dei dati, per un coefficente di silhouette del 30%. Un risultato particolare si ottiene applicando DBSCAN sull'indicatore del prodotto interno lordo andando però ad invertire righe e colonne, applicando quindi il clustering sul tempo utilizzando i paesi come feature.

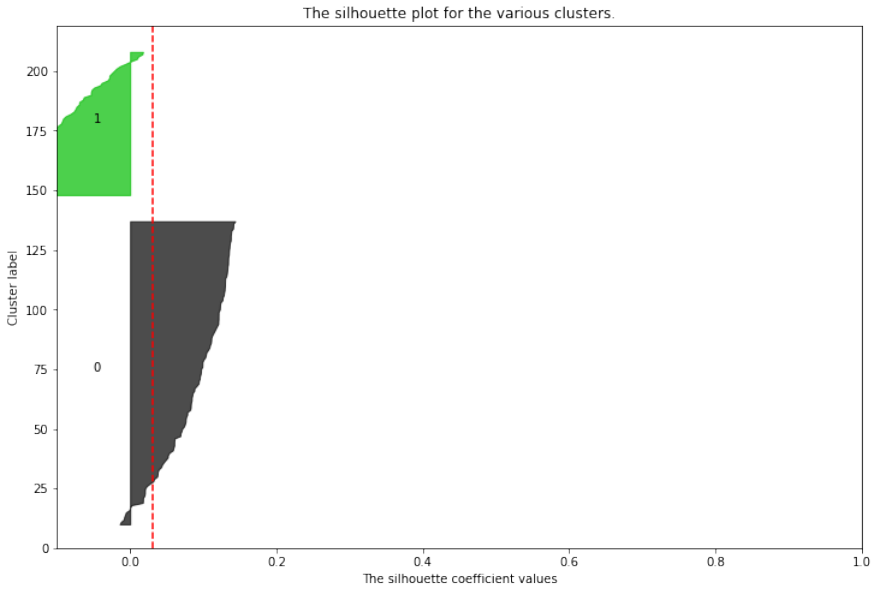


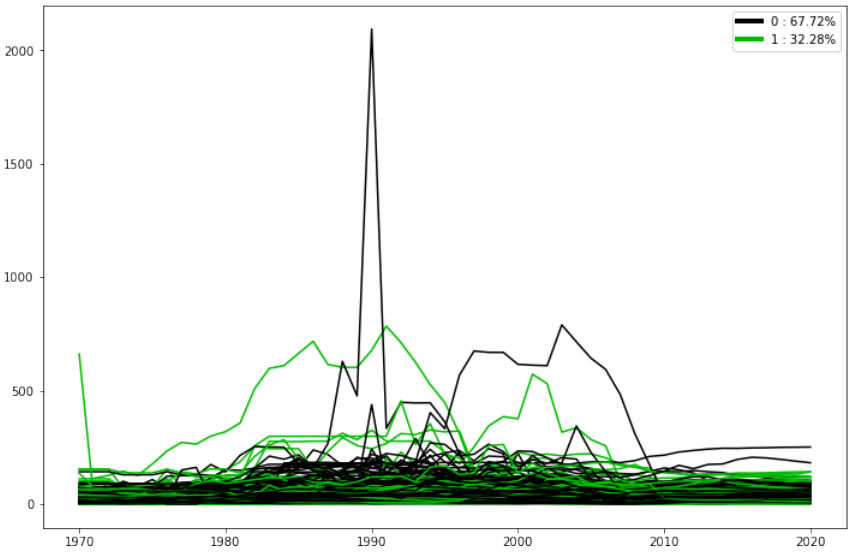


Una prima occhiata dei grafici potrebbe farci pensare di aver raggiunto un buon risultato, avendo trovato 3 cluster con dei buoni valori del coefficente di silhouette che arriva ad una media dell'87%. Andando però ad analizzare gli elementi taggati come outlier si nota che il modello ha utilizzato come dati per il clustering solo quelli creati artificialmente dalla KNN, valori che rimangono stabili da 1800 al 1960, andando invece a taggare come rumorosi i dati effettivamente avvalorati della serie temporale (zona evidenziata in rosso).

3.3 Approccio multidimensionale tramite wavelet transform

In questa fase vengono applicati gli algoritmi KMEANS e Agglomerative sulla serie temporale trasformata in un array di coefficenti tramite la haar wavelet transform. Anche in questo caso l'algoritmo non ha prodotto risultati soddisfacenti con l'algoritmo agglomerative che non supera il 3% del coefficente medio di silhouette al variare di tutti i suoi parametri e con l'algoritmo KMEANS che produce risultati peggiori dell'algoritmo applicato al dataset originale, trovando un unico cluster principale che contiene il 99.50% dei dati.

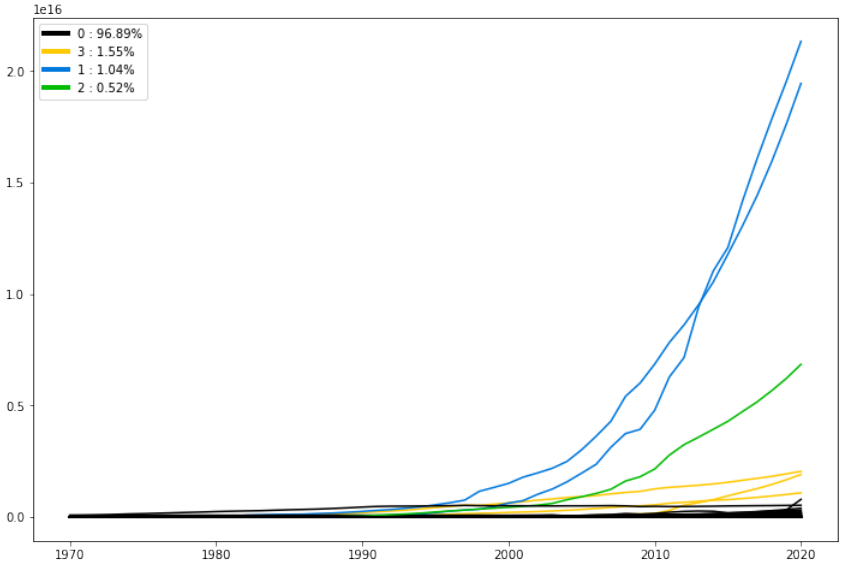




3.4 Approccio tramite dynamic time warping

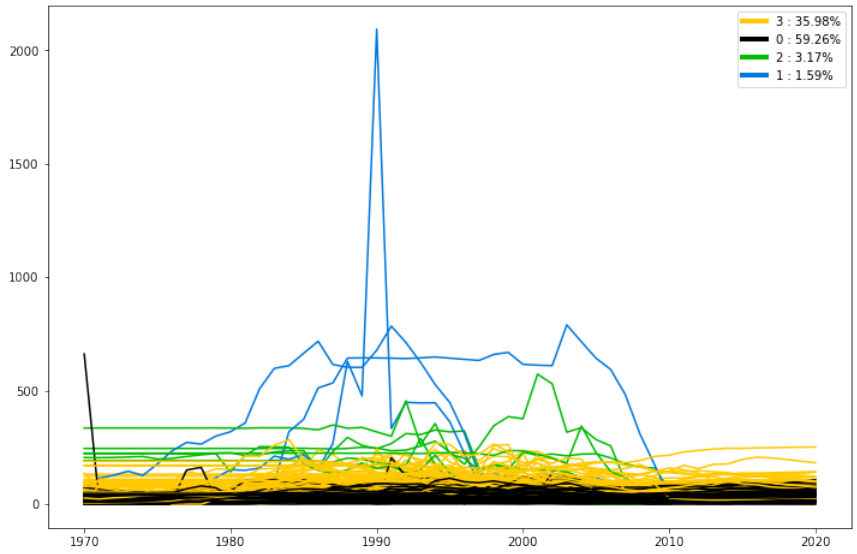
Viene utilizzato il dynamic time warping come misura di distanza delle serie temporali.

3.4.1 Time Serires KMeans



La prima immagine mostra i risultati del modello KMeans con 4 cluster, KNN per l'imputing sull'indicatore del prodotto interno lordo sul periodo di tem-

po 1970-2020. I risultati ottenuti risultano in linea con i risultati del kmeans applicato sul dato multidimensionale.



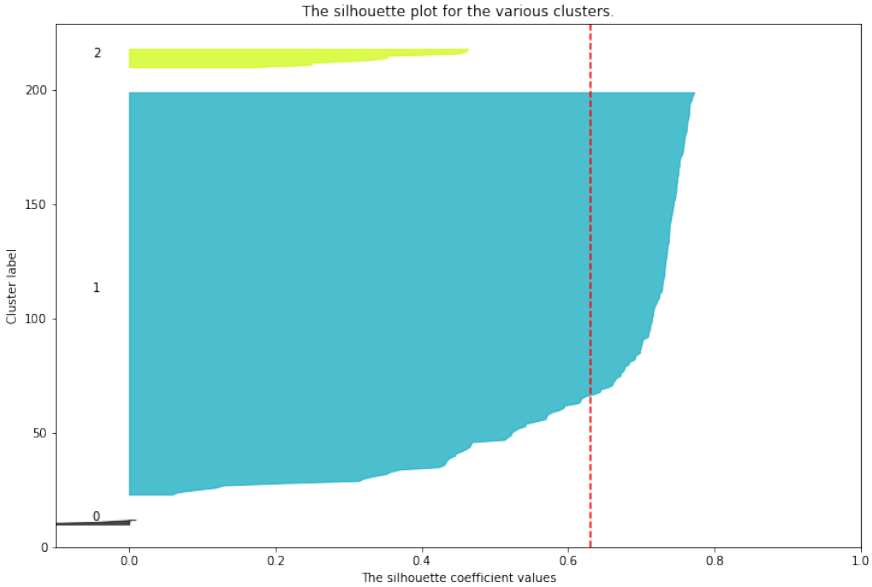
Anche quando applicato sull'indicatore del rateo con gli stessi parametri del modello precedente i risultati sono solo leggermente superiori all'algoritmo k-means applicato con l'approccio multidimensionale, portando però alla creazione di cluster con un numero di elementi più bilanciato (non è presente un singolo cluster con più del 90% dei dati del dataset). Applicando lo stesso algoritmo al dataset modificato con imputing NOCB con 3 cluster si nota un aumento del coefficente di silhouette al 56% ma si torna alla situazione di un solo cluster principale con il 93% dei dati del dataset.

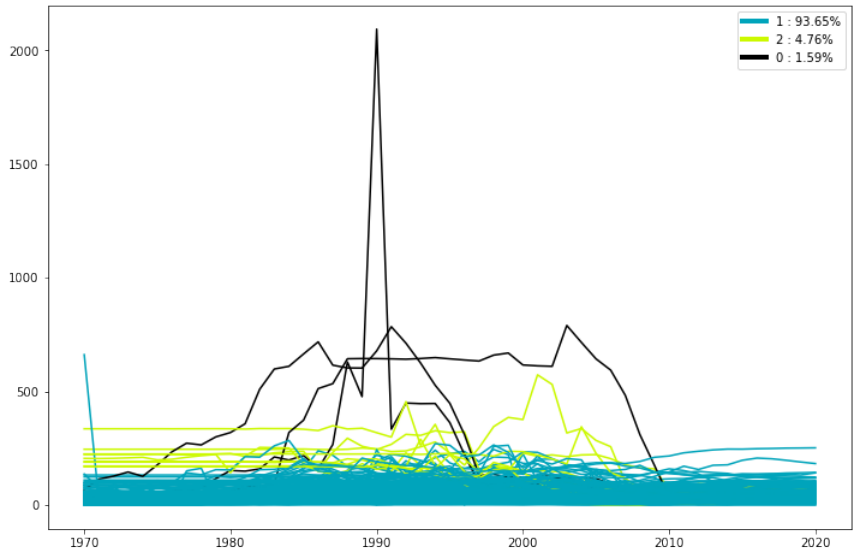
Modello e parametri	**Silhouette**
Time series KMEANS, 1970-2020, k=4, KNN, PROD	*0.9565*
Time series KMEANS, 1970-2020, k=4, KNN, RATE	*0.3766*
Time series KMEANS, 1970-2020, k=3, NOCB, RATE	0.5606

3.4.2 Agglomerative su matrice di distanze DTW

L'algoritmo di clustering agglomerativo permette di fornire direttamente una matrice delle similarità iniziali precomputata dall'utente, per poi eseguire su di essa i merge successivi che generano i cluster finali, viene quindi creata una matrice di similarità dei dati calcolata con la misura del dynamic time warping, questa viene poi data in input all'algoritmo di clustering. A causa dell'elevato tempo computazionale richiesto per computare la matrice questo modello è stato applicato soltanto al periodo di tempo 1970 - 2020 utilizzando il metodo

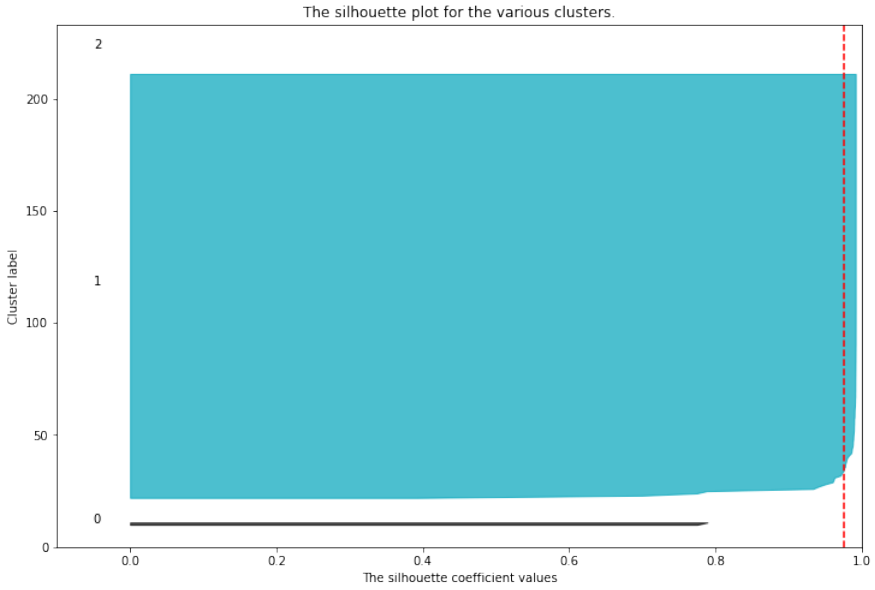
di imputing KNN, che è risultato il più efficente nei modelli precedenti. Avendo utilizzato una matrice di similarità precomputata l'algoritmo permette di modificare solo il parametro del linkage, saranno selezionabili i valori "single", "complete", "average".

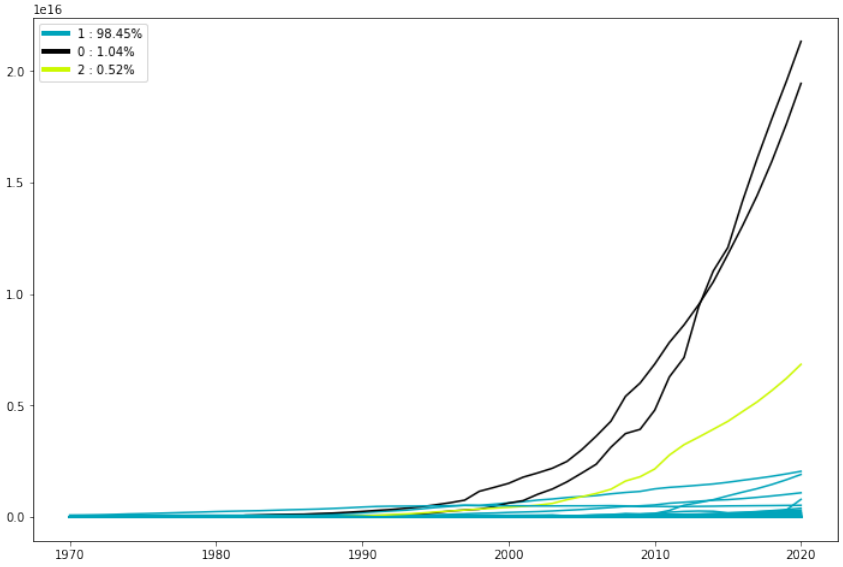




Il modello migliore risulta il modello con linkage completo su tre cluster applicato all'indicatore del rateo, nonostante presenti anche questo un singolo cluster con la maggior parte dei dati l'analisi delle serire temporale sembra aver

clusterizzato i dati sulla base della forma e del valore del rateo. I modelli single e average producono dei risultati del coefficente silhouette più alto ma producendo un singolo cluster principale.





Lo stesso algoritmo applicato sull'indicatore del prodotto presenta risultati simili con tutti i tipi di linkage con un risultato di clustering quasi identico al risultato ottenuto con il kmeans con l'approccio multidimensinoale

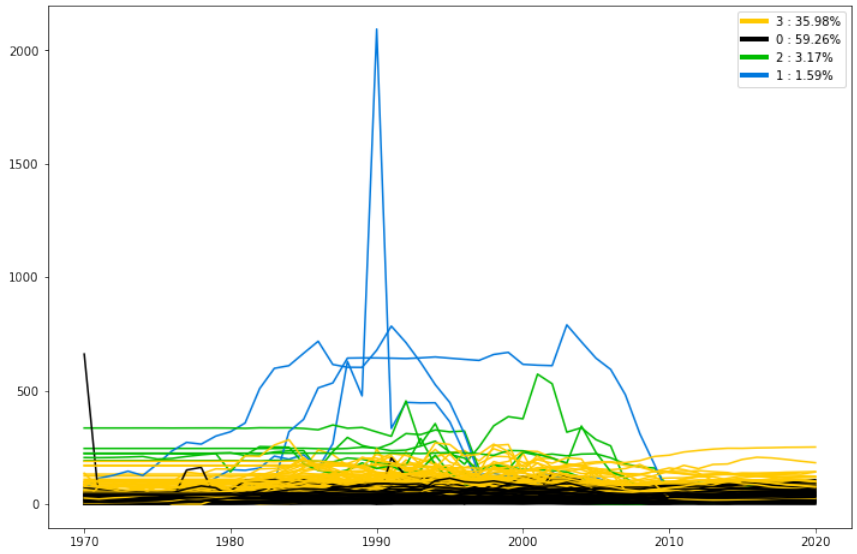
Modello e parametri	Silhouette
Agglomerative DTW, 1970-2020, k=3, KNN, RATE, COMPLETE	*0.6510*
Agglomerative DTW, 1970-2020, k=3, KNN, RATE, AVERAGE	*0.8156*
Agglomerative DTW, 1970-2020, k=3, KNN, RATE, SINGLE	0.4278
Agglomerative DTW, 1970-2020, k=3, KNN, PROD, SINGLE	0.9748

4 Pattern Mining

In questa fase vengono ricercati gruppi di peasi che presentano simili percentuali di incremento annuale attraverso l'utilizzo dell'algoritmo APRIORI. Le informazioni dell'incremento annuale sono rappresentate in forma di probabilità, vengono quindi effettuate delle binarizzazioni dei dati in modo da poter applicare gli algoritmi di pattern mining, i principali indicatori analizzati sono:

- Incremento positivo
- Incremento negativo

L'analisi viene poi eseguita anche in suddivisioni del range di valori positivo e negativo (ad esempio su un incremento che va dal 25% al 50%). Nell'ultima fase di analisi viene applicato sui gruppi di paesi trovati nella fase di clustering. Nello specifico vengono utilizzate le informazioni dei cluster creati dall'algoritmo TimeSeriesKMeans sul rateo. Verranno esclusi da questa analisi i cluster che presentano un numero sufficente di paesi (cluster 1 e 2 nell'immagine).



4.1 Pattern sul prodotto interno lordo

4.1.1 Incremento negativo

	support	itemsets	length
0	0.156863	(Bahamas, The)	1
1	0.156863	(Cameroon)	1
2	0.176471	(Central African Republic)	1
3	0.196078	(Gabon)	1
4	0.156863	(Oman)	1
5	0.196078	(Brunei Darussalam)	1
6	0.176471	(Ecuador)	1
7	0.254902	(Kuwait)	1
8	0.176471	(United Arab Emirates)	1
9	0.196078	(Congo, Republic of)	1
10	0.274510	(Zimbabwe)	1
11	0.196078	(Montserrat)	1
12	0.176471	(Timor-Leste, Dem. Rep. of)	1
13	0.215686	(Qatar)	1
14	0.176471	(Greece)	1
15	0.294118	(Guinea-Bissau)	1
16	0.294118	(Libya)	1
17	0.196078	(Saudi Arabia)	1
18	0.156863	(Japan)	1
19	0.176471	(Qatar, Kuwait)	2
20	0.156863	(Qatar, United Arab Emirates)	2
21	0.176471	(Saudi Arabia, United Arab Emirates)	2
22	0.176471	(Qatar, Saudi Arabia)	2
23	0.156863	(Libya, Saudi Arabia)	2
24	0.156863	(Qatar, Saudi Arabia, United Arab Emirates)	3

4.1.2 Incremento maggiore del 15%

	support	itemsets	length
0	0.627451	(Lao P.D.R.)	1
1	0.627451	(Costa Rica)	1
2	0.784314	(Argentina)	1
3	0.764706	(Democratic Republic of the Congo)	1
4	0.764706	(Sudan)	1
5	0.725490	(Venezuela)	1
6	0.627451	(Malawi)	1
7	0.666667	(Iran)	1
8	0.666667	(Zambia)	1
9	0.725490	(Turkey)	1
10	0.627451	(Sierra Leone)	1
11	0.843137	(Ghana)	1
12	0.647059	(Tanzania)	1
13	0.607843	(Sudan, Argentina)	2
14	0.647059	(Ghana, Argentina)	2
15	0.647059	(Sudan, Democratic Republic of the Congo)	2
16	0.627451	(Zambia, Democratic Republic of the Congo)	2
17	0.666667	(Democratic Republic of the Congo, Turkey)	2
18	0.686275	(Democratic Republic of the Congo, Ghana)	2
19	0.607843	(Venezuela, Sudan)	2
20	0.607843	(Zambia, Sudan)	2
21	0.607843	(Sudan, Turkey)	2
22	0.686275	(Sudan, Ghana)	2
23	0.686275	(Venezuela, Ghana)	2
24	0.627451	(Zambia, Ghana)	2
25	0.647059	(Ghana, Turkey)	2
26	0.607843	(Democratic Republic of the Congo, Ghana, Turkey)	3
27	0.607843	(Venezuela, Sudan, Ghana)	3

4.1.3 Incremento dal 25% al 50%

	support	itemsets	length
0	0.215686	(Chile)	1
1	0.431373	(Colombia)	1
2	0.215686	(Lao P.D.R.)	1
3	0.235294	(Iceland)	1
4	0.274510	(Democratic Republic of the Congo)	1
5	0.215686	(Myanmar)	1
6	0.313725	(Venezuela)	1
7	0.215686	(Mexico)	1
8	0.294118	(Nigeria)	1
9	0.411765	(Iran)	1
10	0.294118	(Turkey)	1
11	0.333333	(Qatar)	1
12	0.411765	(Ghana)	1
13	0.215686	(Paraguay)	1
14	0.235294	(Tanzania)	1
15	0.215686	(Iran, Colombia)	2
16	0.215686	(Paraguay, Colombia)	2

4.1.4 Incremento dal 50% al 100%

	support	itemsets	length
0	0.117647	(Belarus)	1
1	0.117647	(Equatorial Guinea)	1
2	0.137255	(Lao P.D.R.)	1
3	0.294118	(Uruguay)	1
4	0.176471	(Democratic Republic of the Congo)	1
5	0.137255	(Mozambique)	1
6	0.117647	(Uzbekistan)	1
7	0.117647	(Venezuela)	1
8	0.117647	(Zambia)	1
9	0.274510	(Turkey)	1
10	0.137255	(Uganda)	1
11	0.117647	(Ghana)	1
12	0.196078	(Peru)	1
13	0.137255	(Uruguay, Turkey)	2
14	0.117647	(Uruguay, Peru)	2
15	0.117647	(Mozambique, Turkey)	2

4.1.5 Incremento maggiore del 100%

	support	itemsets	length
0	0.117647	(Belarus)	1
1	0.294118	(Brazil)	1
2	0.215686	(Angola)	1
3	0.294118	(Argentina)	1
4	0.196078	(Democratic Republic of the Congo)	1
5	0.117647	(Venezuela)	1
6	0.117647	(Nicaragua)	1
7	0.117647	(Israel)	1
8	0.117647	(Vietnam)	1
9	0.117647	(Peru)	1
10	0.117647	(Belarus, Angola)	2
11	0.196078	(Brazil, Argentina)	2
12	0.117647	(Nicaragua, Brazil)	2
13	0.117647	(Brazil, Vietnam)	2
14	0.117647	(Brazil, Peru)	2
15	0.156863	(Angola, Democratic Republic of the Congo)	2
16	0.117647	(Peru, Argentina)	2
17	0.117647	(Brazil, Peru, Argentina)	3

4.2 Pattern sul rateo

4.2.1 Incremento negativo fino al -5%

	support	itemsets	length
34	0.235294	(Canada, Switzerland)	2
35	0.254902	(Belgium, Canada)	2
36	0.235294	(Belgium, Switzerland)	2
37	0.235294	(India, Switzerland)	2
38	0.235294	(Italy, Euro area)	2
39	0.274510	(Italy, Belgium)	2
40	0.235294	(Austria, Sri Lanka)	2
41	0.254902	(Belgium, Euro area)	2
42	0.235294	(Belgium, Austria)	2
43	0.235294	(India, Belgium)	2
44	0.254902	(India, Bangladesh)	2
45	0.235294	(Italy, Belgium, Euro area)	3

4.2.2 Incremento negativo dal -10% al -50%

	support	itemsets	length
26	0.235294	(Chile, Paraguay)	2
27	0.254902	(Chile, Ghana)	2

4.2.3 Incremento positivo fino al 15%

	support	itemsets	length
38	0.450980	(Italy, Canada)	2
39	0.411765	(Canada, United States)	2
40	0.411765	(Italy, France)	2
41	0.411765	(Italy, Belgium)	2
42	0.411765	(Italy, United States)	2
43	0.509804	(St. Lucia, Japan)	2
44	0.411765	(Dominica, Japan)	2
45	0.470588	(France, Japan)	2
46	0.411765	(France, United States)	2
47	0.411765	(Japan, United States)	2
48	0.470588	(Japan, Germany)	2
49	0.411765	(Belgium, United States)	2
50	0.411765	(Bahamas, The, United States)	2
51	0.470588	(Barbados, United States)	2

4.3 Pattern sui cluster del rateo

4.3.1 Cluster 0 pattern positivi e negativi

	support	itemsets	length
38	0.509804	(Dominican Republic, El Salvador)	2
39	0.509804	(Germany, Fiji)	2
40	0.509804	(Bahamas, The, Barbados)	2
41	0.529412	(Cyprus, Germany)	2

	support	itemsets	length
39	0.450980	(Oman, Indonesia)	2
40	0.450980	(Philippines, Botswana)	2
41	0.450980	(New Zealand, Indonesia)	2
42	0.450980	(New Zealand, Iran)	2
43	0.450980	(New Zealand, Botswana)	2
44	0.470588	(Australia, New Zealand)	2
45	0.470588	(New Zealand, Denmark)	2

4.3.2 Cluster 3 pattern postitivi e negativi

	support	itemsets	length
25	0.509804	(Canada, Greece)	2
26	0.568627	(Honduras, Japan)	2
27	0.529412	(Italy, Greece)	2
28	0.509804	(Japan, Mali)	2
29	0.568627	(Japan, Ethiopia)	2
30	0.509804	(Japan, Bolivia)	2
31	0.529412	(Japan, United States)	2
32	0.568627	(Japan, Greece)	2
33	0.509804	(Gambia, The, Japan)	2
34	0.509804	(Japan, Democratic Republic of the Congo)	2
35	0.509804	(United States, Greece)	2

	support	itemsets	length
32	0.431373	(Comoros, Panama)	2
33	0.411765	(Comoros, Côte d'Ivoire)	2
34	0.431373	(Belgium, Ireland)	2