UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

STATISTICA E ANALISI DEI DATI

Analisi Temporale delle Aree Marine Protette: Studio basato su Dati OECD e Utilizzo di R per l'Esplorazione e l'Analisi Statistica

Università degli Studi di Salerno

Irene Gaita

Matricola: 0522501839

Anno Accademico 2023-2024

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

Più del **70%** della superficie del nostro pianeta è occupata dagli oceani che comprendono ecosistemi eterogenei e offrono una vasta gamma di servizi fondamentali per la società umana, la salute e le economie [1]. In diversi paesi, le attività industriali legate agli oceani hanno registrato una crescita senza adeguata considerazione per l'ambiente, mettendo a rischio le risorse naturali e i servizi ecosistemici marini fondamentali che costituiscono la base delle economie e del benessere umano.

Le **zone marine protette** (MAPs) sono generalmente descritte come porzioni dell'ambiente marino, interne o adiacenti, che sono **riservate dalla legislazione** o da altri mezzi efficaci per **garantire un livello superiore di protezione alla biodiversità marina e/o costiera**. Tali zone hanno il potenziale di preservare e ristabilire habitat e le specie, assicurando nel contempo che gli ecosistemi marini e costieri continuino a svolgere un ruolo cruciale nel fornire protezione contro tempeste ed erosione, sequestro di carbonio, pesca, opportunità ricreative e turistiche.

Il riconoscimento dell'importanza delle reti di aree protette per la biodiversità marina e per i servizi ecosistemici è riflesso nell'Obiettivo 3 del **Quadro Globale per la Biodiversità** , il quale stabilisce l'obiettivo di ripristinare almeno il 30% delle aree

[1] https://www.protectedplanet.net/en/thematic-areas/marine-protected-areas

degradate degli ecosistemi terrestri, acque interne, costieri e marini entro il 2030 [2].

Questo studio si propone di contribuire in modo significativo alla comprensione della gestione e conservazione delle risorse marine attraverso un'analisi approfondita dell'**evoluzione delle aree marine protette (MAPs) nel periodo 2000-2022**. Particolare attenzione verrà dedicata alla distinzione tra paesi appartenenti all'Organizzazione per la **Cooperazione e lo Sviluppo Economico** (OECD) al fine di identificare possibili disuguaglianze nell'impegno per la conservazione marina. La valutazione di trend significativi nelle dimensioni delle MAPs non solo fornirà una panoramica del progresso nel raggiungimento degli obiettivi globali di conservazione, ma **permetterà anche di esaminare la coerenza delle politiche di protezione ambientale adottate dai diversi paesi nel tempo**. Questo studio si propone inoltre di esplorare le disparità regionali nelle MAPs, mettendo in luce eventuali **pattern geografici rilevanti**, nonchè di fornire una visione complessiva sull'efficacia delle **attuali iniziative di conservazione marina**, delineando il loro impatto sulla biodiversità e sui servizi ecosistemici marini per orientare futuri sviluppi e decisioni politiche. L'approccio metodologico si avvarrà di strumenti statistici forniti dal linguaggio di programmazione R, garantendo una robustezza analitica e facilitando la comparabilità tra le nazioni.

1.1 Descrizione dataset

Il dataset presentato nella figura 1.1 è stato acquisito mediante il sito web ufficiale delle statistiche dell'**Organizzazione per la Cooperazione e lo Sviluppo Economico**. Questa procedura ha coinvolto l'esplorazione dettagliata delle risorse online fornite dall'Organizzazione, con la finalità di recuperare in modo accurato e completo il dataset utilizzato nel nostro studio di ricerca.

Il dataset presentato offre una visualizzazione temporale dell'espansione delle aree marine protette, con gli anni compresi **tra il 2000 e il 2022** sull'asse delle ascisse,

[2]Con il Global Biodiversity Framework si dovranno eliminare gradualmente o riformare entro il 2030 i sussidi che danneggiano la biodiversità per un valore di 500 miliardi di dollari all'anno, aumentando gli incentivi positivi per la conservazione e l'uso sostenibile della biodiversità.

https://www.aics.gov.it/oltremare/articoli/pianeta/cop15-cosa-significa-il-global-biodiversity-framework-per-la-cooperazione/.

e le nazioni suddivise tra quelle facenti parte dell'OECD [3] e quelle che non ne fanno parte sull'asse delle ordinate. Tuttavia, l'attenzione di questa analisi statistica è focalizzata sulle nazioni che sono **membri dell'OECD** e che nel corso degli anni hanno **mostrato cambiamenti o trend di particolare interesse**. A tal ragione, sono state scelte 25 nazioni: **Australia, Belgio, Canada, Cile, Colombia, Costa Rica, Danimarca, Estonia, Francia, Germania, Grecia, Irlanda, Italia, Giappone, Core, Messico, Olanda, Nuova Zelanda, Norvegia, Polonia, Portogallo, Spagna, Svezia, Regno Unito e Stati Uniti**.

Al fine di migliorare la confrontabilità, la facilità di visualizzazione e la stabilità dei modelli, tutti i valori sono stati arrotondati a **zero cifre decimali**. Questa scelta è stata compiuta per garantire una rappresentazione chiara e concisa dei dati, mantenendo al contempo la precisione sufficiente per analizzare le tendenze nel lungo periodo. La semplificazione dei valori consente una migliore comprensione delle dinamiche delle aree marine protette in relazione alle nazioni dell'OECD nel periodo considerato.

Variable	Total marine protected area, sqkm																
Unit	Square kilometres																
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	201
Country																	
Australia	376 896.5	377 198.5	399 905.8	399 923	402 051.8	406 363.8	412 437.8	417 115.8	417 560.3	442 185.3	442 910.5	448 550	829 988.8	830 412	830 600.8	830 606	847 5
Belgium	55.25	55.25	55.25	58.25	58.25	350.25	350.25	350.25	350.25	350.25	350.25	350.25	1 248.25	1 248.25	1 249.75	1 249.75	1 24
Canada	24 661.75	24 718.75	24 918.5	28 145	30 039.5	32 196.5	32 534	32 740.75	41 618.75	42 122	50 992	51 081.5	51 087.75	52 431	52 434.5	53 968.75	56 2
Chile	8 849.75	8 849.75	8 849.75	8 868.75	10 068.5	10 177	10 218.5	10 219.75	10 219.75	10 219.75	160 674.3	160 674.3	160 674.3	160 674.5	160 675.3	160 896.3	461 3
Colombia	29 421.75	29 421.75	29 421.75	29 421.75	29 421.75	60 890.5	60 906.5	60 906.5	60 910.5	60 959.75	61 416.5	61 425.5	61 426.5	63 135.5	63 829	63 829	63 83
Costa Rica	58 358.75	58 371	58 371	58 371	58 371	58 371	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 570.75	58 57
Denmark	7 678	7 678	7 678.25	9 451.25	11 913.75	12 282.5	12 283.75	13 020.5	13 020.5	18 667	18 667	18 676.25	18 676.25	18 676.25	18 677.25	18 677.25	19 08
Estonia	580.5	580.5	580.5	580.5	6 473.25	6 529.5	6 530	6 544	6 544	6 575	6 725.5	6 725.5	6 725.5	6 725.5	6 725.5	6 725.5	6 7
Finland	7 172.5	7 222.25	7 223.25	7 225	7 251	7 448.5	7 457.5	7 461.5	7 467.25	7 470	7 472	8 126	8 132.25	8 133.75	8 156.75	8 868.75	8 8
France	78 823	78 831	78 835.5	78 876.5	78 877.5	80 907.75	81 150.5	84 700.75	91 109.5	95 382.25	96 512	97 744	114 213.8	115 989.5	118 465.5	121 208.5	123
Germany	9 102.5	9 226.5	9 232.75	9 237.5	17 468.75	17 487.25	17 493.75	22 524.75	25 010	25 062.5	25 155.75	25 157.25	25 157.75	25 157.75	25 161.25	25 170	25 54
Greece	13 845.25	16 108.25	17 255.25	17 255.5	17 255.75	17 256	17 270.25	17 270.25	17 595	17 628.75	18 242.5	22 610.5	22 628.25	22 628.25	22 628.25	22 628.25	22 62
Iceland	2 772.5	2 771.5	2 772	2 772	2 772	2 772	2 848.75	2 849.5	2 921.5	2 927.25	2 927.25	2 942.75	2 956.75	2 956.75	3 261.5	3 261.5	3 2
Ireland	2 784	2 784	2 786	2 786	2 899.25	2 902.75	2 911.75	2 918.75	2 926.25	5 727.5	5 739.5	5 762	5 762	5 762	5 762	5 762	8 29
Israel	0	0	5.75	6.5	7.25	7.5	7.5	7.5	9	10.5	10.5	10.5	10.5	10.5	10.5	10.5	
Italy	29 997	29 997.5	30 024.75	30 044.5	30 294.75	34 585.75	36 834.25	37 290	37 290.75	37 505.25	37 505.25	37 505.25	37 506.25	37 506.5	37 506.5	37 575.5	37 63

Figura 1.1: Dataset di partenza [1]

[3] Organizzazione per la cooperazione e lo sviluppo economico

CAPITOLO 2

Background

2.1 Serie temporali

Le serie temporali sono impiegate nella visualizzazione grafica dei dati al fine di esaminarne l'andamento nel corso di un determinato periodo. In questo contesto, il **tempo** assume un ruolo cruciale come variabile significativa, poiché i dati vengono organizzati in sequenza cronologica.

Nel contesto del set di dati preso in considerazione e come mostra anche la figura 2.1, l'inizio del periodo temporale sarà rappresentato dall'anno 2000, mentre la fine del periodo sarà rappresentata dall'anno 2022. In questa analisi, considereremo ciascun anno come una singola frequenza nel corso di questo intervallo temporale specifico.

```
Time Series:
Start = 2000
End = 2022
Frequency = 1
 [1]  376896  377198  399906  399923  402052  406364  412438  417116  417560  442165  442910
[12]  448550  829989  830412  830601  830606  847507  855715 3030763 3036163 3036163 3036163
[23] 3036163
```

Figura 2.1: Serie temporale in R

Attraverso la funzione plot() possiamo visualizzare l'andamento dei valori assunti dalle varie nazioni nel corso degli anni.

Consideriamo dunque ad una ad una le nazioni del dataset e vediamo come è variato nel tempo il valore che rappresenta la totalità del territorio marino protetto.

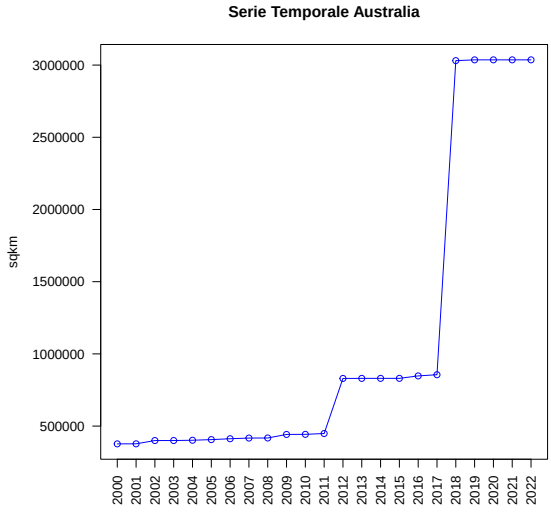

Figura 2.2: Serie temporale Australia

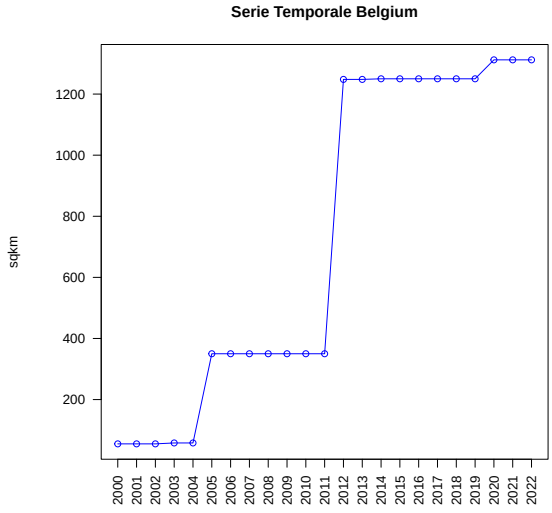

Figura 2.3: Serie temporale Belgio

Nel contesto **australiano**, i vibranti colori dei coralli, risultanti dalla presenza delle zooxantelle, possono fungere da indicatore della salute degli ecosistemi corallini. Tuttavia, la variazione della temperatura, anche di pochi gradi, può scatenare il fenomeno del sbiancamento corallino. Un esempio eclatante di questa dinamica è rappresentato dagli eventi di sbiancamento che hanno interessato la **Grande Barriera Corallina**[1] negli ultimi venticinque anni, con un picco preoccupante nel 2016, quando oltre il 90% della sua superficie è stata colpita, portando a una perdita del 22% del corallo. Le aree maggiormente colpite hanno sperimentato un tasso di mortalità del corallo superiore all'80%.

Tuttavia, nonostante questi impatti negativi, la superficie delle regioni settentrionali della Grande Barriera Corallina ha mostrato una crescita tra il 2018 e il 2022, indicando una certa capacità di resilienza dell'ecosistema. La resilienza, intesa come la capacità di un ecosistema di ripristinare il suo stato di equilibrio dopo eventi avversi, è evidente nei periodi di minor stress ambientale [2]. Il grafico rappresentato nella Figura 2.2 riflette la dinamica della crescita delle aree marine protette in Australia.

Nel grafico rappresentato dalla Figura 2.2, si nota che il valore minimo di chilometri quadrati (sqkm) in **Australia** è stato registrato nel 2000. Da quel punto fino al 2011, la crescita è stata costante, caratterizzata da incrementi brevi. Nel 2012, si è verificato un notevole aumento di 381439 sqkm, e la crescita è rimasta costante fino al 2018. Tuttavia, a partire dal 2018, si è osservato un aumento repentino della crescita, mantenendo una continuità fino al 2022.

Andamento analogo è rilevato dal **Belgio** ed evidenziato nel grafico della Figura 2.3. Si distingue per la presenza di due picchi di crescita, uno nel 2005 e l'altro nel 2013. Nel periodo esteso dal 2013 al 2022, si osserva una tendenza all'incremento praticamente lineare.

[1] https://it.wikipedia.org/wiki/Grande$_b$*arriera*$_c$*orallina*

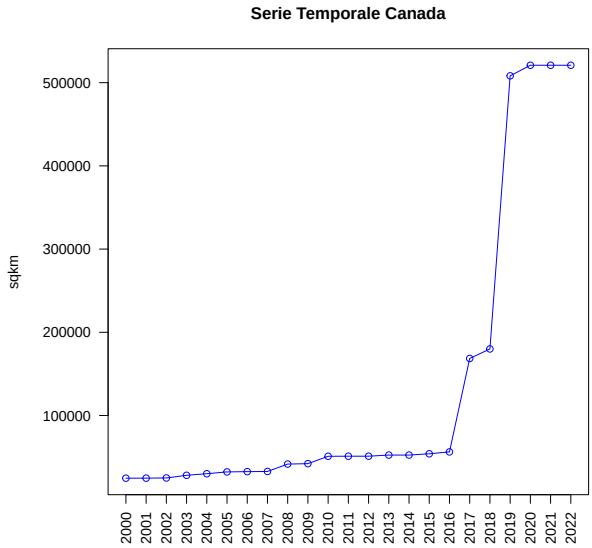

Figura 2.4: Serie temporale Canada

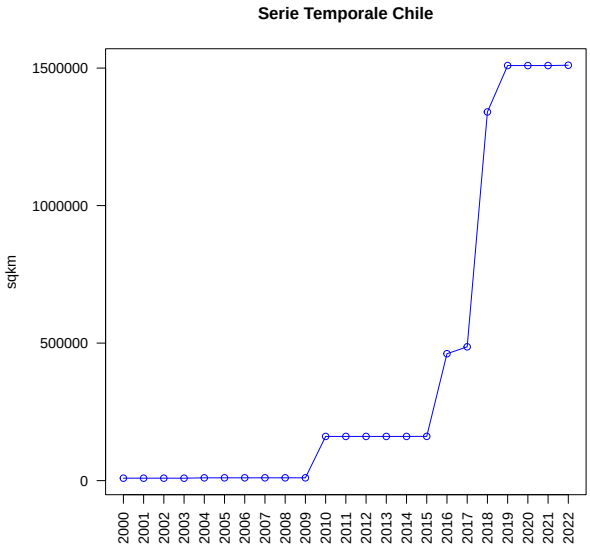

Figura 2.5: Serie temporale Cile

L'istituzione del 27 giugno 2018 dello **Scott Islands Marine National Wildlife Refuge** rappresenta una pietra miliare nella conservazione marina del **Canada**. Quest'area marina protetta, la prima ad essere istituita ai sensi del Canada Wildlife Act, ha come obiettivo principale la salvaguardia delle diverse specie di uccelli marini che popolano la costa del Pacifico. La sua creazione riflette l'impegno del Canada nel proteggere e preservare la ricca biodiversità marina presente nella regione [3].

Riguardo all'andamento della crescita delle aree marine protette in Canada,visibile in Figura 2.4 si osserva un trend pressoché lineare e costante. Nel periodo compreso tra il 2018 e il 2019, si evidenzia un significativo aumento delle estensioni delle aree protette, mantenendo successivamente una costanza fino al 2022. Questa crescita costante riflette l'attenzione e gli sforzi del Canada nel rafforzare la rete di aree marine protette, contribuendo così alla conservazione e al mantenimento degli ecosistemi marini in un contesto di rapido cambiamento ambientale. La creazione dello Scott Islands Marine National Wildlife Refuge potrebbe aver giocato un ruolo significativo in questa tendenza positiva, fornendo un esempio tangibile dell'impatto positivo delle politiche di conservazione e della creazione di nuove aree protette.

Nel contesto del **Cile**, è importante notare che la crescita economica, come illustrato nei grafici rappresentati dalle Figure 2.5, segue un trend pressoché lineare e costante. Questo andamento è caratterizzato da brevi episodi di incremento tra

il 2010 e il 2015. Tuttavia, emerge un notevole aumento delle crescite nel periodo compreso tra il 2018 e il 2019, mantenendo successivamente una stabilità fino al 2022. È interessante notare che durante questo periodo di crescita economica, la presidente uscente **Michelle Bachelet** ha contribuito alla tutela dell'ambiente marino attraverso l'istituzione di nove aree marine protette.

La legislazione introdotta da Bachelet ha sottolineato l'importanza di preservare le acque al largo delle coste cilene, comprese le isole, come le Diego Ramírez, che fungono da habitat per leoni marini, balene e foreste di alghe, nota come kelp. Questa iniziativa mira a proteggere la ricchezza ecologica dell'ecosistema marino cileno. Tuttavia, con l'insediamento del nuovo presidente **Sebastián Piñera**, si è posto un focus aggiuntivo sulla tutela degli animali domestici e sulla trasformazione dello zoo di Santiago in un ecoparco. L'obiettivo principale di garantire la ricerca scientifica, il benessere animale e la conservazione delle specie in via d'estinzione sembra essere stato affrontato con successo [4].

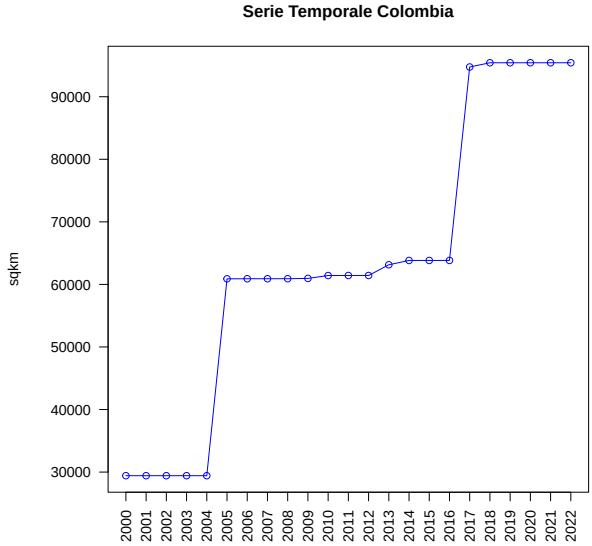

Figura 2.6: Serie temporale Colombia

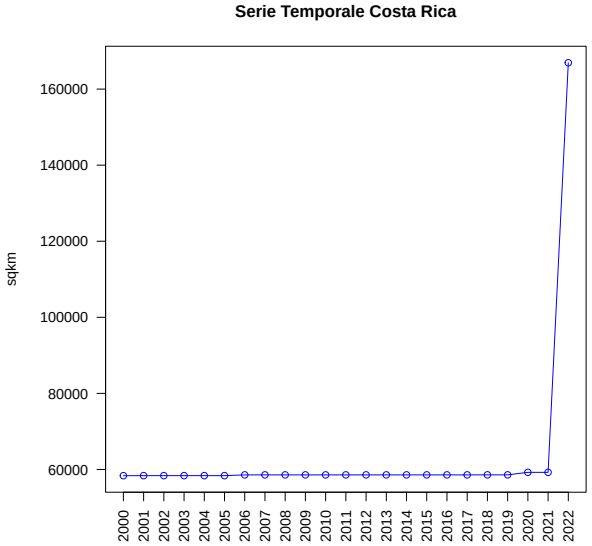

Figura 2.7: Serie temporale Costa Rica

Anche relativamente al grafico illustrato nella Figura 2.6, si riscontrano due picchi di crescita nei chilometri quadrati (sqkm) in **Colombia**. Nel 2005, si è verificato un incremento di 31.468 sqkm, seguito da un altro picco nel 2017 con un aumento di 30.921 sqkm. Per gli anni successivi, la crescita presenta brevi incrementi, ma mantiene una costante tendenza alla crescita.

Il grafico relativo alla **Costa Rica**, rappresentato nella Figura 2.7, mostra una singolare dinamica. Nel periodo dal 2001 al 2017, non si osserva alcuna variazione, con un valore costante di 58.571 km². Tuttavia, negli anni 2018 e 2019 si registra un incremento di soli 6 km². Questo aumento è associato a decisioni politiche e iniziative mirate ad espandere le aree marine protette della nazione. Gli anni successivi, 2020 e 2021, mostrano una crescita moderata, raggiungendo la cifra di 59.239 km². Nel 2022 si evidenzia un repentino aumento di 107.681 km². Questo incremento significativo è associato a eventi o decisioni eccezionali, come l'adesione a nuovi **accordi internazionali**, come quello discusso durante la COP26, o la **firma della dichiarazione per la protezione delle isole Cocos, Galapagos, Malpelo e Coiba**. Questa iniziativa potrebbe aver contribuito in modo significativo all'espansione delle aree marine protette della Costa Rica [5].

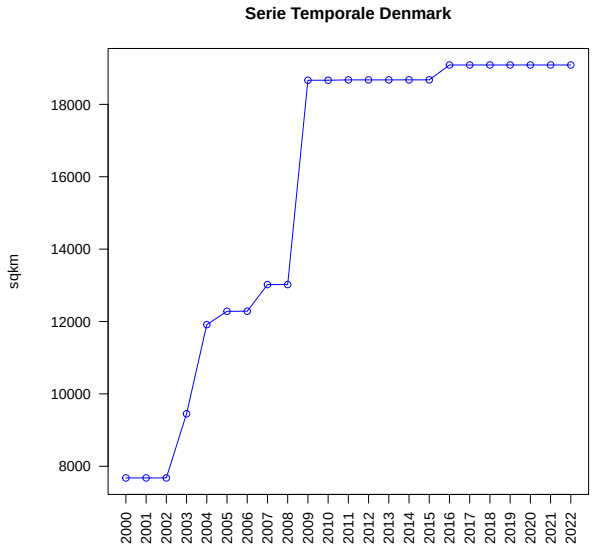

Figura 2.8: Serie temporale Danimarca

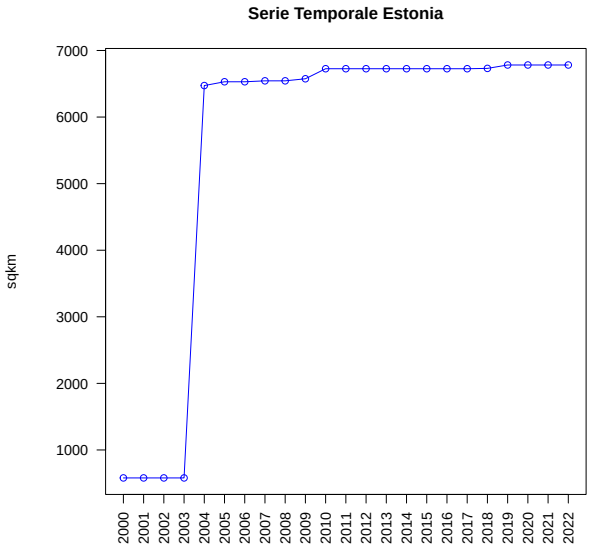

Figura 2.9: Serie temporale Estonia

Nella Figura 2.8 inerente all'andamento della **Danimarca**, si osserva che nel periodo compreso tra il 2000 e il 2002, il valore rimane invariato a 7678 sqkm. Un

notevole picco di crescita è evidenziato nel 2009, mentre nei periodi successivi fino al 2022, si manifestano aumenti caratterizzati da brevi incrementi.

Per quanto concerne il grafico raffigurato nella Figura 2.9, che illustra il totale delle aree marine protette dell'**Estonia**, si evidenzia un incremento di 5893 sqkm nel 2004 rispetto agli anni precedenti, dal 2000 al 2003, i quali registrano una superficie di 580 sqkm. Nel periodo successivo, ovvero dal 2005 al 2022, la crescita manifesta una costanza mantenendosi nell'intervallo compreso tra 6000 e 7000 sqkm.

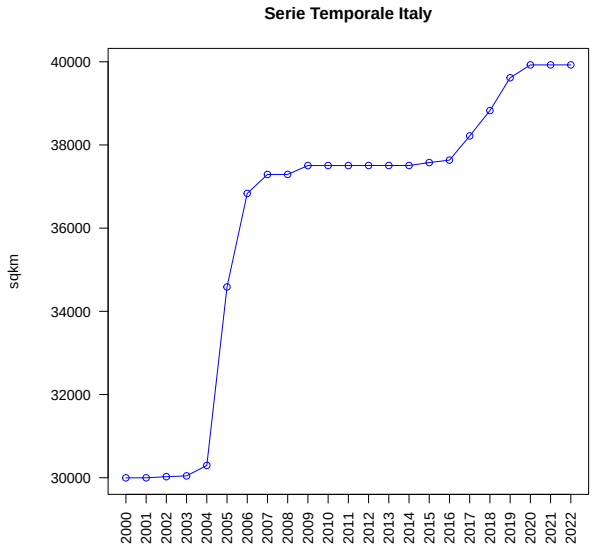

Figura 2.10: Serie temporale Italia

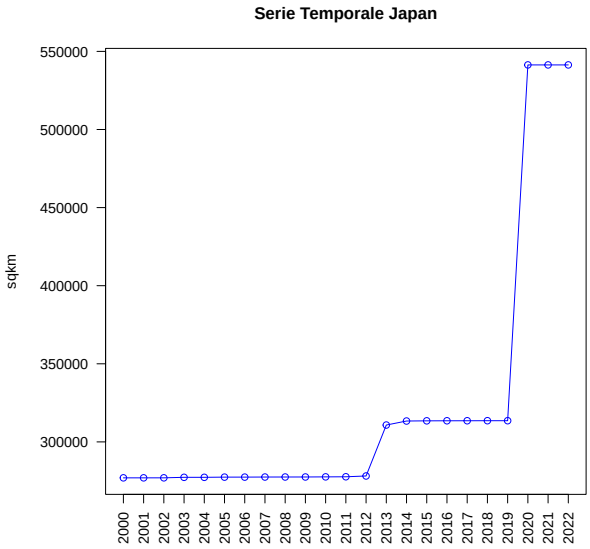

Figura 2.11: Serie temporale Giappone

In riferimento al grafico presentato nella Figura 2.10, l'entità complessiva delle aree marine protette in **Italia** ha manifestato una marcata tendenza di crescita costante fino al 2020, mantenendosi stabile con una superficie di 39924 sqkm fino al 2022.

Seppure il grafico del **Giappone** in Figura 2.11 sia simile a quello della Costa Rica in Figura 2.7, occorre notare che nel 2013 il Giappone manifesta un incremento di 32611 sqkm, un dato che distingue la sua dinamica territoriale.

La distribuzione delle variazioni dei chilometri quadrati (sqkm) nei territori marini protetti mostra similitudini notevoli tra alcune nazioni. Ad esempio, i grafici di **Francia**, **Germania**, **Grecia** e **Irlanda** presentano un andamento simile a quelli di Danimarca in Figura 2.8. Allo stesso modo, le tendenze in **Corea del Sud** sono analoghe a quelle osservate in Italia in Figura 2.10, **Messico**, Giappone in Figura 2.11, **Spagna**, **Polonia** e **Nuova Zelanda**.

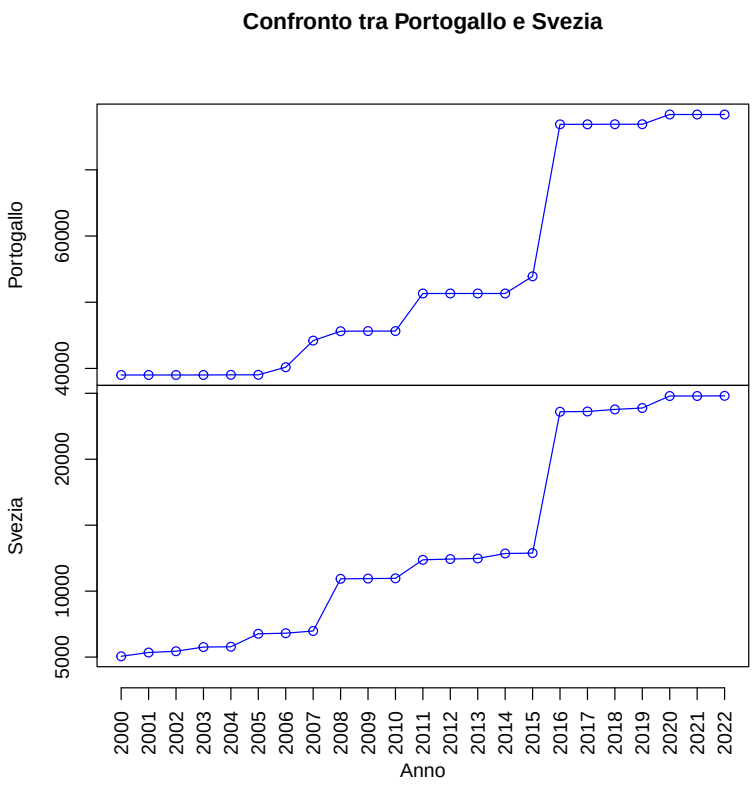

Figura 2.12: Confronto tra la Serie Temporale del Portogallo e della Svezia

In Figura2.12 è possibile osservare un esempio di confronto tra due serie temporale di nazioni presenti nella lista appena menzionata: il **Portogallo** e la **Svezia**. L'andamento di crescita risulta essere simile per entrambe le nazioni con un picco di crescita nel 2016 ed un beve incremento nel 2020 che resta costante fino al 2022.

2.2 Diagramma a barre - Barplot

Il diagramma a barre costituisce uno strumento visivo impiegato per confrontare informazioni tra diverse categorie.

In questo contesto, il diagramma a barre sarà impiegato per rappresentare graficamente il totale delle aree marine protette di **ciascuna nazione per anno**. L'obiettivo è individuare in modo rapido e intuitivo le nazioni con la massima, minima e simile estensione delle aree marine protette. Per agevolare la comprensione del lettore, saranno aggiunte delle linee al grafico che aiuteranno a individuare le nazioni con la massima e la minima estensione delle aree marine protette per l'anno specifico

considerato nel diagramma a barre. Sull'asse delle ascisse saranno indicate le nazioni, mentre sull'asse delle ordinate verranno rappresentati i chilometri quadrati di territorio marino protetto (sqkm).

2.2.1 Barplot Anno 2000

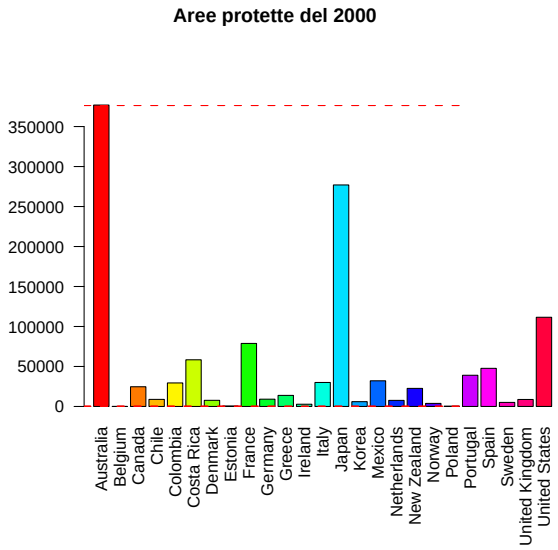

Figura 2.13: Barplot Anno 2000

Dal grafico in Figura 2.13, possiamo osservare che il **valore massimo** è associato all'**Australia** con 376896 sqkm, mentre il **valore minimo** è attribuito al **Belgio**.

Anche se l'**Italia** e il **Messico** sembrano avere lo stesso totale delle aree marine protette, vi è una differenza di 668827 sqkm. La seconda nazione con il totale in sqkm delle aree marine risulta essere il Giappone con 276987 sqkm.

2.2.2 Barplot Anni 2001-2002-2003-2005

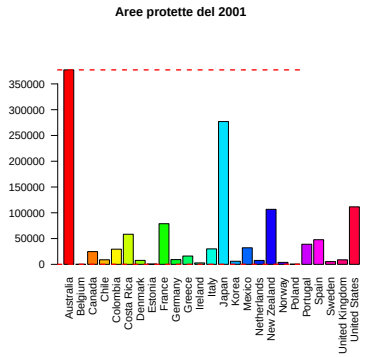

Figura 2.14: Barplot Anno 2001

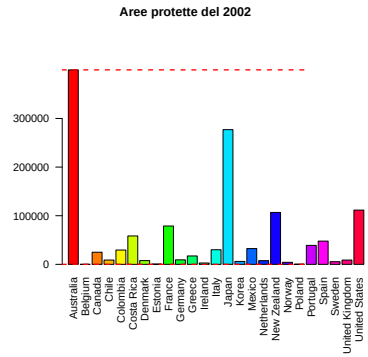

Figura 2.15: Barplot Anno 2002

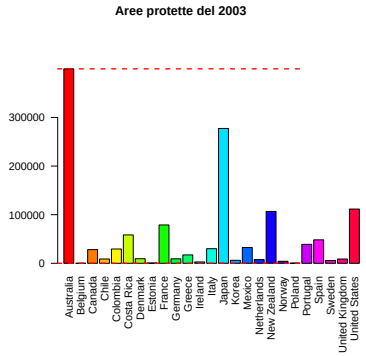

Figura 2.16: Barplot Anno 2003

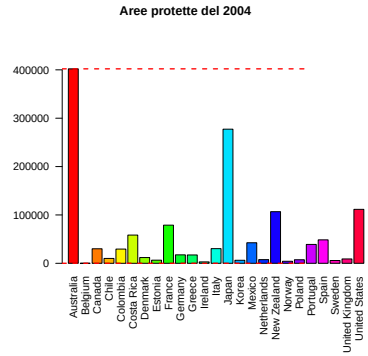

Figura 2.17: Barplot Anno 2004

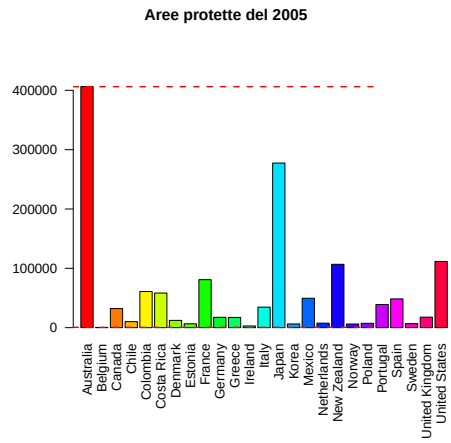

Figura 2.18: Barplot Anno 2005

I grafici nelle Figure 2.14, 2.15, 2.16, 2.17 relativi agli anni 2001, 2002, 2003 e 2004, risultano essere simili tra di loro. Il **valore massimo** risulta essere attribuito sempre all'**Australia** che nel 2005 ha riscontato un incremento di 24854 sqkm rispetto al 2001 cioè di 6.62%. Le nazione hanno mantenuto quindi in questo arco temporale una crescita lineare e costante.

2.2.3 Barplot Anno 2006

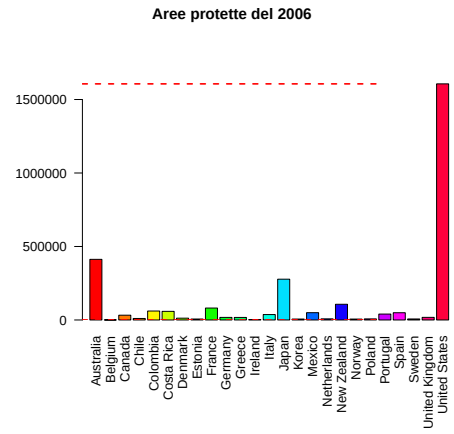

Figura 2.19: Barplot Anno 2006

L'analisi del diagramma a barre, rappresentato nella Figura 2.19, relativo all'anno 2006, rivela una significativa transizione nella distribuzione delle aree marine protette.

Gli **Stati Uniti** emergono come la nazione preminente, evidenziando il massimo valore di chilometri quadrati (sqkm) dedicati a tali aree. Questo sorpasso indica un cambiamento rilevante da parte degli Stati Uniti rispetto agli altri Paesi considerati. È interessante notare come rispetto l'anno precedente, è stato registrato un aumento di ben 1494833 sqkm.

In confronto all'**Australia**, il divario osservato nei valori del diagramma a barre ammonta a 1193967 sqkm.

2.2.4 Barplot Anni dal 2007 al 2017

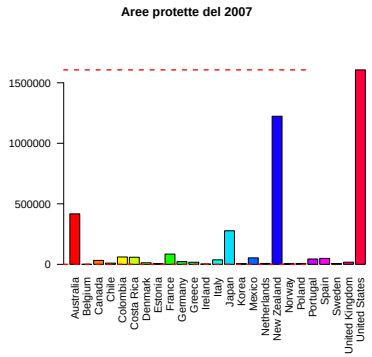

Figura 2.20: Barplot Anno 2007

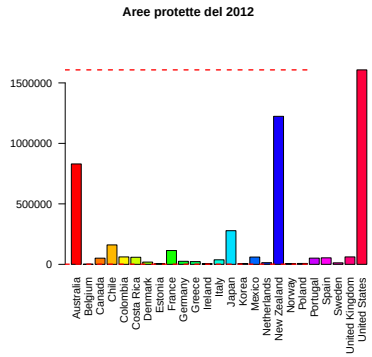

Figura 2.21: Barplot Anno 2012

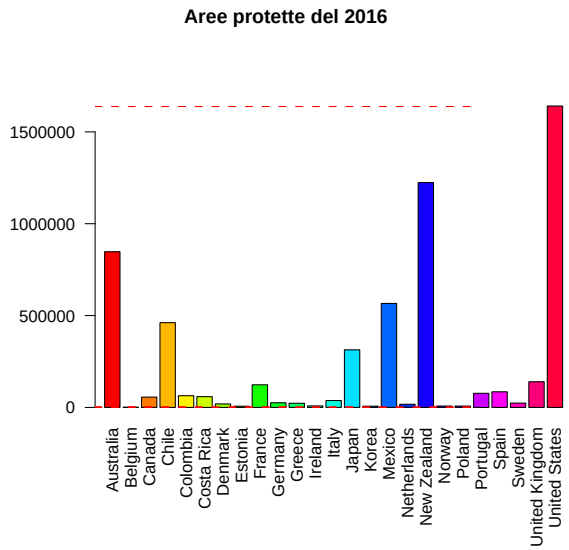

Figura 2.22: Barplot Anno 2017

Nelle Figure 2.20, 2.21 è possibile osservare i diagrammi a barre relativi agli anni 2007 e 2012.

Nell'arco di questo periodo i grafici risultano essere affini. Rispetto all'anno 2006, è rilevante notare come la **Nuova Zelanda** registri un incremento di 1116798 chilometri quadrati (sqkm) nel 2007 dell' area marina protetta. Tale aumento indica una notevole

espansione degli sforzi di conservazione marina da parte della Nuova Zelanda nel periodo considerato.

È importante notare il crescente aumento della superficie da parte degli Stati Uniti. Il 21 luglio 2008 il Presidente George W. Bush firmò con la legge H.R. 802 il Maritime Pollution Prevention Act del 2008, che permise agli Stati Uniti di diventare Stato Parte della Convenzione MARPOL. Quest'ultimo stabilisce che «large diesel-powered, ocean-going vessels such as container ships, tankers, cruise ships and bulk carriers must limit their emission of nitrogen oxides (NOx) and use cleaner- burning fuels to reduce their sulfur dioxide (SO2) emissions»[2][6].

Dal 2012 al 2017 in Figura 2.22, l'andamento generale dei grafici mantiene una similitudine con netti imcrementi per il **Cile** e per il **Messico**.

2.2.5 Barplot Anni dal 2018 al 2022

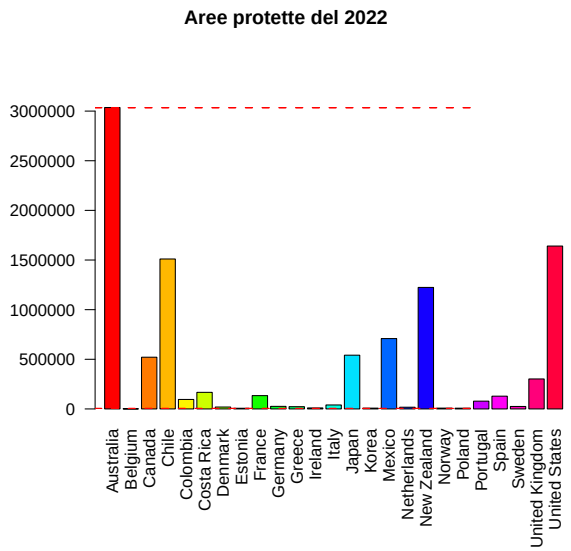

Figura 2.23: Barplot Anno 2022

[2]Le grandi navi d'alto mare alimentate a diesel, come le navi portacontainer, le petroliere, le navi da crociera e le navi portarinfuse, devono limitare le loro emissioni di ossidi di azoto (NOx) e utilizzare combustibili più puliti per ridurre le emissioni di anidride solforosa (SO2).

Esaminando il diagramma a barre nella Figura 2.23, emerge chiaramente come l'**Australia**, nel confronto con il periodo precedente, riprende la prima posizione con un massimo valore di 3036163 chilometri quadrati (sqkm). Nel medesimo intervallo temporale, gli **Stati Uniti** mantengono una notevole stabilità nei loro valori, con una variazione pressoché trascurabile: nel 2017 il dato è di 1640636 sqkm, mentre nel 2022 si attesta a 1640681 sqkm.

Questa tendenza indica un'andatura costante per gli Stati Uniti nel mantenimento delle aree marine protette nel periodo in esame.

2.3 Visualizzazione delle variazioni territoriali

Durante l'analisi condotta, è stato possibile creare due rappresentazioni cartografiche utilizzando **ggplot2**, come mostrato nelle Figure 2.24 e 2.25. Queste mappe offrono una visualizzazione chiara delle variazioni dell'estenzione del territorio marino protetto delle nazioni facenti parte dell'OECD per gli anni 2000 e 2022.

L'impiego di colori e forme distintive ha permesso di evidenziare le differenze significative a livello globale, consentendo una comprensione approfondita delle tendenze che emergono dai dati analizzati.

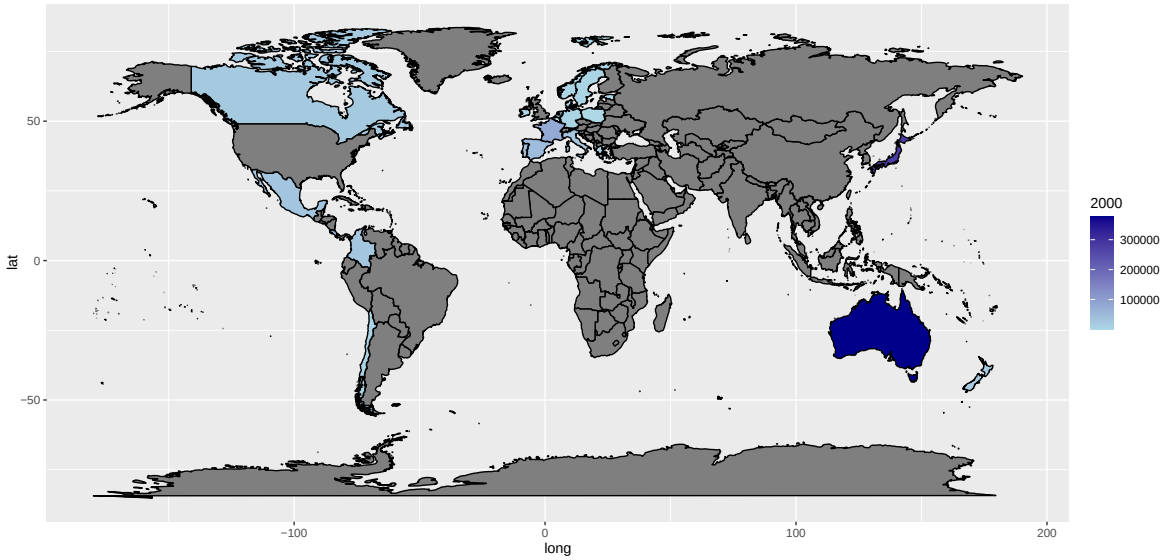

Figura 2.24: Mappa anno 2000

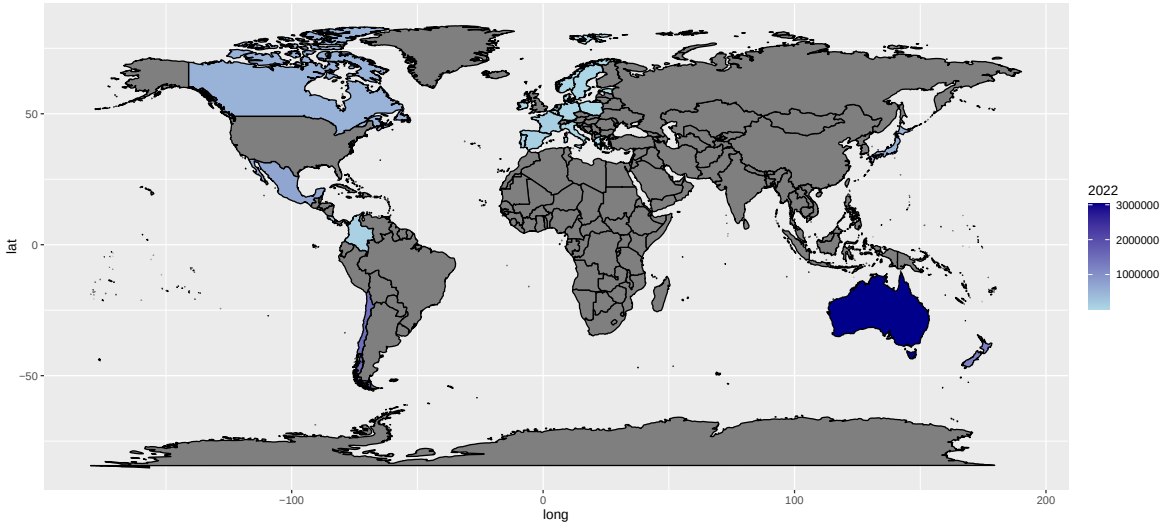

Figura 2.25: Mappa anno 2022

2.4 Distribuzione di frequenza

Un metodo comune per organizzare i dati è utilizzare la distribuzione di frequenza, che consiste nel calcolo della frequenza assoluta e relativa delle diverse modalità che una variabile X può presentare. Questo approccio fornisce una visione immediata dell'intero set di dati, aiutando a valutare la significatività delle diverse osservazioni e la loro distribuzione sulla scala considerata.

Prima di costruire una tabella di frequenza, è essenziale comprendere la natura dei dati.

Nel nostro caso, trattandosi di dati quantitativi vasti, la scelta più appropriata è suddividerli in **classi** in quanto è difficile che si verifichino molteplici occorrenze dello stesso valore.

La scelta sulle classi è stata fatta dunque nel seguente modo: 0-500, 500-5000, 5000-50000, 50000-500000, 500000-5000000.

La **frequenza assoluta** rappresenta il numero di volte che un determinato valore appare nell'insieme di dati.

La **frequenza relativa** invece rappresenta la percentuale di volte che un determinato valore appare nel campione rispetto al totale degli elementi.

Le frequenze assolute sono state calcolate e rappresentate attraverso **barplot**, mentre le frequenze relative sono state calcolate e visualizzate tramite **grafici a torta**

per ogni anno, considerando gli intervalli di superficie delle aree marine protette in chilometri quadrati.

2.4.1 Frequenza Assoluta e Relativa anno 2000

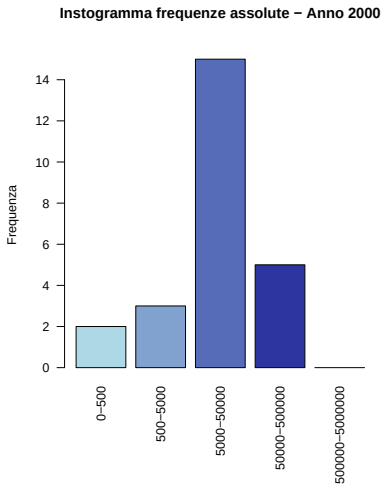

Figura 2.26: Barplot frequenza assoluta anno 2000

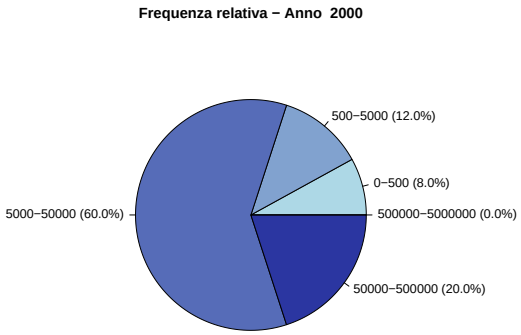

Figura 2.27: Grafico a torta frequenza relativa anno 2000

L'esame dei grafici nella Figura 2.26 e 2.27 rivela che l'intervallo di superficie **più frequente** è compreso tra **5000-50000** chilometri quadrati (sqkm), coinvolgendo 15 su 25 nazioni e rappresentando il 60% del totale. L'intervallo **minimo** è da **500000-5000000** con nessuna nazione all'interno di questo range. L'intervallo 0-500 registra due nazioni : il **Belgio** e la **Polonia**.

2.4.2 Frequenza Assoluta e Relativa dal 2001 al 2003

L'analisi della frequenza assoluta e relativa delle superfici delle aree marine protette, rappresentata nell'istogramma e nel grafico a torta nelle Figure 2.26 e 2.27 del 2000, ha rilevato che le distribuzioni sono rimaste sostanzialmente **invariate** fino al 2003. La continuità dei pattern osservati nei grafici di riferimento suggerisce una stabilità nella distribuzione delle superfici delle aree marine protette durante il periodo considerato.

2.4.3 Frequenza Assoluta e Relativa anno 2004

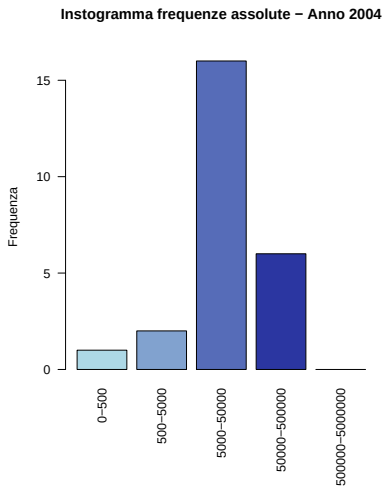

Figura 2.28: Barplot frequenza assoluta anno 2004

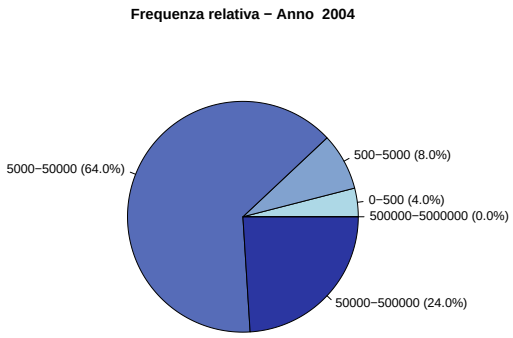

Figura 2.29: Grafico a torta frequenza relativa anno 2004

In Figura 2.28 è possibile notare l'**incremento** dell'intevallo **5000-50000** sqkm di una nazione ossia della **Polonia** e l'**incremento** dell'intervallo da 50000-500000 di **6 nazioni** rispetto gli anni precedenti (5 su 25, 20%).

Dall'esame del barplot in Figura 2.28 e del grafico a torta in Figura 2.29, emerge un **decremento** di 1 nell'intervallo di superficie compreso tra 500-5000 chilometri quadrati (sqkm). Tale diminuzione è riflessa nella frequenza relativa, con una **riduzione** del 4% rispetto ai dati precedenti e nell'**aumento** della porzione 50000-500000.

2.4.4 Frequenza Assoluta e Relativa anno 2005

Dall'osservazione dell'istogramma nella Figura 2.30, si evidenzia una diminuzione nell'intervallo di superficie compresa tra 500 e 5000 chilometri quadrati (sqkm) per una nazione nel 2005. Questa diminuzione è chiaramente rappresentata dalla redistribuzione di questa nazione nell'**intervallo successivo** di superficie, ovvero da 5000-50000 sqkm, come evidenziato nel grafico a torta associato. La nazione in questione è la **Norvegia**.

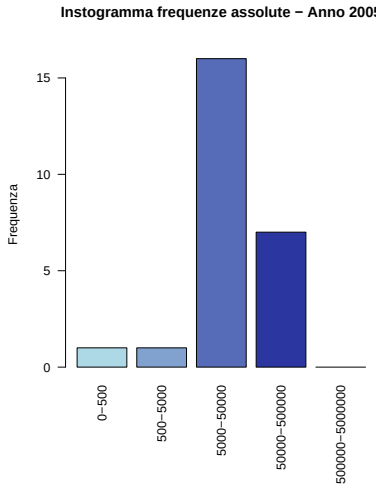

Figura 2.30: Barplot frequenza assoluta anno 2005

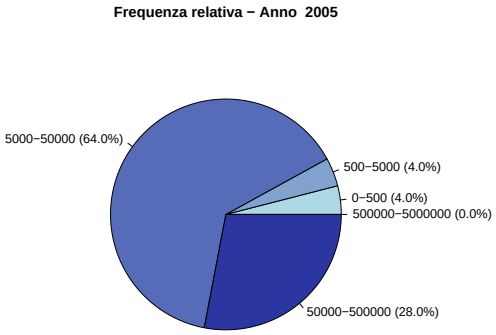

Figura 2.31: Grafico a torta frequenza relativa anno 2005

2.4.5 Frequenza Assoluta e Relativa dal 2005 al 2008

Durante l'analisi dei dati relativi al periodo compreso tra il 2005 e il 2008, è emerso che non si sono verificate variazioni significative nelle frequenze assolute e relative delle superfici delle aree marine protette rispetto l'anno 2004. L'esame generale della distribuzione delle dimensioni di tali aree, nel contesto di diverse nazioni, **non mostra cambiamenti rilevanti** durante il suddetto intervallo di tempo. Questa stabilità è indicata dalla continuità delle frequenze nel periodo considerato, suggerendo che dal 2004 al 2008 non sono intervenute alterazioni sostanziali nella configurazione delle dimensioni delle aree marine protette a livello complessivo.

2.4.6 Frequenza Assoluta e Relativa anno 2009 e 2010

L'esame dell'istogramma in Figura 2.32 rivela un significativo cambiamento nella distribuzione delle dimensioni delle aree marine protette. In particolare, l'intervallo di superficie compreso tra 500-5000 chilometri quadrati non presenta nazioni rappresentate, come evidenziato anche nel grafico a torta associato in Figura 2.33 con una percentuale di **0%**. Questo indica un **ampliamento** delle aree marine protette da parte della nazione in esame, che è ora inclusa nell'intervallo successivo.È degno di nota il fatto che due nazioni rientrino nell'intervallo compreso tra 500000 e 5000000,

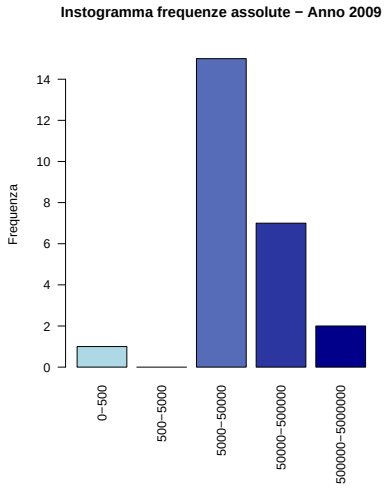

Figura 2.32: Barplot frequenza assoluta anno 2009

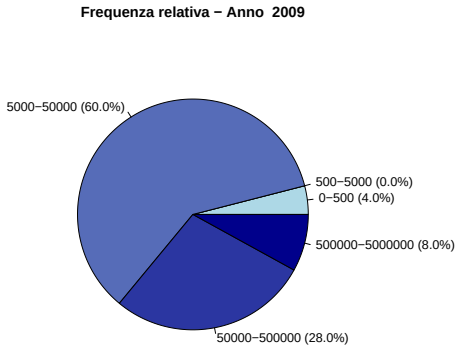

Figura 2.33: Grafico a torta frequenza relativa anno 2009

rappresentando l'8% del totale.

Nel 2010 non sono stati rilevati ulteriori cambiamenti significativi rispetto i grafici in questione.

2.4.7 Frequenza Assoluta e Relativa anno 2011

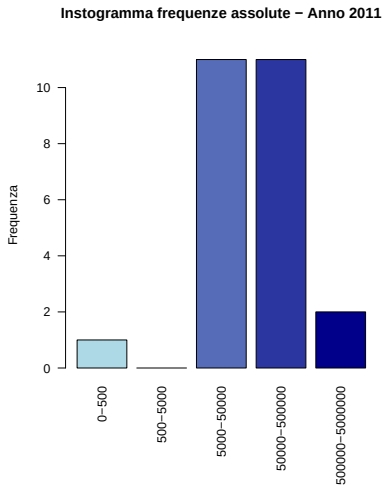

Figura 2.34: Barplot frequenza assoluta anno 2011

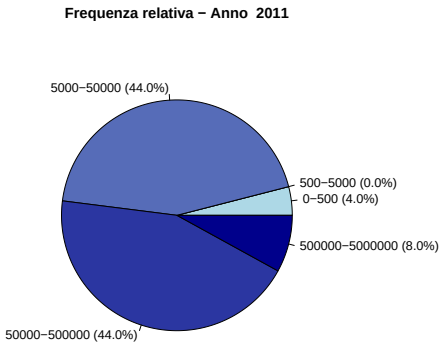

Figura 2.35: Grafico a torta frequenza relativa anno 2011

L'esame dell'barplot per la frequenza assoluta, rappresentato nella Figura 2.34,

evidenzia un incremento significativo nell'intervallo di superficie compreso tra 50000-500000 chilometri quadrati rispetto all'anno precedente. In particolare, si osserva un aumento di uno nell'istogramma, indicando che una nazione ha ampliato notevolmente le sue aree marine protette rientrando in quest'intervallo.

Il grafico a torta per la frequenza relativa, visualizzato in Figura 2.35, riflette tale aumento in maniera proporzionale, rappresentando ora il 47.8% rispetto al 30.4% dell'anno precedente.

2.4.8 Frequenza Assoluta e Relativa dal 2012 al 2022

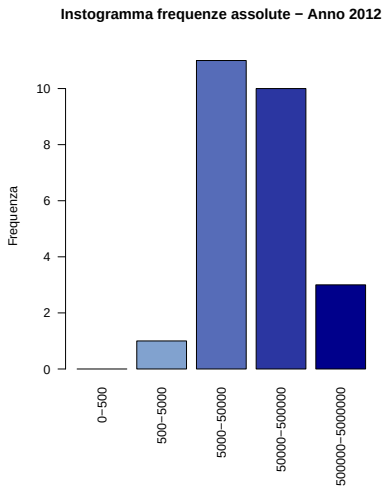

Figura 2.36: Barplot frequenza assoluta anno 2012

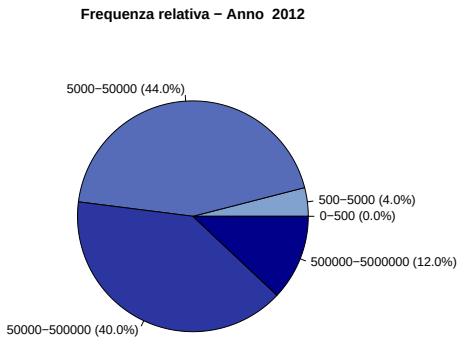

Figura 2.37: Grafico a torta frequenza relativa anno 2012

Durante il periodo dal 2012 al 2022, sono emersi **significativi cambiamenti** nella distribuzione delle nazioni rispetto agli intervalli specifici. In particolare, si è notato un spostamento di nazioni dall'intervallo di occorrenze compreso tra 50000-500000 a quello tra 500000-5000000.

Questi cambiamenti sono chiaramente riscontrabili sia nell'istogramma delle frequenze assolute delle Figure 2.36 e 2.38 che nel grafico a torta delle frequenze relative nelle Figure 2.37 e 2.39. Nel grafico a torta, si è passati dal 40% al 25% nell'intervallo 50000-500000, indicando una diminuzione percentuale. Al contempo, si è osservato un aumento dal 12% al 28% nell'intervallo da 500000-5000000, rappresentando un aumento significativo in termini percentuali. Le nazioni facenti parte di questo ultimo

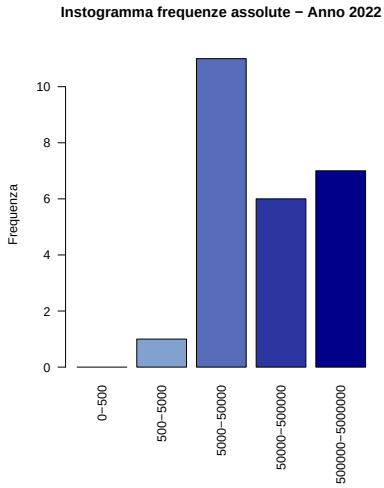

Figura 2.38: Barplot frequenza assoluta anno 2022

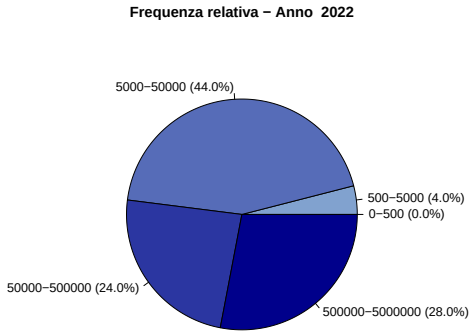

Figura 2.39: Grafico a torta frequenza relativa anno 2022

intervallo sono: **Australia, Canada, Cile, Giappone, Messico, Nuova Zelanda e gli Stati Uniti**.

2.5 Boxplot

Consideriamo un campione $(x_1, ..., x_n)$ dei valori assunti da una variabile quantitativa X. Per capire come sono distribuiti questi valori, ordiniamo la lista dal più piccolo al più grande. Ora, possiamo identificare alcuni dati importanti utilizzando la funzione quantile().

- Il primo quartile (**Q1**) è il valore sotto il quale si trova il **25%** dei dati, mentre il **75%** è sopra.
- La mediana (**Q2**) è il valore al **centro** del nostro set di dati, con il 50% dei dati sopra e il 50% sotto.
- Il terzo quartile (**Q3**) è il valore sotto il quale si trova il **75%** dei dati, mentre il **25%** è sopra.
- I valori **Q0** e **Q4** rappresentano il **minimo** e il **massimo** del nostro set di dati.

Per visualizzare queste informazioni, possiamo creare un boxplot, chiamato anche scatola con baffi.

Questo disegno ha una scatola con estremi Q1 e Q3, tagliata da una linea orizzontale in corrispondenza di Q2 (la **mediana**).

Ci sono anche due linee orizzontali chiamate baffi, dove il **baffo inferiore** è legato al valore più piccolo che è maggiore o uguale a $Q1 - 1.5 * (Q3 - Q1)$, e il **baffo superiore** è legato al valore più grande che è minore o uguale a $Q3 + 1.5 * (Q3 - Q1)$.

La distanza tra Q1 e Q3 è chiamata **intervallo interquartile**, che ci dà un'idea della dispersione dei dati. Se tutti i dati rientrano nell'intervallo $(Q1 - 1.5 * (Q3 - Q1)$, $Q3 + 1.5 * (Q3 - Q1))$, i baffi si estendono fino ai valori minimo e massimo del nostro set di dati. Altrimenti, i valori al di fuori di questo intervallo vengono mostrati come punti, chiamati valori anomali o **outlier**.

2.5.1 Boxplot anno 2000

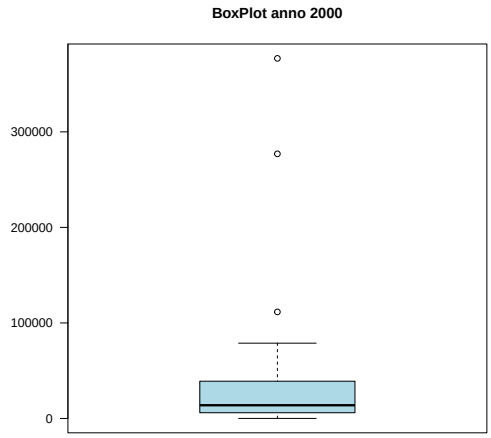

Figura 2.40: Boxplot anno 2000

Analizzando il boxplot rappresentato nella Figura 2.40 relativo al 2000, possiamo osservare la presenza di **tre** outlier, individuati come valori che cadono al di fuori dei baffi del boxplot. **Australia** (376896), **Giappone** (276987) e **Stati Uniti** (111539). Inoltre, notiamo che la forma della distribuzione è asimmetrica, poiché la distanza tra il secondo quartile (Q2) e il terzo quartile (Q3) è significativamente **maggiore** rispetto alla distanza tra il secondo quartile (Q2) e il primo quartile (Q1). Questa **asimmetria nella distribuzione** dei dati è indicativa di una dispersione non uniforme intorno alla mediana.

2.5.2 Boxplot anni dal 2001 al 2006

Dal 2001 al 2006, i boxplot mostrano una distribuzione dei dati che è simile a quella osservata nel 2000 in Figura 2.40. Ciò suggerisce una certa **stabilità** o consistenza nelle caratteristiche statistiche del campione durante questo periodo.

2.5.3 Boxplot anni dal 2007 al 2017

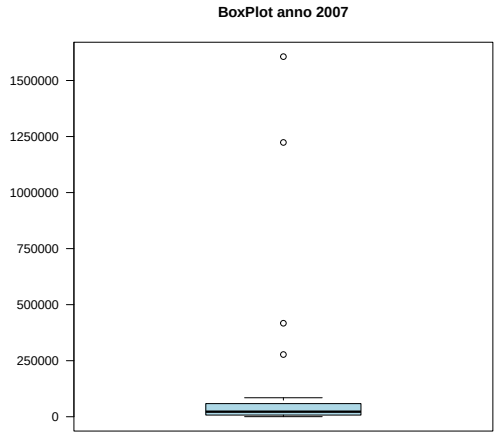

Figura 2.41: Boxplot anno 2007

Il boxplot in Figura 2.41 rivela una distribuzione compatta dei dati nel 2007, come evidenziato dalla breve distanza tra Q1 (primo quartile) e Q3 (terzo quartile). Tuttavia, sono presenti **quattro** outlier, individuati come valori notevolmente distanti dalla maggior parte del campione, un numero superiore rispetto agli anni precedenti. Questi valori anomali sono situati rispettivamente a 277508 (**Giappone**) 417116 (**Australia**), 1223604 (**Nuova Zelanda**) e 1606470 (**Stati Uniti**) , mostrando una marcata deviazione rispetto alla mediana (Q2) e alla maggior parte del dataset.

La Nuova Zelanda si è attivata nel 2007 dal punto di vista ambientale nell'oceano Pacifico meridionale, essendo firmataria di diversi importanti accordi sulla conservazione degli habitat marini. Fra gli accordi ambientali ratificati vi sono poi quelli relativi al **Protocollo Ambientale Antartico**, alla biodiversità, al cambiamento del clima (**Protocollo di Kyoto**), alla **Convenzione sul commercio internazionale di specie in pericolo** (CITES), alle modifiche dell'ambiente, ai rifiuti tossici (Basilea),

allo scarico dei rifiuti in mare, all'eliminazione dei test nucleari, al **Trattato per il legname tropicale**, 28% (Ramsar) ed alla caccia alla balena [7].

Il boxplot Figura 2.41 fino al 2017, suggerisce una **crescita costante** e mantiene una distribuzione relativamente stabile dei dati nel tempo.

2.5.4 Boxplot anni dal 2017 al 2022

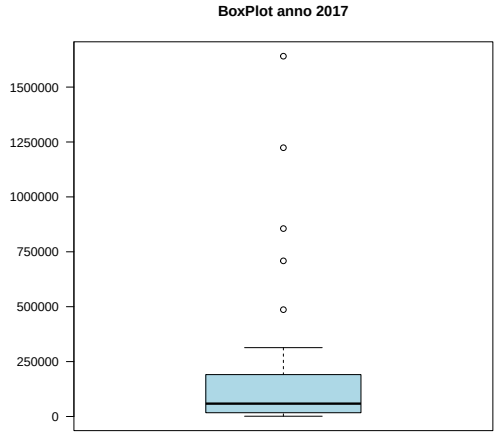

Figura 2.42: Boxplot anno 2017

Analizzando il Boxplot rappresentato nella Figura 2.42, osserviamo che fino al 2017 la distribuzione dei dati mostra una **tendenza crescente**. Tuttavia, è importante notare che nel 2017 si verificano cambiamenti significativi nei parametri del Boxplot. In particolare, la distanza tra Q2 e Q3 è **notevolmente aumentata** rispetto alla distanza tra Q2 e Q1, indicando una **maggiore presenza di valori più elevati** nella distribuzione. Osservando il Boxplot in Figura2.42 come rappresentativo, possiamo dire che fino al 2022 la distribuzione dei dati segue una c**rescita costante** nel corso degli anni.

2.5.5 Confronto Boxplot anni 2000 e 2022

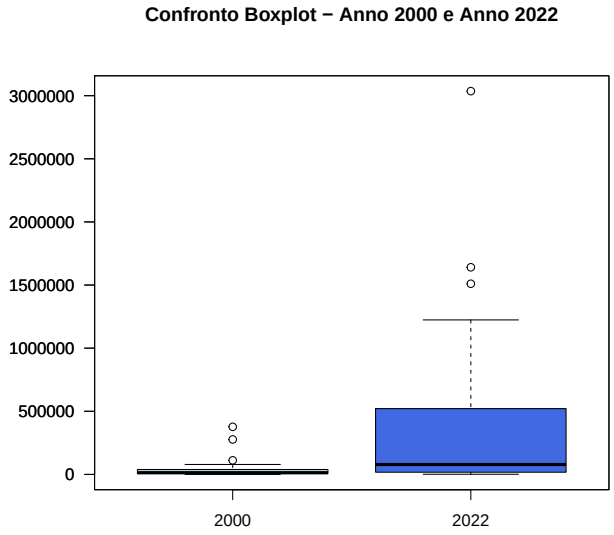

Figura 2.43: Confronto Boxplot anni 2000 e 2022

Comparando i boxplot del 2000 e del 2022 in Figura 2.43, emergono alcune differenze significative. Nel boxplot del 2000 come in quello del 2022, osserviamo la presenza di **tre** outlier, indicativi di dati eccezionali. Nel 2022, la distribuzione dei dati mostra una **tendenza crescente nettamente evidente**, con una maggiore distanza tra Q2 e Q3 rispetto al 2000. Il confronto diretto tra il 2000 e il 2022 conferma la distanza crescente negli anni tra Q2 e Q3 e suggerisce un **incremento generale**, mostrando una maggiore presenza di valori più elevati nel 2022.

2.6 Diagramma di Pareto

Il diagramma di Pareto si presenta sotto forma di un grafico a barre, in cui le barre rappresentano le diverse variabili in ordine decrescente di importanza. La variabile più significativa occupa la posizione più alta sulla barra, seguita da altre variabili in **sequenza decrescente**. Questa rappresentazione visiva facilita l'identificazione delle variabili più rilevanti e consente di concentrare gli sforzi sugli elementi chiave che influenzano i risultati del fenomeno in esame. Nell'ambito di questa analisi statistica, il diagramma di Pareto sarà impiegato tenendo conto delle percentuali relative osservate nella sezione precedente (Sezione 2.4).

2.6.1 Diagramma di Pareto anno 2000

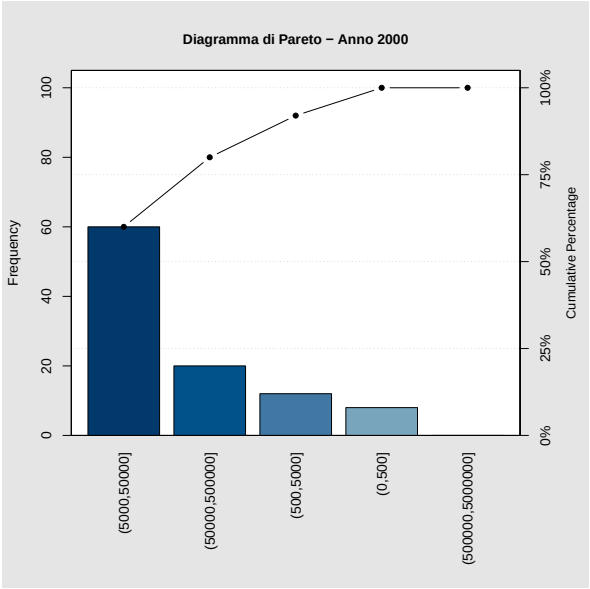

Figura 2.44: Diagramma di Pareto anno 2000

Nel diagramma di Pareto in Figura 2.44 relativo alle dimensioni delle nazioni, emerge chiaramente che l'**intervallo più significativo** nel 2000 è compreso tra **5000-50000** sqkm, rappresentando il 60% del totale delle nazioni. Questo indica che la maggior parte dei paesi ha una superficie compresa in questo intervallo, indicando una distribuzione concentrata di dimensioni territoriali.

Al contrario, gli intervalli **meno significativi** sono quelli da **0-500** e da **500-5000** sqkm, che insieme, costituiscono il 20% delle nazioni. Questi intervalli, sebbene

rappresentino una parte relativamente piccola del totale, indicano la presenza di nazioni con dimensioni molto ridotte.

È interessante notare che non vi è alcun impatto, o **zero presenza di nazioni**, nell'intervallo da **500000-5000000** sqkm.

2.6.2 Diagramma di Pareto anni dal 2001 al 2006

Nel periodo compreso tra il 2001 e il 2006, è interessante notare che i diagrammi di Pareto mostrano una notevole somiglianza rispetto a quello del 2000 in Figura 2.44. Le distribuzioni delle frequenze relative nelle categorie analizzate sembrano mantenere una **coerenza nel tempo** durante questo intervallo di anni.

2.6.3 Diagramma di Pareto anni dal 2007 al 2017

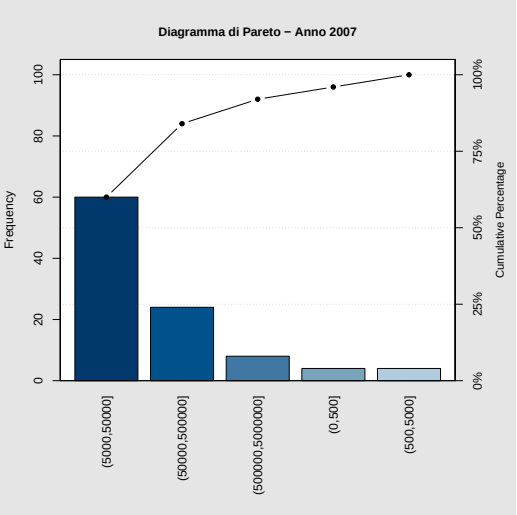

Figura 2.45: Diagramma di Pareto anno 2007

Analizzando attentamente il diagramma di Pareto in Figura 2.45, si nota una certa **continuità nelle tendenze** rispetto agli anni precedenti. Come in passato, l'intervallo più rappresentativo è ancora da 5000-50000 sqkm, evidenziando la predominanza di nazioni con dimensioni comprese in questo range. L'intervallo da 50000-500000 sqkm, che passa dal 20% nel 2000 al 24% nel 2007. Questo indica un aumento del numero di nazioni che rientrano in questa fascia di dimensioni, suggerendo una dinamica territoriale in evoluzione. Allo stesso tempo, due intervalli mantengono una costante

presenza nel tempo, ciascuno contribuendo al 4% del totale: da 0-500 sqkm e da 500-5000 sqkm. Un aspetto degno di nota è l'**incremento** osservato nell'intervallo da 500000-5000000 sqkm, passando dallo **0%** nel 2000 all'attuale **8%**, suggerendo una dinamica territoriale in evoluzione. Il Diagramma di Pareto Figura 2.45 fino al 2017, suggerisce una crescita costante e mantiene una distribuzione relativamente stabile dei dati nel tempo.

2.6.4 Diagramma di Pareto anni dal 2017 al 2021

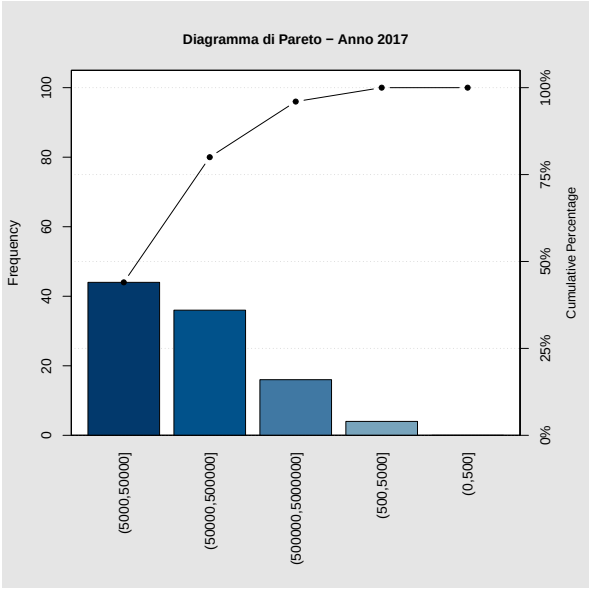

Figura 2.46: Diagramma di Pareto anno 2017

Il Diagramma di Pareto evidenzia che l'**intervallo più ampio**, da **5000-50000** sqkm, continua a essere preponderante, rappresentando il 44% del totale. Questa costante predominanza suggerisce una **stabilità** nella distribuzione delle dimensioni delle nazioni in questa fascia nel corso del tempo. Un notevole incremento è osservato nell'intervallo da **50000-500000** sqkm, che rappresenta ora il 36% del totale. Questo aumento significativo rispetto agli anni precedenti indica una tendenza verso la presenza di un numero maggiore di nazioni con dimensioni comprese in questa fascia territoriale. Anche l'intervallo da 500000-5000000 sqkm registra **un incremento**, raggiungendo il 16% rispetto al 2007, quando era al 8%. Al contrario, l'intervallo da **0-500** sqkm mostra una presenza del **0%**.

2.6.5 Diagramma di Pareto anno 2022

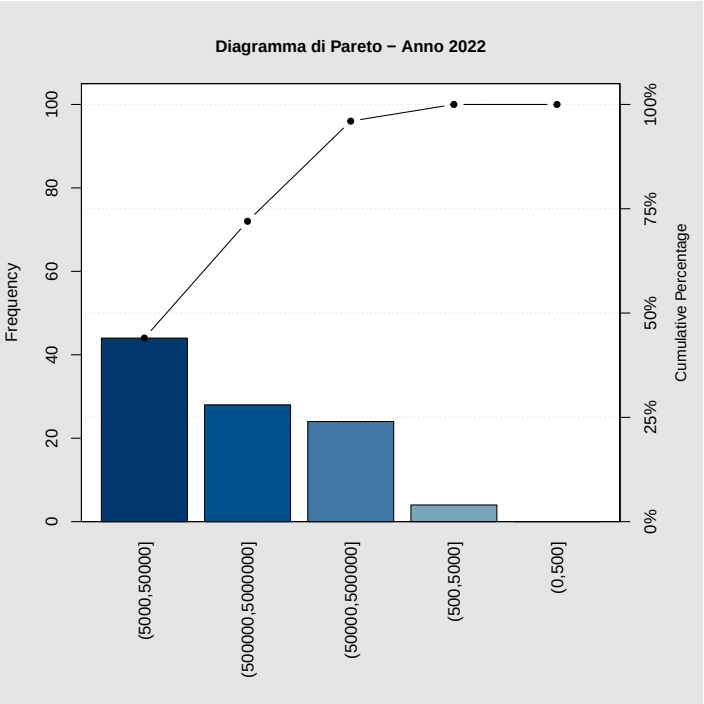

Figura 2.47: Diagramma di Pareto anno 2022

L'analisi dell'ultimo anno in Figura 2.47 disponibile mostra una **costante predominanza** dell'intervallo da **5000-50000** sqkm, che conferma il 44% del totale, mantenendo una continuità con gli anni precedenti. Gli intervalli meno significativi, da **0-500** sqkm e da **500-5000** sqkm, mantengono la **stessa proporzione** degli anni precedenti, con 0% e 4% rispettivamente. Un notevole **aumento** è osservato nell'intervallo da **500000-5000000** sqkm, passando dal 16% degli anni precedenti al 24%. Questo indica un incremento significativo nella presenza di nazioni con dimensioni estese, riflettendo possibili cambiamenti nelle dinamiche territoriali globali. D'altra parte, si registra una **diminuzione** nell'intervallo da **50000-500000** sqkm, passando dal **36%** al **24%**. Questa variazione può indicare una tendenza verso una riduzione della presenza di nazioni con dimensioni intermedie rispetto agli anni precedenti.

CAPITOLO 3

Statistica Descrittiva

La statistica descrittiva univariata si occupa di analizzare la distribuzione di una singola variabile. Questo processo include la valutazione di indicatori centrali come la **media, mediana e moda**, insieme a quelli non centrali, come **quartili, decili e percentili**. Inoltre, vengono considerati gli indicatori di dispersione, tra cui **varianza, deviazione standard e coefficiente di varianza**, che quantificano quanto i dati si discostano dalla media. La forma della distribuzione viene analizzata tramite indicatori di **skewness e curtosi**, che forniscono informazioni sulla simmetria e sulla forma delle code della distribuzione.

3.1 Media Compionaria

La media campionaria viene ottenuta calcolando la media aritmetica di un insieme di valori. Essa corrisponde alla **somma** di tutti i valori divisa **per il numero totale di osservazioni**, fornendo un indicatore del valore medio nella distribuzione. Tuttavia, è importante interpretare con cautela la media, poiché può essere sensibile a valori particolarmente estremi, noti come valori anomali. Questi valori possono avere un impatto significativo sulla media, rendendola suscettibile a variazioni improvvise.

Di seguito, saranno riportate le medie campionarie per ogni anno, offrendo una panoramica del valore medio della popolazione in ciascun anno.

Analizzando la tabella in Figura 3.1 si osserva un **andamento crescente**, indicato da medie annuali in costante aumento. Dal **2001 al 2006**, le medie mantengono un **livello simile**, indicando una certa stabilità. Tuttavia, a partire **dal 2007**, si verifica un **notevole incremento**, con le medie che superano consistentemente i valori precedenti. Il periodo dal **2016 al 2018** testimonia un notevole picco, riflettendo un'**accelerazione nella crescita** della grandezza in questione. Nel **2019 e 2020**, le medie rimangono elevate, confermando un **trend positivo**. Infine, nel **2021 e 2022**, non si osservano significative variazioni rispetto all'anno precedente, attestando la **stabilità** raggiunta dopo un periodo di crescita sostenuta.

	2000	2001	2002	2003	2004	2005
MEDIA	48100.56	51594.52	52591.36	52851.56	54442.76	57006.44

	2006	2007	2008	2009	2010	2011
MEDIA	117249.96	162840.80	164111.04	166045.76	172572.28	174669.36

	2012	2013	2014	2015	2016	2017
MEDIA	191073.44	193075.92	196884.64	197209.16	233625.16	248472.68

	2018	2019	2020	2021	2022
MEDIA	372461.80	396971.84	406778.52	406778.52	411129.68

Tabella 3.1: Media campionaria dal 2000 al 2022

3.2 Mediana

Si definisce mediana campionaria il valore che **bipartisce i dati** in due gruppi di **uguale numerosità** di un certo insieme dei valori ordianti, in maniera tale che lo stesso numero di dati cada sia a sinistra che a destra della mediana stessa.

Nel periodo dal **2000 al 2005**, si osserva un **aumento graduale delle mediane**. Successivamente, si evidenzia un notevole incremento nei successivi anni, con **picchi nel 2010** (37505) e nel **2011** (51082). Da qui in poi, le mediane mantengono valori elevati, raggiungendo il massimo nel 2022 con un valore di 78358. Questo andamento suggerisce una crescita significativa dei dati nel periodo considerato, con particolare enfasi sui notevoli aumenti registrati negli **anni successivi al 2005**, e in particolare nel **decennio successivo al 2010**. La tabella in Figura 3.2 fornisce una panoramica chiara e ordinata delle mediane nel corso degli anni.

	2000	2001	2002	2003	2004	2005
MEDIANA	13845	16108	17255	17256	17469	17764

	2006	2007	2008	2009	2010	2011
MEDIANA	17783	22525	25010	25062	37505	51082

	2012	2013	2014	2015	2016	2017
MEDIANA	51088	51322	51322	56240	58571	58577

	2018	2019	2020	2021	2022
MEDIANA	58577	58577	59239	59239	78358

Tabella 3.2: Mediana dal 2000 al 2022

3.2.1 Confronto tra media e mediana

Il confronto tra media e mediana nella Tabella 3.3 offre un quadro approfondito sulla distribuzione e l'andamento temporale dei dati. L'aumento costante della media nel corso degli anni suggerisce una crescita globale dei valori, con picchi pronunciati, soprattutto dopo il 2010. Tuttavia, l'analisi della mediana fornisce un'immagine più equilibrata, indicando che **la metà centrale dei dati è meno influenzata da variazioni estreme.**

L'osservazione di picchi nella media, non sempre rispecchiati dalla mediana, suggerisce la **presenza di valori estremi in specifici anni che influenzano la media**. Questo può essere interpretato come periodi di significativo aumento dei dati che non sono uniformemente distribuiti nel campione. La stabilità della mediana suggerisce che, nonostante queste fluttuazioni, la distribuzione centrale dei dati rimane relativamente costante.

	2000	2001	2002	2003	2004	2005
MEDIA	48100.56	51594.52	52591.36	52851.56	54442.76	57006.44
MEDIANA	13845	16108	17255	17256	17469	17764

	2006	2007	2008	2009	2010	2011
MEDIA	117249.96	162840.80	164111.04	166045.76	172572.28	174669.36
MEDIANA	17783	22525	25010	25062	37505	51082

	2012	2013	2014	2015	2016	2017
MEDIA	191073.44	193075.92	196884.64	197209.16	233625.16	248472.68
MEDIANA	51088	51322	51322	56240	58571	58577

	2018	2019	2020	2021	2022
MEDIA	372461.80	396971.84	406778.52	406778.52	411129.68
MEDIANA	58577	58577	59239	59239	78358

Tabella 3.3: Confronto tra Media e Mediana dal 2000 al 2022

3.3 Moda

La moda rappresenta il **valore che si presenta con la maggiore frequenza** (assoluta o relativa) in un campione. Nel caso in cui ci siano più valori con la stessa frequenza massima, ciascuno di essi viene considerato come valore modale.

Va notato che, a differenza della media e della mediana, la moda può essere calcolata anche quando si tratta di dati qualitativi.

Nel caso di qust'analisi statistica il calcolo della moda **non fornisce alcun informazione utile** in quanto è improbabile avere più occorrenze dello stesso valore nel corso dello stesso anno.

3.4 Varianza, deviazione standard e coefficiente di variazione

Gli indici di posizione forniscono informazioni importanti, ma non considerano la variabilità dei dati. Anche se media e mediana campionaria possono coincidere, la diversa dispersione dei dati è cruciale.

Nel nostro caso, la notevole discordanza tra media e mediana in ogni anno indica già una conoscenza dell'**assimetria** e della **variabilità dei dati**.

	2000	2010	2022
MEDIA	48100.56	172572.28	411129.68
MEDIANA	13845	37505	78358
DEVIAZIONE STANDARD	89286.23	390750.76	727531.53

Tabella 3.4: Confronto tra Media, Mediana, Deviazione Standard e varianza anni 2000,2010 e 2022

Gli indici di dispersione calcolati nella Tabella 3.4 fanno emerge chiaramente una notevole variabilità rispetto alla media, sottolineando la presenza di valori che si **discostano** significativamente dalla tendenza centrale rappresentata dalla media e

dalla mediana. È interessante notare come la deviazione standard sia stata preferita rispetto alla varianza nell' analisi.

Questa scelta è motivata dal fatto che la varianza può essere fortemente influenzata dalla scala dei dati, mentre la deviazione standard offre un **indice più robusto e interpretabile**, permettendo di valutare in modo più accurato la dispersione dei dati indipendentemente dalla scala di misura.

3.5 Forma della distribuzione di frequenza

Gli indici trattati fino ad ora ci hanno già suggerito un'asimmetria della distribuzione dei dati considerati. Tuttatia, esistono degli indici che permettono di misurare precisamente la simmetria della funzione di distribuzione e la piccatezza: skewness e curtosi campionaria.

3.6 Skewness

La skewness misura l'**asimmetria** della distribuzione dei dati rispetto alla media.

- Se la skewness è **zero**, la distribuzione è simmetrica rispetto alla media.
- Se la skewness è **positiva**, la coda della distribuzione è più lunga verso **destra**.
- Se la skewness è **negativa**, la coda della distribuzione è più lunga verso **sinistra**.

	2000	2010	2022
SKEWNESS	2.627725	2.727824	2.178419

Tabella 3.5: Skewness anni 2000,2010 e 2022

Come è possibile osservare nella Tabella 3.5 tutti e tre gli anni presentano valori positivi, indicando una **forte asimmetria verso destra**.

Nel 2000, il valore di 2.627725 suggerisce una distribuzione con una coda più lunga sulla destra, caratterizzata dalla presenza di valori superiori alla media.

La stessa tendenza si osserva nel 2010 con una skewness di 2.727824, e sebbene il 2022 mostri una leggera diminuzione a 2.178419, rimane evidente una distribuzione asimmetrica verso destra.

Questi risultati indicano la presenza di fattori che contribuiscono a **valori notevolmente superiori alla media**, influenzando la forma complessiva della distribuzione.

3.7 Curtosi

La curtosi misura quanto le code di una distribuzione differiscono da quelle di una distribuzione normale.

Una distribuzione normale è caratterizzata da $\beta_2 = 3$ e indice di curtosi $\gamma_2 = 0$.

In particolare:

- Se $\beta_2 < 3$ e $\gamma_2 < 0$, la distribuzione è definita **platicurtica**, indicando che è più **piatta** di una normale.
- Se $\beta_2 > 3$ e $\gamma_2 > 0$, la distribuzione è definita **leptocurtica**, indicando che è **più appuntita** di una normale.
- Se $\beta_2 = 3$ e $\gamma_2 = 0$, la distribuzione è definita **normocurtica**, seguendo la **curva** di una normale.

	2000	2006	2022
CURTOSI	9.898337	19.408528	8.056695

Tabella 3.6: Curtosi anni 2000,2010 e 2022

In Tabella 3.6, è possibile notare come nel 2000, il valore di curtosi elevato (9.898337) suggerisce una distribuzione con code molto pesanti e un **picco accentuato**, indicando una concentrazione significativa dei dati intorno alla media.

Nel 2006, la curtosi raggiunge un valore ancora più elevato (19.408528), indicando una distribuzione **estremamente appuntita con code molto pesanti**, segnalando una concentrazione ancora più marcata dei dati vicino alla media.

Nel 2022, pur mantenendo una distribuzione **leptocurtica** con un valore di curtosi pari a 8.056695, si osserva una **leggera diminuzione** rispetto al 2006.

Complessivamente, questi dati suggeriscono che la distribuzione dei dati in questi anni è caratterizzata da una concentrazione notevole intorno alla media, con code della distribuzione **più pesanti** rispetto a una distribuzione normale

CAPITOLO 4

Statistica descrittiva bivariata

La statistica bivariata rappresenta il settore della statistica dedicato ai metodi grafici e statistici volti a descrivere le **relazioni tra due variabili quantitative**. L'analisi sarà condotta considerando gli anni estremi del dataset in esame: il 2000 e il 2022. Si opterà per l'utilizzo del **coefficiente di correlazione campionario** al posto della **covarianza campionaria** per valutare le possibili relazioni tra le due distribuzioni di dati, poiché il coefficiente di correlazione è una misura normalizzata. Il valore della covarianza infatti, dipende dalle unità di misura delle variabili coinvolte. Di conseguenza, può essere difficile interpretare la forza della relazione e confrontarla tra diverse analisi.

4.1 Coefficiente di correlazione campionario

Il coefficiente di correlazione campionario costituisce una misura quantitativa che valuta l'**intensità del legame lineare** tra le variabili considerate. Tale coefficiente fornisce indicazioni sulla **disposizione dei punti rispetto a una retta interpolante**, evidenziando se esista una retta che allinea tutti i punti. I suoi valori rientrano nell'intervallo **[-1, 1]**.

- Se il coefficiente è **>** 0, sottolinea una relazione positiva, indicando che all'aumentare di una variabile, l'altra **aumenta proporzionalmente**.
- Se il coefficiente è **<** 0, denota una correlazione negativa, suggerendo che all'aumentare di una variabile, l'altra **diminuisce proporzionalmente**.
- Se il coefficiente è **=** 0, indica l'**assenza di una correlazione lineare**.

Per quanto riguarda il coefficiente di correlazione campionario tra gli anni 2000 e 2022, il valore ottenuto è **0.7034462**, molto vicino a 1.

Tale risultato indica una **buona correlazione positiva**, suggerendo una relazione lineare positiva tra le due variabili. Ciò implica che, in questo contesto, l'incremento di una variabile è seguito proporzionalmente dall'incremento dell'altra, manifestando una tendenza lineare.

Tale relazione può essere accuratamente approssimata da una **retta di regressione lineare** che riflette l'allineamento dei dati.

```
## Coefficiente di correllazione tra gli anni 2000 e 2022:
## [1] 0.7034462
```

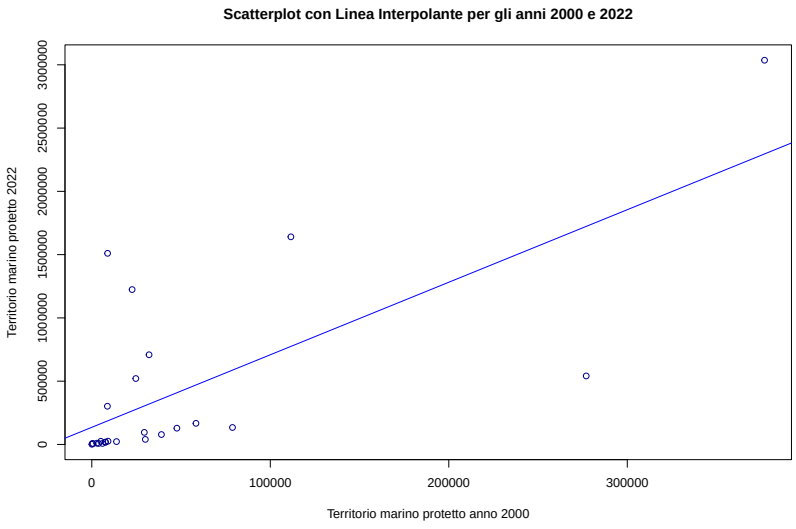

Figura 4.1: Scatterplot con linea interpolare per gli anni 2000 e 2022

In Figura 4.1 è possibile osservare lo **scatterplot** applicato alle due distribuzioni, con la retta di regressione visibile in colore blu. Nello scatterplot, è possibile

vedere una **tendenza positiva**, ovvero una disposizione generale dei punti in direzione ascendente da sinistra a destra. Osservandolo, è possibile notare una notevole **consistenza nella parte inferiore sinistra**, dove la maggior parte dei punti sembra essere concentrata. Questa densità suggerisce una relazione coerente in quella regione, indicando che la **variazione delle variabili** coinvolte è relativamente **stabile e prevedibile**. Allo stesso tempo, la presenza di outlier evidenti nella parte superiore attira l'attenzione. Questi punti singolari si discostano notevolmente dalla tendenza principale indicando **casi speciali o condizioni eccezionali** che hanno un impatto significativo sulla relazione tra le variabili.

4.1.1 Regressione lineare semplice

Nel contesto della regressione lineare, le connessioni tra le variabili vengono modellate tramite funzioni di **predizione lineari**, le cui variabili del modello sono stimate attraverso i dati. Questi modelli sono comunemente noti come modelli lineari. Il modello di regressione lineare semplice può essere rappresentato dall'**equazione di una retta**, la quale si adatta meglio alla distribuzione dei dati nello scatterplot rispetto ad altre possibili rette. Questo modello è spesso impiegato per interpretare, descrivere o anticipare trend futuri basandosi sulla relazione tra una variabile dipendente, indicata come Y, e una variabile indipendente, indicata come X.

L'equazione del modello lineare si esprime come segue: $Y = \alpha + \beta * X$

Dove:

- α rappresenta l'**intercetta**.
- β è il **coefficiente angolare**, indicante la pendenza della retta.

Una volta individuati i valori di α e β, è sufficiente sostituire la variabile X nell'equazione per ottenere il valore predetto di Y.

In sintesi, il modello di regressione lineare fornisce uno strumento per comprendere e utilizzare la relazione tra le variabili coinvolte, contribuendo a interpretare o prevedere il **comportamento futuro della variabile dipendente sulla base dei dati osservati**.

```
## Valori alpha e beta:
## [1] 135422.358113      5.731894
```

Esaminando la distribuzione dei dati nel 2000 come variabile X e la distribuzione dei dati nel 2022 come variabile Y, osserviamo che l'**intercetta** (α) della retta di regressione è stimata a **135422.358113**, indicando il valore previsto di Y quando la variabile X è zero. Nel frattempo, il **coefficiente angolare** (β) è valutato a **5.731894**, suggerendo che per ogni unità di variazione nella distribuzione del 2000 (X), ci si aspetta un **aumento** di circa 5.73 unità nella distribuzione del 2022 (Y), mantenendo costanti tutte le altre variabili.

Lo stesso risultato ci viene restituito dalla funzione $lm(y\ x)$ fornita da R in cui gli argomenti indicano che Y dipende da X.

```
## Call:
## lm(formula = sea22[["2022"]] ~ sea22[["2000"]])

## Coefficients:
##    (Intercept)  sea22[["2000"]]
##     135422.358            5.732
```

L'equazione è quindi :

$Y = 135422.358113 + 5.731894 * X$

4.1.2 Residui

Dopo aver identificato la retta di regressione, è possibile valutare la **discrepanza tra i valori osservati e quelli stimati**. Ciò è dovuto al fatto che i valori predetti dalla retta di regressione non sempre coincidono esattamente con quelli reali, generando delle discrepanze.

I **residui** forniscono un'**indicazione** di quanto i valori osservati si discostino dai valori predetti attraverso la retta di regressione. Essi rappresentano le **differenze tra i valori reali e quelli stimati**, contribuendo a evidenziare la precisione del modello e a individuare eventuali pattern o deviazioni significative nella relazione tra le variabili considerate.

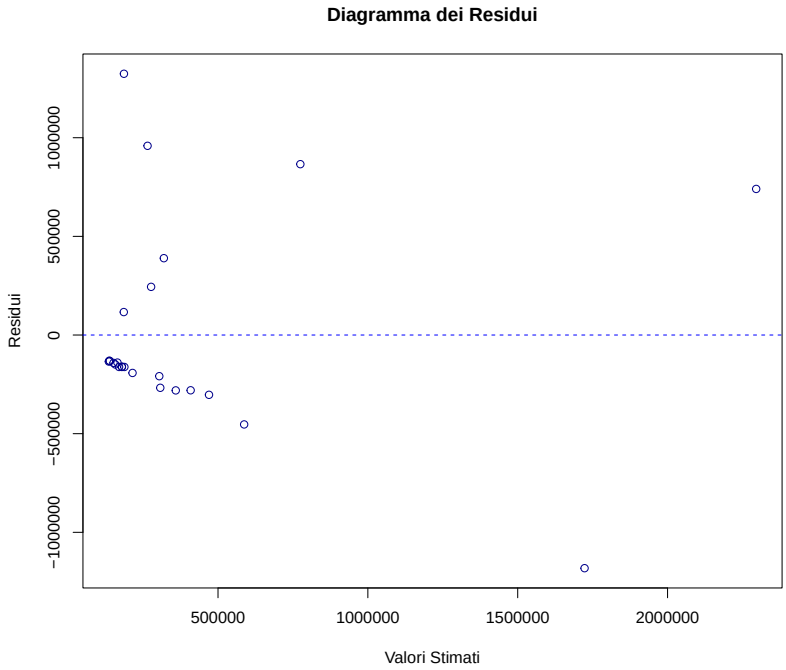

Figura 4.2: Diagramma dei residui

In Figura 4.2, sembra che la maggior parte dei residui si concentri in una **gamma più stretta**, indicando una certa **coerenza** nella discrepanza tra i dati osservati e quelli previsti dal modello. Tuttavia, è evidente la presenza di alcuni valori residui estremi come -1181742.5 e 1324150.4.

Il valore predetto che si avvina di più a quello reale è quello della **Colombia** con 95430 .

Il residuo maggiore con 1324150.4 è quello della **Nuova Zelanda**.

CAPITOLO 5

Analisi dei cluster

L'analisi dei cluster rappresenta un metodo che consente di raggruppare **singole entità** appartenenti a un **insieme** più grande in sottoinsiemi noti come cluster. Tale raggruppamento è basato sulla somiglianza, mirando a far sì che gli elementi all'interno dello **stesso cluster** siano il **più simili possibile**, mentre quelli in **cluster differenti** siano il **più possibile dissimili**.

Per valutare il grado di somiglianza tra due elementi, è possibile utilizzare metriche come i coefficienti di similarità o misure di distanza. I coefficienti di similarità presentano valori compresi tra 0 e 1, mentre le misure di distanza possono assumere qualsiasi valore maggiore o uguale a 0. Esistono due categorie principali di approcci all'analisi dei cluster: quelli **gerarchici e quelli non gerarchici**.

5.0.1 Clustering gerarchico

Il clustering **gerarchico** rappresenta un metodo di raggruppamento che **evita** la necessità di **determinare in anticipo il numero di cluster** o di scegliere parametri per la loro automatica determinazione. Offre una visione completa dell'insieme in termini di distanza o similarità, ma presenta l'inconveniente di non poter riassegnare gli individui già classificati a un livello precedente dell'analisi.

L'obiettivo finale dei metodi gerarchici consiste nel conseguire una sequenza di partizioni che possono essere visualizzate mediante una struttura ad albero nota come **dendrogramma**. In tale rappresentazione, gli **ordini verticali indicano i livelli di distanza**, mentre gli **ordini orizzontali riportano gli individui singoli**. Il dendrogramma fornisce una panoramica completa della struttura dell'insieme in termini delle misure di distanza tra gli individui.

5.0.2 Scelta delle metriche e dendogramma

Il set di dati considerato, come precedentemente indicato, include una nazione con un territorio marino protetto molto vasto nel 2022: gli Stati Uniti. Questa nazione costituisce un punto atipico all'interno della distribuzione dei dati e, insieme ad altre nazioni che presentano valori elevati di popolazione, contribuisce a rendere la distribuzione dei dati fortemente asimmetrica. Di conseguenza, è stata adottata la metrica di **Canberra** per calcolare le misure di distanza, poiché risulta essere **meno sensibile all'asimmetria delle distribuzioni di dati** e alla presenza di eventuali valori anomali. Inoltre, per il metodo gerarchico agglomerativo, è stato scelto il metodo del **legame completo** in quanto rappresenta un approccio standard. In Figura 5.1 è possibile osservare il Dendogramma generato.

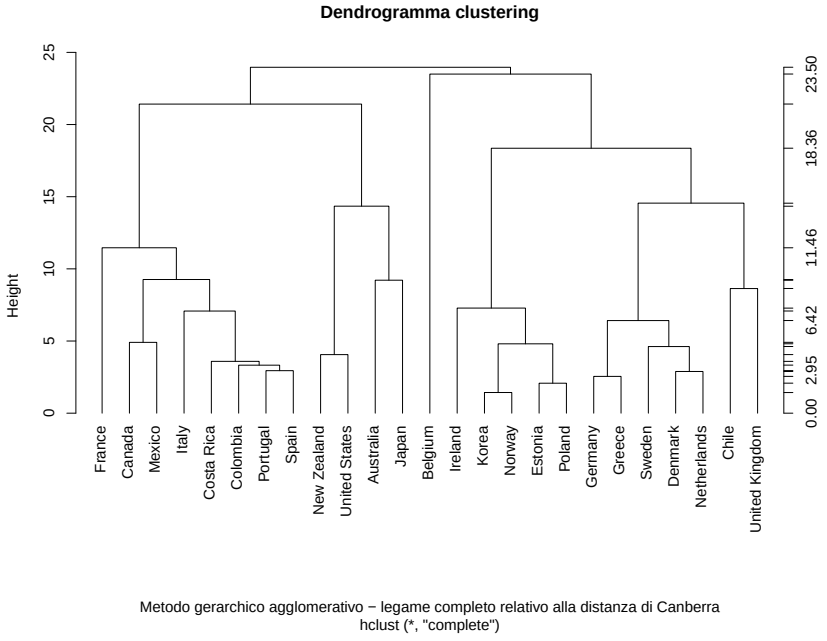

Figura 5.1: Dendogramma clastering

5.0.3 Screeplot

Al fine di analizzare la struttura di clustering e decidere il numero appropriato di cluster da utilizzare, è stato utilizzato lo Screeplot. Il grafico visualizza la "scarpata" o il "gomito" nella curva delle altezze delle fusioni nel dendrogramma, aiutando a determinare il **numero di cluster**. In Figura 5.2 è possibile osservare sull'asse delle ordinate i numeri di gruppi ottenibili con il clustering gerarchico, mentre sull'asse delle ascisse le distanze a cui avvengono le successive aggregazioni tra i gruppi. Se, nella transizione da k gruppi a k-1 gruppi, si osserva un notevole aumento della distanza di aggregazione, è suggerito effettuare il taglio del dendrogramma mantenendo k gruppi. In questo caso, sembra che tra le altezze 4 e 8 ci sia un cambiamento nella pendenza della curva. Il numero di cluster consigliato, cioè dove vi è la **maggiore distanza di aggregazione**, risulta essere **5**.

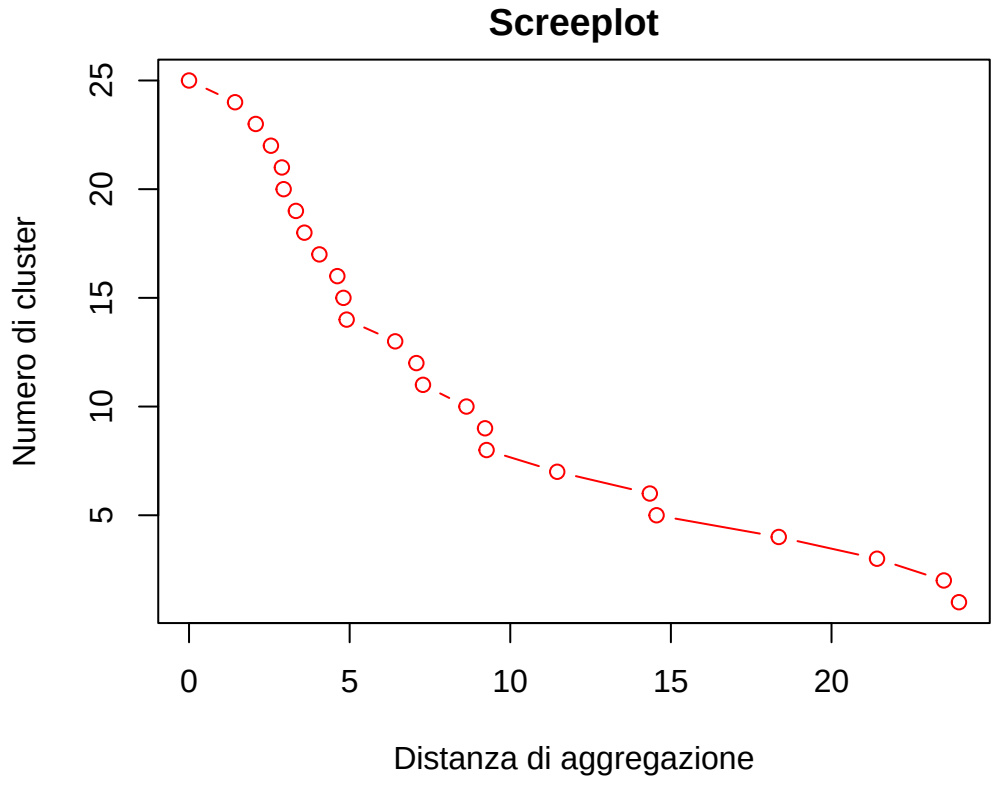

Figura 5.2: Screeplot cluster gerarchico

In Figura 5.3 è possibile osservare il dendogramma con **cinque** cluster definiti e racchiusi tra rettangoli rossi. Le nazioni che si trovano all'interno dello **stesso rettangolo rosso** fanno parte dello **stesso cluster**.

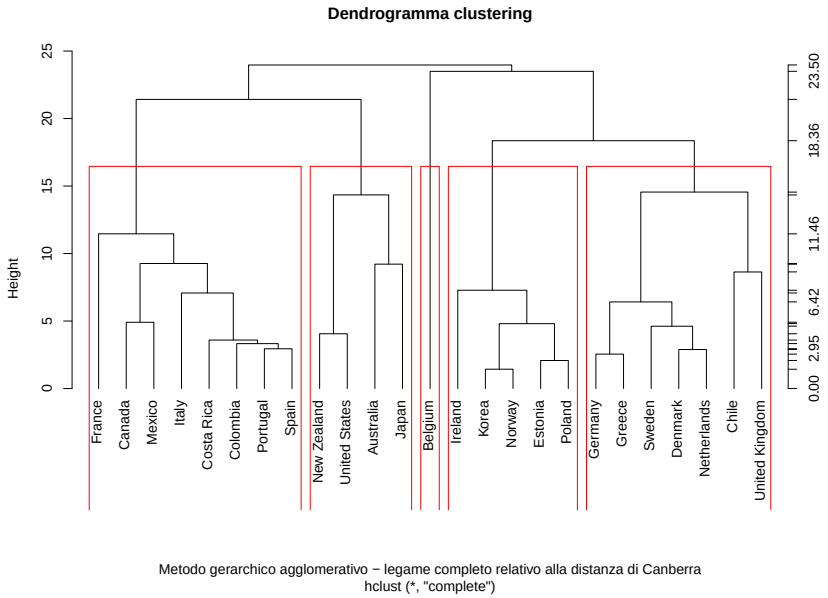

Figura 5.3: Dendogramma con 5 cluster

5.0.4 Misure di non omogeneità statistiche e individuazione dell'outlier

Quello che vogliamo ottenere è che gli individui appartenenti allo stesso cluster siano quanto più possibile omogenei tra loro e il più possibile differenti da quelli appartenenti agli altri cluster individuati. Quello che facciamo allora è considerare una misura di non omogeneità interna ai cluster (**within**) e una misura di non omogeneità tra cluster (**between**). I risultati sono:

- La misura di non omogeneità **totale** risulta essere di **110107317977796**.
- La misura di non omogeneità statistica del primo gruppo è di **31123262232802.8**.
- La misura di non omogeneità statistica del secondo gruppo è **NA** (in quanto è presente solo il Belgio).
- La misura di non omogeneità statistica del terzo gruppo è di **2665345445946.5**.

- La misura di non omogeneità statistica del quarto gruppo è di **9300300755509.1**.
- La misura di non omogeneità statistica del quinto gruppo è di **192654725.6**.

La misura di non omogeneità statistica all'interno dei cinque gruppi (**within**) è : **43089101088984**

La misura di non omogeneità tra i cluster (**between**) è : **67018216888812**

In conclusione, la misura di non omogeneità all'interno dei gruppi (within) è quindi **piccola** rispetto la misura di non omogeneità tra i cluster (between).

Inoltre: $trB/trHI$ = **0.60866269490214**

Questo indica che i cluster hanno una rilevanza significativa nella spiegazione della variazione complessiva nei dati. La presenza di un valore superiore al 50% suggerisce che la struttura dei cluster ha un impatto sostanziale sulla distribuzione complessiva dei dati, e che le differenze tra i gruppi sono statisticamente significative.

Dendogramma senza outlier

Come chiaramente evidenziato nel diagramma a dispersione presente nella Figura 5.1, il **Belgio** si presenta come un valore anomalo. Nel paragrafo successivo, esamineremo l'andamento della clusterizzazione sia gerarchica che non gerarchica, escludendo il Belgio dalla nostra analisi.

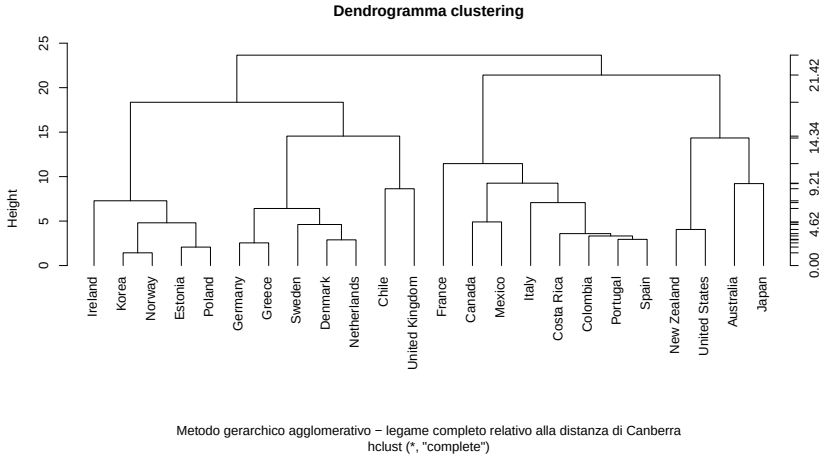

Figura 5.4: Dendogramma senza il Belgio come outlier

Screeplot

Lo screeplot in Figura5.5 ci suggerisce la divisione in k cluster con **k = 4**

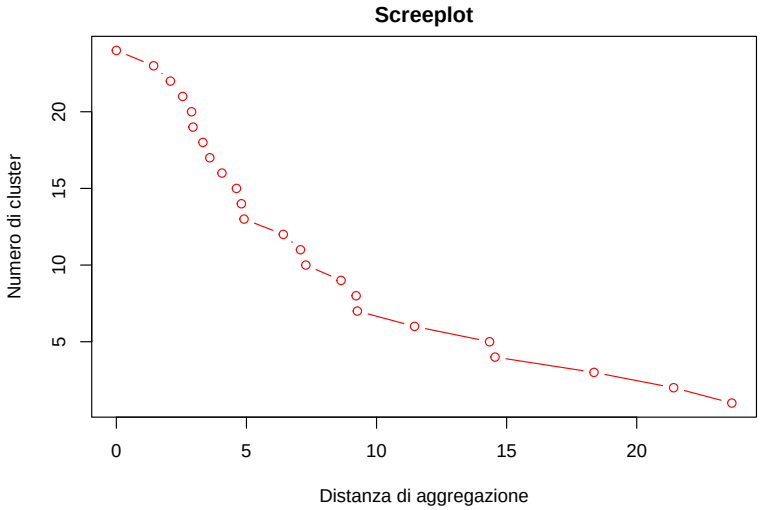

Figura 5.5: Screeplot senza outlier

Misure di non omogeneità

Quello che facciamo allora è considerare una misura di non omogeneità interna ai cluster (within) e una misura di non omogeneità tra cluster (between) con **quattro** cluster.

I risultati sono:

- La misura di non omogeneità totale risulta essere di **96301055423196.5**.
- La misura di non omogeneità statistica del primo gruppo è di **77533041309.8**.
- La misura di non omogeneità statistica del secondo gruppo è di **125058425086.5**.
- La misura di non omogeneità statistica del terzo gruppo è di **122419794670.3**.
- La misura di non omogeneità statistica del quarto gruppo è di **81607827356.5**.

La misura di non omogeneità statistica all'interno dei sei gruppi (within) è : **406619088423.3**

La misura di non omogeneità tra i cluster (between) è : **95894436334773.4**

In conclusione, la misura di non omogeneità all'interno dei gruppi (within) è quindi **piccola** rispetto la misura di non omogeneità tra i cluster (between).

Inoltre: $1 = \frac{trS}{trT} + \frac{trB}{trT}$

406619088423.3/96301055423196.5 + 95894436334773.4/96301055423196.5

0.0042223741644 + 0.9957776258356=1

Questo ulteriore supporto numerico **conferma la validità della decomposizione** della varianza totale e sottolinea l'importanza delle differenze tra i cluster nel contribuire alla variabilità complessiva nei dati analizzati.

Diagramma con cluster evidenziati

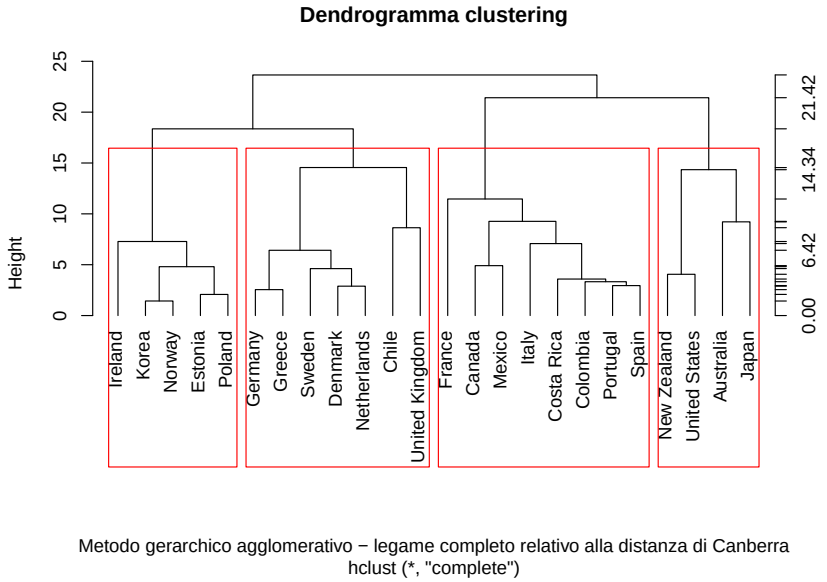

Figura 5.6: Dendogramma con 4 cluster evidenziati

5.0.5 Clustering non gerarchico con il metodo del k-means

Gli algoritmi di tipo **non gerarchico** procedono, data una prima partizione, a **riallocare gli individui nel gruppo con centroide piu' vicino**, fino a che per nessun individuo si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui esso appartiene, il metodo piu' utilizzato prende il nome di **k-means**.

Da segnalare che l'algoritmo potrebbe convergere ad un ottimo locale e non globale, il che significa che se si inizia con un diverso insieme di punti di riferimento si può giungere ad una differente partizione finale.

Inoltre, il metodo del k-means offre tre scelte iniziali:

1. **Scelta casuale** dei punti di riferimento.
2. **Ripetizione** della procedura di scelta casuale dei punti di riferimento.
3. Scelta dei **centroidi** come punto di riferimento.

L'analisi procederà considerando la terza scelta, cio'e la scelta dei centroidi come punto di riferimento.

Risultati ottenuti

- La misura di non omogeneità totale risulta essere di **96301055423196.4**.
- La misura di non omogeneità interna ai cluster (within) è di **6390635796051.97**.
- La misura di non omogeneità tra i cluster (between) è di **89910419627144.4**.

Quindi:
(6390635796051.97/96301055423196.4)+(89910419627144.4/96301055423196.4) = **1.22934**

Questo valore **superiore a 1** suggerisce che la scelta dei centroidi come punto di riferimento potrebbe aver portato a una varianza spiegata superiore al totale della varianza, il che potrebbe essere indicativo di una certa **instabilità nell'algoritmo o della presenza di effetti non desiderati**.

5.0.6 Conclusioni

Come precedentemente specificato i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between). Sono state applicate entrambe le metodologie di clustering: gerarchico e non gerarchico.

La prima **metodologia gerarchica** risulta essere la migliore, infatti confrontando i risultati tra entrambe le metodologie si ha che:

- Il valore di within è minore: **406619088423.3** < **6390635796051.97**
- Il valore di between è maggiore: **95894436334773.4** > **89910419627144.4** La misura di non omogeneità statistica tra i quattro cluster suggerisce una sovrapposizione tra i cluster nel cluster non generico. Infatti risulta : **1** < **1.22934**.

Analizzando i risultati appena presentati, emerge che il **metodo gerarchico** dimostra una tendenza a *minimizzare* la misura di non omogeneità statistica all'interno dei cluster (within) e, simultaneamente, *massimizzare* la misura di non omogeneità statistica tra i gruppi (between).

Questo indica che **il processo gerarchico ha successo nel formare cluster più omogenei internamente e al contempo separarli in maniera più distinta rispetto agli altri gruppi**.

La misura di non omogeneità statistica tra i quattro cluster è pari a 1 nel contesto del clustering generico, suggerendo un'evidente e netta separazione tra i cluster. Un valore di 1 indica che ciascun individuo appartiene esclusivamente a un singolo cluster, senza sovrapposizioni o ambiguità. Questo risultato riflette la **robustezza e l'efficacia del metodo gerarchico** nell'identificare e distinguere pattern distinti all'interno del dataset.

CAPITOLO 6

Inferenza statistica

L'inferenza statistica è una metodologia che consente di generalizzare e **trarre conclusioni** sulla base di dati raccolti da un gruppo rappresentativo (**campione**) di una popolazione più ampia. Cioè, si studia una popolazione descritta da una variabile aleatoria X la cui funzione di distribuzione ha una forma nota ma contiene un parametro non noto. Per ottenere informazioni su questo parametro non noto della popolazione, si può fare uso dell'inferenza statistica considerando un campione estratto dalla popolazione e effettuando su tale campione delle opportune misure.

6.1 Criterio del chi-quadrato

Il test del chi-quadrato viene impiegato per determinare se un determinato campione osservato possa **derivare** da una popolazione descritta da una **variabile aleatoria** X con una **specifica funzione di distribuzione** FX(x).

Analizzando gli scatterplot presenti nella Sezione 2.1, emerge chiaramente che l'andamento dei dati per le diverse nazioni prese in considerazione e la distribuzione temporale di tali dati sono connotati da un pattern a scalini. Questo modello di distribuzione non può essere adeguatamente descritto né da una distribuzione normale, né da una distribuzione di Poisson.

Alla luce di questa peculiarità, l'applicazione del test del chi-quadrato a tali distribuzioni risulta **inappropriata**. Entrambe le distribuzioni, normale e di Poisson, non saranno in grado di adattarsi efficacemente al modello a scalini osservato nei dati.

Da un punto di vista metodologico, è interessante sottolineare che, al momento attuale, la letteratura scientifica **non offre metodologie consolidate e applicabili per il calcolo di una variabile aleatoria che rispecchi adeguatamente il modello a scalini** osservato nei dati analizzati.

6.2 Distribuzione di Bernoulli

Una prova di Bernoulli è un esperimento casuale caratterizzato da **due soli possibili risultati**, interpretabili l'uno come **successo** e l'altro come **insuccesso**.

La variabile aleatoria X che descrive il risultato di una prova di Bernoulli assume soltanto due valori: 1 (indicante il successo) con **probabilità p** e 0 (indicante l'insuccesso) con **probabilità 1 p**.

Nel contesto specifico dell'analisi, la distribuzione di Bernoulli è stata impiegata per modellare se l'area marina protetta dell'Australia ha una alta probabilità di **ridursi** o di **aumentare** in termini di chilometri quadrati.

La necessità di sviluppare questo modello è emersa nel contesto di un articolo che mette in guardia sui potenziali rischi per l'ecosistema marino dell'Australia.

Questa esigenza si è manifestata a seguito di un potenziale avvertimento riguardante il possibile danno inflitto all'arcipelago di Rowley Shoals, che ospita una delle barriere coralline più vitali del mondo, a causa dei test sismici impiegati nella ricerca di gas e petrolio [8].

6.3 Distribuzione binomiale

Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche, ed assumiamo che in ogni prova i risultati di interesse siano sintetizzabili nel verificarsi dei seguenti due eventi necessari ed incompatibili: A (interpretabile come successo) e B (interpretabile come insuccesso),

con P(A) = p (0 < p < 1). Un siffatto esperimento si dice costituito da n prove ripetute indipendenti di Bernoulli.

Nel nostro contesto, abbiamo considerato la variabile aleatoria come il **numero di casi in cui si verifica la crescita dei dati**, rappresentando l'evento A nelle n prove. La variabile aleatoria, denota quindi il numero totale di anni in cui si osserva un incremento nell'estensione dell'area marina protetta dell'Australia. Questa è stata modellata utilizzando una distribuzione **binomiale**, dove **ciascun anno costituisce una prova di Bernoulli**.

La **probabilità di successo**, indicata da p, rappresenta la probabilità di registrare un aumento nell'area marina protetta in un dato anno.

Attraverso la funzione di distribuzione visibile in Figura 6.3, abbiamo calcolato la probabilità cumulativa di avere al massimo il numero osservato di successi nei dati raccolti (Figura 6.2).

In tal modo, siamo stati in grado di valutare statisticamente la frequenza della crescita dell'area marina protetta e comprenderne la probabilità cumulativa rispetto al periodo di osservazione specificato.

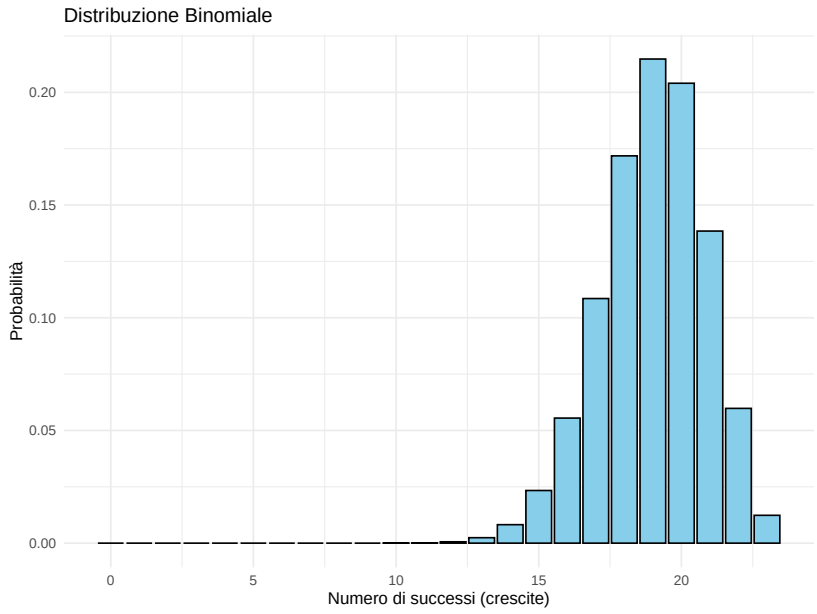

Figura 6.1: Funzione di probabilità binomiale

I valori calcolati indicano che, in media, si osservano **19 anni con crescita** nell'area marina protetta dell'Australia, considerando il periodo di riferimento dei dati forniti.

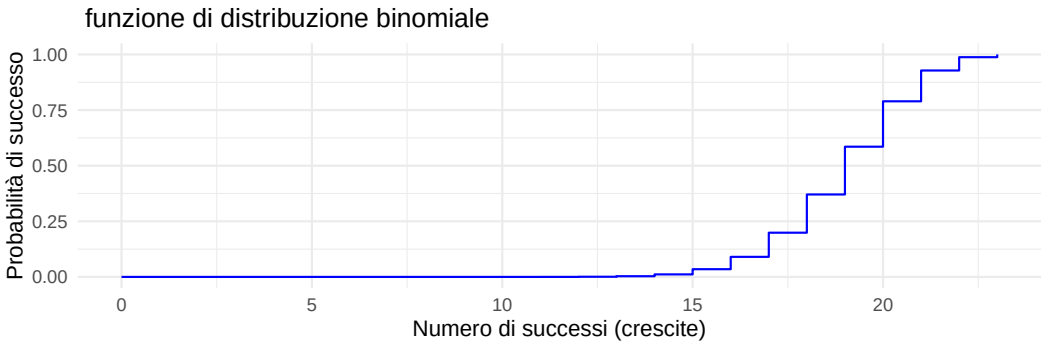

Figura 6.2: Funzione di distribuzione binomiale

La **varianza** di 364.3043 suggerisce una certa variabilità nei numeri di anni di crescita rispetto alla media di 19.

La **deviazione standard** di circa 1.817787 quantifica questa variabilità in modo più interpretabile, indicando quanto gli anni di crescita tendano a deviare dalla media.

Il **coefficiente di variazione** del 9.567297% fornisce una misura percentuale della variabilità relativa alla media. In questo caso, il coefficiente di variazione è **relativamente basso**, indicando che la variabilità nei dati rispetto alla media è moderata.

Per verificare la fattibilità di questo risultato sono stati generati **1000 valori casuali** in linea con la media e la deviazione standard degli incrementi nell'area marina protetta dell'**Australia**.

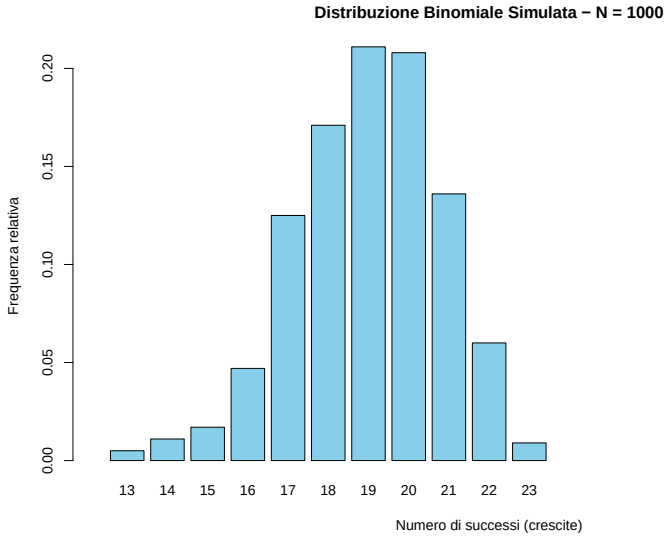

Figura 6.3: Generazione di 100 numeri pseudocasuali mediante la funzione

I risultati ottenuti dalla simulazione con 1000 numeri generati confermano e ampliano le conclusioni derivate dall'analisi dei dati osservati.

La **media stimata** di circa 18.981 anni con crescita nell'area marina protetta dell'Australia **riflette una coerenza** con il valore medio di 19 anni calcolato dai dati effettivi.

La **varianza** della distribuzione simulata, pari a 3.259899, suggerisce una variabilità nei numeri di anni di crescita rispetto alla media, in linea con l'osservazione di una varianza di 364.3043 nei dati reali.

La **deviazione standard** di circa 1.805519 quantifica questa variabilità, indicando quanto gli anni di crescita possano deviare in modo interpretabile dalla media.

Il **coefficiente di variazione simulato**, attestato al 9.512244%, è coerente con il valore del 9.567297% calcolato dai dati osservati.

Questo coefficiente di variazione relativamente basso suggerisce che la variabilità nei dati simulati rispetto alla media è moderata, confermando così l'osservazione fatta sulla distribuzione reale.

In entrambi i casi, la relativa **stabilità della tendenza di crescita** nell'area marina protetta dell'Australia **è supportata sia dai dati osservati che dalla simulazione**, offrendo una prospettiva coerente sulla variabilità e sulla consistenza dei risultati.

Metodo dei momenti

Il valore stimato di 0.966759 con il metodo dei momenti, rappresenta la **probabilità stimata di successo** (crescita dell'area marina protetta) basata sull'**analisi della distribuzione binomiale**. In contesti probabilistici, il valore stimato di probabilità può variare da 0 a 1, dove 0 indica nessuna probabilità di successo e 1 indica certezza di successo.

Intervallo di confidenza

La stima dell'intervallo di confidenza di grado 1 α= 0.95 per il parametro p risulta essere**(0.9428046, 0.9808843)**. Si nota che la stima puntuale del parametro p, ossia $\hat{p} = 0.966759$, è contenuta nell'intervallo.

Verifica delle ipotesi

Si consideri l'ipotesi nulla $H_0 : p \geq 0.9$, che sostiene la crescita o il mantenimento dell'area marina protetta. L'ipotesi alternativa è $H_1 : p < 0.9$, indicando la possibilità di una riduzione dell'area marina protetta.

La decisione finale dipenderà dai risultati del test statistico.

Si desidera verificare quindi l'ipotesi $H_0 : p \geq 0.9$ in alternativa a $H_1 : p < 0.9$ con un livello di significatività $\alpha = 0.05$.

Occorre considerare un **test unilaterale destro**.

In questo caso: il valore critico z associato a un livello di significatività del 5% ($\alpha = 0.05$) è -1.644854.

La statistica del test z è -1.181582.

Poiché la statistica del test è maggiore del valore critico z, **non c'è abbastanza evidenza statistica per rifiutare l'ipotesi nulla** ($p \geq 0.9$) a un livello di significatività del 5%.

In altre parole, con i dati forniti, non puoi affermare in modo significativo che la proporzione sia inferiore a 0.9.

In questo contesto, non c'è abbastanza evidenza statistica per rifiutare H_0 , siamo inclini a mantenere l'idea che la proporzione di crescita (p) sia almeno del **90%**.

Bibliografia

[1] "Sustainable ocean economy," https://stats.oecd.org. (Citato alle pagine iv e 3)

[2] "La grande barriera corallina australiana tra sbiancamento e resilienza agli stress ambientali," https://www.geopop.it/la-grande-barriera-corallina-australia/: :text=Scoperta (Citato a pagina 6)

[3] "Riserve naturali marine in canada," https://lestresorsderable.com/it/blog/post/riserve-naturali-marine-in-canada.html. (Citato a pagina 7)

[4] "Cile, nascono nove aree marine protette lungo 6.400 chilometri di coste," https://www.lifegate.it/cile-nove-aree-marine-protette-6400-chilometri. (Citato a pagina 8)

[5] "Un'area marina protetta di 500.000 km2 tra ecuador, colombia, panama e costa rica," https://greenreport.it/news/aree-protette-e-biodiversita/unarea-marina-protetta-di-500-000-km2-tra-ecuador-colombia-panama-e-costa-rica/. (Citato a pagina 9)

[6] "L'italia, gli stati uniti d'america e il diritto internazionale del mare," http://dspace.unive.it/bitstream/handle/10579/3378/811740-1155924.pdf?sequence=2. (Citato a pagina 17)

[7] "Ambiente neo zelandese," https://www.goaustralia.it/ambiente-neo-zelandese/. (Citato a pagina 28)

[8] "Il 2022 potrebbe iniziare molto male per l'ecosistema marino," https://www.lasvolta.it/314/il-2022-potrebbe-iniziare-molto-male-per-lecosistema-marino. (Citato a pagina 57)