

ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

---

SCUOLA DI SCIENZE

Corso di Laurea in Informatica Per il Management

**Data analytics applicato alle  
conferenze scientifiche:  
una valutazione sperimentale**

**Relatore:**  
**Chiar.mo Prof.**  
**Marco Di Felice**

**Presentata da:**  
**Altea De Giuseppe**

**Sessione II**  
**Anno Accademico 2023/2024**



# Indice

<b>Introduzione</b>	<b>i</b>
<b>1 Stato dell'arte</b>	<b>1</b>
1.1 Dati . . . . .	1
1.2 Data science . . . . .	3
1.3 Data analytics . . . . .	4
1.3.1 Fasi della data analytics . . . . .	5
1.3.2 Esempi di utilizzo . . . . .	7
1.4 Data analytics per analisi della letteratura scientifica . . . . .	8
<b>2 Progettazione</b>	<b>13</b>
2.1 Obiettivi dell'analisi . . . . .	13
2.2 Tecnologie usate . . . . .	14
2.2.1 Python . . . . .	14
2.2.2 Pandas . . . . .	15
2.2.3 Matplotlib . . . . .	16
2.2.4 Numpy . . . . .	16
2.2.5 Seaborn . . . . .	17
2.2.6 WordCloud . . . . .	17
2.2.7 Pycountry . . . . .	17
2.2.8 Plotlyexpress . . . . .	18
2.2.9 Jupyter . . . . .	18

---

<b>3</b>	<b>Implementazione</b>	<b>19</b>
3.1	Raccolta dati . . . . .	19
3.2	Algoritmi . . . . .	21
3.2.1	Cardinalità e Medie . . . . .	21
3.2.2	Distribuzione citazioni . . . . .	22
3.2.3	WordCloud . . . . .	23
3.2.4	Rappresentanze e Mappe . . . . .	24
3.2.5	Riferimenti AI e Blockchain . . . . .	27
3.2.6	Correlazioni . . . . .	28
<b>4</b>	<b>Risultati</b>	<b>33</b>
4.1	Cardinalità . . . . .	33
4.2	Provenienza articoli e autori . . . . .	34
4.3	Citazioni medie . . . . .	42
4.4	WordCloud . . . . .	51
4.5	Riferimenti AI e Blockchain . . . . .	53
4.6	Correlazioni . . . . .	59
<b>5</b>	<b>Conclusioni e Sviluppi Futuri</b>	<b>69</b>
	<b>Appendice</b>	<b>71</b>
5.1	Globecom . . . . .	71
5.2	Infocom . . . . .	109
5.3	Percom . . . . .	139
	<b>Bibliografia</b>	<b>170</b>





# Introduzione

Negli ultimi decenni, l'analisi dei dati si è affermata come uno strumento fondamentale in numerosi ambiti della ricerca scientifica e tecnologica, grazie alla sua capacità di estrarre conoscenze utili da grandi quantità di informazioni. In particolare, nel contesto delle conferenze scientifiche, l'applicazione di tecniche di data analytics potrebbe offrire l'opportunità di individuare tendenze, analizzare la distribuzione delle pubblicazioni e valutare le collaborazioni accademiche. Questo approccio consente di comprendere meglio l'evoluzione di specifiche discipline scientifiche e di supportare i ricercatori nella pianificazione e valutazione dei loro lavori.

Questa tesi si propone di applicare metodologie di analisi dati alle conferenze scientifiche nell'ambito delle reti di comunicazione, un settore in continua evoluzione. Il progetto ha come obiettivo principale quello di analizzare dati relativi a tre conferenze di rilievo internazionale proposte dalla Institute of Electrical and Electronics Engineers <sup>1</sup>, ovvero la Global Communications Conference <sup>2</sup>, la International Conference on Computer Communications <sup>3</sup> e la International Conference on Pervasive Computing and Communications <sup>4</sup>. L'intento è quello di individuare delle correlazioni o tendenze tematiche e valutare metriche come la distribuzione geografica di articoli e autori.

Nel primo capitolo verrà delineato lo stato dell'arte, con una panoramica sulle principali tematiche di data analytics e i loro utilizzi nel contesto delle conferenze

---

<sup>1</sup>IEEE

<sup>2</sup>Globecom

<sup>3</sup>Infocom

<sup>4</sup>Percom

scientifiche, nonché sui temi di ricerca emergenti nell'ambito delle reti di comunicazione, e si vanno a descrivere più dettagliatamente quelle che sono le conferenze trattate.

Nel secondo capitolo saranno descritte le finalità e gli obiettivi dell'analisi, approfondendo, inoltre, le tecnologie scelte e le motivazioni alla base del loro utilizzo.

Per quanto riguarda, poi, la prima fase operativa del progetto, ovvero la raccolta e pulizia dati e la descrizione degli algoritmi usati sarà analizzata nel dettaglio nel capitolo tre che parla appunto dell'implementazione.

Infine nel capitolo "Risultati", nonché l'ultimo, si presenteranno i risultati ottenuti con una rappresentazione grafica che evidenzia i trend più significativi o caratteristiche interessanti.

Questa analisi sperimentale non solo contribuisce a fornire una visione più dettagliata dello stato attuale delle conferenze analizzate, ma rappresenta anche un modello replicabile per future valutazioni di impatto scientifico in altri ambiti di ricerca.

# Capitolo 1

## Stato dell'arte

### 1.1 Dati

I dati sono elementi fondamentali di informazione, costituiti da simboli, che di per sé, non hanno un significato intrinseco e richiedono di essere elaborati per diventare utili. Privati del loro contesto, i dati risultano spesso poco significativi, in quanto non offrono un quadro chiaro o interpretabile della realtà. Per esempio un dato è un marker temporale, un nome o un numero, che però senza un contesto o altri dati collegati, non si può dire a che cosa si riferisca.

Al contrario, un'informazione si ottiene quando i dati vengono organizzati, interpretati e contestualizzati, permettendo di ricavare una conoscenza più chiara e affidabile su fatti, situazioni o fenomeni. Questa distinzione tra dato grezzo e informazione elaborata costituisce la base del processo di analisi e trasformazione che consente di estrarre valore e significato dei dati, rendendoli uno strumento essenziale per la comprensione del mondo e per la presa di decisioni. Per esempio prendendo il marker temporale come dato se ci integriamo un nome e un paese si può intuire che quella sia la data di nascita di una persona e gli altri siano altri dati anagrafici.

Si consideri, inoltre, che i dati hanno anche un valore di tipo economico, infatti come viene evidenziato da un articolo del Fondo Monetario Internazionale, anche le piattaforme o i servizi che possono sembrare gratuiti in realtà si prendono in

autonomia in cambio i dati degli utenti che usufruiscono del servizio. [1]

Negli anni i sistemi in grado di generare dati digitali crescono esponenzialmente tanto da far crescere di conseguenza anche il loro valore di mercato.

A provare la crescita continua dell'uso dei dati, Federica Laricchia, una ricercatrice specializzata in ambito tecnologia e comunicazione, ha condotto una ricerca sul mercato dei dati in Italia ed emerge che nel 2015 il mercato aveva un valore di 790 milioni di euro, mentre nel 2020 ne valeva 1.8 miliardi di euro. [2]

Questi dati poi sono la parte fondamentale dell'analisi dati e di tutti quei campi che utilizzano macchine di apprendimento per andare a sviluppare strumenti o piattaforme che producono output personalizzati.

L'utilizzo dei dati richiede innanzitutto strumenti quali database, ossia tool che contengono i dati in maniera ordinata per logica del loro significato, il data engineer va a creare dei modelli che rappresentano il database per l'organizzazione dei dati, poi crea la vera e propria struttura che li conterrà. Questa struttura è composta da tabelle che sono formate dalle colonne che rappresentano gli attributi ovvero il significato della singola variabile o dato e le righe che rappresentano delle osservazioni, ovvero un'insieme di dati relativo ad una cosa. Riprendendo l'esempio di prima, il nome "Sara" è una variabile o dato, che si troverà nella colonna del "nome", mentre l'intero insieme di dati "Sara", "11/07/2001", "Roma"; è un'osservazione che viene memorizzata come riga.

ID	Nome	Cognome	Età
0	Marco	Rossi	18
1	Luca	Verdi	25
2			

Figura 1.1: Esempio tabella database

## 1.2 Data science

Con il tempo, sempre più dati sono stati creati e studiati, di conseguenza c'è stato bisogno di creare un ambito specifico che si specializzasse nella loro lavorazione e interpretazione, a questo ambito si è dato il nome di: Data Science.

La scienza dei dati comprende principi metodologici, processi e tecniche provenienti da metodologie scientifiche volte a comprendere fenomeni ed estrarre conoscenza dai dati tramite la fase di analisi eseguita da parte di un esperto chiamato "data scientist".

Le metodologie su cui si basa derivano da molti campi di studio tradizionali, come la matematica, la statistica, l'informatica e le scienze sociali; con particolare importanza per la statistica che è fondamentale con le sue formule e calcoli per ottenere risultati rilevanti dalla scienza dei dati.

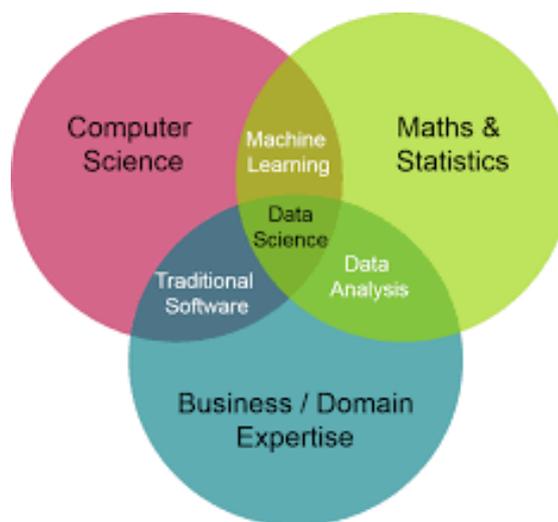


Figura 1.2: Data science e i campi da cui deriva

Il fine della scienza dei dati è quello di migliorare il processo decisionale che risulta essere di fondamentale interesse per le aziende, per migliorare la loro organizzazione e i loro risultati.[3] Le prime volte in cui viene coniato il termine è intorno agli anni '70 da Peter Neur, un informatico danese, che specifica che è "la scienza dei trattamenti dei dati". Ma poichè un po' generica questa definizione, tutt'ora, si prende più in considerazione la definizione di data science scritta nel

paper pubblicato, nel 2001, da William Cleveland, un informatico americano, che sostiene che la data science tocchi anche altri campi come: modelli e metodi per i dati, sistemi computazionali, valutazione dei tool ed altri. [4] Difatti è simile alla definizione data poche righe prima.

### 1.3 Data analytics

La data analytics è una disciplina fondamentale all'interno del più ampio settore della data science e si occupa di elaborare, interpretare e trasformare i dati in informazioni utili e comprensibili. I dati al giorno d'oggi sono molto presenti e soprattutto disponibili e facilmente accessibili grazie all'evoluzione tecnologica e a soluzioni di archiviazione sempre più capienti e poco costose.

Questo incremento è agevolato non solo dalla grande memoria disponibile e dai costi bassi che ha la tecnologia, ma anche dal suo più frequente utilizzo nei vari contesti lavorativi e non, che scelgono sempre di più l'utilizzo di alternative digitali per ottenere insight utili.

Più precisamente la data analytics si occupa di andare a trasformare dei dati grezzi, spesso privi di significato se presi singolarmente, in informazioni strutturate da cui poter trarre delle conclusioni.

Questo è un processo reso possibile dagli avanzamenti di algoritmi, tecniche statistiche e strumenti di machine learning che sono alla base delle innovazioni che stanno rivoluzionando il modo di vivere e lavorare.

Per ottenere risultati migliori e più orientati ad ogni tipo di settore per cui si usa l'analisi dati, è possibile andare a utilizzare le quattro tipologie di analisi, ovvero:

- **Analisi descrittiva**, va ad analizzare quelli che sono i dati passati chiedendosi cosa è successo prima e quindi va poi a descrivere i risultati. Un esempio possono essere gli stati che hanno pubblicato più articoli scientifici nelle conferenze.[5]
- **Analisi diagnostica**, va a cercare il perché si sono verificati determinati eventi o tendenze. Ad esempio perché alcuni stati sono più dominanti di altri.

- Analisi predittiva, utilizza formule e modelli statistici per cercare di predire quello che avverrà in futuro tramite i dati passati. Per esempio si stima l'influenza di un paese in una conferenza futura.
- Analisi prescrittiva, unisce le precedenti analisi per andare a consigliare quelli che dovrebbero essere i passi successivi per andare a migliorare una situazione. Per esempio come dare più visibilità ad un paese che di solito ne ha poca.

Le quattro tipologie di analisi lavorano spesso insieme per permettere di ottenere i risultati adatti a prendere decisioni informate e strutturate.

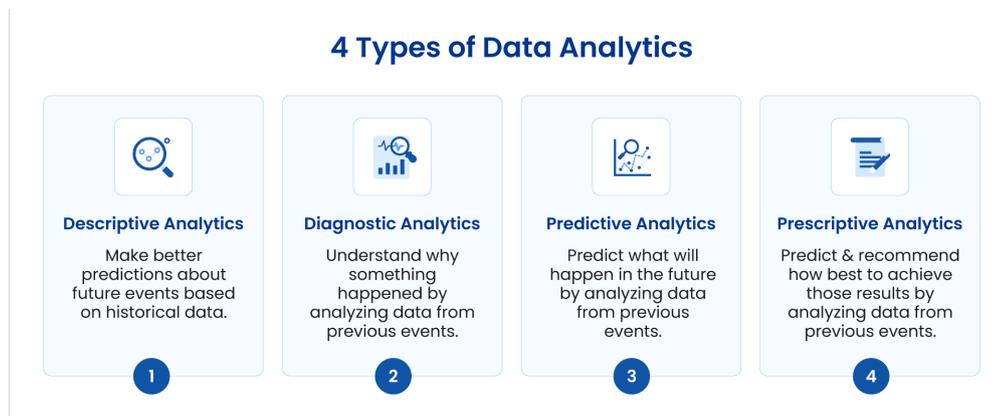


Figura 1.3: Quattro tipi di data analytics

### 1.3.1 Fasi della data analytics

Per poter avere un buon riscontro dall'esplorazione dei dati bisogna seguire cinque fasi.

Che, esse, consistono in:

- raccolta dati,
- storage dei dati,
- elaborazione dei dati,

- pulizia dati,
  
- analisi dei dati.

La prima riguarda la raccolta dei dati dalle fonti di origine, un processo fondamentale per garantire che i dati siano pertinenti e di qualità. Si dedicano a ciò principalmente le aziende o gli enti di competenza che utilizzano strumenti appositi come siti e software. La seconda fase prevede il salvataggio nel database progettato per gestire in modo efficiente le specifiche caratteristiche di quei dati. La scelta del database più adatto è cruciale per ottimizzare i successivi processi di analisi. Successivamente, la terza fa riferimento a come questi dati verranno organizzati per essere facilmente accessibili e interpretabili. Questo include quindi schemi E-R<sup>1</sup> e modelli logici con le varie entità e attributi.

Infine, la quarta e la quinta fase sono quelle di cui ci si occupa anche nel progetto, ovvero la pulizia dei dati, dunque una selezione dei dati più conformi all'analisi, ma anche pulire i dataset da dati errati o nulli. Tale processo, realizzato tramite delle righe di codice che si occupano di ciò, assicura che i dati selezionati siano conformi all'analisi e pronti per l'elaborazione senza la possibilità di compromissione dei risultati. L'analisi dei dati, la quinta fase, va a trasformare i dati in informazioni utilizzabili, grazie ad algoritmi che permettono di estrarre conoscenze e risposte a specifiche domande di ricerca. [6]

---

<sup>1</sup>Entity-Relationship, è un modello teorico per la rappresentazione concettuale e grafica dei dati ad un alto livello di astrazione. È formato dalle entità che sono le tabelle che rappresentano i soggetti del problema, dagli attributi che sono gli oggetti delle entità e dalle relazioni che è il legame tra due o più entità



Figura 1.4: Fasi della data analytics

### 1.3.2 Esempi di utilizzo

Le conclusioni che si traggono dall'analisi dati trovano applicazione in una vasta gamma di settori, rappresentando uno strumento fondamentale per prendere decisioni informate, ottimizzare processi e migliorare le performance in diversi contesti. Per esempio, in ambito commerciale, viene utilizzata per individuare quelle che possono essere le strategie più adatte a favorire la crescita della propria azienda, andando a studiare le preferenze dei consumatori, i trend di mercato e i comportamenti di acquisto.

È altrettanto utile in tutte le scienze come la medicina che viene usata, per esempio, per studiare le cartelle cliniche dei pazienti in modo da scovare pattern o arrivare a diagnosi precoci, o anche in ambiti socio-demografici per comprendere tendenze demografiche supportando decisioni politiche e socio-economiche o ancora nella finanza.

In sintesi è un elemento chiave che permette di guidare il progresso in numerosi ambiti, consentendo alle organizzazioni e professionisti di conoscere informazioni e ottenere soluzioni pratiche che generano un valore aggiunto.

## 1.4 Data analytics per analisi della letteratura scientifica

### Conferenze scientifiche

Le conferenze scientifiche sono un importante canale di scambio fra ricercatori in ambito accademia ed industria. È stato provato diverse volte che la rete di letteratura scientifica è complessa e in continuo sviluppo rendendone la ricerca più difficile. Per questo potrebbe risultare utile andare ad eseguire un'analisi dati che dia una visualizzazione e un aiuto alle persone per comprendere cose complesse su questo argomento. Dato anche il continuo incremento di informazioni a riguardo, così da avere sempre una panoramica sulle dinamiche della letteratura scientifica. [7]

La comunicazione scientifica è un processo che comprende meccanismi e strategie volte a promuovere e diffondere idee scientifiche, sia all'interno della comunità accademica che verso un pubblico più ampio. Si tratta di un'attività essenziale per garantire la circolazione della conoscenza scientifica sulle realtà che ci circonda, sfruttando una vasta gamma di canali e strumenti dedicati alla comunicazione. Questo mira a favorire lo scambio di idee e dialoghi tra ricercatori, enti ed aziende interessate. È utile, quindi, andare a capire le dinamiche che caratterizzano il panorama scientifico, analizzando aspetti come i paesi maggiormente coinvolti, la provenienza accademica o aziendale degli articoli e le aree tematiche di riferimento delle pubblicazioni. [8]

Nella tesi si va ad esaminare quelli che sono i dati provenienti da alcune conferenze scientifiche e gli articoli pubblicati.

Nello specifico si prendono in considerazione tre conferenze promosse dalla IEEE. La IEEE (Institute of Electrical and Electronics Engineers, Inc.) è la più grande

società professionale tecnica del mondo. Pubblica il 30% della letteratura mondiale nei campi dell'ingegneria elettrica ed elettronica e dell'informatica e ha sviluppato quasi 900 standard industriali attivi. [9]



Figura 1.5: Logo IEEE

Nasce nel 1963 dall'unione di due istituzioni precedenti l'AIEE <sup>2</sup> e la IRE <sup>3</sup> ed oggi è composta da 46 società di informatici ed ingegneri di ogni genere. L'organizzazione sponsorizza ogni anno più di 850 conferenze in tutto il mondo. Tra queste ci sono la Global Communications Conference <sup>4</sup>, la International Conference on Computer Communications <sup>5</sup> e la International Conference on Pervasive Computing and Communications <sup>6</sup>, ovvero le tre conferenze che si prenderanno in considerazione da questo punto in poi.

- La GLOBECOM è una delle conferenze maggiori dell'IEEE, un evento a livello mondiale che promuove l'innovazione di quasi tutti gli ambiti della comunicazione. Ogni anno partecipano più di 3000 ricercatori scientifici, ma anche professionisti e i management di industrie del settore. Comprende documenti tecnici, workshop e altro, progettati per l'avanzamento di tecnologie, sistemi e infrastrutture per lo sviluppo delle telecomunicazioni. L'obiettivo è di favorire il progresso delle telecomunicazioni globali, introducendo soluzioni in grado di trasformare il panorama tecnologico e sociale,

---

<sup>2</sup>American Institute of Electric Engineers, 1884

<sup>3</sup>Institute of Radio Engineers, 1912

<sup>4</sup>Globecom

<sup>5</sup>Infocom

<sup>6</sup>Percom

guidando lo sviluppo di sistemi di comunicazione sempre più veloci, accessibili ed efficienti, offrendone un accesso in modo economico e su scala globale. [10]

- La INFOCOM è una conferenza di alto livello sul networking che permette ai ricercatori di presentare idee significative e innovative in questo campo e aree correlate.

La conferenza pone un forte accento sulla ricerca, esplorando temi complessi e fondamentali nel campo del networking, a differenza di eventi come la Globecom che abbracciano un panorama più ampio. Grazie a questo aspetto INFOCOM si è consolidata come essenziale per favorire il progresso tecnologico, stimolare la collaborazione tra esperti e ampliare sempre di più le conoscenze in ambito di reti di comunicazione. [11]

- La PERCOM (International Conference on Pervasive Computing and Communications) è un'altra conferenza proposta dalla IEEE, rappresenta il principale appuntamento accademico annuale nel campo dell'informatica e delle comunicazioni pervasive, costituendo un punto di riferimento per la comunità scientifica e tecnologica globale. Il calcolo pervasivo, ovvero l'integrazione delle capacità computazionali e comunicative in oggetti e ambienti della vita quotidiana, si colloca all'avanguardia nella ricerca sui sistemi mobili e rappresenta un elemento chiave per lo sviluppo di tecnologie innovative. Negli ultimi anni, questa disciplina ha trovato numerose applicazioni concrete in ambito commerciale, grazie ai significativi progressi ottenuti in un'ampia gamma di settori. Tra questi spiccano le reti wireless, il calcolo mobile e distribuito, i sistemi di sensori, l'intelligenza ambientale e la progettazione di dispositivi intelligenti. [12]

Gli ambiti di ricerca e sviluppo che trattano queste conferenze, non solo contribuiscono a migliorare la qualità della vita, ma servono anche per un futuro sempre più interconnesso e tecnologicamente avanzato.

Per cui in un mondo in cui la tecnologia è sempre più avanzata e a farla sviluppa-

re sono proprio queste conferenze scientifiche e tecnologiche è interessante anche capire le caratteristiche e le tendenze della letteratura scientifica. [13]

### **Analisi data applicata alle conferenze scientifiche**

L'analisi dati si applica in molti ambiti, prettamente in quelli in cui il risultato porta ad un miglioramento dell'efficienza del prodotto o servizio per cui si sta utilizzando. Si prendano in considerazione gli esempi di prima, per la medicina si usa per il miglioramento di cure e prestazioni, per il commercio si usa per offrire un servizio migliore ai clienti e così via.

Ma un ambito in cui è ancora poco utilizzata è la letteratura scientifica, nel quale però può essere utile per andare a vedere una possibile correlazione fra variabili per ottenere per esempio più citazioni. Oppure uno stato può andare a vedere se ha una forte influenza nella ricerca scientifica o anche se vengono premiati di più gli articoli proposti da università oppure da aziende.

Al giorno d'oggi sono presenti tecniche che permettono di semplificare e rendere più efficiente l'analisi dati andando a garantire livelli di scalabilità e ripetibilità mai raggiunta prima. Tuttavia, la costruzione di una visione d'insieme rappresenta una sfida tipica della visualizzazione dei dati. Scegliere le fonti più adatte, analizzare e rappresentare le informazioni in modo chiaro e significativo e interpretare correttamente ciò che emerge dal quadro complessivo, sono passaggi fondamentali.[14] Si andrà quindi a spiegare nei prossimi capitoli quelli che sono i passi per ottenere i risultati sperati, tramite illustrazioni, grafici e codici.



# Capitolo 2

## Progettazione

Nel terzo capitolo si spiega il progetto, fornendo una descrizione delle sue finalità oltre che evidenziare quali sono gli obiettivi principali dell'analisi svolta. Vengono, inoltre, approfondite quali tecnologie sono state preferite, spiegandone i motivi e come sono state utilizzate.

Questa tesi si basa sull'analisi dati di conferenze scientifiche, questa analisi è stata eseguita seguendo le fasi precedentemente descritte, ovvero, raccolta dei dati, pulizia, trascrizione in dataframe ed effettiva analisi sia descrittiva che statistica. I dati sono stati raccolti dal sito [Ieee/xplore.org](http://Ieee/xplore.org), ottenendo dei file csv da cui poi sono stati estrapolati i dati e messi in dataframe selezionando solo gli attributi utili all'analisi. Dopo di che si è proceduto con il rispondere agli obiettivi principali che ci si è posti.

### 2.1 Obiettivi dell'analisi

L'obiettivo di questo progetto consiste nell'andare ad analizzare i dati delle conferenze introdotte nel capitolo precedente, per trarne informazioni varie. Alcune di queste informazioni sono innanzitutto le cardinalità dei dataset, quindi il numero di articoli portati alle conferenze, poi per esempio le provenienze sia degli autori sia degli articoli, quali sono gli stati che contribuiscono maggiormente e quali sono gli stati che ottengono più citazioni.

Per quanto riguarda il numero di citazioni, sono state fatte delle considerazioni riguardo alle citazioni medie in base anche ad alcune caratteristiche tipo lo stato di provenienza e il tipo di affiliazione. Oppure anche la distribuzione degli articoli sul numero di citazioni, quindi quanti articoli hanno ottenuto quel numero di citazioni.

Altre informazioni possono riguardare il contenuto dei paper pubblicati, infatti si vedrà uno studio sulla crescita del numero di articoli che fanno riferimento ad argomenti come l'intelligenza artificiale e la blockchain.

Si possono, inoltre, analizzare le correlazioni tra lo stato di provenienza dell'articolo e il numero di citazioni ricevute, oppure tra il tipo di affiliazione dell'autore, accademica, aziendale o mista, e l'impatto dell'articolo in termini di citazioni.

Questo per quanto riguarda le conferenze prese separatamente. Sono state fatte poi delle analisi incrociate che vanno a porsi come obiettivo il confronto tra le conferenze come la media delle citazioni che ottengono gli articoli proposti da esse.

## 2.2 Tecnologie usate

Di seguito vengono descritti gli strumenti, linguaggi e librerie utilizzati nella stesura di questo progetto di Tesi, con un'analisi approfondita delle loro caratteristiche principali. Verranno inoltre illustrati i motivi che hanno portato alla scelta di queste tecnologie piuttosto che di altre, evidenziandone i vantaggi offerti in termini di efficienza, flessibilità e compatibilità con gli obiettivi del progetto. Questo capitolo intende fornire una panoramica chiara e completa delle basi tecnologiche che hanno sostenuto il lavoro svolto.

### 2.2.1 Python

Python è il linguaggio di programmazione usato in questo progetto. Si tratta di un linguaggio ad alto livello, tipato dinamicamente e orientato agli oggetti. Viene apprezzato per la sua semplicità e rapidità d'uso. Grazie alla sua facilità di apprendimento è ideale sia per chi ha già esperienza sia per chi è alle prime armi con l'informatica.

Si può utilizzare per ogni tipo di applicazione, ma spesso si usa Python nella data analysis in quanto, oltre ad essere semplice da capire, offre numerose librerie che permettono di usare formule molto utilizzate nella statistica e creare grafici di ogni tipo che rappresentino i risultati.

Si possono vedere qua di seguito le librerie. [15]



Figura 2.1: Logo Python

Nel seguito vengono descritte le librerie utilizzate.

### 2.2.2 Pandas

Pandas è una delle librerie più popolari e potenti del linguaggio Python, ampiamente utilizzata nell'ambito della data science, dell'analisi dei dati e dell'elaborazione statistica.

Una delle principali ragioni per cui Pandas è così diffusa è la sua capacità di fornire strumenti avanzati per la manipolazione e l'analisi dei dati in modo semplice ed intuitivo. Questo è possibile grazie alle strutture dati "Dataframe" e "Series", che offrono flessibilità e facilità d'uso. Inoltre consente anche di leggere e scrivere facilmente dati provenienti dai vari formati di file come CSV, Excel, Txt ed altri, dando poi la possibilità di lavorarci tramite le strutture dati citate prima.

Nel progetto in questione, i dati vengono estrapolati da un file excel e vengono poi letti ed importati in un DataFrame, che permette di manipolare, analizzare e visualizzare i dati in maniera efficiente e ottenerne informazioni rilevanti. [16]



### 2.2.3 Matplotlib

Matplotlib è una libreria che permette di creare i grafici di qualsiasi tipo con Python, utilizzando i dati e i risultati ottenuti dalle analisi.

Verrà utilizzata in tutte le analisi per creare i grafici che sono descritti in questa tesi.

Ogni grafico mostrato è stato creato con la libreria matplotlib usando, più o meno sempre, le seguenti righe di codice:

```
plt.bar(dfIntero['bin'], dfIntero['somma'], color='green')
plt.xlabel('Numero citazioni')
plt.ylabel('Numero di paper che hanno quel numero di
           citazioni')
plt.title('Numero di paper che hanno quel numero di
          citazioni anno: ' + str(x) )
plt.xticks(dfIntero['bin'], rotation=90)
plt.savefig('grafico_distribuzioni' + str(x) + '.pdf', dpi
           =300, bbox_inches='tight')
plt.show()
```

Quindi si esplicita il tipo di grafico, in questo caso "bar", si mettono all'interno della funzione le variabili da mettere rispettivamente nell'asse delle x e delle y e il colore del grafico. Poi si settano tutte le etichette e il titolo. Per agevolare il lavoro si usa la funzione "savefig" che va a salvare in automatico il grafico nell'estensione che specifichi all'interno, qua si nota che è in "pdf".

Infine si mostra il grafico con la funzione "show()".

### 2.2.4 Numpy

Si parla di un'altra delle librerie più utilizzate tra i data scientist e non solo. Permette di lavorare con dati numerici e array multidimensionali, di utilizzare tool per il calcolo scientifico e l'elaborazione numerica.

In questo lavoro la libreria viene utilizzata per andare a classificare i tipi di affiliazione degli autori degli articoli, per esempio se un articolo ha tutti gli autori

affiliati ad università viene posta l'affiliazione di tipo 1 per questo articolo, 2 se sono tutte aziende e 0 se sono misti.

### 2.2.5 Seaborn

Seaborn è una libreria per la visualizzazione dei dati in Python, costruita sulla base di matplotlib. Il suo obiettivo principale è di semplificare la creazione di grafici statistici, offrendo un'interfaccia di alto livello che rende più intuitiva e accessibile la visualizzazione delle informazioni. Inoltre consente di creare grafici non solo esteticamente accattivanti, ma anche altamente informativi, permettendo agli utenti di evidenziare relazioni, tendenze e distribuzioni nei dati in modo efficace.

Nel lavoro svolto viene utilizzato per andare a creare grafici che includessero elementi statistici come la mediana e gli intervalli di confidenza, utilizzati nella visualizzazione delle statistiche del numero di citazioni.

### 2.2.6 WordCloud

Permette di realizzare dei grafici che vanno ad evidenziare le parole più frequentemente utilizzate in un gruppo di parole o in un testo. Per la realizzazione, bisogna installare le librerie matplotlib e wordcloud e dopo di che si va a passare in input il testo da analizzare, nel caso di queste conferenze si prende l'abstract dell'articolo.

#### Stopwords

Da usare insieme a wordcloud c'è la libreria stopwords che va a bloccare tutte quelle parole che sono di consuetudine usare, come i verbi, che altrimenti comparirebbero ovviamente come le parole più grandi.

### 2.2.7 Pycountry

E' una piccola libreria python che permette di identificare i continenti di cui fanno parte gli stati citati nel dataset. In questa analisi si usa per identificare da quali continenti deriva il maggior numero di articoli.

### 2.2.8 Plotlyexpress

Plotly.express è un modulo che fa parte della libreria plotly e permette di creare delle figure e delle grafiche diverse dal solito. Viene utilizzata in questo caso, infatti, per creare le heatmap che rappresentano delle mappe che vengono colorate in base al valore della variabile data, che in questo caso si tratta del numero di articoli proveniente da un determinato stato.

### 2.2.9 Jupyter

Jupyter è un editor, che permette di eseguire codice Python ed è completamente open source.

E' stato scelto per scrivere ed eseguire il codice dell'analisi in quanto permette di lavorare in modo semplice e intuibile, anche grazie alla sua suddivisione in celle che offre la possibilità di eseguire un pezzo di codice alla volta.

# Capitolo 3

## Implementazione

In questo capitolo si va ad illustrare quella che è la seconda fase della analisi dati, ovvero la raccolta dei dati e di seguito il loro pre-processing, quindi la loro pulizia, che è invece la terza fase.

Vengono successivamente mostrati e descritti gli algoritmi principali che hanno permesso di ottenere i risultati desiderati.

### 3.1 Raccolta dati

I dataset utilizzati per l'analisi sono stati presi dalla banca dati dell'IEEE, tramite il sito "IEEE/xplore.org". [17]

Come già scritto riguardano le conferenze citate nella sezione 2.3.

Per la conferenza INFOCOM si hanno circa 300 elementi, mentre per la GLOBECOM si ha una cardinalità di circa 900 articoli all'anno, mentre la PERCOM conta all'anno in media circa 30 articoli. Per tutte e tre le conferenze gli attributi sono gli stessi, citando alcuni dei più utilizzati nell'analisi troviamo: il numero di citazioni dell'articolo, l'anno di pubblicazione, il titolo della conferenza, il titolo dell'articolo, l'abstract, i nomi degli autori e i nomi delle affiliazioni degli autori.

Inizialmente, utilizzando la libreria pandas, vengono caricati i file csv precedentemente scaricati. Come mostrato nel codice, gli anni di riferimento e il nome della cartella contenente i dati vengono specificati manualmente. Successivamente,

attraverso un ciclo for, si legge ciascun file di tipo csv, estraendo dalla prima riga i nomi degli attributi dei dataset e prendendo ogni elemento separato dalla virgola.

```
anno1=2014
anno2=2023
cartella="globecom_data"
dfIntero = pd.DataFrame()
for x in range (anno1, anno2+1):
    percorso="C:\\Users\\Altea\\Desktop\\universita\\tesi\\"
        + cartella + "\\\" + str(x) + "_f.csv"
    read = pd.read_csv(percorso, sep=",", header=0)
```

I dati dopo sono stati puliti e per ogni analisi è stato creato un dataframe con le colonne di interesse.

In questa analisi non sono stati tolti i valori null poichè o non andavano a perturbare i risultati e quindi non era necessario cancellarli oppure sono stati sostituiti con degli zero con la funzione "fillna()"

```
df=pd.DataFrame(read, columns=['Publication Title', '
    Publication Year'])
```

In questo caso si va a creare un nuovo dataframe, df, prendendo le colonne utili a calcolare la cardinalità del dataframe. Con più precisione si contano il numero di articoli presentati ogni anno alla conferenza.

Per ottenere ulteriori risultati è stato fatto anche un lavoro di geo-localizzazione per quanto riguarda l'estrapolazione dello stato di provenienza degli autori.

Per ottenere la colonna "fixcountries", ovvero quella con gli stati di provenienza degli autori, si è proceduto inizialmente con il parse dell'università di appartenenza, se non si riusciva in questo modo, si proseguiva con delle funzioni di geo-localizzazione, usando la libreria "geopy.geocoders" di Python. Altrimenti l'ultima soluzione era quella di inserirlo manualmente.

## 3.2 Algoritmi

In questa sezione si mostrano alcuni frammenti di codice che caratterizzano le classi principali del progetto.

### 3.2.1 Cardinalità e Medie

La prima cosa che di solito si fa in un'analisi dati è la parte descrittiva e quindi si va a calcolare in primis la cardinalità.

Di seguito vediamo il codice per calcolare il numero degli articoli di ogni anno proposti alle conferenze.

```
dfIntero = pd.DataFrame()
for x in range (anno1, anno2+1):
    percorso="C:\\Users\\Altea\\Desktop\\universita\\tesi\\"
        + cartella + "\\\" + str(x) + "_f.csv"
    read = pd.read_csv(percorso, sep=",", header=0)
    df=pd.DataFrame(read, columns=['Publication Year', '
        Article Citation Count'])
    dfIntero=pd.concat([dfIntero, df], ignore_index=True )
numArticoli=dfIntero.groupby("Publication Year").size().
    reset_index(name='Numero articoli')
```

Si crea innanzitutto un dataframe "dfIntero" nuovo, si esegue il ciclo for per ogni anno presente nella cartella contenente i file con i dati, poi si crea un dataframe "df" per ogni file prendendo le poche colonne necessarie per fare il conteggio. Tecnicamente basta anche solo quella dell'anno poichè successivamente si andrà ad utilizzare la funzione count() raggruppando per l'attributo dell'anno, ma siccome, come si vedrà anche dopo, si calcola anche la media delle citazioni è stata aggiunta anche la colonna 'Article Citation Count' per utilizzare poi lo stesso dataframe. Una volta creato il DataFrame relativo a ciascun anno, questo viene aggiunto al DataFrame inizializzato fuori dal ciclo, consentendo di unire tutti i dati di tutti gli anni tramite la funzione "pd.concat".

Infine fuori dal ciclo prendo dfIntero e vado a raggrupparlo per la variabile 'Publi-

cation Year' che indica l'anno e uso la funzione "count()" per contare il numero di articoli di ogni anno.

Un altro dato interessante è dato dalla media delle citazioni degli articoli di ogni anno, questo lo ottengo facendo un groupby sempre sullo stesso dataframe "dfIntero" e chiamando la funzione "mean()"

```
media=dfIntero.groupby("Publication Year").mean('Article
Citation Count').reset_index()
```

### 3.2.2 Distribuzione citazioni

Con questo codice, dopo aver fatto la prima parte di creazione del dataframe con le colonne d'interesse, si fa un groupby per il numero di citazioni per ogni anno. Quindi se tre articoli hanno ottenuto 20 citazioni per la voce "20 citazioni" ci sarà un conteggio di 3.

Successivamente si va a fare un ciclo for che prende una variabile "y" che va da 0 a 330 (numero massimo di citazioni) con un passo di 10.

In questo modo si creano dei bin che contengono il numero di articoli che hanno ottenuto un numero di citazioni da y a y+10. Difatti nelle righe si fa un count con questa selezione e poi si crea un dataframe che contiene una colonna con i bin e una colonna con il conteggio degli articoli.

```
anno1=2014
anno2=2023
cartella="globecom_data"
dfIntero = pd.DataFrame()
for x in range (anno1, anno2+1):
    percorso="C:\\Users\\Altea\\Desktop\\universita\\tesi\\"
        + cartella + "\\ " + str(x) + "_f.csv"
    read = pd.read_csv(percorso, sep=",", header=0)
    df=pd.DataFrame(read, columns=['Publication Title', '
Publication Year', 'Article Citation Count'])
    count=df.groupby('Article Citation Count').size().
        reset_index(name='conteggio')
```

```

dfIntero = pd.DataFrame()
for y in range(0, 330, 10):
    bin=count[(count['Article Citation Count'] > y) & (
        count['Article Citation Count'] <= y+10)]
    somma=bin.sum()
    dfbin = pd.DataFrame({'bin':[str(y)+'-'+str(y+10)],
        'somma':[somma['conteggio']]})
    dfIntero=pd.concat([dfIntero, dfbin], ignore_index=
        True )
print(dfIntero)

```

### 3.2.3 WordCloud

Di seguito viene presentato il codice che permette di creare il grafico WordCloud con le parole presenti nell'abstract e nel titolo di ogni articolo.

Si va come per ogni codice a creare il dataframe con gli attributi necessari, dopo di che per andare ad analizzare sia il titolo che l'abstract, in quanto sono due attributi separati, si fa un join tra queste due colonne unendole come unico testo in una variabile.

Si vanno anche a togliere i verbi fra le parole da prendere in considerazione, tramite la libreria stopwords in modo da non invadere con parole comuni come i verbi il grafico che invece dovrebbe considerare altre parole.

Infine si crea il WordCloud con il testo filtrato e si ottengono i grafici.

```

df=pd.DataFrame(read, columns=['Document Title', '
    Publication Year', 'Abstract'])
testo = " ".join((str(abstract) + " " + str(title)) for
    abstract, title in zip(df['Abstract'], df['Document Title
    ']))
doc = nlp(testo)
parole_senza_verbi = [token.text for token in doc if token.
    pos_ != "VERB"]
testo_filtrato = ' '.join(parole_senza_verbi)

```

```
wordcloud = WordCloud(width=800, height=400,
    background_color='white').generate(testo_filtrato)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
```

### 3.2.4 Rappresentanze e Mappe

Le seguenti porzioni di codice rappresentano diversi modi per creare grafici che mettono in evidenza le analisi relative alla rappresentanza degli autori in ciascun stato.

In questo caso si prendono in considerazione i singoli stati con l'attributo "Fix-countries".

Per prima cosa si crea una colonna dividendo gli stati di tutti gli autori in ogni riga con la funzione split e poi con la funzione explode si duplicano le righe dividendo gli autori.

Nell'ultima riga si vanno poi a raggruppare gli stati con il group by e si esegue un count per avere il numero degli autori che rappresentano quello stato.

```
df=pd.DataFrame(read, columns=['Publication Title', '
    Publication Year', 'Fix_Countries'])
    df['Affiliazioni'] = df['Fix_Countries'].str.split(';')
    dfAffiliazioni = df.explode('Affiliazioni')
    numStati=dfAffiliazioni.groupby("Affiliazioni").size().
        reset_index(name='Numero stati')
```

Di seguito invece si vanno a creare delle funzioni che consentono di creare dei gruppi di stato per capire quale parte del mondo è più presente nelle conferenze. Questi gruppi sono gli Stati Uniti, la Cina, l'Europa, il resto del mondo, ma poi vengono presi in considerazione anche combinazioni di questi gruppi, come Cina+Europa, Cina+Stati Uniti e Stati Uniti + Europa.

Ciò avviene tramite una serie di passaggi descritti di seguito:

1. Viene utilizzata una prima funzione che va a fare l'explode degli articoli andando a moltiplicare l'articolo dividendo gli stati, quindi un articolo con tre stati dopo risulterà tre volte quell'articolo con ogni riga uno stato.
2. Dopo si va ad individuare di quale gruppo "primario" fa parte lo stato passato in input alla funzione che decifra gli stati.
3. Dopo di ch  si riuniscono gli articoli
4. Infine si va a vedere di quale gruppo "primario" fanno parte gli stati, se un articolo per esempio ha tutti stati del gruppo China, sar  un articolo del gruppo China, se ne ha due della China e uno dell'Europa sar  "China+Europa" e cos  via.

```
def stato_a_gruppo(stato):
    stati_us = ['US', 'United States']
    stati_china = ['China']
    stati_europa = ['Italy', 'Germany', 'France', 'Spain', '
        UK', 'United Kingdom', 'Belgium', 'Norway', 'Greece', '
        Hungary', 'Sweden',
        'Switzerland', 'Luxembourg', 'Netherlands', 'Ukraine'] #
        Aggiungi altri stati europei
    if stato in stati_us:
        return 'US'
    elif stato in stati_china:
        return 'China'
    elif stato in stati_europa:
        return 'Europa'
    else:
        return 'Altro'

def determina_gruppo(provenienza):
    stati = provenienza.split(';')
    gruppi = set([stato_a_gruppo(stato.strip()) for stato in
        stati])
```

```

gruppi_ordinati = sorted(gruppi)
gruppo_combinato = '+'.join(gruppi_ordinati)

gruppi_accettati = ['China', 'US', 'Europa', 'China+US',
                    'China+Europa', 'Europa+US']

if gruppo_combinato not in gruppi_accettati:
    return 'Altri'
return gruppo_combinato

dati_percentuali_anni = []

for x in range (anno1, anno2+1):
    percorso="C:\\Users\\Altea\\Desktop\\universita\\tesi\\"
        + cartella + "\\\" + str(x) + "_f.csv"
    read = pd.read_csv(percorso, sep=",", header=0)
    df=pd.DataFrame(read, columns=['Publication Title', '
        Publication Year', 'Fix_Countries'])
    df['Gruppo'] = df['Fix_Countries'].apply(
        determina_gruppo)
    conteggio_gruppi = df['Gruppo'].value_counts()
    percentuali_gruppi = (conteggio_gruppi / len(df)) * 100
    percentuali_gruppi.name = str(x)
    dati_percentuali_anni.append(percentuali_gruppi)

```

Come ultima analisi sulle rappresentanze si vede un codice che permette di creare delle grafiche a mappa con tono di colore in base al numero di volte che uno stato viene rappresentato.

Si usa innanzitutto la libreria plotly.express che permette di fare queste grafiche, poi nella funzione choropleth si inseriscono i dati da prendere in considerazione, le etichette e la gestione del colore.

```

fig = px.choropleth(
    numStati,

```

```
locations="Affiliazioni",
locationmode='country names',
color="Numero rappresentanze",
hover_name="Affiliazioni",
color_continuous_scale=px.colors.sequential.Plasma,
title='Numero di stati rappresentati nel ' + str(x)
)
fig.write_image('mappa_rappresentanze_' + str(x) + '.pdf
')
fig.show()
```

### 3.2.5 Riferimenti AI e Blockchain

Per ottenere la frequenza di articoli che parlano di un determinato argomento, come in questo caso Artificial Intelligence e Blockchain, si è sviluppato questo codice che va a selezionare solo le righe del dataset che contengono nell'abstract delle parole specifiche contenute in un'array tramite la funzione "contains()". Dopo di ch  si conta quante sono le righe selezionate, poi si contano le righe totali del dataset e si calcola la percentuale. Questo viene fatto per ogni anno con un ciclo for e viene inserita la percentuale di ogni anno in un dataframe che verr  utilizzato per fare il grafico che mostra le percentuali di ogni anno.

```
dfperAnno=dfIntero.groupby("Publication Year").size().
reset_index(name='Totale articoli') #per avere il numero
totale di articoli
par=['learning', 'artificial intelligence', 'machine
learning', 'deep learning']
parole = "|".join(par)
risultato=dfIntero['Abstract'].str.contains(parole, case=
False, na=False)
righe=dfIntero[risultato]
print(righe)
```

```
numArticoli=righe.groupby("Publication Year").size().
    reset_index(name='Numero articoli')
print(numArticoli)
join= numArticoli.join(dfperAnno.set_index('Publication Year
    '), on='Publication Year', how='inner')
print(join)
join['percentuale']=(join['Numero articoli']*100)/join['
    Totale articoli']
```

Per la blockchain il codice è uguale ma viene messa questa lista di parole, più rilevanti in questo ambito, da cercare: `par=['blockchain', 'smart contract', 'smart contracts', 'bitcoin', 'cryptovalute']`.

### 3.2.6 Correlazioni

Per questo paragrafo si prendono in considerazioni due coefficienti statistici, ovvero Pearson e Kendall.

Il primo indica un'eventuale relazione di linearità tra due variabili. Ha un valore compreso tra +1 e -1, dove +1 corrisponde alla perfetta correlazione lineare positiva, 0 corrisponde a un'assenza di correlazione lineare e -1 corrisponde alla perfetta correlazione lineare negativa. Per calcolare Pearson è necessario che le variabili siano quantitative ed appaiate sullo stesso caso. [18]

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Figura 3.1: Formula del coefficiente di Pearson

Qualora una variabile, o entrambe, non fosse quantitativa, si può provare a prendere in considerazione il coefficiente di Kendall che invece permette di usare anche variabili qualitative ordinali e valori outliers. [19]

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)$$

Figura 3.2: Formula del coefficiente di Kendall

Nella prima riga del codice seguente si va a chiamare la funzione "corr()" specificando come metodo "Pearson" e associando la funzione agli attributi 'Numero stati' e 'Article Citation Count' del dataframe "dfStatiCitazioni" costruito apposta. Questa funzione si usa per andare a calcolare la matrice di correlazione che avrà come dati appunto i due attributi specificati e come valori 1 e 1 per la diagonale principale e i valore della correlazione per la diagonale secondaria.

Successivamente si va a prendere la matrice e con la funzione "loc[]" si selezionano solo i valore della correlazione.

Si è fatta la stessa cosa per calcolare il coefficiente di Kendall, sostituendo nel metodo la voce da Pearson a Kendall.

Infine, è stato creato un dataframe con all'interno l'anno e i due coefficienti ottenuti per poterne trarre poi dei grafici.

```

correlation_Pearson = dfStatiCitazioni[['Numero stati', '
    Article Citation Count']].corr(method='pearson')
pearson_corr = correlation_Pearson.loc['Numero stati', '
    Article Citation Count']
correlation_kendall = dfStatiCitazioni[['Numero stati', '
    Article Citation Count']].corr(method='kendall')
kendall_corr = correlation_kendall.loc['Numero stati', '
    Article Citation Count']
riga = {'Year': x, 'Pearson': pearson_corr, 'Kendall':
    kendall_corr}
dfCorrelazioni = pd.concat([dfCorrelazioni, pd.DataFrame([
    riga])], ignore_index=True)

```

Il codice appena visto è stato preso dallo script per calcolare la correlazione tra numero di stati che vengono rappresentati da un articolo e il suo numero di citazioni.

Per il programma che va a cercare delle correlazioni tra il numero di citazioni e il numero di autori si eseguono gli stessi passaggi.

Mentre il sottostante codice va a fare un lavoro più complesso, ovvero, innanzitutto definisce i tipi di affiliazioni degli articoli, quindi se un articolo ha autori provenienti interamente da università oppure fanno tutti parte di aziende o anche se sono misti.

```
dfCorrelazioni=pd.DataFrame()
def assegna_valore(gruppo):
    if all(gruppo == 0):
        return 0 # Tutte universit\a
    elif all(gruppo == 1):
        return 1 # Tutte aziende
    else:
        return 2 # Miste
for x in range (anno1, anno2+1):
    percorso="C:\\Users\\Altea\\Desktop\\universita\\tesi\\"
        + cartella + "\\\" + str(x) + "_f.csv"
    read = pd.read_csv(percorso, sep=",", header=0)
    df=pd.DataFrame(read, columns=['Publication Year',
        Document Title', 'Author Affiliations', 'Article
        Citation Count'])
    df2=pd.DataFrame(read)
    df['affiliazioni'] = df['Author Affiliations'].str.split
        (';')
    dfAffiliazioni = df.explode('affiliazioni')
    par=['University', 'School', 'Institute', 'Department']
    parole = "|".join(par)
    dfAffiliazioni['IsCorporate'] = np.where(dfAffiliazioni[
        'affiliazioni'].str.contains(parole, case=False, na=
        False), 0, 1)
    dfAffiliazioni['AffiliationType'] = dfAffiliazioni.
        groupby('Document Title')['IsCorporate'].transform(
        assegna_valore)
```

```
dfFinale = dfAffiliazioni.drop_duplicates(subset='
    Document Title')
print(dfFinale[['Document Title', 'AffiliationType', '
    Article Citation Count']])
```

Partendo dal ciclo for si va, come per ogni altro codice a prendere il file di interesse e creare il dataframe con solo le colonne necessarie. Dopodiché si vanno a dividere le righe per ogni affiliazione, per esempio se un articolo ha due autori, quindi due affiliazioni, le righe si duplicheranno una con un autore e una con l'altro. Successivamente si cerca tramite la funzione "contains()" se queste affiliazioni sono accademiche o aziendali, specificando che se contengono nel nome una delle parole dichiarate, sono accademiche, altrimenti aziendali.

A seguire si vanno a raggruppare nuovamente per il titolo e si passa il risultato alla funzione "assegna\_valore" definita nelle prime righe dello script.

Questa funzione va ad assegnare un valore tra 0, 1 e 2 se, rispettivamente, gli articoli hanno come valore delle affiliazioni tutti 0, quindi tutte università, tutti 1, ovvero completamente aziendali, oppure viene assegnato il valore 2 se ci sono entrambi i valori.

Infine si va a calcolare come negli altri casi la correlazione, ma in questo caso solo di Kendall poiché è presente una variabile qualitativa, ovvero il tipo di affiliazione.



# Capitolo 4

## Risultati

Ora si introdurranno i test effettuati e i risultati più significativi ottenuti dall'analisi, tramite grafici che li rappresentino.

In alcuni casi si raggruppano i grafici delle tre conferenze sui dati dell'ultimo anno, ovvero il 2023. Mentre in altri casi verranno mostrati separatamente.

In appendice sono collocati tutti i risultati prodotti dallo studio.

### 4.1 Cardinalità

Per prima cosa, nell'analisi, è stata esaminata la cardinalità dei dataset utilizzati, ovvero il numero di articoli pubblicato ogni anno nelle tre conferenze considerate.

Il grafico a barre riportato di seguito illustra appunto questi dati affiancando le barre che riportano il numero di articoli delle tre conferenze per ogni anno.

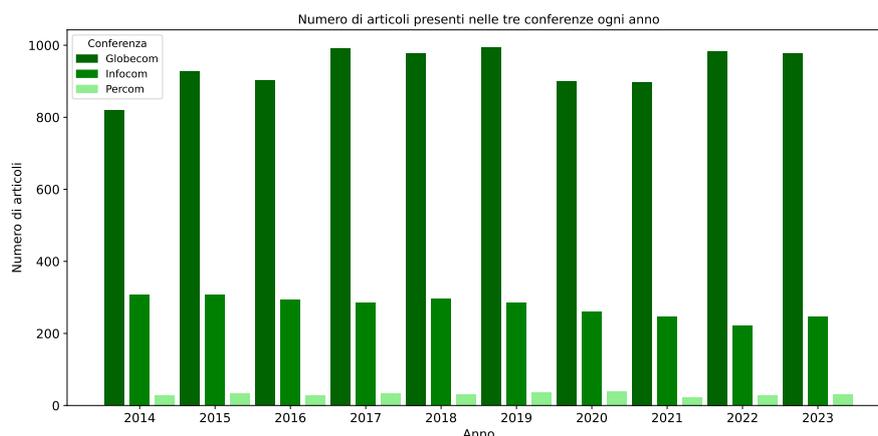


Figura 4.1: Numero di articoli presenti ogni anno nelle tre conferenze

Nel grafico sicuramente si nota la differenza di articoli presentati tra le tre conferenze, la Globecom ha una cardinalità di quasi 1000 articoli, la Percom di circa 30 e la Infocom è una via di mezzo con circa 300 articoli all'anno.

Non sono evidenti particolari trend temporali, ma si può notare come nel 2020 e nel 2021 ci sia un numero leggermente minore di articoli presentati, ovviamente per una causa abbastanza riconoscibile, ovvero la pandemia da Coronavirus. Infatti oltre ai blocchi che ha causato che già si conoscono, la pandemia ha bloccato anche molti ricercatori nel fare il loro lavoro causando un decremento per questi due anni.

[20]

## 4.2 Provenienza articoli e autori

I primi risultati che si possono andare ad evidenziare sono quelli che vanno a studiare come sono cresciuti negli anni gli stati per quanto riguarda la pubblicazione e promozione degli articoli.

Si visualizza come prima cosa una "heat-map" che va a rappresentare un planisfero che si colora in base al numero di articoli provenienti da un determinato stato. Si considera solo l'anno 2023.

Numero di stati rappresentati nel 2023

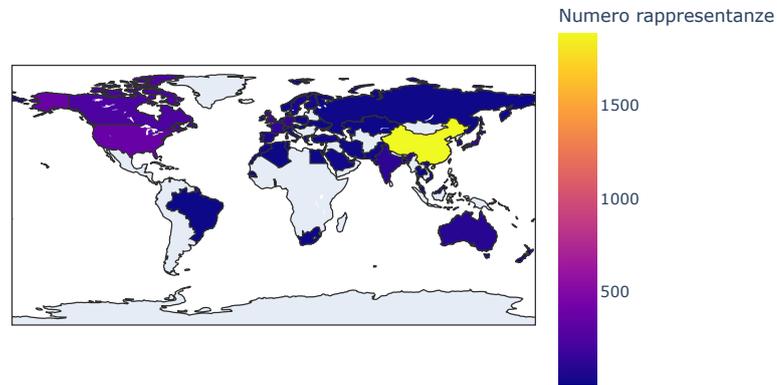


Figura 4.2: Mappa provenienze articoli Globecom nel 2023

Sicuramente la cardinalità descritta prima si noterà anche nella presenza mondiale, poi si noterà di più il fatto che la Globecom sia una conferenza più internazionale rispetto le altre due. Infatti vediamo più aree del mondo colorate come il nord America, il Brasile, gran parte dell'Europa, la Russia e l'Asia. Addirittura la Cina risulta gialla, ovvero ha più di 1500 presenze.

Numero di stati rappresentati nel 2023

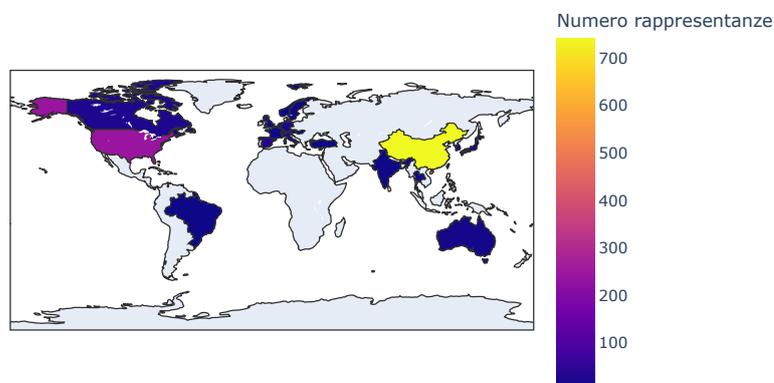


Figura 4.3: Mappa provenienze articoli Infocom nel 2023

La conferenza Infocom vede una distribuzione geografica meno ampia e i numeri complessivi sono significativamente più bassi, riflettendo quindi una partecipazione più circoscritta. Di conseguenza si vedono meno parti del mondo colorate. Tuttavia c'è una differenza minore tra le presenze registrate per la Cina e quelle degli Stati Uniti, quest'ultima si aggira intorno alle 300 presenze mentre la prima intorno alle 700. Questo dato suggerisce un maggiore equilibrio tra i due Paesi rispetto ad altre conferenze, pur mantenendo la Cina in una posizione predominante in termini di contributi complessivi.

Numero di stati rappresentati nel 2023

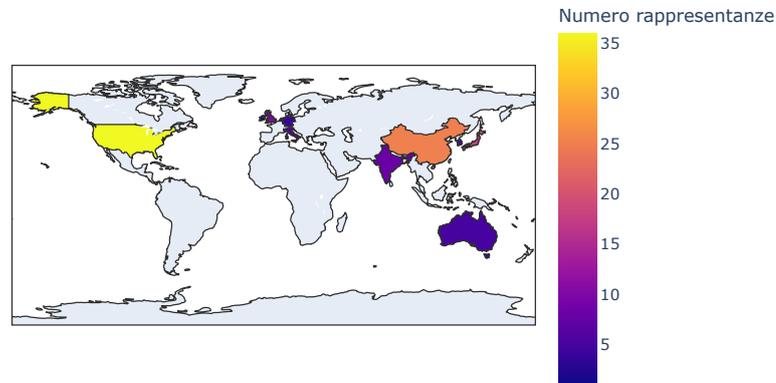


Figura 4.4: Mappa provenienze articoli Percom nel 2023

Per quest'ultimo grafico si nota invece una maggiore presenza degli Stati Uniti e un po' meno della Cina, mentre l'Europa è sempre quella meno presente, in questo caso vede addirittura solo cinque stati, ovvero l'Italia, la Germania, i Paesi Bassi, la Gran Bretagna e l'Irlanda.

Si prendano ora in considerazione i continenti e si illustra di seguito tramite il grafico la provenienza degli articoli nel 2023 delle tre conferenze.

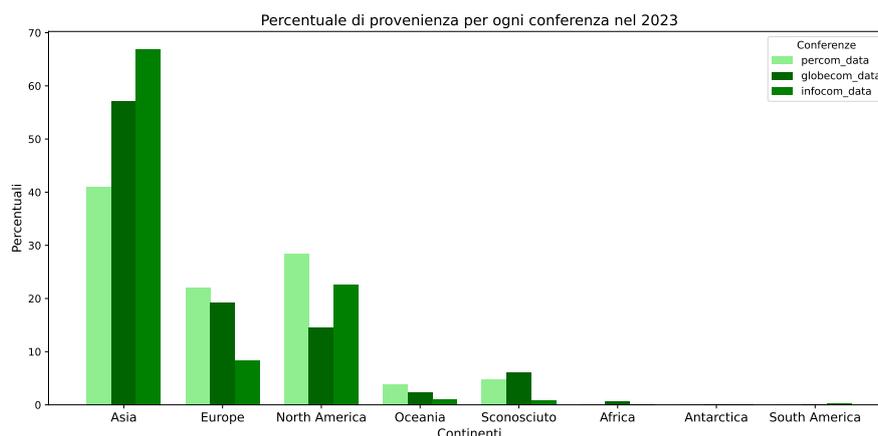


Figura 4.5: Grafico con la provenienza degli articoli delle tre conferenze

Come si è notato dai precedenti risultati l'Asia, più in particolare la Cina, negli ultimi anni è cresciuta parecchio, infatti nel 2023 si nota la netta differenza di percentuale di articoli presenti nelle tre conferenze, rispetto agli altri continenti. Seguono come da aspettative le tre barre del Nord America e quelle dell'Europa. Interessante notare che per la Globecom, a differenza delle altre due conferenze, l'Europa nel 2023 abbia avuto un numero maggiore di articoli rispetto al Nord America.

Inoltre sono molti gli articoli che sono stati classificati come "Sconosciuto", mentre un'altra piccola parte di articoli proviene dall'Oceania. Mentre l'Africa è presente solo nella conferenza più grande e per quanto riguarda l'Antartide e il Sud America non è presente alcun articolo.

Si dividono ora gli stati di provenienza in macro gruppi, ovvero: China, Stati Uniti, Europa, China+US<sup>1</sup>, China+Europa, Europa+US e Altri. Questo dato che come anche visto in precedenza, sono i gruppi che più influenzano le conferenze scientifiche.

---

<sup>1</sup>United States

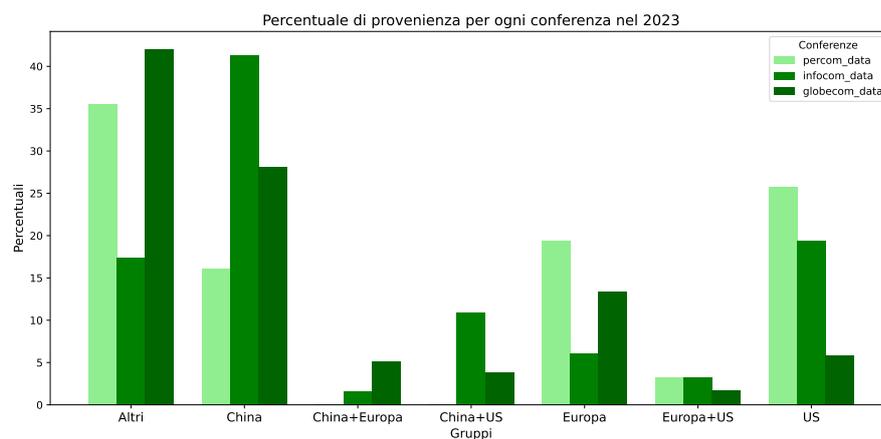


Figura 4.6: Grafico con la provenienza degli articoli divise per macro-gruppi delle tre conferenze

Come previsto, osserviamo che i gruppi "primari", ovvero quelli considerati più influenti, emergono con maggiore evidenza. I gruppi composti da autori di due nazioni risultano meno numerosi, mentre tutti gli altri articoli composti da autori di diverse nazionalità rappresentano la categoria più ampia nella maggior parte dei casi. Tuttavia, un'eccezione è data per la Infocom, dove il numero più elevato di articoli è attribuito al gruppo con autori interamente provenienti dalla Cina.

Si illustra di seguito il cambiamento che hanno fatto negli anni questi gruppi di nazioni presi in considerazione fin'ora tramite uno stacked-bar. I grafici sono divisi per conferenza, si parte con la Globecom.

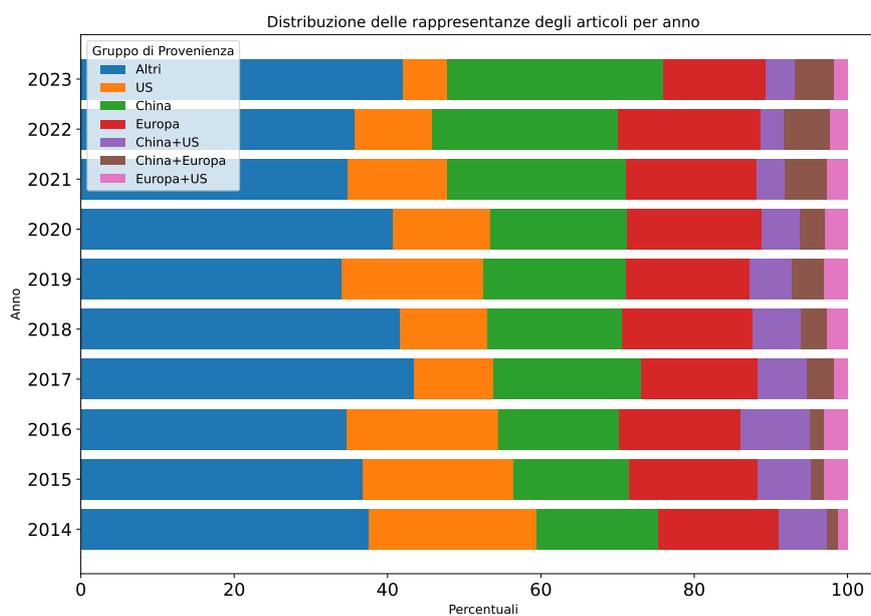


Figura 4.7: Stacked bar GLOBECOM presenze articoli per stato

Da questa prima conferenza si nota una netta presenza degli stati che non siano Cina, Stati Uniti o Europa, ciò era prevedibile per il fatto che si parla di una conferenza molto internazionale.

L'Europa è un continente che è sempre presente, sebbene in misura limitata, senza mostrare una chiara tendenza di crescita o diminuzione. Mentre si notano delle fette minori per China e Stati Uniti che nel tempo vedono crescere la China e diminuire gli Stati Uniti.

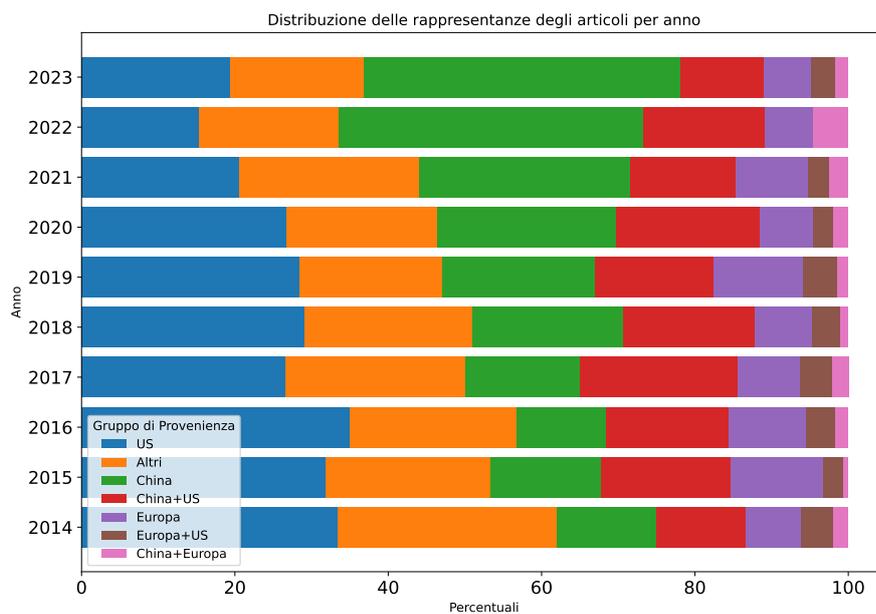


Figura 4.8: Stacked bar INFOCOM presenze articoli per stato

Per la Infocom si nota una notevole crescita della Cina che da circa il 15% di articoli portati nel 2014 arriva ad una percentuale di quasi 40 nel 2023, in parallelo si nota il calo degli Stati Uniti che passano da un oltre 30% ad un poco più del 20%.

Il resto del mondo ha una fetta importante, ma ovviamente si parla di tanti stati per cui è un risultato abbastanza attendibile.

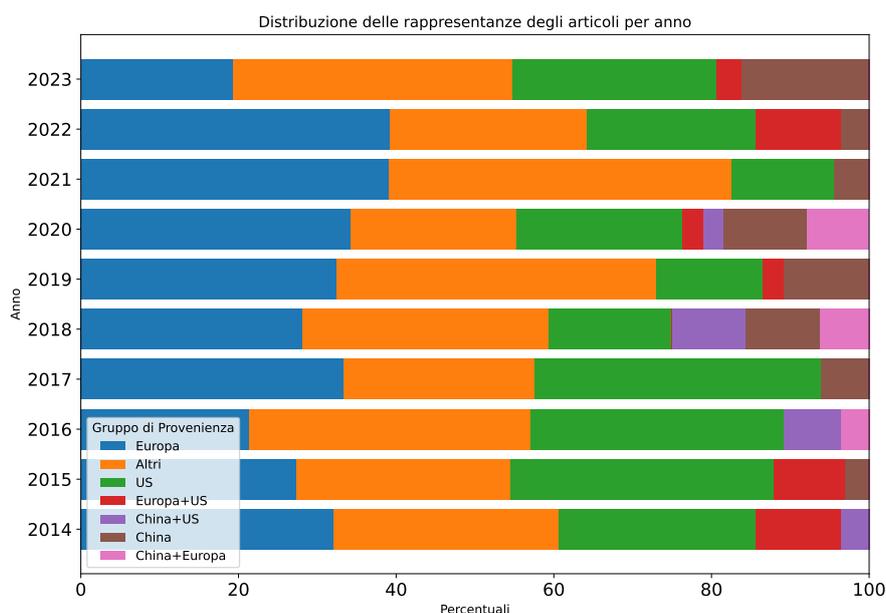


Figura 4.9: Stacked bar PERCOM presenze articoli per stato

Per quanto riguarda la Percom si vede un cambiamento di trend, padroneggia la maggior parte di articoli l'Europa fino al 2022, nel 2023 si inizia a vedere un calo. Gli Stati Uniti hanno anche loro una parte importante anche se non sempre, mentre la Cina in questa conferenza viene a mancare un po' con alcuni anni anche di assenza totale. Sono molto presenti invece altri stati non classificati nei gruppi principali.

### 4.3 Citazioni medie

Nel grafico sottostante si illustrano le citazioni medie per ogni articolo presentato nelle tre conferenze prese in considerazione nel progetto, per ogni anno, con il loro intervallo di confidenza, ovvero l'ampiezza dei suoi valori, quindi dove si trovano più o meno il minimo e il massimo.

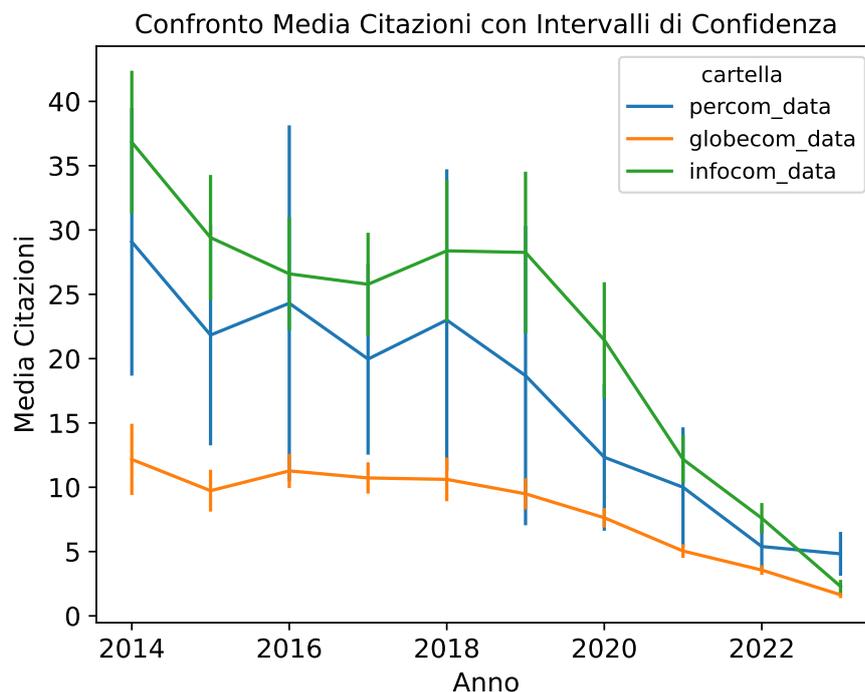


Figura 4.10: Citazioni medie per ogni conferenza con confidenza

Emerge da questo grafico che la Globecom ha delle citazioni medie particolarmente basse. Questo può essere spiegato, almeno in parte, dal numero di articoli pubblicati elevato: con una produzione così ampia, non tutti gli articoli riescono a ottenere citazioni significative, contribuendo così ad abbassare la media complessiva. Inoltre, guardando i valori minimi e massimi, si nota che sono molto vicini tra loro quindi si hanno dei numeri di citazioni generalmente contenuti e raramente elevati.

La Infocom al contrario ha la media più alta quasi ogni anno, si inverte con la Percom solo nel 2023, il suo intervallo di confidenza inoltre è relativamente ampio per cui vuol dire che ci sono articoli anche con un numero di citazioni più elevato. Mentre la Percom ha valori medi delle medie ma il range fra i suoi minimi e massimi è molto ampio il che vuol dire che ci sono articoli che hanno ottenuto molte citazioni, infatti si vedono alcuni articoli che in certi anni arrivano ad avere un massimo di quasi 40 citazioni.

## Distribuzione delle citazioni

Si vada ora a prendere il numero di articoli che hanno quel determinato numero di citazioni, in questo caso il "determinato numero di citazioni" viene raggruppato in dei bin<sup>2</sup> da 10, partendo da zero, quindi si raggruppano tutti gli articoli che hanno avuto da zero a 10 citazioni, poi da 10 a 20 e così via. Si ottiene così una distribuzione, le vediamo di seguito divise per conferenza. In questo paragrafo si descrivono i grafici che fanno riferimento agli articoli del 2014 poichè il numero di citazioni aumenta col passare degli anni, di conseguenza, prendendo i dati dell'anno meno recente, si va ad analizzare un campione più strutturato.

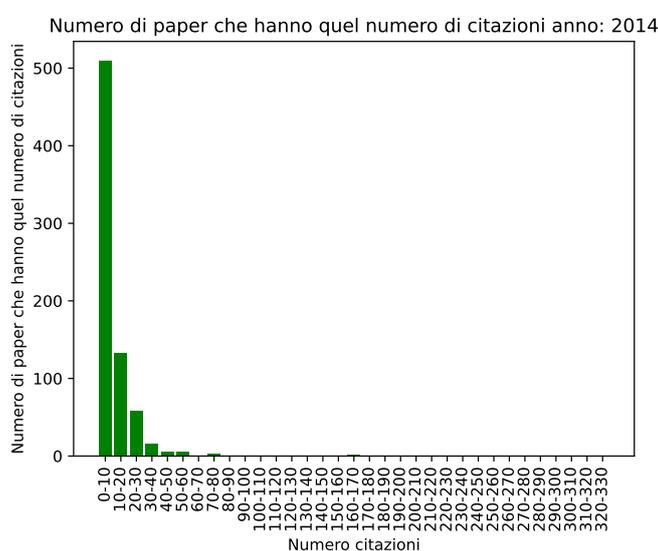


Figura 4.11: Distribuzione del numero di citazioni per la Globecom

La Globecom ha più di un terzo degli articoli che hanno dalle 0 alle 10 citazioni e decisamente meno articoli hanno ottenuto tra le 10 e le 40 citazioni. Anche qui sono presenti dei valori estremi che hanno ottenuto tra i 50 e le 80 citazioni, ma anche 160.

<sup>2</sup>Intervallo utilizzato per raggruppare e organizzare i dati

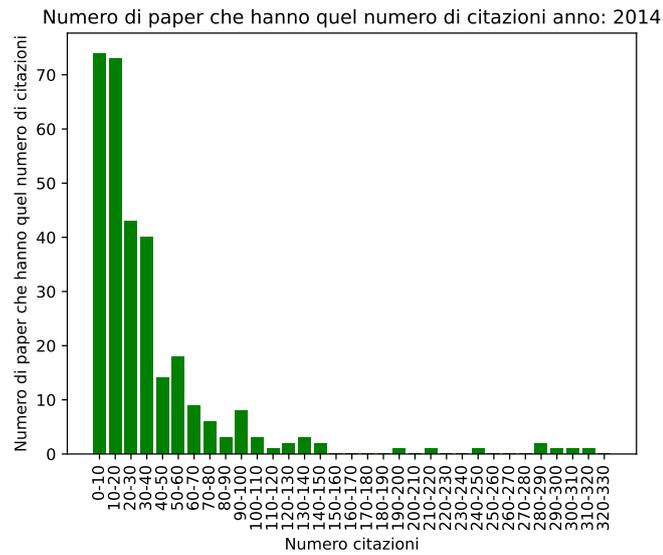


Figura 4.12: Distribuzione del numero di citazioni per la Infocom

La Infocom con i suoi 300 articoli vede una maggioranza per citazioni dallo 0 al 10 e dal 10 al 20, quasi a pari merito. Si vede poi un numero medio di articoli che hanno ottenuto un numero di citazioni dal 20 al 40. Un ristretto numero di articoli ha ottenuto citazioni fino a 150 e infine vediamo alcuni outliers che hanno ottenuto anche fino a 310 citazioni.

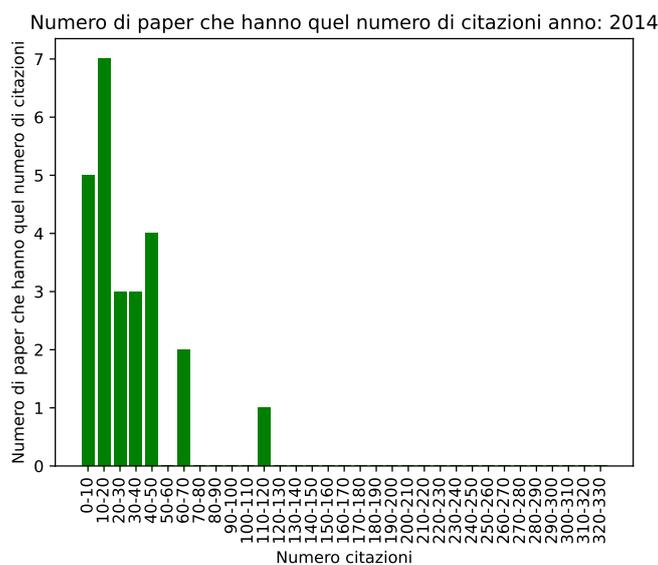


Figura 4.13: Distribuzione del numero di citazioni per la Percom

Per la Percom, a differenza delle altre conferenze, si vede un numero più alto per il bin che va da 10 a 20, proporzionalmente al numero di articoli proposto alla conferenza. Al secondo posto c'è comunque il range di 0-10 citazioni e poi la restante metà degli articoli ha ottenuto un numero di citazioni che va da 20 a 50, poi da 60 a 70 e infine c'è un outliers che ha ottenuto intorno alle 115 citazioni.

### Citazioni medie e tipo di affiliazioni

Si vanno ora, a prendere in considerazione le citazioni medie ottenute ogni anno da tutti gli articoli, suddivisi per tipologia di affiliazione. Ovvero le affiliazioni possono essere di tipo 1 se tutti gli autori che hanno preso parte alla scrittura del paper, sono sotto un'università o un'istituzione accademica. Se tutti gli autori provengono da un'azienda, si classificherà il tipo come 2, mentre per articoli con affiliazione mista viene messo lo 0.

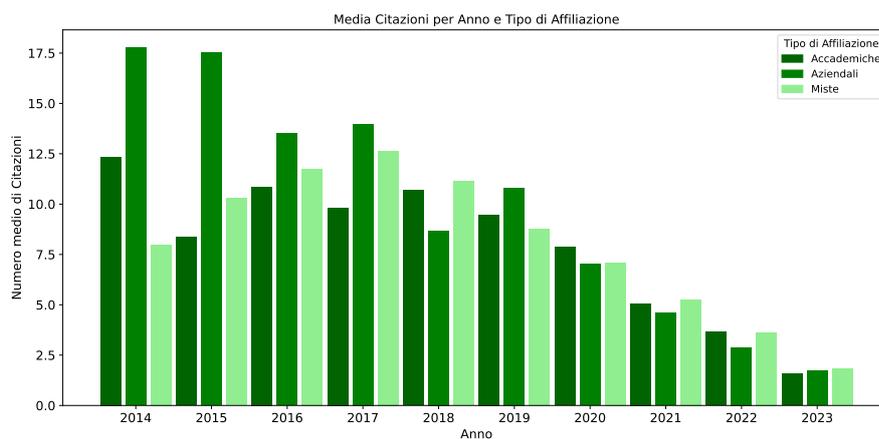


Figura 4.14: Conferenza GLOBECOM citazioni medie per tipo di affiliazioni

Per questa prima conferenza, invece possiamo notare due distinzioni. La prima metà degli anni vede un maggior numero di citazioni medie per i paper di tipo 2, mentre per la seconda metà si vede una crescita delle citazioni medie delle affiliazioni di tipo 0, quindi miste.

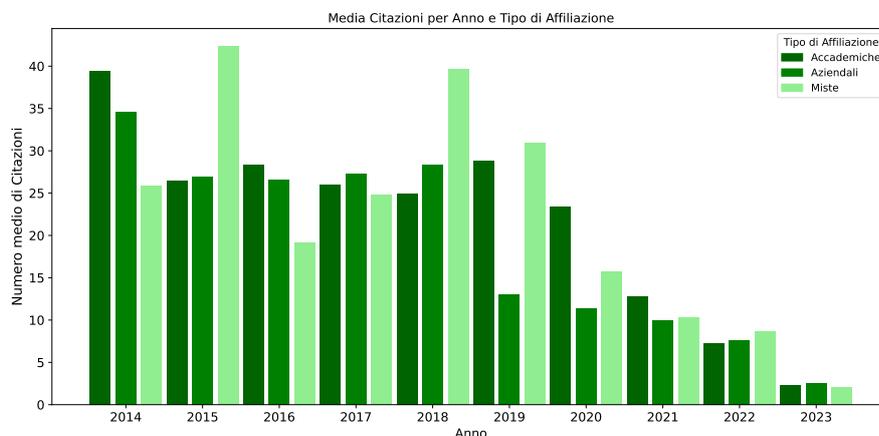


Figura 4.15: Conferenza INFOCOM citazioni medie per tipo di affiliazioni

Per la Infocom, per solo due anni, tra questi dieci presi in considerazione, sono più alte le citazioni medie degli articoli con autori provenienti interamente da aziende; mentre per gli altri otto anni se li dividono equamente le altre due

tipologie di affiliazioni. Inoltre non è presente nessuna tendenza di anni, nel senso che non è i primi quattro anni di tipo 1 e gli ultimi 4 di tipo 2.

Si vede infine la Percom come sono disposte le citazioni medie in base al tipo di affiliazione degli articoli.

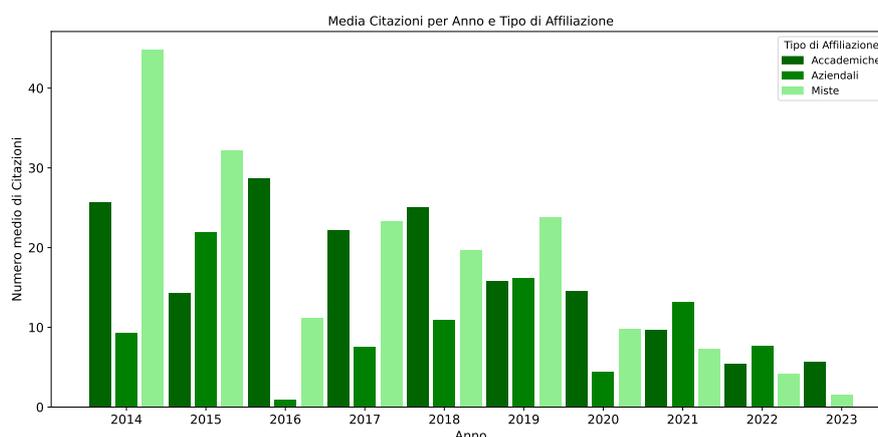


Figura 4.16: Conferenza PERCOM citazioni medie per tipo di affiliazioni

In questo caso non si vede un trend o una distinzione netta fra le classificazioni, si può notare che in quattro anni su dieci prevalgono gli articoli con tipologia di affiliazione mista, per altri quattro anni prevale quella accademica e solo due anni, ovvero nel 2021 e nel 2022 sono presenti più articoli con affiliazione di tipo aziendale, da notare che invece nel 2023 sono completamente assenti quelli di quest'ultima aggregazione.

### Citazioni medie e provenienza dell'articolo

Un altro incrocio che si può prendere in considerazione tra le variabile è quello del numero medio di citazioni e la provenienza degli articoli.

Per individuare al meglio questo tipo di analisi si è scelto come grafico per visualizzare la risposta un boxplot che vada a mostrare i quartili, la mediana, e gli outliers. In questo modo possiamo farci un'idea ben strutturata di quello che risulta.

Più precisamente un boxplot è composto dalla riga più meno spesso al centro che

rappresenta la mediana, il rettangolo rappresenta quella che è la parte più alta della campana di un grafico di distribuzione, ovvero il range che va dal primo al terzo quartile, i "baffi" ovvero le lineette sopra e sotto che rappresentano gli estremi dei dati quindi il minimo e il massimo ed infine ci sono i pallini distaccati che sono gli outliers che spesso non vengono presi in considerazione poiché potrebbero perturbare il risultato.

Qui di seguito si illustra il risultato ottenuto dal dataset della Globecom.

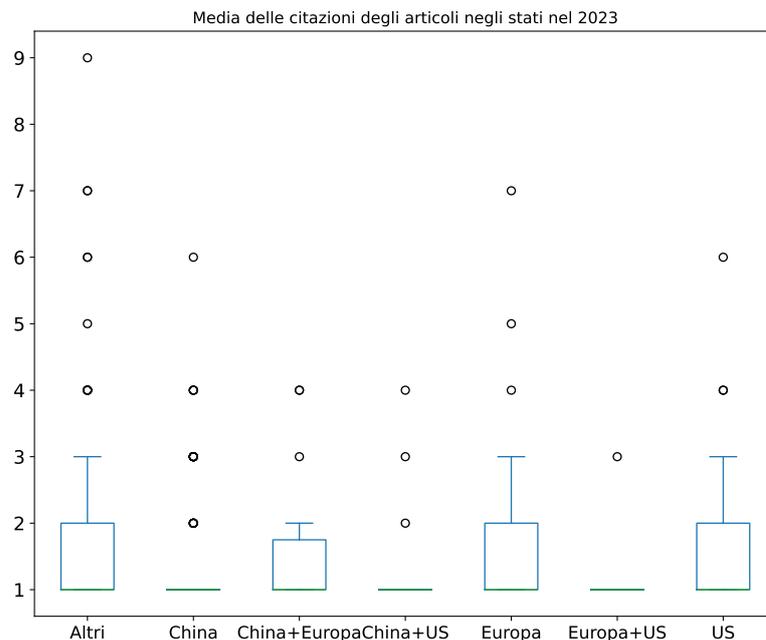


Figura 4.17: Conferenza GLOBECOM citazioni medie per provenienza

Si possono notare degli intervalli di citazioni medie molto bassi, intorno all'uno. Analizzando ogni componente, le mediane sono tutte sull'uno, il range tra il secondo e il terzo quartile per i gruppi "China + Europa", "Europa", "US" e "Altri" arriva a due e per gli ultimi tre gruppi il valore massimo arriva a tre. In tutti i gruppi sono poi presenti dei valori outliers che nel caso di alcuni stati non appartenenti a dei gruppi specifici è presente un outliers a 9, mentre gli altri elementi atipici si trovano tutti al di sotto del 7.

È un risultato aspettabile in quanto si era già visto nel paragrafo con i grafici sulle

distribuzioni che il maggior numero di articoli aveva un numero di citazioni tra lo 0 e il 10, probabilmente la maggior parte ne ha zero o uno.

Ora si mostra il grafico ottenuto dai dati della conferenza Infocom.

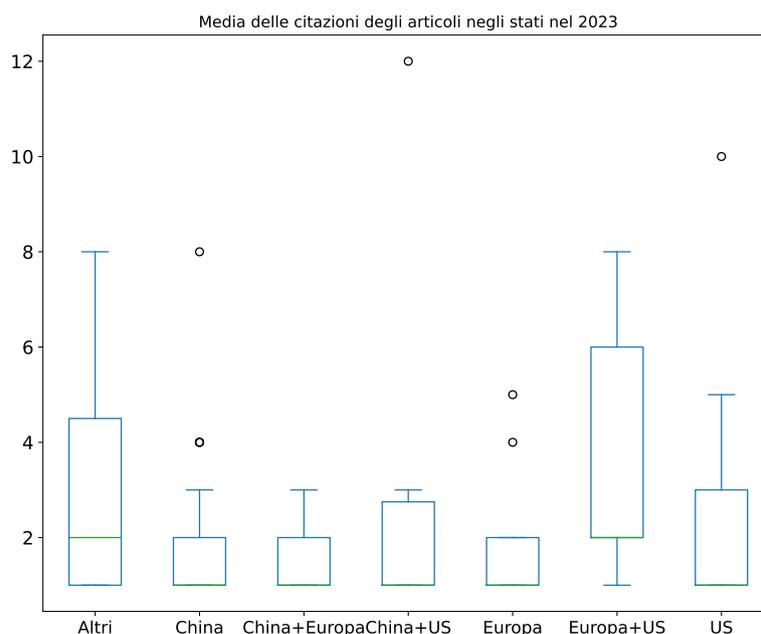


Figura 4.18: Conferenza INFOCOM citazioni medie per provenienza

In questo caso si vede un andamento diverso, ovvero ci sono dei massimi che arrivano anche ad otto, ovvero quelli del gruppo "Europa+Stati Uniti" e del gruppo "Altri". Anche la loro mediana si trova a 2 a differenza degli altri che è a uno. I valori estremi sono pochi in questo caso, vediamo un articolo che ha ottenuto dodici citazioni che fa parte del gruppo "China+US".

Infine l'ultimo boxplot rappresenta il numero medio di citazioni per paese della terza conferenza, ovvero la Percom.

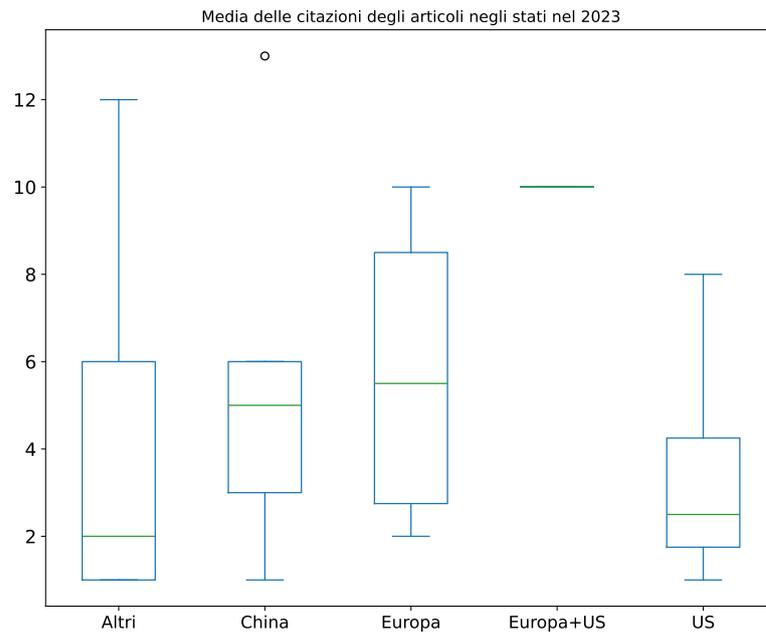


Figura 4.19: Conferenza PERCOM citazioni medie per provenienza

Si consideri che per questo grafico è stato utilizzato un campione ristretto di dati, poichè questa conferenza ha la media di circa 30 articoli all'anno presentati, per cui non sono molti, inoltre le citazioni aumentano col passare degli anni, quindi vien da se capire che i risultati di questo boxplot non possano essere del tutto strutturati.

Si commenta comunque notando delle mediane alte per tutti i gruppi, con anche degli intervalli di confidenza ampi come già si era notato nella Figura 5.1, tranne per il gruppo di articoli che hanno autori provenienti da Europa e Stati Uniti. Questo vuol dire che i pochi articoli proposti sono molto citati.

## 4.4 WordCloud

I WordCloud sono grafiche utili a mostrare quali sono gli argomenti più citati e trattati in un testo o ambiente di lavoro. E va ad evidenziare questi argomenti rendendo di dimensione maggiore le parole citate con maggiore frequenza.

In questo sottoparagrafo sono mostrati i tre wordcloud delle tre conferenze che

fanno riferimento all'edizione del 2023 e prendono in considerazione l'abstract e il titolo degli articoli.

Si vede prima la Globecom.



Figura 4.20: Conferenza GLOBECOM grafico word cloud

Come già scritto la Globecom è una conferenza sul networking, difatti la parola più grande è la parola "Network", poi seguono, come da aspettativa, le parole "System" e "Communication".

Dopo si considerino gli articoli della Infocom.

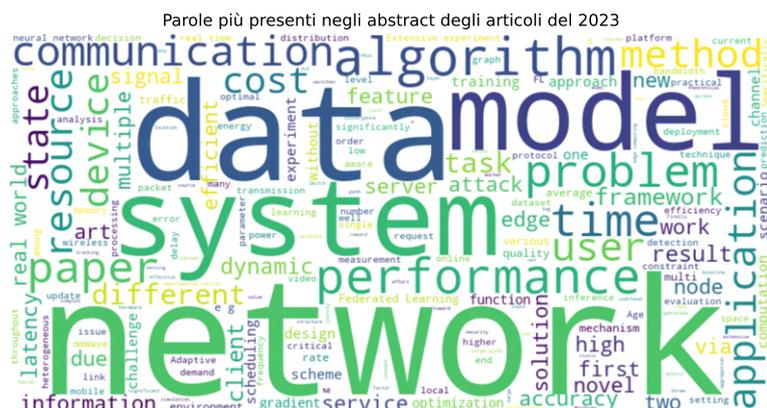


Figura 4.21: Conferenza INFOCOM grafico word cloud

La Infocom vede più frequentemente nei suoi articoli parole come "Data", "Network", "system" e "Model".

Per ultima viene mostrata la rilevanza degli argomenti degli articoli della Percom.

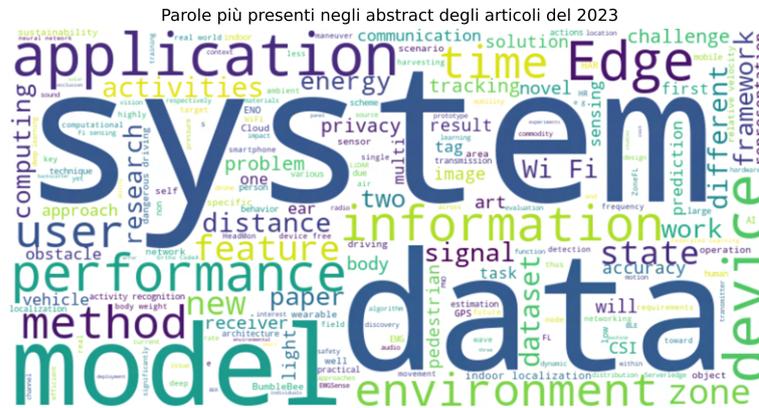


Figura 4.22: Conferenza PERCOM grafico word cloud

La Percom vede una netta presenza della parola "System", poi di dimensioni maggiori ci sono anche le parole "data" e "model".

I risultati di questi grafici sono attendibili in quanto le conferenze riguardano ambiti di networking e reti di comunicazione.

## 4.5 Riferimenti AI e Blockchain

### Artificial Intelligence

I sottostanti grafici vanno a sottolineare come negli anni sia incrementato il numero di riferimenti all'Intelligenza Artificiale

L'Intelligenza artificiale è presente ormai da più di 60 anni. La prima definizione è data dal Test di Turing come l'abilità delle macchine di comunicare con gli umani. Marvin Minsky diceva che è intelligenza artificiale se gli umani ritenevano intelligente quello che faceva quella macchina, altri pensano che dipenda dalle operazioni che esegue. [21]

Ovviamente le definizioni sono tante e varie, ma sicuramente più si va avanti e

più l'intelligenza delle macchine migliora ed è sempre più d'aiuto nelle attività quotidiane dell'uomo, grazie anche allo sviluppo tecnologico in generale.

Quindi è naturale che negli ultimi anni sia incrementato il numero di articoli che parlano di questo, come vediamo anche confermare dai sottostanti grafici.

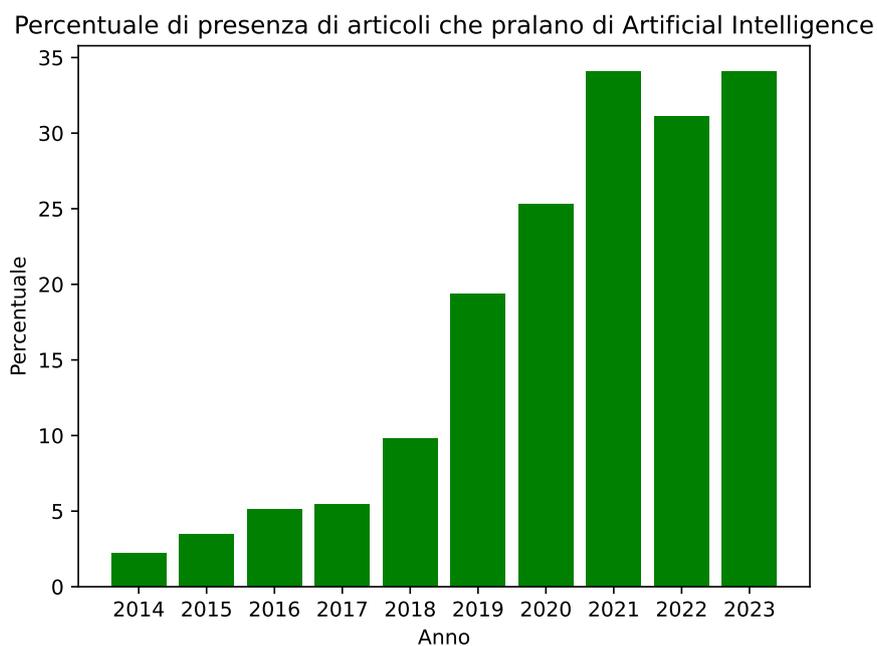


Figura 4.23: Conferenza GLOBECOM riferimento Artificial Intelligence

In questo grafico a barre si possono vedere i risultati dell'analisi eseguita sui dati della conferenza GLOBECOM e mostrano come effettivamente ci si aspetta, un incremento che va dal 5% in più, ogni anno, di articoli che parlano dell'intelligenza artificiale. Arrivando al giorno d'oggi a contare il 35% circa di articoli che riguardano l'argomento tratto nel paragrafo.

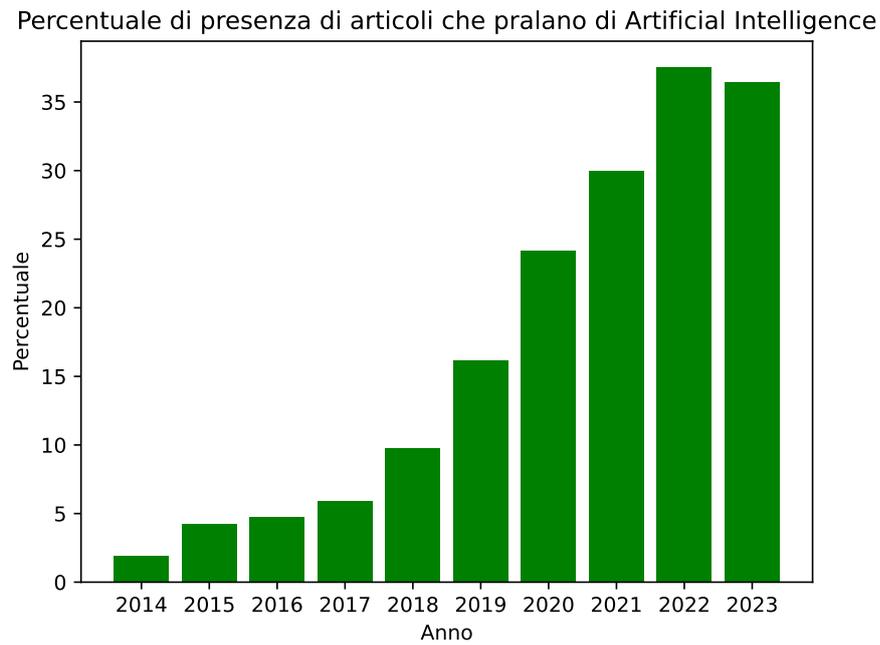


Figura 4.24: Conferenza INFOCOM riferimento Artificial Intelligence

Nella conferenza INFOCOM si può notare un andamento analogo. Arriva nel 2023 con circa il 35% di articoli che parlano dell'Intelligenza Artificiale, partendo dal 2014 con il 2%.

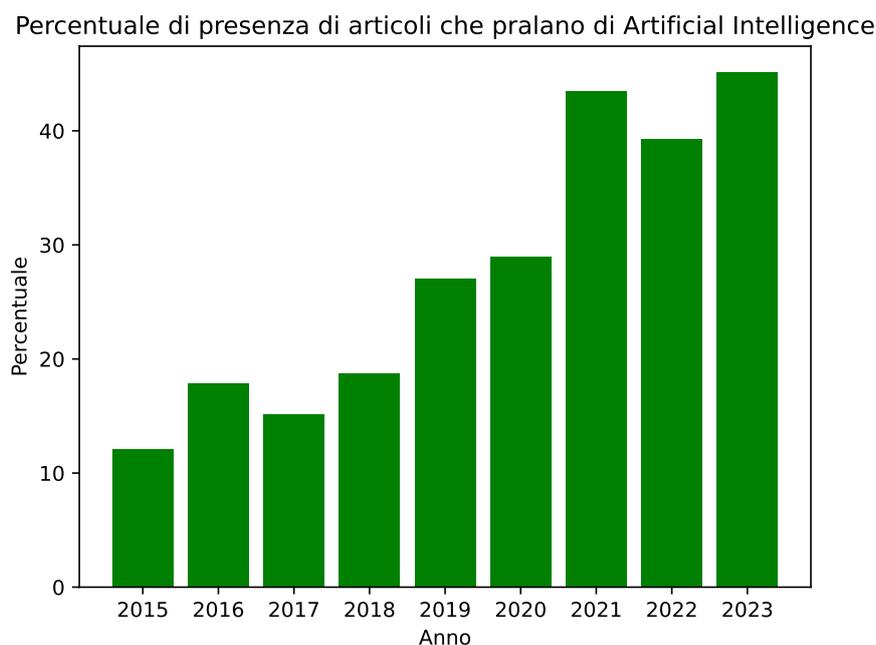


Figura 4.25: Conferenza PERCOM riferimento Artificial Intelligence

La Percom parte dal 2015 con una percentuale più alta rispetto alle altre due conferenze e mantiene poi lo stesso andamento circa, ottenendo nel duemilaventitré una percentuale del 40, ottenendo quindi la percentuale maggiore fra le tre conferenze.

## Blockchain

Si considera ora la Blockchain, una tecnologia che, come si vede anche dai seguenti grafici, si sviluppa di recente, o meglio si inizia ad utilizzare con l'avvenimento delle criptovalute. Infatti la Blockchain si occupa di quelle che sono le piattaforme per il trading di crypto e degli smart contracts. [22]

Le criptovalute sono delle valute, appunto, in forma virtuale e scambiabile solo per via telematica. Come si intuisce anche dal nome sono delle valute che vengono crittografate per far sì che possano essere visualizzate ed utilizzate solo da chi possiede la chiave, ovvero un codice informatico, per decrittografarla. [23]

Gli smart contracts sono dei codici sorgenti che vengono compilati in modo da avere dei contratti efficienti e senza ambiguità contrattuali fra le parti. [24]

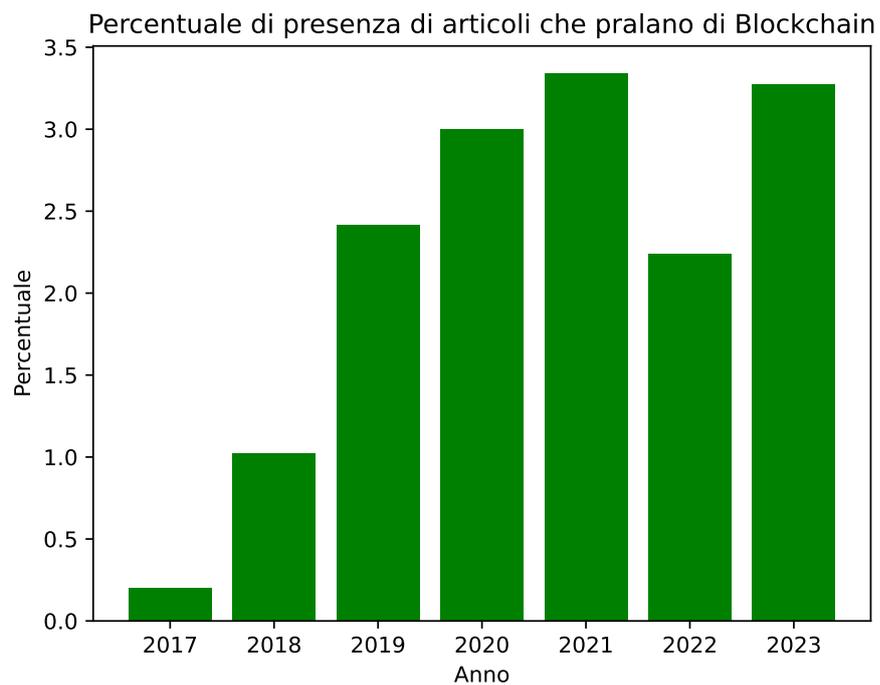


Figura 4.26: Conferenza GLOBECOM riferimento Blockchain

Nella GLOBECOM i primi articoli che discutono della blockchain compaiono nel 2017 e non si fermano, se non nel 2022 che c'è un calo di numeri, ma viene recuperato poi nell'anno successivo.

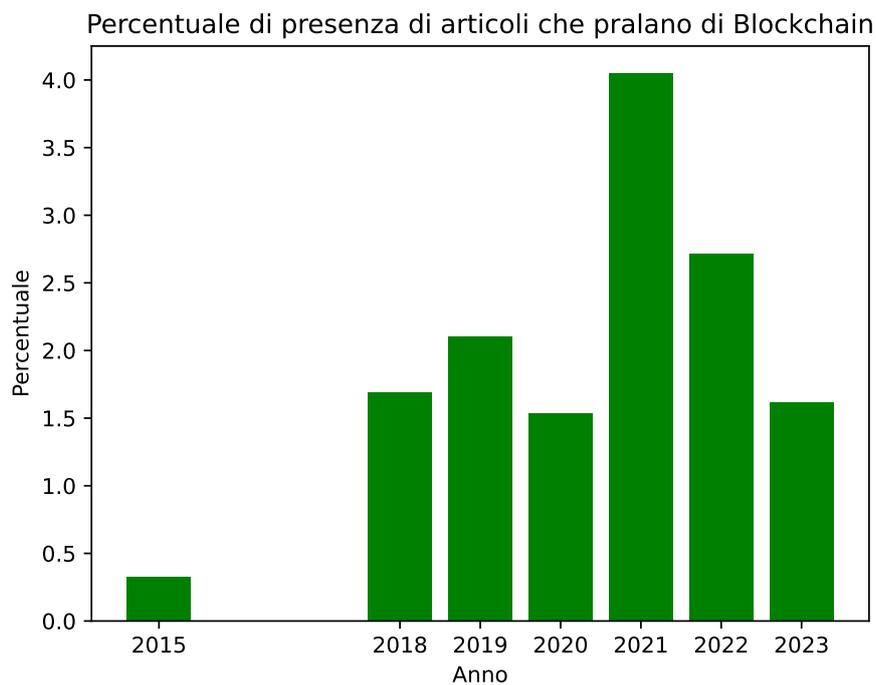


Figura 4.27: Conferenza INFOCOM riferimento Blockchain

In questa conferenza invece si cita per la prima volta la blockchain nel 2015, dopodiché non verrà nominata per due anni, fino al 2018, da qui in poi verrà sempre presa in considerazione anche se poco.

Si vede un picco del 4 per cento degli articoli promossi dalla INFOCOM che parla di blockchain.

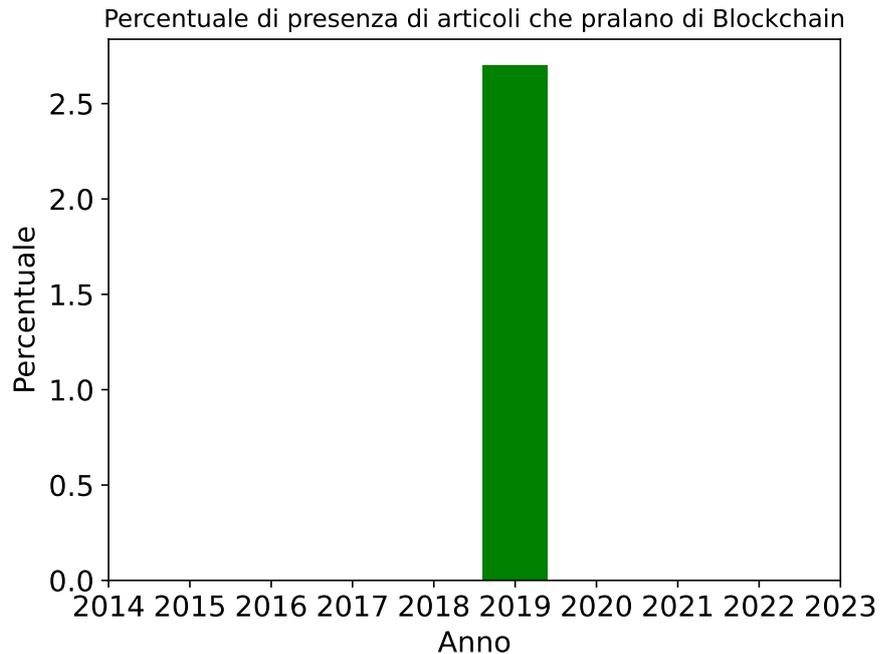


Figura 4.28: Conferenza PERCOM riferimento Blockchain

Il grafico sovrastante mostra che alla Percom solo nel 2019 sono stati proposti articoli che hanno parlato di Blockchain, questi sono circa poco più del 2,5%. Considerando che ci sono circa 33 articoli presentati alla Percom nel 2023, il 2,5% equivale ad un solo articolo, per cui solo un articolo parla della Blockchain in tutte le edizioni, prese in considerazione in questa tesi, della Percom.

## 4.6 Correlazioni

Data la precedente definizione si è provato, tramite l'analisi, ad individuare se fosse presente qualche correlazione. Più precisamente si prova a rispondere alle domande: - il numero di autori influisce sul numero di citazioni? - se un articolo ha autori di diversi stati permette di avere più citazioni? - il tipo di affiliazione degli autori aiuta nell'avere più citazioni? Da qui, dopo che si è studiata la possibilità di quali correlazioni calcolare, Pearson o Kendall, si è proceduto ottenendo i seguenti grafici.

Verranno mostrati però solo quelli ottenuti con il coefficiente di Kendall, poichè è l'unico che può mostrare dei risultati.

### Numero di citazioni e stato di provenienza

In primis guardiamo la correlazione calcolata tra il numero di citazioni e lo stato di provenienza.

Come già scritto in precedenza il valore del coefficiente di correlazione di Kendall va da +1 a -1 con una correlazione perfettamente positiva se il valore è vicino al +1, una concordanza perfettamente negativa se è vicino a -1 e nessuna correlazione se il valore è zero.

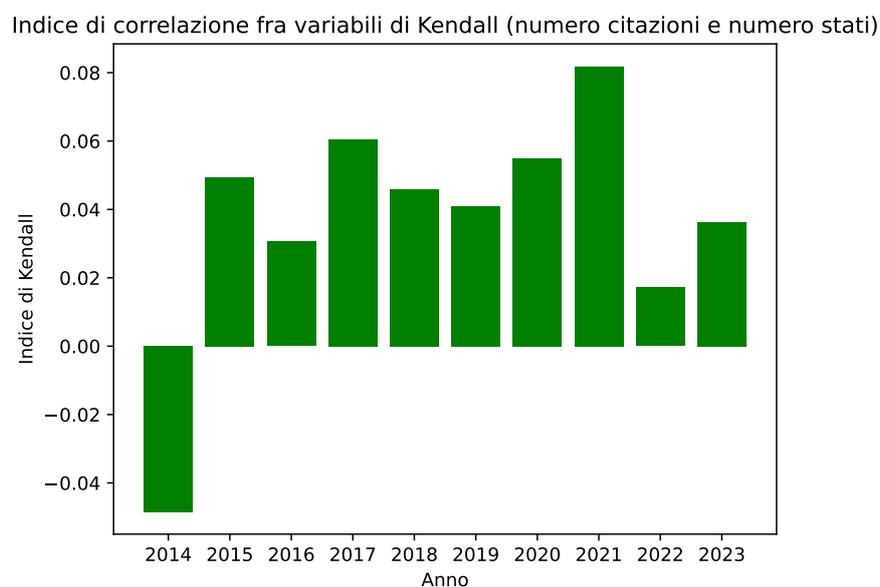


Figura 4.29: Grafico correlazione di Kendall tra numero citazioni e stato di provenienza, conferenza GLOBECOM

Per cui tenendo conto delle indicazioni dei valori, il grafico sopra riportato, che fa riferimento alla conferenza Globecom, mostra in linea generale dei valori molto vicini allo zero per cui esiste molta poca concordanza tra le due variabili.

Prendiamo ora la Infocom.

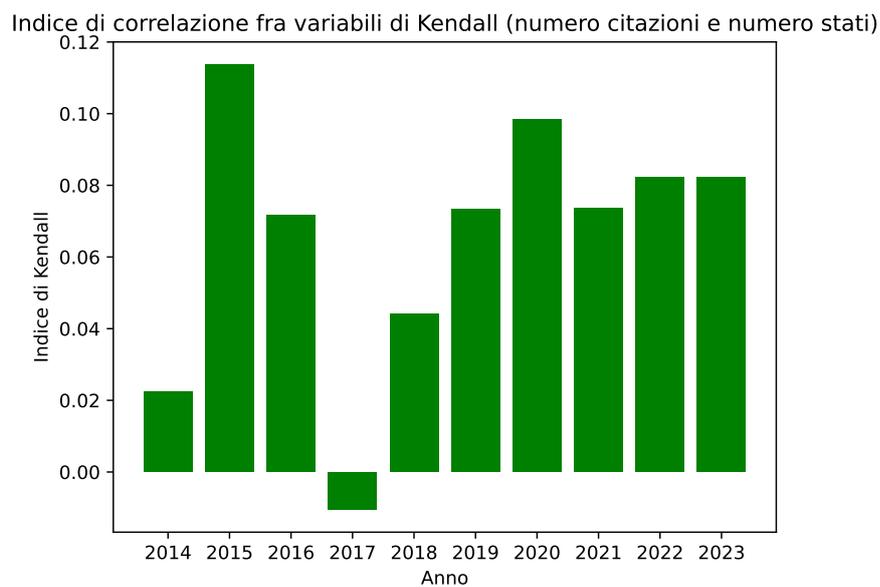


Figura 4.30: Grafico correlazione di Kendall tra numero citazioni e stato di provenienza, conferenza INFOCOM

In questo caso, non ci sono valori comunque molto alti, però si arriva a superare lo 0,1 in alcuni anni, ma questo non porta a dire che esista effettivamente una correlazione tra le due variabili.

Infine si vede la Percom.

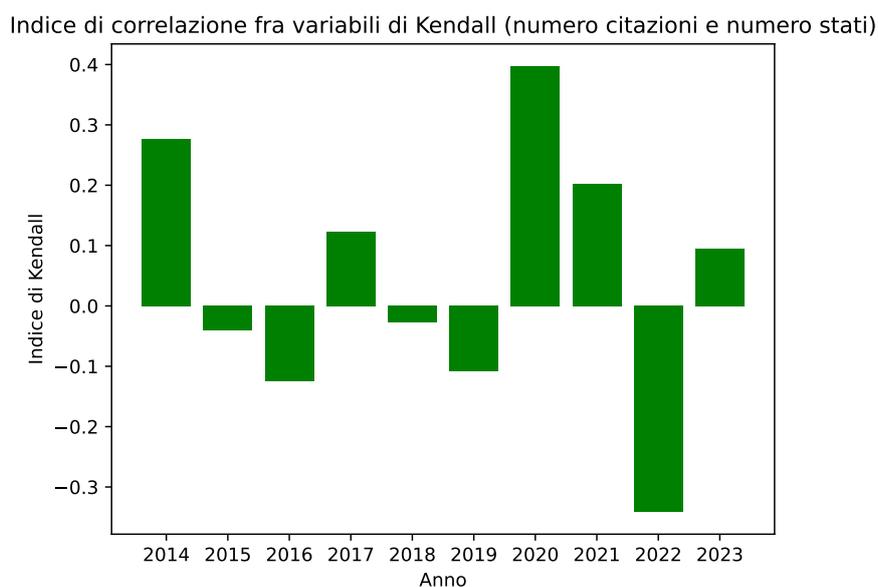


Figura 4.31: Grafico correlazione di Kendall tra numero citazioni e stato di provenienza, conferenza PERCOM

In questo grafico a barre si può notare un andamento molto discontinuo negli anni sui valori della correlazione, ciò è attendibile anche per il fatto che il numero di articoli è piccolo per cui con un campione di circa 30 articoli non si riesce a trovare un valore stabile. Comunque vediamo che in alcuni anni arriva ad un valore positivo di 0.4 e altri fino a -0.3, mentre altri ancora rimangono intorno allo zero.

In conclusione non si può dire che esista per nessuna delle tre conferenze una correlazione fra il numero di citazioni che ottiene un articolo e la sua provenienza.

### Numero di citazioni e numero di autori

Si risponde ora alla domanda: "Il numero di autori che partecipano alla stesura di un articolo, influisce sul numero di citazione che esso ottiene?".

Si procede nello stesso ordine di sempre per le tre conferenze.

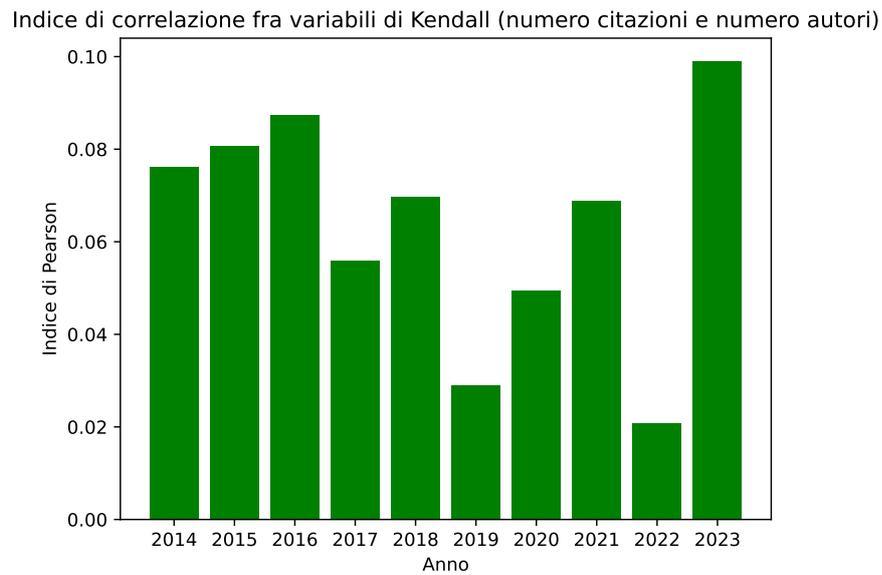


Figura 4.32: Grafico correlazione di Kendall tra numero citazioni e numero di autori, conferenza GLOBECOM

Per la Globecom si notano innanzitutto sempre valori positivi per tutti i dieci anni presi in considerazione, ma sono comunque tutti valori molto bassi. Non superano mai lo 0.1, di conseguenza non esiste un'effettiva concordanza fra queste due variabili.

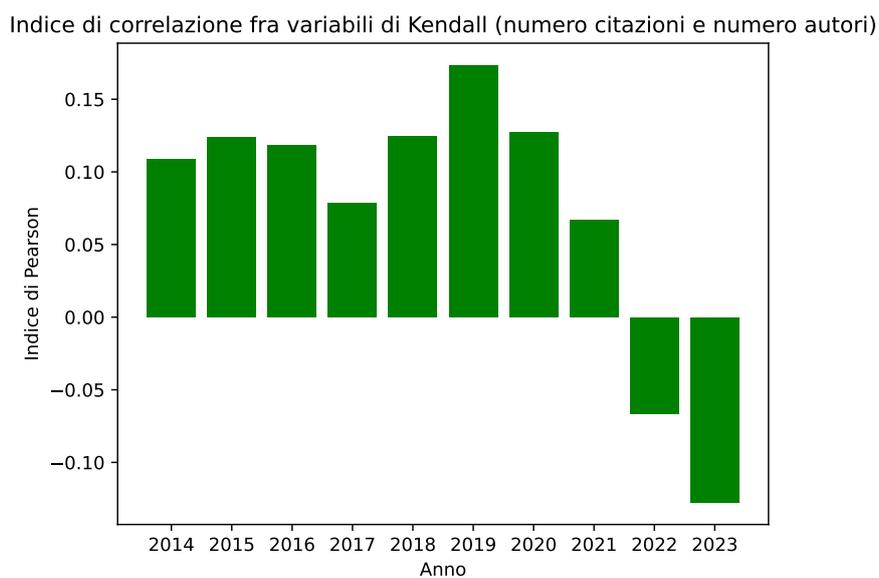


Figura 4.33: Grafico correlazione di Kendall tra numero citazioni e numero di autori, conferenza INFOCOM

Per la seconda conferenza si tornano a vedere valori negativi ma sempre vicini allo zero, per cui anche qui non si può parlare di correlazione.

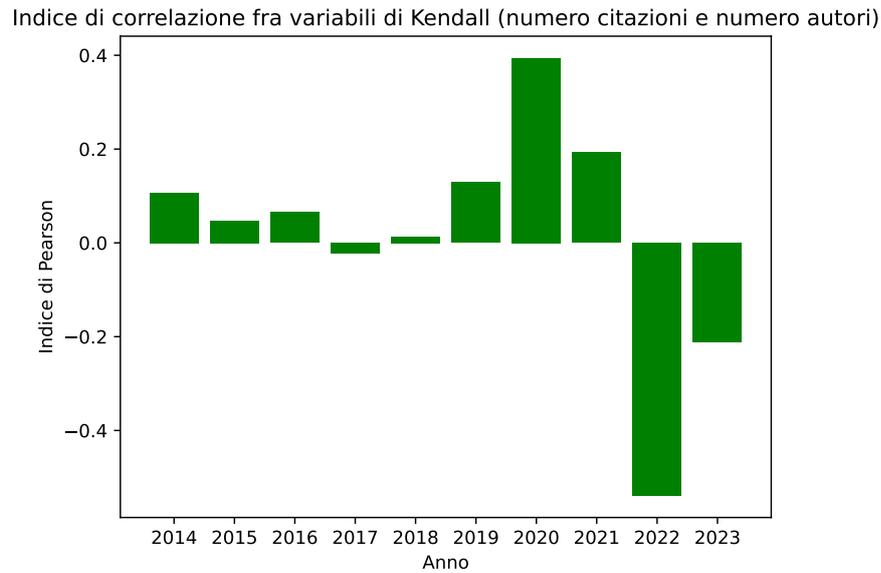


Figura 4.34: Grafico correlazione di Kendall tra numero citazioni e numero di autori, conferenza PERCOM

Mentre in quest'ultimo grafico riguardante la Percom, vediamo valori arrivare fino a 0.4 e -0.4, ma vale lo stesso discorso fatto poche righe sopra, ovvero che il campione di riferimento è piccolo e il risultato non è del tutto attendibile.

### Numero di citazioni e tipo di affiliazione

Terzo tipo di analisi di correlazione, ovvero quella che prende come variabili sempre il numero di citazioni e stavolta il tipo di affiliazione, cioè se gli autori dell'articolo in questione hanno tutti un'affiliazione accademica, puramente aziendale oppure mista. In questo caso il coefficiente sarà per forza di Kendall, al contrario delle due precedenti che è stato scelto volutamente di mostrare solo quello, qui è stata l'unica opzione, poiché con il tipo di affiliazione si parla di una variabile qualitativa e Pearson non accetta variabili che non siano prettamente quantitative.

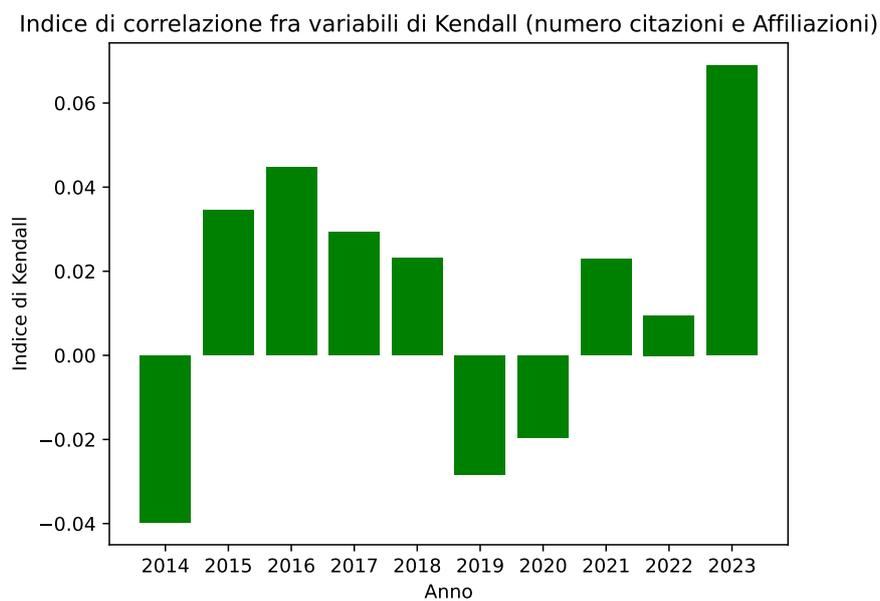


Figura 4.35: Grafico correlazione di Kendall tra numero citazioni e tipo di affiliazione, conferenza GLOBECOM

Il grafico mostra ciò che già ci si aspettava, ovvero che, anche qui come per le altre volte, si mantengono dei valori intorno allo zero per cui non vi è nessuna dipendenza fra le due variabili.

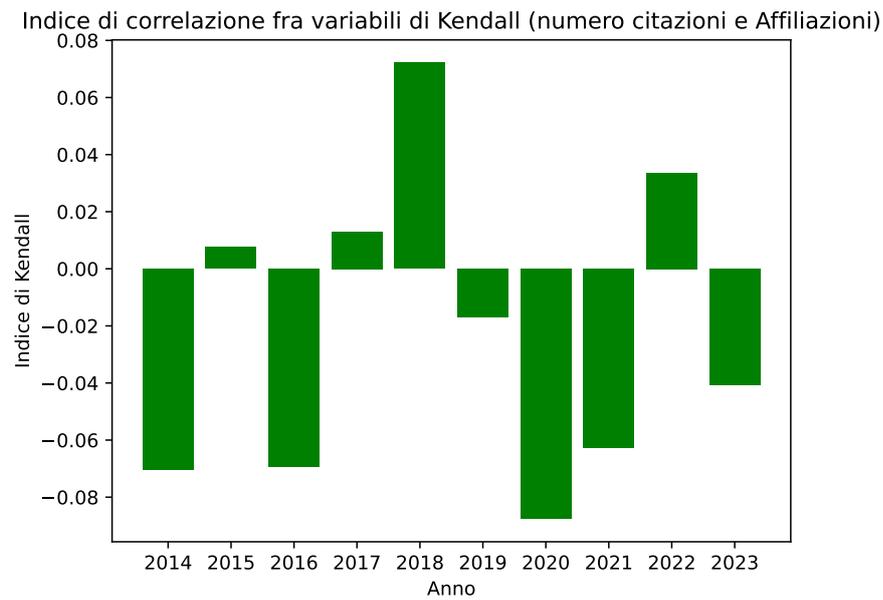


Figura 4.36: Grafico correlazione di Kendall tra numero citazioni e tipo di affiliazione, conferenza INFOCOM

Vale la stessa cosa per il grafico che utilizza i dati della Infocom, i valori oscillano in un range che va dal -0.8 al +0.8, anche qui non si può dire esserci una rilevante corrispondenza fra numero di citazioni e tipo di affiliazione.

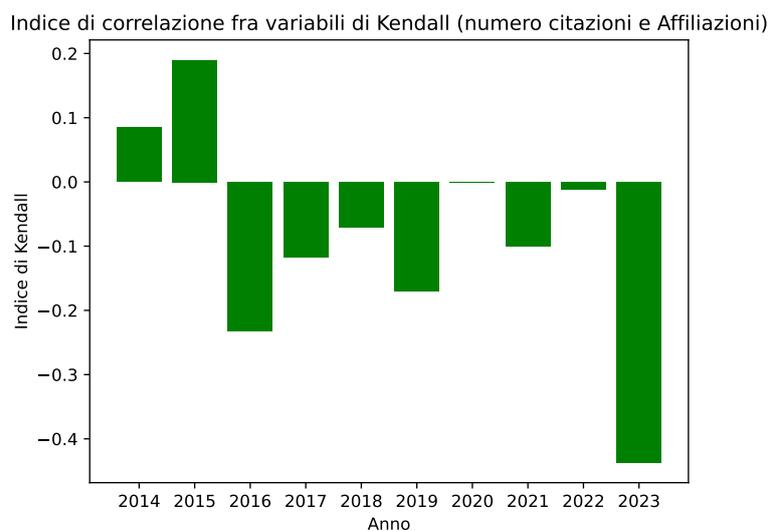


Figura 4.37: Correlazione di Kendall per tipo di affiliazione, conferenza Percom

Non fa differenza per questa conferenza, il risultato, i valori sono tutti intorno allo zero per cui non esiste una correlazione fra le variabili.

Queste mancanze di correlazione possono essere dovute da vari fattori come un campione non abbastanza grande oppure una semplice distribuzione casuale che non porta ad alcuna correlazione.

# Capitolo 5

## Conclusioni e Sviluppi Futuri

La tesi si è occupata di analisi dati relativi alle conferenze scientifiche in ambito comunicazione.

Si sono, quindi, potute visionare alcune caratteristiche di queste conferenze scientifiche promosse dalla IEEE. Per esempio è emerso che fino a dieci anni fa il paese che più promuoveva articoli nelle tre conferenze erano gli Stati Uniti, mentre oggi quella con un numero maggiore di articoli pubblicati è la Cina.

Prima di tutto è fondamentale ricordare che si tratta di tre conferenze con cardinalità molto diverse tra loro, la Globecom è quella più grande con più di mille articoli, ma in questa analisi ne vengono presi solo mille come campione. La Infocom arriva a circa 300 pubblicazioni all'anno. Mentre la Percom conta solo intorno ai 30 articoli, per questa ragione non abbiamo considerato il valore assoluto ma quello percentuale.

Un altro risultato interessante ottenuto è che col passare degli anni, come ci si aspettava, il numero di articoli che discute sull'Intelligenza Artificiale cresce; mentre il numero di quelli che trattano come argomento la Blockchain è variabile, anche se sicuramente viene più toccato negli ultimi anni che in quelli precedenti. Un'analisi che poteva risultare interessante era quella sulle correlazioni, per esempio trovare qualcosa che influenzasse il numero di citazioni che ottiene un articolo. I risultati indicano che nessuna variabile come stato di provenienza, tipo di affiliazione o numero di autori è correlato al numero di citazioni che ottiene un articolo.

Mentre invece si è notato come sia più alto il numero medio di citazioni ottenuto dalle conferenze più "piccole" come Infocom e Percom, piuttosto che negli articoli pubblicati in incontri più grandi come la Globecom che per via di cose ha molti articoli con zero citazioni che abbassano la media, oltre al fatto che quelli che ottengono citazioni ne ottengono tra le 0 e le 10.

Come sviluppi futuri si potrebbe fare un'analisi più approfondita sotto altri aspetti, per esempio incrociando il paese di provenienza con il tipo di affiliazione dell'articolo.

Inoltre si potrebbero prendere in considerazione altre conferenze sia della IEEE che non, anche per capire quanto pesa il contributo della IEEE effettivamente in tutto il mondo della ricerca scientifica.

# Appendice

## 5.1 Globecom

Citazioni stati

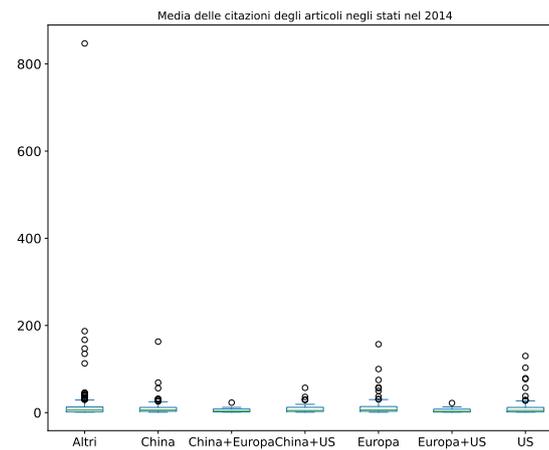


Figura 5.1: Numero di citazioni per gruppo, 2014

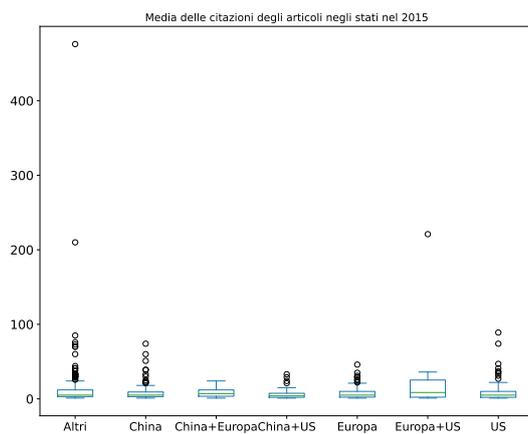


Figura 5.2: Numero di citazioni per gruppo, 2015

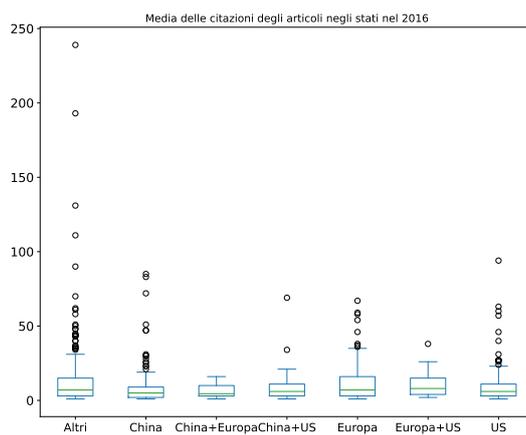


Figura 5.3: Numero di citazioni per gruppo, 2016

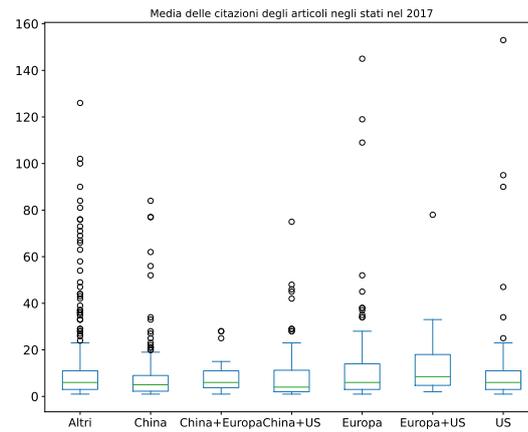


Figura 5.4: Numero di citazioni per gruppo, 2017

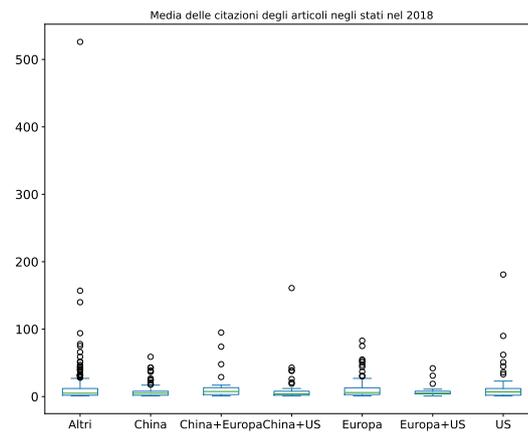


Figura 5.5: Numero di citazioni per gruppo, 2018

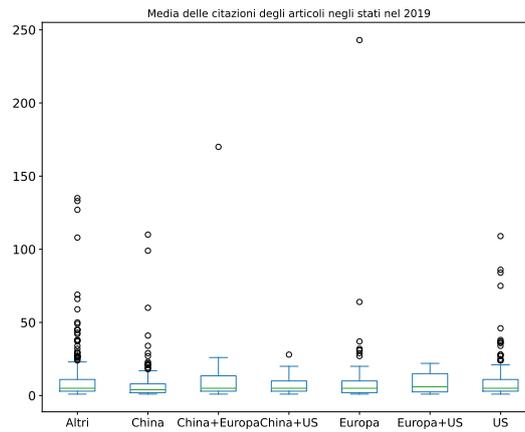


Figura 5.6: Numero di citazioni per gruppo, 2019

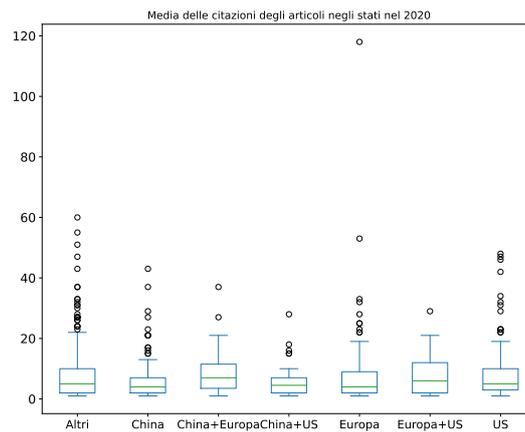


Figura 5.7: Numero di citazioni per gruppo, 2020

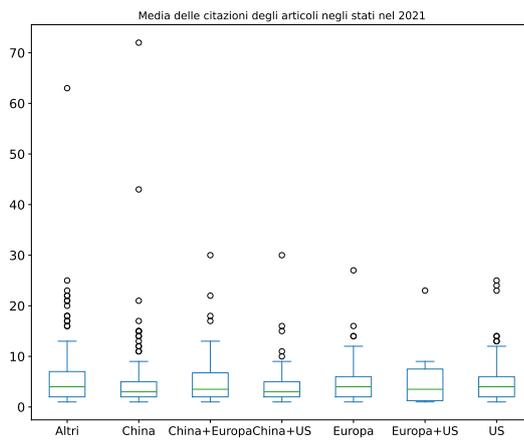


Figura 5.8: Numero di citazioni per gruppo, 2021

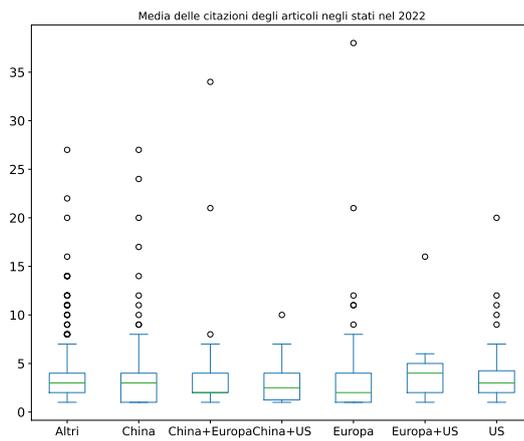


Figura 5.9: Numero di citazioni per gruppo, 2022

## Distribuzioni

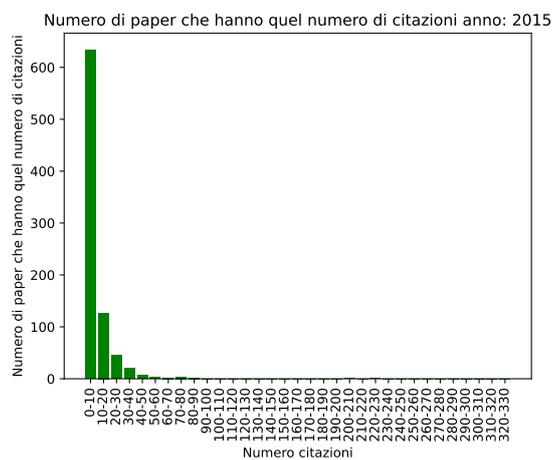


Figura 5.10: Distribuzione del numero di citazioni, 2015

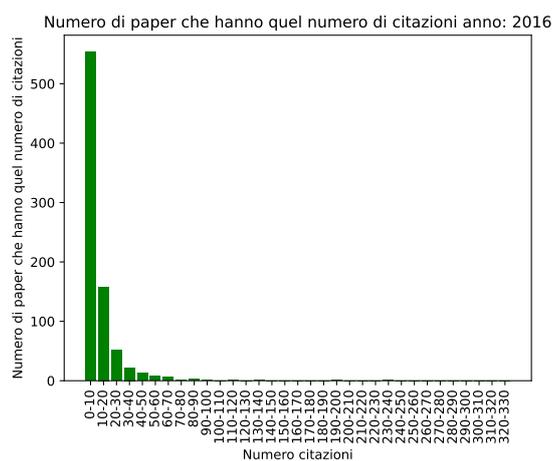


Figura 5.11: Distribuzione del numero di citazioni, 2016

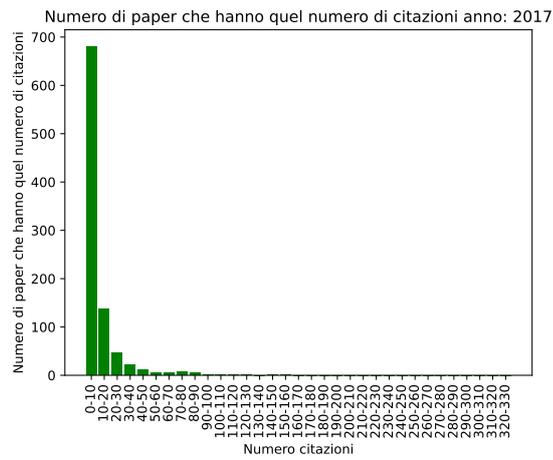


Figura 5.12: Distribuzione del numero di citazioni, 2017

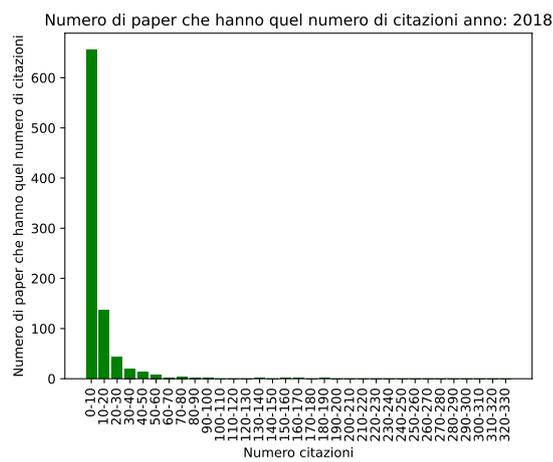


Figura 5.13: Distribuzione del numero di citazioni, 2018

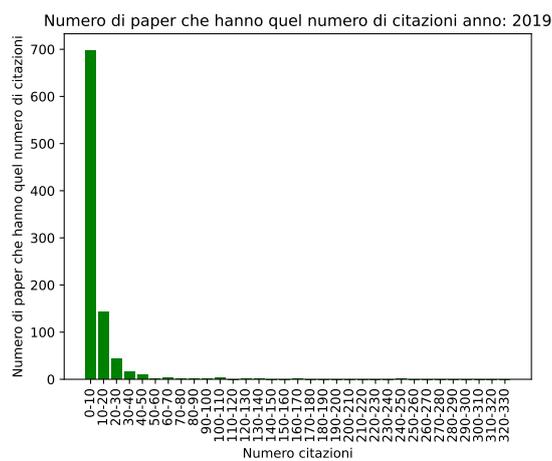


Figura 5.14: Distribuzione del numero di citazioni, 2019

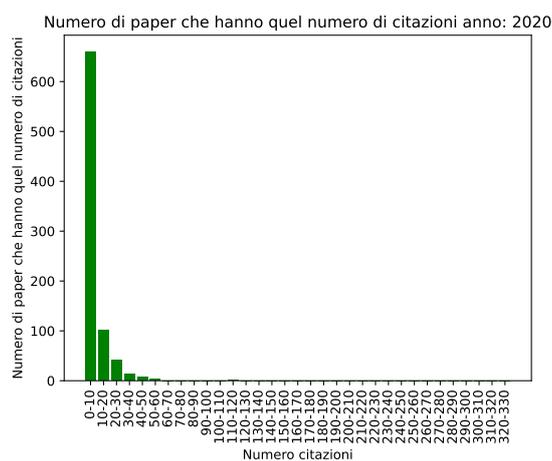


Figura 5.15: Distribuzione del numero di citazioni, 2020

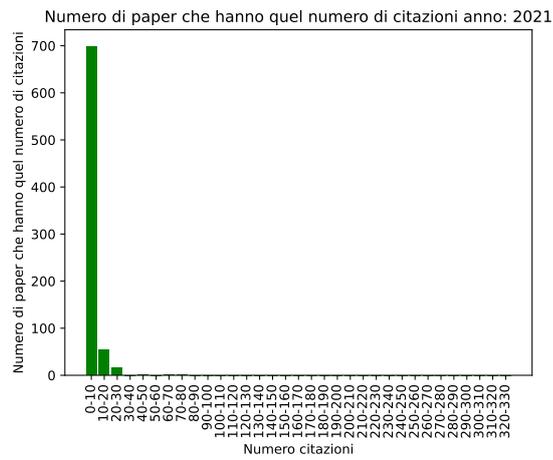


Figura 5.16: Distribuzione del numero di citazioni, 2021

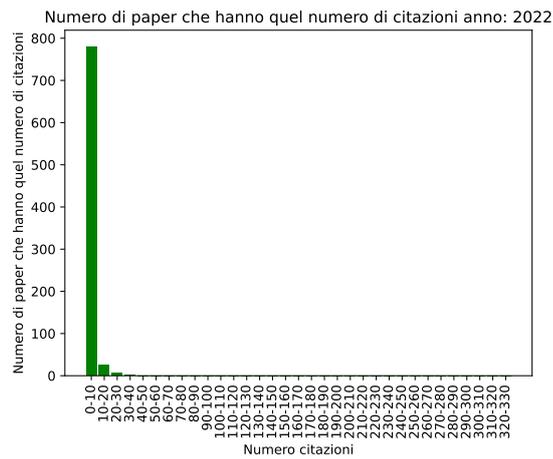


Figura 5.17: Distribuzione del numero di citazioni, 2022

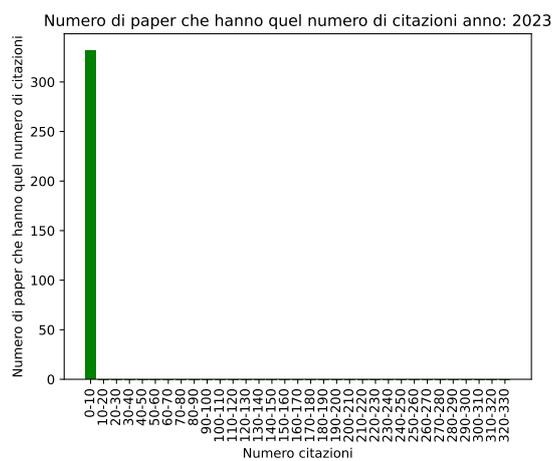


Figura 5.18: Distribuzione del numero di citazioni, 2023

## Provenienze

Numero di stati rappresentati nel 2014

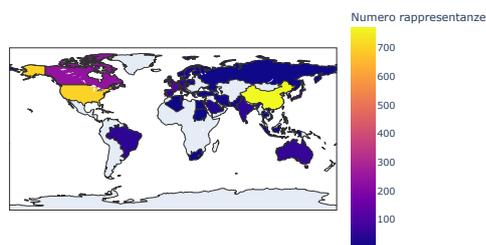


Figura 5.19: Mappa provenienze articoli, 2014

Numero di stati rappresentati nel 2015

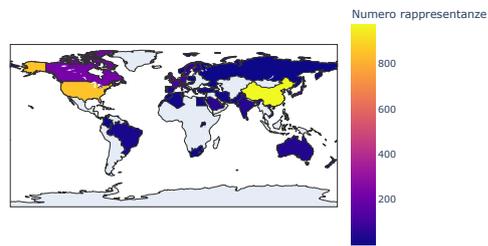


Figura 5.20: Mappa provenienze articoli, 2015

Numero di stati rappresentati nel 2016

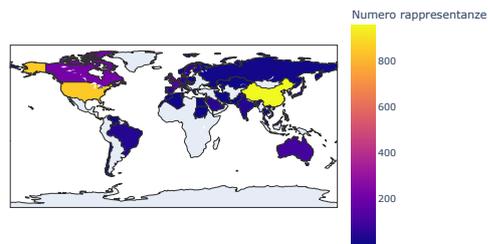


Figura 5.21: Mappa provenienze articoli, 2016

Numero di stati rappresentati nel 2017

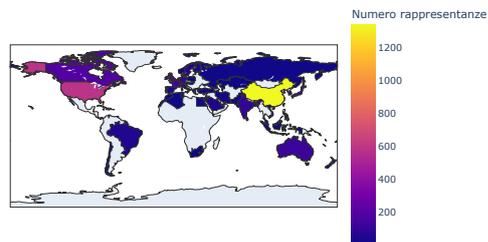


Figura 5.22: Mappa provenienze articoli, 2017

Numero di stati rappresentati nel 2018

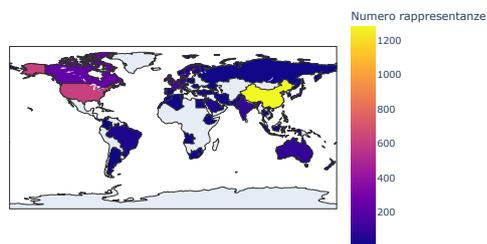


Figura 5.23: Mappa provenienze articoli, 2018

Numero di stati rappresentati nel 2019

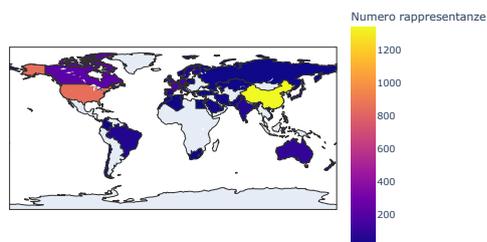


Figura 5.24: Mappa provenienze articoli, 2019

Numero di stati rappresentati nel 2020

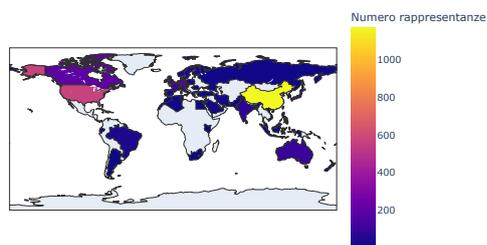


Figura 5.25: Mappa provenienze articoli, 2020

Numero di stati rappresentati nel 2021

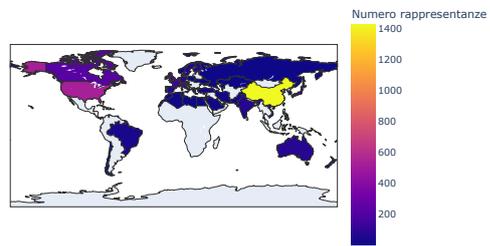


Figura 5.26: Mappa provenienze articoli, 2021

Numero di stati rappresentati nel 2022

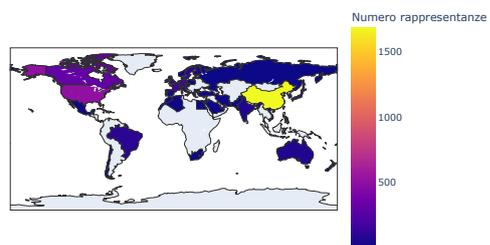


Figura 5.27: Mappa provenienze articoli, 2022

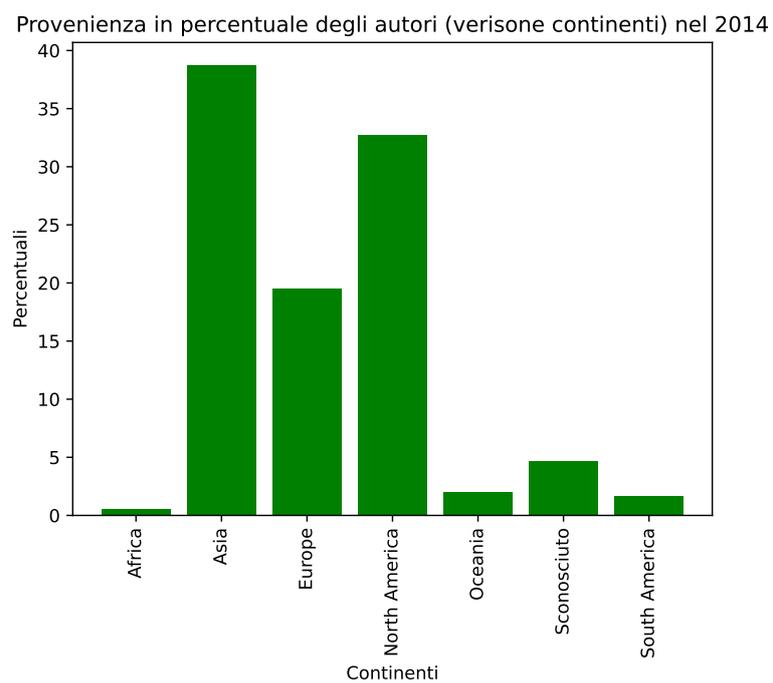


Figura 5.28: Percentuali provenienze articoli, continenti, 2014

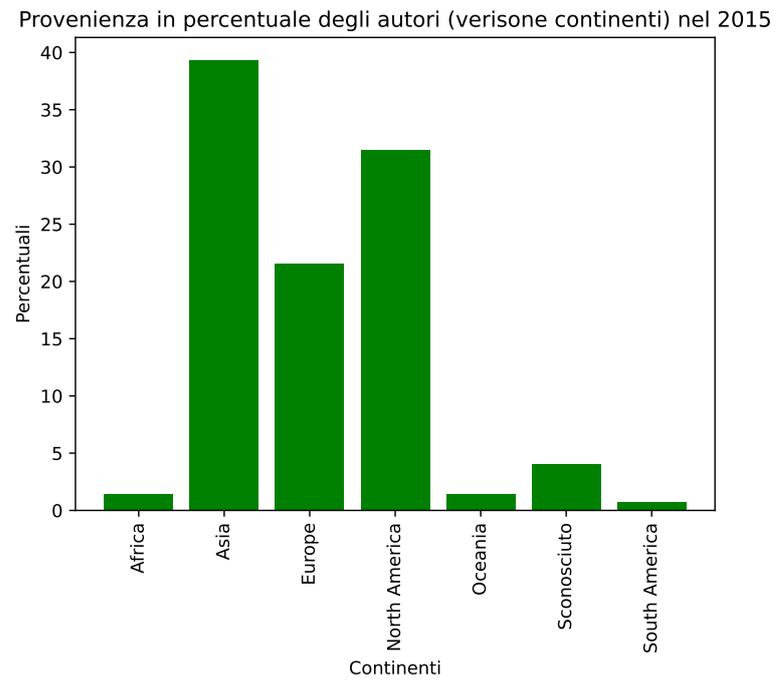


Figura 5.29: Percentuali provenienze articoli, continenti, 2015

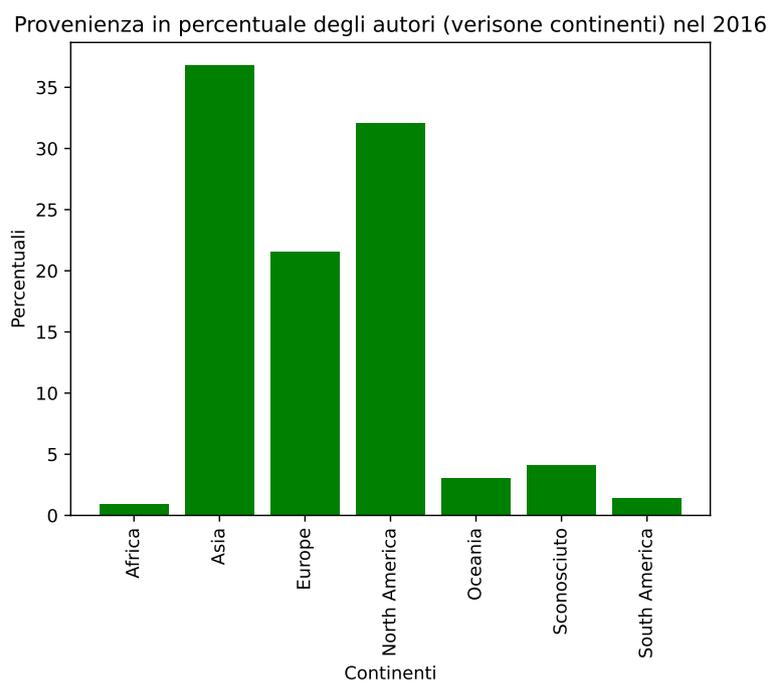


Figura 5.30: Percentuali provenienze articoli, continenti, 2016

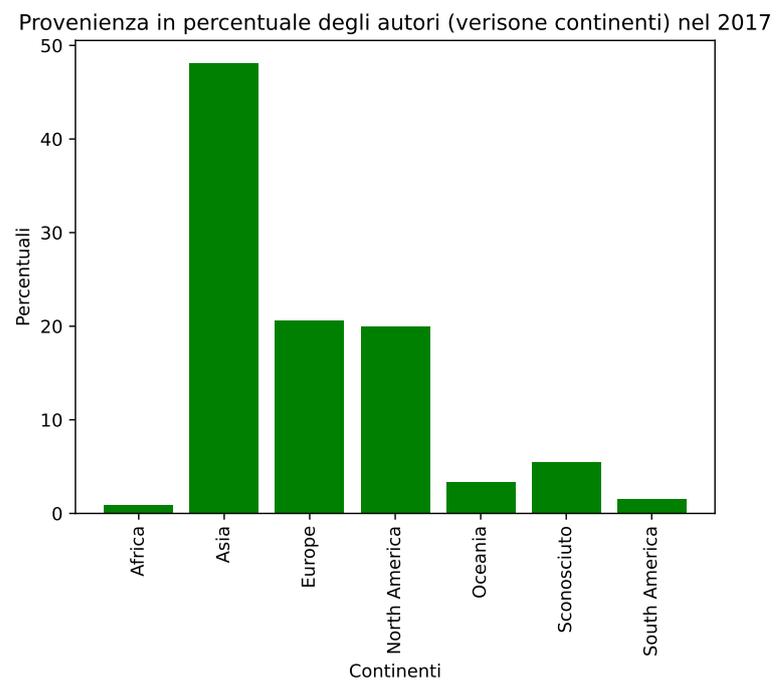


Figura 5.31: Percentuali provenienze articoli, continenti, 2017

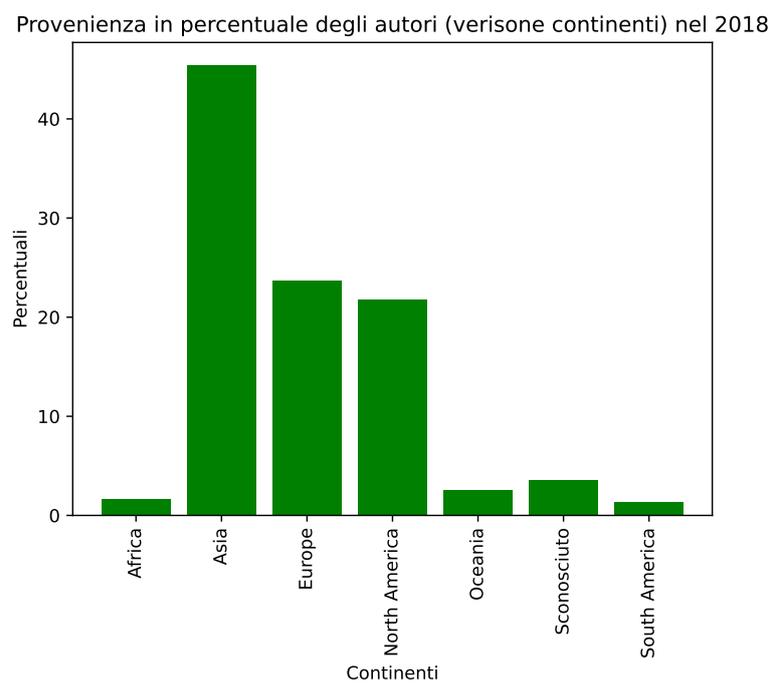


Figura 5.32: Percentuali provenienze articoli, continenti, 2018

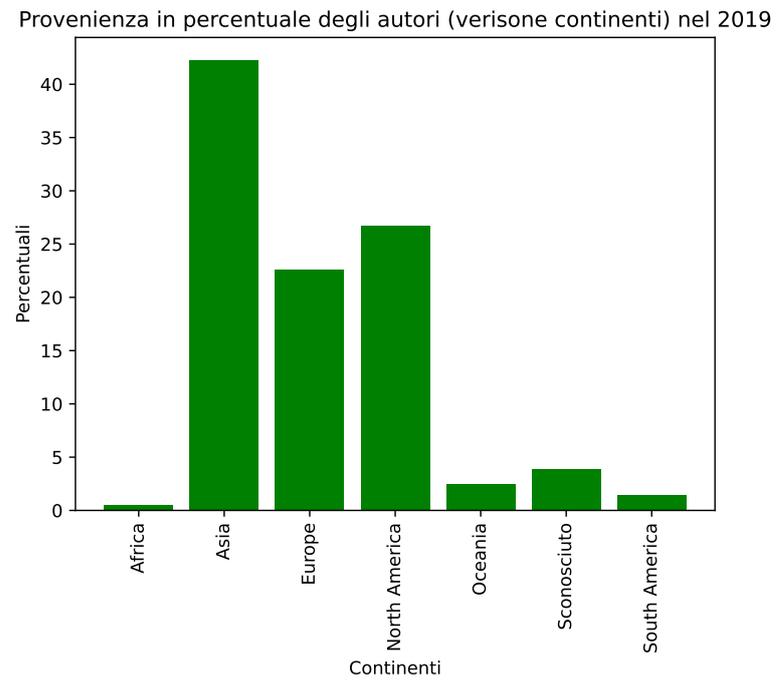


Figura 5.33: Percentuali provenienze articoli, continenti, 2019

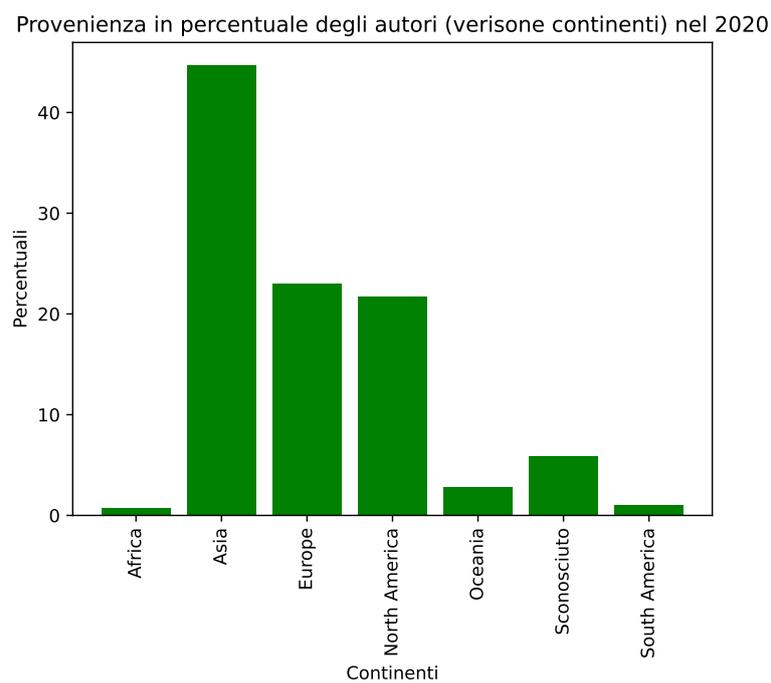


Figura 5.34: Percentuali provenienze articoli, continenti, 2020

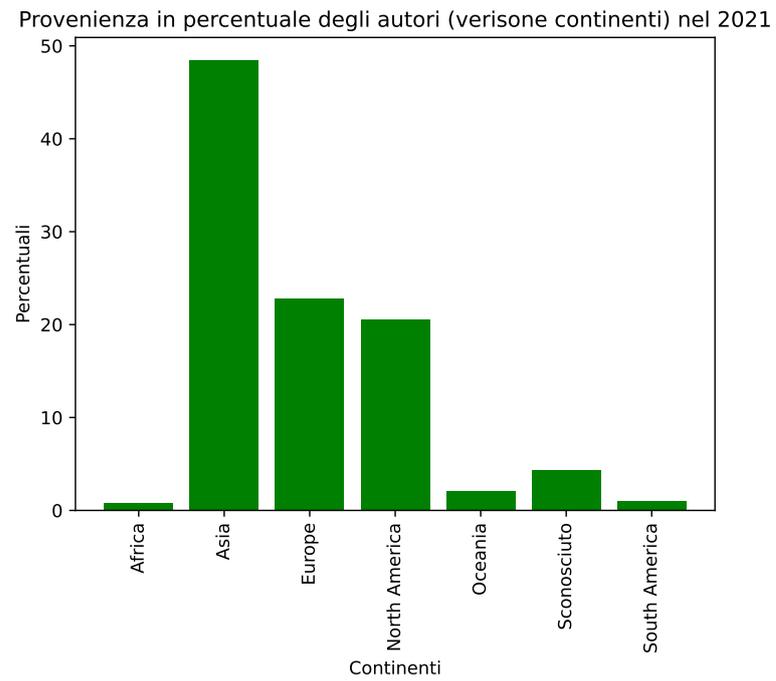


Figura 5.35: Percentuali provenienze articoli, continenti, 2021

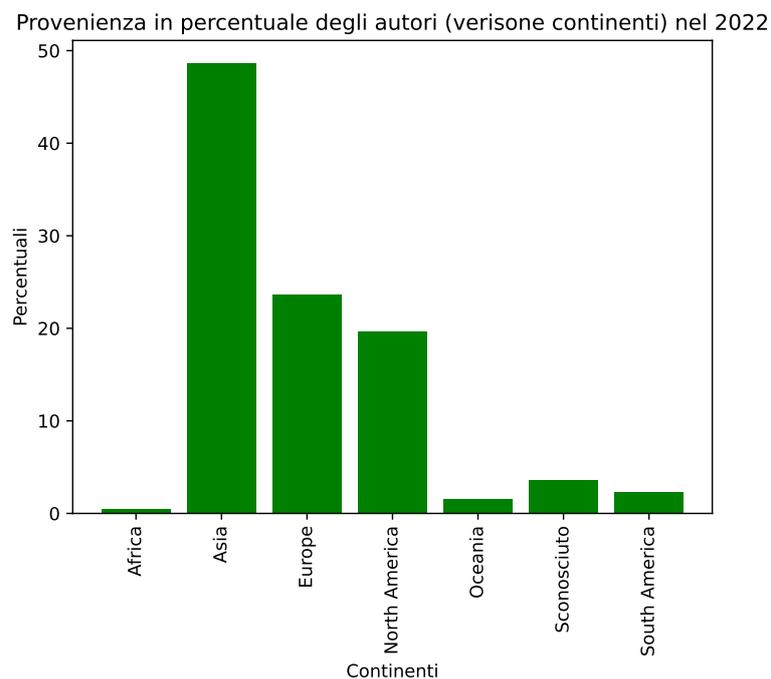


Figura 5.36: Percentuali provenienze articoli, continenti, 2022

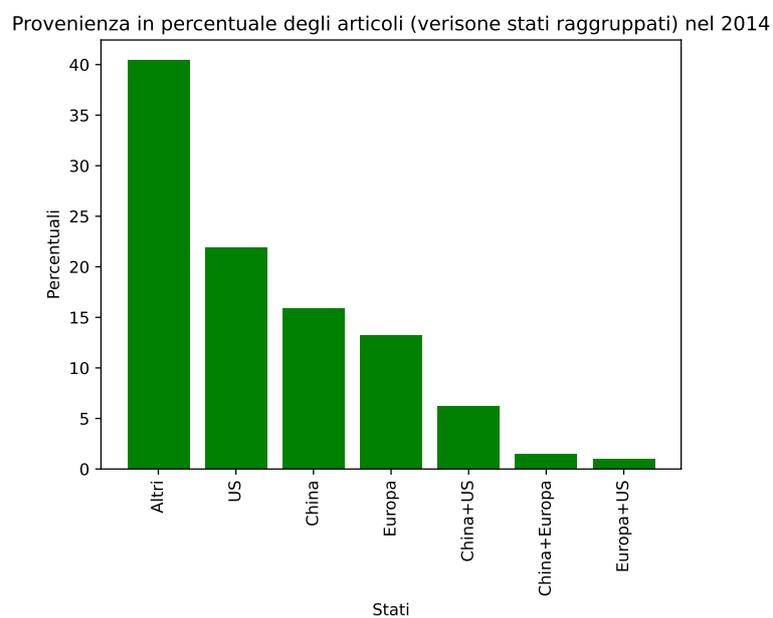


Figura 5.37: Percentuali provenienze articoli a gruppi, 2014

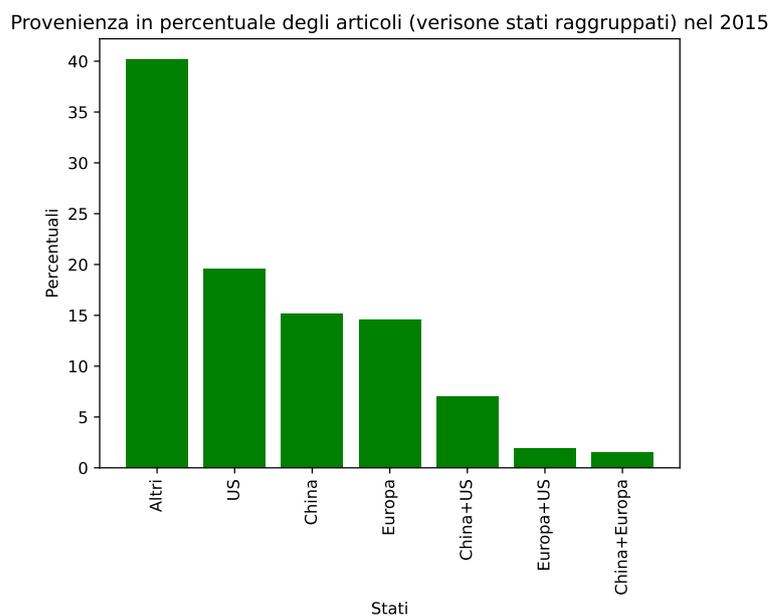


Figura 5.38: Percentuali provenienze articoli a gruppi, 2015

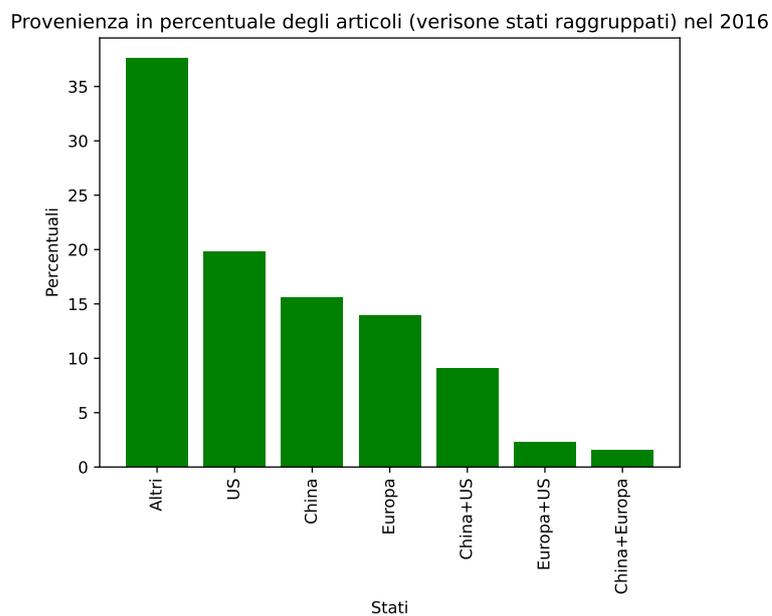


Figura 5.39: Percentuali provenienze articoli a gruppi, 2016

Provenienza in percentuale degli articoli (verisone stati raggruppati) nel 2017

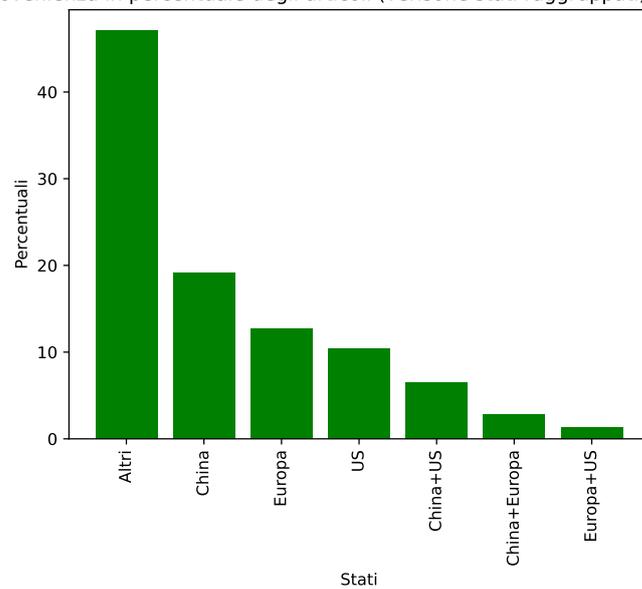


Figura 5.40: Percentuali provenienze articoli a gruppi, 2017

Provenienza in percentuale degli articoli (verisone stati raggruppati) nel 2018

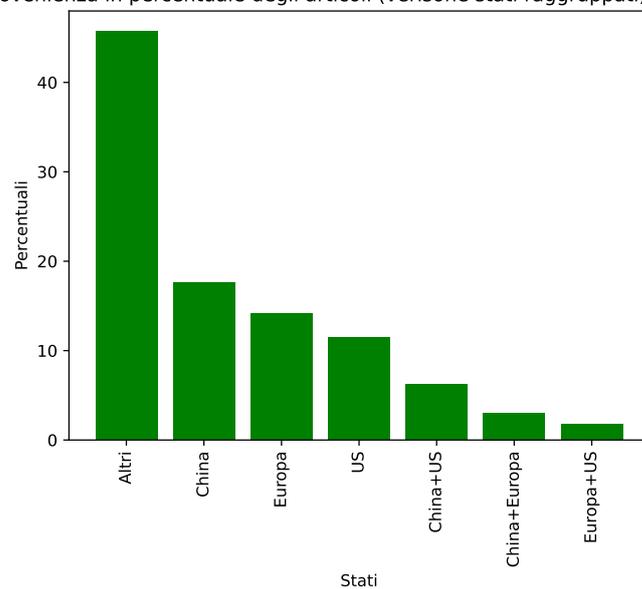


Figura 5.41: Percentuali provenienze articoli a gruppi, 2018

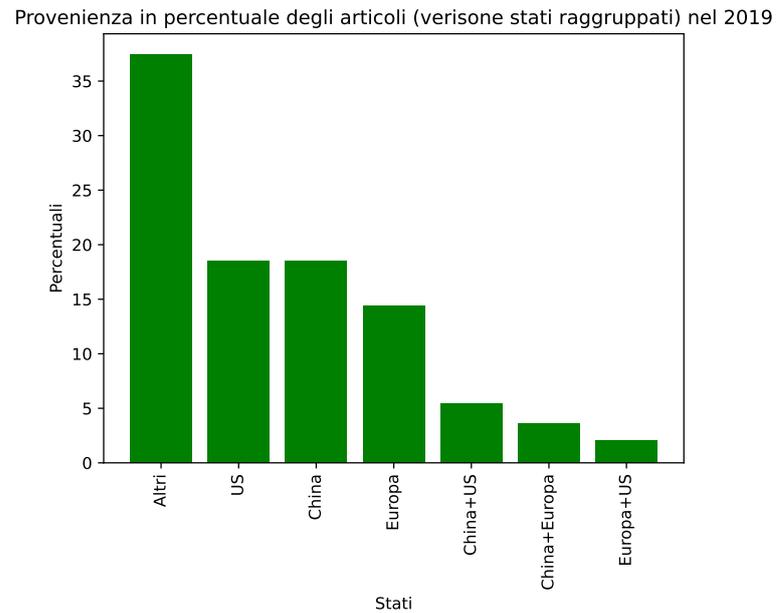


Figura 5.42: Percentuali provenienze articoli a gruppi, 2019

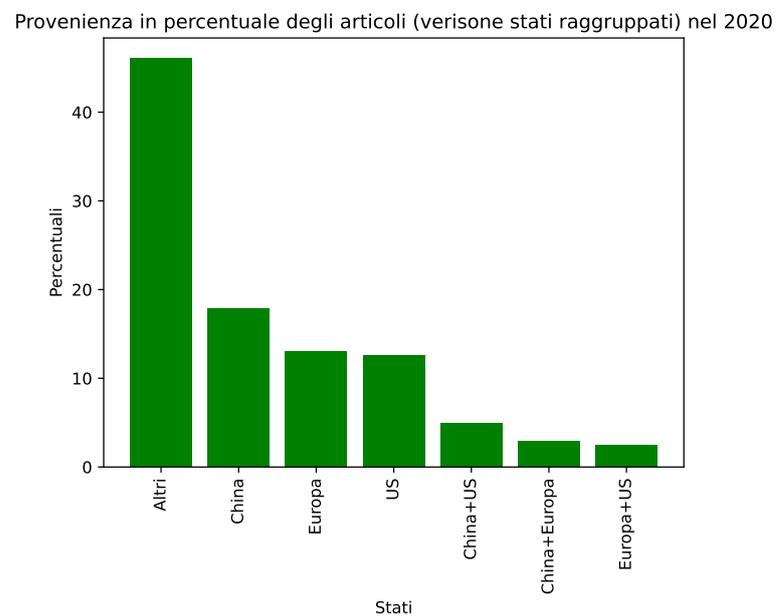


Figura 5.43: Percentuali provenienze articoli a gruppi, 2020

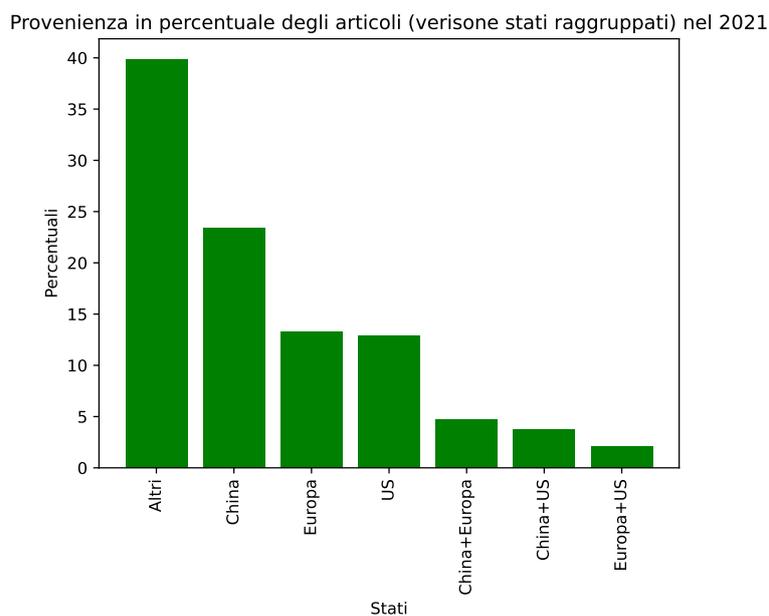


Figura 5.44: Percentuali provenienze articoli a gruppi, 2021

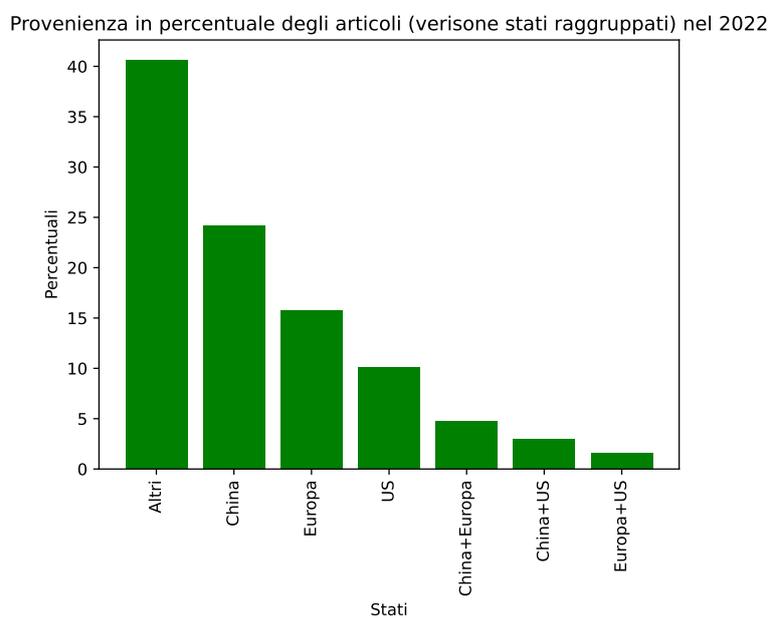


Figura 5.45: Percentuali provenienze articoli a gruppi, 2022

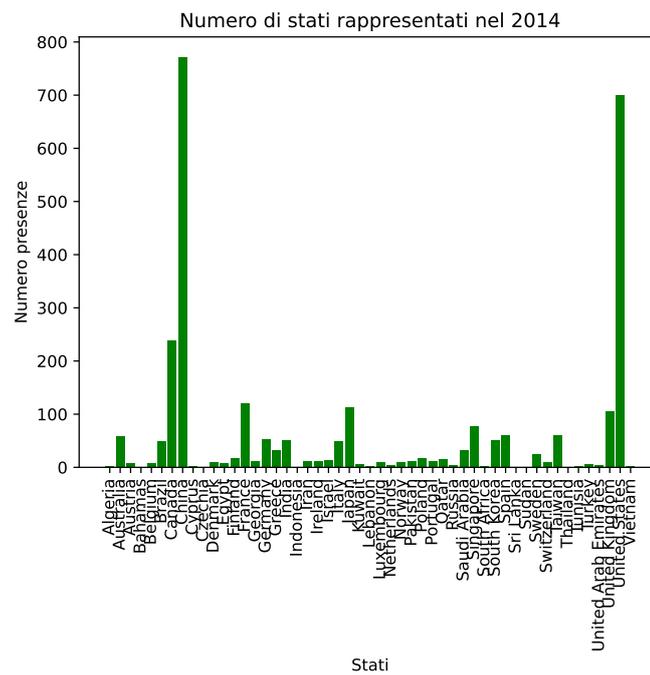


Figura 5.46: Provenienze autori, 2014

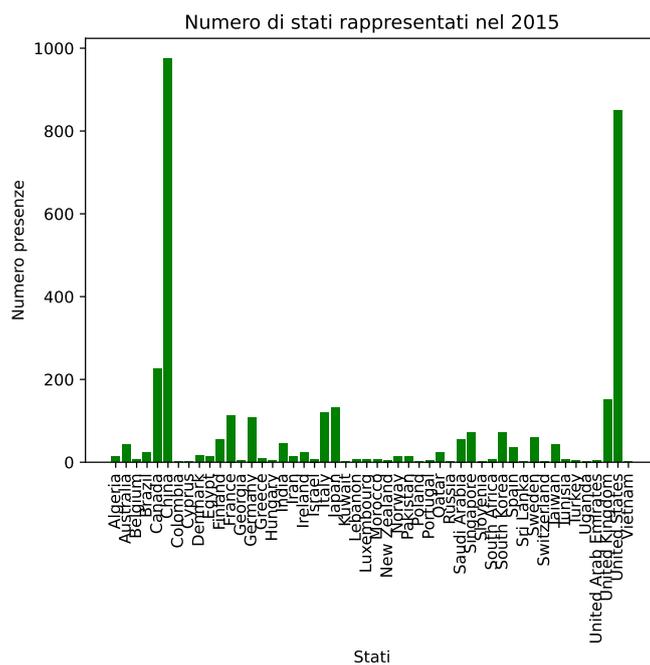


Figura 5.47: Provenienze autori, 2015

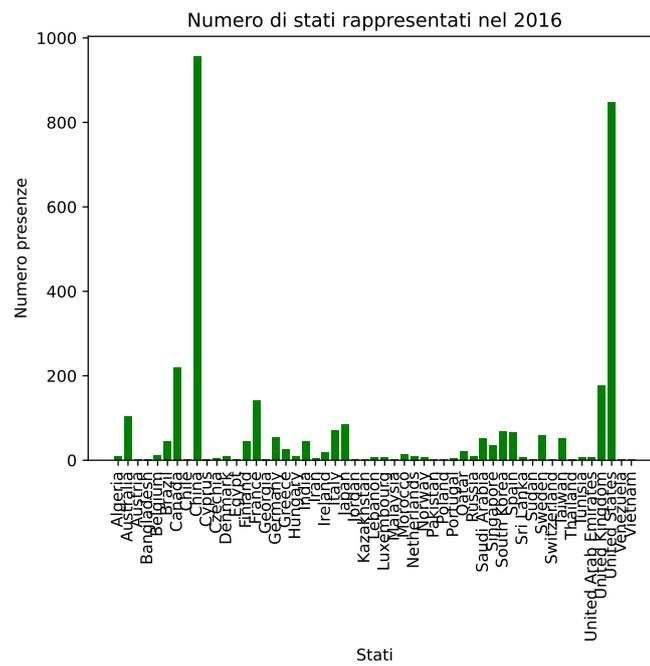


Figura 5.48: Provenienze autori, 2016

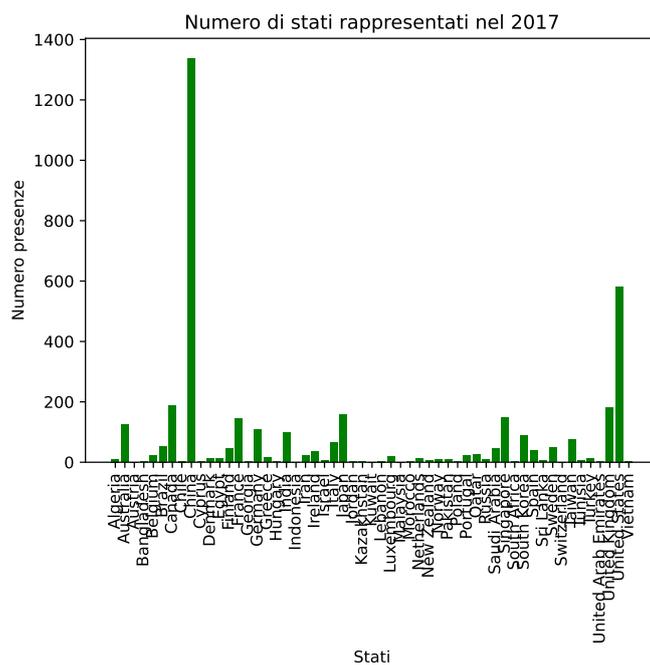


Figura 5.49: Provenienze autori, 2017



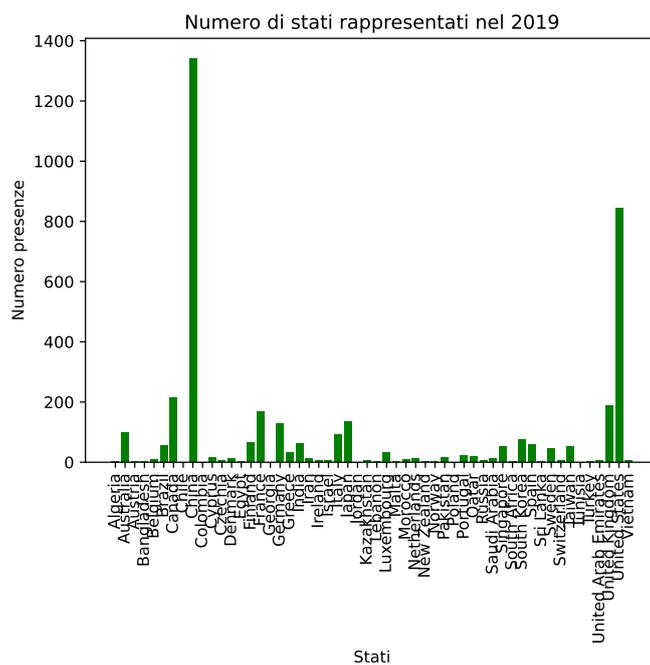


Figura 5.51: Provenienze autori, 2019

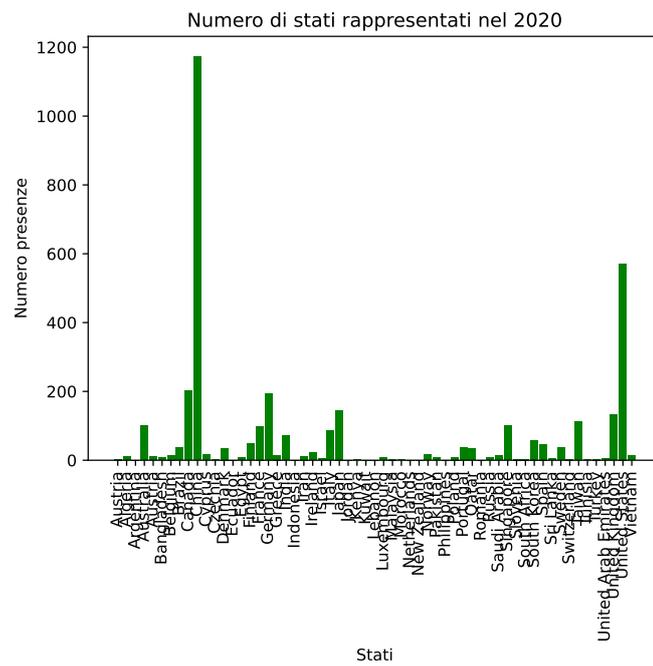


Figura 5.52: Provenienze autori, 2020

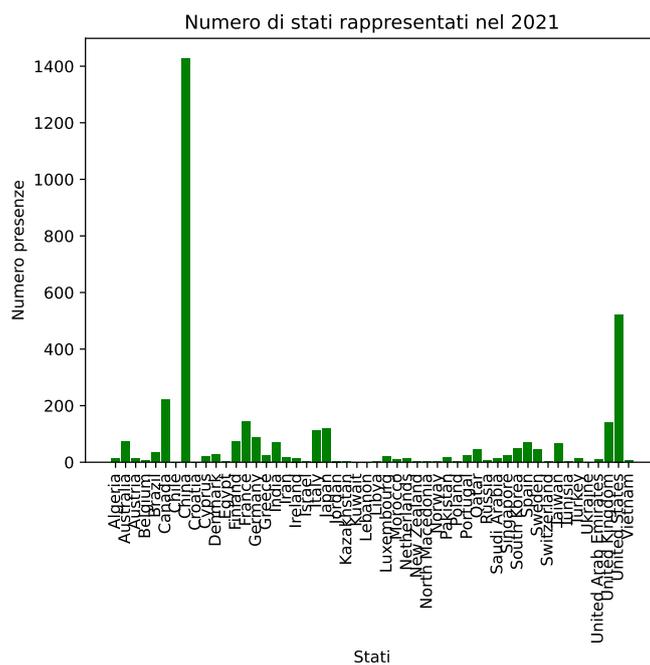


Figura 5.53: Provenienze autori, 2021









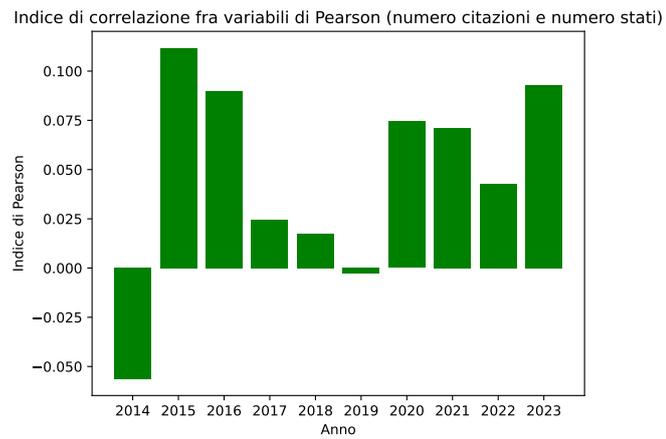


Figura 5.65: Correlazione di Pearson numero citazioni e numero di stati

## 5.2 Infocom

### Citazioni stati

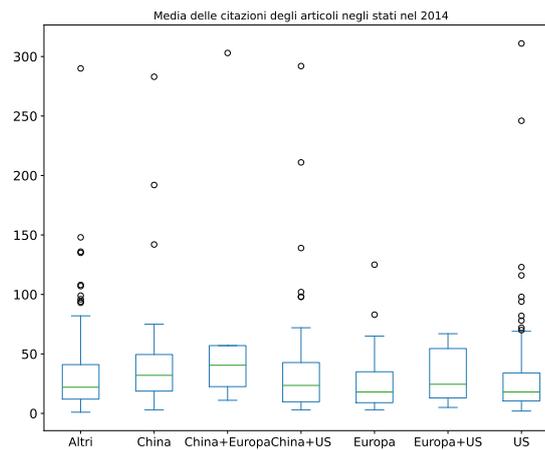


Figura 5.66: Numero di citazioni per gruppo, 2014

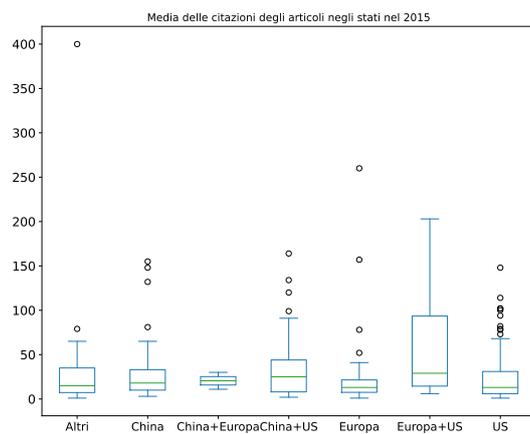


Figura 5.67: Numero di citazioni per gruppo, 2015

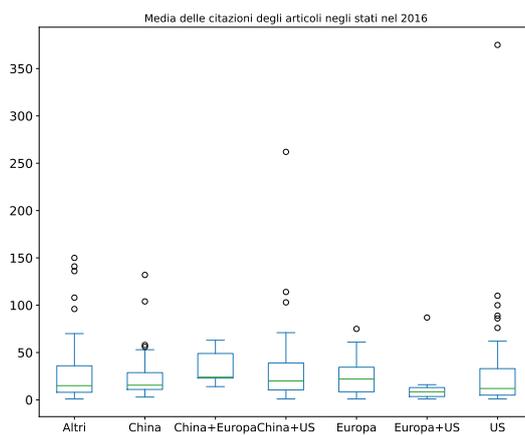


Figura 5.68: Numero di citazioni per gruppo, 2016

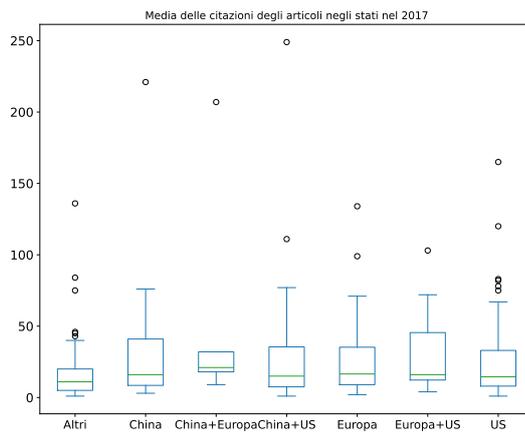


Figura 5.69: Numero di citazioni per gruppo, 2017

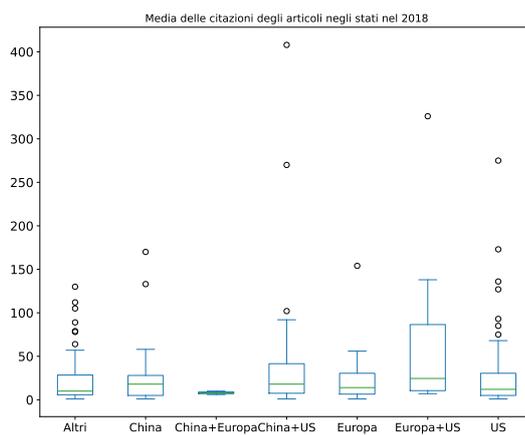


Figura 5.70: Numero di citazioni per gruppo, 2018

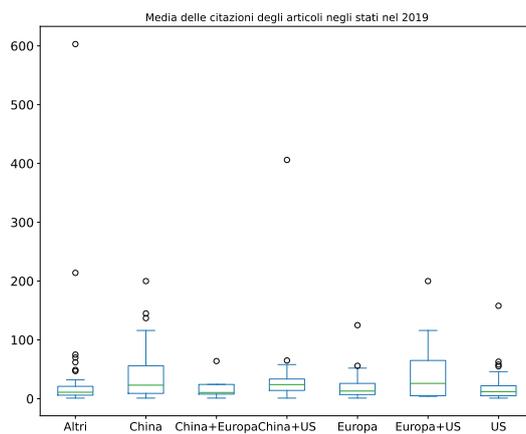


Figura 5.71: Numero di citazioni per gruppo, 2019

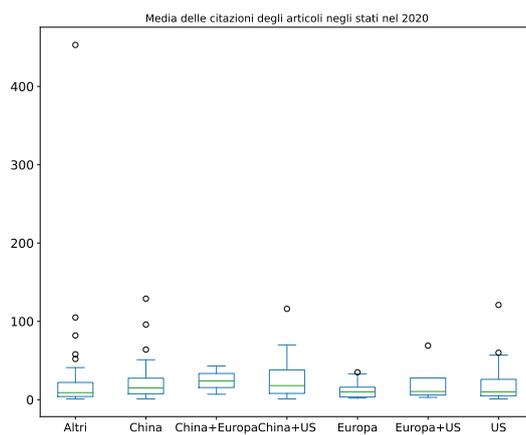


Figura 5.72: Numero di citazioni per gruppo, 2020

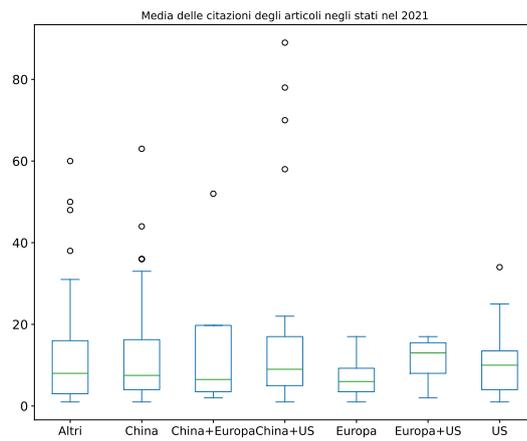


Figura 5.73: Numero di citazioni per gruppo, 2021

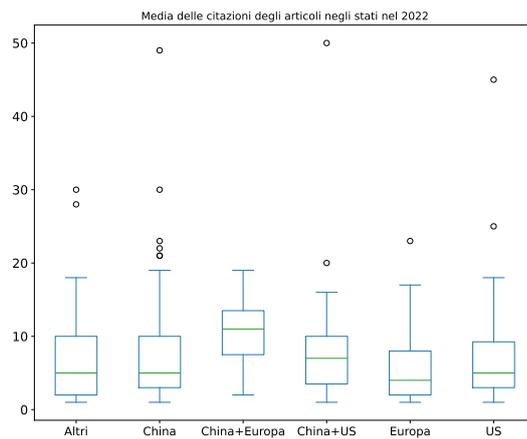


Figura 5.74: Numero di citazioni per gruppo, 2022

## Distribuzioni

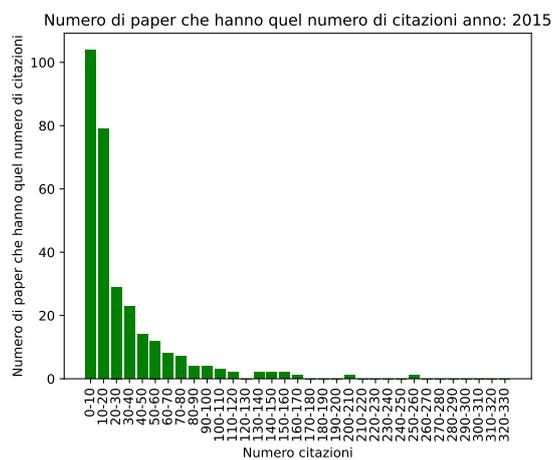


Figura 5.75: Distribuzione del numero di citazioni, 2015

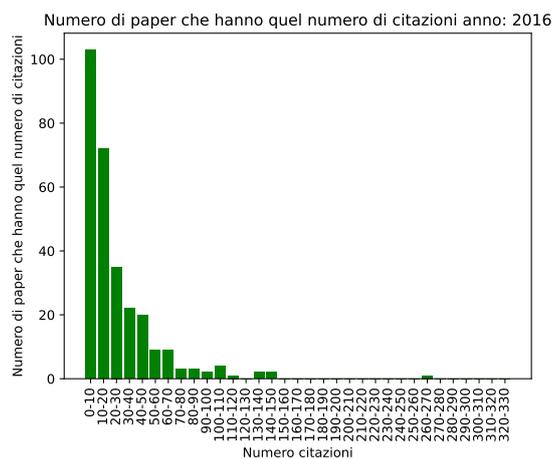


Figura 5.76: Distribuzione del numero di citazioni, 2016

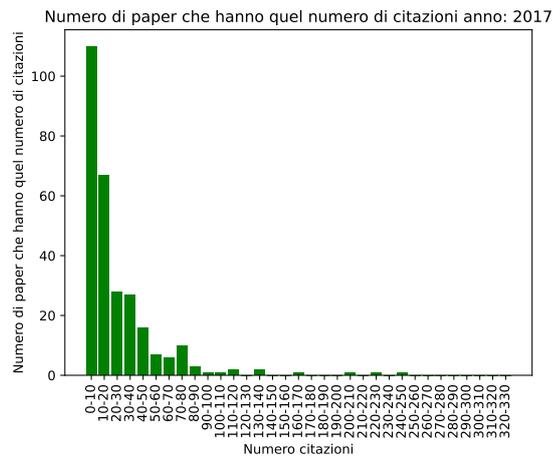


Figura 5.77: Distribuzione del numero di citazioni, 2017

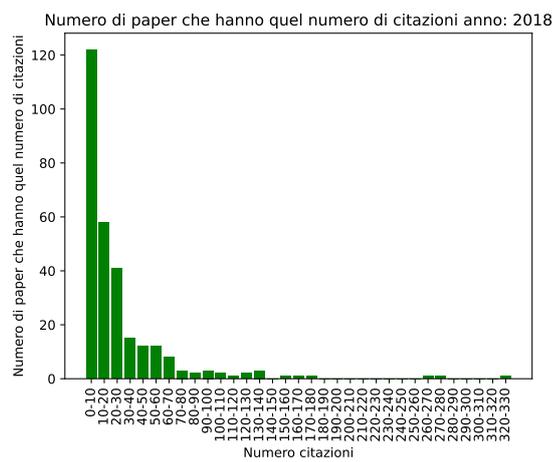


Figura 5.78: Distribuzione del numero di citazioni, 2018

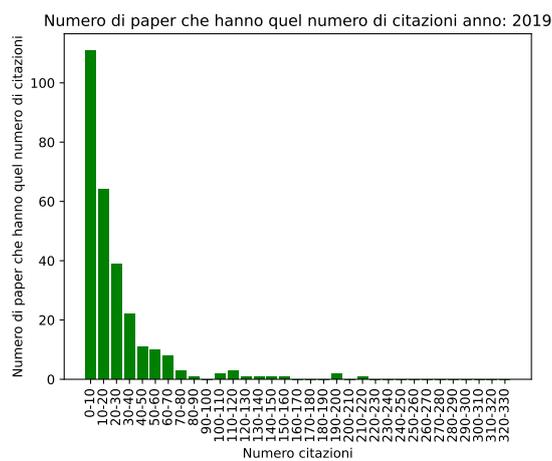


Figura 5.79: Distribuzione del numero di citazioni, 2019

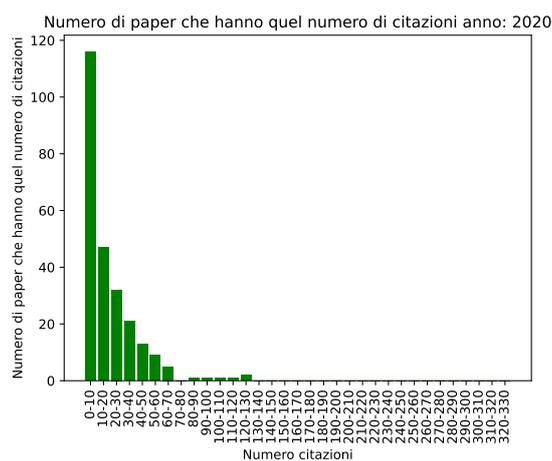


Figura 5.80: Distribuzione del numero di citazioni, 2020

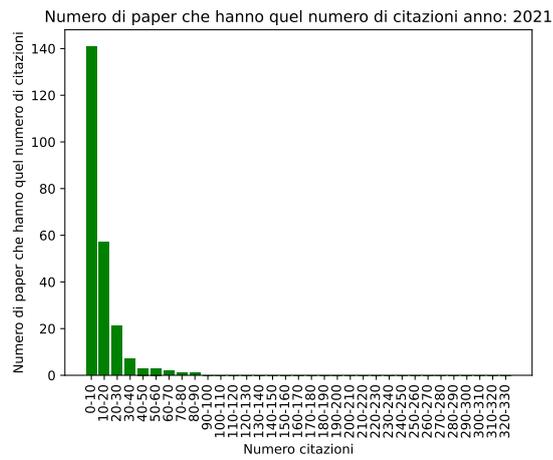


Figura 5.81: Distribuzione del numero di citazioni, 2021

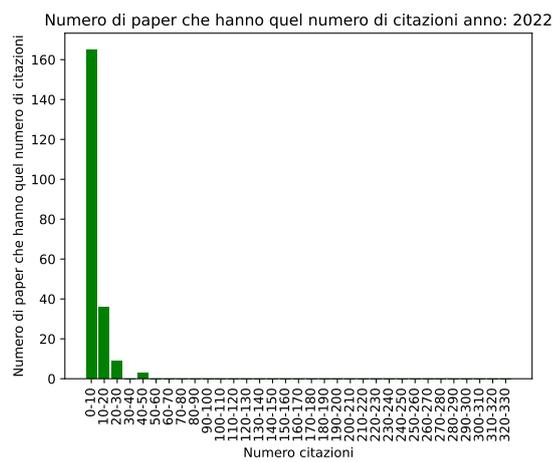


Figura 5.82: Distribuzione del numero di citazioni, 2022

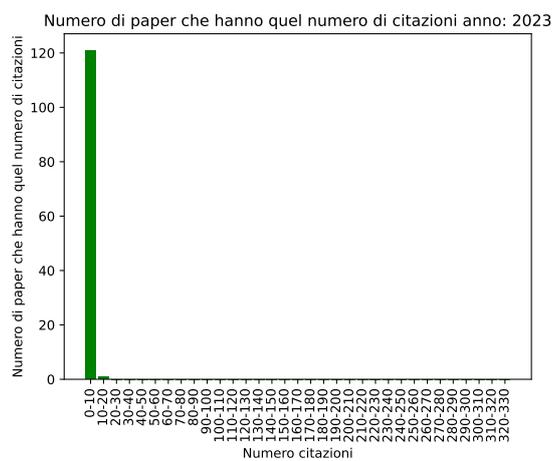


Figura 5.83: Distribuzione del numero di citazioni, 2023

## Provenienze

Numero di stati rappresentati nel 2014

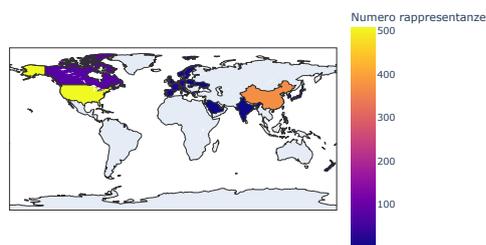


Figura 5.84: Mappa provenienze articoli, 2014

Numero di stati rappresentati nel 2015

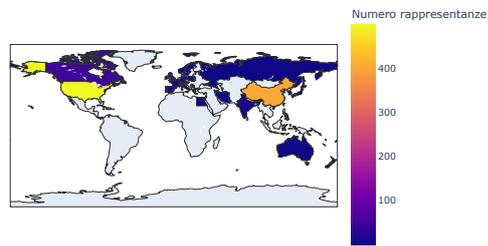


Figura 5.85: Mappa provenienze articoli, 2015

Numero di stati rappresentati nel 2016

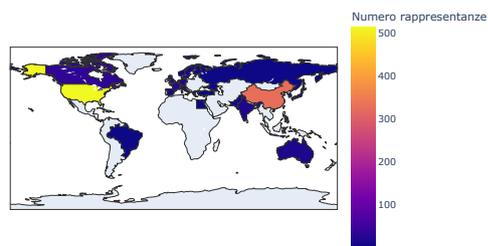


Figura 5.86: Mappa provenienze articoli, 2016

Numero di stati rappresentati nel 2017

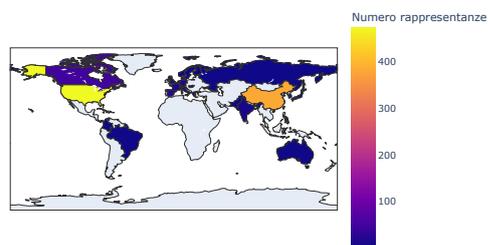


Figura 5.87: Mappa provenienze articoli, 2017

Numero di stati rappresentati nel 2018

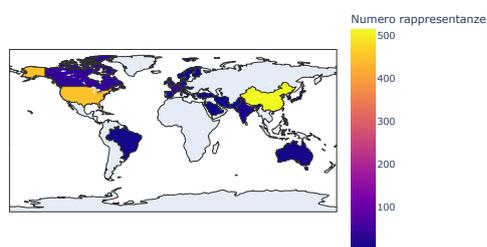


Figura 5.88: Mappa provenienze articoli, 2018

Numero di stati rappresentati nel 2019

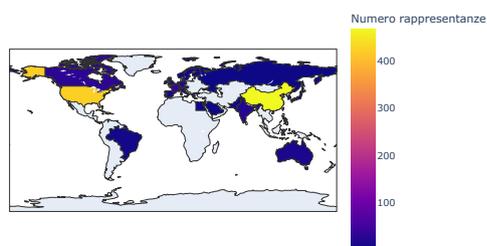


Figura 5.89: Mappa provenienze articoli, 2019

Numero di stati rappresentati nel 2020

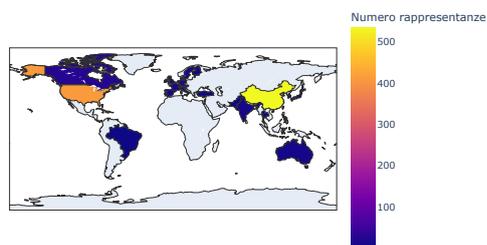


Figura 5.90: Mappa provenienze articoli, 2020

Numero di stati rappresentati nel 2021

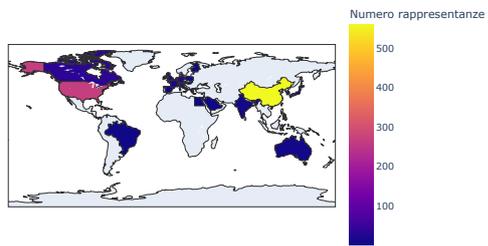


Figura 5.91: Mappa provenienze articoli, 2021

Numero di stati rappresentati nel 2022

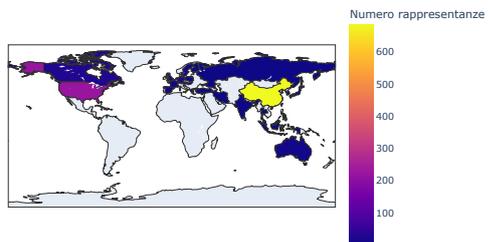


Figura 5.92: Mappa provenienze articoli, 2022

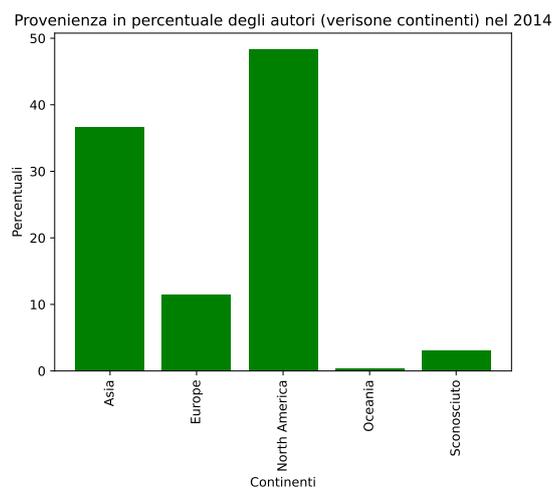


Figura 5.93: Percentuali provenienze articoli, continenti, 2014

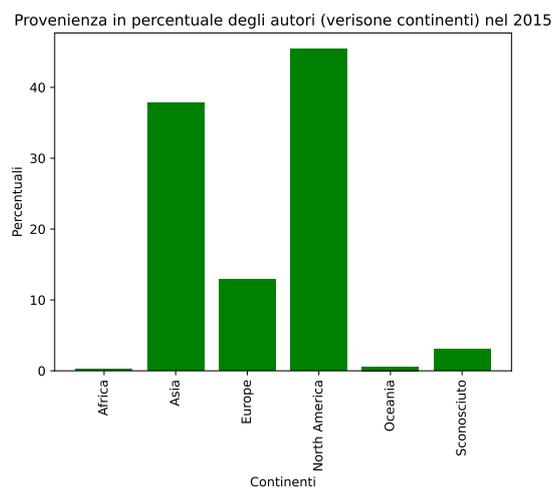


Figura 5.94: Percentuali provenienze articoli, continenti, 2015

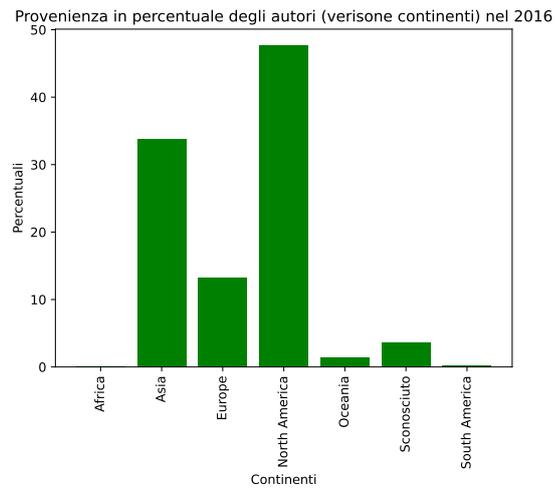


Figura 5.95: Percentuali provenienze articoli, continenti, 2016

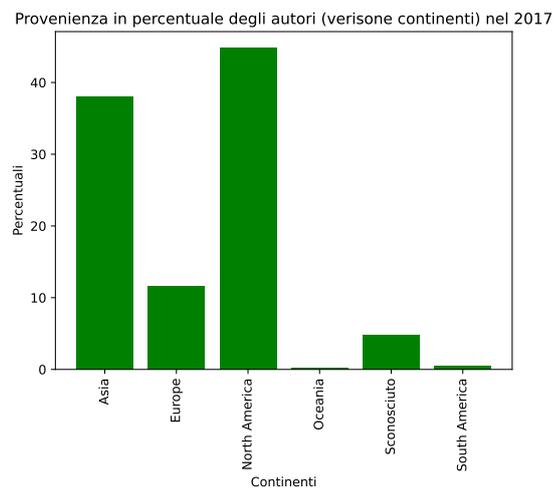


Figura 5.96: Percentuali provenienze articoli, continenti, 2017

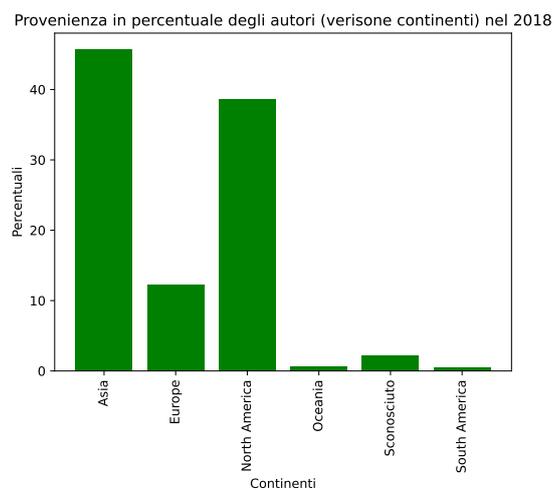


Figura 5.97: Percentuali provenienze articoli, continenti, 2018

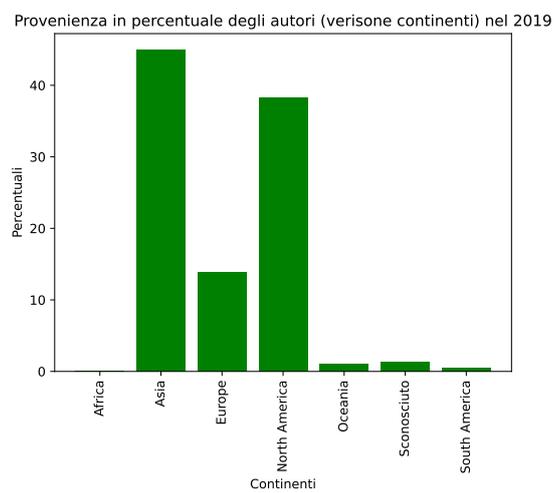


Figura 5.98: Percentuali provenienze articoli, continenti, 2019

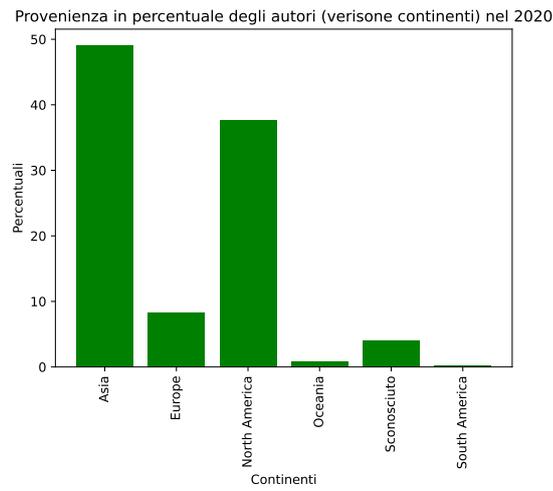


Figura 5.99: Percentuali provenienze articoli, continenti, 2020

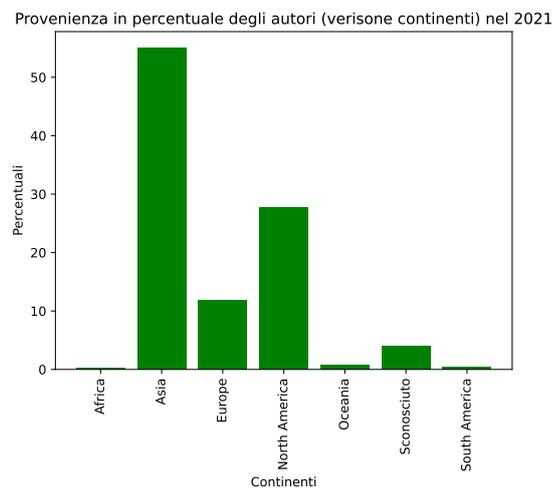


Figura 5.100: Percentuali provenienze articoli, continenti, 2021

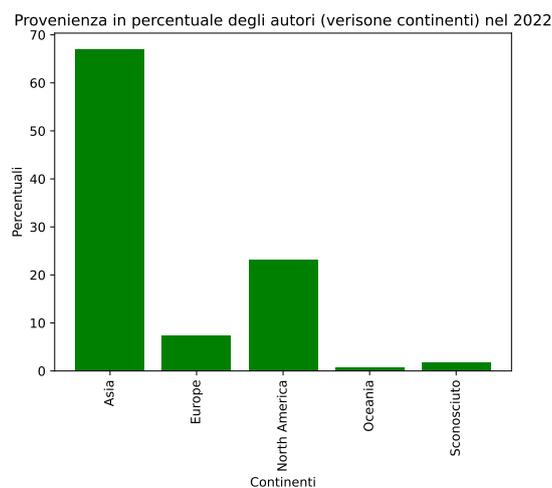


Figura 5.101: Percentuali provenienze articoli, continenti, 2022

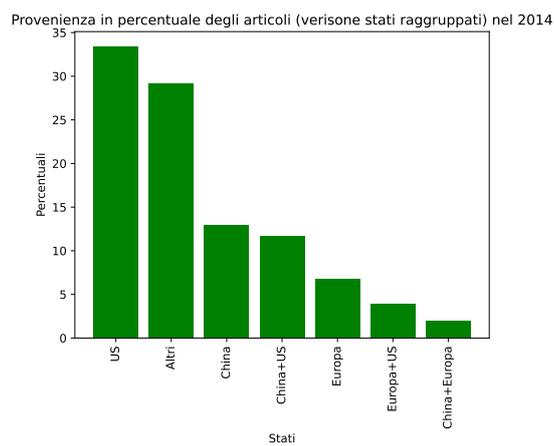


Figura 5.102: Percentuali provenienze articoli a gruppi, 2014

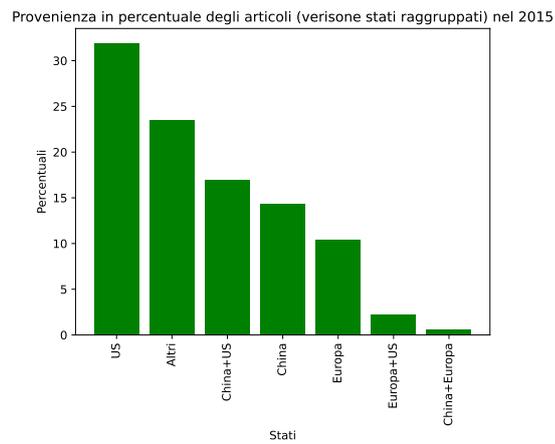


Figura 5.103: Percentuali provenienze articoli a gruppi, 2015

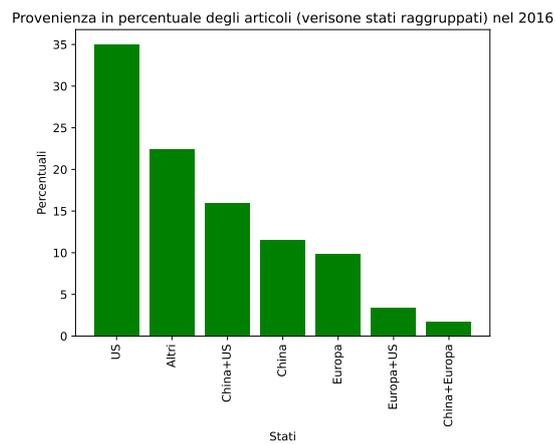


Figura 5.104: Percentuali provenienze articoli a gruppi, 2016

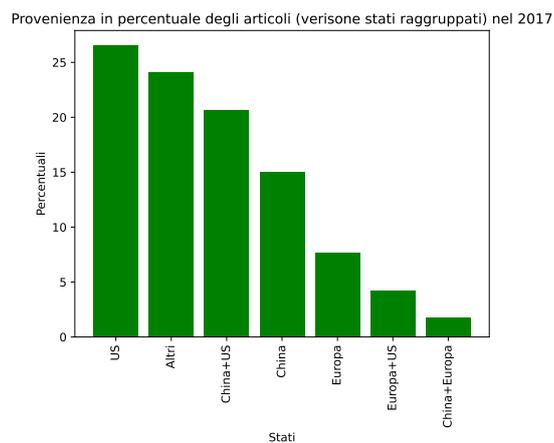


Figura 5.105: Percentuali provenienze articoli a gruppi, 2017

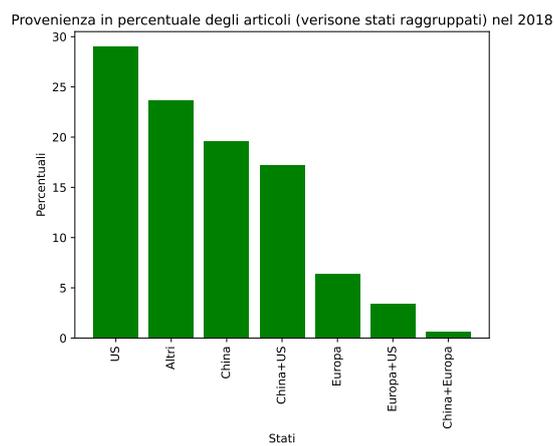


Figura 5.106: Percentuali provenienze articoli a gruppi, 2018

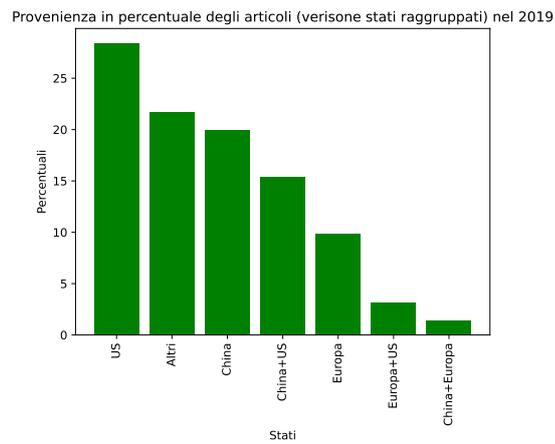


Figura 5.107: Percentuali provenienze articoli a gruppi, 2019

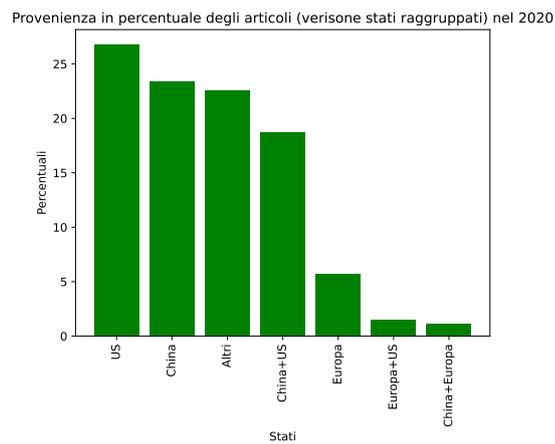


Figura 5.108: Percentuali provenienze articoli a gruppi, 2020

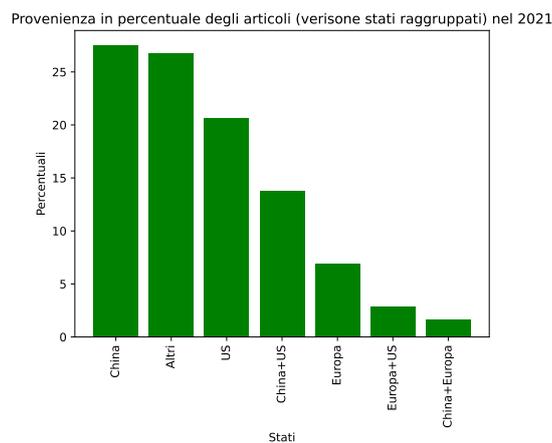


Figura 5.109: Percentuali provenienze articoli a gruppi, 2021

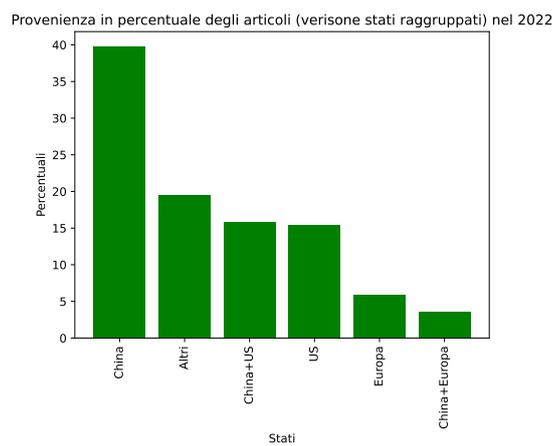


Figura 5.110: Percentuali provenienze articoli a gruppi, 2022

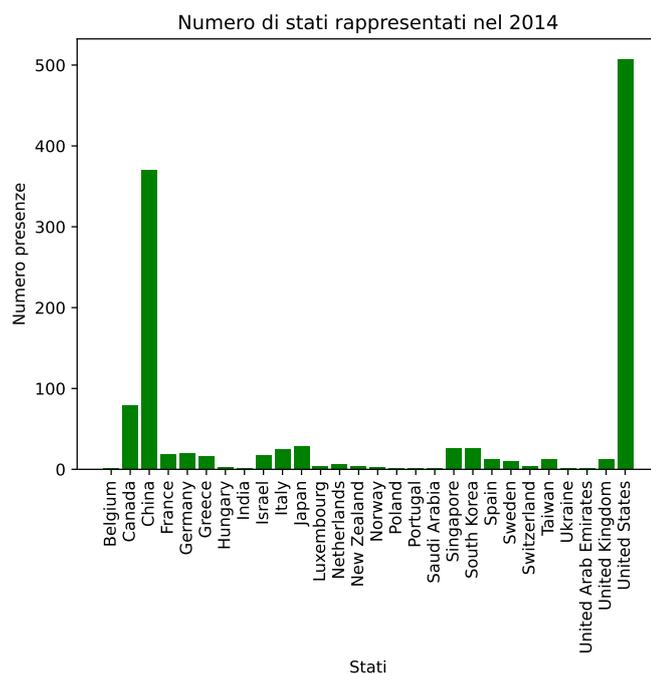


Figura 5.111: Provenienze autori, 2014

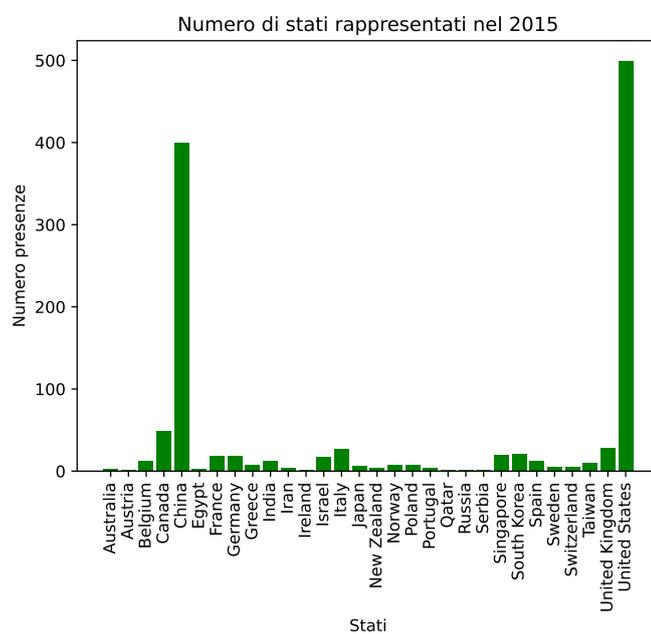


Figura 5.112: Provenienze autori, 2015

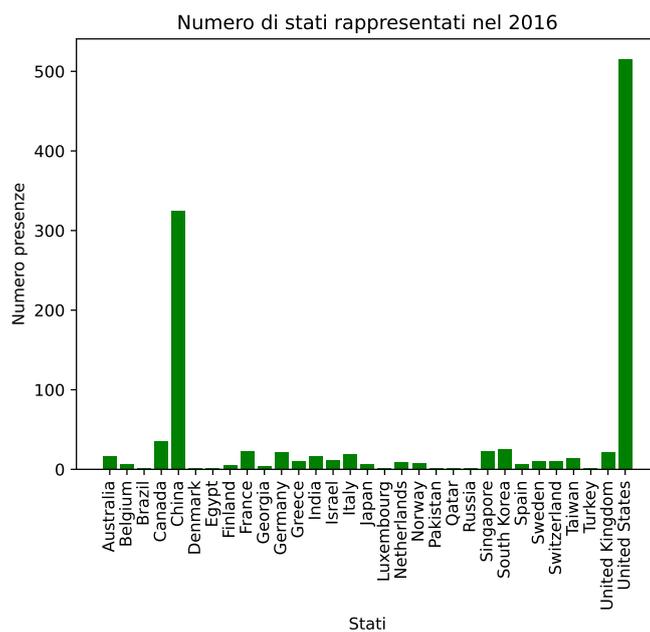


Figura 5.113: Provenienze autori, 2016

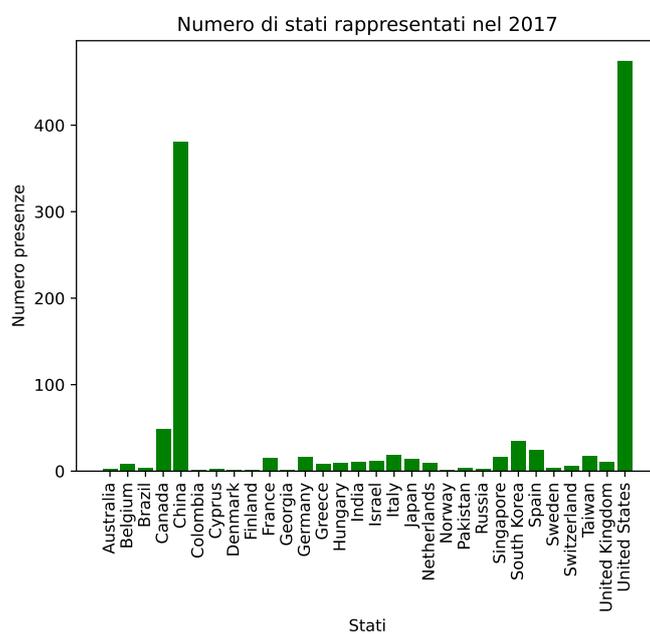


Figura 5.114: Provenienze autori, 2017

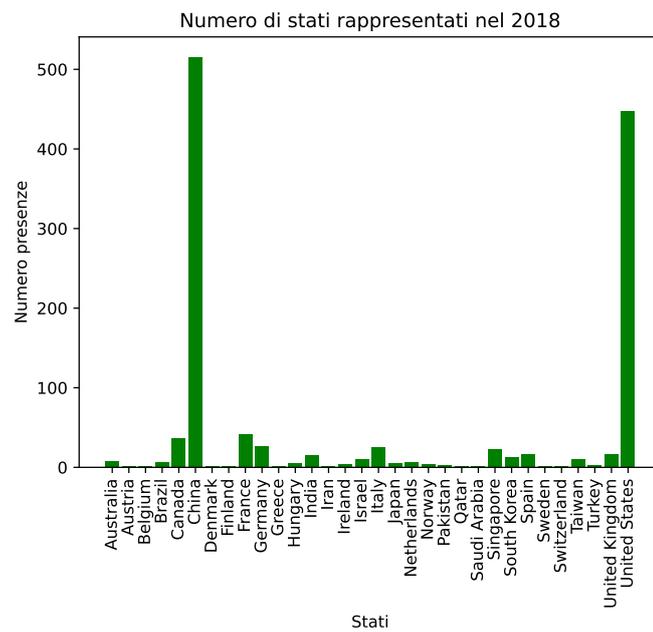


Figura 5.115: Provenienze autori, 2018

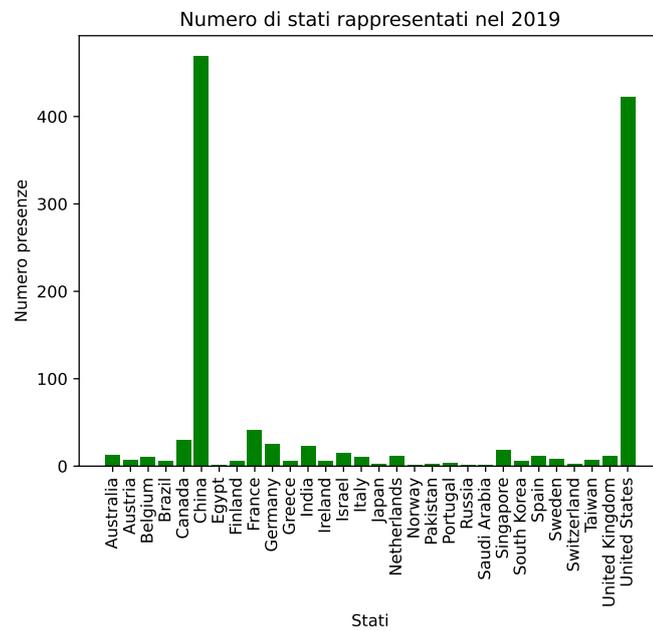


Figura 5.116: Provenienze autori, 2019

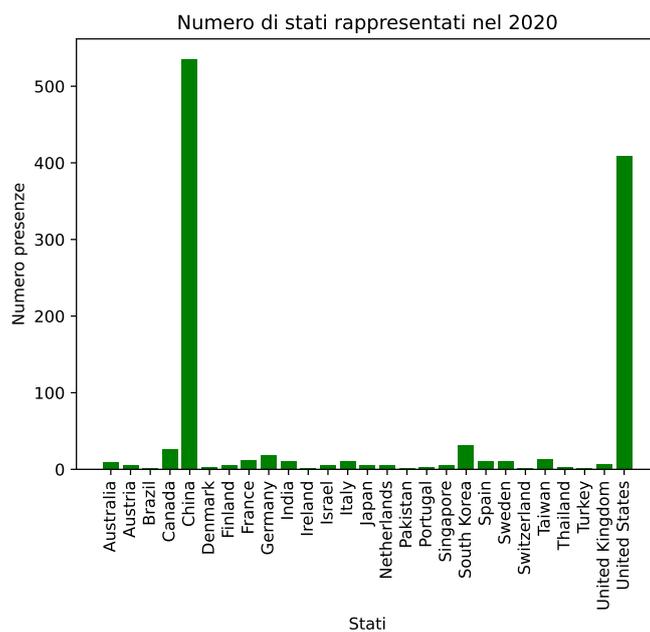


Figura 5.117: Provenienze autori, 2020

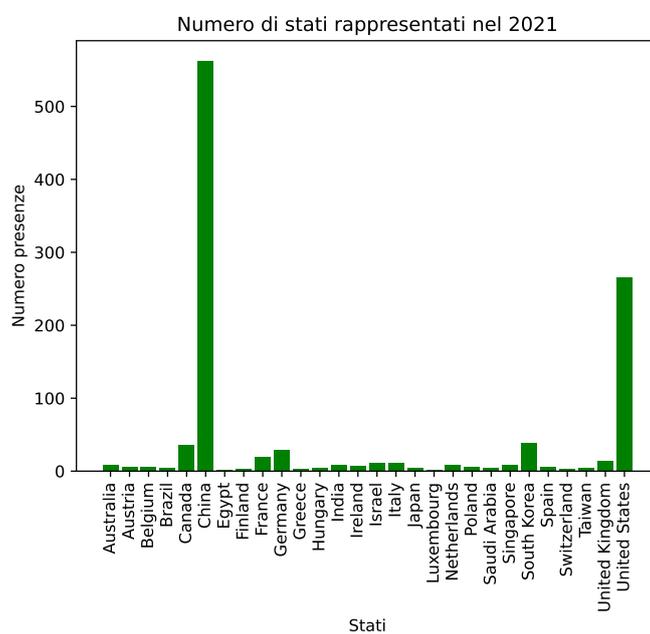


Figura 5.118: Provenienze autori, 2021

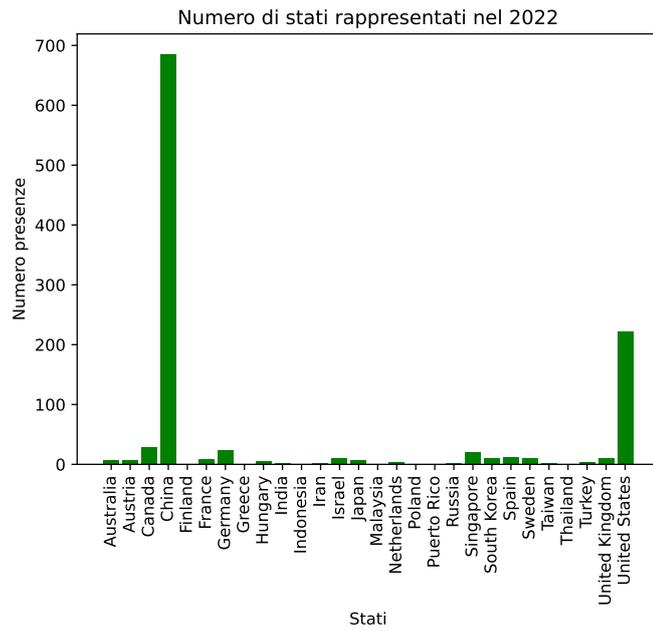


Figura 5.119: Provenienze autori, 2022

Wordcloud

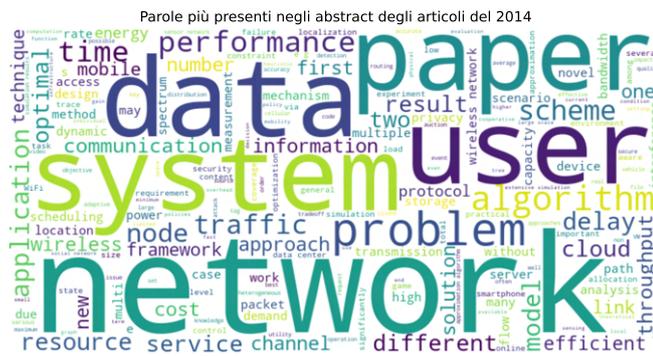


Figura 5.120: Wordcloud abstract degli articoli, 2014







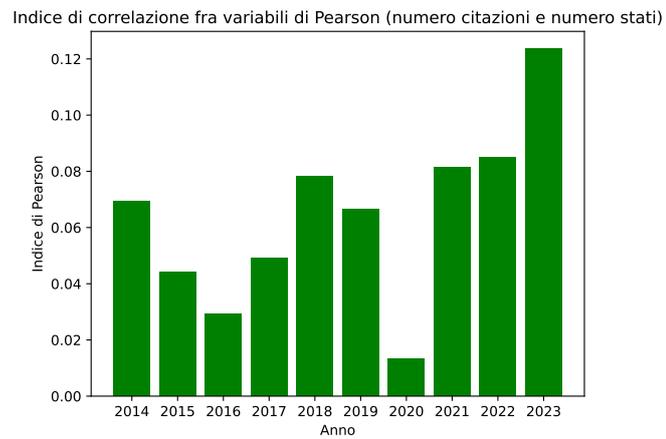


Figura 5.130: Correlazione di Pearson numero citazioni e numero di stati

## 5.3 Percom

### Citazioni stati

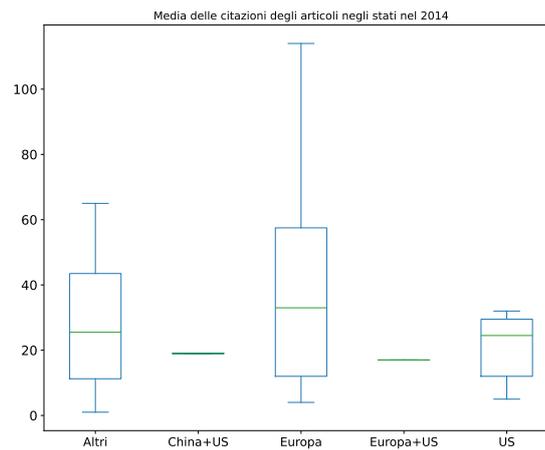


Figura 5.131: Numero di citazioni per gruppo, 2014

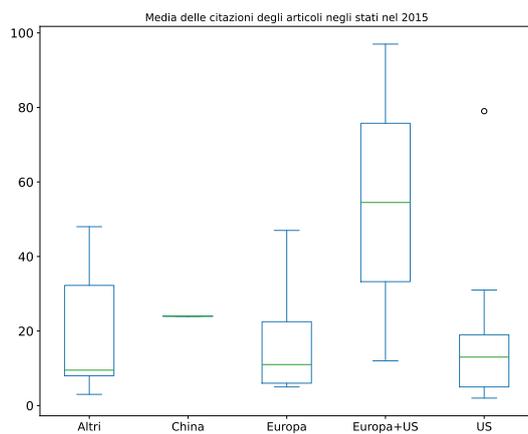


Figura 5.132: Numero di citazioni per gruppo, 2015

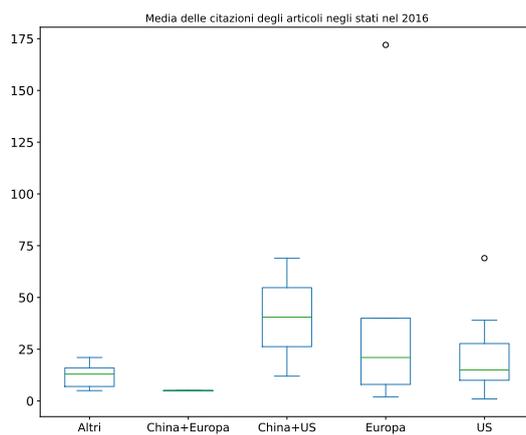


Figura 5.133: Numero di citazioni per gruppo, 2016

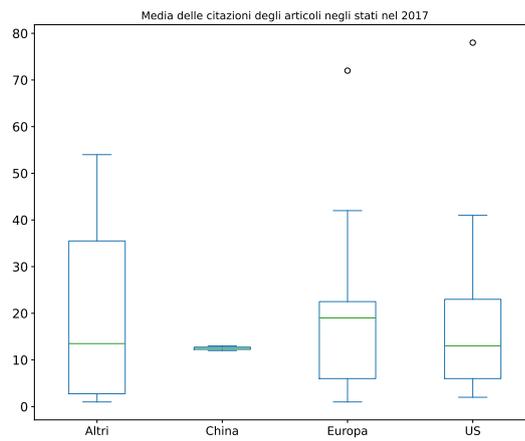


Figura 5.134: Numero di citazioni per gruppo, 2017

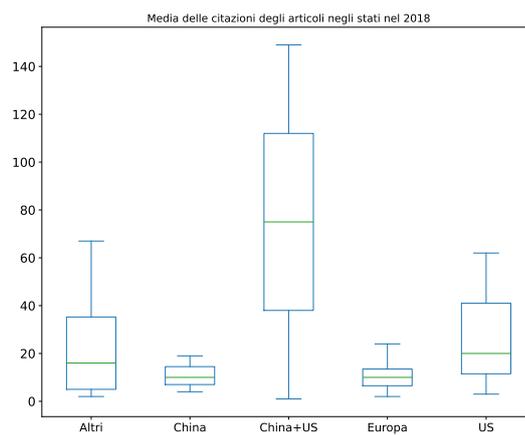


Figura 5.135: Numero di citazioni per gruppo, 2018

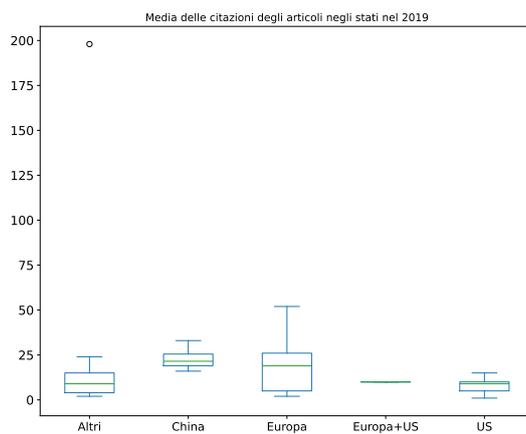


Figura 5.136: Numero di citazioni per gruppo, 2019

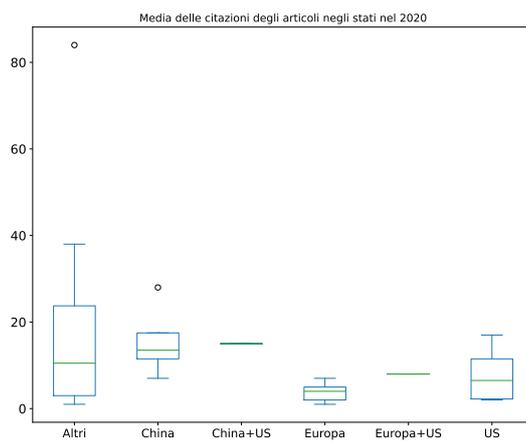


Figura 5.137: Numero di citazioni per gruppo, 2020

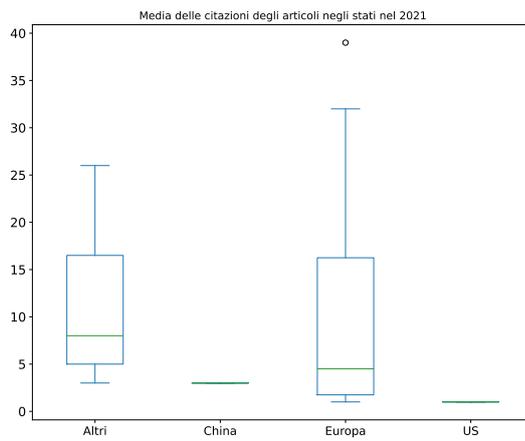


Figura 5.138: Numero di citazioni per gruppo, 2021

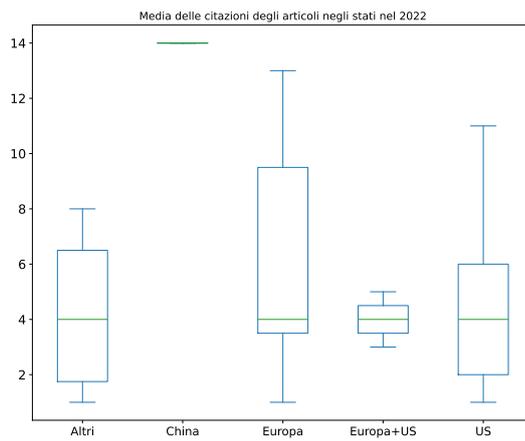


Figura 5.139: Numero di citazioni per gruppo, 2022

## Distribuzioni

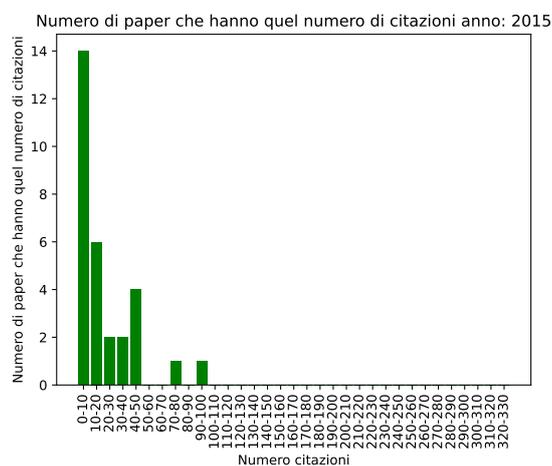


Figura 5.140: Distribuzione del numero di citazioni, 2015

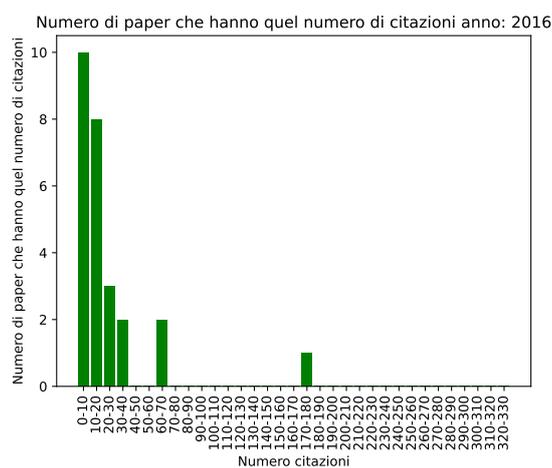


Figura 5.141: Distribuzione del numero di citazioni, 2016

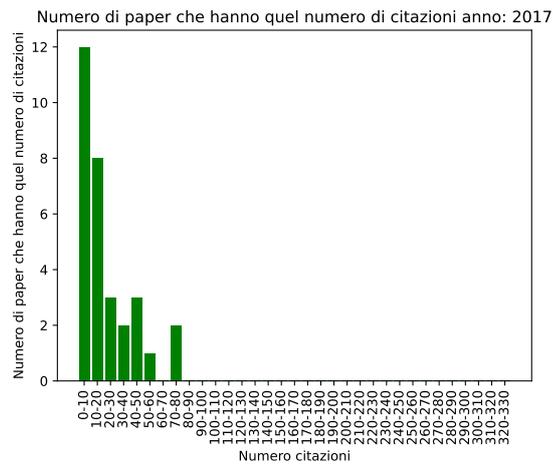


Figura 5.142: Distribuzione del numero di citazioni, 2017

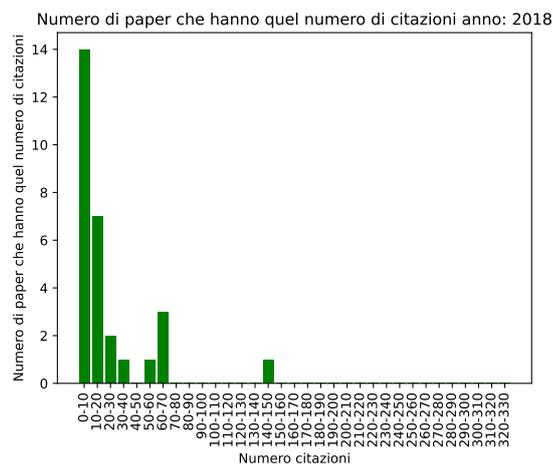


Figura 5.143: Distribuzione del numero di citazioni, 2018

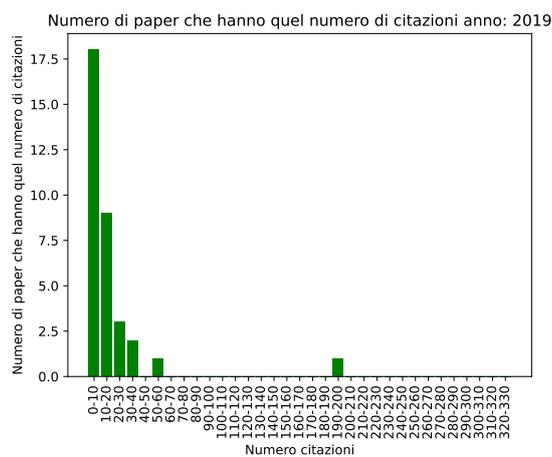


Figura 5.144: Distribuzione del numero di citazioni, 2019

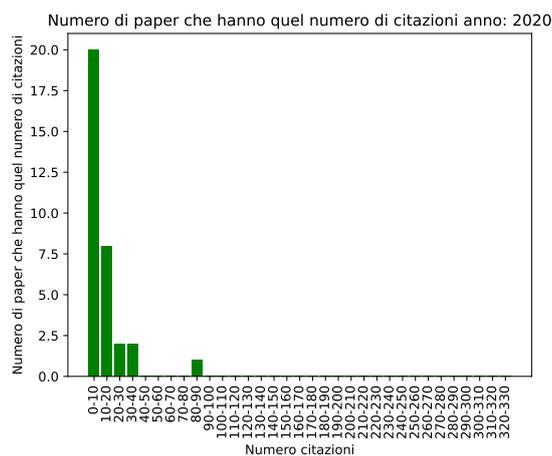


Figura 5.145: Distribuzione del numero di citazioni, 2020

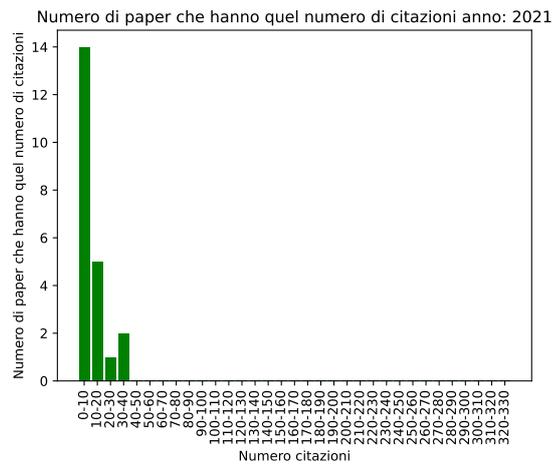


Figura 5.146: Distribuzione del numero di citazioni, 2021

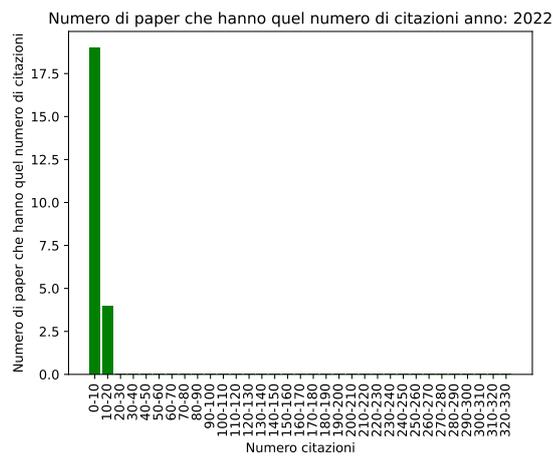


Figura 5.147: Distribuzione del numero di citazioni, 2022

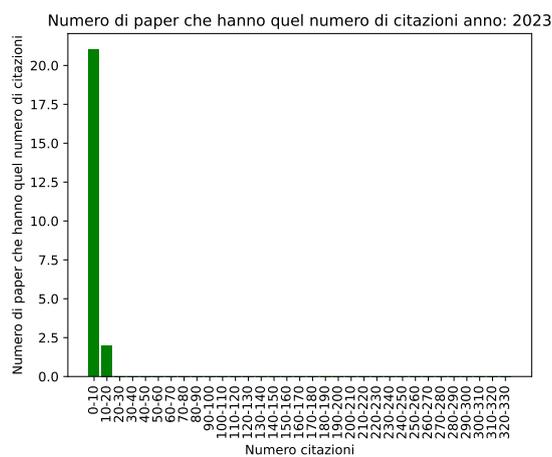


Figura 5.148: Distribuzione del numero di citazioni, 2023

## Provenienze

Numero di stati rappresentati nel 2014

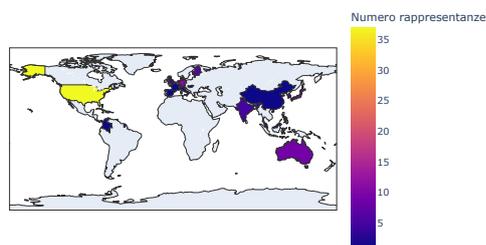


Figura 5.149: Mappa provenienze articoli, 2014

Numero di stati rappresentati nel 2015

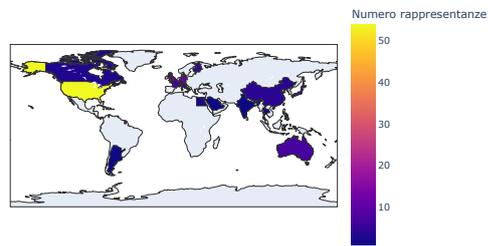


Figura 5.150: Mappa provenienze articoli, 2015

Numero di stati rappresentati nel 2016

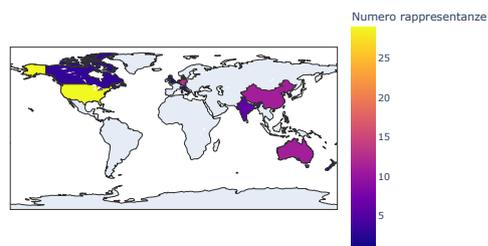


Figura 5.151: Mappa provenienze articoli, 2016

Numero di stati rappresentati nel 2017

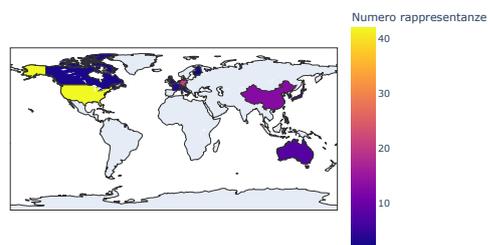


Figura 5.152: Mappa provenienze articoli, 2017

Numero di stati rappresentati nel 2018

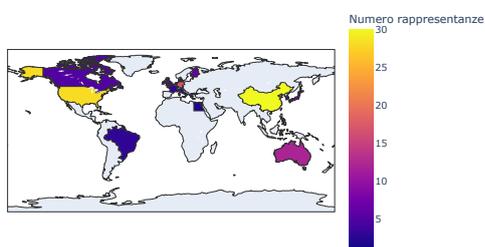


Figura 5.153: Mappa provenienze articoli, 2018

Numero di stati rappresentati nel 2019

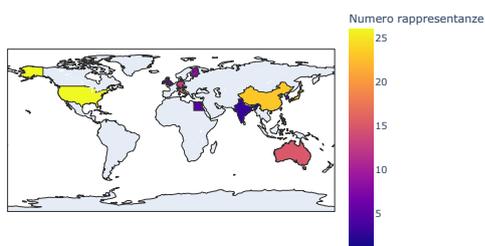


Figura 5.154: Mappa provenienze articoli, 2019

Numero di stati rappresentati nel 2020

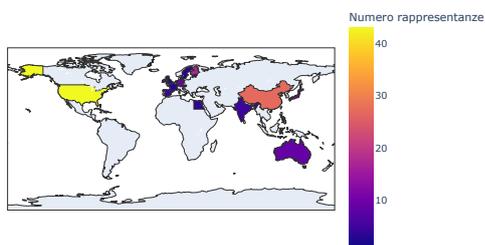


Figura 5.155: Mappa provenienze articoli, 2020

Numero di stati rappresentati nel 2021

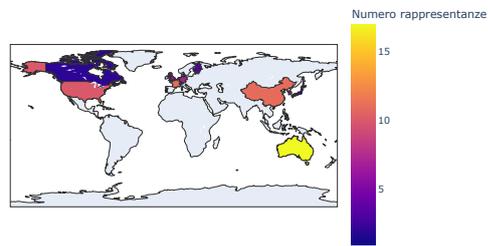


Figura 5.156: Mappa provenienze articoli, 2021

Numero di stati rappresentati nel 2022

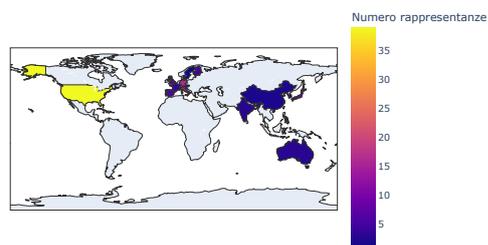


Figura 5.157: Mappa provenienze articoli, 2022

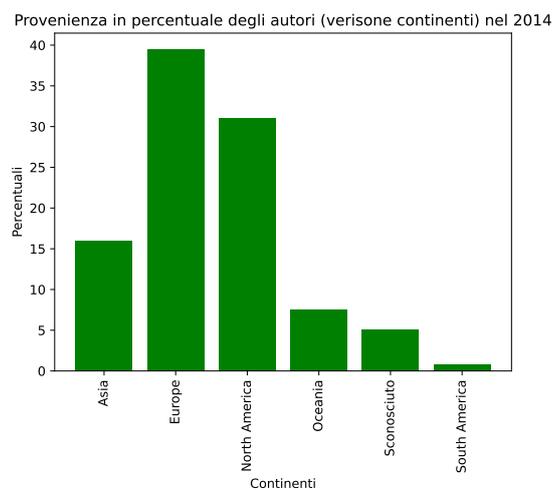


Figura 5.158: Percentuali provenienze articoli, continenti, 2014

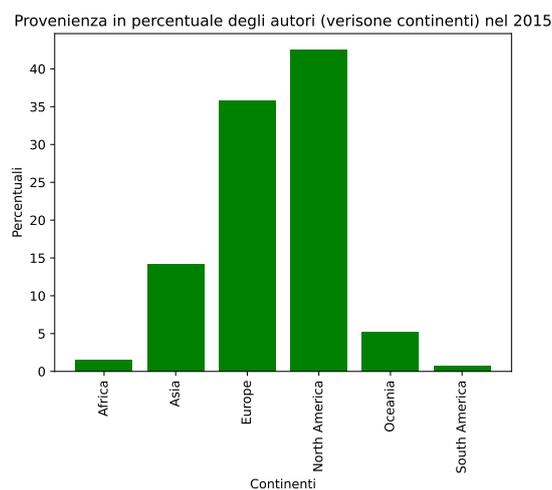


Figura 5.159: Percentuali provenienze articoli, continenti, 2015

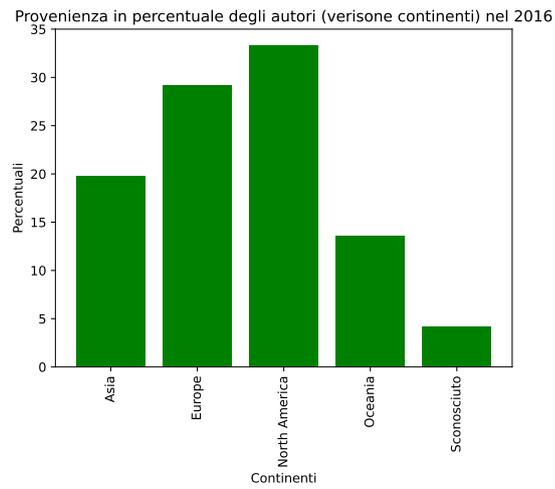


Figura 5.160: Percentuali provenienze articoli, continenti, 2016

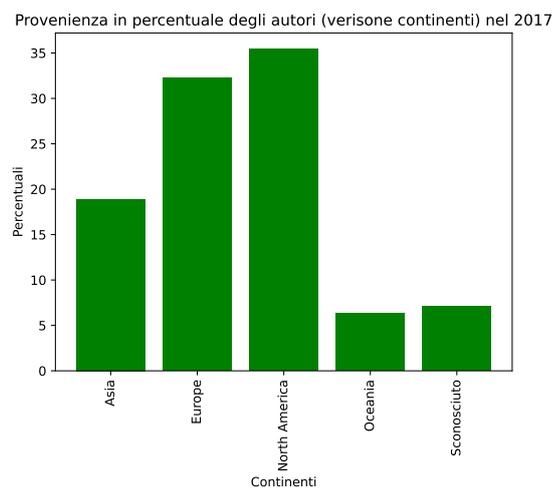


Figura 5.161: Percentuali provenienze articoli, continenti, 2017

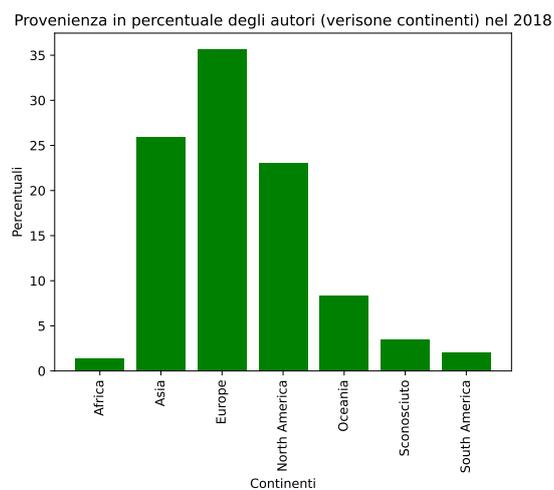


Figura 5.162: Percentuali provenienze articoli, continenti, 2018

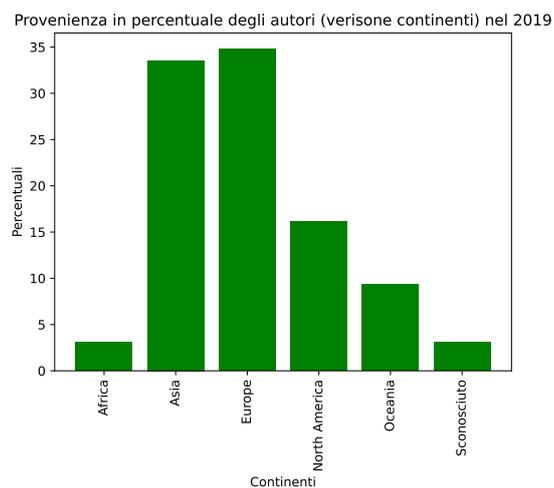


Figura 5.163: Percentuali provenienze articoli, continenti, 2019

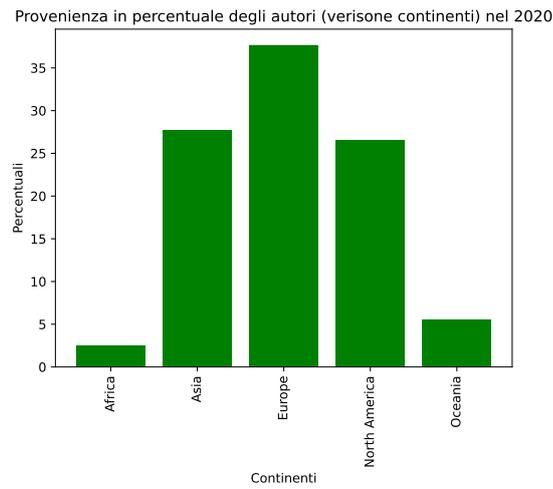


Figura 5.164: Percentuali provenienze articoli, continenti, 2020

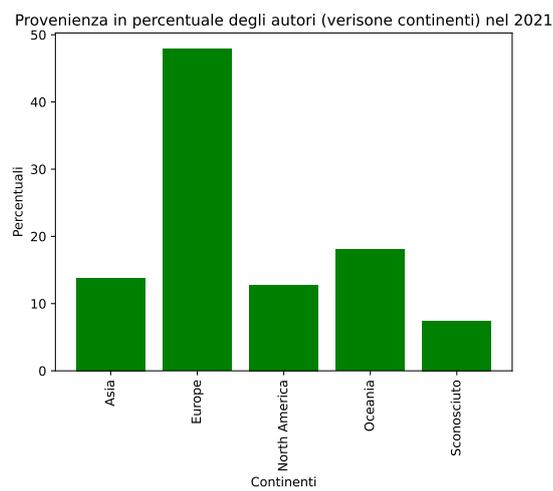


Figura 5.165: Percentuali provenienze articoli, continenti, 2021

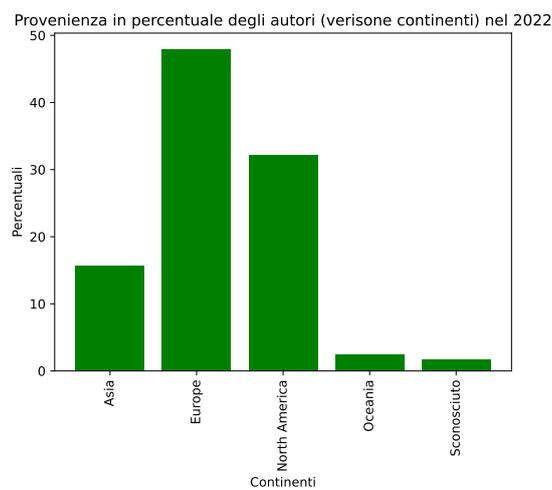


Figura 5.166: Percentuali provenienze articoli, continenti, 2022

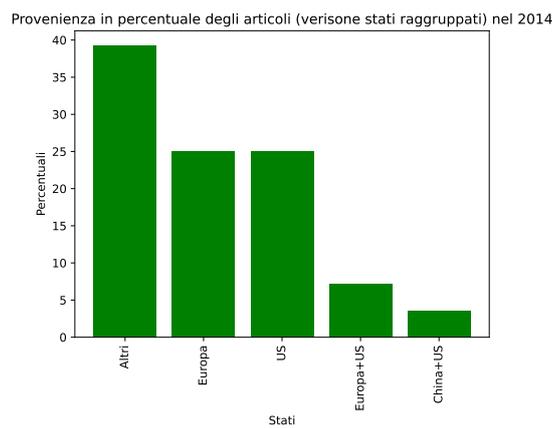


Figura 5.167: Percentuali provenienze articoli a gruppi, 2014

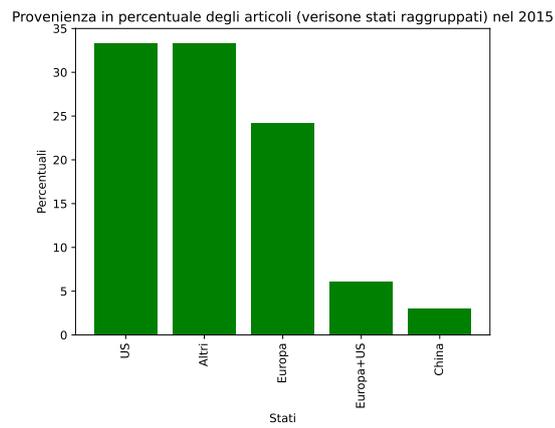


Figura 5.168: Percentuali provenienze articoli a gruppi, 2015

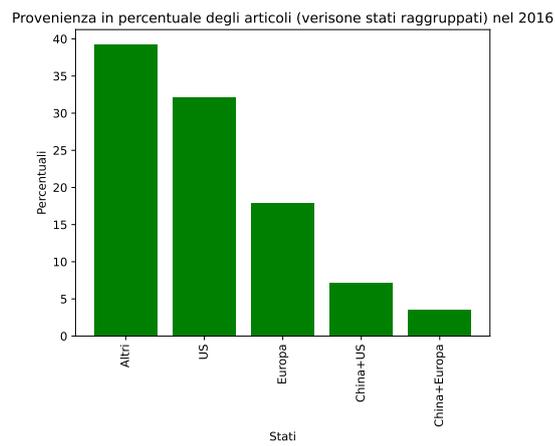


Figura 5.169: Percentuali provenienze articoli a gruppi, 2016

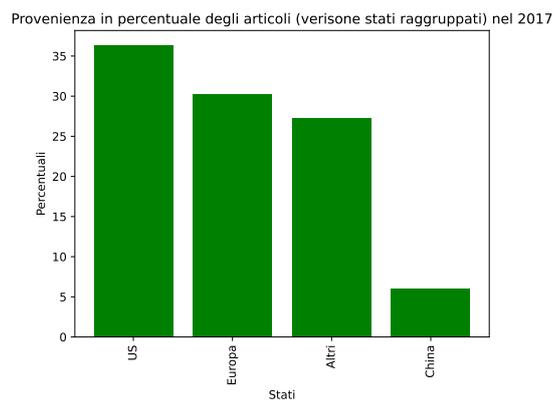


Figura 5.170: Percentuali provenienze articoli a gruppi, 2017

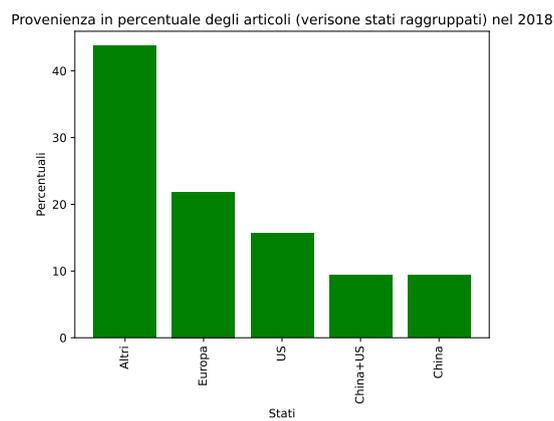


Figura 5.171: Percentuali provenienze articoli a gruppi, 2018

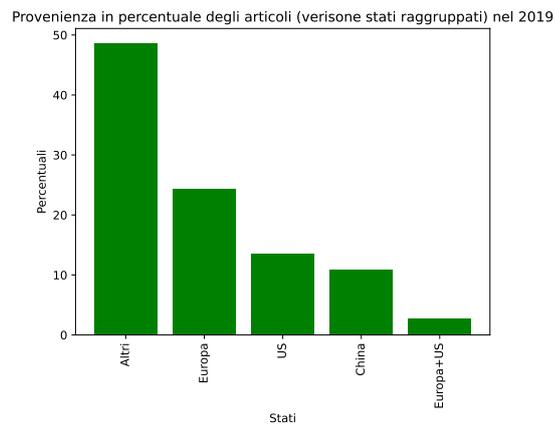


Figura 5.172: Percentuali provenienze articoli a gruppi, 2019

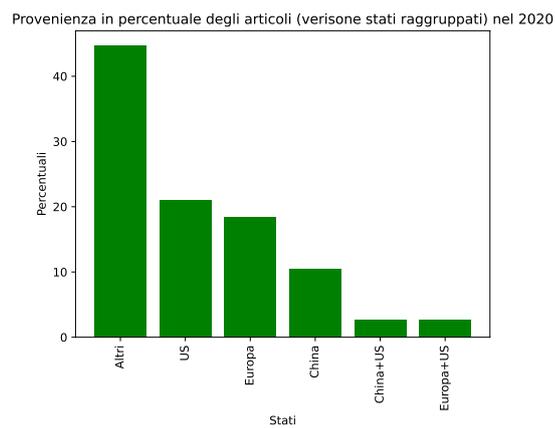


Figura 5.173: Percentuali provenienze articoli a gruppi, 2020

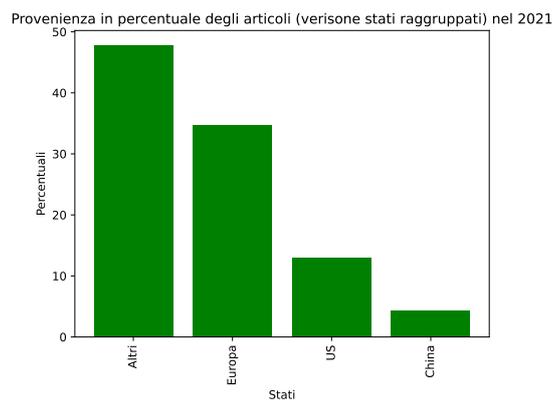


Figura 5.174: Percentuali provenienze articoli a gruppi, 2021

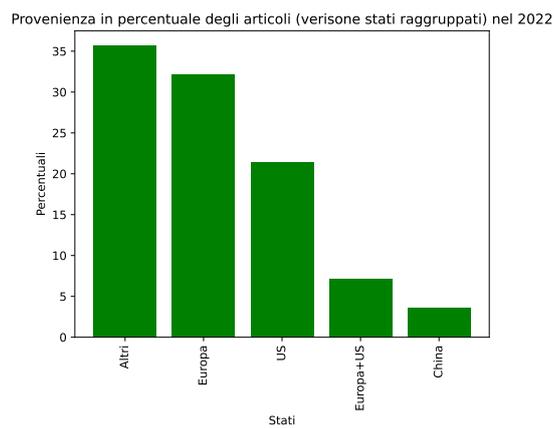


Figura 5.175: Percentuali provenienze articoli a gruppi, 2022

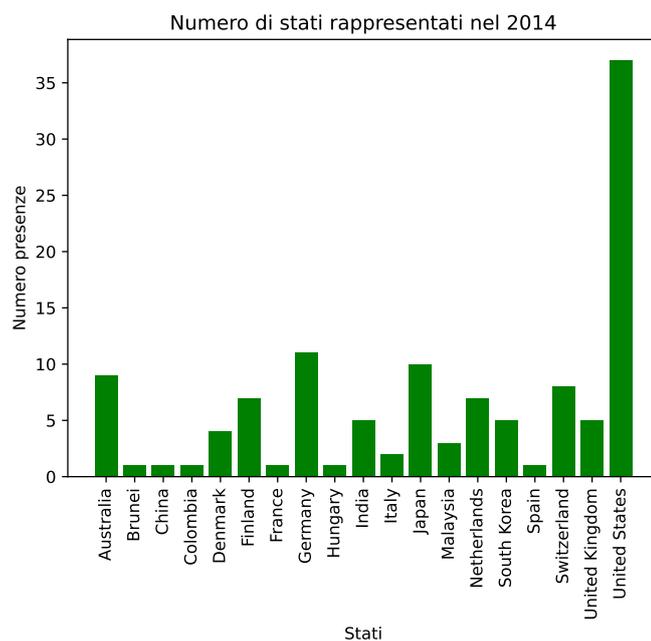


Figura 5.176: Provenienze autori, 2014

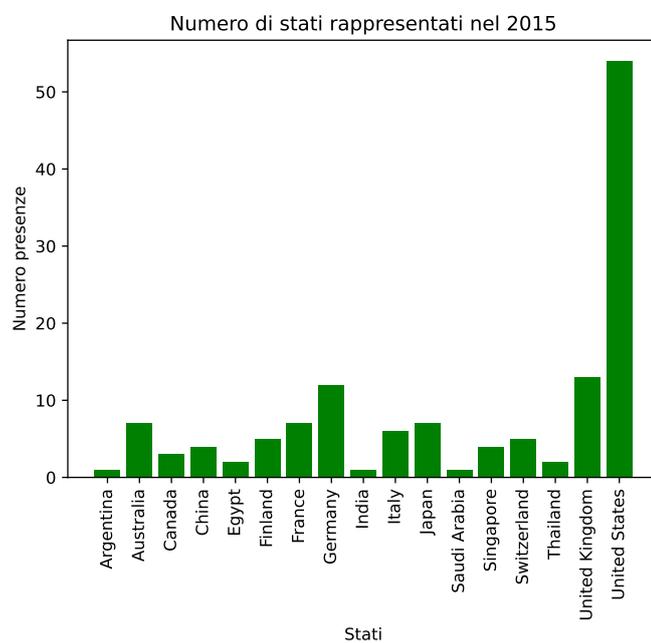


Figura 5.177: Provenienze autori, 2015

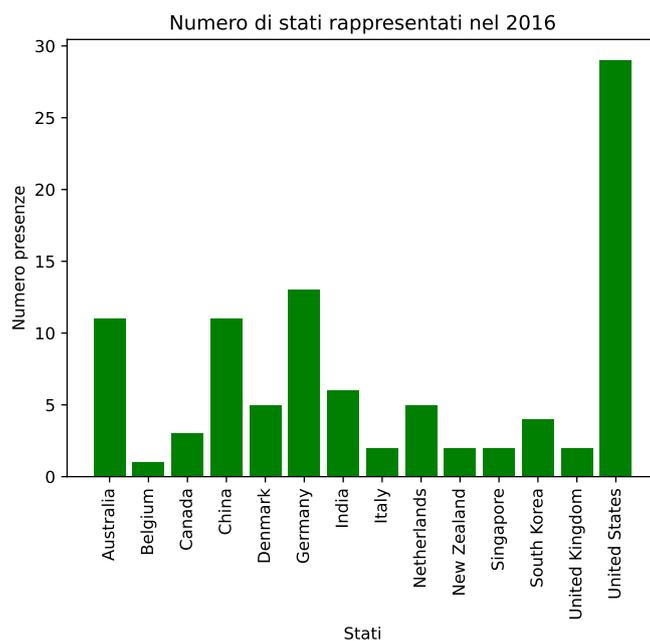


Figura 5.178: Provenienze autori, 2016

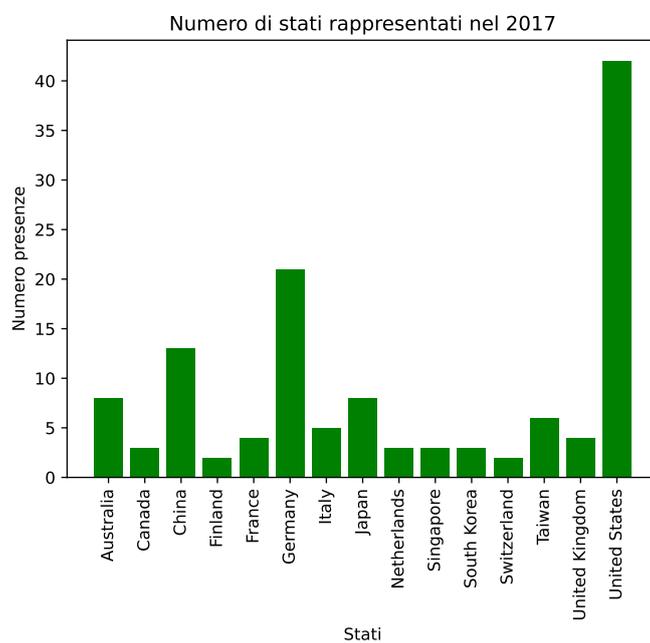


Figura 5.179: Provenienze autori, 2017

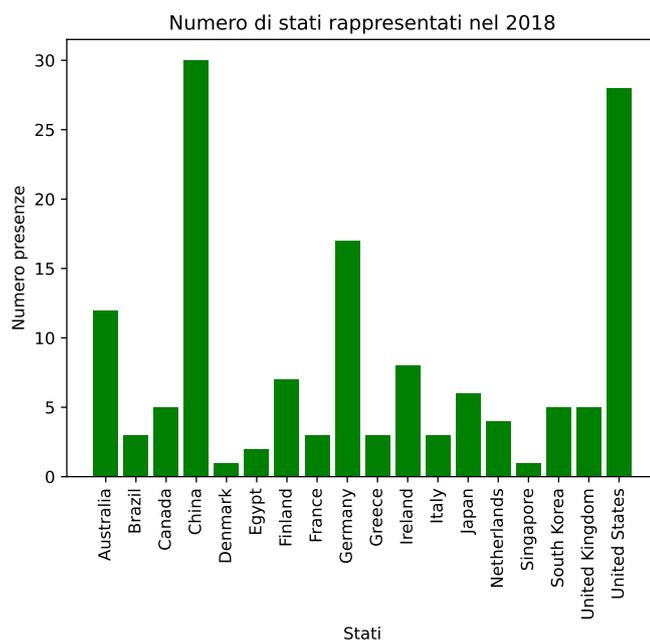


Figura 5.180: Provenienze autori, 2018

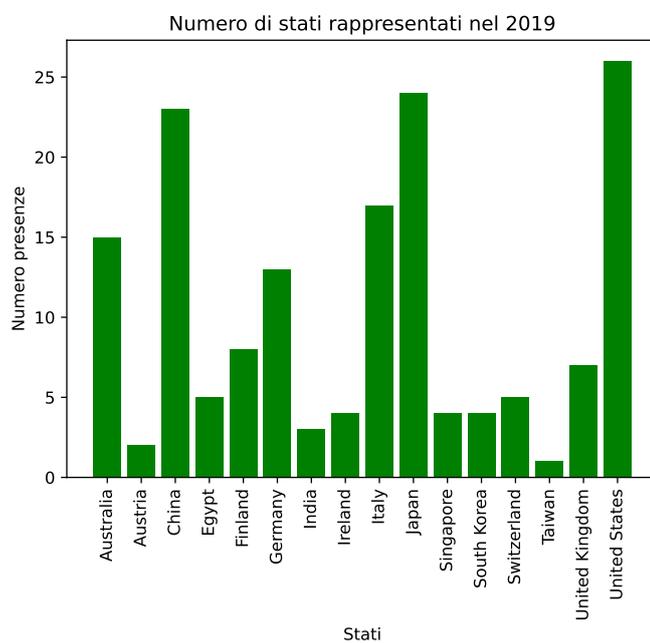


Figura 5.181: Provenienze autori, 2019

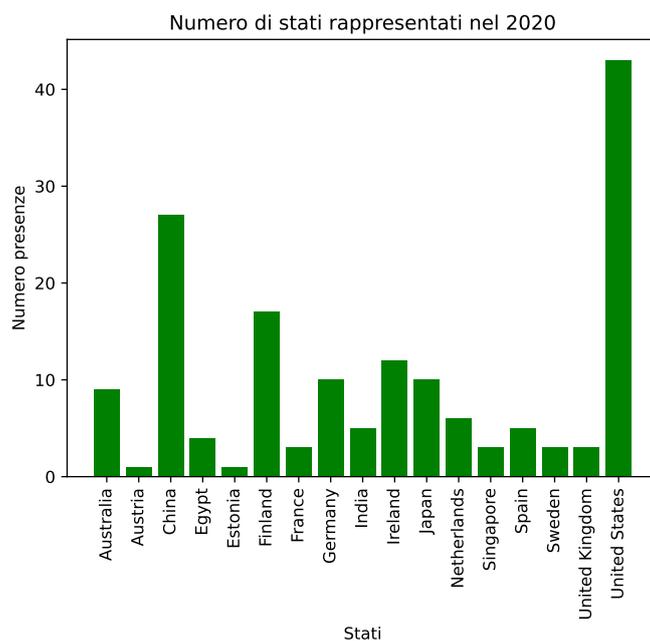


Figura 5.182: Provenienze autori, 2020

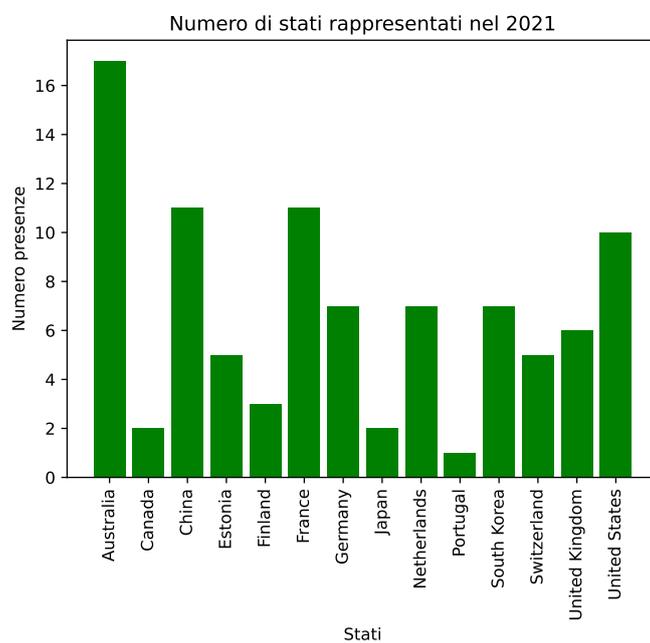


Figura 5.183: Provenienze autori, 2021

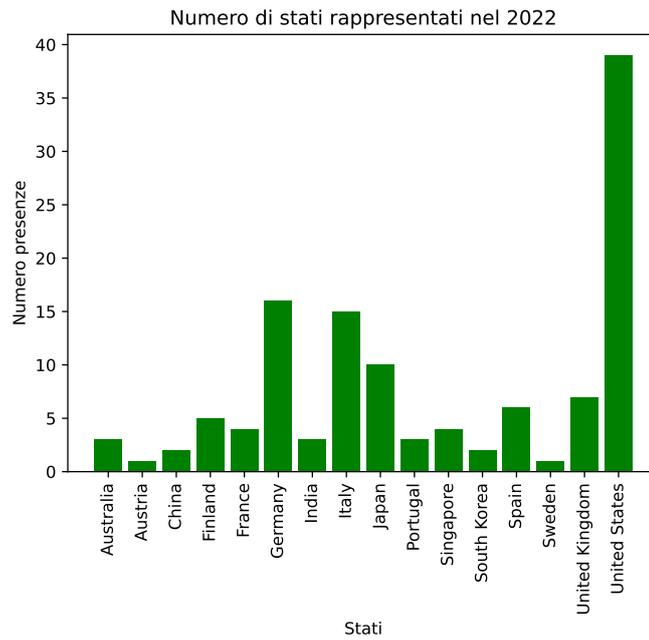


Figura 5.184: Provenienze autori, 2022

Wordcloud

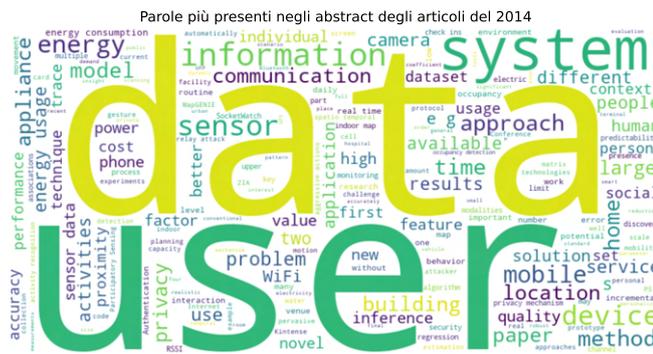


Figura 5.185: Wordcloud abstract degli articoli, 2014





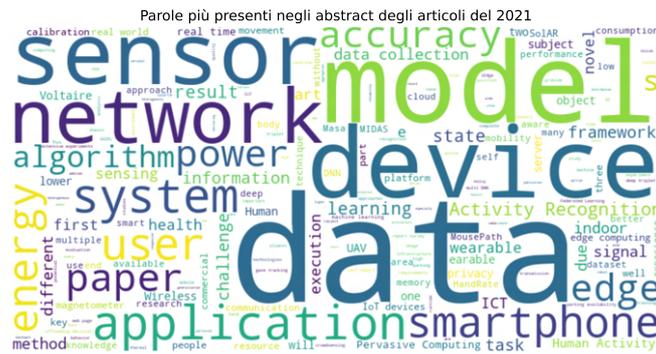


Figura 5.192: Wordcloud abstract degli articoli, 2021

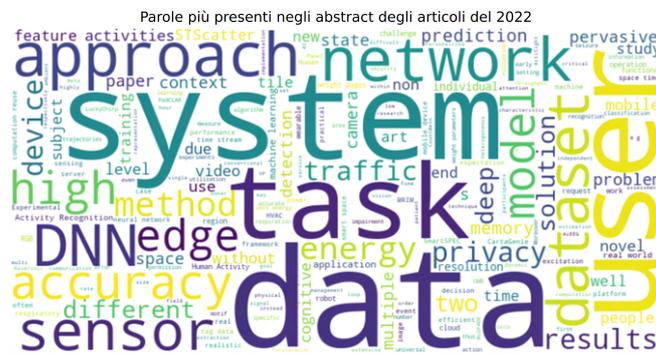


Figura 5.193: Wordcloud abstract degli articoli, 2022

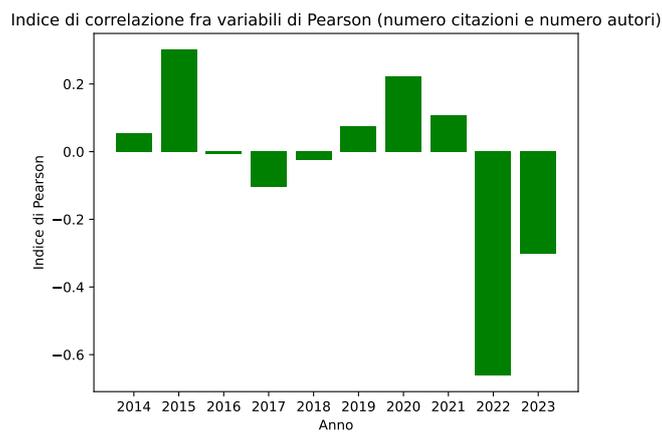


Figura 5.194: Correlazione di Pearson numero citazioni e numero di autori

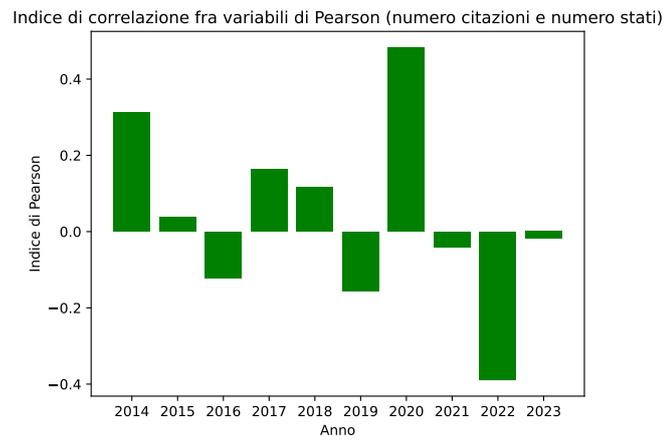


Figura 5.195: Correlazione di Pearson numero citazioni e numero di stati



# Bibliografia

- [1] K. Barefoot, D. Curtis, W. Jolliff, J. R. Nicholson, and R. Omohundro. Defining and measuring the digital economy, us department of commerce bureau of economic analysis. page 210, 2018.
- [2] Federica Laricchia. Big data analytics market value in italy 2015-2020. 2022.
- [3] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013. PMID: 27447038.
- [4] Jay Liebowitz, editor. *Data Analytics and AI*. Auerbach Publishers, Incorporated, 2020. ProQuest Ebook Central.
- [5] Debashish Roy, Rajeev Srivastava, Mansi Jat, and Mustafa Said Karaca. *A Complete Overview of Analytics Techniques: Descriptive, Predictive, and Prescriptive*, pages 15–30. Springer International Publishing, Cham, 2022.
- [6] What is data analytics?, <https://aws.amazon.com/it/what-is/data-analytics/>.
- [7] Huanqin Ji and Wei Gan. Data visualization for making sense of scientific literature. In *2020 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS)*, pages 870–873, 2020.
- [8] Jānis Kampars and Jānis Grabis. Near real-time big-data processing for data driven applications. In *2017 International Conference on Big Data Innovations and Applications (Innovate-Data)*, pages 35–42, 2017.

- 
- [9] Ieee - about, <https://www.ieee.org/about/index.html>.
- [10] Ieee globecom - about, <https://globecom2024.ieee-globecom.org/about-ieee-globecom-2024>.
- [11] Ieee infocom - about, <https://infocom2025.ieee-infocom.org/about>.
- [12] Percom conference, <https://www.percom.org/>.
- [13] Yanyan Wang, Junsheng Zhang, and Xiaodong Qiao. An event-centric scientific literature information organization method. In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 418–425, 2019.
- [14] AL Barabási, R Albert, and H Jeong. Scale-free characteristics of random networks:: the topology of the world-wide web. *PHYSICA A*, 281(1-4):69–77, JUN 15 2000.
- [15] About python, <https://www.python.org/about/>.
- [16] About pandas, <https://pandas.pydata.org/about/>.
- [17] Ieee dataport, <https://ieee-dataport.org/>.
- [18] P. Sedgwick. Pearson’s correlation coefficient. *BMJ : British Medical Journal (Online)*, 345, 2012.
- [19] D. Valencia, R.E. Lillo, and J. Romo. A kendall correlation coefficient between functional data. *Adv Data Anal Classif*, 13:1083–1103, 2019.
- [20] Fatemeh Pariafsai, Manish Kumar Dixit, and Sherecce Fields. Impact of covid-19 pandemic on international research collaboration: A pilot survey results. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–5, 2023.
- [21] Y. Jiang, X. Li, and H. et al. Luo. Quo vadis artificial intelligence? *Discov Artif Intell*, 2:4, 2022.

- 
- [22] M. Di Pierro. What is the blockchain? *Computing in Science and Engineering*, 19(5):92–95, 2017.
- [23] Criptovalute, <https://www.consob.it/web/investor-education/criptovalute>.
- [24] P. De Filippi, C. Wray, and G. Sileno. Smart contracts. *Internet Policy Review*, 10(2), 2021.
- [25] Correlazione, <https://www.treccani.it/enciclopedia/correlazione>.