



UNIVERSITÀ DI PISA

DIPARTIMENTO DI INFORMATICA

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

**Valutazione dell'utilizzo di Crunchbase per analizzare la
migrazione umana altamente qualificata**

Relatori:

Alina Sîrbu

Laura Pollacci

Candidato:

Paul M. Magos

Anno Accademico 2021/2022

Paul M. Magos

**Valutazione dell'utilizzo di Crunchbase per
analizzare la migrazione umana altamente
qualificata**

TESI DI LAUREA TRIENNALE

DIPARTIMENTO DI INFORMATICA

1343

Università di Pisa

22 Luglio 2022

A Janos e Andreea,

migrati nel 2004

Indice

| | | |
|----------|--|-----------|
| 1 | Introduzione | 1 |
| 2 | Studio della migrazione | 4 |
| 2.1 | Dati e approcci tradizionali | 5 |
| 2.1.1 | Migrazione altamente qualificata | 8 |
| 2.2 | Approcci e dati non convenzionali | 9 |
| 2.2.1 | Migrazione altamente qualificata | 10 |
| 3 | Metodologia | 12 |
| 3.1 | Collezione dei dati | 13 |
| 3.1.1 | Query Builder | 14 |
| 3.1.2 | Crunchbase Academic Research Access | 20 |
| 3.2 | Preprocessing ed estrazione delle informazioni | 23 |
| 3.3 | Validazione dei flussi e scorte | 28 |
| 3.4 | Metodi e strumenti utilizzati | 31 |
| 3.4.1 | Correlazione | 31 |
| 3.5 | Conclusione | 33 |
| 4 | Analisi | 34 |
| 4.1 | Confronto tra metodi di collezione | 35 |

| | | |
|----------|--|-----------|
| 4.2 | Descrizione del dataset collezionato | 36 |
| 4.3 | Analisi delle scorte | 38 |
| 4.3.1 | Dati di Crunchbase | 39 |
| 4.3.2 | Confronto Crunchbase con UN | 42 |
| 4.3.3 | Caso di studio: Italia | 43 |
| 4.3.4 | Caso di studio: Gran Bretagna | 45 |
| 4.3.5 | Caso di studio: Europa e nord America | 47 |
| 4.3.6 | Discussione | 49 |
| 4.4 | Analisi dei flussi | 49 |
| 4.4.1 | Dati di Crunchbase | 50 |
| 4.4.2 | Confronto Crunchbase con UN ed Eurostat | 52 |
| 4.4.3 | Confronto UN ed Eurostat con stati aggregati | 55 |
| 4.4.4 | Flussi aggregati UN unito Eurostat | 58 |
| 4.4.5 | Caso di studio: Italia | 59 |
| 4.4.6 | Caso di studio: Gran Bretagna | 63 |
| 4.4.7 | Discussione | 66 |
| 5 | Conclusione | 67 |

Introduzione

Lo studio della migrazione riguarda numerosi campi di studio data la sua importanza economica, sociale e per lo sviluppo di nuove politiche. Una categoria di questo fenomeno in particolare è quella della migrazione altamente qualificata ovvero delle persone con un bagaglio di conoscenze di alto livello. Lo studio di questo fenomeno può portare diversi benefici, come l'analisi dell'integrazione dei migranti altamente qualificati con i nativi della stessa categoria. Ad esempio i dati sugli impieghi possono indicare le aree nelle quali i migranti di questa categoria si specializzino maggiormente [18, 21]. Altri esempi sono lo studio delle competenze dei migranti altamente qualificati in Italia [16], o gli effetti di avvenimenti di grande portata come la Brexit [8].

Tradizionalmente, gli studi si sono affidati a dati ufficiali raccolti da istituzioni e enti. Tuttavia, spesso, i dati ufficiali sono disponibili con ritardo e non sono confrontabili tra stati a causa della mancanza di uno standard nella raccolta [6, 24]. Recentemente, la disponibilità dei Big Data ha portato la ricerca a sperimentare la possibilità di usare nuove fonti di dati per aggirare e colmare le limitazioni tradizionali. Tra le nuove risorse di dati, i cosiddetti social-big data includono diversi tipi di tracce digitali prodotte dalle persone mediante dispositivi mobili, servizi online e piattaforme di social networking [30]. Data la complessità del fenomeno migratorio e nonostante gli sforzi della ricerca, molti aspetti possono essere ulteriormente indagati. In particolare, grazie a piattaforme di social networking orientate

all’ambito lavorativo, come LinkedIn¹ e Crunchbase², è possibile osservare informazioni sugli spostamenti dei professionisti attraverso dati non convenzionali.

Questa tesi propone una metodologia per la collezione e l’estrazione di flussi e scorte migratorio da dati provenienti da Crunchbase, per l’analisi della migrazione altamente qualificata. Inoltre, lo studio mira a valutare l’attendibilità dei dati estratti da Crunchbase mediante il confronto con dati ufficiali, provenienti da Eurostat e United Nations.

Per la raccolta dati confrontiamo due metodologie diverse. La prima si basa su una procedura di web scraping semi-automatica sul motore di ricerca di Crunchbase, e raccoglie dati aggregati e anonimizzati. La seconda utilizza il Academic Access per ottenere dati su utenti Crunchbase. Per determinare i flussi e le scorte sono state effettuate alcune assunzioni sulla nazionalità possibile di un utente, basandosi sui suoi titoli di studio e sui lavori passati. Infine, la validazione è stata effettuata attraverso il confronto di scorte e flussi con i rispettivi dati presenti nel Multi-aspect Integrated Migration Indicators (MIMI) dataset [12].

Il lavoro svolto ha permesso di acquisire nuove competenze legate all’analisi e la manipolazione del traffico dei siti web, alla programmazione in Python, alla realizzazione di grafici per visualizzare dati complessi come flussi migratori, e al calcolo di correlazioni per i dati visualizzati. Inoltre, sono stati appresi concetti come i flussi migratori e le scorte di migranti, e sono state comprese varie domande di ricerca studiate in ambiti diversi dall’informatica.

Il resto della tesi è strutturato come segue. Nel Capitolo 2 vengono introdotti gli argomenti centrali dello studio, con particolare attenzione alla migrazione altamente qualificata. Il capitolo presenta dati e approcci tradizionale (Sezione 2.1) e, successivamente, i dati e metodi non convenzionali (Sezione 2.2). Il Capitolo 3 illustra le metodologie e le tecnologie utilizzate per la collezione dei dati, per l’invio di richieste ai server Crunchbase, l’organiz-

¹LinkedIn <https://www.linkedin.com/>

²Crunchbase: <https://www.crunchbase.com/>

zazione dei dati ottenuti attraverso l'accesso accademico e, infine, la validazione dei dati collezionati. Il Capitolo 4 si concentra sulle analisi effettuate, in primo luogo sull'utenza Crunchbase e la sua provenienza. In seguito vengono mostrati e discussi tutti i casi di studio di confronto tra i dati Crunchbase e i dati del MIMI. Alcuni casi di studio si concentrano su alcuni paesi di interesse, come l'Italia e il Regno Unito. Infine, il Capitolo 5 conclude l'elaborato discutendo i risultati ottenuti e le limitazioni incontrate, insieme ai possibili studi futuri.

Capitolo 2

Studio della migrazione

Questo capitolo descrive il fenomeno migratorio e le metodologie per studiarlo presenti in letteratura. Il capitolo tratta sia approcci tradizionali che non convenzionali, trattando con maggior attenzione il contesto della migrazione altamente qualificata.

Il fenomeno migratorio è sempre stato una costante nella storia dell'uomo [30]. Durante i secoli ci sono stati diversi flussi migratori, dettati in molti casi da guerre e povertà, come le migrazioni dal continente Africano verso l'Europa [17]. Di conseguenza, lo studio della migrazione ha interessato diversi settori di studio, e.g. sociologico, economico, demografico.

La migrazione è un fenomeno importante per il modo in cui rimodella la società, rendendola di natura molteplice. Inoltre, lo studio permette di determinarne diverse sfumature del fenomeno come le migrazioni temporanee dei lavoratori o i migranti che si stabiliscono in altri stati [19].

La migrazione è tipicamente analizzata tramite la quantificazione di flussi e scorte migratorie. I flussi migratori definiscono il numero di migranti che si spostano attraversano un confine, in un determinato periodo di tempo [4]. Invece, le scorte migratorie definiscono il numero di migranti presenti in una zona in un determinato momento [4]. Il termine im-

migrante viene usato per definire un non residente che intende stabilirsi in un luogo per un periodo superiore a 12 mesi[4]. Viene definito emigrante una persona che sposta la propria residenza in uno stato estero in cui intende rimanere per un periodo superiore a 12 mesi [4].

Ad oggi, sono stati impiegati sia dati e modelli tradizionali che non convenzionali al fine di indagare le questioni ancora aperte in tema di migrazione [30]. Per dati tradizionali si intendono dati ottenuti attraverso istituzioni nazionali o internazionali (censimenti, sondaggi, indagini, registri della popolazione). I dati non convenzionali sono, in generale, *Big Data*, ad esempio dati da social network e reti mobili.

2.1 Dati e approcci tradizionali

Tradizionalmente, la migrazione è studiata mediante dati ufficiali derivanti da diverse fonti tra cui sondaggi, registri della popolazione [30] e censimenti [2]. Questi dati sono molto utili nella ricerca del fenomeno della migrazione, ma soffrono di alcuni problemi. I dati vengono raccolti da vari enti privati o pubblici. Per esempio, in Italia i censimenti vengono effettuati dall'Istituto nazionale di statistica (ISTAT) ogni 10 anni. Tuttavia, questi enti possono registrare la popolazione in modo diverso da stato a stato [24], producendo delle incongruenze nei dati. Un altro esempio è lo studio demografico delle scorte per United Nation (UN) che viene stimato tramite modelli migratori ¹.

Alcuni stati non riescono a tener conto dei nativi che rientrano in patria, come ad esempio la Germania che sfrutta la cittadinanza come nazionalità e non il luogo di nascita portando a sovrastimare i numeri di migranti [10]. Un altro problema è il fatto che non esiste una definizione univoca di migrante e diversi dataset e analisi usano terminologie differenti [1].

¹UN Migration Stock Documentation: https://www.un.org/en/development/desa/population/migration/data/estimates2/docs/MigrationStockDocumentation_2019.pdf.

Un altro svantaggio dei dati tradizionali è il tempo necessario per raccoglierli. La tempestività per stabilire se una persona si è trasferita, per quanto riguarda l’Europa, potrebbe richiedere anche due anni [6]. Molti paesi Europei non hanno statistiche sugli emigranti a causa del fatto che gli emigranti non vengono incentivati nel segnalare il loro status alle amministrazioni di provenienza. Di conseguenza, i dati vengono raccolti con ritardo e possono non essere comparabili tra paesi.

Ad oggi, esistono numerosi set di dati relativi alla migrazione, tra cui ISTAT², Eurostat³, OECD⁴, United Nations⁵ e il MIMI. I dataset attualmente disponibili differiscono per copertura geografica e temporale, accesso e livello di dettaglio. I dati Eurostat riguardano principalmente il continente Europeo, ad esclusione della Gran Bretagna, per la quale non riporta dati. Al contrario, UN fornisce informazioni sui migranti relativi a 200 stati ottenuti attraverso censimenti della popolazione e indagini demografiche⁶. In entrambi i dataset, le scorte di immigrati sono conteggiate su base quinquennale. Livelli di aggregazione di questo tipo possono comportare la perdita di informazioni dettagliate, complicando ulteriormente la ricerca. Tuttavia, ad oggi diversi metodi sono stati proposti per arginare queste problematiche [33].

Il MIMI dataset è un’integrazione di dati da vari fonti tradizionali e non, che useremmo nel nostro lavoro per validare i dati estratti da Crunchbase. Per quanto riguarda il livello di dettagli forniti sui migranti, come genere, età, residenza e istruzione, il MIMI dataset contiene informazioni relative ai migranti senza distinzioni di educazione. Al contrario, EU-

²Statistiche Istat: <http://dati.istat.it/>

³Eurostat Database: <https://ec.europa.eu/eurostat/data/database>

⁴International Migration Database - OECD: <https://stats.oecd.org/Index.aspx?DataSetCode=MIG>

⁵Global Migration Database - United Nations: https://population.un.org/unmigration/index_sql.aspx

⁶UN Migration Stock Documentation: https://www.un.org/en/development/desa/population/migration/data/estimates2/docs/MigrationStockDocumentation_2019.pdf.

ROSTAT ha pubblicato il Labour Force Survey (LFS) [7] che include informazioni relative alle migrazioni di persone altamente formate dai 15 anni in su, dal 1983 al 2020. Tuttavia, l'accesso ai dati dell'Labour Force Survey richiede un'attesa variabile⁷. Inoltre i dati sono disponibili solo per un insieme di stati limitato⁸. Il MIMI dataset comprende dati ottenuti da Eurostat e United Nation, insieme al Social Connectedness Index di Facebook⁹. I dati del MIMI includono:

- Paese di provenienza/nazionalità;
- Paese di destinazione;
- Scorte migratorie ottenute da UN su base quinquennale;
- Flussi migratori annuali ottenuti sia da UN che da Eurostat basati sui cittadini e sui residenti.

Ai fini del lavoro proposto in questa tesi, è stato usato il MIMI dataset in quanto comprende sia dati UN che Eurostat. Tuttavia, il dataset impone alcune limitazioni, che provengono dalle fonti incluse in MIMI:

- I dati UN sulle scorte sono quinquennali;
- I dati Eurostat sono limitati quasi totalmente al continente europeo;
- I dati UN riguardano 200 stati in particolare e rappresentativo per sud est asiatico;
- I dati sono riferiti a tutta la popolazione senza dettagli relativi ai livelli di istruzione.

⁷LFS data: <https://ec.europa.eu/eurostat/web/microdata/overview>.

⁸LFS States: <https://bit.ly/LFSstates>.

⁹Social Connectedness Index Link: <https://dataforgood.facebook.com/dtg/tools/social-connectedness-index>.

Ai fini del lavoro proposto in questa tesi, sono prese in considerazione le scorte relative al 2010, 2015 e 2020. Per i flussi invece vengono considerati quelli tra il 2010 ed il 2020.

2.1.1 Migrazione altamente qualificata

L'analisi delle migrazioni di persone altamente qualificate, ad esempio con titoli accademici e lavori ad alto profilo, ha riscosso un crescente interesse negli ultimi anni, data la sua importanza per la produttività e l'educazione [30]. Un recente studio di Impicciatore et al. [16] studia il fenomeno migratorio internazionale degli studenti italiani laureati negli anni 2007, 2011 e 2015. I dati utilizzati provengono dall'ISTAT e lo studio mostra che gli studenti che decidono di emigrare sono prevalentemente quelli della classe agiata o che hanno ottenuto un voto di laurea elevato. Inoltre, tendono a spostarsi di più gli studenti delle discipline STEM¹⁰ con l'obiettivo di migliorare la loro situazione occupazionale.

L'analisi sull'impatto della Brexit effettuata da Falkingham et al. [8] è focalizzata sulle migrazioni in Europa ed utilizza i dati del Survey of Graduating International Students (SoGIS) [9]. Lo studio, condotto su dati relativi all'anno 2017, confronta i dati precedenti e successivi alla data di inizio effettivo della Brexit (29 marzo 2017) mostrando che eventi di questa portata portano a generare incertezze nei piani sulla migrazione degli studenti in alcuni stati dell'Unione Europea.

Sheffer et al., [29] hanno analizzato le migrazioni interne dei medici brasiliani, dal 1980 al 2014. Lo studio prende in considerazione genere ed età dei migranti e mostra che il 57,7% dei medici nello studio ha migrato, il 93,4% dei medici che hanno studiato in una città con meno di 100,000 abitanti ha migrato in altre città ed infine la percentuale di migranti di sesso femminile (54,2%) è inferiore rispetto a quella maschile (60%).

¹⁰STEM: all'inglese science, technology, engineering and mathematics, indica le discipline di ambito scientifico-tecnologico, come scienza, tecnologia, ingegneria e matematica.

Utilizzando dati sulle migrazioni da OECD, Eurostat, United Nations e dall'LFS, Rainer Münz [20] ha effettuato uno studio sulla dimensione della popolazione migrante Europea. Lo studio mostra come dal 2010 in poi l'Europa avrebbe dovuto competere in misura maggiore rispetto a prima con Australia, USA e Canada per essere scelta come meta dai lavoratori.

Sebbene esistano sondaggi specifici a livello internazionale come il Gallup World Poll¹¹, l'International Social Survey Program (ISSP)¹² e il Pew Global Attitudes Survey¹³, questi ricoprono solo un insieme di stati (Gallup World Poll 160, ISSP 44, Pew 69).

Altre studi possibili sono stati fatti attraverso dati dal programma Erasmus per studiare gli spostamenti degli studenti [25], limitatamente ad alcuni stati dell'Europa.

Nonostante gli sforzi della ricerca, le limitazioni imposte dai dati tradizionali, come copertura, accesso, metodi e tempistiche di raccolta non standardizzate tra paesi, lasciano molte domande aperte nello studio della migrazione umana e, in particolare, di quella altamente specializzata.

2.2 Approcci e dati non convenzionali

Le limitazioni poste dai dati e dall'analisi delle migrazioni tradizionale porta a ricercare metodi alternativi [30]. Esistono diverse fonti di dati non convenzionali che sono state proposte in letteratura, come ad esempio reti mobili [3, 13], la geolocalizzazione degli indirizzi IP [23] e dati dei social network, come Facebook, Twitter [34, 27].

Diversi lavori utilizzano fonti non convenzionali derivate dai social network per analizzare la migrazione umana. In [34] è utilizzata la geolocalizzazione dei post di Twitter. Invece,

¹¹Gallup World Poll: <https://www.gallup.com.analytics/318875/global-research.aspx>.

¹²ISSP: <https://issp.org/>

¹³Pew Global Attitudes Survey: <https://www.pewresearch.org/global/database/>

in [27], gli autori usano la rete di Facebook per stimare le scorte migratorie di alcuni paesi Europei.

L'accesso ai dati non convenzionali avviene di solito tramite API¹⁴ specializzate, facilitando la raccolta di grandi quantità di dati. Allo stesso tempo, le API introducono sfide diverse, legate alla necessità di utilizzare linguaggi di programmazione e al dinamismo delle API stesse. Per queste e altre ragioni l'accesso ai dati di social network, talvolta, può essere arduo seguendo le vie abituali che richiedono l'utilizzo di un'API. Esistono vie alternative per accedere ai dati ad esempio dei social network, come mostrato in [11]. Gli autori hanno sfruttato tecnologie dedicate all'estrazione di dati da piattaforme Web su LinkedIn. Queste tecnologie possono essere descritte in maniera generale come algoritmi di web scraping. Lo scraping è in genere una simulazione dell'interazione umana con un entità web.

2.2.1 Migrazione altamente qualificata

Tra le piattaforme di social networking alcune sono dedicate ai rapporti professionali, come ad esempio LinkedIn¹⁵ e Crunchbase¹⁶. I dati di queste piattaforme possono includere anche informazioni aggiuntive, come le esperienze lavorative e il percorso educativo.

In [31] vengono sfruttati i dati ottenuti da LinkedIn, per studiare la migrazione dei professionisti durante il periodo compreso tra il 2000 ed il 2012. Tra i risultati si denota come gli Stati Uniti si confermino la destinazione più quotata per questo tipo di migranti, sebbene nel periodo di studio (2000-2012) sia diminuita la percentuale di persone che considerano gli Stati Uniti come meta per migrare. Inoltre, per lo stesso periodo, viene osservata una crescita da parte del continente asiatico come meta per le migrazioni altamente qualificate. Perrotta

¹⁴Application Programming Interface

¹⁵LinkedIn: <https://it.linkedin.com/>.

¹⁶CruncBase <https://www.crunchbase.com/>.

et al. [22] hanno collezionato i dati relativi agli utenti dalla piattaforma per reclutatori di LinkedIn, nel periodo Ottobre 2020 - Settembre 2021. I dati sono stati utilizzati per determinare l'utilità e le limitazioni dell'utilizzo di LinkedIn nello studio delle intenzioni migratorie dei professionisti in Europa. Tra i risultati ottenuti viene indicato che gli stati del nord e dell'ovest Europa sono i più quotati da chi tra gli utenti LinkedIn è aperto a trasferimenti legati al lavoro. I dati collezionati hanno permesso quindi di identificare dei potenziali futuri migranti. Tuttavia, questo dataset da solo non è sufficiente per collegare chi ha espresso un desiderio di migrare con chi effettivamente migra.

Per l'accesso ai dati di LinkedIn e Crunchbase viene messa a disposizione una API. Tuttavia, LinkedIn nel 2015 ha limitato l'accesso alla propria API¹⁷ e permette agli utenti di scaricare esclusivamente la propria rete. Crunchbase contiene informazioni su diverse entità come ad esempio organizzazioni, persone e scuole. Inoltre, include le relazioni tra le varie entità come gli investimenti, le relazioni lavorative e i percorsi di studio delle persone. A differenza di LinkedIn, Crunchbase, fornisce l'accesso ai dati per fini di ricerca o attraverso la sottoscrizione di un abbonamento.

Questa tesi si colloca nell'ambito di ricerca della migrazione altamente qualificata attraverso fonti di dati non convenzionali, in particolare lavorando con dati di Crunchbase. Inoltre, questi dati sono stati validati con dati ufficiali contenuti nel MIMI dataset. L'analisi viene effettuata su scala globale e copre 10 anni, dal 2010 al 2020. Al meglio della nostra conoscenza, non è stato prodotto nessuno studio della migrazione umana utilizzando dati provenienti da Crunchbase. Quasi la totalità degli studi si sviluppano sulle organizzazioni ed i loro investimenti, sulle *startup* ed il modo in cui si evolvono [5].

¹⁷Developer Program Changes 2015 LinkedIn:<https://developer.linkedin.com/blog/posts/2015/developer-program-changes>.

Capitolo 3

Metodologia

In questo capitolo sono descritti nel dettaglio gli algoritmi, le strutture, le librerie ed il linguaggio utilizzato per la collezione, la manipolazione dei dati di Crunchbase¹. Infine, viene descritto l'approccio alla validazione dei dati Crunchbase che vengono confrontati con i dati presenti nel MIMI.

Come mostrato in Figura 3.0.1, la metodologia proposta è divisa in tre fasi: collezione, pre-processing e validazione dei dati. La collezione dei dati è stata effettuata attraverso due diversi metodi:

- Custom Query Builder
- Crunchbase Academic Research Access

Il Custom Query Builder è stato utilizzato per la raccolta dei dati fino a che non è stata ottenuta la possibilità di accedere ai dati tramite Academic Research Access. I dati collezionati da Crunchbase seguendo l'approccio descritto in Sezione 3.1.2, sono preprocessati al fine di estrarre le informazioni relative agli utenti, ai flussi e alle scorte migratorie. Dopo la collezione e il preprocessing, è stata affrontata la validazione dei dati. Con questo obiettivo,

¹Crunchbase: <https://www.crunchbase.com/>.



Figura 3.0.1: Diagramma della metodologia

le scorte ed i flussi calcolati a partire dai dati collezionati da Crunchbase sono stati confrontati con quelli nel MIMI mediante un analisi di correlazione. Inoltre, i risultati sono stati visualizzati mediante *chord diagram*², seguendo diversi livelli di aggregazione geografica.

3.1 Collezione dei dati

La collezione dei dati è avvenuta attraverso due metodi: il Custom Query Builder e l'Academic Research Access. Il Custom Query Builder è stato realizzato mediante lo studio del meccanismo di *query* offerto dalla piattaforma Crunchbase. In un secondo momento è stato ottenuto l'accesso accademico ai dati che ha permesso di prelevare i dati ufficiali da Crunchbase.

Per determinare la validità dei dati ottenuti si confrontano con i dati del MIMI dataset, per scorte e flussi. Ogni confronto con i dati ufficiali consiste nel calcolo della correlazione e la visualizzazione dei dati su grafici di dispersione.

²Il chord diagram, in italiano, diagramma a corda rappresenta i flussi e le connessioni tra diverse entità, chiamate nodi. Ciascuna entità è rappresentata da un segmento del perimetro circolare. Gli archi sono tracciati tra entità connesse. Inoltre, la dimensione dell'arco è proporzionale all'importanza o dimensione dell'entità misurata (scorte o flussi)³

3.1.1 Query Builder

Per ottenere i dati si è scelto, nella fase iniziale, di sfruttare il Query Builder offerto dalla piattaforma Crunchbase. Il Query Builder permette di filtrare in base alle entità, tra cui persone, compagnie, investitori, acquisizioni. Dato che l'obiettivo di questa tesi è studiare la migrazione delle persone altamente specializzate abbiamo limitato la collezione dei dati alle persone. Il Query Builder permette di filtrare ulteriormente la ricerca sulla base di varie informazioni degli utenti, tra cui la carriera lavorativa, i titoli di studio, eventi ai quali hanno partecipato e il genere.

Una volta eseguita una ricerca tramite il Query Builder, il sito pubblico di Crunchbase restituisce i primi 5 profili che soddisfano la query più il numero totale di profili. Il nostro approccio non utilizza i profili individuali, ma salva solamente il numero totale di profili, che saranno interpretati come la dimensione delle scorte e flussi migratori. Questo approccio rispetta la privacy degli utenti e risulta in numeri aggregati e anonimizzati.

Per studiare la migrazione abbiamo costruito dei query che selezionano i soli profili utenti di persone che hanno conseguito la laurea in un set di paesi selezionati⁴ (lo stato di laurea sarà considerato un proxy per la nazionalità). Per ottenere le scorte annuali è stato estratto il numero di persone che lavoravano il 1° Gennaio di ogni anno, dal 2010 al 2020, in ciascuno degli stati selezionati. Infine, per ottenere i flussi, è stato estratto il numero di persone che hanno cambiato impiego durante ciascun anno, dal 2010 al 2020, sposandosi tra gli stati selezionati.

Al fine di automatizzare il processo di collezione abbiamo deciso di applicare tecniche di *reverse engineering*⁵. Grazie alla comprensione del meccanismo di comunicazione del sito

⁴Stati Uniti, Canada, India, Filippine, Pakistan, Australia, Cina, Singapore, Brasile, Giappone e continente europeo.

⁵Ingegneria/ingegnerizzazione inversa.

è stato realizzato il Custom Query Builder (Sezione 3.1.1), che genera automaticamente le richieste da inviare, le invia e ne gestisce le risposte.

Custom Query Builder

Per realizzare un programma che interagisse in modo diretto con i server di Crunchbase, è stato analizzato il codice html del sito web. Questo ha permesso di comprendere la fonte dei dati rappresentati sulla piattaforma web. L'analisi ha portato a comprendere che Crunchbase sfrutta un API interna, con richieste/risposte JSON via http per il trasferimento dati. Il Codice 3.1.1 mostra la struttura delle richieste. Sulla base di questa struttura di esempio, è possibile realizzare una lista di query personalizzata. Le query essendo trasferimenti di tipo http necessitano di un header; nel nostro caso la copia dell'header generato dal browser. L'header contiene informazioni, tra cui quelle relative ai linguaggi accettati con il relativo ordine di preferenza (nel nostro caso `it-IT, it;q=0.9, en-US; q=0.8, en; q=0.7`), al tipo di contenuti accettati e inviati (nel nostro caso `application/json` e `text/plain`) e se c'è già stato un collegamento vi saranno dei cookie⁶.

La codifica in Python delle richieste è stata generata dal software Insomnia⁷. Insomnia permette di trasformare in codice Python le richieste http come quella nel Codice 3.1.1, previa estrazione delle richieste inviate nel browser. Nonostante l'utilizzo di Insomnia, la richiesta deve essere comunque adattata per ogni singola query. Inoltre, dopo un certo numero di richieste il sito di Crunchbase richiede un captcha, che nel nostro caso abbiamo eseguito manualmente. Questa procedura è abbastanza scalabile per la quantità di richieste che abbiamo dovuto eseguire per questa analisi.

⁶Piccoli file generati dai siti web salvati nella memoria del computer dell'utente, per mantenere informazioni sulle connessioni effettuate in precedenza.

⁷Insomnia: <https://github.com/Kong/insomnia>.

```

1   {
2     "query": [
3       {
4         "type": "sub_query",
5         "collection_id": "organization.has_alumni.reverse",
6         "query": [
7           {
8             "field_id": "location_identifiers",
9             "operator_id": "includes",
10            "values": ["9dd7a8c5-7b7f-7785-90f4-fed17fa5a6ff"]
11          }
12        ]
13      },
14      {
15        "type": "sub_query",
16        "collection_id": "job.has_past_job.forward",
17        "query": [
18          {
19            "type": "sub_query",
20            "collection_id": "organization.has_past_position.reverse",
21            "query": [
22              {
23                "field_id": "location_identifiers",
24                "operator_id": "includes",
25                "values": ["f110fca2-1055-99f6-996d-011c198b3928"]
26              }
27            ]
28          },
29          {
30            "field_id": "started_on", "operator_id": "lte",
31            "values": ["1/1/2010"]
32          },
33          {
34            "field_id": "ended_on", "operator_id": "gte",
35            "values": ["1/1/2010"]
36          }
37        ]
38      }
39    ]
40  }

```

Codice 3.1.1: Esempio di codice JSON delle richieste che il Query Builder invia ai propri server

Il Codice JSON in 3.1.1 è parte di una query tipica che Crunchbase invia ai propri server per richiedere le scorte migratorie di nazionalità italiana negli Stati Uniti d'America per il 2010. La prima porzione della query (righe 4 - 12) filtra le persone che hanno studiato in un istituto la cui posizione è in Italia (riga 10). La seconda parte della query (righe 15 - 38) impone che i lavori effettuati dalle persone filtrate dalla prima sub query (righe 4 - 12) siano per aziende negli Stati Uniti d'America (riga 25). Inoltre, ogni rapporto lavorativo deve essere iniziato prima e terminato dopo il 1° Gennaio 2010 (righe 29 - 36).

Crunchbase rappresenta ogni stato attraverso un codice UUID⁸. Di conseguenza, è stata realizzata manualmente una lista delle coppie Stato-UUID analizzando il codice html dei luoghi suggeriti nella pagina di compilazione della query.

```

1 def queries_lite_create():
2     Countries = countries_get("SUBSET COUNTRIES")      # Coppie Stato_UUID
3     QUERIES = read_json("queries_lite")                  # Carico le query
4     combination = product(Countries, repeat=2)          # Permutazioni di stati
5     for combo in combination:
6         if combo[0] not in QUERIES:                      # Se la combinazione non è
7             QUERIES[combo[0]] = {}                         # presente la aggiungo
8         if combo[1] not in QUERIES[combo[0]]:
9             QUERIES[combo[0]][combo[1]] = {}
10        for YEAR in range(2010, 2022):
11            QUERIES[combo[0]][combo[1]][str(YEAR)] = {}
12            for gender in {"male", "female"}:
13                payload = payload_update_lite(
14                    Countries[combo[0]]["uuid"],
15                    Countries[combo[1]]["uuid"],
16                    YEAR,
17                    gender
18                )
19                QUERIES[combo[0]][combo[1]][str(YEAR)][gender] = []
20                QUERIES[combo[0]][combo[1]][str(YEAR)][gender].append(payload)
21
22    write_json(QUERIES, "queries_lite.json")           # Scrivo le nuove query
23                                                # calcolate nel file JSON
24                                                # delle query

```

Codice 3.1.2: Funzione `queries_lite_create()` che genera un file JSON contenente le query relative alle scorte migratorie da inviare a Crunchbase

⁸Universally Unique Identifier, in italiano identificativo univoco universale.

La Funzione `queries_lite_create()` nel Codice 3.1.2, realizza un documento in formato JSON (riga 22) che contiene tutte le query relative alle scorte migratorie che, successivamente, saranno inviate a Crunchbase. Utilizza la libreria `json` per la serializzazione e la de serializzazione dei dati (righe 2, 3, 22) e la funzione `product` della libreria `Itertools` (riga 4) per generare un insieme di combinazioni di ordine 2 degli stati selezionati. All'interno del ciclo `for` (righe 5 - 21) vengono generate le query ed aggiunte alla struttura in locale (righe 13, 20).

```

1 def send_queries():
2     dataset = functions.read_json("SavedCountries_lite")
3     link = "/v4/data/searches/people?source=custom_query_builder"
4     ed_country, work_country, year, \
5     gender, query, queries = functions.get_query_lite(dataset)
6
7     while ed_country and work_country and year and \
8         gender and query and queries:
9         try:
10             conn.request("POST",
11                         link,
12                         query,
13                         new_header())
14
15             res = conn.getresponse()
16             data = json.loads(res.read())["count"]
17         except:
18             print("Request Error")
19             return
20
21         if ed_country not in dataset:
22             dataset[ed_country] = {}
23         if work_country not in dataset[ed_country]:
24             dataset[ed_country][work_country] = {}
25         if year not in dataset[ed_country][work_country]:
26             dataset[ed_country][work_country][year] = {}
27
28         dataset[ed_country][work_country][year].update({gender: data})
29
30         functions.write_json(dataset, "SavedCountries_lite.json")
31         functions.write_json(queries, "queries_lite.json")
32
33         time.sleep(random.randint(8, 12))
34         ed_country, work_country, year, \
35         gender, query, queries = functions.get_query_lite(dataset)
36

```

Codice 3.1.3: Funzione `send_queries()` per l'invio delle query.

La Funzione `send_queries()` nel Codice 3.1.3 invia le query contenute nel file creato dalla Funzione `queries_lite_create()` nel Codice 3.1.2. Genera un file di tipo JSON con la struttura del Codice 3.1.4, contenente il risultato ricevuto dai server di Crunchbase (riga 31) aggiornandolo di volta in volta con il risultato della nuova query eseguita (righe 10, 35) nel ciclo `while` (righe 8 - 36). Il Codice JSON 3.1.4 contiene la provenienza delle scorte (riga 2) riferite ad un determinato paese (riga 3). Si ha poi un esempio di alcuni anni a cui si riferiscono le scorte suddivise per genere (righe 4, 15).

```

1  {
2      "China": {
3          "China": {
4              "2021": {
5                  "female": 0,
6                  "male": 0
7              },
8              "2010": {
9                  "female": 39,
10                 "male": 176
11             },
12             "2011": {
13                 "female": 36,
14                 "male": 155
15             }
16         }
17     }
18 }
```

Codice 3.1.4: Codice JSON di esempio delle scorte

Il metodo proposto ha permesso di ottenere tutte le scorte migratorie separate per nazionalità, residenza, anno e genere per i paesi selezionati. In termini di tempo, ottenere le risposte per tutte le scorte ha richiesto $\simeq 11$ ore, incluso il tempo necessario per generare e cambiare l'header quando non era più valido.

Il Custom Query Builder ha delle limitazioni, la più importante di queste è il fatto che i filtri nelle query (Sezione 3.1.1) devono valere sempre tutti. Quindi il Custom Query Builder conta le persone varie volte, una per ogni città in cui hanno studiato. Inoltre, considera i

generi maschili e femminili, ma gli utenti possono scegliere di non menzionare il genere portando il Custom Query Builder ad ignorarli nelle richieste.

3.1.2 Crunchbase Academic Research Access

L'accesso all'API accademica di Crunchbase è possibile sottomettendo una domanda a Crunchbase che viene accettata dall'azienda dopo una valutazione del progetto e un colloquio. Una volta ottenuto l'accesso si possono inviare richieste ai server Crunchbase. Le richieste inviabili attraverso l'API condividono in parte la struttura di quelle nel Codice 3.1.1, così come alcune limitazioni (Sezione 3.1.1). Tuttavia, l'accesso all'API accademica consente di scaricare in locale tutte le entità più significative del sito.

Tra le entità a disposizione⁹ ci focalizziamo sulle organizzazioni (*organizations*), i titoli (*degrees*), i lavori (*jobs*) e le persone (*people*). L'entità *degrees* è costituita dai titoli acquisiti da un utente presso una determinata sede. L'entità *jobs* è costituita invece dai lavori presenti e passati di un utente preso determinate organizzazioni. La struttura dei dati è organizzata in diversi file csv, similmente alle tabelle di un database relazionale. Ogni file rappresenta un tipo di entità con i propri attributi come chiave primaria, chiavi referenziate e le informazioni relative ad un elemento dell'entità come la posizione.

Questi dati sono stati usati per calcolare le scorte di migranti e i flussi migratori.

Le informazioni ottenute sono state organizzate in file JSON dove, per ogni utente è presente la posizione dichiarata, il genere, la lista dei lavori effettuati con informazioni relative al luogo e al periodo di lavoro e lista dei diplomi ottenuti con relativa posizione.

La Funzione `totalCreate()` nel Codice 3.1.5 si occupa di organizzare i dati nella struttura JSON descritta nel Codice 3.1.6. Vengono caricati i dati dalle quattro entità interessate

⁹Documentazione di Crunchbase: <https://data.crunchbase.com/docs/daily-csv-export>.

```

1 def totalCreate( bulk_dir=None, save=False):
2
3     all_people = people_parsing(bulk_dir=bulk_dir)
4     activities = organization_parsing(bulk_dir=bulk_dir)
5     jobs = jobs_parsing(all_people=all_people,
6                         activities=activities, bulk_dir=bulk_dir)
7     degrees = degree_parsing(all_people=all_people,
8                               activities=activities, bulk_dir=bulk_dir)
9     # COMMON STRINGS
10    p_loc = "person_location"
11    p_gen = "person_gender"
12    jperson_uuid = "job_person_uuid"
13    dperson_uuid = "person_degree_uuid"
14
15    my_persona_db = {}
16    with alive_bar(len(jobs) + len(degrees), title="Total", \
17                   force_tty=True, spinner="classic") as Total_bar:
18
19        for each in jobs:
20            if jobs[each][jperson_uuid] not in my_persona_db:
21                my_persona_db[jobs[each][jperson_uuid]] = {
22                    p_gen: jobs[each][p_gen],
23                    p_loc: jobs[each][p_loc],
24                    'jobs': [],
25                    'degrees': []
26                }
27            jobp = jobs[each]
28            my_persona_db[jobp[jperson_uuid]]['jobs'].append(jobp)
29            Total_bar()
30
31
32        for each_deg in degrees:
33            if degrees[each_deg][dperson_uuid] not in my_persona_db:
34                my_persona_db[degrees[each_deg][dperson_uuid]] = {
35                    p_gen: degrees[each_deg][p_gen],
36                    p_loc: degrees[each_deg][p_loc],
37                    'jobs': [],
38                    'degrees': []
39                }
40            degp = degrees[each_deg]
41            my_persona_db[degp[dperson_uuid]]['degrees'].append(degp)
42            Total_bar()
43
44
45    if save:
46        with open("Crunchbase Data/total.json", "w+") as outFile:
47            outFile.write(json.dumps(my_persona_db, indent=4))
48    return my_persona_db

```

Codice 3.1.5: Funzione che genera la struttura degli utenti con relative informazioni

(righe 3 - 8). Attraverso un ciclo `for` (righe 19 - 30) vengono prese le informazioni relative ai lavori e vengono memorizzate in una struttura locale (righe 28). Lo stesso processo viene applicato per i titoli di studio (righe 32 - 43). Viene utilizzata una barra di progresso per visualizzare lo stato del processo (righe 17, 29, 42). Infine, se si passa il parametro `save` (riga 45) vero, verrà serializzato tutto il contenuto della struttura realizzata in locale. Viene in ogni caso restituita al chiamante tutta la struttura (riga 48).

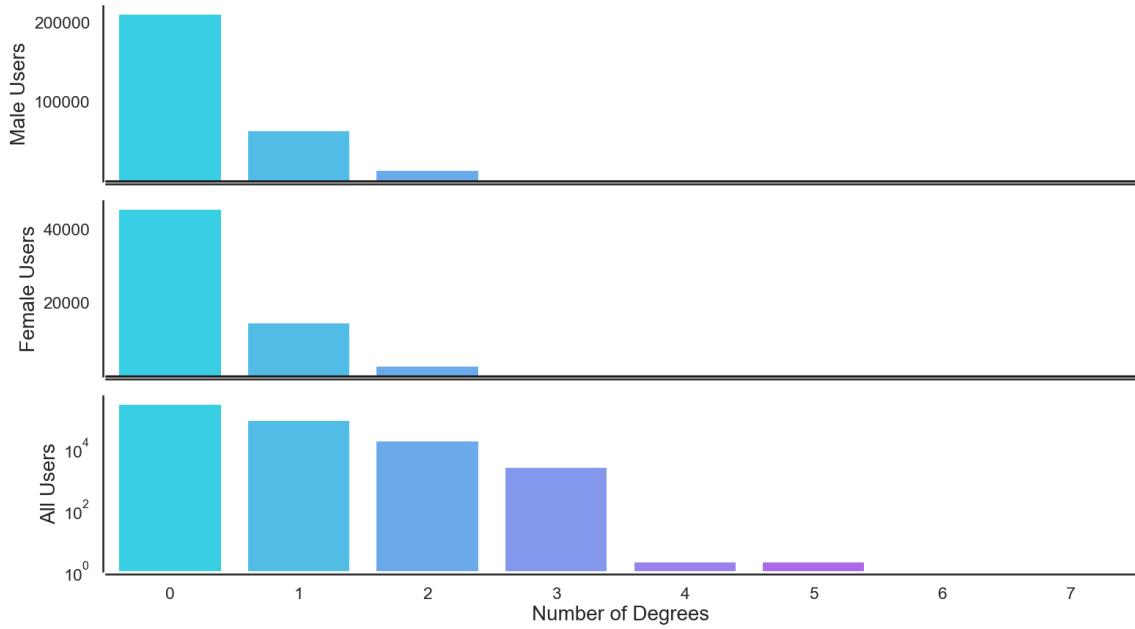
```

1  {
2      "person_UUID": {
3          "person_gender": "male",
4          "person_location": "USA",
5          "jobs": [
6              {
7                  "job_organization_uuid": "ORG_UUID",
8                  "job_start_date": "2005-10-01",
9                  "job_end_date": "2014-06-01",
10                 "job_is_current": false,
11                 "job_location": "USA"
12             },
13             {
14                 "job_organization_uuid": "ORG_UUID",
15                 "job_start_date": "2001-01-01",
16                 "job_end_date": "2002-01-01",
17                 "job_is_current": false,
18                 "job_location": "GBR"
19             }
20         ],
21         "degrees": [
22             {
23                 "university_degree_uuid": "UNI_UUID",
24                 "degree_completed": true,
25                 "university_location": "USA"
26             }
27         ]
28     },
29 }
```

Codice 3.1.6: Codice JSON di esempio per un utente

Il Codice 3.1.6 rappresenta l'esempio di come vengono organizzate le informazioni relative ad un utente, nella struttura generata dalla Funzione `totalCreate()` nel Codice 3.1.5. Per ogni utente, rappresentato dal suo *UUID* (riga 2), si ha il suo genere (riga 3), la location che dichiara (riga 4), la lista dei lavori (righe 5 - 20) e dei titoli di studio (righe 21 - 27).

3.2 Preprocessing ed estrazione delle informazioni

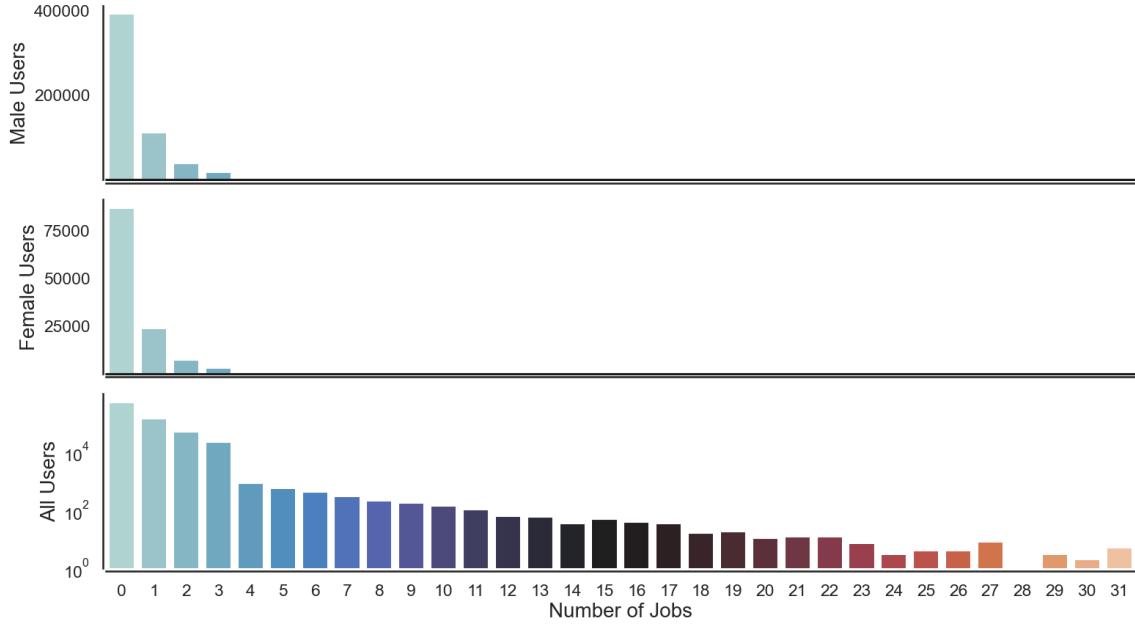


(a) Istogramma dei titoli di studio

La fase di *preprocessig* ed estrazione delle scorte e dei flussi dai dati Crunchbase, ha avuto fondamento grazie un’analisi degli utenti Crunchbase. Questa ha determinato quale fosse la distribuzione dei titoli di studio e dei lavori sulla piattaforma. Il grafico in Figura 3.2.1a mostra che la maggior parte degli utenti ha tra uno e tre titoli di studio. Dalla Figura 3.2.1b si nota che maggior parte degli utenti ha avuto/ha almeno un lavoro.

Abbiamo definito la “nazionalità” come il paese più ricorrente tra le posizioni delle istituzioni presso le quali la persona ha ottenuto titoli di studio¹⁰. Se la nazionalità non è deducibile dai titoli di studio, ad esempio per mancanza della localizzazione delle istituzioni, si utilizza la posizione dell’azienda per cui l’utente ha lavorato meno di recente. La ”residenza” è stata invece dedotta dai trasferimenti lavorativi degli utenti, definendola attraverso la posizione dell’azienda con la quale l’utente ha terminato il rapporto lavorativo.

¹⁰A parità di numeri viene considerato il titolo di studio conseguito meno di recente.



(b) Istogramma dei lavori

La Funzione `extractStocks(total=None, save=False, counter=0)` nel Codice 3.2.7 scorre i profili utente che possono essere passati in forma di dizionario oppure vengono caricati dalla memoria se presenti (riga 3)). Per ogni utente si controlla che abbia avuto un impiego lavorativo e si estrapola la nazionalità dai titoli di studio (riga 14). Successivamente si scorre la lista dei suoi lavori (righe 17 - 42), e in questo ciclo si ottengono la posizione dell’azienda in cui ha lavorato e le date relative al rapporto lavorativo (righe 20 - 32). Infine, si incrementano i valori delle scorte (in locale) in base ai dati che vengono incontrati (riga 41). Come la Funzione `totalCreate()` nel Codice 3.1.5, anche questa permette di usare il parametro `save` per salvare il risultato; in ogni caso restituisce al chiamante le scorte calcolate.

Il preprocessig dei flussi viene effettuato mediante un processo identico a quello della Funzione `extactStocks()` nel Codice 3.2.7, con un ciclo in più all’interno del ciclo dei lavori (righe 17 - 42) per scorrere gli altri lavori dell’utente e confrontarne il periodo d’inizio e fine. Questo permette di stabilire se durante un determinato anno l’utente ha cambiato

```

1 def extractStocks(total=None, save=False, counter=0):
2     my_db = NestedDict()
3     if total is None:
4         total = read_total()
5     if counter == 0:
6         counter = count_jobs(total)
7     with alive_bar(counter, title="Stocks",
8                     force_tty=True, spinner="classic") as stocks_bar:
9
10    for person in total:
11        ed_country := get_nationality(total[person]['degrees'],
12                                         total[person]['jobs'])
13        if not len(total[person]['jobs']) or
14            not ed_country:
15            continue
16
17    for person_job in total[person]['jobs']:
18        stocks_bar()
19        if math.isnan(float(person_job['job_location'])) or
20            math.isnan(float(person_job['job_start_date'])):
21            continue
22
23        job_loc = person_job['job_location']
24        job_start = dt.strptime(person_job['job_start_date'],
25                                "%Y-%m-%d")
26        if math.isnan(float(person_job['job_end_date'])):
27            if person_job['job_is_current']:
28                job_end = dt.strptime("2023", "%Y")
29            else:
30                continue
31        else:
32            job_end = dt.strptime(person_job['job_end_date'],
33                                  "%Y/%m/%d")
34
35    for year in range(2010, 2021):
36        year_t = dt.strptime(str(year), "%Y")
37        if year_t < job_start or year_t > job_end:
38            continue
39
40        gender = genderPrs(total[person]['person_gender'])
41        my_db[ed_country][job_loc][str(year)][gender] += 1
42
43
44    if save:
45        fileName = "Crunchbase Query Results/query_stocks_results.json"
46        stockFile = open(fileName, "w+")
47        stockFile.write(json.dumps(my_db, indent=4))
48    return my_db

```

Codice 3.2.7: Codice per l'estrazione delle scorte

residenza.

```
1   {
2     "USA": {
3       "USA": {
4         "2010": {
5           "male": 22,
6           "female": 511082,
7           "unknown": 1
8         },
9       },
10      "ITA": {
11        "2015": {
12          "male": 0,
13          "female": 11,
14          "unknown": 0
15        },
16        "2016": { ... }
17      },
18    },
19    "CAN": { ... }
20 }
```

Codice 3.2.8: Codice di esempio scorte Academic Research Access

Il metodo proposto ha premesso di ottenere le scorte di migranti organizzate nella medesima struttura di quelle nel MIMI dataset. Un esempio è mostrato nel Codice 3.2.8. Il primo stato (e.g. righe 2, 19) indica la nazionalità delle scorte, il secondo (e.g. righe 3, 10) indica il paese in cui sono state conteggiate. Per ogni coppia si ha poi la lista degli anni con i relativi valori suddivisi per genere (righe 4-8, 11-16).

La Funzione ricorsiva `common_items(d1, d2)` nel Codice 3.2.9 effettua l'intersezione di due dict passati come parametro. Per ogni chiave `k` presente in entrambi i dizionari (6), se la chiave è istanza di dict viene richiamata ricorsivamente la funzione sui valori contenuto nei dizionari alla chiave `k`; altrimenti se la chiave è un numero compreso tra 2010, 2015 e 2020 (gli anni delle scorte in UN) viene ritornato il valore delle scorte presente in entrambi i dizionari. La Funzione `stockIntersection()` nel Codice 3.2.10 carica i dati delle scorte (UN e Crunchbase) in locale (righe 9, 10). Se chiamata senza parametri restituisce il risultato della funzione `common_items(d1, d2)` in un dataframe (righe 11, 28, 29). La Funzione

Codice 3.2.10: Codice che interseca le scorte ufficiali con quelle Crunchbase

`stockIntersection()` accetta quattro parametri come filtri: nazionalità, stato scorte, continente della nazionalità, e continente delle scorte. Tutti i filtri vengono controllati all'interno del ciclo che scorre i dati ottenuti dall'intersezione di UN e Crunchbase (righe 14 - 26).

Il preprocessing descritto in questa sezione ha permesso di selezionare tutte le coppie di paesi (nazionalità - paese scorte) per cui sono presenti i dati sia nei dati ufficiali che in quelli Crunchbase. Ciò ha permesso di utilizzare i dati per la validazione delle scorte direttamente dalla struttura generata dalla Funzione `stockIntersection()` nel Codice 3.2.10.

Per ottenere l'intersezione dei flussi Crunchbase con i flussi dei dati ufficiali è stato utilizzato un codice simile a quello per le scorte con un ciclo `for` ulteriore per scorre le residente degli utenti. Dato che per i flussi è stata usata la medesima struttura delle scorte, generata in fase di *preprocessing* (Sezione 3.2). Con una leggera modifica che annida un ulteriore paese nel Codice Json 3.2.8 formando così una tripla cittadinanza-residenza-destinazione.

3.3 Validazione dei flussi e scorte

In questa sezione viene descritto il processo di validazione delle scorte e dei flussi collezionati da Crunchbase. Per visualizzare i dati migratori sono generati grafici di dispersione nei quali l'asse delle X rappresenta i dati di Crunchbase ottenuti mediante Academic Research Access e l'asse delle Y dai dati di confronto che possono essere dati Eurostat, UN o anche l'unione dei due. Ogni grafico presenta nella parte inferiore destra i valori delle correlazioni per validare i dati Crunchbase.

Il Codice 3.3.11 genera un grafico a dispersione con i valori Crunchbase sull'asse orizzontale e i valori dei dati ufficiali sull'asse verticale (righe 11 - 15). Inoltre, solo per le scorte, genera una linea che indica la tendenza che hanno i valori delle scorte (riga 8). Infine, aggiunge il riquadro delle correlazioni al grafico (righe 19 - 23). La correlazione viene

```

1 df_matplot # ->> DATI INTERSECATI
2
3 # GENERO TESTO CON TUTTE LE CORRELAZIONI
4 string_on_plot = correlation_calc(df_matplot)
5
6 # SCATTERPLOT
7 p = sns.scatterplot()
8 if stock: sns.regplot(x='Crunchbase',
9                         y='Official',
10                        data=df_matplot)
11 plt.scatter(data=df_matplot,
12             x="Crunchbase", y="Official",
13             c="Crunchbase", cmap=color_map,
14             edgecolors="black", linewidths=1,
15             alpha=0.6, s=100)
16 # RIGHT BAR
17 cbar = plt.colorbar()
18
19 # CORRELATIONS TEXT BOX
20 ob = offsetbox.AnchoredText(string_on_plot, loc="lower right",
21                             borderpad=2.5, prop=dict(size=15))
22 ob.patch.set(boxstyle='round, pad=0.6')
23 p.add_artist(ob)

```

Codice 3.3.11: Codice per realizzare i grafici di dispersione

calcolata dalla Funzione `correlation_calc()` (Codice 3.3.12).

La Funzione `correlation_calc()` nel Codice 3.3.12 genera una stringa contenente le correlazioni di Pearson e Spearman per i dati all'interno del dataframe, parametro della funzione. La funzione è strutturata per adattarsi al calcolo di correlazioni sia per stock che per flussi. Infatti, attraverso il parametro `flows` è possibile dire se il dataset passato si riferisce a flussi migratori (di default si calcolano le scorte). Il Codice, per le scorte, si aspetta di trovare nel dataframe passato colonne con i nomi "Crunchbase" (riga 2) e la stringa passata nel parametro `source_`(riga 3). Vengono mascherati i dati invalidi in entrambi i dataset con valori pari a 0 (righe 12, 13), e successivamente vengono calcolare le varie correlazioni (righe 17 - 30). Se il parametro `flows` è True, le colonne cercate nel dataframe saranno composte dal nome delle fonti (Crunchbase o fonte ufficiale) e da "`_cit`" o "`_res`" (riga 6), ed inoltre tutti i calcoli verranno eseguiti sia per i cittadini che per i residenti (riga 7).

```

1 def correlation_calc(df, source_="UN", flows=None):
2     crunch = "Crunchbase"
3     source = source_
4     out = {}
5     PearsLogOf = "PearsonLog" + source
6     el = "_cit" if flows else ""
7     while True:
8         if df[crunch + el].any() and df[source + el].any():
9             np.seterr(divide='ignore')
10
11         # MASK INVALID VALUES
12         maCruch = np.ma.masked_invalid(df[crunch + el])
13         maOff = np.ma.masked_invalid(df[source + el])
14         logMaCruch = np.log(maCruch)
15         logMaOff = np.log(maOff)
16
17         # PEARSON CORR CALC
18         out["Pearson" + el] = round(
19             np.ma.corrcoef(maCruch, maOffDF)[0][1], 2)
20         out["PearsonLog" + el] = round(
21             np.ma.corrcoef(logMaCruch, logMaOff)[0][1], 2)
22         out[PearsLogOf + el] = round(
23             np.ma.corrcoef(maCruch, logMaOff)[0][1], 2)
24         out["PearsonLogCB" + el] = round(
25             np.ma.corrcoef(logMaCruch, maOffDF)[0][1], 2)
26
27         # SPEARMAN CORR CALC
28         df_corr_spearman = df.corr(method='spearman')
29         out["Spearman" + el] = round(
30             df_corr_spearman[crunch + el][source + el], 2)
31
32         if flows:
33             if el == "_cit":
34                 el = "_res"
35                 continue
36             break
37
38         # TO STRING
39         if flows:
40             new_d = "\n".join("{}: {!r}".format(k, out[k])
41                             for k in sorted(out, key=len, reverse=False)
42                             if k.endswith("_cit"))
43             new_d2 = "\n".join("{}: {!r}".format(k, out[k])
44                             for k in sorted(out, key=len, reverse=False)
45                             if k.endswith("_res"))
46             return new_d, new_d2
47
48         return "\n".join("{}: {!r}".format(k, out[k])
49                         for k in sorted(out, key=len, reverse=False))

```

Codice 3.3.12: Codice per il calcolo delle correlazioni

La correlazione di Pearson ci dice se vi è una relazione lineare tra due variabili. Anche se il coefficiente di correlazione di Pearson indica che non ci sia correlazione lineare potrebbe esistere una relazione di altro tipo. Per questo motivo, i dati vengono sottoposti anche al calcolo del coefficiente di correlazione per ranghi di Spearman, che determina relazioni monotone. Quando i due indici sono vicini abbiamo una conferma della relazione

3.4 Metodi e strumenti utilizzati

Tutto il codice è stato scritto in Python e sviluppato utilizzando l'IDE PyCharm. E' stata utilizzata la versione *education* fornita agli studenti dall'Università di Pisa, che ha svolto le operazioni di installazione delle librerie in modo automatico. Le librerie Python utilizzate riguardano:

- lettura di file JSON: json;
- manipolazione dei dati: dict, pandas¹¹, alive-progress [26] usata per determinare lo stato del processo;;
- realizzazione di grafici: matplotlib [15], seaborn [32]. Inoltre, plotapi¹² è stata usata per visualizzare in grafici circolari le scorte ed i flussi;
- calcolo delle correlazioni: NumPy [14] per Pearson, e pandas per Spearman

3.4.1 Correlazione

L'Indice di correlazione è una misura che ci permette di stabilire quanto due insiemi di dati siano correlati, ovvero quanto i valori di uno dipendono dall'altro. I due indici di

¹¹<https://pandas.pydata.org/docs/>

¹²Plotapi: <https://plotapi.com/>

interesse per questo elaborato sono l'indice di Pearson per correlazioni lineari [28] e l'indice di Spearman per correlazioni non lineari [35]. In questo elaborato abbiamo utilizzato l'indice di correlazione per validare i flussi e scorte migratorie estratte da Crunchbase.

Pearson Pearson impone alcune condizioni per i dati:

- Le due variabili devono essere entrambe di tipo quantitativo;
- Le due variabili devono avere dati che si riferiscono allo stesso caso;
- Le due variabili a confronto devono avere una crescita lineare, se questo non avviene si trasformano una o entrambe le variabili secondo una scala logaritmica;
- Non devono esserci Outliers tra i dati¹³;
- Entrambe le variabili devono avere una distribuzione normale, che può essere verificata per campioni minori ai 5000 valori tramite il test di Shapiro-Wilk¹⁴.

Il coefficiente di correlazione di Pearson per due variabili random si calcola partendo dalla covarianza delle variabili :

$$\rho(X, Y) = \frac{COV(XY)}{\sigma(X)\sigma(Y)}$$

Dove $COV(XY)$ indica la covarianza tra le due variabili e $\sigma(X)$ e $\sigma(Y)$ indicano la deviazione standard rispettivamente per X e per Y.

Spearman I criteri che permettono di sfruttare questo coefficiente si fermano ai primi due step dell'indice di Pearson, il coefficiente di Spearman si può infatti definire come un caso specifico di indice di Pearson in cui si trasformano i dati in ranghi prima di calcolare il

¹³Un dato che si discosta di molto rispetto agli altri

¹⁴Test per verificare la normalità

coefficiente.

$$\rho_s = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2} \sqrt{\sum_i (s_i - \bar{s})^2}}$$

In applicazioni pratiche si utilizza una formula semplificata: $\rho_s = 1 - \frac{6\sum_i D_i^2}{N(N^2-1)}$ dove D_i rappresenta la differenza tra i ranghi $r_i - s_i$ delle due variabili confrontate.

Interpretazione degli indici:

- Se l'indice $\rho > 0$ si definiscono X e Y come variabili correlate positivamente.
- Se l'indice $\rho = 0$ non c'è correlazione tra le due variabili.
- Se l'indice $\rho < 0$ si definiscono X e Y come variabili correlate negativamente.

3.5 Conclusione

In questo capitolo è stato descritto l'approccio utilizzato per accedere, collezionare e analizzare i dati. L'analisi del funzionamento della piattaforma Crunchbase (Sezione 3.1.1), ha permesso di sviluppare un algoritmo di collezione dei dati (Sezione 3.1.1). Mediante l'accesso diretto ai dati di Crunchbase (Sezione 3.1.2) è stato reso più rapido il calcolo delle scorte e dei flussi.

Da un analisi dei dati si determina che buona parte degli utenti ha almeno 1 titolo di studio ed ha avuto più di due lavori. Queste proprietà sono fondamentali affinché attraverso l'assunzione fatta per stabilirne la nazionalità si possa procedere con l'analisi ed il confronto con dati veri. L'analisi dei dati, ottenuti attraverso l'Academic Research Access, è stata svolta mediante librerie Python (Sezione 3.4) che hanno permesso di manipolare i dati (Sezioni 3.2), realizzare i grafici e calcolare le varie correlazioni (Sezioni 3.3).

Capitolo 4

Analisi

Questo capitolo illustra i dati collezionati e i risultati ottenuti tramite il processo di analisi proposto per studiare la migrazione altamente specializzata a partire da fonti di dati non convenzionali. La prima parte del capitolo descrive il dataset collezionato attraverso Academic Research Access e propone lo studio dell'utenza Crunchbase. In seguito, il capitolo descrive i risultati delle analisi svolte sulle scorte (Sezione 4.3) e sui flussi (Sezione 4.4). In entrambe i casi, viene prima analizzato l'insieme di dati collezionato per determinare le zone più rappresentate. Per le scorte vengono analizzati alcuni casi di studio specifici, le scorte in Italia e gli emigrati italiani, le scorte in Gran Bretagna e gli emigrati britannici ed infine le scorte di nazionalità Europea in nord America e le scorte di nazionalità nord americana in Europa. I flussi vengono analizzati nella loro totalità e su diversi piani di aggregazione per zone geografiche, con due casi di studio dedicati all'Italia e alla Gran Bretagna.

Per ogni analisi affrontata è stato calcolato:

- Pearson: Correlazione di Pearson calcolata sugli insiemi forniti.
- Spearman: Indice di Spearman calcolato sugli insiemi forniti.

- PearsonLog: Correlazione di Pearson calcolata sugli insiemi forniti in scala logaritmica.
- PearsonLog(Risorsa): Correlazione di Pearson calcolata sugli insiemi forniti, e l'insieme “Risorsa” è in scala logaritmica.

Il campo ”Risorsa” può avere quattro valori ovvero CB per Crunchbase, ESTAT per Eurostat, UN per United Nations e Official per l'unione di UN e Eurostat.

4.1 Confronto tra metodi di collezione

Le scorse dal 2010 al 2020 ottenute attraverso l'impiego del Custom Query Builder (CQB) - Sezione 3.1.1 - e i dati ufficiali ottenuti attraverso l'Academic Research Access (Sezione 3.1.2) vengono confrontati nella Figura 4.1.1. Le nazionalità per cui i dati sono stati confrontati sono Austria, Belgio, Francia, Germania, Grecia, Irlanda ed Italia. Vengono calcolate anche le correlazioni. Osserviamo sia dal grafico che dalle correlazioni che le scorse ottenute tramite il CQB sono molto diverse da quelle dell'accesso accademico, con valori molto minori tramite CQB. Quindi il resto delle analisi vengono eseguito sui dati da Academic Research Access.

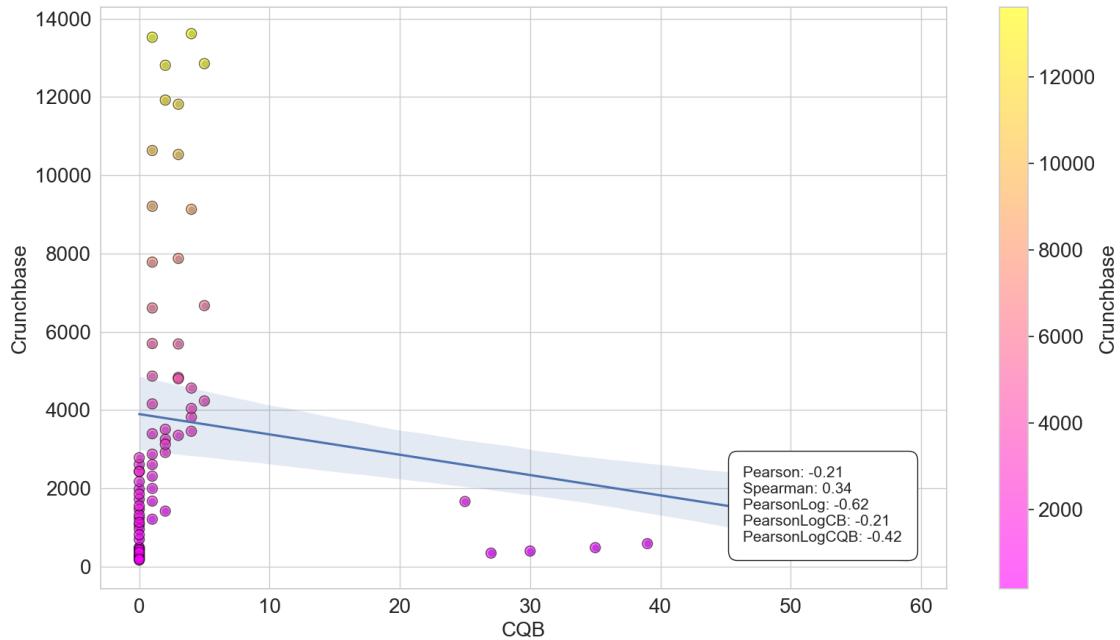


Figura 4.1.1: Confronto delle scorte ottenute dal Custom Query Builder e dall’Academic Research Access per i paesi selezionati.

4.2 Descrizione del dataset collezionato

L’intero dataset generato attraverso l’Academic Research Access contiene 1.288.646 profili utente. Viene effettuata un’analisi per determinare gli stati per cui Crunchbase è più rappresentato. La prima fase di analisi viene effettuata attraverso una mappa di calore dell’utenza di Crunchbase per visualizzare quali sono le zone geografiche per cui la piattaforma è più utilizzata (Figure 4.2.1). La seconda fase prevede la creazione di un grafico che indica l’utenza nel tempo per i primi dieci stati per numero di utenti (Figura 4.2.2). Entrambi i grafici sono stati realizzati usando i dati relativi a tutta l’utenza Crunchbase altamente qualificata per periodo compreso tra il 2010 ed il 2020.

La Figura 4.2.1 mostra l’utenza annuale di Crunchbase (dal 2010 al 2020) in base alla nazionalità (inferita seguendo la definizione in Sezione 3.2). Si noti che le nazionalità sono state aggregate per sub continenti. La crescita dell’utenza si osserva per tutti i paesi, in tutti



Figura 4.2.1: Utenza annuale di Crunchbase nel periodo 2010-2020 aggregata per zone geografiche

i dieci anni osservati. Tuttavia, si nota che l’incremento nel numero di utenti è fortemente dipendente dalle zone geografiche. L’America del nord è la zona con più utenti, seguita, seppur con larga differenza, dal nord e ovest Europa, e dal sud e ovest Asia.

La Figura 4.2.2 mostra l’utenza Crunchbase dei dieci stati con più utenti sulla base della nazionalità. Per tutti e dieci i paesi si osserva una linea di valori monotona crescente, che indica che nel tempo la piattaforma è sempre più usata.

I dati relativi a scorte e flussi estratti durante la fase di preprocessing (Sezione 3.2) sono stati analizzati per comprenderne la cardinalità. Il numero di scorte collezionate attraverso l’accesso accademico a Crunchbase è di 26.432. Ogni scorta è rappresentata da una tripla nazionalità-statoScorte-anno.

I flussi, invece, hanno una cardinalità di 6716 combinazioni rappresentate da una tripla origine-destinazione-anno. Inoltre, ogni combinazione ha un valore per i residenti ed uno per i cittadini.

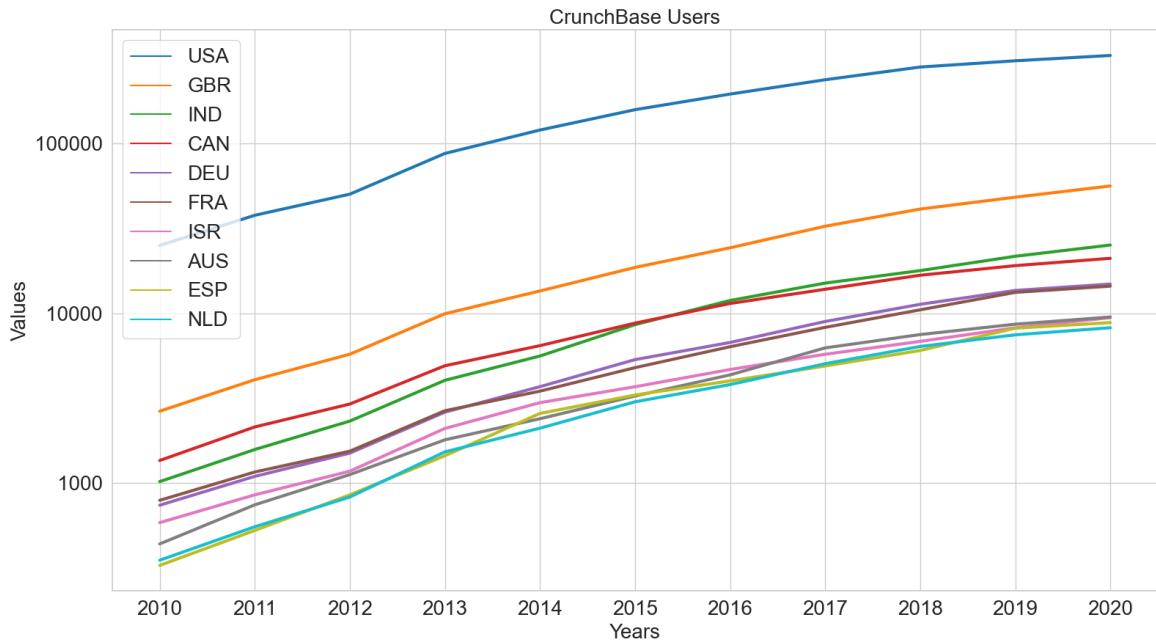


Figura 4.2.2: Trend utenza annuale di Crunchbase dal 2010 al 2020 per primi dieci paesi per numero di utenti totali al 2020.

4.3 Analisi delle scorte

In questa sezione vengono analizzate le scorte di migranti di Crunchbase. Durante la fase iniziale, abbiamo analizzato le scorte lungo tutto l’arco temporale coperto dai dati (dal 2010 al 2020). Le scorte di Crunchbase sono state confrontate, mediante il calcolo della correlazione di Pearson e di Spearman, con le scorte in UN per gli anni 2010, 2015 e 2020. Per la visualizzazione dei risultati sono stati realizzati grafici di dispersione.

Per approfondire l’analisi, sono stati realizzati tre casi di studio. I primi due studi riguardano gli emigrati e immigrati italiani e britannici rispettivamente. Infine, il terzo caso analizzato si concentra sugli stati europei e nord americani.

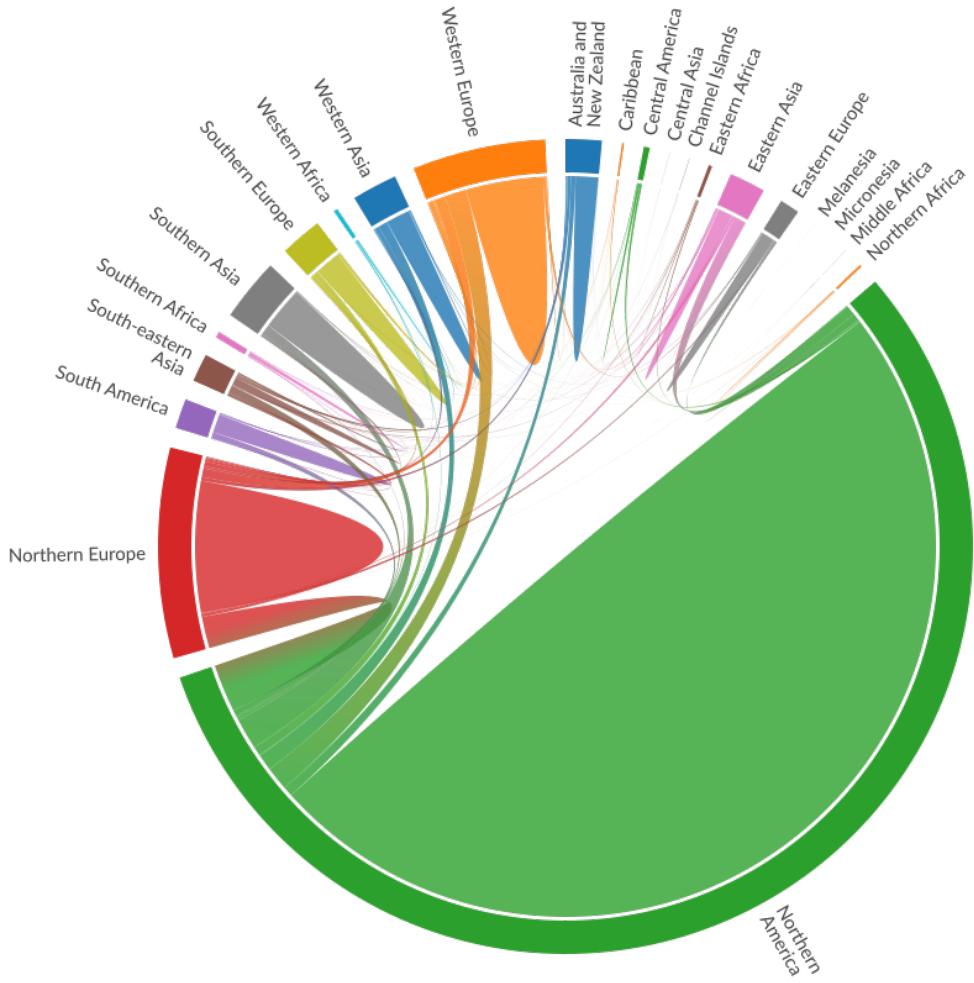


Figura 4.3.1: Scorte di migranti ottenute da Crunchbase sommate per il periodo 2010-2020.

4.3.1 Dati di Crunchbase

Le scorte migratorie in Figura 4.3.1 sono in accordo con l’analisi dell’utenza Crunchbase effettuata in Sezione 4.2. Ogni sezione del perimetro rappresenta una zona geografica, ogni arco rappresenta un numero di scorte di quella nazionalità rilevate nell’altro paese. Gli archi hanno dimensione differente in base alla dimensione delle scorte rilevate. La maggior parte degli utenti è rappresentato da nativi nord-americani residenti in nord America (non migranti). Osservando le singole zone, gli emigrati nord americani risiedono nell’ovest e nel nord

Europa, e nel sud e est Asia. Inoltre, il nord America sembra accogliere il maggior numero di immigrati provenienti da tutte le zone, seppur in diversa misura. Gli utenti del nord Europa e Europa dell'ovest sono rispettivamente il secondo e terzo gruppo più rappresentato. In entrambi i casi, così come per le altre zone, la maggior parte degli utenti è rappresentata da nativi. Nonostante ciò, parte degli utenti del nord Europa sembra essere emigrata in ovest Europa, e viceversa.

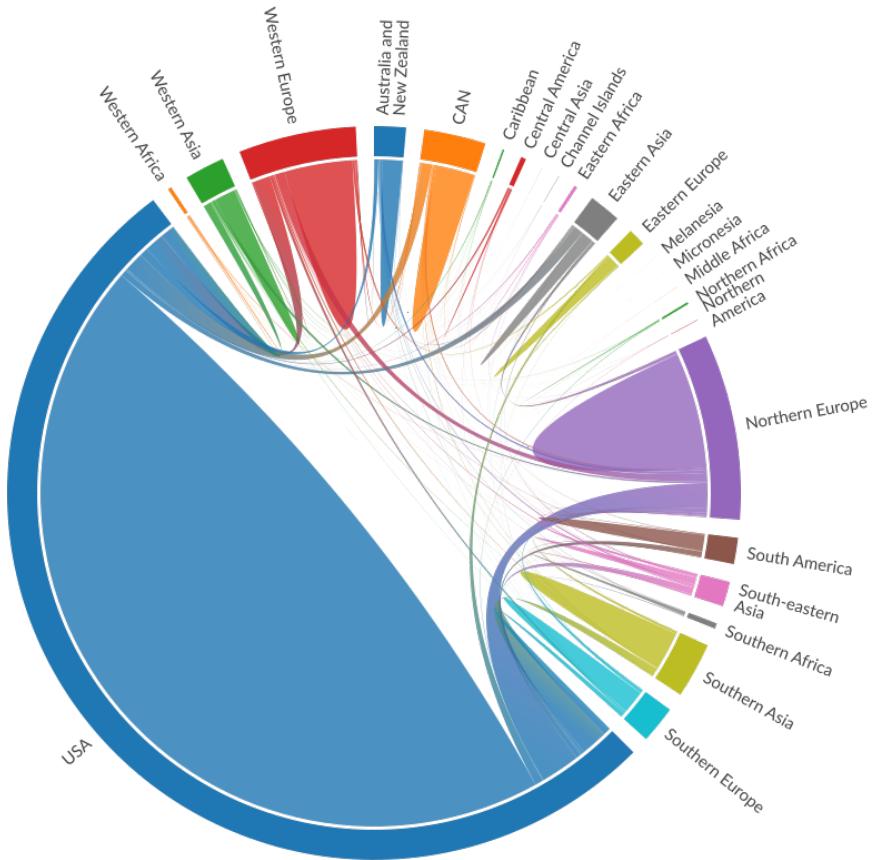


Figura 4.3.2: Scorte migratorie aggregate per zone più Stati Uniti d’America e Canada. Le scorte sono sommate per il periodo dal 2010 al 2020.

La Figura 4.3.2 rappresenta le scorte migratorie, sommate dal 2010 al 2020, considerando Stati Uniti d’America e Canada come zone a sé stanti. Ogni sezione del perimetro del grafico

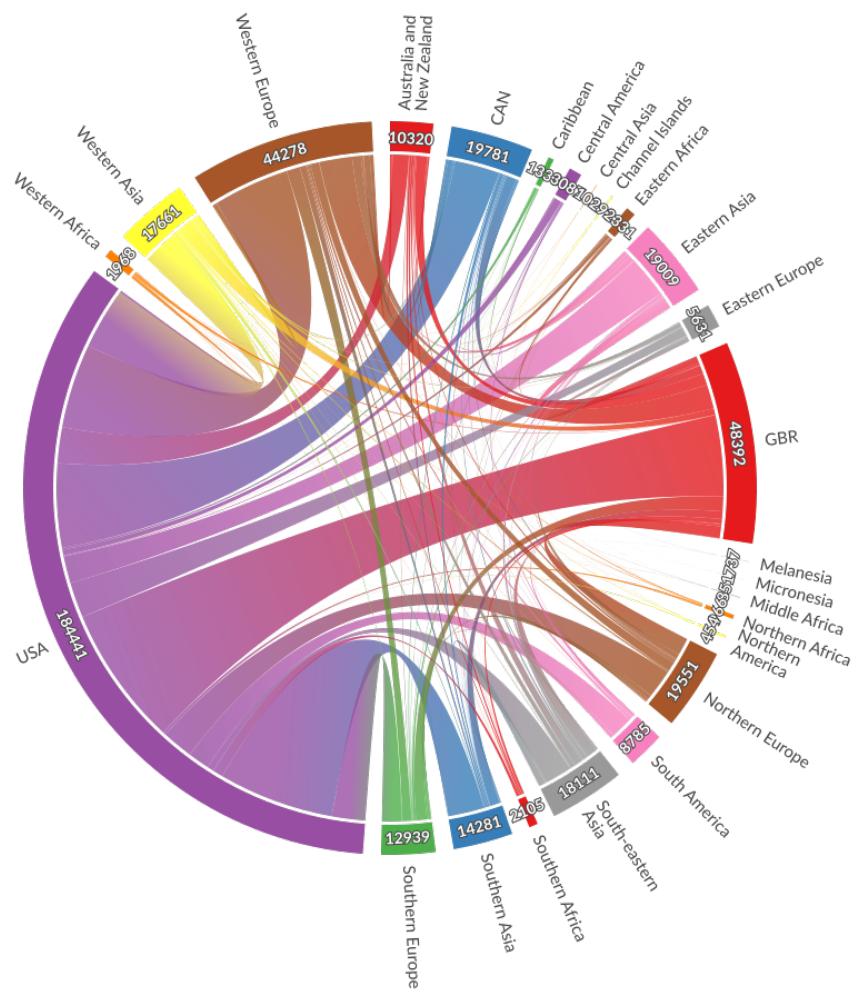


Figura 4.3.3: Crunchbase scorte con Stati Uniti, Canada and Gran Bretagna separati dall'aggregazione.

rappresenta uno degli stati (USA o CAN) o la zona geografica. Sebbene il grafico sia molto simile a quello in Figura 4.3.1, sottolinea la differenza nel numero di immigrati/emigrati tra Stati Uniti d'America e Canada e tutte le altre zone.

La Figura 4.3.3 rappresenta tutte le scorte migratorie internazionali considerando Stati Uniti d'America, Canada e Gran Bretagna come zone a sé stanti, sommate dal 2010 al 2020. Gli archi hanno dimensione differente in base alla dimensione delle scorte rilevate su Crunchbase. A differenza delle Figure 4.3.2 e 4.3.1 i dati relativi a persone non migranti (e.g. Stati Uniti a Stati Uniti) non vengono considerati in questo grafico. Dalla figura si può notare la forte presenza di statunitensi in tutte le altre zone. Lo scambio di migranti più evidente è presente tra Stati Uniti e Gran Bretagna, a seguire ci sono in ordine il sud Asia, l'ovest dell'Europa e il Canada. L'Asia in tutte le sue coordinate presenta valori valori di scorte comparabili a quelli del nord Europa e del Canada.

4.3.2 Confronto Crunchbase con UN

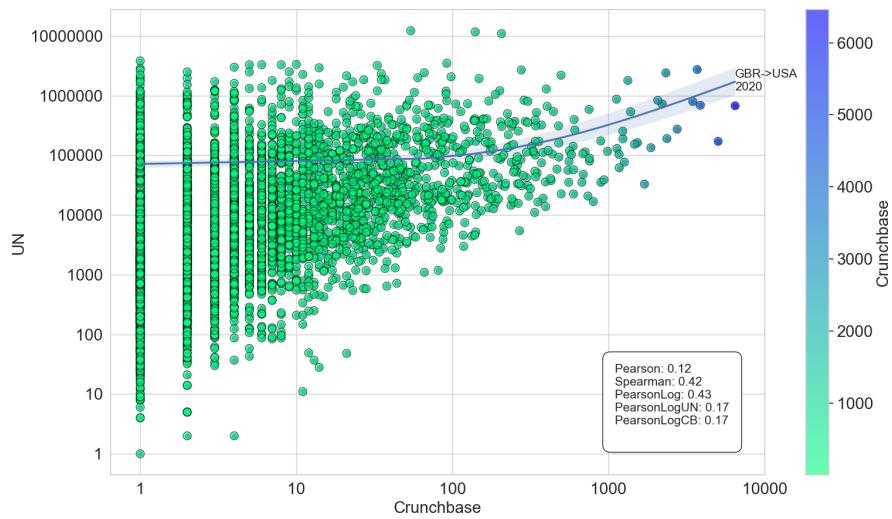


Figura 4.3.4: Confronto tra scorte di migranti in Crunchbase e UN. I dati sono per gli anni 2010, 2015 e 2020.

Il confronto delle scorte di Crunchbase con quelle di UN comporta la selezione delle combinazioni comuni di paesi. Inoltre, come descritto in precedenza (Sezione 2.1), UN dispone di dati solo di dati quinquennali, nel nostro caso abbiamo selezionato gli anni 2010, 2015 e 2020. Questo duplice processo di selezione (coppie comuni e anni) riduce le combinazioni a 5295 (-21137). Il grafico in Figura 4.3.4 mostra la correlazione tra le scorte di Crunchbase e di UN, utilizzando una scala logaritmica per entrambi gli assi. Sull'asse x si hanno i dati Crunchbase, sull'asse y i dati UN. Il valore massimo ottenuto per le scorte Crunchbase è rappresentato accanto al punto relativo al suo valore, con la forma *Nazionalità->Paese delle scorte Anno*. Nel riguardo in basso a destra sono presenti le varie correlazioni calcolate. La correlazione di Pearson è di 0.12, mentre l'indice di Spearman è di 0.42. Il coefficiente di Spearman ci indica che sia presente una correlazione debole tra le scorte in Crunchbase ed UN. Se si trasformano i dati delle scorte (UN e Crunchbase) in scala logaritmica per il calcolo della correlazione di Pearson, il valore è di 0.43 confermando il risultato ottenuto con il coefficiente di Spearman.

4.3.3 Caso di studio: Italia

In questo studio vengono analizzate le scorte di migranti di nazionalità italiana nel mondo e le scorte di migranti in Italia. Lo studio è svolto per gli anni 2010, 2015 e 2020. La Figura 4.3.5 mostra il confronto delle scorte di emigranti di nazionalità italiana tra UN e Crunchbase per gli anni 2010, 2015 e 2020. A destra la barra che indica i valori relativi ai vari colori, e il riquadro contenente le correlazioni calcolate. In basso il riquadro contenente le varie correlazioni calcolate. Come per il confronto precedente (Sezione 4.3.2), il valore massimo ottenuto per le scorte Crunchbase è rappresentato accanto al punto relativo al suo valore, con la forma *Nazionalità->Paese delle scorte Anno*. La maggior parte delle scorte di italiani

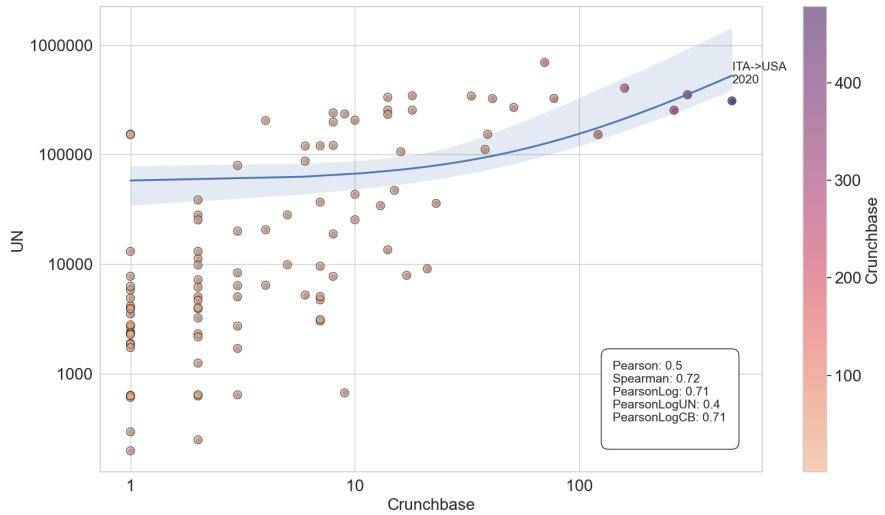


Figura 4.3.5: Emigrati italiani nel mondo, confronto per gli anni 2010, 2015 e 2020 tra Crunchbase e UN.

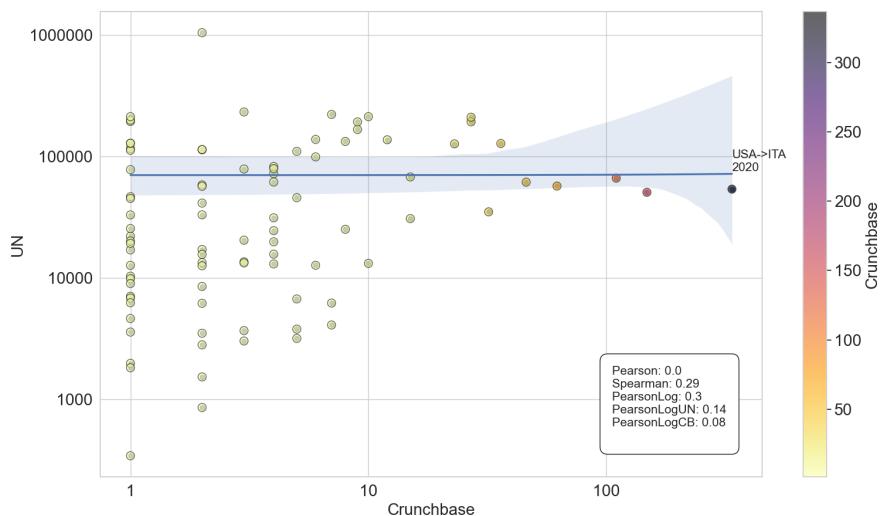


Figura 4.3.6: Immigrati in Italia, confronto per gli anni 2010, 2015 e 2020 tra Crunchbase e UN.

altamente qualificati per Crunchbase risiede negli Stati Uniti. La correlazione di Pearson è 0,5 e l'indice di Spearman 0.72. La correlazione di Pearson, si egualia all'indice di Spearman (0.71) se le scorte in ingresso (UN e Crunchbase) sono poste in scala logaritmica. Questo ci indica una forte correlazione tra i dati UN e Crunchbase per le scorte di italiani fuori sede. Inoltre, il grafico in Figura 4.3.6 mostra il confronto tra le scorte di immigrati in Italia da Crunchbase e UN. La correlazione di Pearson in questo confronto ha valore zero, cioè indica che non c'è correlazione. La correlazione di Pearson vede un incremento positivo fino a 0.3, se poniamo le scorte delle fonti (UN e Crunchbase) in scala logaritmica. L'indice di Spearman invece ha un valore di 0.29. Sebbene i valori dei coefficienti incrementino, non è sufficiente per determinare una correlazione debole.

4.3.4 Caso di studio: Gran Bretagna

Le scorte relative a cittadini di nazionalità britannica emigrati, in Figura 4.3.7, mostrano il valore massimo per gli Stati Uniti(circa 6.000). Entrambi gli indici di Pearson e Spearman mostrano correlazioni positive, di 0.45 e 0.71, rispettivamente. La correlazione di Pearson calcolata con entrambe le variabili in scala logaritmica restituisce valori intorno a 0.73, denotando un buon livello di correlazione.

Per quanto riguarda le scorte migratorie della Gran Bretagna, il grafico in Figura 4.3.8 mostra una correlazione di Pearson di 0.13 e dei valori più alti rispetto a quelli osservati per l'Italia (Figura 4.3.6). L'indice di Spearman ha un valore di 0.51, quasi identico a quello di Pearson (0.52) quando le scorte sono poste in scala logaritmica. Le scorte più numerose in Gran Bretagna nel 2020 erano di nazionalità statunitense. Il valore massimo delle scorte di migranti in Gran Bretagna (5000) è tra i più altri osservati.

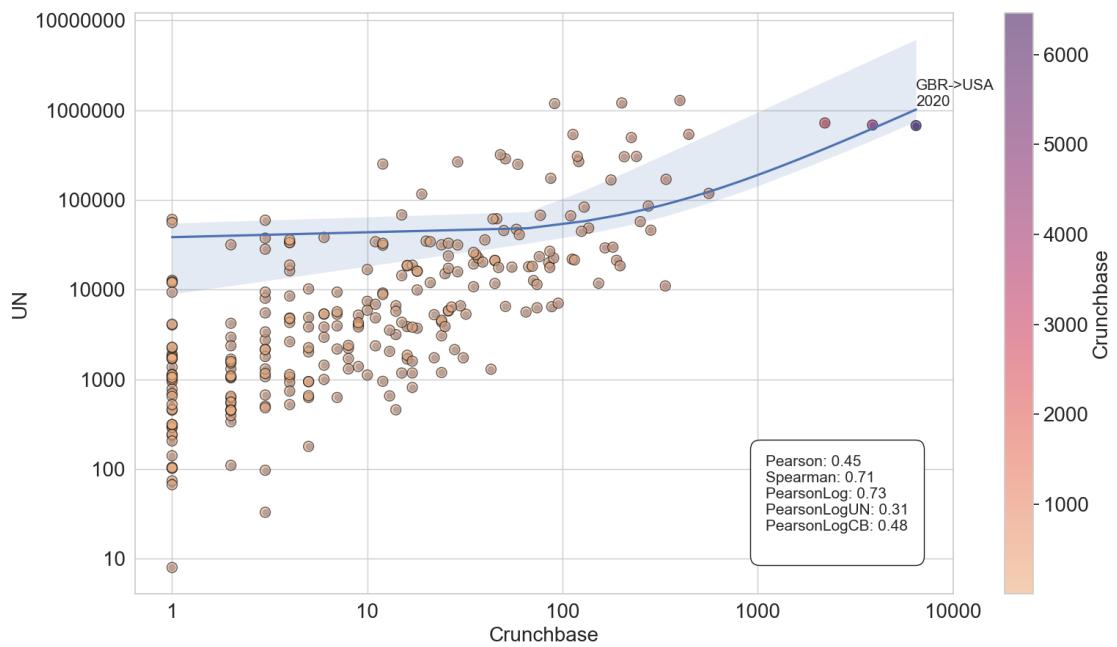


Figura 4.3.7: Emigrati britannici nel mondo, confronto per gli anni 2010, 2015 e 2020 tra Crunchbase e UN.

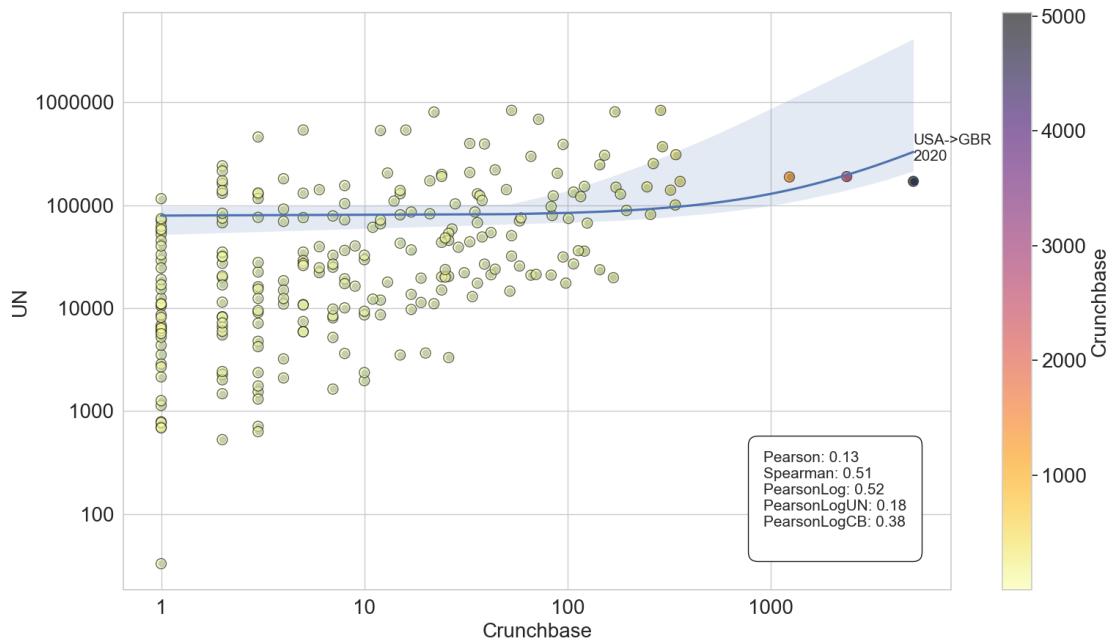


Figura 4.3.8: Immigrati in Gran Bretagna, confronto per gli anni 2010, 2015 e 2020 tra Crunchbase e UN.

4.3.5 Caso di studio: Europa e nord America

Viene effettuato un filtro sui dati Crunchbase e UN per il continente Europa e il subcontinente nord americano. Il grafico in Figura 4.3.9 analizza le scorte migratore di persone di nazionalità europea in nord America. Sull'asse delle x vengono posti i valori Crunchbase, sull'asse delle y i valori UN. In questo caso, la correlazione di Pearson è di 0.57 e l'indice di Spearman di 0.63. La correlazione di Pearson ponendo le scorte in scala logaritmica si egualgia all'indice di Spearman (0.62). Questi risultati suggeriscono quindi dei buoni livelli di correlazione. Il valore massimo di scorte di migranti altamente qualificati per Crunchbase, indicato nel grafico accanto al punto, è di britannici in nord America. In Figura 4.3.10 viene mostrato il confronto delle scorte di nazionalità nord americana in Europa. La correlazione di Pearson è di 0.75 e l'indice di Spearman di 0,65, indicando una correlazione positiva forte per Pearson. Il valore di scorte maggiori per Crunchbase si ottiene per i migranti statunitensi in Inghilterra nel 2020. Come già mostrato durante l'analisi dell'utenza di Crunchbase (Sezione 4.2), Europa e Nord America sono le zone più rappresentate.

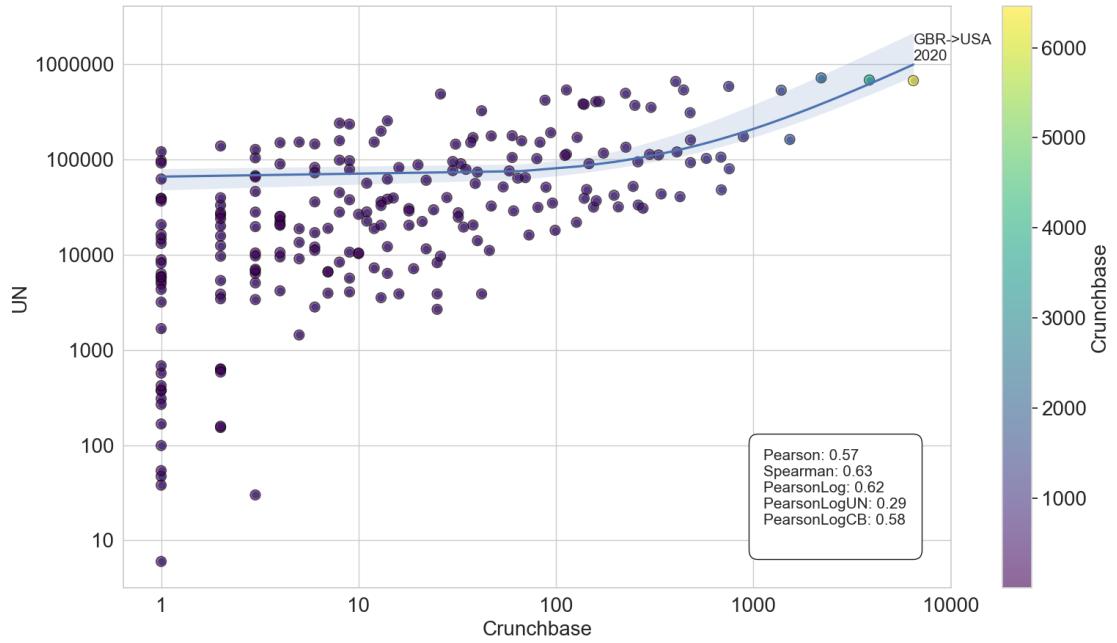


Figura 4.3.9: Scorte di migranti di nazionalità Europea in nord America (2010, 2015 e 2020).

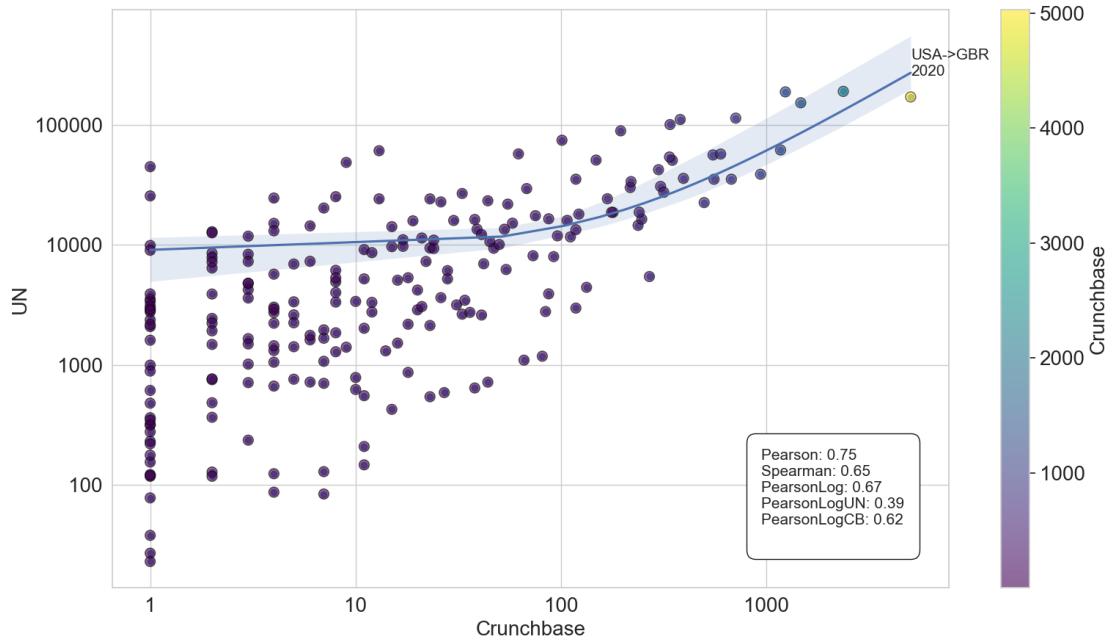


Figura 4.3.10: Scorte di migranti di nazionalità nord americana in Europa (2010, 2015 e 2020).

4.3.6 Discussione

Dato che a differenza di Crunchbase, UN dispone di sole scorte quinquennali, sono stati analizzati dati relativi al 2010, 2015 e 2020. Le scorte migratorie di Crunchbase presentano una correlazione debole con UN nel caso generale (Sezione 4.3.2). Il caso di studio dell'Italia (Sezione 4.3.3) mostra correlazioni discrete per gli emigrati dall'Italia, ma nessuna correlazione per gli immigrati in Italia. Per il caso di studio della Gran Bretagna (Sezione 4.3.4) sia le scorte di emigrati che di immigrati hanno correlazione discreta con UN. L'ultimo caso di studio si focalizza sugli stati del continente europeo e sul nord America (Sezione 4.3.5). La correlazione è discreta per gli emigrati europei in nord America. Inoltre, Pearson indica una forte correlazione per gli immigrati nord americani in Europa. I risultati ottenuti suggeriscono che a seconda dei casi e soggetti di studio, i dati di Crunchbase potrebbero essere impiegati per lo studio delle scorte di migranti.

4.4 Analisi dei flussi

In questa sezione vengono analizzati flussi di migranti di Crunchbase aggregati per zona geografica. Oltre allo studio dei flussi aggregati, per la Gran Bretagna è stato analizzato il periodo connesso alla Brexit (Sezione 4.4.1). I flussi di Crunchbase sono stati confrontati, mediante il calcolo della correlazione di Pearson e di Spearman, con i flussi in UN, Eurostat e con l'unione dei due, per gli anni dal 2010 al 2020. I risultati dello studio dei flussi aggregati per paesi lungo tutto l'arco temporale coperto dai dati (dal 2010 al 2020) sono stati visualizzati mediante chord diagram. Per il caso studio relativo all'Italia è stato impiegato il solo dataset dei flussi Eurostat. Per la Gran Bretagna è stata usata l'unione dei flussi in UN ed Eurostat. Come per le scorte (Sezione 4.3.1) la visualizzazione dei risultati avviene

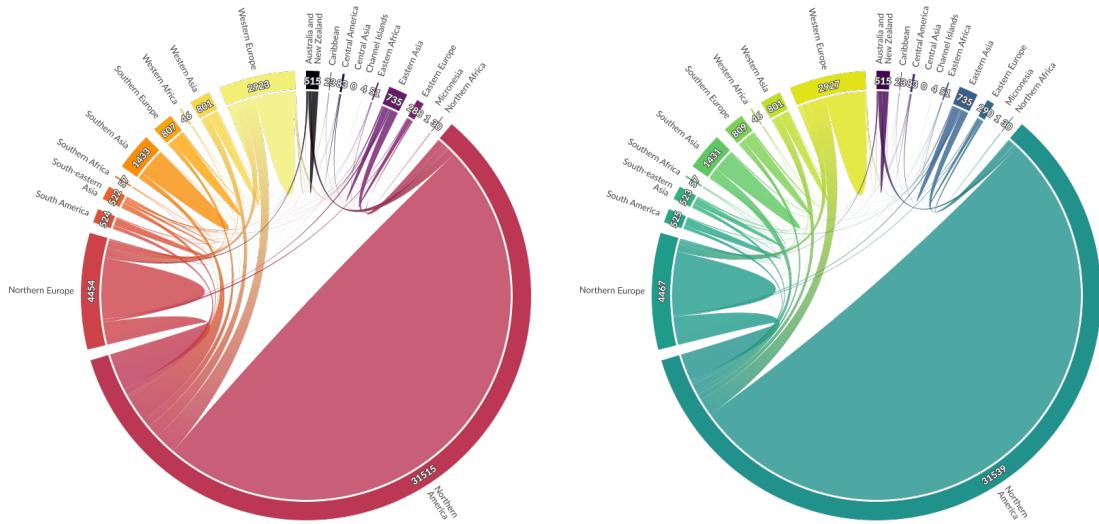


Figura 4.4.1: Flussi di migranti da Crunchbase aggregati per zone geografiche, dal 2010 al 2020. Il numero relativo alla zona indica il numero di flussi uscenti.

attraverso grafici di dispersione.

4.4.1 Dati di Crunchbase

I flussi in Figura 4.4.1a relativi a tutti i cittadini (utenti che sono residenti nello stato della nazionalità) presenti su Crunchbase sono maggiori per il nord America, il nord Europa e l’Europa dell’ovest. Buona parte dei flussi sono rappresentati da persone che si spostano in paesi della stessa zona, in particolare per il nord America.

La Figura 4.4.1b rappresenta i flussi migratori dei residenti osservabili dai dati di Crunchbase. I flussi dei residenti sono simili a quelli dei cittadini (Sezione 4.4.1). I flussi del nord America si presentano, anche in questo caso, i più numerosi.

Inoltre, i risultati mostrano che la distribuzione dei flussi migratori su Crunchbase è maggiore per il subcontinente nord americano e per l’Europa. Tuttavia, sono presenti flussi

relativi anche agli altri sub-continenti, sebbene siano minori.

Caso di studio: Gran Bretagna e la Brexit

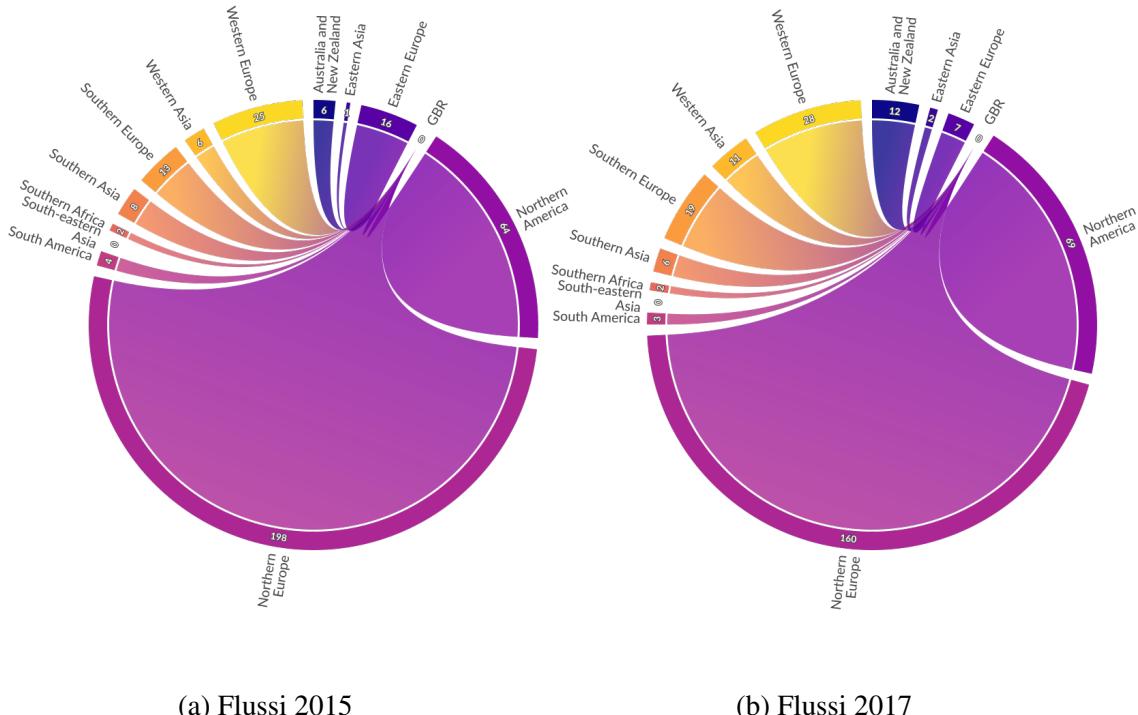


Figura 4.4.2: Flussi di immigrati in Gran Bretagna da Crunchbase, per gli anni 2015 e 2017.

La Figura 4.4.2a mostra i flussi di immigrati, in Crunchbase, che interessano la Gran Bretagna nell'anno 2015. Gli stati vengono aggregati per zona, il numero relativo alla zona indica il numero di flussi uscenti. I flussi di emigrati dalla Gran Bretagna sono stati ignorati lasciando spazio ai flussi di immigrati nel grafico. I flussi maggiori provengono dal nord Europa, il nord America e l'Europa dell'ovest. La Figura 4.4.2b mostra i flussi di migranti, da Crunchbase, verso la Gran Bretagna nel 2017. La Figura mostra leggere differenze nei flussi rispetto al 2015. I flussi provenienti dal nord Europa sono diminuiti di circa il 20% (da 198 a 160). Al contrario, sono aumentati i flussi verso l'Australia e la Nuova Zelanda (da 6 a 12), e verso il sud Europa (da 13 a 18). Il nord America ha avuto una variazione di sole 5 persone, da 64 a 69.

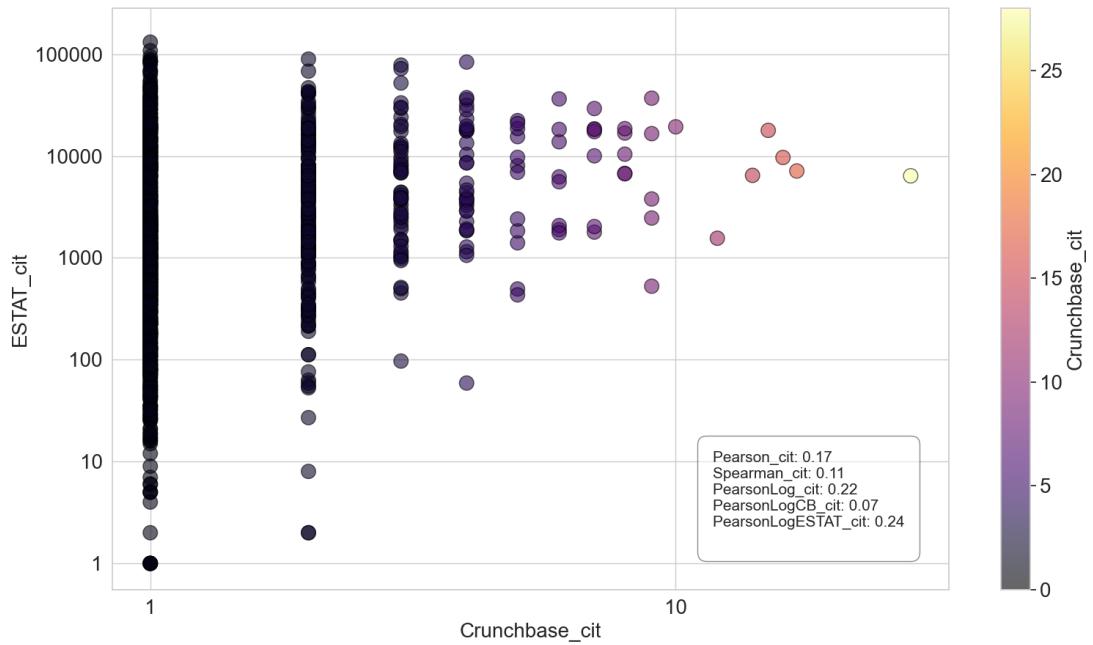
I risultati ottenuti mostrano che il nord America, il nord Europa e l'ovest Europa sono le zone geografiche più coinvolte nei flussi di migrazione di utenti altamente qualificati. Queste osservazioni sono in linea con quanto osservato nello studio dell'utenza Crunchbase (Sezione 4.2) e nell'analisi delle scorte (Sezione 4.3.1).

4.4.2 Confronto Crunchbase con UN ed Eurostat

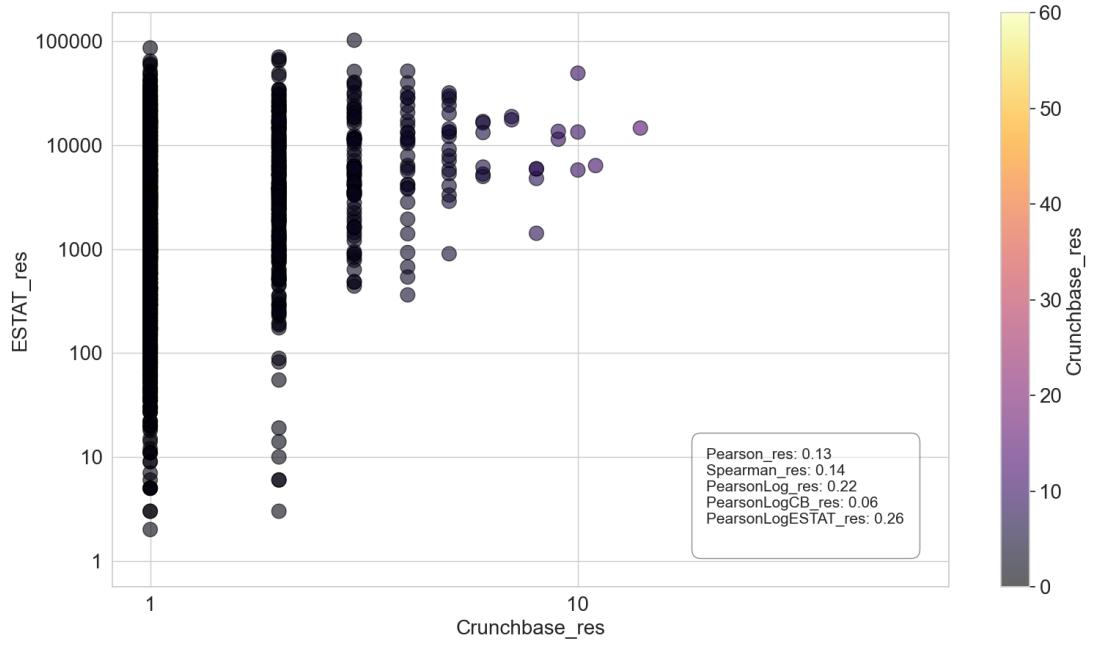
Come per i dati delle scorte, l'insieme dei dati relativi ai flussi di Crunchbase viene intersecato con quelli dei flussi presenti negli altri dataset. Per il confronto con UN, i flussi Crunchbase vedono una riduzione fino a 5.532 combinazioni (-1184). Per il confronto con Eurostat la diminuzione dei dati è nettamente superiore, arrivando a 2.208 combinazioni (-4508). Ogni combinazione fa riferimento sia al caso di cittadini che al caso di residenti.

Eurostat La Figura 4.4.3a mostra il confronto tra i flussi di Crunchbase ed di Eurostat per i cittadini che migrano dal paese di nazionalità. La correlazione di Pearson è di 0.17, simile all'indice di Spearman (0.11), indicando una scarsa relazione fra i due set di dati. Il grafico in Figura 4.4.3b presenta i flussi dei residenti emigranti in Crunchbase e in Eurostat. Sebbene ci sia un incremento del valore massimo di migranti rispetto al caso dei cittadini, le correlazioni sono molto basse (Pearson 0.13, Spearman 0.14).

UN La Figura 4.4.4a mostra i flussi dei cittadini che hanno migrato in confronto a UN. La Figura 4.4.4b mostra i flussi dei residenti. Entrambi i casi non presentano correlazione per Pearson e Spearman. Il valore massimo di flussi di migranti per Crunchbase in confronto a UN non supera le 40 unità. Il confronto dei dati relativi ai flussi di Crunchbase con i flussi di UN (Sezione 4.4.2) ed i flussi di Eurostat (Sezione 4.4.2) non ha mostrato correlazioni significative.

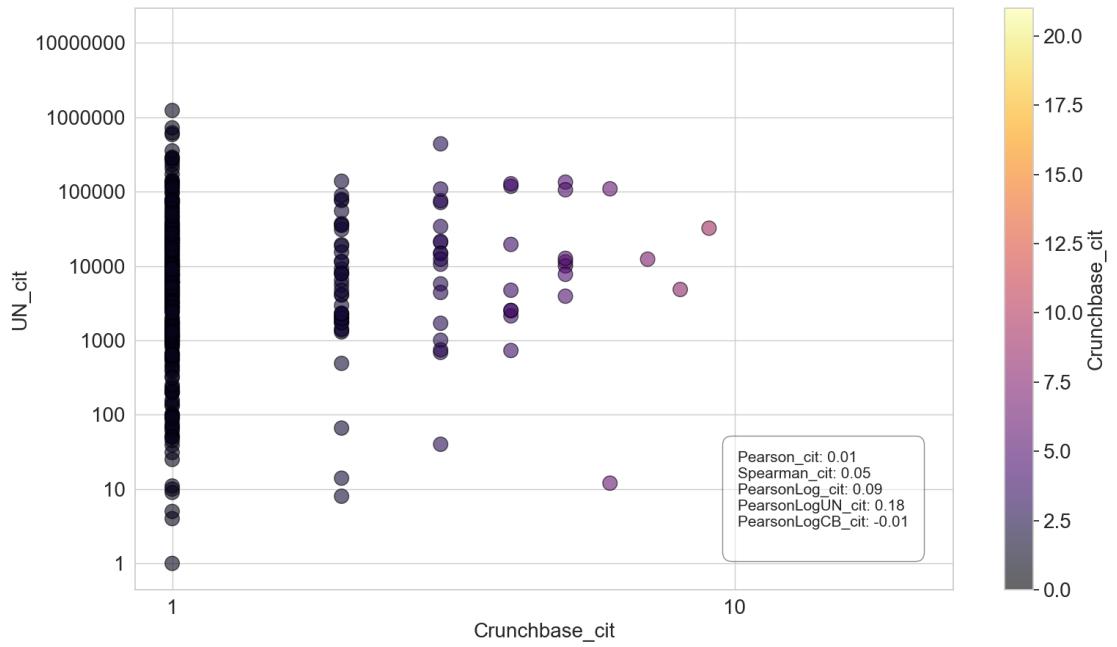


(a) Flussi cittadini

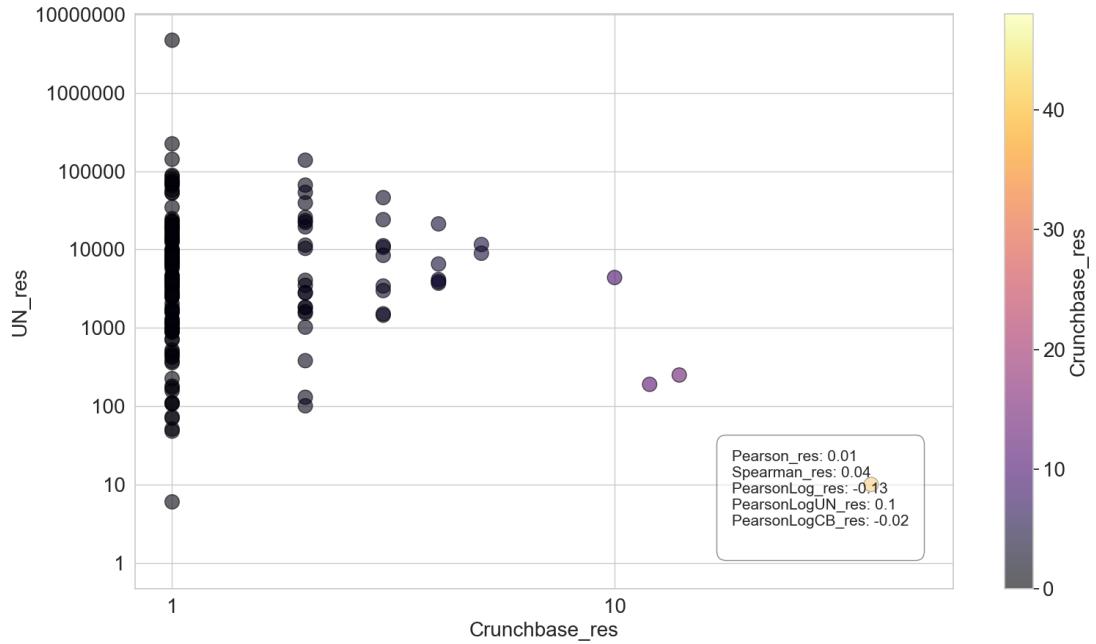


(b) Flussi residenti

Figura 4.4.3: Confronto dei flussi Crunchbase con i flussi Eurostat (2010 al 2020).



(a) Flussi cittadini



(b) Flussi residenti

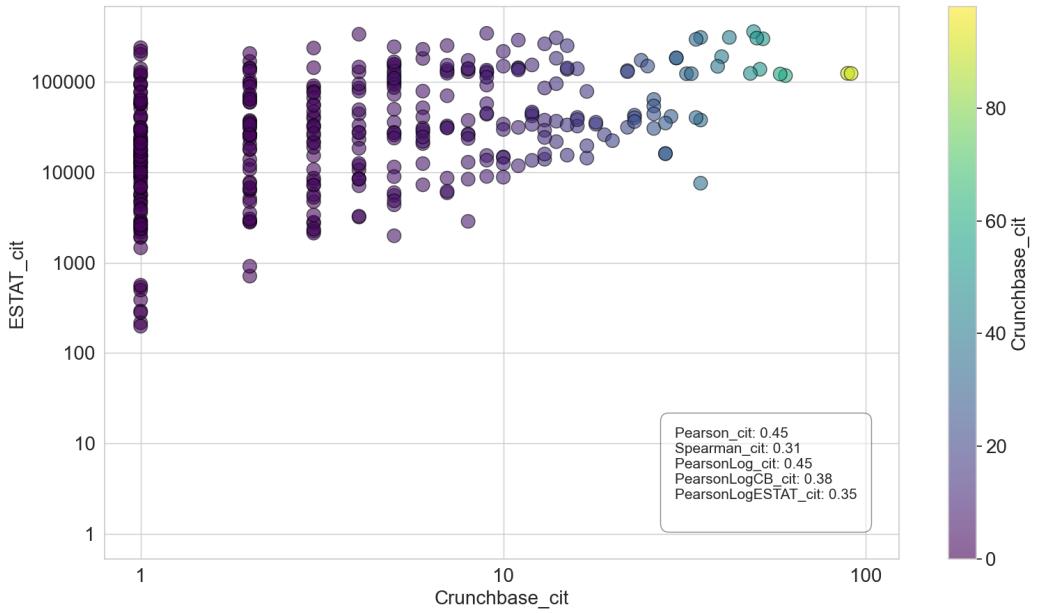
Figura 4.4.4: Confronto dei flussi in Crunchbase con UN dal 2010 al 2020.

4.4.3 Confronto UN ed Eurostat con stati aggregati

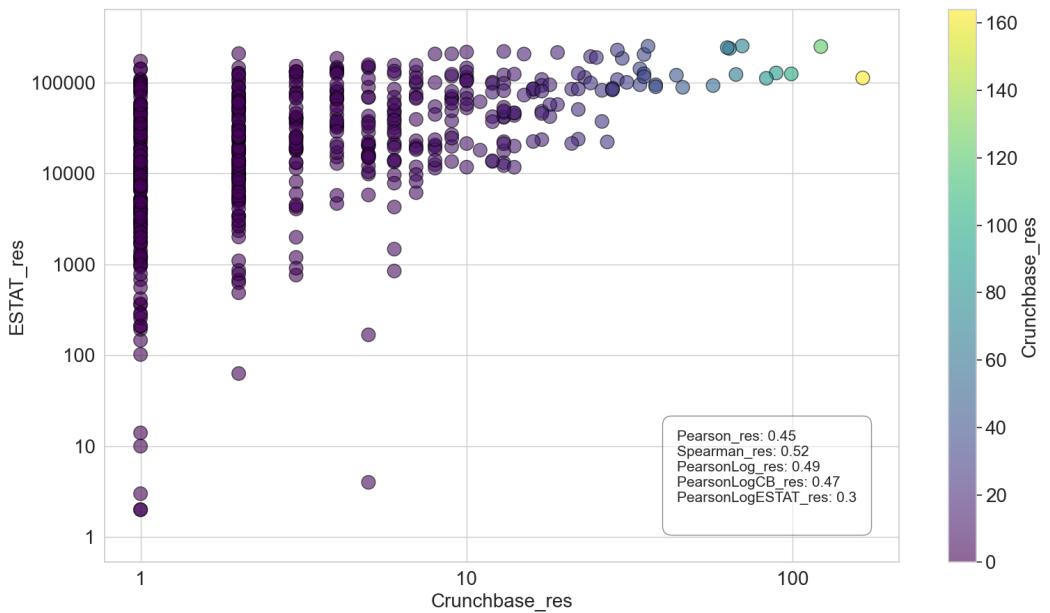
Eurostat con stati aggregati Analizziamo i flussi Crunchbase aggregando (comando) i flussi degli stati per zone geografiche. L’analisi di dati aggregati porta ad avere informazioni meno specifiche per gli stati in dettaglio, ma permette di comprendere quali sono le zone più interessate dai flussi in ingresso ed in uscita. Nella Figura 4.4.5a viene mostrato il grafico di confronto tra Eurostat e Crunchbase dei cittadini emigrati aggregati per zone geografiche, dal 2010 al 2020. La correlazione di Pearson è di 0.45, a differenza dell’indice di Spearman che indica un valore di 0.31. Il coefficiente di Pearson suggerisce che esista una correlazione debole tra i dati dei flussi aggregati di Crunchbase e UN.

La Figura 4.4.5b mostra il confronto tra Eurostat e Crunchbase dei residenti migrati aggregati per zone geografiche. La correlazione di Pearson (0.45) indica una correlazione debole, l’indice di Spearman (0.52) invece indica una correlazione discreta. La correlazione di Pearson calcolata con i flussi Crunchbase in scala logaritmica ottiene un valore superiore (0.47) alla correlazione normale, ma pur sempre basso.

UN con stati aggregati In Figura 4.4.6a vengono mostrati i flussi dei cittadini aggregati per zone geografiche in UN e Crunchbase. Il confronto presenta una correlazione di Pearson di 0.13 ed un indice di Spearman di 0.29. Il valore della correlazione di Pearson è di 0.34 utilizzando una scala logaritmica per i soli flussi dei cittadini in UN. Nella Figura 4.4.6b sono presenti i flussi dei residenti aggregati per zone geografiche in UN e Crunchbase. La correlazione di Pearson ha un valore di 0.04 e l’indice di Spearman è di 0.12. I flussi migratori di UN sono più significativi per il sudest asiatico, come visto in [12]. Come vediamo nell’analisi del flussi di Crunchbase (Sezione 4.4.1), il sudest asiatico non è rappresentato. Nel caso dei cittadini i risultati indicano una correlazione bassa ponendo i dati Crunchbase



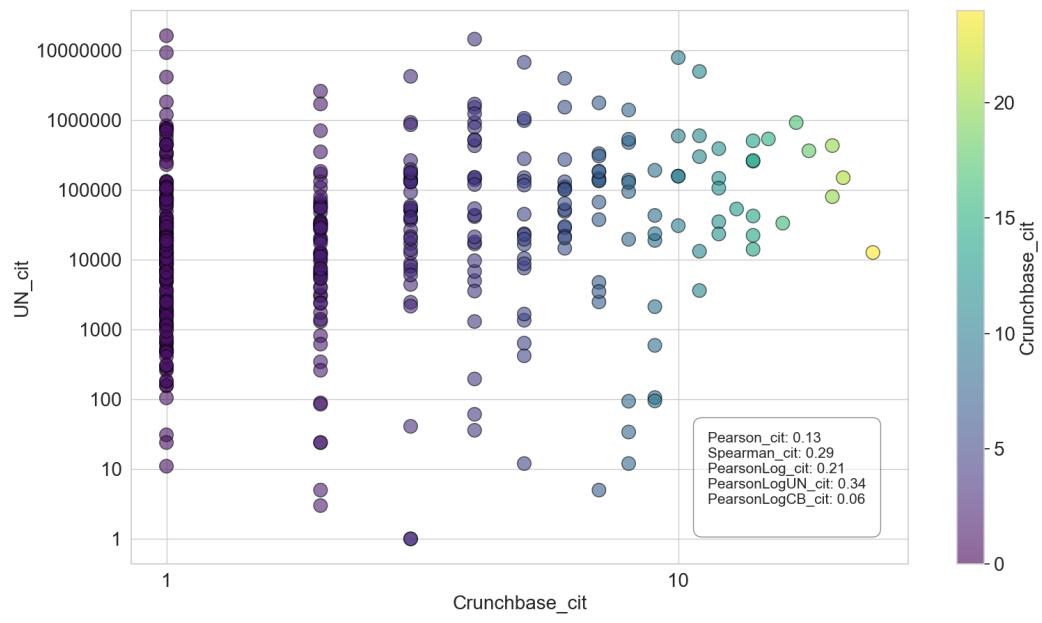
(a) Flussi cittadini



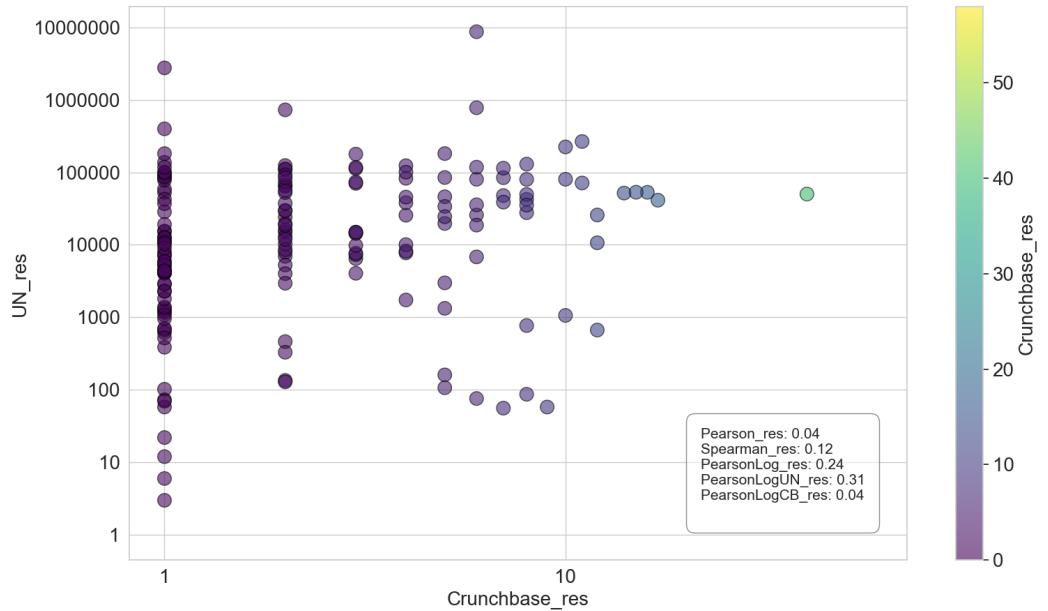
(b) Flussi residenti

Figura 4.4.5: Confronto dei flussi Crunchbase con i flussi in Eurostat aggregati per zone geografiche dal 2010 al 2020

in scala logaritmica. Di contro, i risultati ottenuti per il confronto dei residenti indicano che non ci sia correlazione.



(a) Flussi cittadini

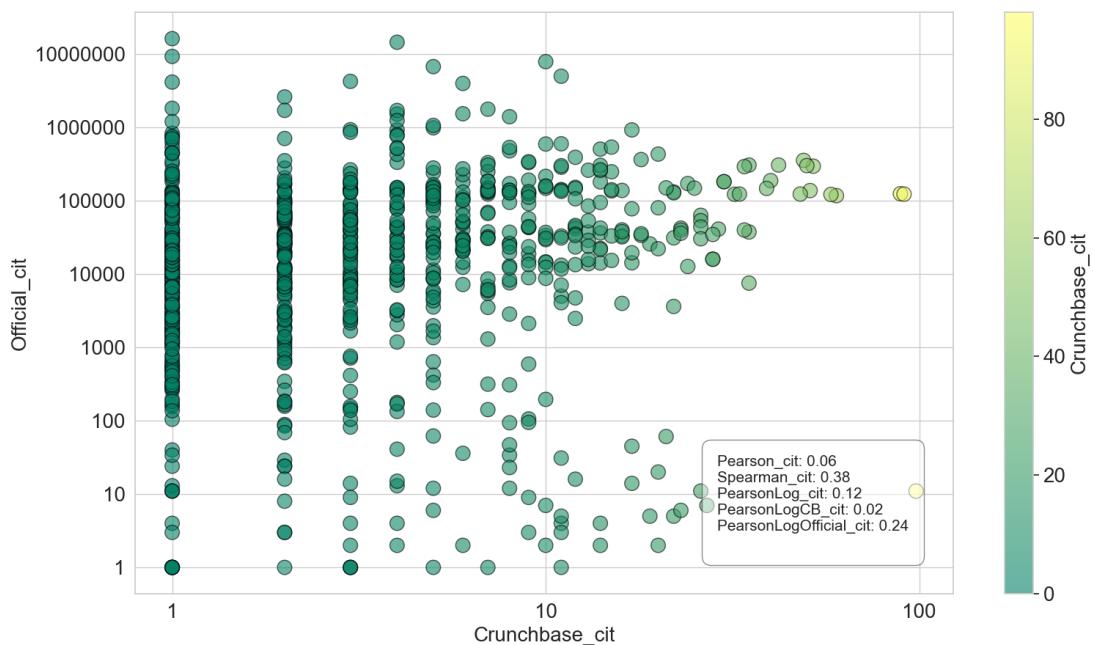


(b) Flussi residenti

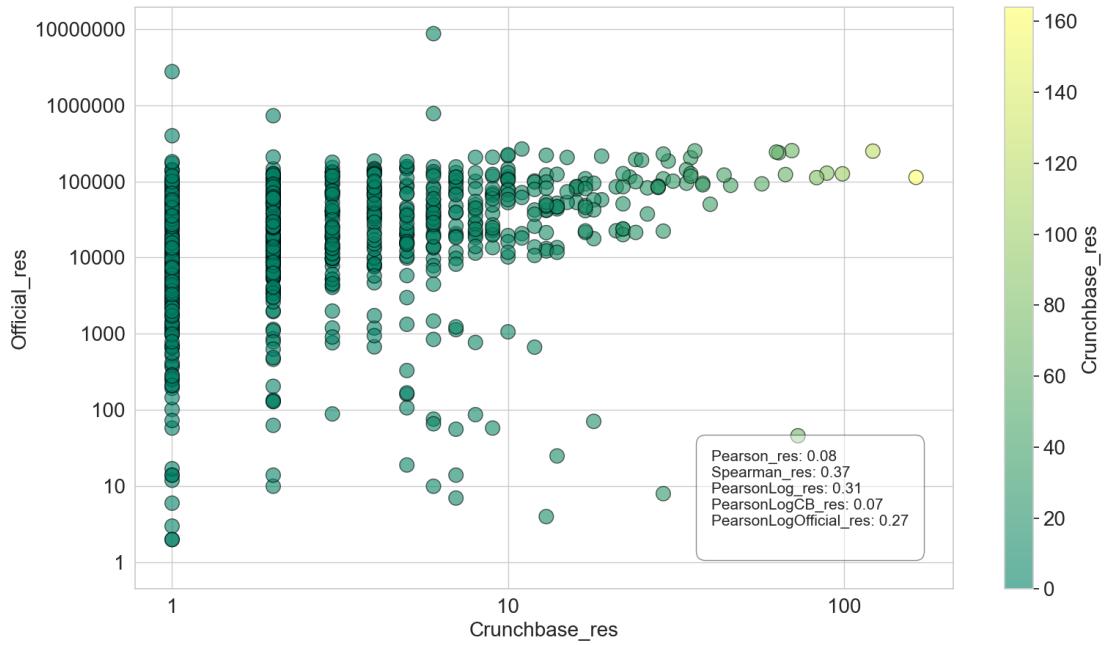
Figura 4.4.6: Confronto dei flussi Crunchbase con i flussi in UN aggregati per zone geografiche dal 2010 al 2020.

4.4.4 Flussi aggregati UN unito Eurostat

Per questo studio i dati di Crunchbase sono stati confrontati con l'unione di UN ed Eurostat, aggregando gli stati per zona geografica. Per l'unione dei due dataset (UN e Eurostat) è stato selezionato, per ogni paese, il flusso di maggior valore tra i due insiemi. Il periodo è come negli altri casi di studio dal 2010 al 2020. Il grafico nella Figura 4.4.7a rappresenta i cittadini. La correlazione di Pearson è di 0.06, mentre l'indice di Spearman è di 0.38. Con i dati ufficiali in scala logaritmica si ha una correlazione di Pearson di 0.24. I valori del coefficiente di Pearson indicano che non sia presente correlazione, a differenza dell'indice di Spearman che determina una correlazione debole. Nella Figura 4.4.7b riferita ai residenti si ottiene una correlazione di Pearson di 0.08 e un indice di Spearman di 0.37. Con i dati ufficiali in scala logaritmica si ha una correlazione di Pearson di 0.27. Come per i cittadini questo confronto ha correlazione debole per Spearman. Inoltre, i valori ottenuti sono inferiori rispetto all'analisi effettuata con i soli dati Eurostat aggregati (Sezione 4.4.3).



(a) Flussi cittadini $UN \cup Eurostat$



(b) Flussi residenti $UN \cup Eurostat$

Figura 4.4.7: Confronto dei flussi Crunchbase con i flussi presenti nell'unione di UN ed ESTAT (stati aggregati per zone geografiche dal 2010 al 2020).

4.4.5 Caso di studio: Italia

In questa sezione vengono analizzati i flussi d'immigrazione ed emigrazione, di residenti e cittadini, che interessano l'Italia. Il periodo dell'analisi è riferito al periodo dell'insieme di dati collezionati (dal 2010 al 2020), ed il dataset di confronto è quello Eurostat.

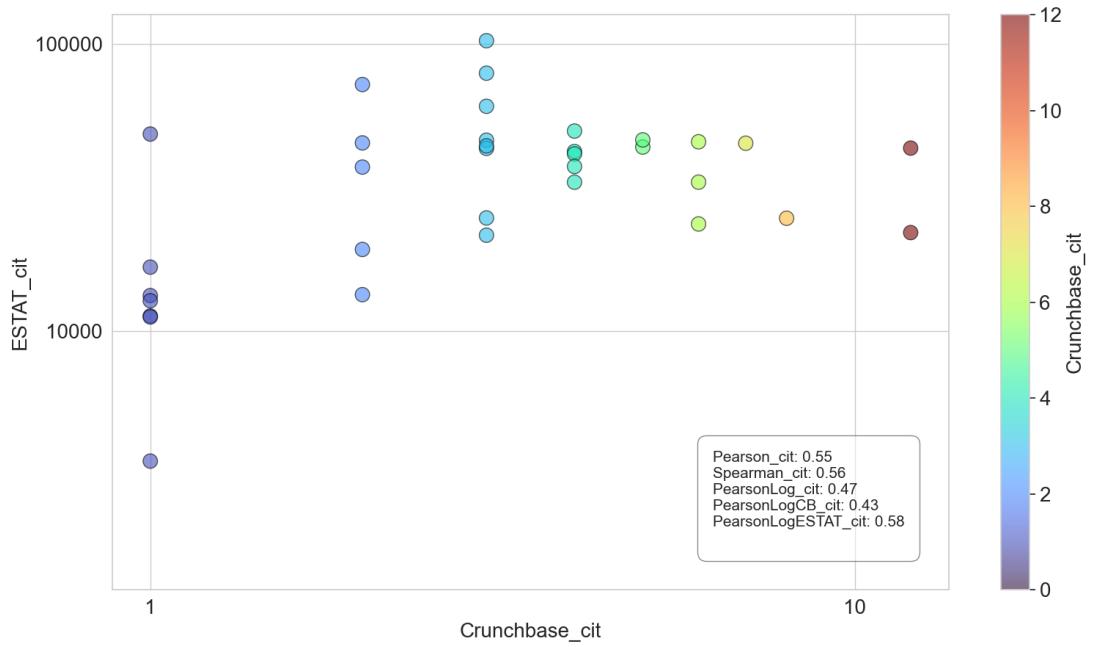
Flussi di emigranti dall'Italia Nella Figura 4.4.8a viene mostrato il confronto dei flussi di emigranti di nazionalità Italiana tra Crunchbase ed Eurostat. Gli stati oltre all'Italia vengono aggregati per zona geografica. Si nota una correlazione di Pearson di 0.55, l'indice di Spearman indica un valore di 0.56. Ponendo i soli dati Eurostat in scala logaritmica la correlazione di Pearson ottiene un valore di 0.58.

La Figura 4.4.8b mostra i flussi migratori dei residenti in Italia. La correlazione di Pearson è di 0.4, invece l'indice di Spearman è di 0.5. Interventi di trasformazione in scala

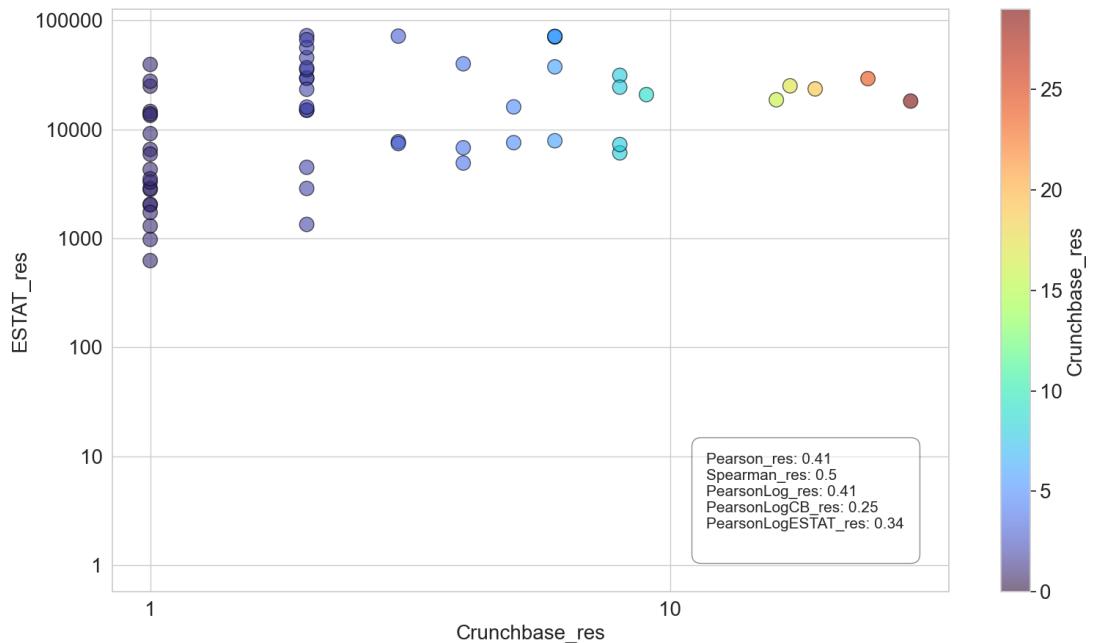
logaritmica di uno dei due insiemi (Eurostat e Crunchbase) portano a correlazioni di Pearson uguali o inferiori (0.41, 0.25, 0.34).

Entrambi i casi di confronto indicano una correlazione discreta tra i flussi di Crunchbase ed Eurostat per gli utenti emigrati dall'Italia. Si nota dalla barra laterale come il valore massimo di emigranti altamente qualificati dall'Italia nei flussi Crunchbase è intorno alle 25 unità.

Flussi d'immigrazione in Italia Nella Figura 4.4.9a viene mostrato il flusso di cittadini esteri che hanno migrato in Italia. La correlazione di Pearson è di 0.16, l'indice di Spearman è di 0.26. Ponendo in scala logaritmica i flussi dei cittadini in Eurostat la correlazione di Pearson incrementa fino a 0.24. Il valore massimo di flussi di cittadini esteri immigrati in Italia incontrato in Crunchbase è di 7 persone. In Figura 4.4.9b è presente il grafico dei flussi dei residenti all'estero che hanno migrato in Italia. Il valore della correlazione di Pearson è di 0.01, ed incrementa fino a 0.18 se poniamo i flussi di Eurostat in scala logaritmica. L'indice di Spearman ha invece un valore di 0.18. Il valore massimo di residenti all'estero immigrati in Italia in Crunchbase è di poco più di 25 persone. Inoltre, i valori dei coefficienti di Pearson e di Spearman per cittadini e residenti indicano che non vi è correlazione tra i dati Crunchbase ed i dati Eurostat per questo caso di studio.

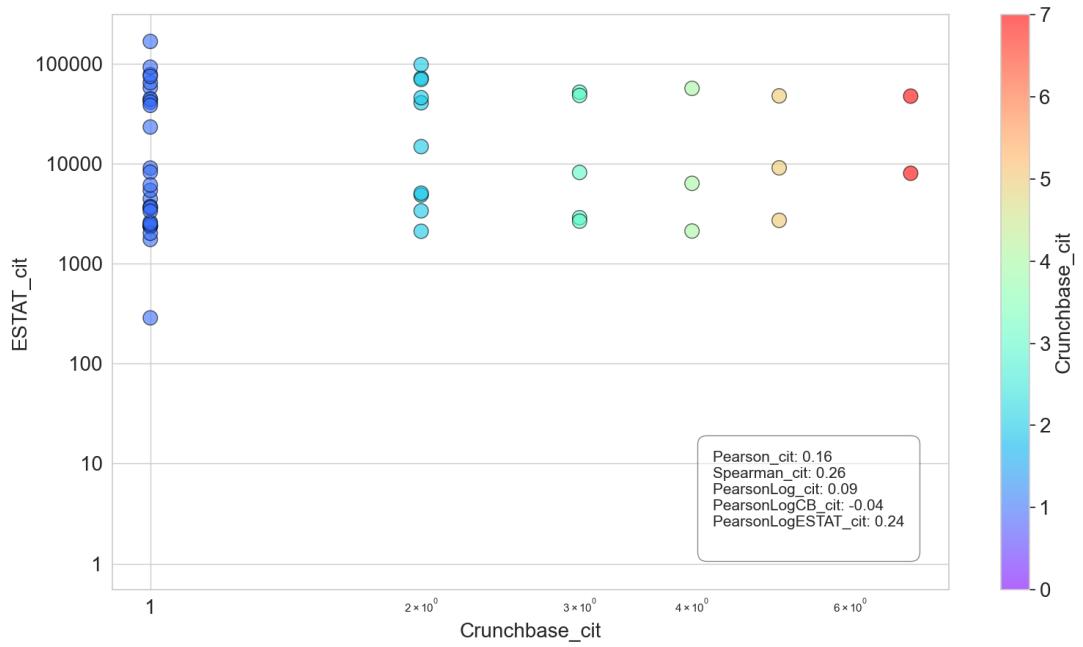


(a) Flussi cittadini

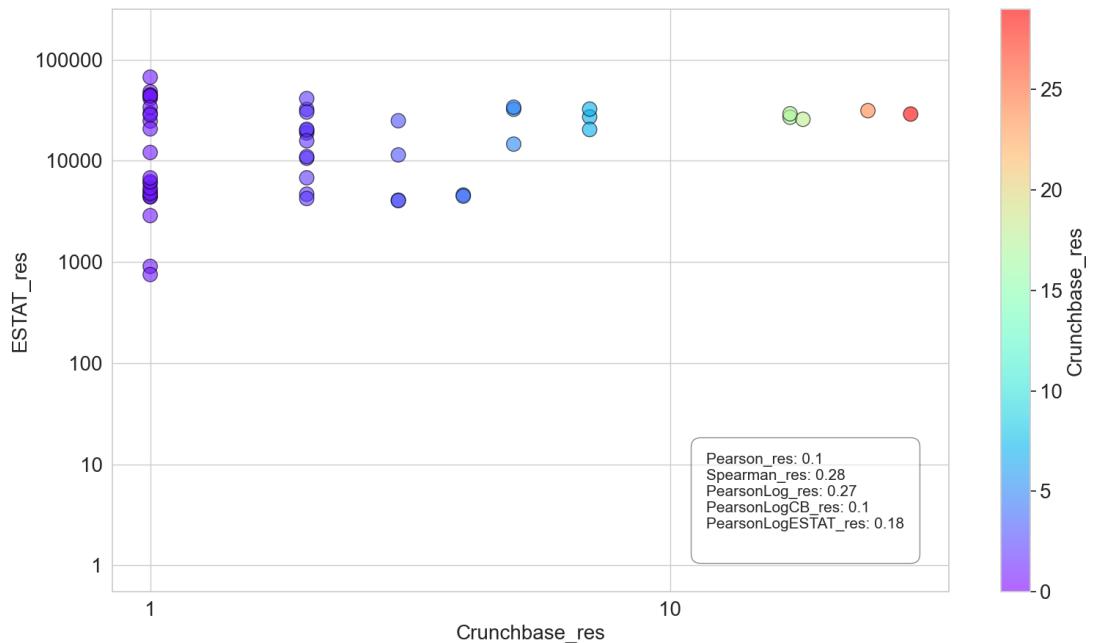


(b) Flussi residenti

Figura 4.4.8: Confronto dei flussi di emigranti, dall’Italia, di Crunchbase con Eurostat (gli altri stati sono aggregati per zone, periodo 2010 al 2020).



(a) Flussi cittadini



(b) Flussi residenti

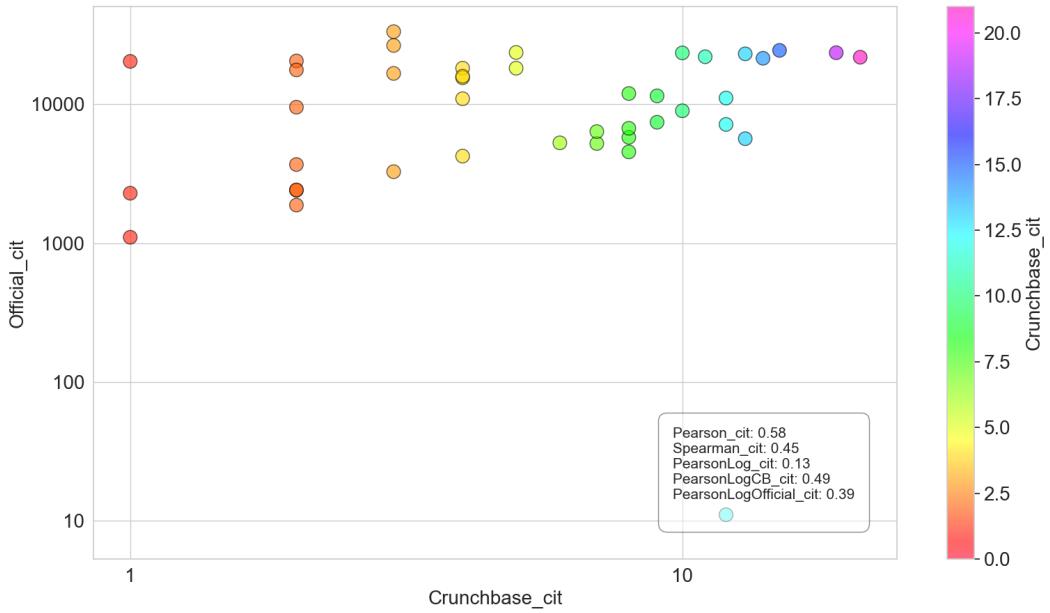
Figura 4.4.9: Confronto dei flussi di immigranti in Italia, tra Crunchbase ed Eurostat (gli altri stati sono aggregati per zone, periodo 2010 al 2020).

4.4.6 Caso di studio: Gran Bretagna

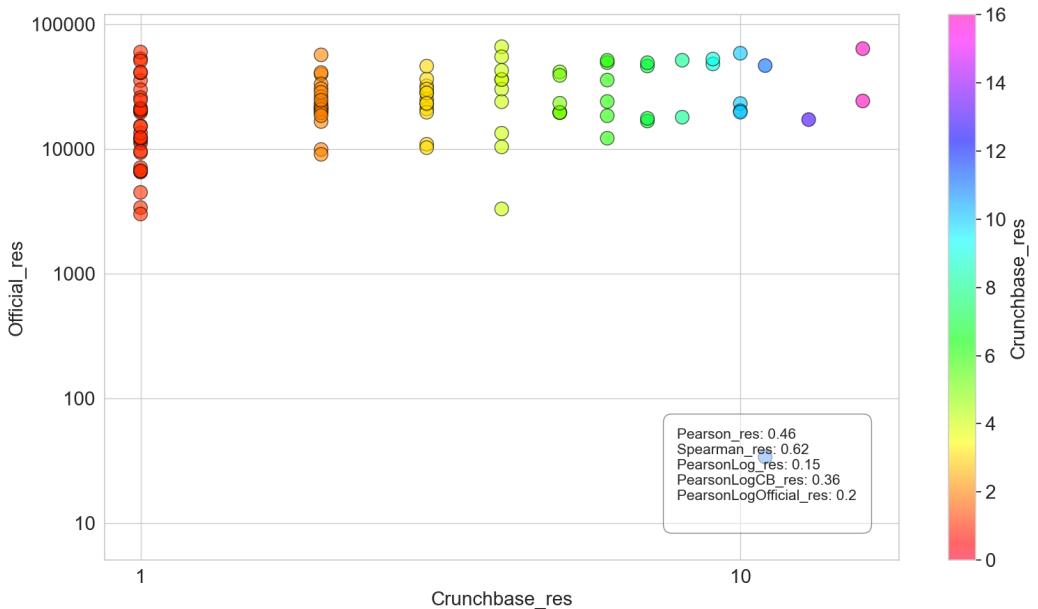
In questa sezione si analizzano in particolare i flussi di migranti da e verso la Gran Bretagna su tutto il periodo dei dati collezionati (dal 2010 al 2020). Il confronto viene fatto con i dati ottenuti da UN unito ad Eurostat.

Flussi di emigranti dalla Gran Bretagna In Figura 4.4.10a viene mostrato il confronto tra i flussi dei cittadini britannici che emigrano, tra Crunchbase e UN unito Eurostat. La correlazione di Pearson ottenuta è di 0.58, invece l'indice di Spearman ha un valore di 0.45. Ponendo gli insiemi in scale logaritmiche la correlazione di Pearson decrementa il suo valore in tutti i casi (0.13, 0.49, 0.39). Il grafico in Figura 4.4.10b mostra i flussi dei residenti in Gran Bretagna, che hanno emigrato. La correlazione di Pearson ha un valore di 0.46 come per il grafico dei cittadini in Figura 4.4.10a. L'indice di Spearman ha invece un valore di 0.62. Inoltre, utilizzare scale logaritmiche sui flussi (Ufficiali o di Crunchbase) porta a valori inferiori alla correlazione normale. Entrambi i coefficienti indicano una correlazione discreta tra i dati Crunchbase e UN unito Eurostat sia per nativi che residenti della Gran Bretagna che hanno emigrato.

Flussi d'immigrazione in Gran Bretagna Nel grafico in Figura 4.4.11a vengono confrontati i flussi di cittadini esteri che migrano in Gran Bretagna, tra Crunchbase e UN unito Eurostat. La correlazione di Pearson è di 0.1, invece l'indice di Spearman ha valore 0.24. Ponendo in scala logaritmica entrambe le fonti dei flussi (UN unito Eurostat e Crunchbase) si ha una correlazione di Pearson negativa di -0.12. Questi valori indicano che non sia presente correlazione tra i dati Crunchbase ed i dati di UN unito Eurostat, per i flussi di cittadini esteri immigrati in Gran Bretagna. La Figura 4.4.11b mostra i flussi degli immigranti in Gran Bretagna. La correlazione di Pearson è di 0.27 l'indice di Spearman è di 0.47. Utilizzando una



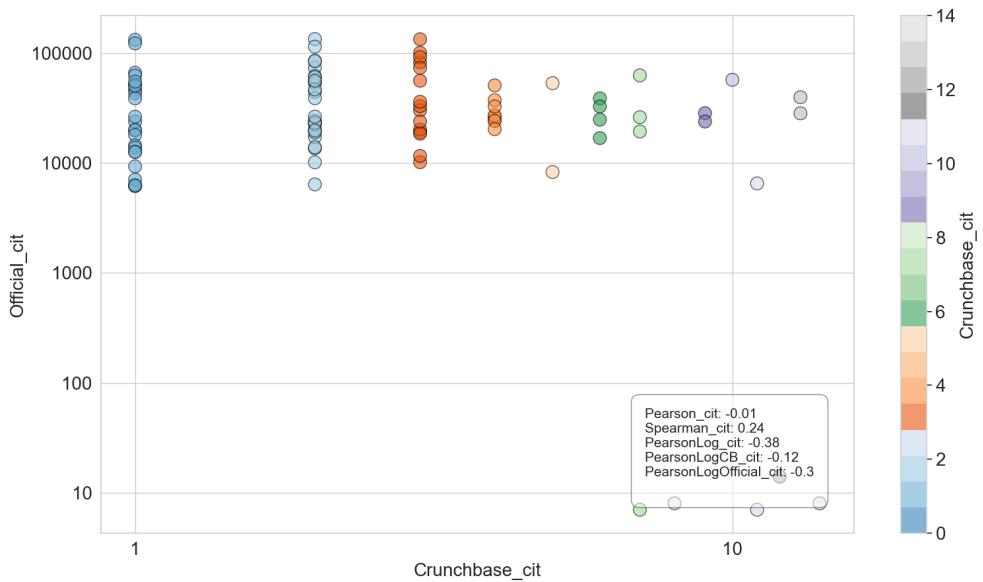
(a) Flussi cittadini



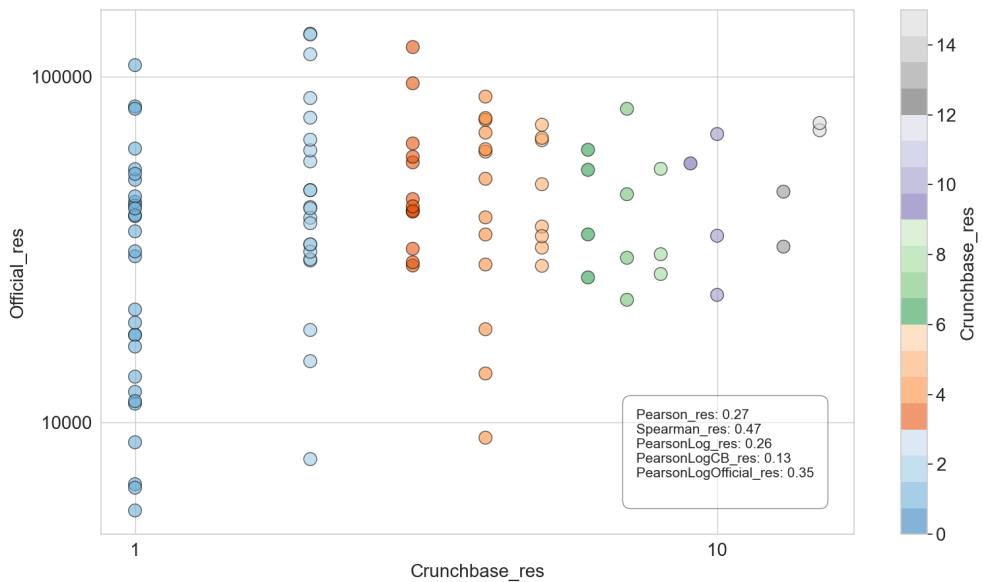
(b) Flussi residenti

Figura 4.4.10: Confronto dei flussi di emigranti britannici, tra Crunchbase e UN unito Eurostat (gli altri stati sono aggregati per zone geografiche, il periodo è 2010-2020).

scala logaritmica per i flussi dei dati ufficiali (UN unito ad Eurostat) si ha una correlazione di Pearson di 0.35. La correlazione di Spearman indica che ci sia una correlazione positiva discreta tra i dati confrontati.



(a) Flussi cittadini



(b) Flussi residenti

Figura 4.4.11: Confronto dei flussi di immigranti in Gran Bretagna, tra Crunchbase e UN unito Eurostat (gli altri stati sono aggregati per zone geografiche, il periodo è 2010-2020).

4.4.7 Discussione

I flussi dei dati collezionati attraverso Crunchbase sono significativi per l’Europa ed il Nord America come visto nella sezione 4.4.1. La correlazione con i flussi in Eurostat è debole, in alcuni casi anche discreta (Sezione 4.4.3) se aggregati, i flussi in Eurostat però non presentano dati per la Gran Bretagna [12]. La correlazione dei flussi con UN non è presente, come abbiamo visto nell’analisi (Sezioni 4.4.2, 4.4.3). L’unione dei flussi UN con i flussi Eurostat non ha correlazione per Pearson nell’analisi effettuata, ma per l’indice di Spearman questa è presente in forma debole (Sezione 4.4.4). Tra i casi di studio affrontati l’emigrazione di cittadini e residenti Italiani presenta una correlazione discreta (Sezione 4.4.5), con un numero massimo di migranti pari a 20. L’immigrazione verso l’Italia non presenta correlazione con i dati Eurostat (Sezione 4.4.5). Per la Gran Bretagna i flussi presentano una correlazione discreta per gli emigranti (Sezione 4.4.6). I flussi di immigranti in Gran Bretagna invece non presentano correlazione per i cittadini di altre nazioni. Per i residenti di altre nazioni che si spostano in Gran Bretagna la correlazione è discreta (Sezione 4.4.6).

Capitolo 5

Conclusione

L’obiettivo della tesi proposta è duplice. Da un lato, è proposta una metodologia per la collezione e l’analisi di dati di Crunchbase. Dall’altro, il lavoro mira a validare i dati di Crunchbase.

Vengono confrontati due metodi di raccolta dati, uno basato sul web scraping e l’altro tramite accesso Accademico all’API. In seguito, l’analisi viene effettuata in primo luogo sull’utenza, per determinare le zone geografiche più rappresentate. Successivamente il focus viene spostato sul confronto dei dati Crunchbase con i dati ufficiali per la validazione. Vengono confrontate le scorte di migranti con i dati UN per gli anni 2010, 2015 e 2020 con diversi casi di studio: l’Europa, il nord America, l’Italia e la Gran Bretagna. I flussi Crunchbase ed i flussi di UN ed Eurostat vengono confrontati per il periodo dal 2010 al 2020, sfruttando diversi livelli di aggregazione al fine di determinare le zone geografiche per cui i flussi Crunchbase sono più significativi. Come per le scorte, vengono dedicati due casi di studio specifici. Il primo riguarda i flussi di migranti in uscita e in entrata dall’Italia, mentre il secondo propone la stessa analisi per la Gran Bretagna.

Il confronto dei dati Crunchbase con i dati collezionati dal Custom Query Builder di-

mostra come lo Scraping pur essendo più immediato può avere delle limitazioni. L’analisi dell’utenza ha permesso di dedurre che Crunchbase è più rappresentato in nord America, nord Europa e ovest Europa. Le correlazioni ottenute analizzando le scorte Crunchbase con le scorte UN, in generale, presentano una correlazione debole. Tuttavia, si ottengono correlazioni forti quando si osservano gli emigrati di nazionalità italiana e britannica. Si ottengono correlazioni medio-alte anche quando si osserva il confronto delle scorte in Europa e in nord America. La validazione dei flussi ha determinato che in generale i flussi Crunchbase non hanno correlazione con i flussi in UN ed Eurostat. Tuttavia, osservando i flussi aggregati in sub continenti si notano correlazioni intorno a 0.5 con Eurostat. L’unione dei due dataset per il confronto dei flussi non apporta miglioramenti particolari alla correlazione con i dati Crunchbase. I casi di studio dedicati all’Italia ed alla Gran Bretagna vedono entrambi valori di correlazione compresi tra 0.5 e 0.6, ma solo per il caso degli emigranti.

I risultati ottenuti indicano che i dati Crunchbase potrebbero essere utilizzati con un certo grado di affidabilità, in studi migratori per gli stati del nord America, e del nord e dell’ovest Europa. In particolare, i casi di studio comuni delle scorte e dei flussi evidenziano una correlazione discreta, se non forte, per gli emigranti italiani e britannici.

Per lavori futuri, si potrebbe analizzare la relazione tra i dati Crunchbase e dati focalizzati sui migranti altamente qualificati, come quelli del Labor Force Survey. Inoltre, dato che Crunchbase detiene informazioni specifiche sugli stati federati degli Stati Uniti (es. Texas, California) si può analizzare la migrazione interna ad essi. Infine, la mobilità degli utenti di Crunchbase potrebbe essere osservata in relazione a eventi particolari (es. Crisi economica, COVID-19) per studiare se e come questa ne venga influenzata.

Ringraziamenti Un ringraziamento viene fatto in particolare a Crunchbase stessa che ha contribuito accogliendo ed approvando la nostra richiesta di accesso ai dati. Si ringraziano inoltre le relatrici Alina Sîrbu e Laura Pollacci per l’ispirazione e la dottrina impeccabile.

crunchbase

Bibliografia

- [1] Bridget Anderson and Scott Blinder. Who counts as a migrant? definitions and their consequences. *Briefing, The Migration Observatory at the University of Oxford*, 2011.
- [2] Martin Bell, Elin Charles-Edwards, Philipp Ueffing, John Stillwell, Marek Kupiszewski, and Dorota Kupiszewska. Internal migration and development: Comparing migration intensities around the world. *Population and Development Review*, 41(1):33–58, 2015.
- [3] Joshua E. Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125, 2012.
- [4] Manola Cherubini, Sebastiano Faro, and Mariasole Rinaldi. Glossario sull’asilo e la migrazione. *CNR Edizioni, Roma*, 2016.
- [5] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. Using crunchbase for economic and managerial research, 2017.
- [6] Joop de Beer, James Raymer, Rob van der Erf, and Leo van Wissen. Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *European Journal of Population / Revue européenne de Démographie*, 26(4):459–481, Nov 2010.

- [7] Eurostat. Eu - labour force survey microdata 1983-2019, release 2020, version 1, 2020.
- [8] Jane Falkingham, Corrado Giulietti, Jackline Wahba, and Chuhong Wang. The impact of brexit on international students' return intentions. *The Manchester School*, 89(2):139–171, 2021.
- [9] Jane Falkingham, Jackline Wahba, Corrado Giulietti, and Wang Chuhong. Survey of graduating international students, 2017-2018.
- [10] Heinz Fassmann. European migration: Historical overview and statistical problems. *Statistics and reality. Concepts and measurements of migration in Europe*, pages 21–44, 2009.
- [11] Marcu Florentina. Web data extraction with robot process automation. study on linkedin web scraping using uipath studio. *Annals of'Constantin Brancusi'University of Targu-Jiu. Engineering Series*, (1), 2020.
- [12] Diletta Goglia, Laura Pollacci, and Alina Sirbu. Dataset of Multi-aspect Integrated Migration Indicators, April 2022.
- [13] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, jun 2008.
- [14] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi,

- Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
- [15] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [16] Roberto Impicciatore and Nazareno Panichella. L'emigrazione dei laureati italiani. un'analisi delle caratteristiche individuali che favoriscono la mobilità internazionale. *Quaderni di Sociologia*, (86-LXV):31–53, 2021.
- [17] Hassène Kassar, Diaa Marzouk, Wagida A. Anwar, Chérifa Lakhoud, Kari Hemminki, and Meriem Khyatti. Emigration flows from north africa to europe. *European Journal of Public Health*, 24, 08 2014.
- [18] Sari Kerr, William Kerr, Caglar Ozden, and Christopher Parsons. High-skilled migration and agglomeration. *Annual Review of Economics*, 9:201–234, 08 2017.
- [19] Russell King. Theories and typologies of migration: An overview and a primer. *Wiley Brandt Series of Working Papers in International Migration and Ethnic Relations*, 12:1–43, 01 2012.
- [20] Rainer Münz. Migration, labor markets, and integration of migrants: An overview for europe. HWWI Policy Paper 3-6, Hamburg, 2007.
- [21] Giovanni Peri and Chad Sparber. Highly-educated immigrants and native occupational choice. *Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London, CReAM Discussion Paper Series*, 50, 01 2008.
- [22] Daniela Perrotta, Sarah C. Johnson, Tom Theile, André Grow, Helga de Valk, and Emilio Zagheni. Openness to migrate internationally for a job: Evidence from linkedin

data in europe. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):759–769, May 2022.

- [23] Andreas Pitsillidis, Yinglian Xie, Fang Yu, Martin Abadi, Geoffrey M. Voelker, and Stefan Savage. How to tell an airport from a home: Techniques and applications. Hotnets-IX, New York, NY, USA, 2010. Association for Computing Machinery.
- [24] Michel Poulain, Nicolas Perrin, and Ann Singleton. *THESIM: Towards harmonised European statistics on international migration*. Presses univ. de Louvain, 2006.
- [25] Carlos Rodríguez González, Ricardo Bustillo Mesanza, and Petr Mariel. The determinants of international student mobility flows: an empirical study on the erasmus programme. *Higher Education*, 62(4):413–430, Oct 2011.
- [26] Rsalmei. Rsalmei/alive-progress: A new kind of progress bar, with real-time throughput, eta, and very cool animations!
- [27] Spyratos S, Vespe M, Natale F, Ingmar W, Zagheni E, and Rango M. Migration data using social media: a european perspective, 2018.
- [28] Peter Samuels and Mollie Gilchrist. Pearson correlation. Technical report, 04 2014.
- [29] Mario Cesar Scheffer, Alex Jones Flores Cassenote, Aline Gil Alves Guilloux, and Mario Roberto Dal Poz. Internal migration of physicians who graduated in brazil between 1980 and 2014. *Human Resources for Health*, 16(1):21, May 2018.
- [30] Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 11(4):341–360, 2021.

- [31] Bogdan State, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. *Migration of Professionals to the U.S.*, pages 531–543. Springer International Publishing, Cham, 2014.
- [32] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
- [33] Frans Willekens, Douglas Massey, James Raymer, and Cris Beauchemin. International migration under the microscope. *Science*, 352(6288):897–899, 2016.
- [34] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14 Companion, page 439–444, New York, NY, USA, 2014. Association for Computing Machinery.
- [35] Jerrold H. Zar. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580, 1972.