POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea

Full-stack analysis su dati demografici e residenziali del Comune di Milano

Relatore
prof. Daniele Apiletti
Tutor aziendali
Nicolas Liguori
Luca Colombo

Candidato
Davide Fioriti

Anno Accademico 2023-2024

Sommario

Questa tesi presenta un'analisi demografica e residenziale approfondita per la città di Milano, sfruttando i dati aperti forniti da Open Data Milano. Il lavoro adotta un approccio full stack, partendo dall'integrazione dei dati tramite Talend Open Studio for Data Integration, per raccogliere e trasformare dataset relativi a nascite, decessi, residenti, costruzioni e informazioni anagrafiche e censuarie. Dopo aver costruito una pipeline di data integration e cleaning, vengono effettuate analisi esplorative e di correlazione per comprendere le tendenze demografiche storiche e estrarre contenuto informativo utile. In seguito, si implementano vari modelli predittivi, tra cui il Simple e il Double Exponential Smoothing, SARIMA e modelli per serie storiche demografiche come quelli di Malthus, Verhulst e Richards, per prevedere l'evoluzione futura della popolazione residente. Le informazioni estratte sono poi proposte in dashboard interattive su Power BI, progettate secondo i principi di percezione visiva Gestalt e le best practice di data visualization. Questo approccio consente di sintetizzare e comunicare efficacemente i risultati dell'analisi a un pubblico eterogeneo, fornendo strumenti decisionali che semplificano la comprensione dei dati considerati.

Premessa

L'elaborato presentato nasce nel contesto di un tirocinio extra-curricolare svolto presso *Exacon S.r.l.*, un'azienda con sede a Milano specializzata in consulenza, progettazione e sviluppo di soluzioni avanzate in ambito IT ed Engineering, con un focus su Big Data, Data Science, Data Engineering, Business Intelligence e Data Warehouse. L'attività di Exacon si colloca nell'ambito della gestione e valorizzazione dei dati aziendali, offrendo competenze che spaziano dall'analisi dei dati alla loro integrazione in architetture complesse. Durante il periodo di tirocinio, è stato possibile approfondire e applicare le conoscenze acquisite nel percorso universitario, con particolare riguardo alla gestione di database e all'analisi dei Big Data. Questo ha consentito di consolidare competenze pratiche in SQL e Python, linguaggi fondamentali nell'ecosistema dei dati. Inoltre, il tirocinio ha rappresentato un'occasione preziosa per approcciare nuove tecnologie non trattate durante gli studi, ma largamente adottate nel settore. Tra queste, PySpark, utilizzato per gestire ed elaborare grandi volumi di dati in modo efficiente, e Talend Open Studio, uno strumento potente per la Data Integration che consente di automatizzare e ottimizzare i processi di estrazione, trasformazione e caricamento dei dati (ETL). Questo percorso formativo ha permesso di sviluppare una visione completa e operativa dell'intero ciclo di vita del dato, offrendo un'esperienza concreta su tecnologie innovative e largamente impiegate in ambito professionale.

Indice

Elenco delle tabelle

Elenco delle figure

Parte I

Contesto di analisi

Capitolo 1

Contesto di analisi

1.1 Introduzione

L'andamento demografico di una città fornisce una chiave di lettura fondamentale per comprendere le sue trasformazioni sociali, economiche e urbane. Analizzare la demografia significa studiare non solo il numero e la composizione della popolazione, ma anche come fattori esterni – l'economia, la migrazione e le politiche urbane – influenzino questi aspetti nel tempo. Nel contesto urbano di una città complessa come Milano, che svolge un ruolo di primo piano nel panorama economico italiano ed europeo, tali analisi diventano essenziali per guidare la pianificazione del futuro. Negli ultimi decenni, Milano ha vissuto profondi cambiamenti. Da un lato, la globalizzazione e lo sviluppo tecnologico hanno accelerato i processi migratori, sia interni che esterni, modificando la composizione demografica della città. L'arrivo di una popolazione sempre più diversificata dal punto di vista culturale ed economico ha trasformato il tessuto sociale di Milano, creando nuove sfide e opportunità. La crescita della popolazione straniera, in particolare, è un fenomeno che ha visto Milano protagonista, facendone un laboratorio di integrazione multiculturale a livello nazionale [15]. Dall'altro lato, il mercato immobiliare ha giocato un ruolo cruciale nella redistribuzione della popolazione tra il centro e le periferie. Le aree centrali della città sono state interessate da una forte riqualificazione urbana, mentre le periferie hanno visto una crescita significativa grazie alla costruzione di nuovi poli residenziali. Questi fenomeni hanno modificato profondamente la geografia sociale della città, con l'espansione di nuove aree urbane e la riqualificazione di quartieri storici. La presenza di aree riqualificate nelle periferie e il miglioramento delle infrastrutture hanno anche incentivato flussi migratori interni da altre regioni italiane verso Milano, rafforzando la posizione della città come centro di attrazione.
Questa tesi si propone di analizzare queste trasformazioni utilizzando i dati messi a disposizione dal Comune di Milano. I dati comprendono informazioni su nascite, decessi, residenti, costruzioni residenziali e dati censuari e anagrafici. Attraverso l'integrazione di questi dataset, sarà possibile descrivere l'andamento demografico della città negli anni, evidenziare particolari pattern nei dati relativi ad alcune categorie della popolazione e interpretare le correlazioni tra le variabili presenti. L'analisi non si limiterà a un livello

puramente descrittivo, ma si cercherà di indagare le relazioni che intercorrono tra le diverse variabili demografiche. L'obiettivo è anche quello di costruire modelli predittivi che siano non solo accurati da un punto di vista descrittivo, ma anche dal lato predittivo, per essere utilizzati da policy maker e urbanisti. Questi modelli potranno essere la base per la pianificazione di interventi futuri e rispondere in modo efficace alle esigenze di una città in continua evoluzione.
La fase di *Data Integration* sarà cruciale per normalizzare e unificare i dati provenienti da diverse fonti, garantendo che siano coerenti e comparabili. Seguirà una fase di *Data Analysis*, dove si cercherà di individuare possibili correlazioni tra le variabili, trend nei dati e fare previsioni. Infine, la fase di *Data Visualization* consentirà di tradurre i risultati dell'analisi in strumenti visivi chiari, attraverso dashboard interattive, facilitando l'interpretazione e l'utilizzo dei dati da parte dei non esperti.

1.2 Problematiche e sfide

Una delle principali problematiche di questo studio risiede nella complessità e nella molteplicità delle fonti di dati che devono essere integrate. La demografia è una scienza dinamica e per essere analizzata correttamente richiede dati aggiornati e coerenti. Tuttavia, i dati demografici spesso provengono da fonti eterogenee, come anagrafe, censimenti, studi urbani, e non sempre sono direttamente comparabili tra loro. Si rende necessario quindi un processo di manipolazione e uniformazione dei dati provenienti dalle sorgenti per renderli accessibili e per poter compiere delle analisi sugli stessi. In un contesto di continua evoluzione dei dati, risulta anche importante avere informazioni aggiornate costantemente, senza dover ripetere ogni volta i procedimenti effettuati già in passato. Quindi non basta proporre le analisi una volta per tutte, relative ad uno specifico periodo temporale, ma si richiede di automatizzare il processo di analisi, in modo tale che sia al passo col tempo, che consideri dati aggiornati e che dia informazioni precise per poter prendere decisioni più sicure. Inoltre si desidera che queste informazioni siano ricavate in tempi brevi e quindi i procedimenti automatizzati devono essere anche efficienti e capaci di restituire risultati nel minor tempo possibile. Chiaramente anche il quantitativo di dati da analizzare in questi contesti è elevato e ciò comporta che l'ottimizzazione del flusso di operazioni da eseguire per l'integrazione delle fonti, l'analisi e la presentazione dei risultati tenga conto di questo fattore. L'efficienza, tuttavia, non può prescindere dalla qualità dei dati. Spesso i dati provenienti da qualsiasi tipo di sorgenti, risultano grezzi, non elaborati, con errori. Una delle problematiche principali in questo contesto è quella di comprendere quali siano i dati "sporchi" da dover processare e quindi la correzione degli stessi. Anche la scelta di modelli predittivi efficaci e coerenti con la struttura dei dati risulta essere di cruciale importanza. Infatti, non basta che il modello descriva bene i dati del passato, ma deve essere capace di fornire previsioni accurate per il futuro, secondo una qualche misura di accuratezza, anch'essa opportunamente selezionata. D'altra parte un modello che fornisce previsioni accurate potrebbe non descrivere bene i dati su cui il modello è stato adattato, indicando che le previsioni ottenute si fondano su basi poco solide. Infine, risulta importante anche la scelta di adeguate modalità di presentazione dei risultati ottenuti. Questa fase del processo di elaborazione dati viene spesso considerata di minore

importanza, ma in realtà è tutt'altro che irrilevante, dal momento che può portare chi visualizzerà i risultati, ad una non comprensione delle informazioni ottenute o, peggio, a interpretazioni non corrette degli stessi. Spesso coloro a cui vengono sottoposte le analisi dati non sono esperti di questi contesti tecnici ed è necessario quindi individuare degli strumenti di sintesi dei risultati che possano essere di facile comprensione anche per loro, senza perdere il contenuto informativo. Quindi non è banale trovare il giusto compromesso tra la semplicità di interpretazione dei risultati e la consegna delle informazioni contenute dalle analisi.

1.3 Impatto e rilevanza dell'analisi

L'analisi demografica è essenziale per la pianificazione urbana e per la formulazione di politiche pubbliche efficaci. I cambiamenti nella popolazione influenzano direttamente le decisioni relative alla costruzione di nuove infrastrutture, alla gestione dei servizi pubblici, e alla distribuzione delle risorse economiche. Per esempio, l'incremento di nascite in alcune aree potrebbe richiedere un ampliamento delle strutture educative, mentre un aumento della popolazione anziana può portare alla necessità di potenziare i servizi sanitari e di assistenza. Nel contesto milanese, la tesi mira a fornire una base di conoscenza per i *policy maker*, evidenziando come l'andamento demografico possa essere utilizzato per supportare uno sviluppo sostenibile della città. Lo sviluppo urbano non può prescindere dalla comprensione delle dinamiche demografiche: solo attraverso un'analisi accurata dei dati è possibile prevedere i bisogni futuri della popolazione e pianificare interventi mirati e sostenibili nel lungo termine. Inoltre, una migliore comprensione dei flussi migratori e della composizione sociale dei vari quartieri può aiutare a ridurre le disuguaglianze e a promuovere uno sviluppo inclusivo, che tenga conto delle diverse esigenze della popolazione. Infine, si cercherà di fornire strumenti utili non solo per il monitoraggio delle trasformazioni in corso, ma anche per la previsione del futuro andamento residenziale, contribuendo così a una pianificazione territoriale più consapevole e lungimirante.

Parte II

Tecniche e metodologie applicate

Capitolo 2

Struttura generale del processo di analisi

In questo elaborato l'analisi delle fonti di informazioni è di tipo *full stack*, ovvero comprendente tutte le fasi di manipolazione dei dati: dal recupero degli stessi, alla presentazione dei risultati ottenuti. L'obiettivo è quello di mostrare le varie fasi del processo che il dato deve subire prima di arrivare a fornire del contenuto utile. Per tale motivo lo studio è stato suddiviso in tre fasi fondamentali comuni alla maggior parte delle analisi di questo tipo: *Data Integration*, *Data Analysis*, *Data Visualization*. La Data Integration riguarda la parte di elaborazione dei dati, dalla loro estrazione dalle fonti considerate, passando per la loro pulizia e controllo di qualità fino ad arrivare al caricamento sul database locale. Tutto ciò sviluppando un flusso di integrazione che sia ripetibile, chiaro ed efficiente. Ripetibile, perché l'idea è quella di pensare ad un contesto reale in cui i dati si aggiornano nel tempo e quindi non possono essere integrati un'unica volta, ma si vuole ripetere questo processo più volte nel futuro senza ricreare da capo ogni volta gli stessi passaggi, ma automatizzandoli. Chiaro, perché in caso di modifica delle sorgenti originali, in futuro si deve essere in grado di andare ad apportare le correzioni al processo elaborato in modo tempestivo. Efficiente, perché l'aggiornamento dei dati deve avvenire nel minor tempo possibile, soprattutto in contesti dove le decisioni sono da prendere in tempi brevi. Una volta elaborati i dati possono passare alla fase di Data Analysis. In una prima fase si esegue un'analisi esplorativa dei dati, individuando gli andamenti delle variabili e le relazioni tra le stesse. In tal modo si può comprendere quali sono le analisi interessanti da poter svolgere e le variabili di particolare interesse. In seguito, si vanno a identificare i modelli statistici capaci di descrivere bene i dati del passato e allo stesso tempo di predire in modo accurato le evoluzioni del futuro. Dal momento che in questa analisi si stanno trattando dati temporali, ovvero serie storiche, i modelli che verranno utilizzati saranno specifici di questo tipo di dati. In particolare l'attenzione sarà focalizzata sulla serie storica dei residenti di Milano. Quindi si procede alla valutazione delle performance di ogni modello adattato alle serie storiche secondo metriche che misurano la qualità di adattamento ai dati e la qualità della previsione. I risultati devono quindi essere organizzati per la presentazione ad un pubblico esterno, nella fase di Data Visualization. In tale fase si deve trovare

il giusto compromesso tra il presentare la totalità dei risultati ottenuti, con i più disparati strumenti di sintesi (con il conseguente rischio di confondere chi deve utilizzare le analisi) e il presentare un troppo sintetico report (che può risultare poco informativo). Anche la scelta del software per la visualizzazione dei risultati ricopre un ruolo non indifferente: delle dashboard interattive possono permettere una migliore comprensione dei risultati e possono guidare in modo più chiaro un non esperto alla consultazione degli stessi, rispetto a dei report statici non interattivi. Di seguito perciò si andranno a descrivere nel dettaglio come queste tre fasi siano state implementate, nel contesto di studio dei dati demografici e residenziali messi a disposizione dal Comune di Milano.

Capitolo 3

Data Integration

In generale la *Data Integration* si riferisce al processo che permette di uniformare e facilitare l'accesso ad un insieme di sorgenti dati autonome ed eterogenee [7]. Questo processo comporta l'identificazione e la risoluzione delle discrepanze tra i dati, la mappatura delle relazioni tra le diverse entità e la creazione di una struttura comune che consenta l'interrogazione e l'analisi dei dati [17]. Le sfide che la Data Integration deve fronteggiare sono ad esempio:

- *Querying e aggiornamento delle fonti nei sistemi di integrazione dei dati*: sicuramente la scelta delle modalità di interrogazione delle fonti di dati rappresenta un aspetto fondamentale nella Data Integration. Tuttavia, anche l'update in modo efficiente di tali fonti in modo continuativo riveste un certo interesse, poiché in contesti moderni i dati a disposizione sono in costante evoluzione e avere risultati accurati e aggiornati in modo veloce e automatizzato è vantaggioso e addirittura cruciale in alcune situazioni;

- *Numero di fonti*: l'integrazione dei dati rappresenta una sfida anche con un numero esiguo di fonti, ma diventa ancora più complessa man mano che il numero di sorgenti cresce, dal momento che assieme ad esso cresce necessariamente anche l'eterogeneità dei dati;

- *Eterogeneità*: le fonti di dati, sviluppate indipendentemente l'una dall'altra, possono provenire da luoghi diversi, dai database ai semplici file. Queste fonti avranno schemi differenti, anche quando modellano gli stessi dati e possono essere strutturate (es. fogli di calcolo `Excel`), semi-strutturate (es. `JSON` o `XML`) o non strutturate (es. file audio, immagini, video). In alcuni contesti si riesce ad identificare un allineamento tra i diversi schemi dei dati provenienti dalle varie fonti, ma in altri ciò non è possibile data la loro struttura estremamente eterogenea. Tuttavia si possono trovare delle modalità per rendere uniforme la lettura degli stessi, agevolando le analisi successive;

- *Autonomia*: le fonti di dati non sempre appartengono alla stessa organizzazione o ente, il che può limitare i diritti di accesso ai dati stessi. Ciò implica che, in alcuni casi, non si ha piena disponibilità dei dati, o che siano necessarie autorizzazioni

specifiche per accedervi. Inoltre, i formati dei dati e i metodi per accedervi possono cambiare senza preavviso, complicando ulteriormente il processo di integrazione, poiché i sistemi devono essere adattati costantemente per gestire queste variazioni, il che richiede un'attenzione continua alla compatibilità tra sistemi.

Uno step fondamentale all'interno della Data Integration è lo studio della *Data Quality*, ovvero tutto ciò che concerne l'identificazione di non idoneità tra i dati, la valutazione del loro impatto sulle analisi, il loro possibile miglioramento o la loro evenutale rimozione. In generale solitamente si parla di "dati" e di "informazioni" come sinonimi, tuttavia nella pratica oggigiorno ci si riferisce alle informazioni come dati processati, ovvero dati controllati, corretti, più sicuri [24]. Si può dire che la Data Quality si riferisce alla misura in cui i dati soddisfano i requisiti di accuratezza, consistenza, tempestività, e completezza. In questo contesto si attua la *Data Cleaning*, ovvero il processo di identificazione e correzione di errori, incoerenze e valori mancanti all'interno di un dataset. La qualità dei dati grezzi raccolti può variare notevolmente e la presenza di anomalie, come valori duplicati o fuori scala, può influenzare negativamente le analisi successive. La pulizia dei dati permette di avere una maggiore accuratezza, completezza e affidabilità delle informazioni. In questa fase, tra le altre, si vanno ad operare le seguenti azioni fondamentali:

- *Rimozione di duplicati*: Identificare e rimuovere record duplicati che possono distorcere i risultati;
- *Gestione dei valori mancanti*: Valutare come trattare i dati `NULL` o mancanti, che possono essere imputati (stimati) o eliminati;
- *Correzione di errori tipografici e incongruenze*: Correggere errori manuali, come formati di date non coerenti o errate categorizzazioni.

In questo capitolo verranno descritti i processi di Data Integration e di Data Cleaning eseguiti sulle fonti dati considerate. Nella sezione 3.1 verrà data una descrizione del software utilizzato per implementare i flussi di integrazione dati, ovvero Talend Open Studio for Data Integration; Nella Sezione 3.2 verranno esposte le sorgenti prese in esame e verranno descritte le modalità di interrogazione della API fornita dal sito del Comune di Milano; nella Sezione 3.3 si descrivono il metodo *Ingestion Delta* impiegato per l'integrazione dei dati, la creazione delle tabelle sul database, la definizione dei metadati relativi alle sorgenti prese in esame, i flussi implementati in Talend relativi all'acquisizione degli stessi; nella Sezione 3.4 vengono riportate le operazioni relative alla Data Quality.

3.1 Talend Open Studio for Data Integration

Talend Open Studio for Data Integration è un software open-source, basato sul linguaggio `Java`, che permette di progettare e automatizzare i processi ETL (Extract, Transform, Load) per la gestione e l'integrazione dei dati. Grazie a una serie di funzionalità intuitive e a una vasta libreria di componenti, Talend Open Studio si rivela uno strumento efficace per chi lavora nell'analisi dei dati, facilitando la gestione e il trasferimento di grandi volumi di dati tra sistemi differenti. La sua interfaccia grafica user-friendly consente agli utenti di

costruire flussi ETL complessi tramite un sistema di drag-and-drop, riducendo il bisogno di programmazione manuale e rendendo l'interfaccia accessibile anche a chi ha competenze limitate in ambito tecnico. Talend Open Studio è stato progettato per connettersi a una vasta gamma di fonti dati, incluse piattaforme SQL e NoSQL, file flat come CSV, sistemi ERP e cloud. L'ampia libreria di componenti Talend facilita l'automazione dei processi ETL, fornendo strumenti di pianificazione che riducono l'intervento manuale e minimizzano il rischio di errori. La disponibilità di componenti predefiniti per diverse operazioni di estrazione, trasformazione e caricamento contribuisce a migliorare l'efficienza e a standardizzare i processi di integrazione, facilitando l'elaborazione di dati provenienti da fonti diverse. Il flusso di lavoro tipico in Talend Open Studio prevede inizialmente la connessione alle fonti di dati, che possono includere database relazionali, database NoSQL, fonti cloud o API RESTful. Una volta stabilita la connessione, è possibile avviare operazioni di estrazione e trasformazione, come la pulizia, la normalizzazione e l'arricchimento dei dati, necessarie per ottenere un dataset coerente e accurato, pronto per essere analizzato o trasferito a sistemi di destinazione. Infine, i dati così trasformati possono essere caricati nella destinazione finale, come un database, un data warehouse o una piattaforma di analisi. Talend è in grado di integrarsi con sistemi di visualizzazione dei dati come Power BI, permettendo di completare il ciclo di gestione dei dati. L'utilizzo di Talend Open Studio presenta numerosi vantaggi nella gestione dei dati. La possibilità di automatizzare i processi ETL consente infatti di ridurre i tempi di esecuzione e di destinare maggiori risorse all'analisi stessa dei dati. Inoltre, Talend è progettato per essere altamente scalabile, capace di gestire grandi volumi di dati e compatibile con piattaforme di big data come Hadoop e Spark, rendendolo ideale per ambienti con grandi volumi di dati. Un ulteriore beneficio è dato dagli strumenti di validazione e pulizia, che permettono di garantire un alto standard di qualità per i dati caricati, e di ottenere analisi basate su dati affidabili. Talend Open Studio si integra facilmente con i principali database relazionali, come PostgreSQL, e offre componenti nativi che facilitano la connessione e l'interazione tramite SQL. Talend supporta inoltre il caricamento incrementale, sincronizzando regolarmente i dati con le fonti originali, e dispone di strumenti di monitoraggio e gestione degli errori per assicurare il controllo e la qualità del flusso di dati.

3.2 Studio sorgenti

Le fonti dati oggetto di questa analisi sono relative ai dati demografici e residenziali provenienti dal database messo a disposizione dal Comune di Milano (https://dati.comune.milano.it/). Tutti i dati (tranne il dataset `costruzioni` che, come si vedrà, ha una struttura a sé) seguono una logica comune, secondo un'idea di stratificazione dei dati, per la quale i record sono presentati come la combinazione di variabili categoriche seguite da un campo che conta il numero di occorrenze di record di questo tipo. Questa modalità è usuale in contesti dove si vogliono preservare le informazioni sensibili relative alla singola persona, ma allo stesso tempo si desidera fornire delle informazioni utili per una analisi, proponendo quindi un dato più generale associato a un gruppo di persone che presentano le stesse caratteristiche. Corrisponde sostanzialmente a compiere una operazione di `GROUP BY` sul dataset comprendente i dati relativi alle singole persone, aggregando su tutte le

variabili categoriche e aggiungendo un conteggio di numerosità. Quindi, ad esempio, se un record del dataset `decessi_2003_2020` presenterà i valori come in Tabella 3.1, significherà che si sono rilevate 3 occorrenze di decessi nel 2013, nel quartiere Cascina Merlata, di maschi, di cittadinanza italiana.

_id	anno_evento	quartiere	genere	cittadinanza_deceduto	numerosita
325	2013	Cascina Merlata	Maschi	Italiana	3

Tabella 3.1. Esempio di record nel dataset `decessi_2003_2020`

A ciascun dataset è stato assegnato un nome identificativo per una più chiara e agevole manipolazione dei dati:

- `nascite_2003_2020`: è il dataset che comprende le nascite nel comune di Milano dal 2003 al 2020, alla data di consultazione della risorsa il numero di record che presenta sono 10.481. Presenta i seguenti campi:
  - _id (codice identificativo del record);
  - Anno_evento (anno della nascita);
  - Quartiere (quartiere di residenza della nascita);
  - Genere (maschile o femminile, relativo alla nascita);
  - Cittadinanza_madre2 (italiana o straniera, relativa alla madre del nascituro);
  - Cittadinanza_padre2 (italiana o straniera, relativa al padre del nascituro);
  - Numerosita (occorrenze rilevate relative all'aggregazione su tutti i campi precedentemente citati);
- `nascite_2021_2023`: è il dataset che comprende le nascite nel comune di Milano dal 2021 al 2023, alla data di consultazione della risorsa il numero di record che presenta sono 2.082. Presenta i campi analoghi a `nascite_2003_2020`;
- `decessi_2003_2020`: è il dataset che comprende i decessi nel comune di Milano dal 2003 al 2020, alla data di consultazione della risorsa il numero di record che presenta sono 4.447. Presenta i seguenti campi:
  - _id (codice identificativo del record);
  - Anno_evento (anno considerato del decesso);
  - Quartiere (quartiere di residenza del decesso);
  - Genere (maschile o femminile, relativo al decesso);
  - Cittadinanza_deceduto (italiana o straniera, relativa al decesso);
  - Numerosita (occorrenze rilevate relative all'aggregazione su tutti i campi precedentemente citati);

- `decessi_2021_2023`: è il dataset che comprende i decessi nel comune di Milano dal 2021 al 2023, alla data di consultazione della risorsa il numero di record che presenta sono 855. Presenta i campi analoghi a `decessi_2003_2020`;
- `residenti`: è il dataset che comprende i residenti nel comune di Milano dal 1999 al 2022, suddivisi per NIL (Nuclei d'Identità Locale) [1], alla data di consultazione della risorsa il numero di record che presenta sono 2.112. Presenta i seguenti campi:
  - _id (codice identificativo del record);
  - ID_NIL (codice identificativo del quartiere di residenza);
  - NIL (Nucleo d'Identità Locale - nome del quartiere di residenza);
  - Anno (anno in cui è stata eseguita la rilevazione);
  - Residenti (numero totale di residenti per quel quartiere e quell'anno);
- `costruzioni`: è il dataset che comprende gli edifici residenziali (ovvero quelli per cui più del 50% della superficie totale è destinata ad uso abitativo) nel comune di Milano dal 2010 al 2023, alla data di consultazione della risorsa il numero di record che presenta sono 882. Presenta i seguenti campi:
  - _id (codice identificativo della costruzione);
  - Municipio (codice identificativo del municipio in cui si trova la costruzione);
  - NIL (quartiere della costruzione);
  - Anno Ritiro (anno di ritiro del permesso per costruire l'edificio);
  - Descr Titolo Abilitativo (descrizione del titolo abilitativo per la costruzione);
  - Numero Abitazioni (numero di abitazioni presenti nella costruzione);
  - Numero Stanze (numero di stanze presenti nella costruzione);
  - Numero Vani Accessori Interni (numero di vani accessori interni);
  - Numero Piani (numero di piani presenti nella costruzione);
  - Volume Totale V/P (volume totale della costruzione);
  - Superficie Utile Abitabile (superficie abitabile utile della costruzione);
  - Superficie Accessori Esterni (superficie accessori esterni della costruzione);
  - Superficie Agricoltura (superficie destinata all'agricoltura);
  - Superficie Attività Produttive (superficie destinata alle attività produttive);
  - Superficie Commercio (superficie destinata al commercio);

[1] Il NIL (Nucleo d'Identità Locale) è una suddivisione territoriale utilizzata dal Comune di Milano per gestire e analizzare le caratteristiche sociali, demografiche e urbanistiche della città. Ogni NIL corrisponde a un quartiere o a una porzione di quartiere, e rappresenta un'unità di riferimento per la raccolta e l'elaborazione dei dati. Viene utilizzato per monitorare l'evoluzione demografica, pianificare i servizi e progettare interventi infrastrutturali in modo mirato e consapevole [21].

  - Superficie Servizi (superficie destinata ai servizi);
  - Specifica Altra Attività (specifiche di eventuali altre attività);
  - Descr_Titolare Titolo Costruire (descrizione del titolare del titolo per costruire);
  - Consumo Energetico (consumo energetico della costruzione);
  - Rapporto Forma (rapporto di forma della costruzione);
  - Fotovoltaico (presenza o meno di impianto fotovoltaico);
  - Solare Termico (presenza o meno di impianto solare termico);
  - Pompe Calore (presenza o meno di pompe di calore);
  - Caldaia a Condensazione (presenza o meno di caldaia a condensazione);
  - Geotermico (presenza o meno di impianto geotermico);
  - Nessun Impianto Tra I Citati (assenza o meno di impianti tra quelli citati);
  - PianoCasa (indica se la costruzione è stata soggetta a interventi edilizi specifici nell'ambito di programmi normativi, con lo scopo di migliorare l'efficienza energetica, aumentare il numero di abitazioni, o riqualificare aree urbane).

- `tipo_fam_res_cens`: è il dataset che riguarda la ricostruzione del numero di famiglie residenti a Milano a partire dal 2003 aggiornato sulla base dei trend anagrafici e dei Censimenti 2001 e 2011, alla data di consultazione della risorsa il numero di record che presenta sono 39.369. Presenta i seguenti campi:

  - _id (codice identificativo del record);
  - Anno (anno di rilevazione);
  - NIL (Nuclei d'Identità Locale) (in questo caso sono la concatenazione del nome del quartiere e l'ID del quartiere);
  - Genere (genere del capofamiglia);
  - Eta_capofamiglia (fascia di età del capofamiglia);
  - Numero componenti (numero di componenti del nucleo familiare);
  - Famiglie (occorrenze rilevate relative all'aggregazione su tutti i campi precedentemente citati);

- `tipo_fam_res_anag`: è il dataset che riguarda il numero di famiglie residenti a Milano e iscritte nell'Anagrafe della popolazione, a partire dal 1999, alla data di consultazione della risorsa il numero di record che presenta sono 1.038.409. Presenta i seguenti campi:

  - _id (codice identificativo del record);
  - Anno (anno di rilevazione);
  - Quartiere (in questo caso sono la concatenazione del nome del quartiere e l'ID del quartiere);
  - Classe_eta_capofamiglia (fascia di età del capofamiglia);

  - Genere_capofamiglia (genere del capofamiglia);
  - Numero_componenti (numero di componenti del nucleo familiare);
  - Tipologia_familiare (tipologia del nucleo familiare);
  - Cittadinanza (cittadinanza del capofamiglia)
  - Famiglie (occorrenze rilevate relative all'aggregazione su tutti i campi precedentemente citati);

- `popolazione_istat`: è il dataset fornito da ISTAT e messo a disposizione da Oper Data Milano, che descrive il movimento naturale e migratorio dal 1880 al 2022 nella popolazione residente di Milano, alla data di consultazione della risorsa il dataset comprende 143 record, uno per ogni anno compreso tra il 1880 e il 2022. Tale dataset è stato preso in considerazione per fornire delle proiezioni sulla serie storica dei residenti nella popolazione di Milano, prendendo in considerazione un periodo più lontano rispetto a quello proposto dal dataset `residenti`, il quale ha informazioni solo dal 1999. Presenta i seguenti campi:

  - Anni (anno di rilevazione);
  - Nati vivi (nati vivi nell'anno di rilevazione);
  - Morti (decessi nell'anno di rilevazione);
  - Immigrati (immigrati a Milano nell'anno di rilevazione);
  - Emigrati (emigrati da Milano nell'anno di rilevazione);
  - Popolazione calcolata Comune di Milano (fine anno) (residenti a Milano nell'anno di rilevazione).

I dati possono essere acquisiti tramite download diretto nel formato desiderato (`.csv`, `.json`, `.xml`) oppure si può interrogare la API fornita dal sito. Le fonti di dati utilizzate sono aggiornate fino a giugno 2024, con l'unica eccezione dei dati relativi ai `residenti`, aggiornati a Novembre 2023. E' importante sottolineare che non è comune avere accesso a fonti così recenti e aggiornate in modo gratuito. Disporre di dati attuali è di fondamentale importanza in quanto consente di ottenere analisi più precise e attendibili. Dati aggiornati riflettono la realtà attuale e garantiscono che i risultati delle analisi siano rilevanti rispetto al contesto contemporaneo. Questo è particolarmente utile quando si effettuano previsioni o si cerca di individuare trend demografici, economici o residenziali, poiché permette di basare le decisioni su informazioni il più possibile vicine alla situazione reale, riducendo il margine di errore e migliorando la qualità complessiva dei risultati.

3.3 Acquisizione dati

3.3.1 Interrogazione API

L'acquisizione dei dati dalle risorse online fornite dal Comune di Milano è avvenuta tramite la *CKAN Data API* (Application Programming Interface) fornita dal sito, ovvero un'interfaccia che permette di interrogare liberamente il database online, seguendo i suoi

Figura 3.1. Screenshot dal sito del Comune di Milano della pagina associata al dataset `nascite1`.

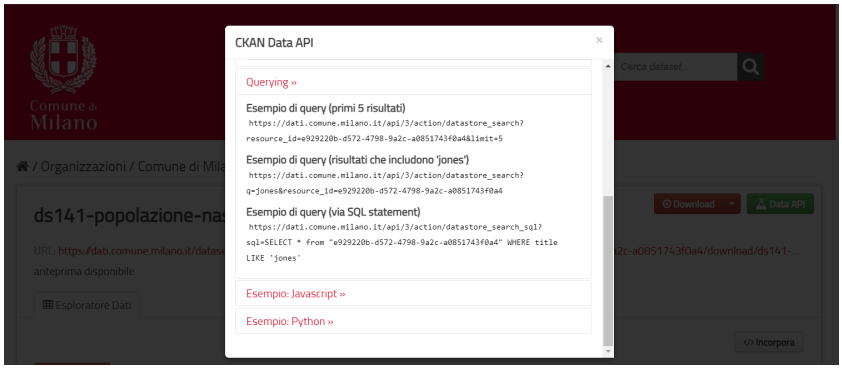

Figura 3.2. Dettaglio sugli esempi di query forniti per interrogare la API.

aggiornamenti nel tempo. In generale, in un contesto di staticità del dataset, ovvero nel quale non si contemplano cambiamenti nel tempo, è meno importante la modalità attraverso cui si procede al download dello stesso. In questo caso invece è naturale supporre che in futuro verranno non solo inseriti nuovi dati, ma anche aggiornati record del passato o magari cancellati. Infatti, può accadere che vengano rilevati degli errori nel dataset che vanno corretti nella maniera adeguata, con cautela ed efficienza. Per tale motivo, tramite l'interrogazione della API fornita si può automatizzare il processo di acquisizione dei dati nel tempo, rendendo più semplice ed efficiente in futuro l'aggiornamento degli stessi. In particolare, questo si può attuare utilizzando le query suggerite dal sito per poter interrogare la API (Figura 3.2) e inserendole nel software di ETL che si utilizza. Il dataset `popolazione_istat`, dal momento che ha dimensioni ridotte e fornisce informazioni già processate, verrà acquisito come semplice file `CSV` e analizzato in Python.

3.3.2 Ingestion Delta

L'approccio utilizzato per l'aggiornamento progressivo nel tempo dei dataset nel database locale è di tipo *Ingestion Delta*. Tale modalità si basa sull'idea di minimizzare il numero di

operazioni di modifica da apportare al dataset presente nel database locale (*tabella target*), nel momento in cui si riscontrino differenze con il dataset corrispondente aggiornato online (*tabella sorgente*), andando ad operare le correzioni solo dove si riscontrano i "delta", ovvero le differenze tra le due tabelle, da qui il nome Ingestion Delta. Le azioni che verranno eseguite (e spiegate più nel dettaglio in seguito) sono le seguenti:

1. *Update*: se un record presente nella tabella target differisce per qualche campo da quello corrispondente nella tabella sorgente, secondo la chiave identificativa del record, si procederà ad un aggiornamento in target prendendo i nuovi valori presenti in sorgente;

2. *Insert*: se è presente un record nella tabella sorgente che non è presente nella tabella target, ovvero ci sono delle chiavi di record in sorgente che non sono presenti in target, significa che è stato aggiunto un nuovo record nella tabella sorgente e quindi verrà inserito nella tabella target;

3. *Delete*: se è presente un record nella tabella target che non è presente nella tabella sorgente, significa che un record che era presente prima è stato rimosso e quindi si procederà alla cancellazione anche dalla tabella target.

Per i rimanenti record, ovvero quelli che risultano invariati tra tabella sorgente e tabella target, non si eseguirà alcuna operazione. Un approccio di questo tipo permette di non ricopiare l'intero dataset ad ogni ciclo di integrazione dei dati, ma di aggiornarlo in modo incrementale con le sole modifiche effettuate, rendendo il processo molto più efficiente e scalabile. Solitamente i record presenti in una tabella sorgente hanno anche un campo relativo al timestamp dell'ultima modifica eseguita sul record. In tal caso l'Ingestion Delta può essere implementato in modo più efficiente andando a selezionare soltanto i record con ultima data di modifica maggiore rispetto all'ultima data di aggiornamento della tabella target che è stato eseguito, o comunque rispetto al massimo dell'ultima data di aggiornamento presente in target. In questo contesto non si ha l'informazione sul timestamp in sorgente, quindi, nel caso di corrispondenza di chiave tra sorgente e target si avrebbe un alto numero di controlli campo per campo nel record per verificare se il record è stato modificato. In contesti dove si ha un elevato numero di campi o valorizzazioni complesse degli attributi, solitamente si aggiunge un campo in coda al dataset contenente l'`hash`, secondo una qualche funzione di Hash, della concatenazione dei valori associati al record. Ciò viene effettuato per velocizzare le operazioni di confronto per le Update: invece di eseguire il controllo campo per campo si può andare ad attuare il confronto unicamente per il campo associato al valore dell'hash. Tuttavia, nel caso di questa analisi, dal momento che il numero di campi è ridotto e non si hanno valorizzazioni complesse, risulta più efficace eseguire il controllo campo per campo piuttosto che l'approccio con hash, poiché risulterebbe più pesante l'applicazione della funzione alla concatenazione di tutti i campi del record. In ogni caso, il flusso di operazioni definito tramite la logica di delta consente di mantenere ordinato il processo di integrazione dei dati, evidenziando le variazioni tra un aggiornamento e l'altro, in modo da poter avere chiara l'evoluzione delle informazioni nel tempo.

3.3.3 Creazione tabelle nel database locale

In questo elaborato è stato utilizzato PosgreSQL come sistema di gestione del database relazionale (RDBMS) e DBeaver come client per la gestione e l'interazione con il database. PostgreSQL è stato scelto per la sua robustezza, scalabilità e supporto per funzionalità avanzate di gestione dei dati, mentre DBeaver ha un'interfaccia utente intuitiva e permette di agevolare l'esplorazione dei dati. Per procedere al caricamento dei dati nel database locale, sono state create le tabelle dando i nomi prima citati e specificando i tipi di dato associati ai campi descritti in ogni dataset. In questa fase si è deciso di rinominare i campi delle tabelle, i quali presentavano upper e lower case, spazi e caratteri speciali, rendendoli omogenei così da facilitare il processo di pulizia dei dati e di analisi (ad esempio il campo `Superficie Attivita' Produttive` è stato rinominato in `superficie_attivita_produttive`). Inoltre si è deciso di introdurre un ulteriore campo chiamato 'ts_lettura' in ogni tabella, con lo scopo di salvare sul database, oltre alle informazioni provenienti dalle sorgenti, anche la data di ultimo aggiornamento del record.

3.3.4 Creazione metadati Talend

In Talend è possibile creare *metadati*, ovvero dei dati che descrivono la struttura e che contengono le informazioni a proposito di altri dati, ma non il contenuto dei dati stessi. Un oggetto di questo tipo risulta utile dal momento che permette di comunicare alla piattaforma di ETL utilizzata quale è la struttura del dato che si sta acquisendo dalla risorsa. E' importante sottolineare che in Talend si distinguono i concetti di metadato e schema relativo a un dataset, concetti che possono risultare simili ma hanno delle differenze nel contesto dell'ambiente di ETL. Per metadato si intende la modalità di lettura della sorgente che viene considerata, che comprende come sono individuate le informazioni (ad esempio tramite `JSONPath` o `XPath`), il carattere separatore, se i dati sono racchiusi tra virgolette, ecc. Per schema invece si vuole indicare la struttura del dataset, quindi il numero di colonne, il tipo di dato associato ad ogni colonna, se ci sono dei vincoli particolari associati ai campi. Questa distinzione permette di utilizzare, in base al contesto, o il metadato o lo schema ad esso associato. Dalla API del sito del Comune di Milano, i dati sono stati acquisiti come file `JSON`, i quali hanno una precisa struttura che segue l'idea dei dizionari chiave-valore, di semplice e chiara interpretazione. In questo caso, si è deciso di creare un metadato associato ad ogni risorsa. Nel creare lo schema per il dataset è possibile specificare anche il nome che si vuole associare al campo ricavato dalla risorsa: in questa fase sono stati rinominati i campi dando una denominazione uniforme come descritto in precedenza.

3.3.5 Costruzione flussi Talend

La struttura dei flussi di operazioni in Talend per l'integrazione dei dati è sostanzialmente la stessa per ciascun dataset, quindi verrà riportato il procedimento utilizzato per la prima sorgente dati `nascite_2003_2020` sottintendendo che per i restanti dataset il flusso implementato è analogo. In Figura 3.3 si riporta un esempio del Job Talend (ovvero il processo di ETL) costruito. I Subjob che si trovano in alto riguardano il Prejob e il

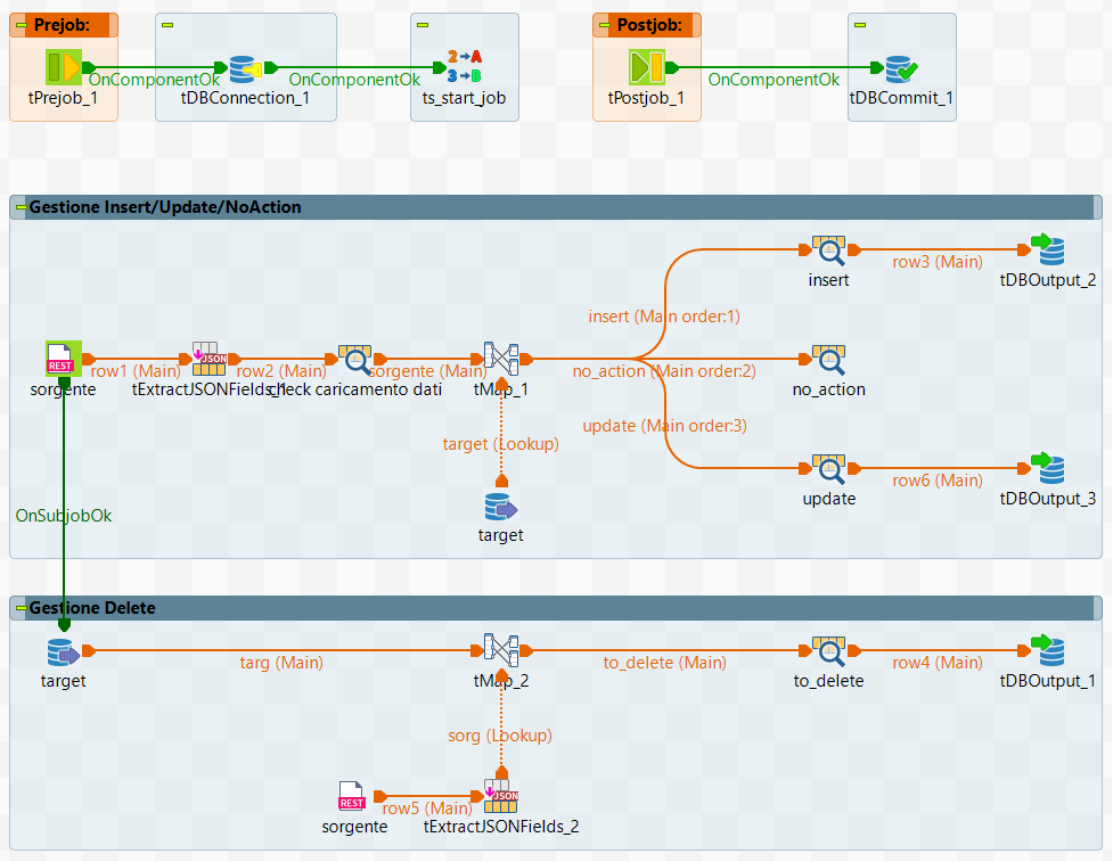

Figura 3.3. Esempio flusso Talend per l'integrazione dei dati.

Postjob, ovvero le operazioni che verranno eseguite all'inizio e alla fine del Job completo. In questo caso come Prejob viene stabilita la connessione con il database locale e viene definita una variabile globale contenente il timestamp attuale, utile per tenere traccia degli aggiornamenti dei dati. Nei dataset sul database locale, infatti, verrà inserito un campo aggiuntivo in ogni dataset contenente il timestamp relativo all'ultimo aggiornamento del record. Come Postjob viene invece indicata l'operazione di commit sul database delle modifiche operate durante il Job. Le operazioni all'interno del flusso sono state suddivise in due parti: la prima riguardante le operazioni di Insert, Update, No Action ovvero di inserimento di nuovi dati, di modifica di dati già esistenti e di non modifica sui dati invariati; la seconda riguardante soltanto l'operazione di Delete, ovvero la rimozione dei dati non più presenti in sorgente. In tal modo si ottimizza il flusso andando a estrarre i dati da sorgente una sola volta sia per il flusso di Insert-Update, sia per il flusso di Delete. Vengono descritte ora le operazioni eseguite da ciascun componente Talend utilizzato nel primo flusso di operazioni (da sinistra verso destra):

- `tRest`: interroga l'API come descritto in precedenza recuperando la sorgente aggiornata. Riceve in ingresso il link richiesta HTTP che interroga il datastore di un dataset pubblico del Comune di Milano;
- `tExtractJSONFields`: estrae i dati dalla sorgente ricevuta in ingresso riconoscendo la struttura del record del dataset fornita dal metadato definito in precedenza;
- `tLogRow`: permette di controllare l'output del componente precedente, fornendo una visualizzazione tabellare del dataset all'interno della finestra di esecuzione di Talend. E' utile per dei controlli intermedi in fase di costruzione del flusso di operazioni;
- `tMap` (Figura 3.4): in generale permette di eseguire mappature nei dati secondo le regole definite dall'utente. In questo caso è utile per eseguire le operazioni di Join tra la tabella sorgente (quella appena caricata) e la tabella target (quella presente in locale) filtrando i dati in base alle condizioni desiderate, per implementare il metodo Ingestion Delta. E' stato impostato come Join Model l'Inner Join, dove la tabella di sinistra è la sorgente, quella di destra è la target e la chiave di Join si ritrova nel campo `_id` di ogni dataset. Per chiave si intende la *primary key* del dataset, ovvero il campo che presenta valori univoci per ogni record, permettendo di identificare il record guardando soltanto al valore assunto dal campo. In realtà il componente `tMap` permette di attuare operazioni diverse in base a delle condizioni desiderate. L'output verrà diviso in tre rami, partizionando nei seguenti modi i dati:
  - Insert (Figura 3.5): in tale output è stata scelta l'opzione *Catch lookup inner join reject: true* poiché si desiderano i record in sorgente che non hanno corrispondenza di `_id` in target. Tali record verranno inviati sul ramo di insert inserendo nel campo relativo al timestamp quello attuale. Si noti che seppure il Join Model scelto è la Inner Join, in realtà in questo caso si è attuata una Left Anti Join dal momento che si è rifiutata l'intersezione;
  - Update (Figura 3.6): per tale output è stata scelta l'opzione *Catch lookup inner join reject: false* perché si desiderano i record che hanno corrispondenza di `_id` ma, come si può notare dalla condizione in rosso, si è aggiunta anche la richiesta che i campi "non chiave" di tali record siano diversi almeno per un campo. Tali record sono i record da aggiornare poiché in almeno un campo è cambiato il valore, verranno perciò inviati nel ramo di update aggiornando anche il campo del timestamp. In questo caso si ha una vera e propria Inner Join tra le tabelle ma con un filtraggio sui dati che hanno almeno un campo con valore diverso;
  - No Action (Figura 3.7): nell'ultimo output si prendono i rimanenti record, ovvero quelli che hanno corrispondenza di chiave tra sorgente e target (*Catch lookup inner join reject: false*) ma che hanno tutti i valori nei campi uguali. Per tali record non verrà apportata alcuna modifica e verranno quindi mandati nel ramo di no action. Anche in questo caso l'operazione eseguita tra le tabelle è una Inner Join.
- `tDBOutput`: gli ultimi componenti riguardano l'operazione da eseguire sulla tabella target in relazione ai record che entrano. In uno si eseguirà la Insert e nell'altro si

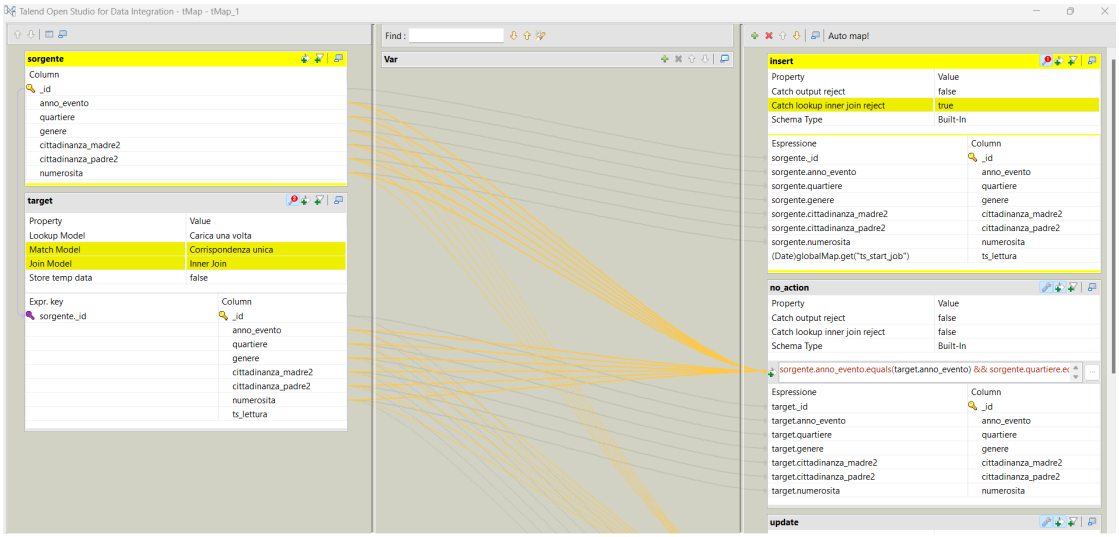

Figura 3.4. Esempio flusso Talend per l'integrazione dei dati - tMap.

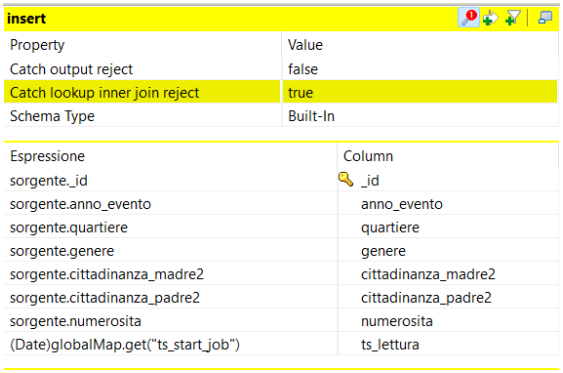

Figura 3.5. Esempio flusso Talend per l'integrazione dei dati - tMap Insert.

eseguirà l'Update. Nel ramo centrale non è presente alcun `tDBOutput` dal momento che non si deve eseguire nessuna operazione per quei dati.

Il secondo flusso di operazioni, come si vede in basso nella Figura 3.3, è relativo alle cancellazioni dalla tabella target di record non più presenti in sorgente. In tal caso è necessario soltanto estrarre gli `_id` dei record in sorgente e in target e attuare una Left Anti Join tra la tabella target e la sorgente, estraendo così gli `_id` dei record da rimuovere (perché presenti in target ma non in sorgente).

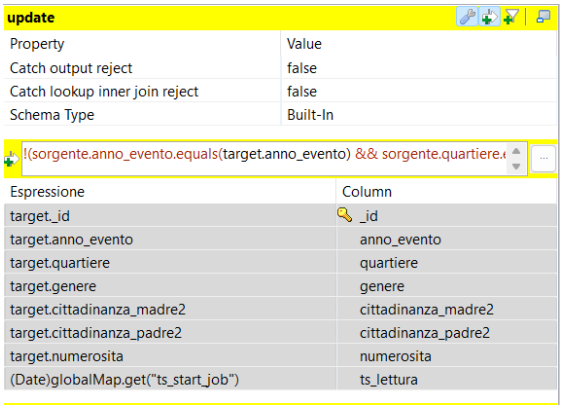

Figura 3.6. Esempio flusso Talend per l'integrazione dei dati - tMap Update.

Figura 3.7. Esempio flusso Talend per l'integrazione dei dati - tMap No Action.

3.4 Data cleaning

Prima di passare alla fase di analisi, è necessario eseguire un controllo sull'integrità e correttezza dei dati acquisiti. Spesso, infatti, accade che i dati siano corrotti a causa di errori in fase di caricamento da parte di utenti o operatori (errori di battitura, confusione tra upper e lower case, ecc.) e per ottenere risultati di analisi significativi e attendibili, è imprescindibile correggere eventuali anomalie e incoerenze nei dati, al fine di garantire una base informativa solida e priva di errori. Tali correzioni, a seconda della casistica di errore rilevata, vengono apportate o al flusso Talend di ETL o direttamente sul database, sui dataset acquisiti. Chiaramente non tutti gli errori possono essere rilevati e non si può mai essere sicuri di manipolare dei dati certi, tuttavia rimuovere gli errori evidenti permette di avere un dataset che si avvicina di più a quello ideale, privo di errori, portando, di conseguenza, anche i risultati ad essere più veritieri. Si riportano gli errori riscontrati nei diversi dataset sia dal punto di vista del tipo di dato o di caricamento, sia dal punto di vista di errori logico/concettuali e le correzioni apportate agli stessi.

- `nascite_2003_2020`:
  - Il campo 'quartiere' presenta molti campi vuoti, per indicare la non presenza dell'informazione. Rispetto ad altri `NULL`, questi sono di maggiore rilevanza, perché rendono il record non utilizzabile, dal momento che l'informazione sul quartiere è molto importante ai fini di una comprensione accurata dell'andamento delle altre variabili a livello locale, nei quartieri di Milano. I record che hanno questo campo vuoto sono 41 in questo caso. La loro eventuale rimozione verrà eseguita direttamente sul database successivamente, nella fase di analisi, poiché nell'ottica di possibili aggiornamenti futuri dei dati è necessario controllare costantemente il numero di campi vuoti per valutare l'impatto della loro cancellazione sulle analisi successive;
  - Inoltre questo campo possiede la quasi totalità dei valori che iniziano con uno spazio; controllando nel file `JSON` originale del dataset sorgente si trova lo stesso pattern nei dati, escludendo un possibile errore in fase di ETL. Tale spazio può quindi essere rimosso da tutti i record, per permettere un confronto con lo stesso campo presente in altri dataset come `costruzioni` nel quale gli spazi non sono presenti. Tale operazione viene eseguita nel flusso Talend, come correzione immediatamente successiva al caricamento del nuovo dataset sorgente.
- `nascite_2021_2023`: si sono attuate le stesse correzioni apportate nel dataset precedente `nascite_2003_2020`, tranne per la gestione dei `NULL` nel campo 'quartiere', che in questo caso non sono presenti;
- `decessi_2003_2020`: sono state attuate correzioni analoghe a quelle attuate in `nascite_2003_2020`. In questo caso i record con campo 'quartiere' vuoto sono 17 e anche in questo caso si rimanda la loro eventuale rimozione alla fase di analisi;
- `decessi_2021_2023`: si sono attuate le stesse correzioni attuate in `nascite_2021_2023`. In questo caso è stato trovato anche un errore nell'inserimento del quartiere 'Niguarda - Caâ Granda - Prato Centenaro - Q.re Fu', il quale non dovrebbe avere il carattere â. Per tale motivo viene rimossa tale lettera dai record che presentano questo valore per uniformare il modo in cui è espresso tale quartiere nelle altre sorgenti;
- `residenti`: in questo caso il campo 'nil', che è associato al campo 'quartiere' presente in altri campi, possiede nel file sorgente originale, in ogni valore presente, uno spazio alla fine. Per tale motivo si procede con l'operazione di *trim* eseguita anche negli altri casi (ovvero la cancellazione di spazi all'inizio e alla fine delle stringhe considerate);
- `costruzioni`: tale dataset presenta un numero notevole di campi e con diversi tipi di dato, per tale motivo è facile aspettarsi un numero più alto di correzioni da effettuare:
  - Sono stati rilevati due campi che presentano nel file sorgente originale un apice nella loro denominazione ("Superficie attivita' produttive" e "Specifica altra attivita'"). Nel processo di acquisizione si hanno difficoltà nell'operazione di parsing del file `JSON` estratto da sorgente, dovute all'apice presente in questi campi. Talend non permette infatti di effettuare l'escape dell'apice in questo

contesto, seppure in altre situazioni ciò è reso possibile. Per tale motivo si decide di procedere alla rimozione dell'apice da questi campi prima del parsing;

- Il campo 'nil' presenta 87 valori `NULL`. In questa fase non si procede alla rimozione di tali record per essere fedeli alla sorgente estratta, ma eventualmente verranno non considerati nella fase di analisi;
- I due campi 'consumo_energetico' e 'rapporto_forma' sono stati acquisiti come di tipo `String`, tuttavia possiedono soltanto dati di tipo `Numeric`. Per tale motivo si procede a una conversione del tipo di dato già nel flusso Talend, portando entrambi i campi al tipo `Numeric`. In tale fase, è stato eseguito un passaggio intermedio di sostituzione delle virgole in punti nei valori numerici e una conversione dei campi stringa vuoti in `NULL`, per poter effettuare la coversione;
- Il campo 'descr_titolo_abilitativo' è qualitativo con livelli e presenta tra gli altri i due valori distinti 'Permesso di costruire' e 'Permesso di Costruire', che indicano la stessa area; perciò sostituiamo la c maiuscola in minuscola;
- Analogamente è stata sostituita 'Impresa / Societ\x85', con 'Impresa / Società';
- Gli ultimi campi del dataset, sono dei campi booleani del tipo sì/NULL, per indicare la presenza di un dato attributo nella costruzione o no. In questi campi si ha disomogeneità per esprimere la presenza del carattere, poiché le formulazioni della presenza sono diverse: 'SI', 's\x8d' e 'si'. Per tale motivo si decide di sostituire tutto con 'SI', per uniformarne la lettura e allo stesso tempo mantenere la struttura originale dell'informazione.
- Sono stati rilevanti anche degli errori di tipo concettuale andando ad osservare più nel dettaglio i dati in questi ultimi campi booleani. I campi riguardano la presenza o assenza di pannelli fotovoltaici, pannelli solari, pompa calore, ecc. e l'ultimo campo presenta la dicitura 'nessun impianto tra i citati'. Un record è risultato avere il flag in quest'ultimo campo e contemporaneamente nel campo 'pompe calore'. Altri 48 record invece hanno il valore vuoto in tutti questi campi. Entrambe queste casistiche non hanno senso dal punto di vista logico, tuttavia si decide per il momento di mantenere questi record perché hanno contenuto informativo negli altri campi, ma eventualmente nella fase di analisi su questi campi booleani verranno esclusi per non distorcere l'analisi.

- `tipo_fam_res_cens`:
  - Il campo 'nil' di tipo `String` presenta i valori che sono la concatenazione dei campi 'nil' e 'id_nil' del dataset `residenti`. Per poter manipolare questi dati agevolmente risulta necessaria una modifica, perciò si decide di rimuovere questo campo e di crearne due nuovi già nel flusso Talend, uno chiamato 'id_nil' che ha come valori la parte finale di questi record (il quale sarà portato al tipo Integer dal momento che anche in `residenti` è di questo tipo) e l'altro chiamato 'nil', contenente i nomi dei quartieri;
  - Nei campi 'eta_capofamiglia' e 'numero_componenti' sono stati rilevati rispettivamente 625 e 635 errori nell'inserimento dei valori; si è indicato nel primo *'80*

anni e pi' e nel secondo *'4 e pi'*. Tali valori sono stati sostituiti rispettivamente con *'80 anni e più'* e *'4 e più'* nella fase di ETL dal momento che sono errori presenti nel file sorgente;

- `tipo_fam_res_anag`: Anche in questo caso è presente un campo `quartiere` che contiene sia il `nil` che l'`id_nil` ad esso associato. Procediamo quindi come per `tipo_fam_res_cens`, dividendo questo campo in due campi distinti;

3.4.1 Correzione flussi Talend dopo l'inclusione delle operazioni di Data cleaning e ottimizzazione dei processi

I flussi Talend descritti nella Sezione 3.3.5, hanno subito necessariamente una variazione a seguito delle considerazioni riportate nel processo di pulizia dei dati. Come si può osservare in Figura 3.8, si è deciso di introdurre un flusso di operazioni che viene eseguito prima del processo effettivo di Insert-Update-Delete, nel quale vengono sostanzialmente operate tutte le modifiche relative alla Data Quality riportate in Sezione 3.4. Il risultato di queste trasformazioni sulla sorgente originale viene salvato su un file temporaneo tramite il componente `tCreateTemporaryFile` e può essere utilizzato nei vari flussi di ETL tramite il componente `tFileOutputDelimited`. Alla fine dell'esecuzione del processo tale file viene quindi eliminato. L'impiego di un file temporaneo, permette di eseguire l'interrogazione della API e la pulizia della sorgente soltanto una volta, riducendo di molto i tempi di esecuzione dei flussi relativi alla Insert-Update e alla Delete.

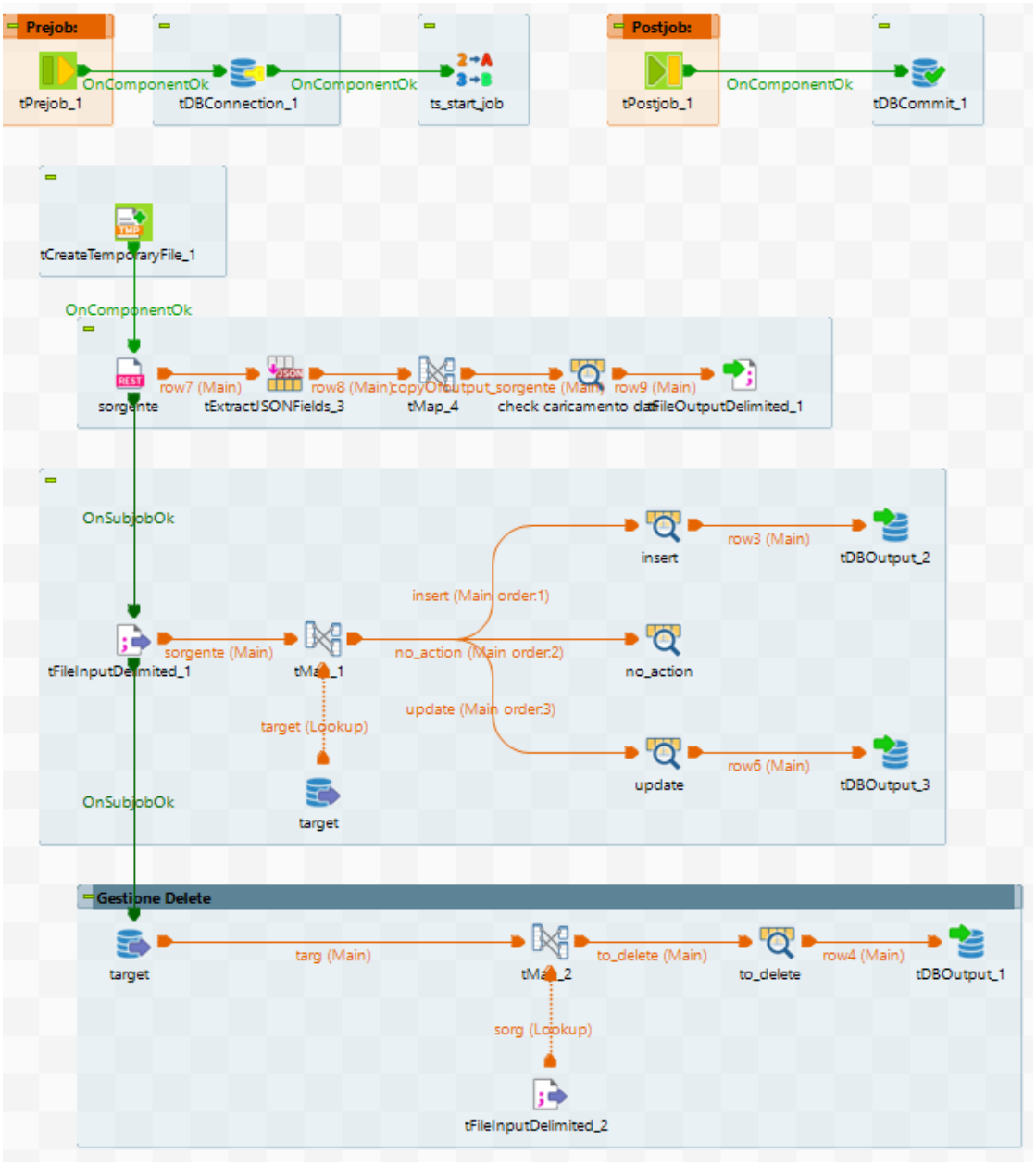

Figura 3.8. Esempio flusso Talend per l'integrazione dei dati ottimizzato, dopo le considerazioni apportate in fase di Data Cleaning.

Capitolo 4

Data Analysis

4.1 Analisi esplorativa

In questa prima fase di analisi si va a comprendere più nel concreto quali siano i dati che si stanno manipolando, quali sono le loro peculiarità e come vanno considerati e inclusi nelle analisi successive. In questa sezione si andrà perciò a fornire una descrizione più precisa dei dataset considerati, per garantire una maggiore comprensione dei dati impiegati e delle motivazioni che porteranno a evenutali esclusioni di alcuni record o attributi, ad esempio. Nell'ambito dell'analisi dei dati, è essenziale garantire l'integrità e la coerenza dei dataset utilizzati. Pertanto, è necessario operare una selezione rigorosa dei dati da includere nello studio. Alcuni record verranno esclusi nelle analisi successive poiché presentano valori non validi o incompleti che potrebbero compromettere l'accuratezza e l'affidabilità delle analisi. Ad esempio, i dati che risultano logicamente incoerenti saranno scartati. Questi possono includere anomalie come valori numerici fuori scala o discrepanze rispetto alle regole logiche o alle aspettative stabilite per l'analisi. Inoltre, sarà data particolare attenzione ai valori `NULL`, ossia campi lasciati vuoti, laddove tali informazioni risultano cruciali. Un caso frequente riguarda il campo 'quartiere' che rappresenta un'informazione fondamentale per analisi demografiche e geografiche. L'assenza di questo dato impedisce un'analisi accurata della distribuzione spaziale dei fenomeni studiati e, di conseguenza, tali record saranno esclusi dall'analisi finale. Si riportano di seguito i record esclusi dalla analisi, suddivisi per dataset:

- In `nascite_2003_2020` si escludono 41 record aventi il campo 'quartiere' vuoto;
- In `nascite_2021_2023` non viene escluso alcun record;
- In `decessi_2003_2020` si escludono 17 record aventi il campo 'quartiere' vuoto;
- In `decessi_2021_2023` non viene escluso alcun record;
- In `residenti` non viene escluso alcun record;
- In `costruzioni` si escludono 85 record aventi il campo 'quartiere' vuoto. Inoltre ci sono delle incongruenze di carattere logico nel blocco di variabili booleane presentate

in precedenza su altri 33 record, relative alla presenza o meno di tecnologie sostenibili. Anche tali record vengono rimossi dall'analisi;

- In `tipo_fam_res_CENS` non viene escluso alcun record;
- In `tipo_fam_res_ANAG` non viene escluso alcun record.

Inoltre, in questo capitolo verrà inserita anche una descrizione del dataset riguardante il movimento naturale e migratorio calcolato da ISTAT, che non è stato integrato nella prima fase poiché è un dataset che non cambia nel tempo, di piccole dimensioni, ma che permette di consegnare delle considerazioni aggiuntive che completano il quadro generale di analisi demografica della popolazione di Milano.

4.1.1 Nascite

Come è stato visto in precedenza, i dati relativi alle nascite sono raccolti in due dataset: `nascite_2003_2020`, associato alle nascite nel periodo temporale che va dal 2003 al 2020; `nascite_2021_2023`, associato al periodo che va dal 2021 al 2023. Dal momento che i due dataset riguardano lo stesso tipo di informazione e oltretutto possiedono la stessa struttura degli attributi, risulta naturale creare una vista materializzata nel database, nella quale viene eseguita l'operazione di UNION tra le due tabelle. La vista materializzata, nel momento in cui viene creata, salva sul database il risultato della query ad essa associata in quel preciso istante, fornendo una "istantanea" della situazione attuale. Se anche le tabelle coinvolte nella vista dovessero aggiornarsi in futuro, la vista materializzata non tiene conto di tali cambiamenti finché non viene eseguito il REFRESH della stessa. Il motivo per cui si crea una vista materializzata piuttosto che una nuova tabella è che, in caso di successivi aggiornamenti dei dati, e quindi di esecuzione dei flussi di ETL visti precedentemente, è necessario eseguire soltanto il REFRESH della vista per ottenere l'aggregazione delle due fonti con i dati aggiornati. Inoltre, invece di creare una semplice vista, si opta per una vista materializzata dal momento che si suppone un contesto in cui si attuano aggiornamenti nei dati distanziati nel tempo, ma nel contempo si vuole utilizzare spesso il risultato della query associato alla vista per le analisi. Ciò significa che, creando una vista semplice, ogni volta che la stessa viene richiamata, la query associata alla vista viene rieseguita e quindi nel caso di operazioni complesse si andrebbe ad appesantire di molto il processo di analisi gravando sui tempi di esecuzione. Andando ad osservare più nel dettaglio la serie storica relativa al numero totale di nascite aggregate per ogni anno, vengono riportate di seguito le statistiche principali (Tabella 4.1). Si noti che per 'conteggio' in questo caso si intende il numero di anni considerati nella serie.

conteggio	media	dev. std.	min	25%	50%	75%	max
21	11.383,62	981,43	9.414	10.831	11.547	12.258	12.518

Tabella 4.1. Statistiche relative alla serie storica delle nascite.

In Figura 4.1 si può notare come l'andamento complessivo della serie abbia trend decrescente: in particolare le nascite nel 2003 sono state 12.258 mentre nel 2023 sono state

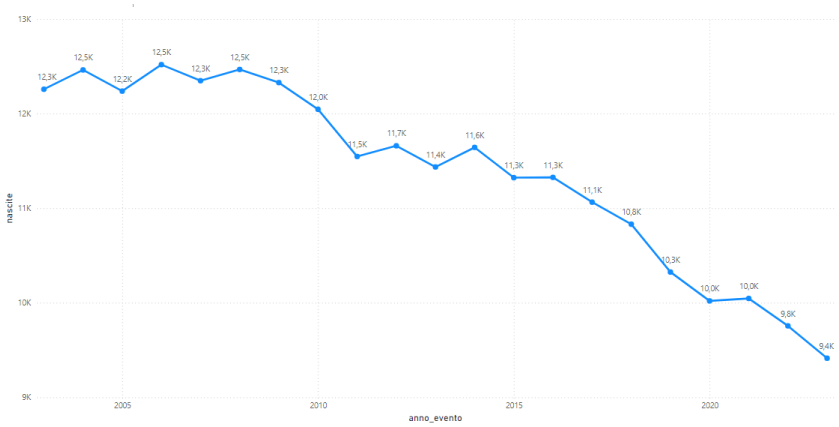

Figura 4.1. Serie storica relativa al numero totale di nascite per ogni anno.

9.414, con una variazione percentuale pari al -23,2%. Si riportano anche le serie storiche delle nascite suddivise per i valori del campo 'cittadinanza_madre' e 'cittadinanza_padre' (Figura 4.2), che sono 'Italiana' e 'Straniera'. In questo plot non sono stati riportati i record che presentano il valore 'n.d.' perché sono in totale rispettivamente 18 e 374 per i due campi, numeri che in relazione al totale dei record che si stanno considerando, non incidono in modo determinante e possono essere trascurati. Si noti che i due campi presentano comportamenti simili confrontando sia le serie per la nazionalità Italiana di madre e padre del nascituro, sia considerando la nazionalità Straniera. Il numero di nascite associate alla nazionalità Straniera del genitore è sempre inferiore a quello relativo alla nazionalità Italiana del genitore. Entrambi queste osservazioni sono spiegate dal fatto che l'89% circa delle nascite considerate in questa analisi ha come genitori o coppie di Italiani o coppie di Stranieri (Tabella 4.2), garantendo quindi una similarità negli andamenti delle serie associate ai due campi 'cittadinanza_madre' e 'cittadinanza_padre'. Inoltre, dal momento che il numero di nascite associate a genitori entrambi Italiani è circa il 65% del totale rispetto al 25% circa associato a genitori entrambi Stranieri, risulta più chiara l'inferiorità dei valori associati alle serie relative a Stranieri rispetto a quelle associate a Italiani. Per le serie filtrate per genere della nascita (Figura 4.3) si ha una prevalenza di nascite di genere maschile in ogni anno considerato nell'analisi e un andamento costante della composizione delle nascite. Per quanto riguarda l'attributo 'quartiere', si riporta in Figura 4.4 l'andamento delle nascite suddiviso per quartiere di residenza del nascituro, per avere un'idea generale. Si può osservare che il quartiere 'Buenos Aires - Porta Venezia - Porta Monforte' presenta il maggior numero di nascite in tutti gli anni seppur abbia un trend decrescente che riflette il comportamento complessivo della serie. Si riportano anche i 5 quartieri che hanno subito la maggiore variazione percentuale in positivo e i 5 quartieri che hanno subito la maggiore variazione in negativo nelle nascite tra il 2003 e il 2023 nella Tabella 4.3. Le variazioni considerate riguardano i primi 50 quartieri di Milano ordinati per numero di nascite del 2003 e i primi 50 quartieri ordinati per numero di nascite del 2023, per poter ottenere dei valori ragionevoli e interessanti, visualizzando

cittadinanza_madre	cittadinanza_padre	nascite	%nascite
Italiana	Italiana	153.764	64,49%
Straniera	Straniera	58.369	24,48%
Straniera	Italiana	16.262	6,82%
Italiana	Straniera	10.032	4,21%

Tabella 4.2. Numero di nascite suddiviso per le combinazioni possibili delle nazionalità dei genitori del nascituro.

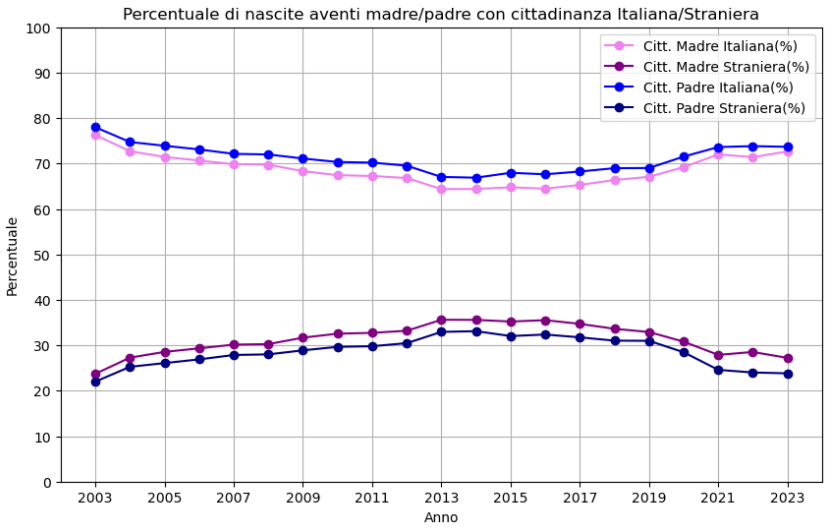

Figura 4.2. Andamenti percentuali delle cittadinanze dei genitori della nascita per ogni anno, per il campo 'cittadinanza_madre' con valori Italiana e Straniera, confrontato con 'cittadinanza_padre' anch'esso con valori Italiana e Straniera.

i quartieri che hanno subito le variazioni più importanti dal 2003 ad oggi. Si è interessati infatti soprattutto ai quartieri che hanno valori elevati in termini di nascite, piuttosto che zone in cui si hanno numeri esigui. Si sottolinea che il NIL 'Maggiore - Musocco - Certosa' dal 2003 ad oggi ha triplicato le nascite, mentre 'Magenta - S.Vittore' le ha sostanzialmente dimezzate. Si può osservare che si hanno variazioni positive nelle nascite in alcuni quartieri, ma confrontando i valori con le variazioni negative si può constatare che queste ultime impattano maggiormente sul totale delle nascite, perché coinvolgono valori maggiori. Ad esempio, tra i peggiori in effetti si rileva il quartiere 'Buenos Aires - Porta Venezia - Porta Monforte' che è anche il quartiere con più nascite a Milano nel periodo 2003-2023.

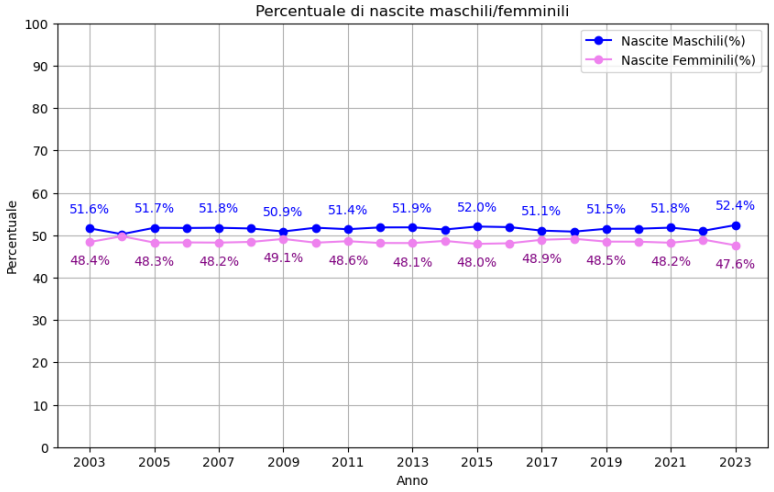

Figura 4.3. Serie storica relativa alla percentuale di nascite per ogni anno suddivisa per il campo 'genere'.

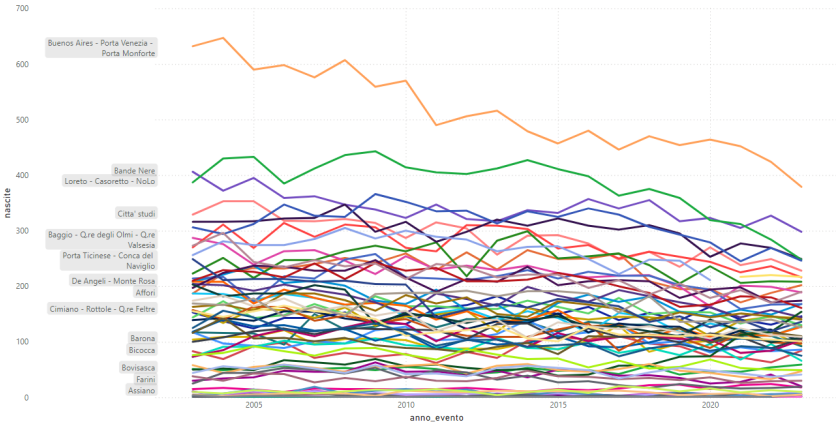

Figura 4.4. Serie storica relativa al numero totale di nascite per ogni anno, suddivisa per il campo 'quartiere'.

Quartiere	Variazione (%)	Nascite 2003	Nascite 2023
Maggiore - Musocco - Certosa	208,00	50	154
Lambrate - Ortica	41,10	73	103
Rogoredo - Santa Giulia	13,83	94	107
Affori	8,02	187	202
Moncucco - San Cristoforo	6,00	100	106
Buenos Aires - Porta Venezia - Porta Monforte	−40,03	632	379
Porta Ticinese - Conca del Naviglio	−45,38	249	136
Ronchetto sul Naviglio - Q.re Lodovico il Moro	−45,97	124	67
Stadio - Ippodromi	−46,22	119	64
Magenta - S.Vittore	−46,80	203	108

Tabella 4.3. Variazioni percentuali nelle nascite per i quartieri più influenti. I primi 5 quartieri sono i migliori per variazione percentuale positiva nei quartieri con più nascite del 2023, mentre gli ultimi 5 sono i peggiori per variazione percentuale negativa nei quartieri con più nascite del 2003.

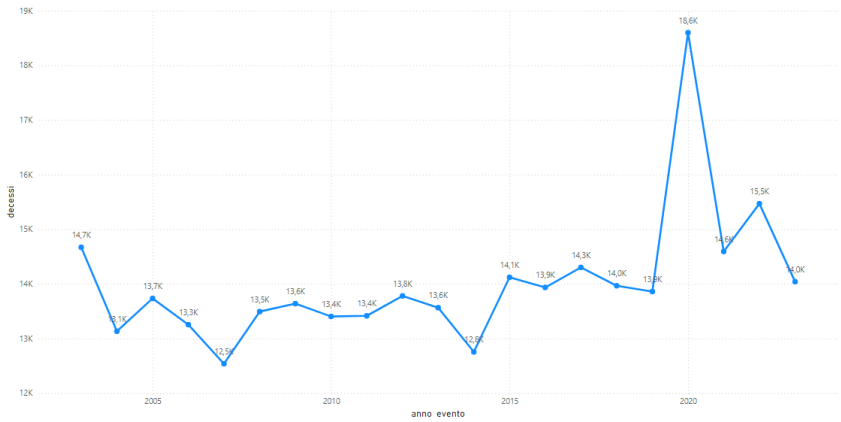

Figura 4.5. Serie storica relativa al numero totale di decessi per ogni anno.

4.1.2 Decessi

Considerando che i dati relativi ai decessi presentano caratteristiche simili a quelli delle nascite, è stata creata una vista materializzata che integra i due dataset `decessi_2003_2020` e `decessi_2021_2023`. In Tabella 4.4 si riportano le statistiche principali della serie.

conteggio	media	dev. std.	min	25%	50%	75%	max
21	14.012,57	1.237,00	12.538	13.416	13.780	14.121	18.597

Tabella 4.4. Statistiche relative alla serie storica dei decessi.

La Figura 4.5 illustra l'andamento generale dei decessi nel corso degli anni. A differenza della serie storica delle nascite, in questo caso non si osserva un trend definito; la serie appare infatti piuttosto regolare, eccezion fatta per il picco registrato nel 2020, attribuibile alla pandemia di COVID-19. Suddividendo i dati in base all'attributo 'cittadinanza_deceduto', si osserva come la maggior parte dei decessi siano relativi a individui di nazionalità Italiana (il 98% circa per ogni anno) (Figura 4.6). La serie storica dei decessi suddivisa per genere presenta una composizione in percentuale costante negli anni (Figura 4.7). Tuttavia, contrariamente alle considerazioni fatte per le nascite, le quali presentavano una costante prevalenza di nascituri maschi negli anni, in questo caso la costante prevalenza dei decessi proviene da soggetti di genere femminile. Anche in questo caso si propone in Figura 4.8 l'andamento dei decessi suddiviso per quartiere, constatando la prevalenza dei decessi proveniente dai quartieri 'Buenos Aires - Porta Venezia - Porta Monforte' e 'Bande Nere'. In Tabella 4.5 si riportano le variazioni percentuali più importanti nei decessi per quartiere tra il 2003 e il 2023. Si riportano nella stessa tabella anche i decessi relativi all'anno 2020, per avere anche delle informazioni sui decessi nell'anno della pandemia, che ha causato il picco evidenziato precedentemente.

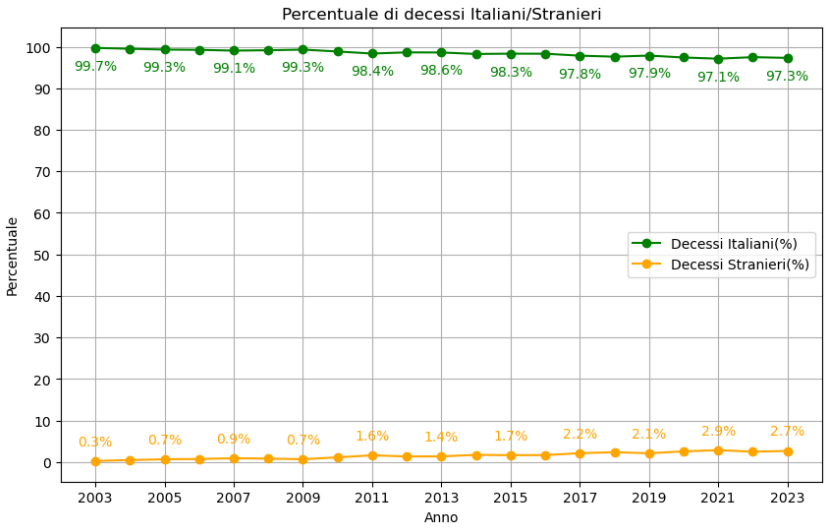

Figura 4.6. Serie storica relativa alla percentuale di decessi per ogni anno, suddivisa per il campo 'cittadinanza_deceduto'.

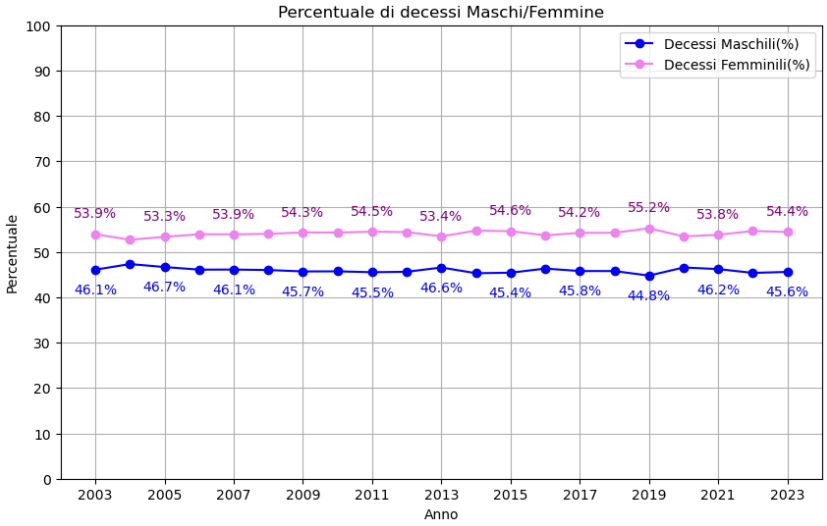

Figura 4.7. Serie storica relativa alla percentuale sul totale di decessi per ogni anno, suddivisa per il campo 'genere'.

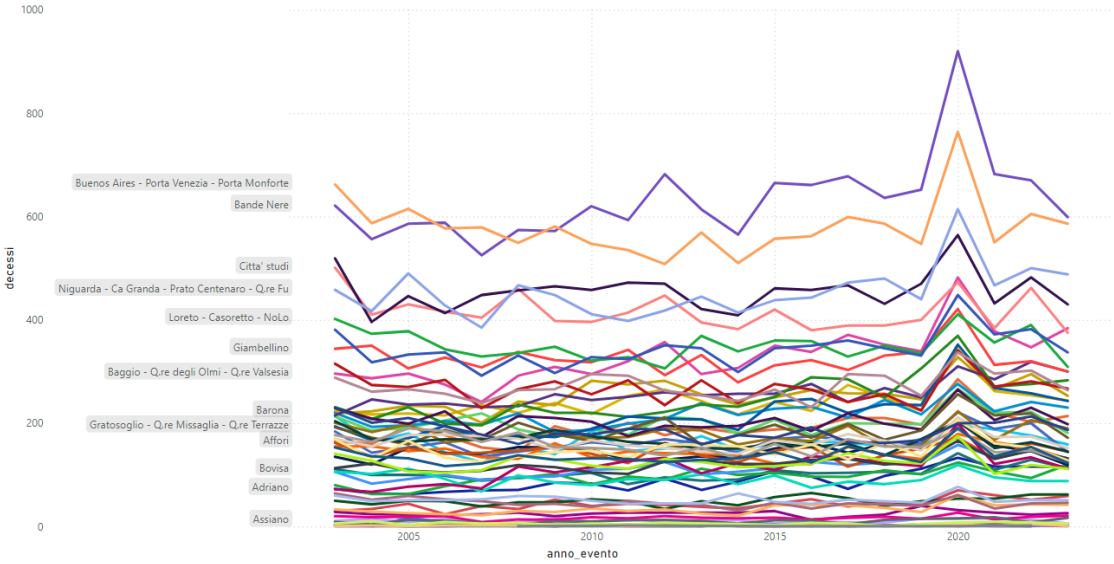

Figura 4.8. Serie storica relativa al numero totale di decessi per ogni anno, suddivisa per il campo 'quartiere'.

Quartiere	Variazione (%)	Decessi 2003	Decessi 2020	Decessi 2023
Bovisasca	50,00	80	124	120
Q.re Gallaratese - Q.re San Leonardo - Lampugnano	41,02	373	673	526
Forze Armate	39,53	215	310	300
Ronchetto sul Naviglio - Q.re Lodovico il Moro	36,17	141	238	192
Vigentino - Q.re Fatima	32,00	100	197	132
Citta' studi	−25,15	501	473	375
Brera	−25,70	214	198	159
Umbria - Molise - Calvairate	−27,82	284	282	205
Magenta - S.Vittore	−28,92	204	204	145
Porta Ticinese - Conchetta	−29,94	177	188	124

Tabella 4.5. Variazioni dei decessi per i quartieri più influenti. I primi 5 quartieri sono i migliori per variazione percentuale positiva nei quartieri con più decessi, mentre gli ultimi 5 sono i peggiori per variazione percentuale negativa.

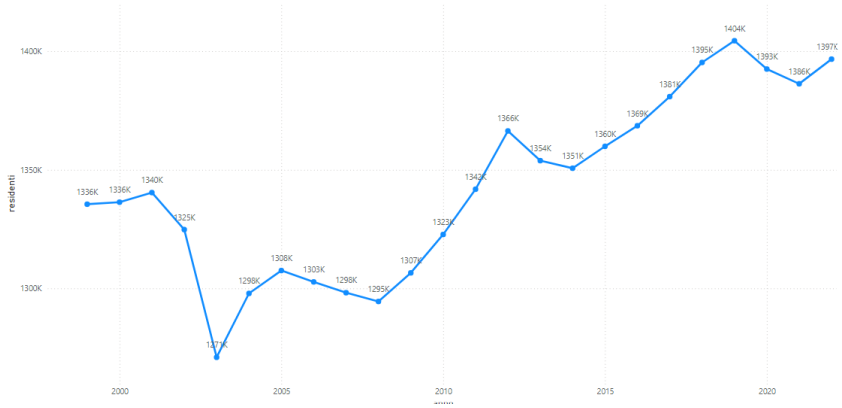

Figura 4.9. Serie storica relativa al numero totale di residenti per ogni anno.

4.1.3 Residenti

Per quanto riguarda il dataset relativo ai residenti si ha che, a differenza dei precedenti relativi alle nascite e ai decessi, in questo caso si hanno informazioni per l'arco temporale che va dal 1999 al 2022, come si può notare dalla serie storica dei residenti totali in Figura 4.9. In Tabella 4.6 ne vengono riportate le statistiche principali.

conteggio	media	dev. std.	min	25%	50%	75%	max
24	1.343.151	38.134,07	1.270.964	1.307.299	1.341.124	1.371.661	1.404.431

Tabella 4.6. Statistiche relative alla serie storica dei residenti.

La serie mostra un forte declino tra 2001 e 2003, mentre dal 2008 in poi si può apprezzare un trend perlopiù crescente con comportamenti altalenanti. Suddividendo la serie per quartiere, si osserva come le zone con il maggior numero di residenti negli anni siano sempre quelle di 'Buenos Aires - Porta Venezia - Porta Monforte', 'Bande Nere' e 'Città Studi'. In particolare per quanto riguarda le variazioni percentuali dei residenti nei quartieri tra il 1999 e il 2022, si può osservare che il quartiere che ha subito la variazione positiva più importante è 'Adriano' che ha incrementato il numero di residenti di circa l'80% rispetto al valore del 1999. D'altra parte il quartiere 'Duomo' ha subito la variazione negativa maggiore, pari a circa il -16% (Tabella 4.7).

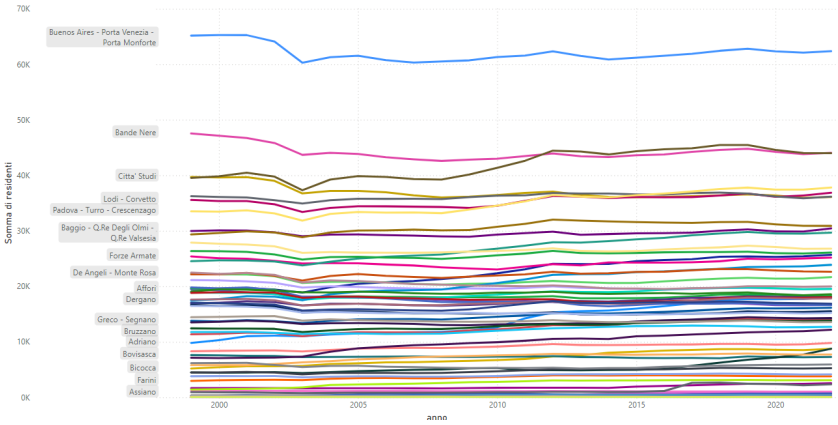

Figura 4.10. Serie storica relativa al numero totale di residenti per ogni anno, suddiviso per quartiere.

Quartiere	Variazione (%)	Residenti 1999	Residenti 2022
Adriano	83,65	9813	18022
Dergano	37,01	17426	23875
Affori	31,62	19502	25669
Bovisa	23,29	11350	13993
Gorla - Precotto	20,97	24524	29666
Citta' Studi	−9,06	39704	36107
Q.Re Gallaratese - Q.Re San Leonardo - Lampugnano	−10,64	36353	32486
Porta Ticinese - Conca Del Naviglio	−11,00	22471	19999
Barona	−14,90	19575	16658
Duomo	−16,19	19817	16608

Tabella 4.7. Variazioni dei residenti per i quartieri più influenti. I primi 5 quartieri sono i migliori per variazione percentuale positiva nei quartieri con più residenti, mentre gli ultimi 5 sono i peggiori per variazione percentuale negativa.

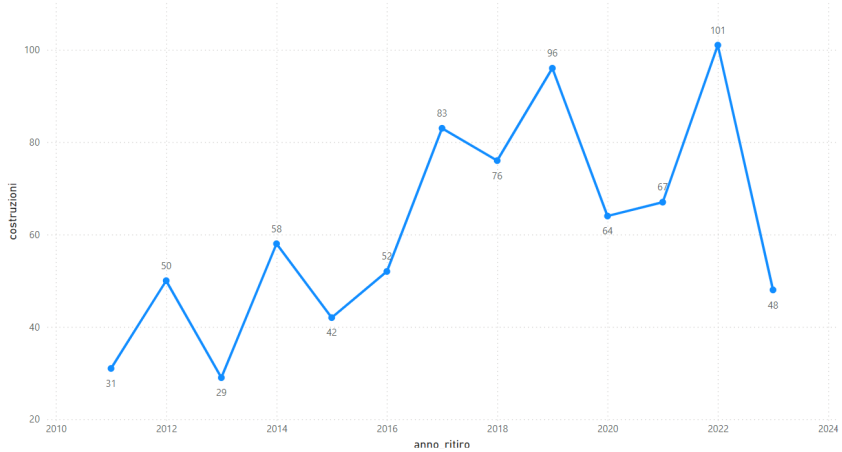

Figura 4.11. Serie storica relativa al numero totale di costruzioni che hanno ottenuto il permesso di costruzione tra il 2011 e il 2023.

4.1.4 Costruzioni

Il dataset relativo alle costruzioni è, tra quelli analizzati, quello che possiede più attributi e l'unico che possiede informazioni sugli edifici costruiti nel periodo 2011-2023. Non presenta una struttura simile al resto dei dati, i quali hanno una naturale interpretazione come serie storiche; in questo caso si hanno le caratteristiche relative al singolo edificio ad uso perlopiù residenziale che è stato costruito. Tuttavia si può pensare di aggregare il conteggio di costruzioni negli anni per avere una prima idea dell'andamento degli edifici realizzati nel periodo considerato (Figura 4.11). L'anno associato ad ogni record è relativo all'anno di ritiro del permesso di costruzione dell'edificio. L'andamento del numero di edifici costruiti negli anni si muove in modo sostanzialmente analogo alle variabili associate alle caratteristiche edilizie dell'edificio stesso, come si può vedere ad esempio in Figura 4.12, che si ritrovano nei campi: Superficie Utile Abitabile, Numero Abitazioni, Numero Stanze, Numero Vani Accessori Interni, Numero Piani, Volume Totale V/p. Ciò può essere osservato anche guardando il grafico di correlazione, costruito sulla base delle variabili numeriche del dataset costruzioni (Figura 4.13). Infatti si può constatare la presenza di una importante correlazione tra le variabili appena citate, che suggerisce un andamento simile negli anni di tali attributi. Per le restanti variabili numeriche non si rilevano particolari osservazioni da apportare in questa fase, se non per la variabile relativa al consumo energetico: prendendo la media per ogni anno, presenta un andamento con trend crescente, indice di un maggiore consumo energetico degli edifici costruiti negli ultimi anni.

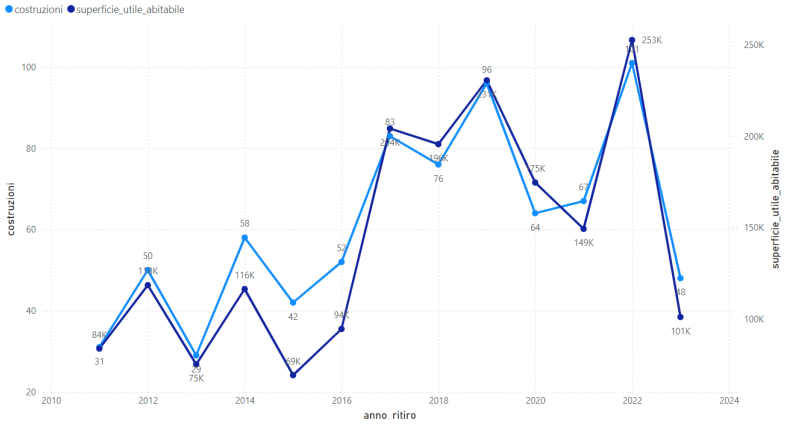

Figura 4.12. Serie storica relativa al numero totale di costruzioni che hanno ottenuto il Permesso di Costruzione tra il 2011 e il 2023 confrontata su scala diversa con il campo superficie_utile_abitabile, per mostrarne il comportamento simile.

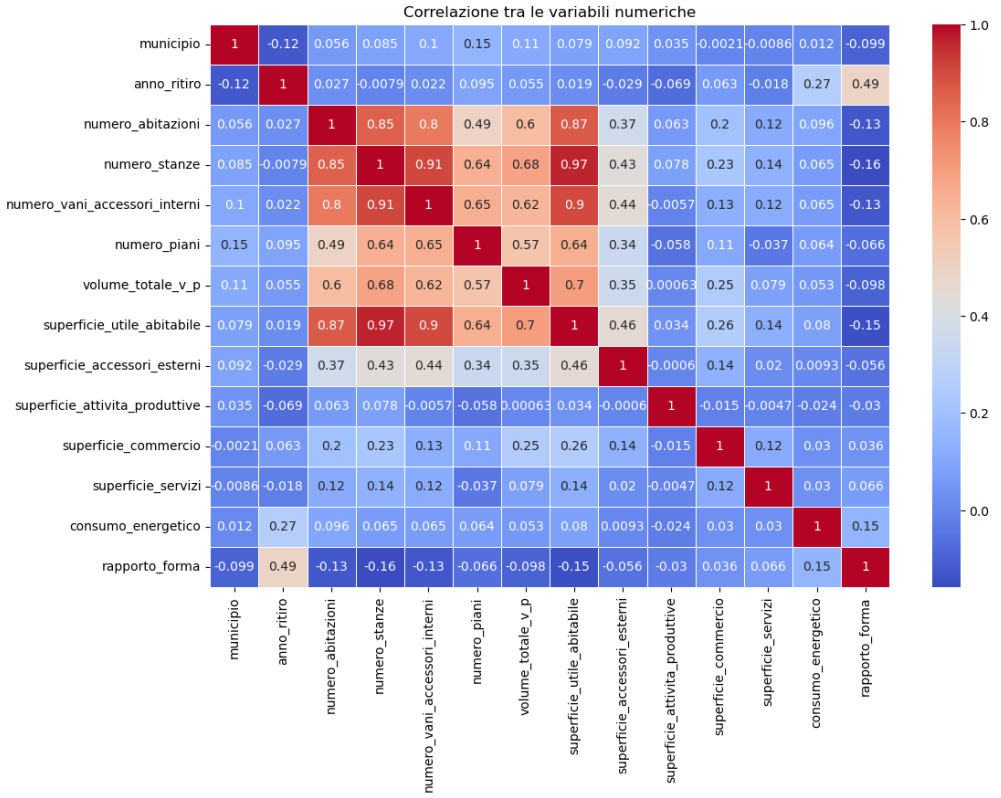

Figura 4.13. Grafico di correlazione associato alle variabili numeriche del dataset `costruzioni`.

Per quanto riguarda le variabili che si possono definire descrittive, ovvero che forniscono informazioni testuali aggiuntive sui record, si danno le seguenti caratterizzazioni:

- descr_titolo_abilitativo: Tale campo presenta i titoli abilitativi edilizi che in Italia regolano la costruzione, la ristrutturazione e la modifica degli edifici. Possiede quattro diversi valori. La DIA (Denuncia di Inizio Attività) permette di iniziare i lavori dopo 30 giorni dalla presentazione al Comune, ed è stata gradualmente sostituita dalla SCIA (Segnalazione Certificata di Inizio Attività), che consente di iniziare i lavori immediatamente dopo la presentazione della documentazione. La SCIA Alternativa permette di avviare lavori complessi che richiederebbero il permesso di costruire, con tempi più rapidi rispetto a quest'ultimo. Infine, il Permesso di Costruire è obbligatorio per interventi edilizi rilevanti come nuove costruzioni o ampliamenti volumetrici e richiede l'approvazione esplicita del Comune, con tempistiche di istruttoria più lunghe. In Figura 4.14 si nota che dal 2013 la DIA non è stata più utilizzata, così come la SCIA dal 2021. La prevalenza delle costruzioni risulta essere con Permesso di costruire o con SCIA alternativa.

- specifica_altra_attivita: è un campo descrittivo, che non possiede informazioni rilevanti per l'analisi e non viene considerato, presenta per la quasi totalità dei record il campo vuoto.

- descr_titolare_titolo_costruire: è un campo con diversi livelli che fornisce informazioni su chi è legalmente responsabile del progetto di costruzione o ristrutturazione. I valori presenti sono "cooperativa edilizia", "ente pubblico", "impresa", "impresa/società", "persona fisica", "altro" e indicano la tipologia giuridica o organizzativa del soggetto che ha ottenuto il titolo abilitativo (ad esempio un permesso di costruire). In Figura 4.15 si evidenzia negli ultimi anni l'aumento di costruzioni associate a "Impresa / Società".

- piano_casa: anche in questo caso è valorizzato con campo vuoto per la quasi totalità dei record. Per i campi in cui è valorizzato presenta una tra le seguenti due normative che fanno parte di un insieme di misure adottate per affrontare problematiche relative all'edilizia abitativa in Italia. L'"Accordo Stato-Regioni del 2009" facilita l'ampliamento o la demolizione e ricostruzione di edifici, per migliorare la qualità abitativa e l'efficienza energetica. Il "Piano nazionale di edilizia abitativa (Dl. 112/2008 e delibera CIPE 2009)", invece, mira a risolvere l'emergenza abitativa costruendo alloggi a prezzi accessibili per categorie vulnerabili, come giovani coppie e famiglie a basso reddito. Anche questo campo non risulta essere di interesse per l'analisi.

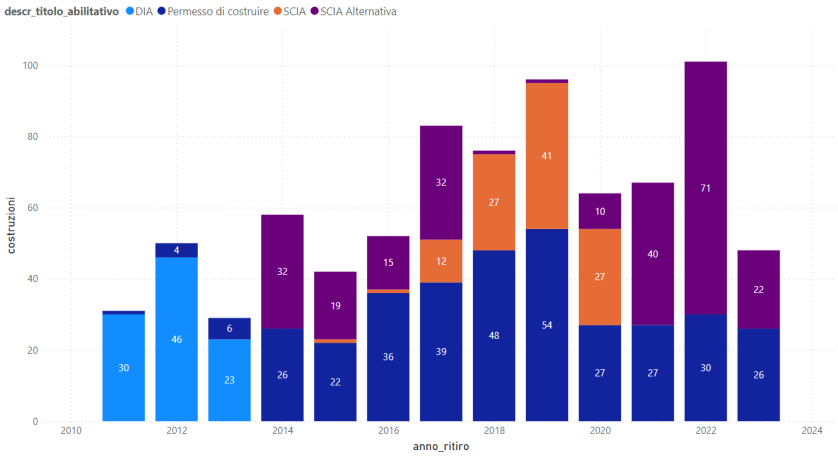

Figura 4.14. Andamento della variabile descr_titolo_abilitativo.

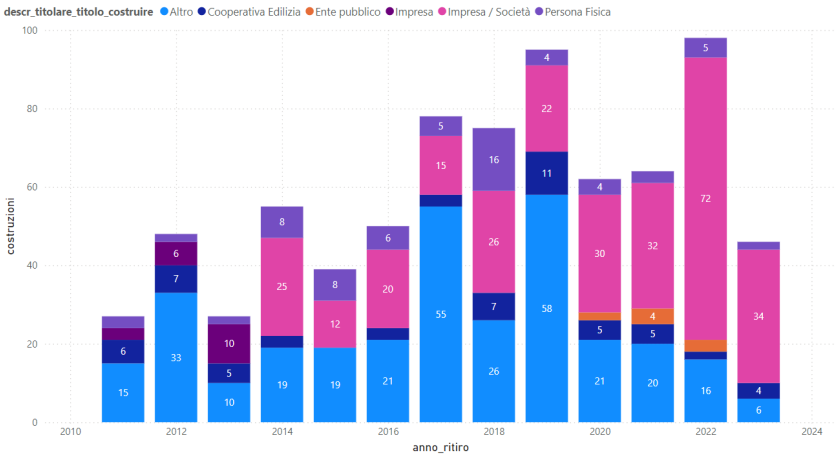

Figura 4.15. Andamento della variabile descr_titolare_titolo_costruire.

Infine si riporta anche una descrizione delle variabili booleane, relative alla presenza o meno di tecnologie sostenibili nelle costruzioni oggetto del dataset:

- Fotovoltaico: la presenza di impianti fotovoltaici nelle strutture permette di trasformare l'energia solare in elettricità, riducendo la dipendenza dalle fonti fossili e abbattendo le emissioni di CO2. Si può osservare in Figura 4.16 come dal 2011 sia incrementata notevolmente la percentuale di costruzioni con impianto fotovoltaico per anno e dal 2018 si apprezza ogni anno almeno l'80% delle costruzioni ad uso prevalentemente abitativo che adottano questa tecnologia.

- Solare Termico: consente l'utilizzo dell'energia solare per riscaldare l'acqua, riducendo il consumo di combustibili fossili e le emissioni. In questo caso non si osserva un particolare trend nei dati ma si mantiene perlopiù costante la percentuale di costruzioni con solare termico, che è nettamente inferiore a quella presentatasi nel fotovoltaico (Figura 4.17).

- Pompe Calore: trasferiscono calore dall'aria, dall'acqua o dal suolo per riscaldare o raffreddare edifici, riducendo l'uso di gas o elettricità tradizionali. Questa variabile sottolinea un impiego importante negli edifici analizzati di pompe di calore negli anni (Figura 4.18), dal 2017 in poi almeno il 70% degli edifici ogni anno comprendono le pompe di calore.

- Caldaia A Condensazione: recupera il calore dai gas di scarico che, nelle caldaie tradizionali, andrebbe disperso, aumentando l'efficienza energetica e riducendo le emissioni. Dal 2011 a oggi si è osservata una drastica diminuzione nella percentuale di edifici che adottano questa tecnologia, come si nota in Figura 4.19.

- Geotermico: sfrutta il calore presente nel sottosuolo per riscaldare edifici e acqua, abbassando l'uso di energia fossile e riducendo l'impatto ambientale. Si può dire che il geotermico venga impiegato nelle costruzioni considerate in questa analisi, e la percentuale di inclusione nei progetti ogni anno rimane costante (Figura 4.20), ma sempre sotto al 50% delle costruzioni.

Per approfondire la relazione tra queste variabili booleane, sono state prese in considerazione le tabelle di contingenza relative alle coppie di valori presenti nelle variabili booleane: 'SI-SI', 'NO-NO', 'SI-NO, 'NO-SI' (Figura 4.21). Si osservi che nelle diagonali delle tabelle 'SI-SI' e 'NO-NO' si hanno le percentuali di costruzioni in cui è presente la tecnologia sostenibile corrispondente, mentre nelle diagonali delle restanti tabelle la percentuale vale 0, poiché in nessun caso può accadere che una costruzione possegga e non possegga una data tecnologia. Nelle restanti celle la percentuale rappresenta la frequenza di volte, rispetto al totale, in cui una costruzione presenta quella combinazione di valori relativi alle variabili booleane. Da queste tabelle si possono ricavare le seguenti considerazioni:

- Nella tabella in alto a sinistra in Figura 4.21, si può apprezzare un'alta frequenza di costruzioni che presentano contemporaneamente fotovoltaico e pompe calore (60,9%), che confrontato con la frequenza di costruzioni con fotovoltaico o con pompe calore (76,0% e 74,3% rispettivamente), risulta essere un valore d'interesse ed

indica che nella maggior parte dei casi dove viene inserito il fotovoltaico vengono anche inserite pompe di calore. In effetti oggigiorno è una scelta comune associare all'impianto fotovoltaico la pompa di calore per rendere la costruzione autosufficiente dal punto di vista energetico;

- D'altra parte invece si osserva che le tecnologie meno impiegate sono solare, caldaia a condensazione e geotermico, come si nota nella tabella in alto a destra, e anche in questo caso si osserva che nella maggior parte degli edifici non sono presenti contemporaneamente: solare e caldaia a condensazione (65,3%), solare e geotermico (50,1%), caldaia a condensazione e geotermico (53,1%).

Dal momento che si ha un numero adeguato di record, si può anche eseguire il test χ^2 di Pearson, in questo caso chiamato anche test di indipendenza, per valutare l'indipendenza tra le coppie di variabili booleane. Tale test ha per ipotesi nulla l'indipendenza delle variabili coinvolte che significa nella pratica che una variabile non influenzi l'altra. Supponiamo di avere due variabili X e Y, entrambe binarie, che possono assumere valori 0 o 1, e i dati raccolti siano organizzati in una tabella di contingenza 2×2:

	$Y = 1$	$Y = 0$	Totale
$X = 1$	n_{11}	n_{10}	$n_{1\cdot}$
$X = 0$	n_{01}	n_{00}	$n_{0\cdot}$
Totale	$n_{\cdot 1}$	$n_{\cdot 0}$	n

dove:

- n_{11} è il numero di osservazioni in cui $X = 1$ e $Y = 1$;
- n_{10} è il numero di osservazioni in cui $X = 1$ e $Y = 0$;
- n_{01} è il numero di osservazioni in cui $X = 0$ e $Y = 1$;
- n_{00} è il numero di osservazioni in cui $X = 0$ e $Y = 0$;
- $n_{1\cdot}$, $n_{0\cdot}$, $n_{\cdot 1}$, e $n_{\cdot 0}$ sono i totali marginali.

Il test di indipendenza χ^2 di Pearson confronta le frequenze osservate nel campione con le frequenze attese sotto l'ipotesi nulla di indipendenza. Le frequenze attese, supponendo l'indipendenza tra le due variabili, sono calcolate come:

$$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

per $i, j \in \{0, 1\}$, cioè il prodotto delle frequenze relative marginali. Pertanto, le frequenze attese nella tabella di contingenza sono:

	$Y = 1$	$Y = 0$	Totale
$X = 1$	$E_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n}$	$E_{10} = \frac{n_{1\cdot} \cdot n_{\cdot 0}}{n}$	$n_{1\cdot}$
$X = 0$	$E_{01} = \frac{n_{0\cdot} \cdot n_{\cdot 1}}{n}$	$E_{00} = \frac{n_{0\cdot} \cdot n_{\cdot 0}}{n}$	$n_{0\cdot}$
Totale	$n_{\cdot 1}$	$n_{\cdot 0}$	n

Il valore della statistica χ^2, nel caso di variabili booleane, si calcola come:

$$\chi^2 = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

cioè, esplicitamente:

$$\chi^2 = \frac{(n_{11} - E_{11})^2}{E_{11}} + \frac{(n_{10} - E_{10})^2}{E_{10}} + \frac{(n_{01} - E_{01})^2}{E_{01}} + \frac{(n_{00} - E_{00})^2}{E_{00}},$$

dove:

- n_{ij} sono le frequenze osservate,
- E_{ij} sono le frequenze attese.

Il valore χ^2 calcolato segue approssimativamente una distribuzione χ^2 con un grado di libertà, dato che la tabella di contingenza è 2×2. Una volta calcolata la statistica χ^2, si può procedere alla seguente valutazione. Calcoliamo il *p-value* associato al valore osservato di χ^2. Il *p-value* rappresenta la probabilità di ottenere un valore della statistica test estremo almeno quanto quello osservato, assumendo che l'ipotesi nulla sia vera. Un p-value basso indica che è improbabile osservare tali differenze se le variabili fossero effettivamente indipendenti.

- Se il p-value è minore del livello di significatività α (tipicamente $\alpha = 0.05$), oppure se il valore di χ^2 osservato è maggiore del valore critico della distribuzione χ^2 per il livello di significatività scelto, allora rifiutiamo l'ipotesi nulla. In questo caso, concludiamo che le variabili X e Y non sono indipendenti, ossia esiste una relazione significativa tra di esse.
- Se invece il p-value è maggiore di α o il valore di χ^2 osservato è minore del valore critico, non possiamo rifiutare l'ipotesi nulla di indipendenza e quindi non ci sono prove sufficienti per affermare che X e Y siano dipendenti.

Osservando la Figura 4.22 dove vengono riportati i p-value ottenuti dai test sulle coppie di variabili, si nota che, solo per la coppia di variabili fotovoltaico e solare termico si ottiene un alto p-value, maggiore della soglia classica di 0,05. Per questa coppia di variabili quindi non si può rifiutare l'ipotesi nulla di indipendenza e quindi non si ha evidenza statistica che tra le due variabili sussista una dipendenza. D'altra parte, le restanti coppie di variabili restituiscono p-value molto bassi, suggerendo quindi una sottostante dipendenza tra queste variabili booleane. Si noti che questo test fornisce solo un livello di significatività della relazione che sussiste tra le variabili e il valore dei p-value potrebbero essere bassi a causa della dimensione piuttosto elevata del campione. Per tale motivo spesso viene anche valutata la statistica *V di Cramér* [16], che è una misura di intensità dell'associazione tra due variabili categoriche, la quale consente di valutare la forza della relazione. Si calcola a partire dalla statistica del test χ^2 ottenuta dalla tabella di contingenza delle due variabili. La formula per calcolare la V di Cramér è la seguente:

$$V = \sqrt{\frac{\chi^2}{N \cdot (k-1)}} \tag{4.1}$$

dove:

- χ^2 è il valore della statistica del chi-quadrato calcolata per il test di indipendenza;
- N è la dimensione del campione (ossia il numero totale di osservazioni nella tabella di contingenza);
- k è il numero di categorie della variabile con il minor numero di livelli, nel caso di variabili booleane $k = 2$.

Il valore di V varia da 0 a 1: $V \approx 0$ indica assenza di associazione tra le variabili, ossia quasi completa indipendenza, mentre $V \approx 1$ indica una quasi completa dipendenza tra le variabili. Per interpretare i valori di V, si utilizzano generalmente le seguenti soglie empiriche:

- $V < 0.2$: associazione molto debole,
- $0.2 \leq V < 0.4$: associazione debole,
- $0.4 \leq V < 0.6$: associazione moderata,
- $V \geq 0.6$: associazione forte.

Tali soglie possono variare in base al contesto e alla disciplina di studio, ma forniscono un'indicazione di massima per interpretare la forza della relazione. È importante notare che, a differenza del p-value del test χ^2, la V di Cramér non è influenzata dalla dimensione del campione, il che la rende particolarmente utile nei casi di campioni di grandi dimensioni, dove il test del chi-quadrato potrebbe risultare significativo anche in presenza di una relazione debole. I risultati del calcolo della V di Cramér per le variabili booleane sono presentati in Figura 4.23, dove si nota in effetti che la relazione che sussiste tra le variabili sembra non avere un alto grado di intensità. Si evidenzia, tuttavia, che il valore più alto consegnato dalla statistica è quello relativo alla coppia fotovoltaico e pompe calore, che, come era stato visto in precedenza, presenta un alto numero di concorrenze nelle costruzioni.

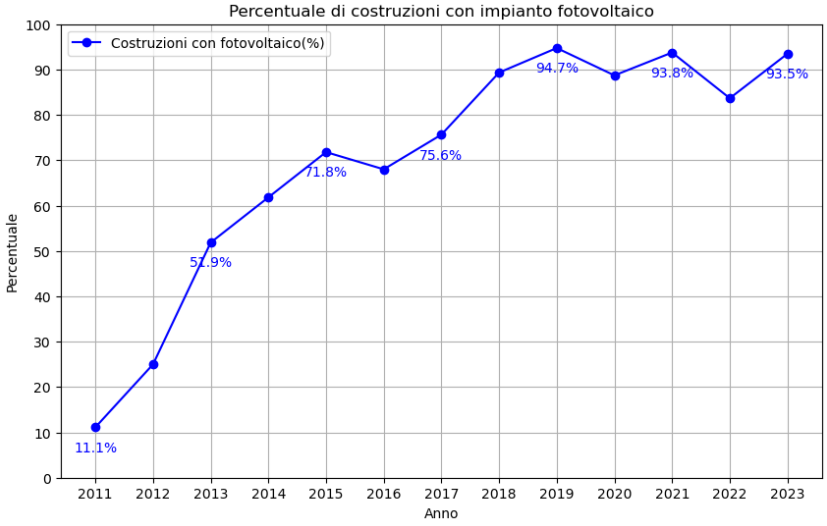

Figura 4.16. Numero di costruzioni con impianto fotovoltaico negli anni.

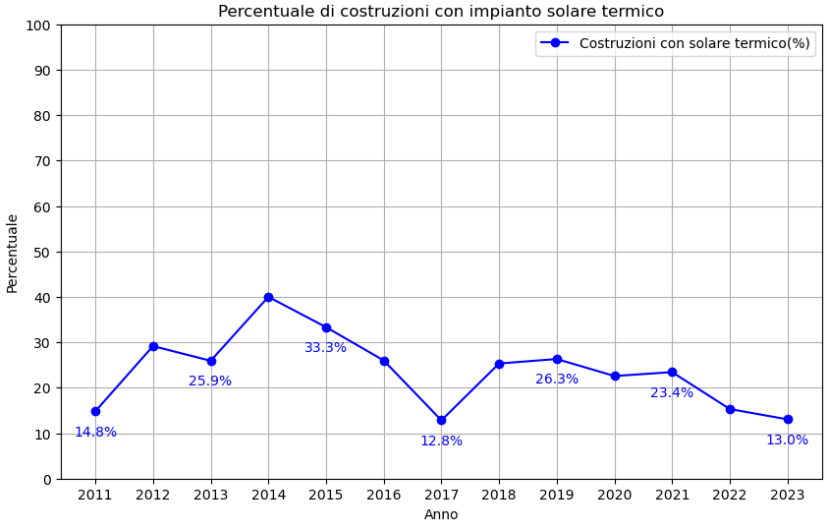

Figura 4.17. Numero di costruzioni con impianto solare termico negli anni.

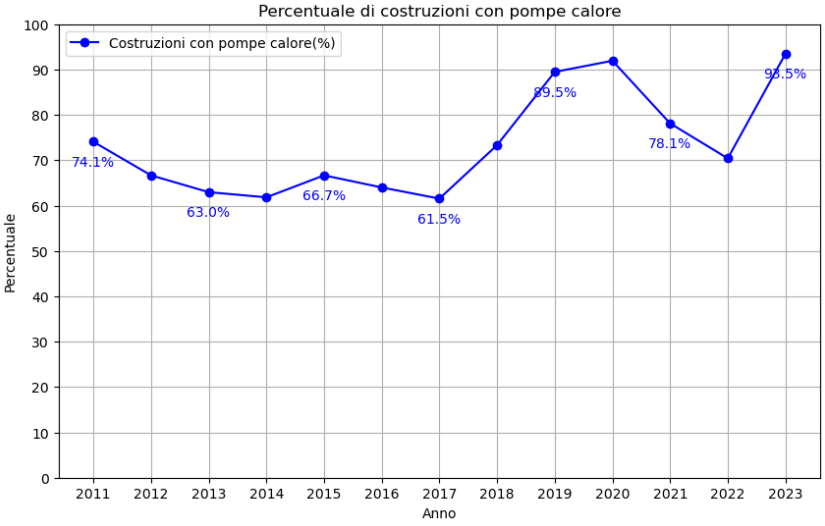

Figura 4.18. Numero di costruzioni con pompe calore negli anni.

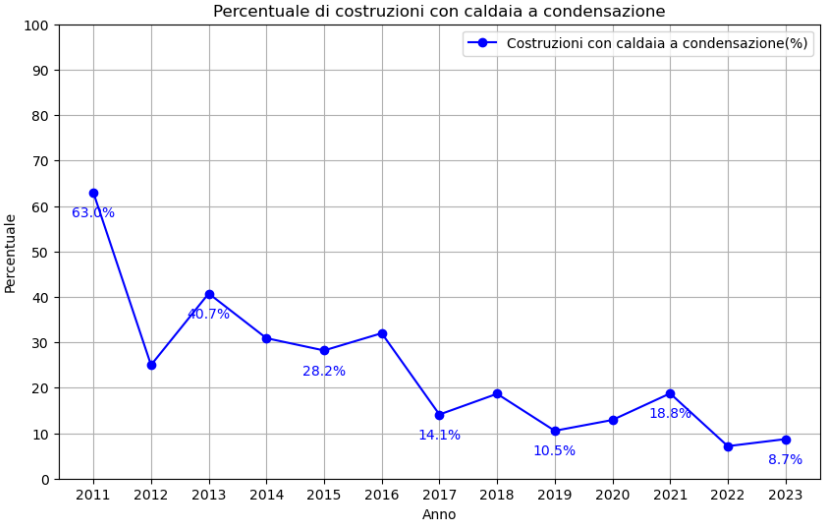

Figura 4.19. Numero di costruzioni con caldaia a condensazione negli anni.

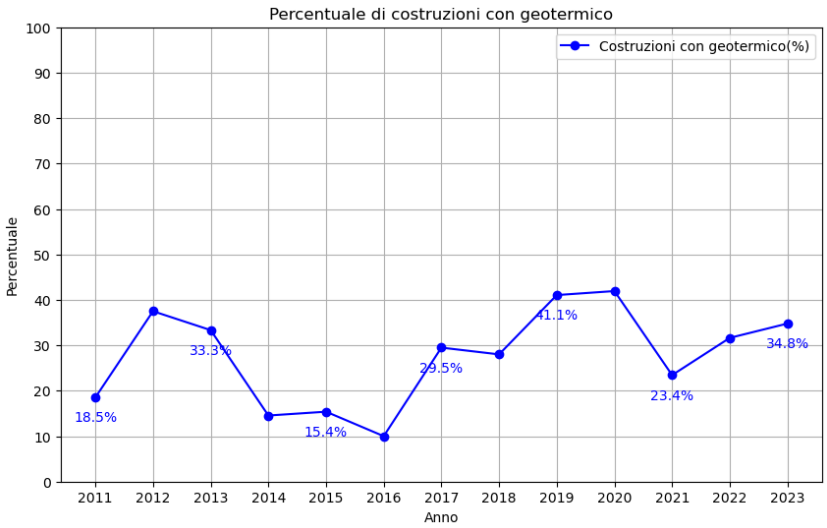

Figura 4.20. Numero di costruzioni con geotermico negli anni.

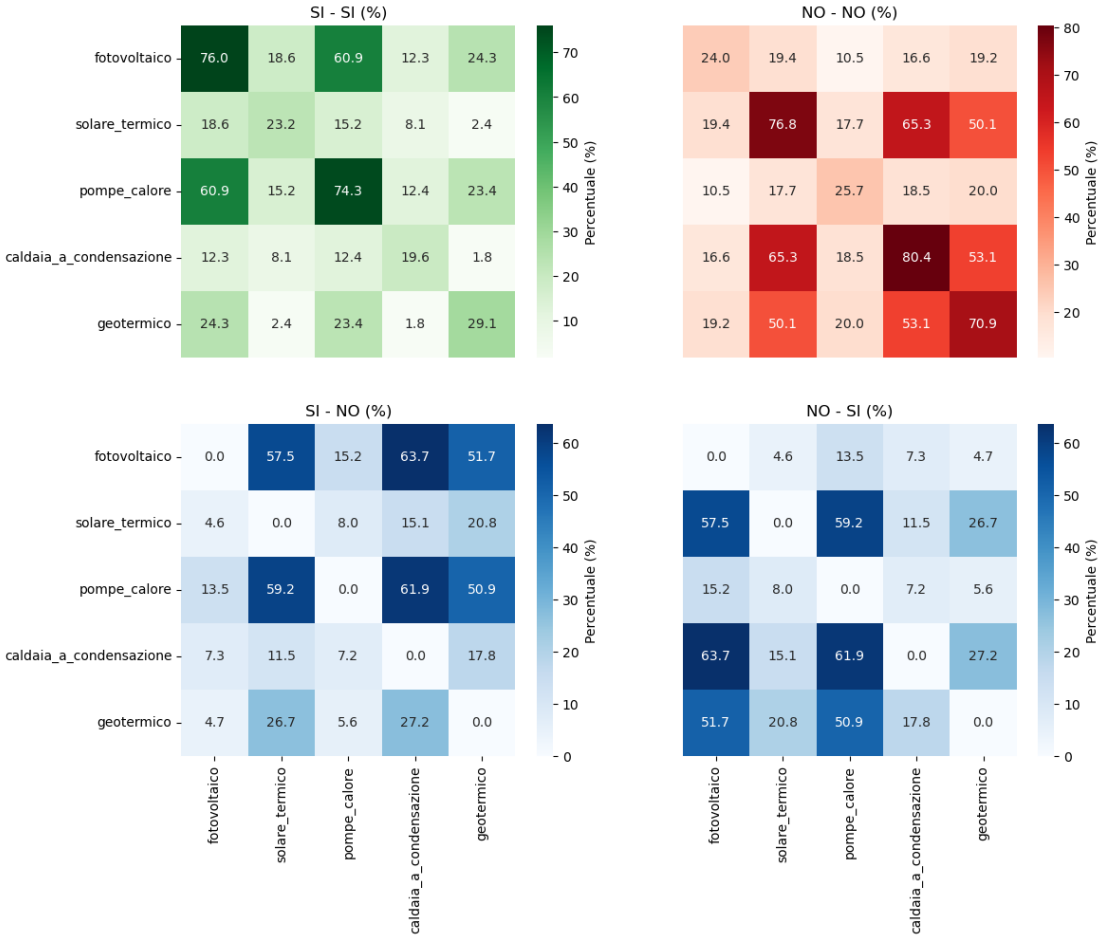

Figura 4.21. Tabelle di contingenza relative alle variabili booleane del dataset `costruzioni`. Nella generica cella (i, j) della tabella 'SI-NO', ad esempio, si legge la percentuale delle costruzioni in cui è presente la tecnologia i-esima ma non la j-esima.

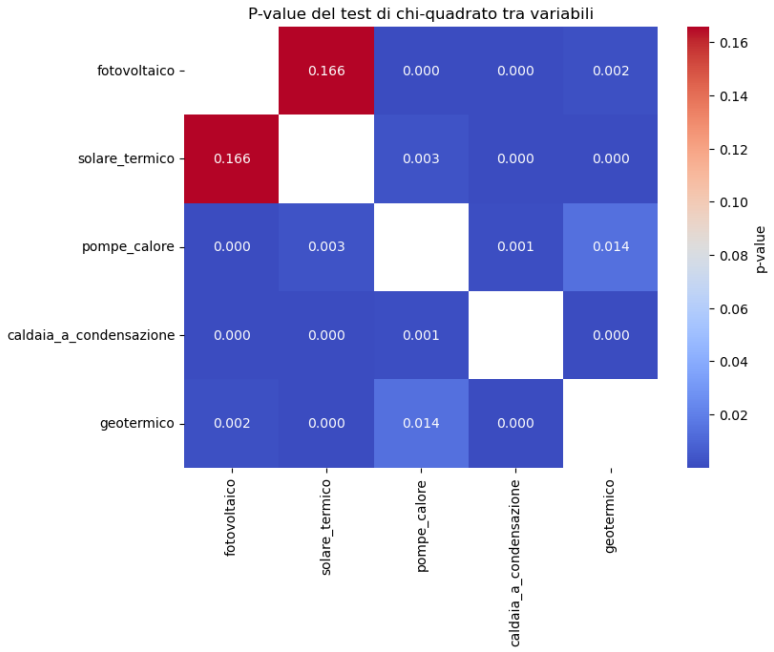

Figura 4.22. P-value associati al Test di indipendenza tra le variabili booleane.

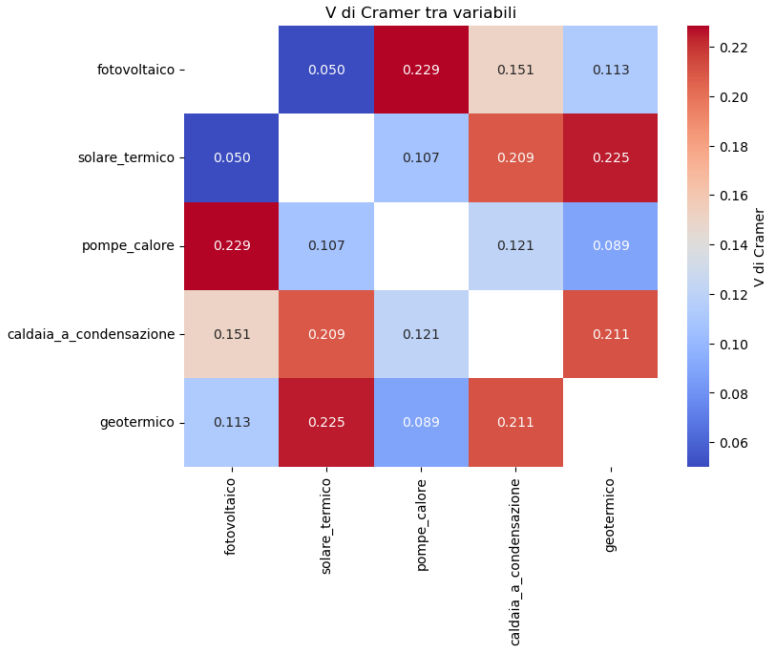

Figura 4.23. Statistiche V di Cramér associate al Test di indipendenza tra le variabili booleane.

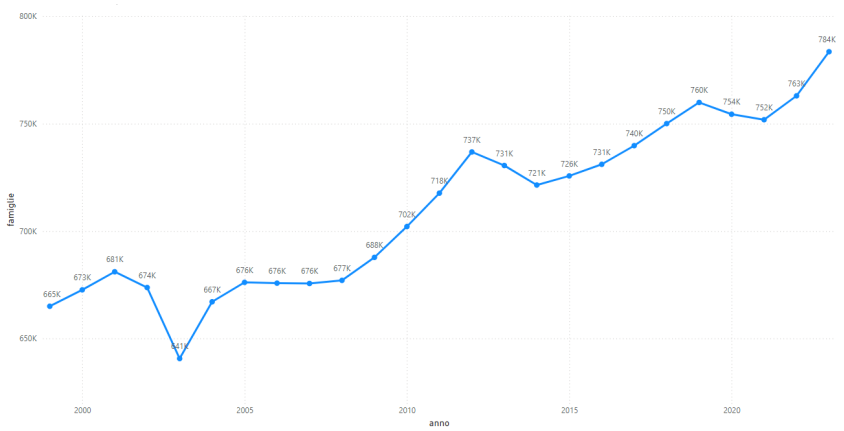

Figura 4.24. Serie storica relativa al numero totale di nuclei familiari presenti a Milano.

4.1.5 Dati anagrafici

Il dataset `tipo_fam_res_ANAG` presenta i dati anagrafici della popolazione di Milano dal 1999 al 2023 suddivisi per nuclei familiari, quindi la metrica di conteggio non è associata al singolo individuo come in nascite, decessi e residenti.

conteggio	media	dev. std.	min	25%	50%	75%	max
25	710.414,72	38.887,99	640.665	675.765	717.671	739.778	783.506

Tabella 4.8. Statistiche relative alla serie storica relativa al dataset `tipo_fam_res_ANAG`.

In Figura 4.24 si può osservare come la serie relativa ai dati anagrafici abbia un comportamento analogo a quello presentato dalla serie storica dei residenti, seppure le metriche di misurazione siano diverse. Suddividendo i dati secondo i valori dell'attributo 'genere_capofamiglia' (Figura 4.25) si ha in generale che i nuclei familiari che hanno come capo famiglia una femmina sono sempre in percentuale inferiore rispetto a quelli con capofamiglia maschio. Si osserva tuttavia un trend crescente della prima serie, mentre un comportamento più stabile nella seconda, evidenziando una sempre minore differenza tra il numero di famiglie con capofamiglia femmina o maschio nel comune di Milano. Anche l'attributo relativo all'età del capofamiglia permette di osservare una netta prevalenza di capofamiglia con età compresa tra i 35 e i 64 anni, seguita dalla fascia di età '65-79 anni', quindi 'meno di 35 anni' e infine '80 anni e più' (Figura 4.26). Per quanto riguarda il numero di componenti appartenenti al nucleo familiare si nota in Figura 4.27 che la maggior parte dei nuclei familiari sin dal 1999 sono monocomponenti seguito da 2, 3, 4, 5 e più componenti in ordine. E' evidente il trend crescente della serie relativa alla percentuale del numero di famiglie composte da un singolo individuo, che dal 1999 al 2023 è passata dal 45% circa a quasi il 60% sul totale dei nuclei familiari. Per quanto riguarda la suddivisione per quartieri si osserva sostanzialmente un comportamento analogo con residenti (Figura

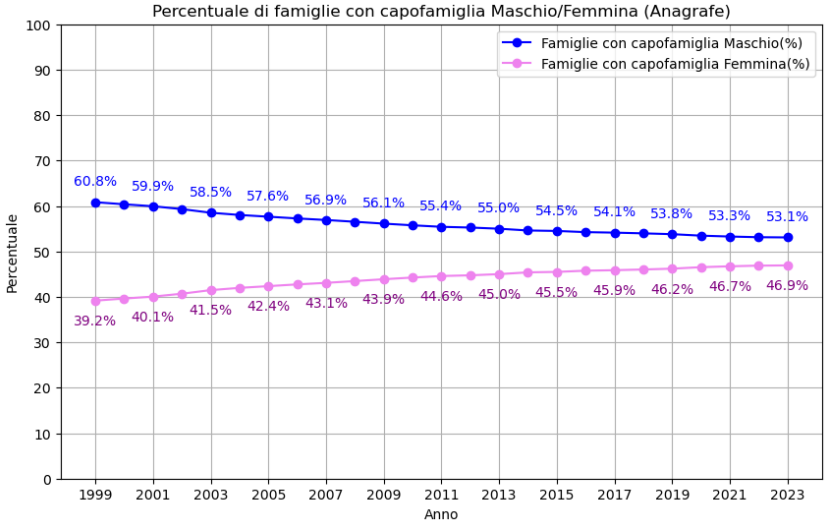

Figura 4.25. Serie storica relativa al numero totale di nuclei familiari presenti a Milano, suddivisa per 'genere_capofamiglia'.

4.28). In Tabella 4.9 si riportano anche le variazioni percentuali del numero di nuclei familiari per quartiere tra il 1999 e il 2023 e si osservano sostanzialmente risultati simili a quelli mostrati per la serie storica dei residenti. La variabile 'cittadinanza' possiede un elevato numero di valorizzazioni ed è difficile individuare una modalità adeguata per descriverla efficacemente. Tuttavia si possono andare ad osservare le prime cittadinanze per numero di occorrenze sul totale, oltre a quella Italiana che rappresenta l'83% circa del totale delle famiglie incluse nel dataset, che sono: l'Egitto (2,30%), le Filippine (2,25%), la Cina (1,38%), il Perù (1,21%), lo Sri Lanka (0,94%). Si riportano le serie storiche relative alle famiglie con queste cittadinanze in Figura 4.29. Per quanto riguarda l'attributo cittadinanza si possono consegnare delle considerazioni anche in merito alle variazioni negli anni delle cittadinanze dei nuclei familiari. In Tabella 4.10 si riportano le variazioni percentuali tra il 1999 e il 2023, del numero di nuclei familiari per le cittadinanze più rilevanti nel contesto sociale di Milano. Quindi si possono osservare le cittadinanze più presenti nel 1999 e nel 2023 e valutarne la variazione, prendendo quelle più significative. In questo caso si hanno variazioni perlopiù positive, eccezion fatta per il Regno Unito. Valutando le variazioni invece nello scorso decennio (2013-2023) si ottengono i risultati in Tabella 4.11, dove si osserva un importante incremento nelle cittadinanze Bangladesh e Cinese, mentre delle decrescite per Perù e Ecuador.

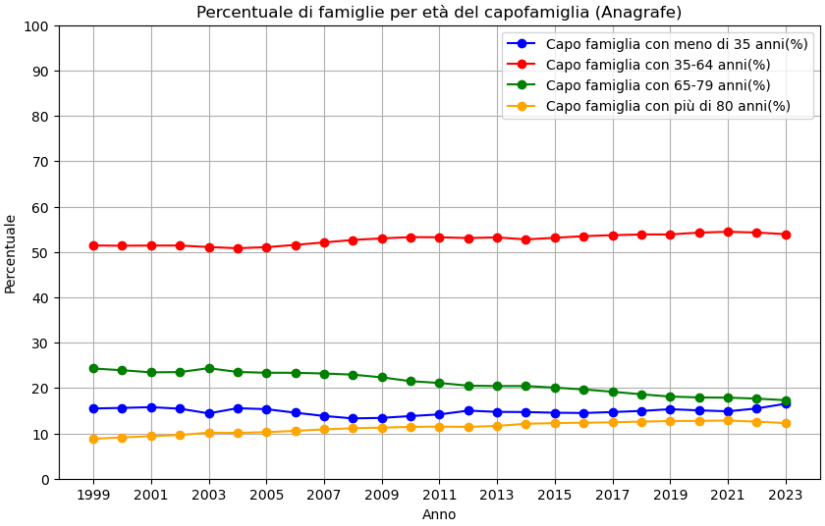

Figura 4.26. Serie storica relativa al numero totale di nuclei familiari presenti a Milano, suddivisa per 'classe_eta_capofamiglia'.

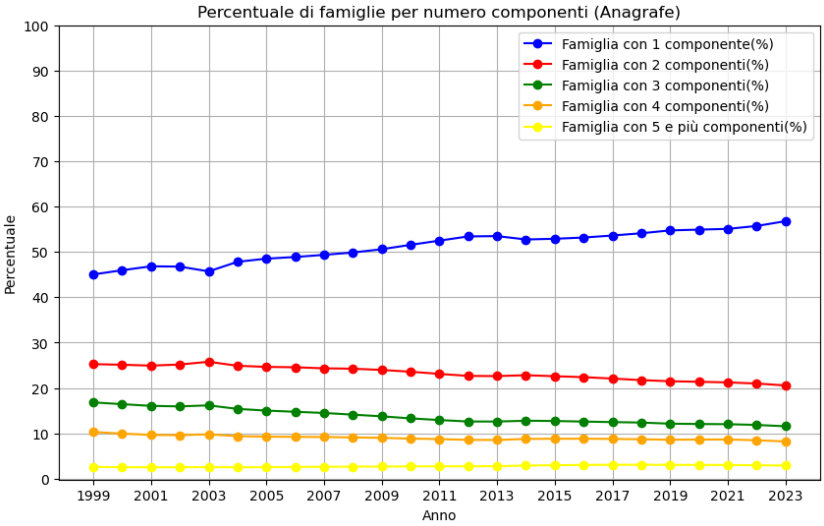

Figura 4.27. Serie storica relativa al numero totale di nuclei familiari presenti a Milano, suddivisa per 'numero_componenti'.

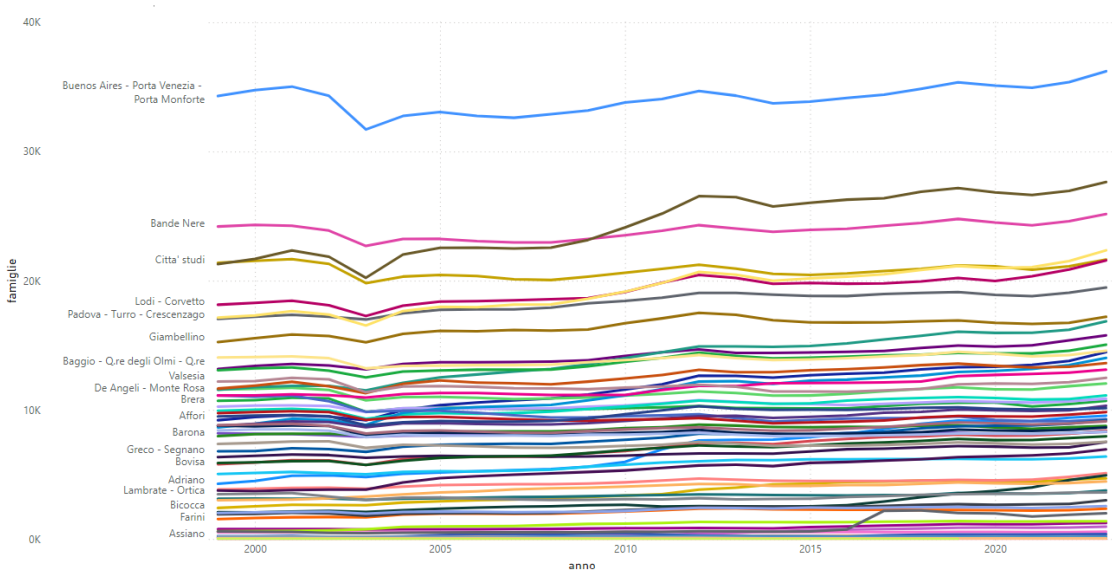

Figura 4.28. Serie storica relativa al numero totale di nuclei familiari presenti a Milano, suddivisa per quartiere.

Quartiere	Variazione (%)	Nuclei Familiari 1999	Nuclei Familiari 2023
Adriano	123,50	4273	9550
Dergano	62,26	8644	14026
Affori	57,26	9200	14468
Gorla - Precotto	45,72	11564	16851
Bovisa	45,03	5794	8403
Pta Romana	0,78	9586	9661
Brera	−0,07	10702	10694
Porta Genova	−1,03	8408	8321
Guastalla	−1,46	8765	8637
Duomo	−16,64	11115	9266

Tabella 4.9. Variazione dei nuclei familiari per i quartieri selezionati tra il 1999 e il 2023. I primi 5 quartieri sono i migliori per variazione percentuale positiva nei quartieri con più nuclei familiari, mentre gli ultimi 5 sono i peggiori per variazione percentuale negativa.

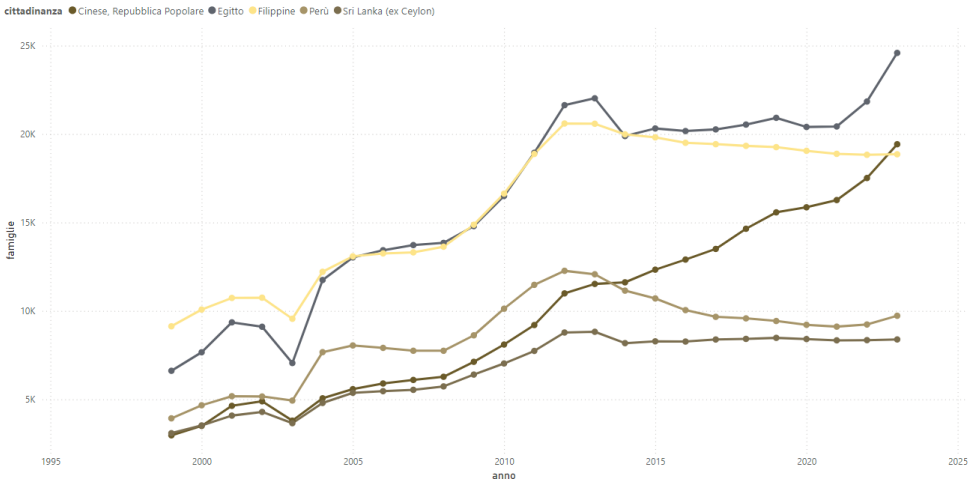

Figura 4.29. Serie storica relativa al numero totale di nuclei familiari presenti a Milano, suddivisa per le prime 5 cittadinanze per numerosità (esclusa l'Italiana).

Cittadinanza	Variazione (%)	Nuclei Familiari 1999	Nuclei Familiari 2023
Ucraina	15.971,40	42	6750
Bangladesh	1.433,04	569	8723
Romania	1.191,06	705	9102
Ecuador	588,78	704	4849
Cinese, Repubblica Popolare	554,70	2967	19425
Senegal	30,63	1554	2030
Marocco	19,87	3829	4590
Francia	16,61	2330	2717
Italia	1,53	601614	610831
Regno Unito	−30,87	2271	1570

Tabella 4.10. Variazione dei nuclei familiari per cittadinanza tra il 1999 e il 2023.

Cittadinanza	Variazione (%)	Nuclei Familiari 2013	Nuclei Familiari 2023
Bangladesh	85,36	4706	8723
Cinese, Repubblica Popolare	68,52	11527	19425
Francia	30,37	2084	2717
El Salvador	28,41	2496	3205
Egitto	11,65	22022	24588
Filippine	−8,38	20583	18858
Brasile	−12,66	2314	2021
Albania	−16,32	2948	2467
Perù	−19,44	12073	9726
Ecuador	−36,75	7667	4849

Tabella 4.11. Variazione dei nuclei familiari per cittadinanza tra il 2013 e il 2023.

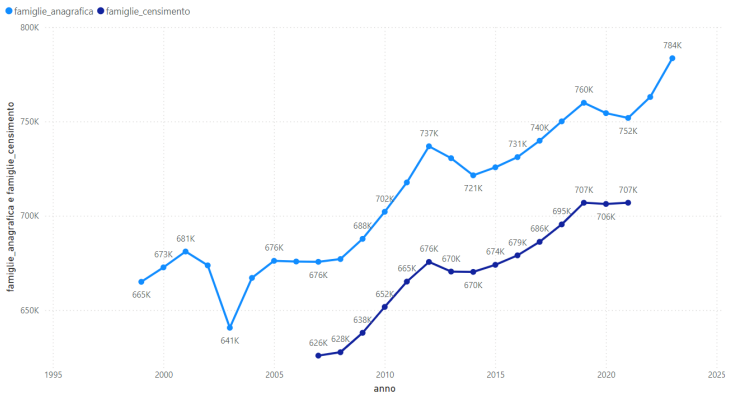

Figura 4.30. Confronto delle serie storiche relative al numero totale di nuclei familiari presenti a Milano, considerando i dati anagrafici e censuari.

4.1.6 Dati censuari

Il dataset `tipo_fam_res_CENS`, relativo ai dati raccolti dai vari censimenti eseguiti nel comune di Milano, ricopre l'arco temporale che va dal 2007 al 2021. In generale, le operazioni di censimento sono utili per correggere il numero dei residenti proveniente dall'anagrafe nelle aree considerate, per avere un'idea più precisa dei residenti realmente presenti a Milano, in questo caso. Ci si aspetta perciò che l'andamento dei residenti anagrafici sia una sovrastima dei residenti forniti dai dati censuari, come in effetti si rileva in Figura 4.30. Si riportano le statistiche principali della serie storica associata al numero totale di famiglie per anno rilevate dal censimento in Tabella 4.12.

conteggio	media	dev. std.	min	25%	50%	75%	max
15	671.956,40	26.915,65	625.885	658.406,5	674.016	690.811,5	706.915

Tabella 4.12. Statistiche relative alla serie storica relativa al dataset `tipo_fam_res_CENS`.

I campi presenti sono analoghi a quelli dei dati anagrafici (tranne l'informazione relativa alla 'cittadinanza' che in questo dataset non viene fornita), perciò si possono consegnare delle considerazioni, confrontando i valori forniti in questo caso con quelli riportati in precedenza. Suddividendo i dati per fascia di età del capo famiglia, si osserva infatti che l'andamento della serie relativa a capofamiglia con età '80 anni e più' è in trend crescente, nel 2015 ha superato quella associata a 'da 18 a 35 anni' e nel 2021 ha raggiunto quella relativa a 'da 65 a 79 anni' (Figura 4.31). In anagrafica tale serie rimane sempre al di sotto di tutte le altre (Figura 4.26). Inoltre in questo caso il numero di famiglie associate a un capo famiglia con età compresa tra i 18 e i 35 anni sembra rimanere costante, mentre i dati anagrafici evidenziano un lieve trend crescente. Per quanto riguarda il genere del capofamiglia si ha un comportamento sostanzialmente analogo a quello presente nei dati anagrafici (Figura 4.32), così come per i vari livelli della variabile 'numero_componenti' (Figura 4.33).

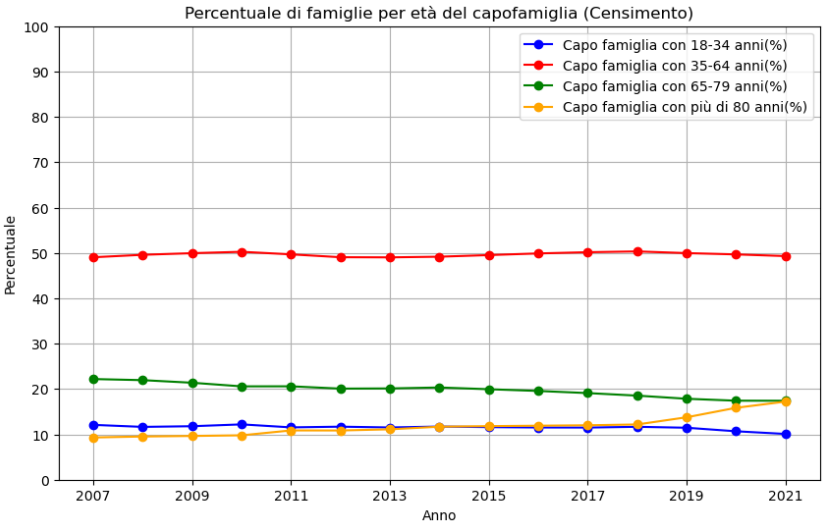

Figura 4.31. Confronto delle serie storiche relative al numero totale di nuclei familiari presenti a Milano, considerando i dati censuari e suddividendo per fascia di età del capo famiglia.

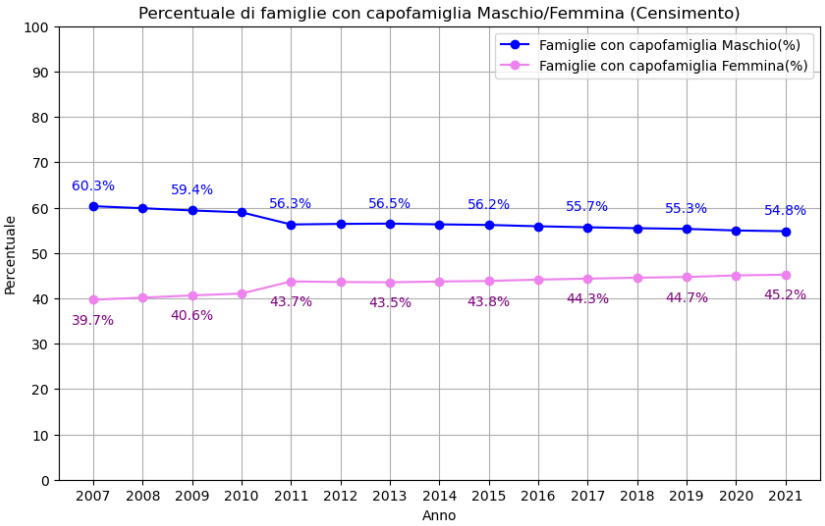

Figura 4.32. Confronto delle serie storiche relative al numero totale di nuclei familiari presenti a Milano, considerando i dati censuari e suddividendo per genere del capo famiglia.

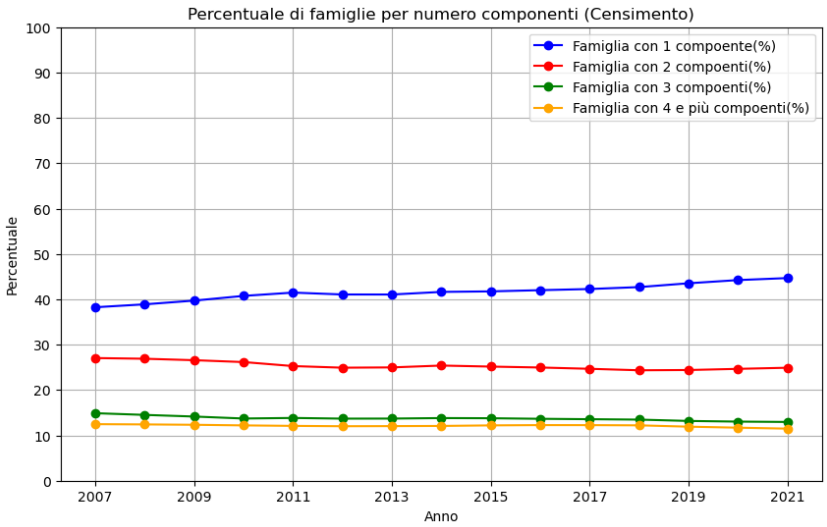

Figura 4.33. Confronto delle serie storiche relative al numero totale di nuclei familiari presenti a Milano, considerando i dati censuari e suddividendo per numero di componenti del nucleo familiare.

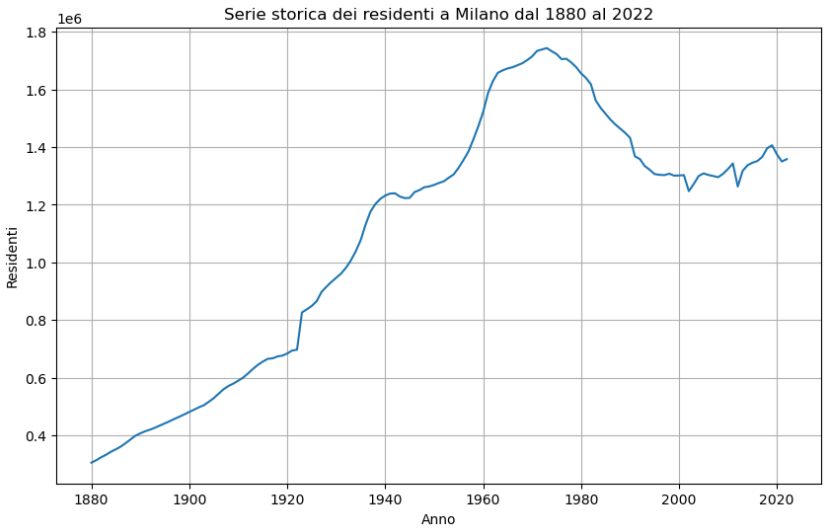

Figura 4.34. Serie storica dei residenti totali a Milano dal 1880 al 2022.

4.1.7 Movimento naturale e migratorio

Si aggiunge una descrizione del dataset relativo alla Popolazione calcolata ISTAT dall'anno 1880 al 2022, comprendente nascite, decessi, immigrati ed emigrati per ogni anno. Tale dataset non è stato integrato nella prima fase poiché ha dimensioni ridotte rispetto ai restanti dataset, ma contiene informazioni più lontane nel tempo, che permettono di comprendere in modo più accurato l'andamento della popolazione. Infatti, se i dataset precedentemente analizzati consentono di avere un'idea più specifica e dettagliata dell'andamento dei singoli NIL, con caratteristiche e informazioni relative a sottocategorie presenti tra i record, d'altra parte non permettono di avere informazioni meno recenti che consentirebbero di dare sostegno a eventuali applicazioni di modelli predittivi per le serie storiche. Il dataset del movimento naturale e migratorio fornisce, invece, informazioni generiche, che non permettono di estrarre informazioni sui singoli NIL, come è stato fatto in precedenza, ma consegna dati annuali più distanti nel tempo. Tale dataset presenta come campi: il numero delle nascite, dei decessi, degli immigrati, degli emigrati e infine dei residenti dal 1880 al 2022. Il numero dei residenti è dato dal valore dell'anno precedente a cui vengono sommati le nascite e gli immigrati e sottratti i decessi e gli emigrati dell'anno corrente.
Per comprendere queste serie storiche, vengono utilizzate le metriche comunemente impiegate da ISTAT per confrontare questo tipo di dati ed estrarre considerazioni interessanti [14]:

- Tasso di natalità: rapporto tra il numero dei nati vivi dell'anno e l'ammontare medio della popolazione residente, moltiplicato per 1.000. In modo analogo si definiscono i tassi di mortalità, immigrazione ed emigrazione;
- Tasso di crescita naturale: differenza tra il tasso di natalità e di mortalità;

- Tasso migratorio totale: differenza tra il tasso di immigrazione e di emigrazione;
- Tasso di crescita totale: somma del tasso di crescita naturale e del tasso migratorio totale.

In Figura 4.35 si osserva un confronto tra i tassi relativi a nascite e decessi ed un confronto tra i tassi relativi a immigrazione ed emigrazione. Si possono dare le seguenti considerazioni:

- Per quanto riguarda nascite e decessi si è osservato dal 1880 un declino del tasso di mortalità che si è stabilizzato dal 1950 circa in poi. Per quanto riguarda il tasso di natalità si ha, dopo una decrescita dal 1880, un comportamento invece meno stabile dal 1950 in poi, che evidenzia in particolare la decrescita negli ultimi anni, come già sottolineato in precedenza.
- Per i tassi di immigrazione e emigrazione si notano valori più alti rispetto ai tassi di natalità e mortalità, che indicano una incisione maggiore sul totale.

Andando ad osservare la Figura 4.36 si possono apportare delle considerazioni sui tassi di crescita:

- Per il tasso di crescita naturale si osservano, eccetto per i due picchi negativi dovuti alle guerre, valori positivi, indicanti un numero di nascite superiore al numero dei decessi. Dal 1965 circa in poi si assiste a un declino del tasso arrivando sotto lo zero al 1975 circa: da lì in poi si ha un tasso negativo fino ad oggi. In particolare il tasso continua a decrescere non a causa dei decessi, poiché stanno rimanendo costanti come si diceva, ma a causa del numero sempre inferiore di nascite.
- Per quanto riguarda il tasso di crescita migratoria si ha da sempre un tasso positivo, eccetto per l'intervallo temporale che va dal 1970 al 1995 circa, e si nota un picco attorno al 2010 circa. Il range di valori coinvolti per questo tasso è molto maggiore rispetto al tasso naturale, e ciò evidenzia la motivazione della crescita demografica degli ultimi anni nella popolazione milanese: seppure il tasso naturale è negativo e continua a decrescere, il tasso di crescita totale è positivo grazie all'alto numero di immigrazioni che incide in modo decisivo sulle emigrazioni e sui decessi.
- Osservando gli ultimi grafici nella figura si riporta il confronto del tasso naturale sul totale e il confronto del tasso migratorio sul totale, per evidenziare come il totale sia fortemente influenzato dal migratorio.

In Figura 4.34 si riporta l'andamento della serie storica del totale dei residenti di Milano tra il 1880 e il 2022, che verrà utilizzata in seguito per consegnare proiezioni dell'andamento demografico futuro della popolazione milanese.

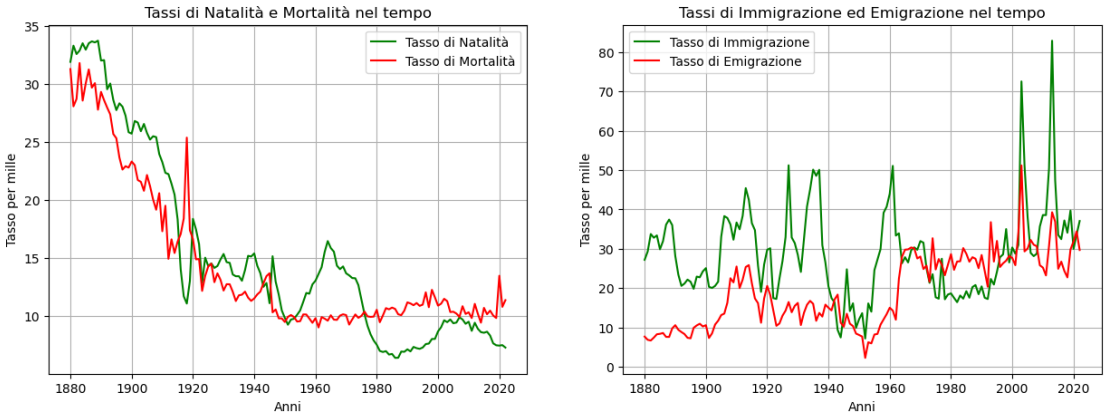

Figura 4.35. Confronto tra i tassi di natalità e mortalità e confronto tra i tassi di immigrazione ed emigrazione.

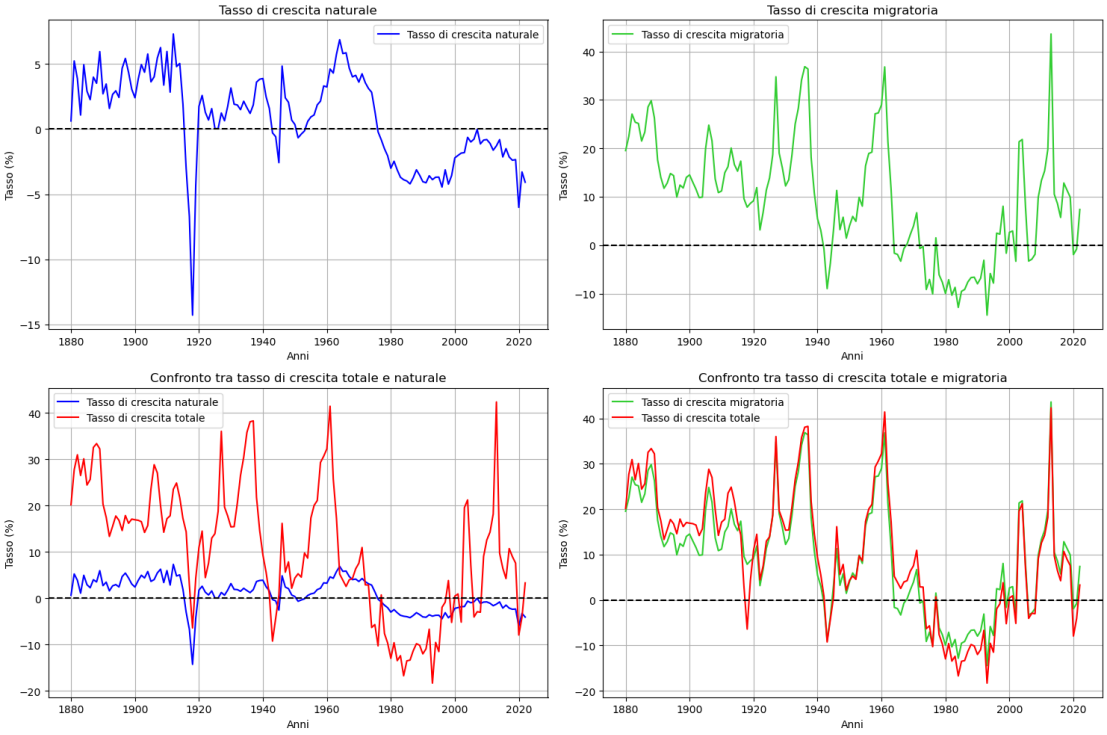

Figura 4.36. Andamenti dei tassi di crescita naturale, tasso di crescita totale, confronto tra il tasso naturale e totale, confronto tra il tasso migratorio e totale.

4.2 Modelli predittivi

L'obiettivo di questa sezione è condurre una valutazione approfondita di diversi modelli predittivi per descrivere l'evoluzione nel tempo del numero di residenti per la popolazione di Milano. Le sorgenti dati integrate e quindi esplorate nei capitoli precedenti riguardano gli anni dal 1999 in poi e hanno come scopo quello di dare una precisa idea della situazione della popolazione con granularità data dal NIL, ovvero i quartieri di Milano, su alcuni specifici attributi. Ciò permette di ottenere un quadro generale relativo a sotto-categorie di nascite, decessi, della tipologia di famiglie, utile per decisori ed esperti che hanno bisogno di informazioni di questo tipo a partire dai dati grezzi. Tuttavia, l'applicazione di modelli predittivi alle serie storiche ottenute dall'aggregazione dei dati rispetto agli anni e ai quartieri, non è ben sostenuta da un quantitativo adeguato di dati, dal momento che si hanno a disposizione soltanto 20 punti circa per ogni anno. Per questo motivo in questa fase si prenderà in considerazione la serie storica dei residenti nel comune di Milano fornita da ISTAT, dal 1880 al 2022 [23], relativa al movimento naturale e migratorio della popolazione. Le analisi predittive su questa serie consentiranno di consegnare osservazioni aggiuntive sulla situazione demografica milanese delineata nei capitoli precedenti. La serie storica è presentata in Figura 4.34, dove si può osservare un trend crescente dal 1880 al 1970 circa, quindi una decrescita fino al 2000, anno in cui la serie sembra tornare ad assumere un comportamento crescente in modo più lento e irregolare. Dal momento che la serie storica dei residenti è influenzata da molteplici fattori che non si riescono ad includere nel modello, l'attenzione si focalizzerà su semplici modelli comunemente impiegati in letteratura ([22], [2]), non troppo complessi, con l'obiettivo di mitigare il rischio di overfitting e garantire una buona capacità generalizzatrice. Si andranno a valutare i modelli Simple Exponential Smoothing (SES), Double Exponential Smoothing (DES), Autoregression Integrated Moving Average Model (ARIMA) adattato ai dati in due modi: il primo applicando la funzione `auto_arima` di `pmdarima`, che possiede un algoritmo di ottimizzazione che consente di individuare la configurazione migliore in termini di AIC; il secondo scegliendo manualmente i parametri sulla base di una analisi dei grafici di autocorrelazione e autocorrelazione parziale relativi alla serie storica. Tali modelli appena descritti, sono classici modelli per serie storiche che tengono conto soltanto dell'andamento della popolazione nel tempo. In molti casi, un approccio di questo tipo può essere adatto, se non addirittura l'unica possibilità quando non si hanno a disposizione altre informazioni. Tuttavia, il contesto relativo alla demografia di una città risulta essere estremamente complesso e influenzato da molteplici fattori che non possono essere considerati trascurabili e vanno inclusi, per quanto possibile, in un modello che ha come obiettivo quello di fornire precise previsioni future. Eventi fuori dall'ordinario come carestie, pandemie, guerre, ma anche fenomeni ordinari come andamenti dei tassi di natalità, di mortalità, flussi migratori, risultano essere aspetti sociali che modelli come un ETS, ad esempio, non sono in grado di catturare. Per tale motivo sono stati presi in considerazione ulteriori modelli presenti in letteratura ([5], [20], [1], [10]), impiegati solitamente nell'ambito di dati demografici per la descrizione e la proiezione futura. Per questi modelli si considera una previsione che rappresenta una tendenza generale della serie, che permette di avere un'idea sul futuro a lungo termine. Tali valutazioni e comparazioni tra modelli avverranno

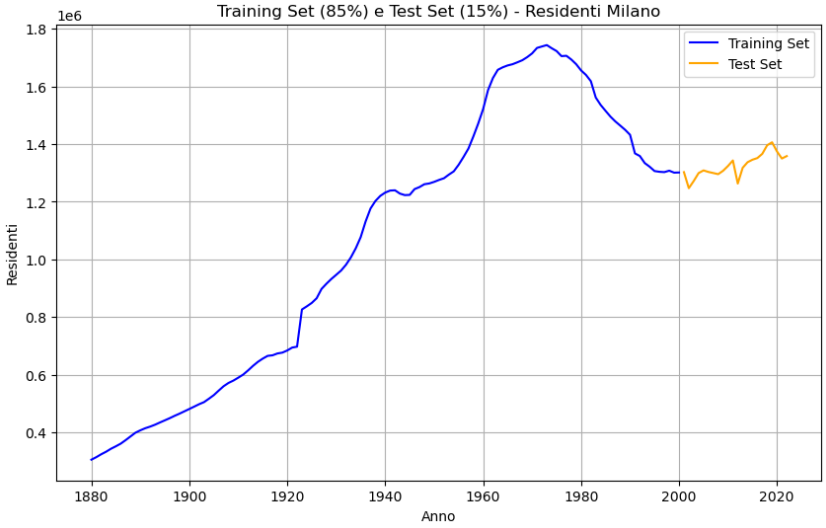

Figura 4.37. Suddivisione della serie storica relativa ai residenti a Milano (dati ISTAT) suddivisa in training e test set.

tramite l'utilizzo di metriche di performance adeguate e ben note in letteratura. I risultati delle performance dei modelli verranno presentati nel Capitolo 6. Per condurre questa analisi è stato impiegato il linguaggio Python, sfruttando la potenza di librerie specializzate come SQLAlchemy per l'interazione con i database, Pandas per la manipolazione e l'analisi dei dati, Matplotlib e Seaborn per la visualizzazione dei risultati, e Statsmodels per l'implementazione di modelli statistici. Per poter valutare le performance dei modelli adattati, si decide di dividere i dati in due insiemi: un training set, che viene impiegato per adattare i modelli sui dati estratti in precedenza; un test set, utilizzato per valutare la bontà del modello in termini di previsione. Si è deciso perciò di considerare le osservazioni dal 1880 al 2000 come training set (circa l'85% dei dati, ovvero 121 osservazioni) e quelle dal 2001 al 2022 come test set (Figura 4.37).

4.2.1 Metriche di performance

Per valutare efficacemente le performance dei modelli dal punto di vista dell'adattamento ai dati è stato utilizzato l'AIC (Akaike Information Criterion) una metrica che valuta la bontà di un modello tenendo conto sia della qualità dell'adattamento che della complessità del modello [3]. Si consideri un contesto in cui si vogliono determinare i parametri di un modello in modo tale da descrivere al meglio i dati che si hanno a disposizione. Un approccio classico è quello di andare a massimizzare la verosimiglianza associata al modello, sui parametri liberi associati allo stesso. Includere molti parametri aumenta i gradi di libertà del modello, rendendolo più flessibile e aumentando le possibilità di adattamento ai dati, tuttavia in tal modo si incrementa anche la complessità del modello, rendendolo sempre meno interpretabile all'aumentare dei parametri e rischiando di renderlo poco efficace in

termini di previsione. L'AIC tiene conto di questi fattori. La formula ad esso associata è

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

dove k è il numero di parametri del modello e $\hat{L}$ è il valore massimo della verosimiglianza del modello ottenuto sostituendo nella sua espressione i parametri ottimi. L'AIC considera in positivo la bontà di adattamento ai dati e penalizza l'inclusione di molti parametri. Dalla formula si comprende che un valore di AIC basso indica un modello migliore. In alcuni casi la formula dell'AIC può essere specializzata, ad esempio se il modello sottostante viene assunto essere del tipo

$$y_t = f(t, \mathbf{q}) + \varepsilon_t,$$

dove $\varepsilon_t \sim N(0, \sigma^2)$ sono variabili indipendenti e identicamente distribuite secondo una normale, rappresentanti l'errore del modello, $\mathbf{q}$ è il vettore, di lunghezza k, dei parametri del modello. Assumendo di aver calcolato le stime di massima verosimiglianza per $\mathbf{q}$, che sono le stime dei minimi quadrati in questo caso, e per σ^2, che viene stimata come la media dei residui del modello (con i parametri stimati) al quadrato, in tal caso l'AIC può essere espresso nel seguente modo:

$$\text{AIC} = 2(k+1) + n \cdot \ln(\text{MSE})$$

dove si hanno $k+1$ parametri poiché si stimano i k parametri del modello e anche la varianza degli errori, n indica il numero di osservazioni del Training Set e dove con MSE si intende il mean squared error, ovvero la stima della varianza degli errori del modello, cioè supponendo di chiamare y_i le osservazioni reali consegnate dai dati e $\hat{y}_i$ i valori corrispondenti ottenuti adattando il modello:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2.$$

Nei modelli che verranno considerati si sottintenderà una contesto di questo tipo, che permette di calcolare agilmente e confrontare facilmente l'AIC tra i modelli. Si sottolinea che l'ipotesi di indipendenza degli errori e di normalità sono ipotesi forti, non sempre supportate dall'osservazione dei risultati delle applicazioni dei modelli ai dati, ma si decide di applicarle per avere un criterio comune di confronto sulla bontà dell'adattamento dei modelli.
Per valutare invece le performance predittive dei modelli, sono state adottate le seguenti metriche:

- MAE (Mean Absolute Error): rappresenta la media degli errori assoluti tra i valori previsti e i valori osservati. La formula del MAE è la seguente:

  $$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

  dove y_i rappresenta il valore osservato, $\hat{y}_i$ il valore previsto, e n il numero totale di osservazioni del Test Set. Questa metrica fornisce una misura intuitiva dell'errore medio di previsione, espressa nelle stesse unità dei dati originali.

- MAPE (Mean Absolute Percentage Error): fornisce la media degli errori in termini percentuali rispetto ai valori osservati. La formula del MAPE è:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

  Questa metrica esprime la precisione del modello come percentuale, rendendola particolarmente utile per confrontare modelli su dataset con scale diverse. In particolare indica la grandezza media degli errori delle previsioni, espressa come percentuale rispetto ai valori effettivi. Un MAPE del 5%, ad esempio, significa che, in media, le previsioni del modello sbagliano del 5% rispetto ai valori reali.

- RMSE (Root Mean Squared Error): è la radice quadrata della media dei quadrati degli errori. La formula è:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$

  L'RMSE penalizza maggiormente gli errori più grandi, essendo sensibile a valori anomali o outlier.

4.2.2 Simple Exponential Smoothing

Il modello di Simple Exponential Smoothing (SES) è utilizzato per serie storiche senza trend e senza stagionalità, in cui ogni previsione è una media pesata dei valori passati [12]. L'obiettivo è minimizzare l'errore quadratico tra i valori previsti e quelli osservati, utilizzando una costante di smoothing α. La formulazione del modello SES è:

$$\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1},$$

dove:

- $\hat{y}_t, \hat{y}_{t+1}$ sono le stime del modello;
- α è il parametro di smoothing, $0 < \alpha < 1$,
- y_t è l'osservazione reale al tempo t.

Il valore previsto per i periodi futuri, è dato dall'ultima osservazione del dataset di training:

$$\hat{y}_{T+h} = \hat{y}_T, \quad h = 1, 2, \ldots \tag{4.2}$$

dove $\hat{y}_T$ è l'ultima previsione calcolata nel periodo di training. Il parametro α viene determinato minimizzando l'errore quadratico medio (MSE) sui dati di training. In `Python`, il pacchetto `statsmodels` permette di stimare α tramite il metodo dei minimi quadrati. Solitamente si addiziona ad ogni variabile del modello un errore normalmente distribuito con media nulla e varianza σ^2, stimando quest'ultima con il Mean Square Residual del modello, ovvero la somma dei quadrati dei residui del modello divisa per il numero di dati di training. Ciò permette di ottenere intervalli di confidenza sulla previsione fornita dal modello.

4.2.3 Double Exponential Smoothing

Il Double Exponential Smoothing (DES), detto anche Holt-Winter model, è una variante del SES che include un termine di trend, rendendolo adatto a serie con trend lineare. La formulazione è data da:

$$\begin{cases} \hat{y}_{t+1} = l_t + b_t, \\ l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}), \\ b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}. \end{cases}$$

dove:

- l_t è il livello stimato al tempo t,
- b_t è il trend stimato al tempo t,
- α e β sono i parametri di smoothing per il livello e il trend, rispettivamente.

Le previsioni future a partire da un tempo T possono essere calcolate come:

$$\hat{y}_{T+h} = l_T + h \cdot b_T, \quad h = 1, 2, \ldots \tag{4.3}$$

dove l_T e b_T si riferiscono, rispettivamente, all'ultimo livello e all'ultimo parametro di trend stimati sul dataset di training. I parametri α e β vengono ottimizzati minimizzando l'errore quadratico medio sui dati di training. In `Python`, `statsmodels` può stimare autonomamente questi parametri. Come nel SES, si associa al modello un set di variabili aleatorie di errore $\varepsilon_t \sim N(0, \sigma^2)$ che sono indipendenti e identicamente distribuite normalmente.

4.2.4 Modello Autoregressivo di ordine 1

Il modello autoregressivo (AR) cerca di spiegare il valore attuale della variabile dipendente in funzione dei suoi valori passati. La forma generale del modello AR(1) a media nulla è:

$$y_t = \alpha y_{t-1} + \varepsilon_t, \quad \forall t$$

dove y_t è la variabile aleatoria del processo al tempo t, α è il parametro del modello compreso tra -1 e 1 che indica quanto la variabile al tempo t sia dipendente dalla precedente e ε_t è l'errore al tempo t (come negli altri casi gli errori si assumono $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$). Questo modello fa parte della più grande famiglia dei modelli ARIMA, che verranno descritti in seguito più nel dettaglio. In generale questi modelli risultano utili quando si riconosce un pattern di autocorrelazione nei dati o comunque si osserva della dipendenza tra i valori presenti nella serie. In questo modo il modello può catturare in modo più efficace la dipendenza che sussiste tra le osservazioni presenti nella serie storica.
Si definisce la funzione di autocorrelazione (ACF) per un processo stocastico stazionario come

$$\rho(l) = \frac{\mathrm{Cov}(y_{t_i}, y_{t_i+l})}{\sqrt{\mathrm{Var}(y_{t_i})\mathrm{Var}(y_{t_i+l})}} = \frac{\gamma(l)}{\gamma(0)}.$$

Assumendo la serie stazionaria, si può calcolare l'autocorrelazione campionaria relativa alla serie temporale e plottarla. Il grafico della ACF è chiamato *correlogramma* e permette di avere delle informazioni sulla dipendenza temporale della serie storica. Infatti il valore della ACF ad ogni lag è compreso tra -1 e 1: se è vicino a 0 significa che le osservazioni a distanza l non sono correlate tra loro; se è vicina a 1 sono fortemente correlate positivamente; se è vicina a -1 sono fortemente correlate negativamente. La ACF può essere stimata nel seguente modo:

$$\hat{\rho}(l) = \frac{\sum_{t=1}^{n-l}(y_t - \bar{y})(y_{t+l} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

dove n rappresenta la lunghezza della serie temporale, $\bar{y}$ è la media campionaria e l è il lag per cui si sta calcolando la stima. Si sottolinea il fatto che più è alto il numero di osservazioni, più il numero di termini considerati nella stima cresce e quindi aumenta la precisione della stima, soprattutto per lag grandi. Con PACF si indica la funzione di autocorrelazione parziale così definita per una serie storica stazionaria:

$$\psi(l) = \frac{\mathrm{Cov}(y_{t_i}, y_{t_i+l} | y_{t_i+1}, y_{t_i+2}, \ldots, y_{t_i+l-1})}{\mathrm{Var}(y_{t_i} | y_{t_i+1}, y_{t_i+2}, \ldots, y_{t_i+l-1})}.$$

Tale funzione ci permette di osservare la correlazione diretta tra due osservazioni (y_{t_i}, y_{y_i+l}) a distanza l tra loro, dopo aver rimosso l'effetto delle osservazioni intermedie y_{t_i+1}, y_{t_i+2}, $\ldots, y_{t_i+l-1}$. Quindi sostanzialmente ci permette di vedere come sono correlate due osservazioni del processo a distanza l depurando il calcolo dalle osservazioni intermedie che potrebbero contribuire ulteriormente. Anche in questo caso esiste il corrispettivo campionario della PACF, il quale diventa tanto più accurato quanto più cresce il numero di osservazioni che si hanno a disposizione nella serie storica. Nel caso della serie storica dei residenti di Milano, assumendola stazionaria e andando a considerare la stima di autocorrelazione, si osserva una forte correlazione positiva per lag maggiori di 1 che decresce nel tempo, mentre una correlazione significativa solo al lag 1 per la PACF (Figura 4.38), mentre per tutti gli altri lag si hanno valori statisticamente nulli. Per "valori statisticamente nulli" si intende indicare i valori stimati della ACF e PACF che cadono all'interno delle bande azzurre nei plot, le quali si riferiscono per l'ACF agli intervalli di confidenza costruiti tramite la formula di Bartlett [6] che permette di approssimare la varianza della stima della autocorrelazione ad ogni lag k nel seguente modo:

$$\mathrm{Var}(\hat{\rho_k}) = \frac{1}{n}\left(1 + 2\sum_{i=1}^{k-1}\hat{\rho_i}^2\right),$$

mentre nel caso della PACF il metodo è semplificato assumendo che a ogni lag si possa stimare la deviazione standard come $\frac{1}{\sqrt{N}}$, ottenendo infatti un intervallo costante per tutti i lag k. Si osservi che il modello sopra definito è a media nulla, ma nulla vieta di introdurre una media nel modello per adattarlo alle esigenze dei dati di training che si hanno. Per costruire un autoregressivo con media μ, basta considerare il processo $z_t = x_t - \mu$ considerando z_t un AR(1) a media nulla e osservare che:

$$x_t - \mu = \alpha(x_{t-1} - \mu) + \varepsilon_t \implies x_t = \mu + \alpha(x_{t-1} - \mu) + \varepsilon_t.$$

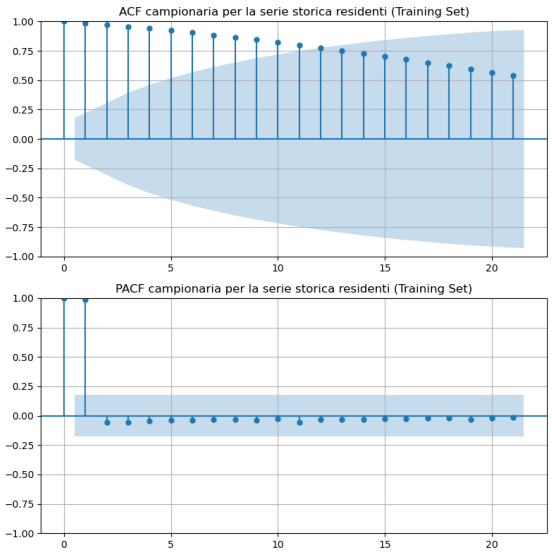

Figura 4.38. Correlogramma e correlogramma parziale per la serie storica di training dei residenti fornita da ISTAT.

I parametri da stimare per il modello AR(1) sono appunto il valore atteso della serie μ, che viene stimato con la media campionaria sui dati di training, il coefficiente α e la varianza σ^2 degli errori che vengono ottenuti tramite il metodo di massima verosimiglianza applicato dalla funzione `ARIMA` della libreria `statsmodels.tsa.arima.model`.
Oltre alla stima dei parametri del modello e alla costruzione di previsioni a partire dallo stesso, è importante calcolare anche gli intervalli di confidenza associati alle previsioni. Infatti, è desiderabile avere un'idea della accuratezza della previsione, fornita da una zona di confidenza che garantisce, con un certo livello di probabilità, che le osservazioni future cadranno al suo interno. Nel caso del modello AR(1), supponendo di aver osservato la serie storica fino al tempo n e di aver stimato il valore atteso μ, il coefficiente autoregressivo α e la varianza degli errori σ^2, si avrà che la distribuzione della previsione al tempo $n+1$ sarà data da:

$$x_{n+1} = \mu + \alpha(x_n - \mu) + \varepsilon_{n+1} \sim N(\mu + \alpha(x_n - \mu), \sigma^2),$$

dove con μ, α e σ si fa riferimento alle loro stime. Per l'osservazione al tempo $n+2$ si avrà sempre una distribuzione normale con valore atteso:

$$E(x_{n+2}|x_n) = \mu + \alpha(E(x_{n+1}|x_n) - \mu) = \mu + \alpha(\alpha x_n - \mu) = \mu + (\alpha^2 x_n - \alpha\mu)$$

e con varianza:

$$Var(x_{n+2}|x_n) = \alpha Var(x_{n+1}|x_n) + \sigma^2 = \alpha^2\sigma^2 + \sigma^2 = \sigma^2(1+\alpha^2).$$

Generalizzando si ottiene, per la previsione al tempo $n+k$, la seguente distribuzione associata:

$$x_{n+k} \sim N\left(\alpha^k x_n, \sigma^2 \sum_{i=1}^{k} \alpha^{2(k-1)}\right),$$

grazie alla quale si possono ricavare gli intervalli di confidenza con il livello desiderato (in questa analisi si considererà un livello di confidenza pari a 0.95). Si osservi che per k elevati, ovvero per previsioni molto distanti dall'ultima osservazione rilevata, si ha una distribuzione che non dipende più dall'osservazione x_n poiché, dal momento che $\alpha \in (-1,1)$ per potenze elevate tende a zero, e quindi il valore atteso della normale tende alla media campionaria stimata dei dati di training. Per quanto riguarda la varianza, si ha che, per valori elevati di k, tende al valore $\frac{\sigma^2}{1-\alpha^2}$, ovvero a un valore costante, chiaramente maggiore di σ^2.

4.2.5 Modello SARIMA

Il modello autoregressivo fa parte di una famiglia più ampia di modelli: gli Autoregressive Integrated Moving Average Models (ARIMA). I modelli ARIMA sono composti da tre parti:

- AR(p): un *modello autoregressivo* di ordine p definito come:

  $$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t.$$

  Gli α_i sono chiamati *coefficienti di autoregressione* e indicano quanto la variabile al tempo t sia influenzata dall'osservazione y_{t-i}; ε_t rappresenta l'errore al tempo t, cioè un rumore bianco di media nulla e varianza σ^2 da stimare (anche in questo caso gli errori sono i.i.d). Tale modello è Markoviano di ordine p, ovvero dipendente solo da variabili aleatorie del processo a distanza p o inferiore, ed è stazionario se tutti i coefficienti autoregressivi sono in valore assoluto minori di uno ($|\alpha_i| < 1$).

- MA(q): un modello *moving average* di ordine q definito come:

  $$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q}.$$

  In tal caso l'osservazione al tempo t non dipende dalle precedenti, ma dalle realizzazioni dei rumori bianchi che si sono osservate nei q tempi precedenti. Tale modello è stazionario per definizione (combinazione lineare di rumori bianchi) e per l'identificabilità del modello si richiede che i coefficienti β_i siano in valore assoluto minori di uno ($|\beta_i| < 1$).

- I(d): un modello integrato di ordine d, scrivibile in termini dell'operatore di backshift B (che ad ogni osservazione associa la precedente: $Bz_t = z_{t-1}$):

  $$(1 - B)^d y_t = \varepsilon_t.$$

  Tale definizione sta a significare sostanzialmente che le differenze finite di ordine d della serie storica corrispondono a un rumore bianco. La differenziazione viene introdotta nel caso in cui la serie non sia stazionaria, per renderla tale.

Un modello definito dalla combinazione di questi tre oggetti viene definito come ARIMA(p, d, q) e, per mezzo dell'operatore B, può essere sinteticamente espresso come segue:

$$\theta_p(B)(1 - B)^d x_t = \phi_q(B) w_t,$$

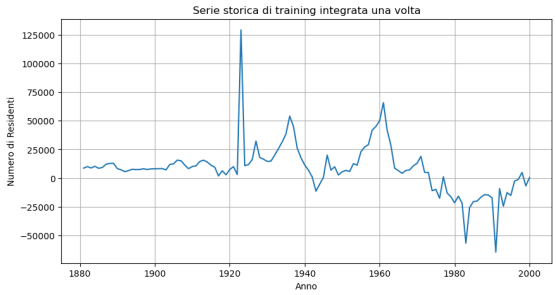

Figura 4.39. Serie storica dei residenti integrata una volta.

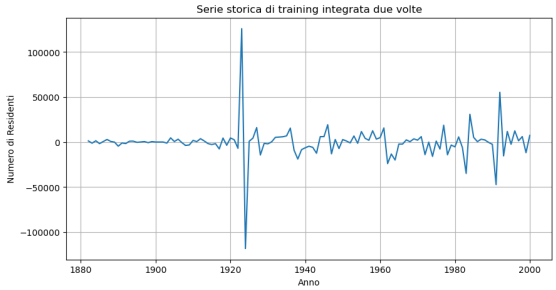

Figura 4.40. Serie storica dei residenti integrata due volte.

dove

$$\theta_p(B) = 1 - \alpha_1 B - \ldots - \alpha_p B^p \quad \text{e} \quad \phi_q(B) = 1 + \beta_1 B + \ldots + \beta_q B^q.$$

Infine, nel caso in cui sia presente una componente stagionale di ordine s nella serie, può essere utile introdurre nel modello una dipendenza dalle osservazioni a distanza s o suoi multipli ed anche integrare la serie secondo un lag pari ad s. Perciò il modello definitivo denominato SARIMA$(p,d,q)(P,D,Q)^s$ potrà essere espresso nel seguente modo:

$$\Theta_P(B^s)\theta_p(B)(1-B^s)^D(1-B)^d x_t = \Phi_Q(B^s)\phi_q(B)w_t.$$

P, D, Q sono i parametri relativi rispettivamente alla componente stagionale AR, alla componente stagionale I e alla componente stagionale MA del modello. Tale modello è chiaramente più complesso di quelli visti finora e necessita della stima di $p+d+q+P+D+Q$ parametri e della stima della varianza dei rumori bianchi.
Precedentemente la serie storica è stata assunta stazionaria, ma tale scelta può essere messa in discussione osservando la serie. Infatti, il concetto teorico di stazionarietà è difficilmente definibile con sicurezza quando si guarda una realizzazione di un processo stocastico. Da una sola realizzazione non si riesce ad affermare con certezza che la serie provenga da un processo stazionario, tuttavia si possono apportare delle ipotesi. Per questo si può pensare di eseguire un test di stazionarietà ben noto in letteratura che permetta di verificare se una serie può essere considerata stazionaria oppure no. Ad esempio il Test di Dickey-Fuller (ADF) applicato alla serie di training ha come ipotesi

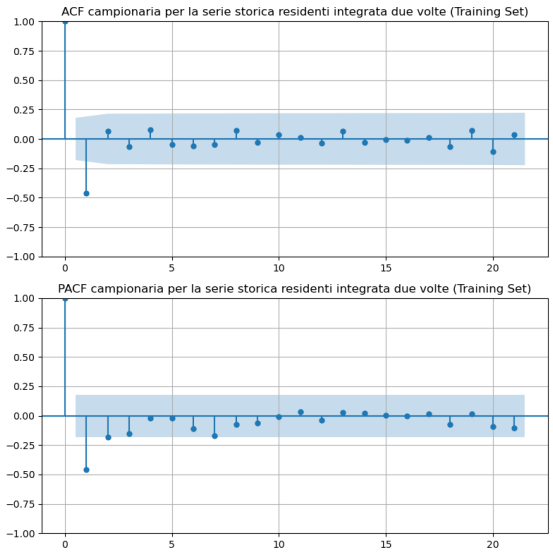

Figura 4.41. Correlogramma e correlogramma parziale per la serie storica di training dei residenti fornita da ISTAT, integrata due volte.

nulla che la serie non è stazionaria e consegna un p-value pari a 0,4926 che non permette di rifiutare l'ipotesi nulla. Se proviamo ad integrare la serie con le differenze finite di ordine 1 otteniamo la serie in Figura 4.39, che sembra ancora non essere stazionaria e in effetti applicando ancora il test si ottiene un p-value pari a 0,0117, buono perché sotto a 0,05, che è il classico livello che viene fissato, ma non fortemente significativo. Perciò si decide di applicare un'altra integrazione ottenendo un p-value $1{,}3 \cdot 10^{-14}$, che permette di rifiutare con molta sicurezza l'ipotesi nulla di non stazionarietà della serie (Figura 4.40). Quindi si procede andando ad analizzare i grafici di ACF e PACF in Figura 4.41 e si osserva che la ACF presenta una correlazione significativa negativa al lag 1, mentre per la PACF si ha significatività per il lag 1 e per i lag 2 e 3, che sono sul limite della zona di confidenza quindi potrebbero non essere così significativi. In ogni caso si decide di adattare un ARIMA(1,2,3), chiamandolo 'Manual ARIMA'.
Verrà anche impiegata la funzione `auto_arima` di `pmdarima` che permette di identificare la miglior configurazione di un SARIMA per i dati di training che vengono forniti alla funzione secondo un algoritmo di ottimizzazione. Anche se non si rileva una stagionalità nei dati, si lascia aperta la possibilità all'algoritmo di ottimizzazione impiegato, di scegliere lui stesso se inserire o meno le componenti stagionali. Ad ogni modo, il modello che viene consegnato dalla funzione è un ARIMA(1,2,1) senza componenti stagionali.

4.2.6 Modello di Malthus

Ora si andranno a descrivere i modelli solitamente impiegati per serie storiche relative a dati demografici. Il modello di crescita di Malthus, uno dei primi modelli matematici per la descrizione della crescita della popolazione, descrive una crescita esponenziale basata

sul presupposto che la popolazione cresca a un tasso proporzionale alla propria dimensione [13]. Il modello si basa sull'equazione differenziale con condizione iniziale fissata:

$$\frac{dP(t)}{dt} = r \cdot P(t) \quad P(0) = P_0,$$

dove P_0 è il valore della popolazione al tempo $t = 0$ e $r = b - d$ è il tasso di crescita intrinseco della popolazione, una costante positiva che rappresenta, nella formulazione originale del modello, la differenza tra il tasso di natalità b e quello di mortalità d, ma che più in generale può essere chiamato *growth rate* e indica il tasso di crescita della popolazione, cioè quanto fortemente cresce o decresce. Tale relazione sottintende che la crescita della popolazione sia direttamente proporzionale alla popolazione stessa, secondo una costante di proporzionalità data dal tasso di crescita. La formulazione del modello di Malthus può essere ricavata esplicitamente risolvendo l'equazione differenziale e ottenendo:

$$P(t) = P_0 \cdot e^{rt}.$$

Se r è positivo il modello prevede una crescita esponenziale, altrimenti una decrescita con tendenza a zero.
La prima osservazione che si può apportare è che il modello originale di Malthus [19], tiene conto, nel tasso di crescita della popolazione, soltanto le nascite e i decessi interni alla popolazione e non di altri fattori come immigrazione ed emigrazione. Vista la scarsa influenza del tasso di crescita naturale sul totale dei residenti della popolazione di Milano, si decide di non stimare r come la differenza del tasso di natalità e mortalità, poiché fornirebbe una descrizione poco accurata dell'andamento della popolazione. In questo contesto r verrà stimato in due modi:

- Si calcolerà la media di tutti i tassi di crescita della popolazione prendendo il valore della popolazione in un anno e nel successivo, tra il tempo iniziale e finale. In questo caso, data la struttura del dataset considerato, si include anche l'effetto di immigrazione ed emigrazione sul rate di crescita della popolazione. Si noti che tale calcolo equivale a prendere il tasso di crescita della popolazione tra il tempo iniziale e finale diviso per il numero di tempi intermedi tra le due osservazioni. Il calcolo del tasso di crescita r si otterrà nel seguente modo, prendendo la formulazione discreta del modello di Malthus, ovvero, considerando i tempi $t = 0,1,\ldots$, come gli anni associati all'osservazione della popolazione:

  $$P_{t+1} = P_0 \cdot e^{r(t+1)} = \ldots = P_t \cdot e^r \implies r = \ln\left(\frac{P_{t+1}}{P_t}\right).$$

  Perciò la stima di r sarà data da:

  $$r = \frac{1}{n}\sum_{t=0}^{n-1} \log\left(\frac{P_{t+1}}{P_t}\right) = \frac{1}{n}\log\left(\frac{P_n}{P_0}\right)$$

  dove P_t indica la reale osservazione che si ha nel training set della popolazione al tempo t.

- Un'altra modalità spesso impiegata è quella di selezionare r attraverso algoritmi che permettono di minimizzare la somma dei residui quadratici del modello (come `curve_fit` in Python), usando le coppie di valori (t, P_t) che si hanno a disposizione nel training set ([1], [11]). Formalmente, in questo caso, si risolve:

$$\min_{r,P_0} \sum_t \left(P_t - P_0 \cdot e^{rt}\right)^2,$$

  dove P_t rappresenta l'osservazione al tempo t che si ha nel dataset di training, mentre $P_0 \cdot e^{rt}$ indica la previsione di P_t ottenuta tramite il modello. Si osservi che la minimizzazione avviene sui due parametri P_0 e r, quindi anche la popolazione al tempo iniziale è presa come parametro libero per poter ottenere un miglior fitting sui dati.

Una seconda osservazione è relativa alla crescita indefinita della popolazione che il modello suppone. Nelle popolazioni reali, la crescita non può essere indefinita a causa di vincoli come la disponibilità di risorse, l'habitat limitato e altri fattori che influenzano la capacità di sostenere individui aggiuntivi. Di conseguenza, la crescita spesso si assume che tenda a rallentare e stabilizzarsi intorno a un valore massimo, noto come capacità portante K (come ad esempio in [4]). Nel contesto dell'andamento della popolazione di una città, risulta naturale includere una capacità portante, che raccoglie l'informazione di vari vincoli sulla disponibilità di risorse come acqua, cibo, energia, ma anche spazio abitativo e infrastrutture, disponibilità di impieghi lavorativi.

4.2.7 Modello di Verhulst

Il modello di Verhulst o modello di crescita logistica, tiene conto di un limite massimo della popolazione, per poter includere un tetto massimo di risorse disponibili per gli individui. Questo modello assume che il tasso di crescita sia proporzionale alla popolazione esistente e all'ammontare di risorse disponibili. Ciò significa che quando la popolazione è bassa, l'evoluzione è simile al modello di Malthus, quindi avrà una crescita esponenziale, ma quando la popolazione si avvicina al confine superiore di capacità il tasso di crescita si abbassa facendo tendere la popolazione lentamente al limite K. Il modello può essere descritto dalla relazione [18]:

$$\frac{dP(t)}{dt} = r_{\max} \cdot P(t) \cdot \left(1 - \frac{P(t)}{K}\right) \quad P(0) = P_0, \tag{4.4}$$

dove con $r_{\max}$ si indica il tasso di crescita massima della popolazione, K indica il limite di capacità della popolazione. Tale modello differisce da quello di Malthus per l'introduzione del termine $\left(1 - \frac{P(t)}{K}\right)$, che agisce su $r_{\max}$ rendendolo variabile. Più la popolazione nel tempo si avvicina a K, più il tasso di crescita massima $r_{\max}$ viene moltiplicato per un valore vicino a zero, dando un tasso di crescita intrinseco vicino a zero. Viceversa, se la popolazione è lontana dal limite massimo K, la capacità percepita dal modello sarà virtualmente infinita, perciò il tasso di crescita intrinseco sarà vicino a quello massimo $r_{\max}$, riproducendo il comportamento del modello di Malthus. Anche in questo caso,

poiché la popolazione è bassa si avrà una crescita comunque lenta per la popolazione. Il modello prevede un punto di flesso, che individua il cambio di concavità dell'andamento della popolazione, nel punto in cui la popolazione assume valore $P^* = \frac{K}{2}$, ovvero al tempo

$$t_{\text{flesso}} = \frac{\ln\left(\frac{K}{P_0} - 1\right)}{r_{\max}}.$$

In tale punto si osserva il massimo tasso di crescita della popolazione cioè il tasso di crescita:

$$\left.\frac{dP}{dt}\right|_{P=P^*} = \frac{r_{\max} K}{4}.$$

Tale modello ha una formulazione esplicita risolvendo l'equazione differenziale 4.4:

$$P_t = \frac{K \cdot P_0}{P_0 + (K - P_0)e^{-r_{\max} \cdot t}},$$

dove si riconosce la forma classica della funzione logistica. In particolare si osserva che il modello ha un comportamento strettamente crescente se $r_{\max}$ è positivo e strettamente decrescente se è negativo. Assumendo un tasso di crescita positivo, per t che tende a infinito si ha che la popolazione tende al valore limite K, mentre al tempo $t = 0$ la popolazione vale P_0. Se r è negativo si avrà invece all'infinito una tendenza a zero. In questo caso i parametri del modello da stimare sono P_0, $r_{\max}$ e K. L'approssimazione di $r_{\max}$ può essere ottenuta calcolando il massimo tasso di crescita nella popolazione osservato nel dataset di training tra due tempi consecutivi, così come K può essere approssimato con il valore massimo della popolazione presentatosi nel dataset di training. Altrimenti, anche in questo caso si possono considerare tutti i parametri liberi e andare ad ottenere le stime di $P_0, r_{\max}, K$ che minimizzano la somma dei quadrati dei residui del modello:

$$\min_{P_0, r_{\max}, K} \sum_t \left(P_t - \frac{K \cdot P_0}{P_0 + (K - P_0)e^{-r_{\max} \cdot t}}\right)^2.$$

4.2.8 Modello di Verhulst generalizzato

In [26] viene proposta una generalizzazione del modello logistico, che permette di renderlo più complesso, introducendo diversi parametri che consentono, tra le altre possibilità, di modificare la posizione del punto di flesso. La derivata della popolazione rispetto al tempo per questo modello è data da:

$$\frac{dP(t)}{dt} = r_{\max} \cdot P(t)^{\alpha} \left[1 - \left(\frac{N(t)}{K}\right)^{\beta}\right]^{\gamma}, \tag{4.5}$$

dove α, β, γ sono assunti positivi. Per questo modello si possono sottolineare le seguenti caratteristiche:

- $\lim_{t\to\infty} N(t) = K$, come per il modello logistico classico;

- Si può dimostrare che in questo modello la popolazione che si ottiene sul punto di flesso del modello è data da

$$P^* = \left(1 + \frac{\beta\gamma}{\alpha}\right)^{-(1/\beta)} K.$$

Da questa formula si ottiene il massimo tasso di crescita del modello:

$$\left.\frac{dP}{dt}\right|_{P=P^*} = r_{\max} \cdot K^\alpha \left(\frac{\alpha}{\alpha+\beta\gamma}\right)^{\alpha/\beta} \left(\frac{\beta\gamma}{\alpha+\beta\gamma}\right)^{\gamma}.$$

Alcuni limiti importanti per la popolazione nel flesso sono i seguenti:

- $\lim_{\gamma\to\infty} P^* = \lim_{\alpha\to 0} P^* = 0$;
- $\lim_{\beta\to 0} P^* = Ke^{-(\gamma/\alpha)}$;
- $\lim_{\beta\to\infty} P^* = \lim_{\alpha\to\infty} P^* = \lim_{\gamma\to\infty} P^* = K$.

Si osservi che, chiaramente, il modello di Verhulst può essere ottenuto da 4.5, sostituendo $\alpha = \beta = \gamma = 1$.

4.2.9 Modello di Richards

Il modello logistico generalizzato offre un'ampio raggio di personalizzazione, ma in questo studio verrà proposta una specializzazione del modello in cui si pongono i parametri $\alpha = \gamma = 1$ e si lascia libero il parametro β, ottenendo

$$\frac{dP(t)}{dt} = r \cdot P(t) \cdot \left[1 - \left(\frac{P(t)}{K}\right)^{\beta}\right].$$

Questo modello è chiamato Modello di Richards [25] e può essere espresso esplicitamente attraverso la formulazione in termini della popolazione al tempo 0:

$$P_t = \frac{P_0 K}{[P_0^\beta + (K^\beta - P_0^\beta)e^{-\beta r t}]^{1/\beta}}.$$

Questo modello permette di controllare la popolazione corrispondente al punto di flesso tramite il parametro β. Infatti si può dimostrare che la popolazione al punto di flesso è data da:

$$P^* = \left(\frac{1}{1+\beta}\right)^{1/\beta} K.$$

e il tempo corrispondente è

$$t_{\text{flesso}} = \frac{1}{\beta r} \ln\left[\frac{1}{\beta}\left(\left(\frac{K}{P_0}\right)^{\beta} - 1\right)\right],$$

che risulta chiaramente dipendente dal parametro β.

Capitolo 5

Data Visualization

La Data Visualization ha assunto un ruolo sempre più centrale nell'ambito dell'analisi dei dati e del decision making aziendale. Il rapido sviluppo delle tecnologie dell'informazione ha generato una quantità enorme di dati, che richiedono strumenti efficaci per essere gestiti e compresi. Tra questi, le dashboard informative sono emerse come una soluzione efficace: esse forniscono, in un'unica schermata, le informazioni più importanti necessarie per svolgere un lavoro, permettendo un monitoraggio immediato. Tuttavia, molte dashboard aziendali non raggiungono il loro pieno potenziale a causa di una scarsa progettazione visiva, non di limiti tecnologici [9]. Per essere efficaci, le dashboard devono presentare molte informazioni in spazi ridotti, in modo chiaro e immediato, sfruttando la percezione visiva per elaborare rapidamente i dati. Questo può essere realizzato solo integrando la progettazione visiva nel processo di sviluppo, basata su una solida conoscenza di ciò che funziona nella percezione visiva. Le competenze necessarie possono essere apprese e la progettazione visiva è cruciale per garantire una comunicazione efficace dei dati. Trasformando dati complessi in rappresentazioni grafiche intuitive, questa disciplina consente di identificare pattern, trend e anomalie che sarebbero difficili da individuare attraverso un'analisi puramente numerica. Gli obiettivi e le sfide di questa disciplina sono principalmente:

- Facilitazione della comprensione: la visualizzazione rende i dati più accessibili a un pubblico più ampio, anche a coloro che non hanno competenze tecniche specifiche. Tuttavia non sempre risulta immediato trovare le giuste modalità di rappresentazione dei dati;

- Supporto al decision making: permette di individuare opportunità e rischi, di valutare diverse opzioni e di prendere decisioni più informate e strategiche;

- Comunicazione efficace: consente di comunicare in modo chiaro e conciso informazioni complesse, sia all'interno che all'esterno dell'organizzazione;

- Ampio spettro di applicazioni: la visualizzazione dei dati trova applicazione in numerosi settori, dalla finanza al marketing, dalla sanità all'industria manifatturiera.

Esistono numerosi strumenti e software per la creazione di visualizzazioni, tra cui Tableau, Infogram, ChartBlocks, Datawrapper, Google Charts e Microsoft Power BI. Ognuno di questi strumenti offre una vasta gamma di funzionalità e si adatta a diverse esigenze. Le tecniche di visualizzazione spaziano da grafici semplici a rappresentazioni più complesse, come mappe, diagrammi di rete e visualizzazioni multidimensionali. La visualizzazione dei dati è uno strumento indispensabile per le organizzazioni moderne. Essa consente di trasformare i dati in conoscenza, di migliorare la comunicazione e di supportare il processo decisionale. Investire in strumenti e competenze in questo campo è fondamentale per rimanere competitivi in un mondo sempre più data-driven.

5.1 Codifica visiva dei dati per una percezione rapida

Nella progettazione delle dashboard, è essenziale sfruttare il processo di percezione rapida, noto come elaborazione preattentiva. Questa fase iniziale della percezione visiva avviene al di sotto del livello di coscienza ed è capace di identificare rapidamente specifici attributi visivi, in contrasto con l'elaborazione attenta, che è sequenziale e molto più lenta. Se un insieme di dati non contiene attributi visivi preattentivi, come nel caso di numeri identici per forma e colore, il cervello richiede più tempo per individuare elementi specifici. Invece, quando un attributo preattentivo come il colore differenzia gli elementi, la loro identificazione avviene molto più velocemente. Secondo Colin Ware [27], gli attributi preattentivi sono 17, che nel contesto di elaborazione delle dashboard possono essere ridotti a 11 essenziali per garantire che i dati vengano percepiti e interpretati rapidamente e con facilità, e possono essere organizzati in quattro categorie principali:

- Colore: tonalità del colore e intensità del colore dell'oggetto visivo. E' interessante sottolineare che la percezione del colore associato ad un oggetto è influenzata dal contesto in cui l'oggetto è inserito. La scelta del colore adeguato dello sfondo della dashboard e, di conseguenza, dell' oggetto visivo diventa quindi importante per una chiara visualizzazione degli oggetti;

- Forma: orientamento, lunghezza della linea, larghezza della linea, dimensione, forma geometrica, segni aggiuntivi di riconoscimento, cornice di inquadramento (enclosure). E' evidente l'importanza di questo gruppo di attributi, dal momento che permettono di sottolineare informazioni di interesse presenti nei dati, le quali, senza l'applicazione di particolari parametri relativi a dimensione, forma, ecc. non risulterebbero così evidenti;

- Posizione spaziale: posizionamento bidimensionale all'interno della dashboard. Questo attributo si riferisce al posizionamento dei dati all'interno di un oggetto visivo, come scatter plot ad esempio di dati quantitativi. Non è un attributo arbitrario nel senso che si fonda sui valori stessi dei dati, ma è chiaramente le differenze nel posizionamento 2-D sono le più semplici da percepire;

- Movimento: sfarfallio, ovvero un attributo associato ad un oggetto, come il colore, che cambia tra due valori continuamente oppure l'intero oggetto compare e scompare ripetutamente. Questo tipo di attributi sono dei potenti strumenti per attirare

l'attenzione degli utenti all'interno di dashboard complesse, ma vanno generalmente evitati o comunque utilizzati con parsimonia perché altrimenti possono risultare anche fastidiosi e confusionari. In alcuni casi, come nei contesti di aggiornamenti continui dei dati real-time, questo tipo di attributi risulta efficace perché sottolinea in dashboard perlopiù uguali all'ultima visualizzazione, quali sono i nuovi dati inseriti o, più in generale, le differenze.

L'uso consapevole di questi attributi permette di migliorare significativamente l'efficacia delle visualizzazioni, rendendo immediatamente accessibili le informazioni più rilevanti senza sovraccaricare l'utente di dettagli complessi.

5.2 Principi di Gestalt per la percezione visiva

Nel 1912, la Scuola di Psicologia della Gestalt ha avviato studi pionieristici per comprendere come percepiamo modelli, forme e organizzazione visiva [9]. Il termine tedesco "gestalt" significa semplicemente "forma" o "configurazione", e riflette l'idea che il nostro cervello non interpreta le informazioni visive come entità isolate, ma cerca sempre di strutturare e organizzare ciò che vede. Questi studi hanno portato alla definizione di una serie di principi della percezione, utili anche oggi per descrivere come tendiamo a raggruppare visivamente gli oggetti. Questi principi sono fondamentali nella progettazione delle dashboard, poiché permettono di organizzare visivamente i dati in modo da facilitare la comprensione da parte dell'utente. In un contesto di data visualization, una dashboard ben progettata non solo rende i dati accessibili, ma li presenta anche in modo tale da rendere immediata la loro interpretazione. Per un pubblico esterno, non sempre familiare con i dati grezzi o con i metodi di analisi, un design intuitivo può fare la differenza tra un'informazione compresa e una trascurata. In questo contesto, i principi della Gestalt forniscono una guida preziosa per migliorare l'efficacia delle visualizzazioni di dati. I sei principi della percezione Gestalt applicati alla visualizzazione dei dati possono essere sintetizzati nei seguenti:

- Prossimità (Proximity): gli elementi che si trovano vicini l'uno all'altro tendono a essere percepiti come un gruppo. Questo principio è utile per aggregare visivamente dati associati alla stessa informazione o informazioni simili su una dashboard. Raggruppando elementi vicini tra loro, si facilita la loro lettura e il confronto. Il principio della prossimità può essere anche utilizzato per guidare l'osservatore in una certa direzione di lettura: posizionare oggetti più vicini in senso orizzontale piuttosto che verticale, incoraggia chi guarda a scorrere orizzontalmente gli oggetti. Un esempio di questo fenomeno si può osservare in Figura 5.1;
- Chiusura (Closure): quando parti incomplete di una figura o forma sono posizionate in modo tale che suggeriscano un contorno chiuso, il nostro cervello tende a completare mentalmente l'immagine. Questo principio può essere utilizzato nelle dashboard per ridurre la complessità visiva, lasciando che l'utente percepisca figure complete anche con rappresentazioni semplificate. In tal modo non si appesantisce con troppi elementi la dashboard che in alcuni casi potrebbe diventare estremamente complessa e confondere l'utente.

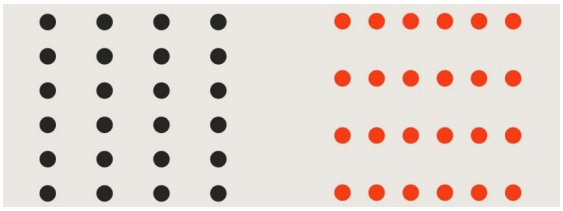

Figura 5.1. Esempio di applicazione del principio di Prossimità di Gestalt. Nella figura di sinistra si è portati in modo naturale a scorrere i punti in senso verticale, mentre nella figura di destra in senso orizzontale. Ciò è dovuto alla vicinanza dei punti in senso verticale nel primo caso ed in senso orizzontale nel secondo.

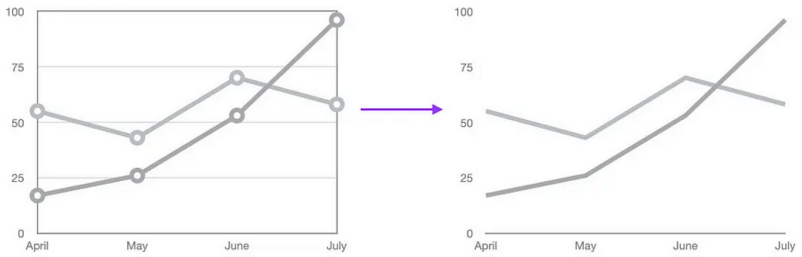

Figura 5.2. Esempio del principio di Chiusura di Gestalt applicato alla Data Visualization. I due grafici forniscono la stessa informazione, ma quello di destra risulta meno complesso e più "leggero" nella visualizzazione.

- Somiglianza (Similarity): gli elementi che condividono caratteristiche simili, come colore, forma o dimensioni, vengono percepiti come associati allo stesso gruppo. Utilizzare colori o forme uniformi per rappresentare tipi simili di dati aiuta a trasmettere immediatamente un messaggio chiaro e coeso.

- Continuità (Continuity): oggetti allineati l'uno con l'altro, o che appaiono in una sorta di continuazione l'uno dell'altro, sono percepiti come appartenenti ad un unico oggetto. Questo principio si applica nella visualizzazione delle tendenze nei dati, come nei grafici a linea, dove la continuità visiva rafforza l'idea di una progressione o di un trend. Si può anche applicare nel contesto di creazione di tabelle riassuntive dei dati: in alcuni casi, si può pensare di eliminare le linee di separazione tra i valori, che spesso appesantiscono la visualizzazione, e raggruppare i dati allineandoli tra loro.

- Involucro (Enclosure): gli oggetti racchiusi all'interno di un contorno o una forma vengono percepiti come appartenenti a un gruppo. L'uso di riquadri o altri elementi visivi per racchiudere gruppi di dati sulla dashboard aiuta a distinguere sezioni separate e a guidare l'attenzione dell'utente. Nel software Power BI, ad esempio,

questo può essere ottenuto creando diverse pagine, ciascuna associata all'insieme di informazioni che si vuole raggruppare.

- Connessione (Connection): quando elementi sono collegati visivamente da linee o altre connessioni grafiche, sono percepiti come associati tra loro. Questo è particolarmente utile per creare connessioni logiche tra elementi diversi di una dashboard, evidenziando le relazioni tra dati differenti. In particolare viene sottolineato il fatto che la percezione prodotta dal raggruppamento via connessione è più forte rispetto a quello ottenuto tramite prossimità o similarità, ma è più debole rispetto a quello prodotto dalla enclosure.

L'applicazione consapevole dei principi della Gestalt permette di creare dashboard visivamente coerenti, dove i dati sono organizzati e comunicati in modo chiaro. Progettare una dashboard efficace non significa solo rappresentare i dati, ma anche facilitare la loro interpretazione. Per il pubblico esterno, spesso non abituato a lavorare con dataset complessi, una visualizzazione basata sui principi della percezione può rendere l'esperienza più intuitiva, garantendo che le informazioni più rilevanti siano evidenziate e facilmente accessibili. In sintesi, una buona dashboard non si limita a mostrare i dati, ma li organizza e li presenta in modo tale da favorire la comprensione e la scoperta di insight, applicando principi psicologici consolidati per ottimizzare l'interazione con l'utente.

5.3 Microsoft Power BI

Nel contesto di questa analisi, è stato scelto Microsoft Power BI come software per la creazione delle dashboard riassuntive delle informazioni estratte. Sebbene originariamente integrato in Excel, Power BI è oggi una piattaforma indipendente che offre una soluzione completa per la Business Intelligence (BI), accessibile anche senza dipendere dalla versione di Microsoft Office [8]. Microsoft ha ascoltato i feedback degli utenti, che desideravano strumenti più flessibili, visualizzazioni avanzate e un'esperienza mobile, sviluppando una piattaforma che permette di condividere report in maniera semplice e intuitiva. Le potenzialità di Power BI si estendono a una vasta gamma di fonti dati, incluse quelle in cloud, facilitando l'integrazione e l'aggiornamento dei dati in tempo reale. Oltre a consentire l'analisi e la visualizzazione dei dati attraverso un'interfaccia intuitiva, Power BI si integra facilmente con linguaggi di programmazione e database come quelli utilizzati in questa tesi, ovvero Python e PostgreSQL. Grazie a queste integrazioni, è possibile arricchire ulteriormente le analisi e sfruttare la potenza dei modelli di machine learning e delle analisi avanzate, connettendo Power BI a database relazionali per estrarre e gestire dati su larga scala.

5.4 Progettazione delle dashboard

Sulla base dei principi esposti in questo capitolo è stata condotta una attenta progettazione delle dashboard riepilogative fondate sulle informazioni estratte dall'analisi. In questa sezione si desidera riassumere le scelte apportate, mentre per la visualizzazione delle dashboard si rimanda al Capitolo 6. I tipi di dati raccolti nelle dashboard riguardano prevalentemente dati storici, associati alle serie temporali di diverse misure, come il numero stesso di nascite, decessi, residenti, ma anche questi numeri filtrati in base a precisi valori degli attributi presenti nei dataset. Per questo motivo gli oggetti visivi più utilizzati sono i grafici a linee, poiché sono i più adatti a descrivere dati temporali, in modo da ottenere confronti chiari e immediati. Tali grafici devono mantenere una coerenza in termini di attributi preattentivi nella rappresentazione tra le diverse sezioni nelle dashboard per aiutare l'osservatore ad orientarsi. Ciò significa associare una certa forma e un determinato colore alla linea relativa ad un certo dataset o a un certo attributo e mantenerla per tutta la dashboard. Dal momento che è stata operata la scelta di suddividere le sezioni per dataset (applicando il principio dell'"Enclosure" visto in Sezione 5.2), chi visualizza avrà chiaro che nella schermata ha tutte le informazioni relative ad un preciso dataset, per questo motivo l'associazione di forma e colore può essere effettuata sulla base degli attributi comuni tra i dataset. Quindi per l'attributo "genere", viene associato il colore blu alla serie storica relativa al genere maschile, mentre il rosa alla femminile, ad esempio. Inoltre, poiché si avrà uniformità nella rappresentazione dei grafici, si può semplificare di molto la struttura per non appesantirla con molte linee, sfruttando le osservazioni fatte per il principio di Continuità. Per quanto riguarda i dati associati alle costruzioni, poiché si distinguono dagli altri dati associati a persone e dal momento che in questo dataset si considera un periodo temporale ridotto, si decide di descrivere la serie storica con un grafico a barre, di più chiara visualizzazione. Per le informazioni di tipo "spaziale" viene impiegata la funzionalità messa a disposizione da Power BI dell'oggetto "Mappa". Questo oggetto permette di ottenere una visualizzazione geografica del Comune di Milano, associando delle bolle ad ogni NIL che hanno dimensioni dipendenti dalla misura che viene fornita. Questo oggetto permette di ottenere un quadro più chiaro della distribuzione delle nascite, ad esempio, rispetto a istogrammi, grafici a linee o altro. Inoltre, questo oggetto dà la possibilità di navigare all'interno della mappa, zoomando su zone precise di interesse.
Per garantire una ordinata navigazione nelle dashboard si è pensato di suddividere la presentazione in due livelli:

- Il primo livello fornisce una overview sulle serie storiche relative alle sorgenti analizzate, una sorta di "home" della dashboard. Le informazioni principali sono ritenute essere, in ordine, prima quelle relative a residenti, quindi nascite e decessi, poi dati anagrafici e censuari ed infine costruzioni. Per questo motivo verrà proposta una evidente rappresentazione della serie dei residenti in alto a sinistra. Quindi si proporrà un oggetto relativo al confronto tra le serie storiche di nascite e decessi in alto a destra. Si è deciso di inserire anche due *callout* relativi al numero di residenti e alla differenza tra nascite e decessi, per dare dei valori numerici in base a un filtro inserito in alto a destra relativo agli anni, ben riconoscibile con bordi neri, nella

forma di un elenco verticale. In particolare sulla scheda relativa alla differenza tra nascite e decessi si è impostata una formattazione condizionale del colore relativo al valore: se positivo assume colorazione verde, se negativo, rossa. Per gli oggetti relativi ai residenti, nascite e decessi è stato impostato uno sfondo grigio per dare maggiore risalto a questi dataset e viene impostato un contorno degli oggetti diverso per residenti e nascite-decessi per renderli immediatamente associabili. In basso si includeranno le informazioni relative al confronto tra numero di famiglie secondo il dataset anagrafico e censuario e l'andamento del numero di costruzioni ad uso abitativo negli anni.

- Il secondo livello si compone di 5 sezioni, nelle quali si proporranno degli insights più specifici relativi ai singoli dataset. A ciascuna sezione è assegnato un colore comune su cui si basa la dashboard, immediatamente associabile all'argomento che riguarda il dataset, per facilitare la navigazione tra le sezioni. Per le nascite il verde, per i decessi il rosso, per i residenti il blu, per le costruzioni il giallo, per dati anagrafici e censuari l'arancione. Le sezioni hanno la seguente struttura:

  - Per le informazioni relative alle nascite si forniranno 4 oggetti visivi: la serie storica delle nascite, l'andamento del genere della nascita negli anni, mappa geografica con le densità di nascite per quartiere. In questa sezione sarà inserito anche un filtro per selezionare anni specifici e interagire con la mappa per visualizzare meglio i dati. Gli attributi relativi alla cittadinanza dei genitori non vengono ritenuti di interesse e non vengono proposti.
  - Per i decessi si agirà in modo analogo alle nascite, anche qui escludendo l'attributo relativo alla cittadinanza dal momento che è poco informativo.
  - Il dataset dei residenti includerà la serie storica, un oggetto che permetta di visualizzare i NIL più popolosi con il numero di residenti relativo (una sorta di tabella interattiva) ed un oggetto geografico anche qui per visualizzare anche sulla mappa la densità dei residenti nei NIL. Anche in questo caso viene proposto un filtro per selezionare anni specifici.
  - Per quanto riguarda le costruzioni si è deciso di riportare 4 oggetti relativi alle informazioni principali, ritenute più importanti: la serie storica della somma di superficie utile abitabile escludendo le altre informazioni fortemente correlate con essa; la serie storica della media del consumo energetico per costruzione; un confronto con grafici a linee che dell'impiego negli anni delle tecnologie rinnovabili individuate dalle variabili booleane nel dataset; un oggetto geografico che mostra la densità di costruzioni per quartiere.
  - I dati anagrafici e censuari vengono raccolti in un'unica sezione, chiamata 'Tipologie Famiglie Residenti', dove si propone un confronto tra le due serie storiche, quindi si vanno ad esibire degli oggetti relativi soltanto ai dati anagrafici degli attributi 'genere capofamiglia', 'numero componenti', per non riportare grafici troppo confusionari, e anche qui una tabella in cui dare le informazioni relative alle cittadinanze, con relativo numero di nuclei familiari residenti e anno. Questa tabella risulta il modo più immediato per ordinare i dati per numero

famiglie, comprendendo quali sono le cittadinanze più frequenti e in quali anni. Infine anche qui si inserisce una mappa interattiva. In questo caso i filtri che vengono proposti sono sia relativi agli anni che alla cittadinanza, così da poter osservare specifici andamenti negli attributi. Per il filtro relativo alla cittadinanza dei nuclei familiari, dato l'alto numero di valori possibili, si è optato per un menù a discesa, nel quale è stata inserita una barra di inserimento per velocizzare le operazioni di ricerca.

Per facilitare la navigazione all'interno delle dashboard sono stati inseriti anche degli intuitivi pulsanti che permettono di passare facilmente da una sezione all'altra.

Parte III

Risultati e conclusioni

Capitolo 6

Risultati

L'analisi condotta in questo elaborato ha consentito di ottenere alcuni risultati di interesse, dapprima rilevati tramite una approfondita analisi esplorativa, e quindi in seguito riscontrati dall'utilizzo di modelli per la descrizione dei dati e per la previsione a partire dagli stessi. In questo capitolo verranno sintetizzati i risultati di maggiore rilevanza, per una più chiara e completa comprensione dell'analisi.

6.1 Risultati dell'analisi esplorativa

Per quanto riguarda l'andamento delle nascite è stato rilevato un importante trend negativo che ha evidenziato una variazione percentuale pari al -23,2% del numero di nascite dal 2003 fino al 2023. La prevalenza delle nascite sull'arco temporale proviene da genitori entrambi italiani (circa il 65%) o entrambi stranieri (circa il 25%) e si hanno più nascite di genere maschile che di genere femminile in ogni anno preso in considerazione della serie storica. I NIL da cui proviene il maggior numero di nascite su tutto l'arco temporale 2003-2023 sono 'Buenos Aires - Porta Venezia - Porta Monforte', 'Bande Nere', 'Loreto - Casoretto - NoLO'. Tra i pochi quartieri che hanno visto un incremento nel numero di nascite tra il 2003 e il 2023, la variazione percentuale positiva maggiore si ritrova per il quartiere 'Maggiore - Musocco - Certosa', con un incremento da 50 a 154 nascite tra il 2003 e il 2023 (+208 %), mentre tra le diminuzioni più importanti si sottolinea quella relativa a 'Buenos Aires - Porta Venezia - Porta Monforte', che come detto è il quartiere che più contribuisce al computo totale delle nascite per la popolazione di Milano, con una decrescita da 632 a 379 nascite (-40,03 %).
Per il dataset relativo ai decessi si è osservato un andamento pressoché costante, con l'eccezione del picco nel 2020, con tutta probabilità attribuibile alla pandemia di COVID-19. La quasi totalità dei decessi si attribuisce a persone con cittadinanza italiana (circa il 98%). Si osserva che i decessi di genere femminile sovrastano quelli di genere maschile su tutto il periodo dal 2003 al 2023 (circa il 53% dei decessi annuali in media proviene dal genere femminile). La maggior parte dei decessi proviene dai NIL 'Buenos Aires - Porta Venezia - Porta Monforte', 'Bande Nere', 'Città studi' e 'Niguarda - Ca granda - Prato Centenaro - Q.re Fu'. L'incremento di decessi più importante tra il 2003 e il 2023

è attribuibile al quartiere 'Bovisasca', passando da 80 a 124 (+50%), mentre la decrescita maggiore si è osservata per 'Porta Ticinese - Conchetta', da 177 a 124 (-29,94%). Per quanto riguarda l'anno 2020, in cui si riscontra il picco associato alla pandemia, si può constatare che il numero di decessi maggiori si ha per i quartieri 'Bande Nere', 'Buenos Aires - Porta Venezia - Porta Monforte' e 'Q.re Gallaratese - Q.re San Leonardo - Lampugnano ', con, rispettivamente, 920, 764 e 673 decessi.
Il dataset dei residenti ha evidenziato un andamento non chiaro dal 1999 al 2007, con un picco negativo nel 2003, ma assume un trend positivo dal 2008 in poi. I quartieri con più residenti sono stati ritrovati in 'Buenos Aires - Porta Venezia - Porta Monforte', 'Bande Nere', 'Città studi' e 'Lodi - Corvetto'. Il quartiere che ha visto la crescita nel numero di residenti decisamente maggiore rispetto al resto dei NIL è 'Adriano', che è passato da una popolazione di 9.813 a 18.022 residenti (+83,65%) seguito da 'Dergano', crescendo da 17.426 a 23.875 (+37,01%) e quindi 'Affori', da 19.502 a 25.669 (+31,62%). D'altra parte, le decrescite percentuali sono in modulo inferiori rispetto alle crescite percentuali, confermando il carattere generale crescente della serie, ad ogni modo i quartieri che hanno subito una decrescita percentuale elevata nel numero di abitanti residenti sono i quartieri 'Duomo', passando da 19.817 a 16.608 residenti (-16,19%), quindi 'Barona', da 19.575 a 16.658 (-14,90%), ed infine 'Porta Ticinese - Conca Del Naviglio', passato da 22.471 a 19.999 (-11,00%).
Sono state estratte delle considerazioni anche dal dataset relativo alle costruzioni dal 2011 al 2023. L'analisi di correlazione tra le variabili di tipo numerico ha evidenziato una forte correlazione positiva tra il gruppo di variabili relativo alle dimensioni della costruzione: numero di abitazioni, numero di stanze, numero di vani accessori interni, numero di piani dell'edificio, volume totale v/p e superficie utile abitabile. Per quanto riguarda il consumo energetico medio annuale si è osservato un trend crescente, che evidenzia un consumo maggiore delle costruzioni più recenti. Per le variabili relative alla "sostenibilità ambientale" è stata osservata una crescita di costruzioni aventi impianto fotovoltaico e pompe calore, un andamento pressoché costante per quanto riguarda l'impianto solare e geotermico, mentre una decrescita nell'impiego di caldaie a condensazione. E' stato osservato inoltre, attraverso tabelle di contingenza, che nel 60,9% delle costruzioni sono presenti contemporaneamente impianto fotovoltaico e pompe di calore. Al contrario si ha che per il 65,3% delle costruzioni non si hanno insieme l'impianto solare termico e la caldaia a condensazione. E' stato inoltre osservato, attraverso il test di indipendenza χ^2, che i dati rifiutano l'ipotesi di indipendenza tra le variabili booleane, tranne per la coppia fotovoltaico e solare termico, per la quale non si può rifiutare l'ipotesi nulla di indipendenza. Tuttavia, valutando la statistica V di Cramér è stato osservato che la relazione sottostante tra le variabili ha un basso livello di intensità, indice di una associazione debole tra le variabili. Ad ogni modo, il grado di associazione più alto si ottiene per la coppia fotovoltaico - pompe di calore, con un valore della statistica pari a 0,229.
I dati anagrafici e censuari, relativi ai nuclei familiari presenti sul territorio di Milano, evidenziano entrambi un trend crescente negli anni, come è chiaro aspettarsi dall'andamento dei residenti, con i dati anagrafici che sovrastimano sempre i dati censuari. E' stato notato che l'andamento del numero di nuclei familiari con capo famiglia femmina è in trend crescente e si sta avvicinando negli ultimi anni a quello con capo famiglia maschio, che rimane costante e sempre sopra il precedente. I nuclei familiari con capo famiglia avente

età compresa tra i 34 e i 65 anni sono in crescita, così come i nuclei familiari composti da un solo componente. I dati censuari, evidenziano, a differenza dei dati anagrafici, un trend crescente nel numero di nuclei familiari aventi capo famiglia con età '80 anni e più' e un andamento costante per '18-35 anni'. Per quanto riguarda le variazioni percentuali del numero di nuclei familiari nei NIL si hanno comportamenti sostanzialmente analoghi a quelli evidenziati nel caso dei residenti per quartiere. Per la nazionalità dei nuclei familiari, è stato osservato che le più frequenti sul totale negli anni sono, dopo l'Italiana, le cittadinanze associate a Egitto, Filippine, Cina, Perù, Sri Lanka. In particolare dal 1999 al 2023 è stata riscontrata la crescita percentuale maggiore nelle nazionalità Ucraina, passata da 42 nuclei familiari a 6.750, Bangladesh, passata da 569 a 8.723, e Romania, passata da 705 a 9.102. D'altra parte, per la nazionalità associata al Regno Unito, è stata osservata una decrescita pari al -30,87%, passando da 2.271 nuclei familiari a 1.570. Si evidenzia in particolare il passaggio dei nuclei familiari di nazionalità Cinese, che presenta valori assoluti maggiori, da 2.967 nel 1999, a 11.527 nel 2013, a 19.425 nel 2023 (+554,70% tra il 1999 e il 2023, +68,52% tra il 2013 e il 2023). Nell'ultimo decennio si riscontra una decrescita percentuale importante per le nazionalità associate a Paesi Sudamericani: Ecuador (-36,75%), Perù (-19,44%), Brasile (-12,66%).
I dati relativi al movimento naturale e migratorio hanno evidenziato che il tasso di crescita naturale per la popolazione di Milano è negativo dal 1975 circa ad oggi e che mostra un trend negativo negli ultimi anni dovuto soprattutto alla diminuzione delle nascite piuttosto che ad un aumento dei decessi. Inoltre è stato rilevato che il tasso migratorio, perlopiù positivo negli ultimi anni, ha un impatto molto più forte sul tasso di crescita totale della popolazione rispetto al tasso di crescita naturale e ciò ha spiegato la crescita della popolazione residente negli ultimi anni nonostante la diminuzione del numero di nascite.

6.2 Risultati dell'applicazione dei modelli predittivi

Per dare un'idea più precisa dell'andamento futuro della demografia milanese, si propongono in questa sezione i risultati ottenuti dall'applicazione di modelli predittivi alla serie storica dei residenti nella popolazione di Milano che viene fornita da ISTAT. Per la serie storica di training sono stati applicati i modelli Simple Exponential Smoothing (SES), Double Exponential Smoothing (DES) o Holt-Winters Model, ARIMA ottenuto tramite la funzione `auto_arima` (Auto ARIMA), che ha fornito un modello ARIMA(1,2,1) e ARIMA con parametri selezionati sulla base di considerazioni apportate analizzando i grafici di autocorrelazione e autocorrelazione parziale (Manual ARIMA), che ha fornito un ARIMA(1,2,3). I risultati delle performance dei modelli vengono presentati nella Tabella 6.1, dove si può osservare che i modelli SES e DES forniscono performance migliori rispetto ai modelli ARIMA in termini di AIC e quindi di adattamento ai dati e forniscono anche previsioni più precise per MAE, MAPE e RMSE. In particolare il modello auto_arima fornisce un AIC migliore rispetto al modello ARIMA manuale ma consegna una previsione peggiore. In Figura 6.1 si possono osservare i modelli adattati con le relative previsioni e intervalli di confidenza sulle previsioni. Si sottolinea che il modello SES fornisce previsioni migliori perché non considera il trend negativo mostrato dalla serie di training degli ultimi anni relativi al dataset di training, mentre gli altri modelli ne tengono conto.

Modello	MAE	MAPE	RMSE	AIC
SES	36.419,05	2,70%	47.035,68	2.434,49
DES	55.258,07	4,10%	66.315,07	2.353,86
Auto ARIMA	119.391,88	8,93%	136.970,61	2.668,68
Manual ARIMA	74.638,90	5,50%	83.438,75	2.844,43

Tabella 6.1. Risultati ottenuti dai modelli applicati alla serie storica dei residenti ottenuti dai dati forniti da ISTAT.

Per quanto riguarda i risultati ottenuti dall'adattamento dei modelli che considerano la natura demografica della serie storica analizzata, si ottengono i risultati in Tabella 6.2. Si noti che per i modelli in cui qualche parametro viene fissato, le stime degli altri parametri sono ottenute sempre tramite le stime di massima verosimiglianza. I risultati in termini di previsione risultano non idonei se si considerano le metriche MAE, MAPE e RMSE, confrontandoli con i risultati ottenuti dai metodi precedentemente citati. Tuttavia, dal momento che questi modelli forniscono delle proiezioni per l'andamento demografico nel lungo termine, dando un'idea della tendenza futura della serie, per questa classe di modelli risulta più utile un confronto grafico e in termini di AIC. Chiaramente la stima tramite la minimizzazione della somma dei residui quadratici per tutti i parametri permette di avere le performance migliori in termini di adattamento ai dati rispetto ai modelli in cui qualche parametro viene scelto in altre modalità. In particolare il modello di Richards fornisce le performance migliori. Si osservi che gli AIC che consegnano i modelli precedenti sono migliori rispetto a tutti quelli dei modelli per serie storiche demografiche, tranne per il modello di Richards, per cui si rileva un AIC più alto rispetto all'ARIMA manuale. In Figura 6.2 si riportano i grafici relativi all'adattamento dei modelli ai dati di training.
In base alle considerazioni sulle performance riscontrate per i modelli, si riportano due proiezioni future per la popolazione residente di Milano per i prossimi 50 anni: la prima proveniente dai modelli classici per serie storiche, utilizzando il modello Double Exponential Smoothing; la seconda appartenente ai modelli basati sulle teorie demografiche classiche, utilizzando il modello di Richards. Entrambi i modelli vengono adattati in questo caso sull'intera serie storica. La previsione fornita dal modello DES evidenzia un trend decrescente che porterà la popolazione dal valore nel 2022 di 1.358.420 ad assumere nel 2072 il valore di 1.305.758, con una variazione percentuale rispetto al 2022 del -3,88% (Figura 6.3). La previsione fornita dal modello di Richards invece consegna una proiezione futura in cui la popolazione avrà un trend crescente tendendo al valore massimo di 1.466.534 residenti nel 2072, con una variazione percentuale pari a +7,96% (Figura 6.4).

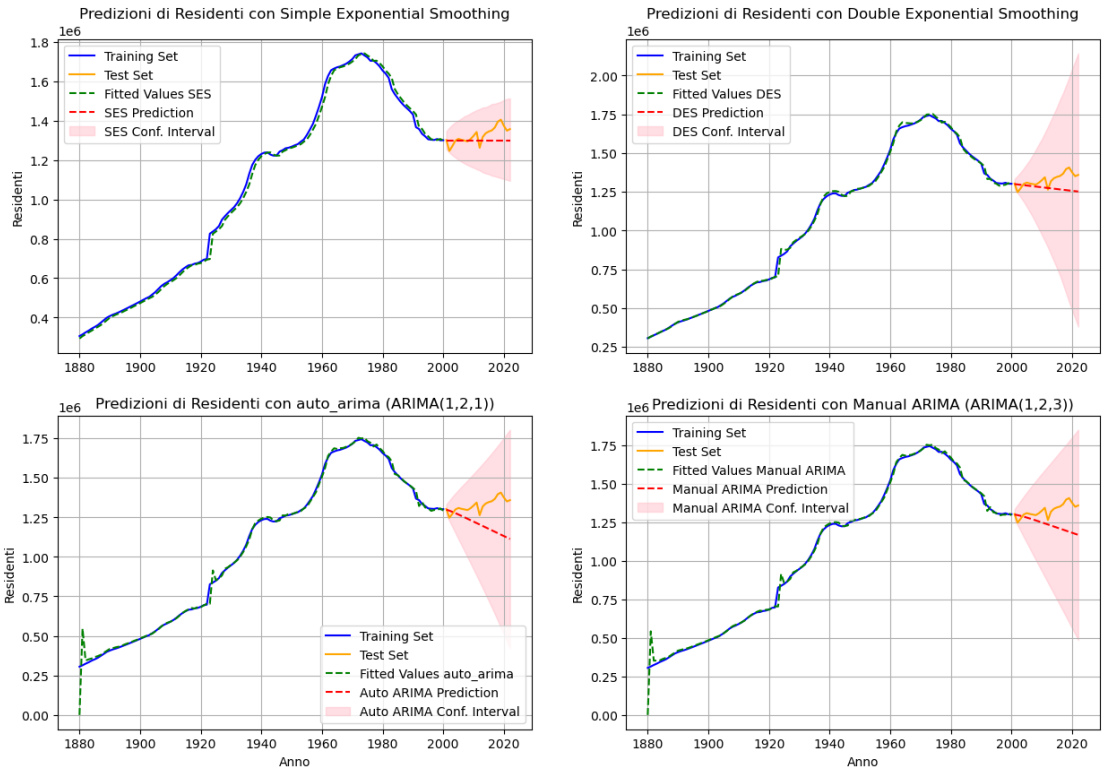

Figura 6.1. Previsioni fornite dai modelli adattati sul dataset di training relativo all'andamento dei residenti a Milano fornito da ISTAT.

Modello	MAE	MAPE	RMSE	AIC
Malthus 1 (least squares)	751.687,23	36,02%	758.743,70	2.997,48
Malthus 2 (avg Growth rate)	949.360,69	41,49%	960.112,02	3.006,20
Verhulst 1 (least squares)	273.542,11	17,11%	275.871,06	2.848,62
Verhulst 2 (max Growth rate)	79.859,57	5,68%	89.377,84	3.019,02
Verhulst 3 (K fixed)	370.917,69	21,87%	372.253,16	2.859,69
Richards 3 (least squares)	218.889,00	14,17%	222.554,24	2.799,68

Tabella 6.2. Risultati ottenuti dai modelli adatti a dati demografici applicati ai dati forniti da ISTAT sui residenti a Milano dal 1880 al 2022.

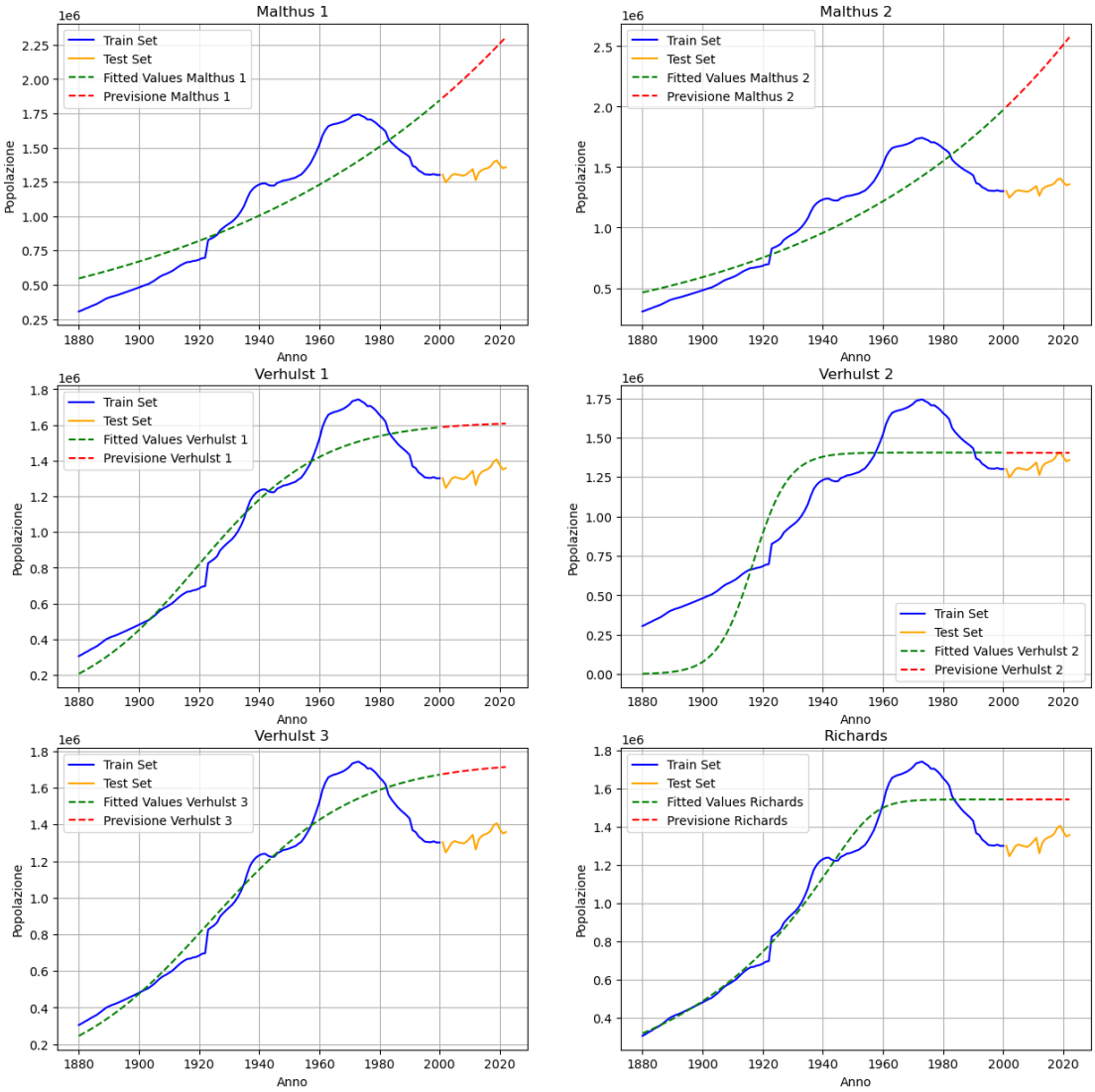

Figura 6.2. Plot dei modelli per dati demografici adattati sul dataset di training.

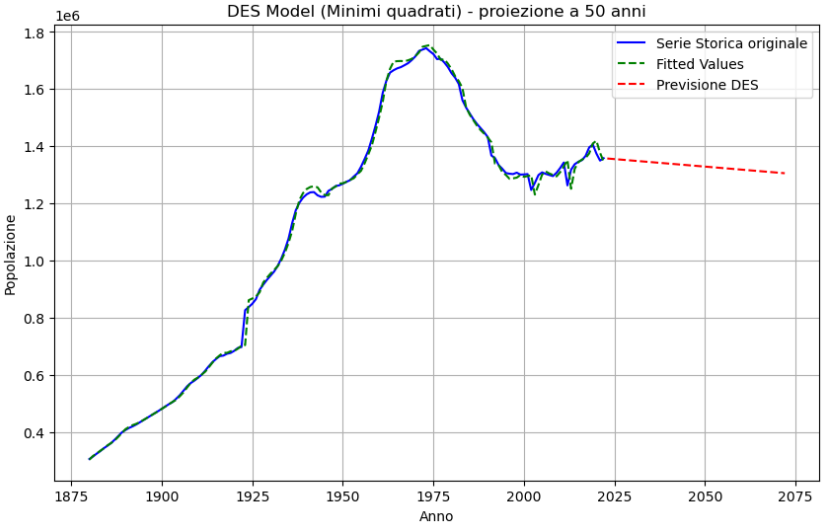

Figura 6.3. Proiezione dell'andamento della popolazione fino al 2072, sulla base del Modello Double Exponential Smoothing.

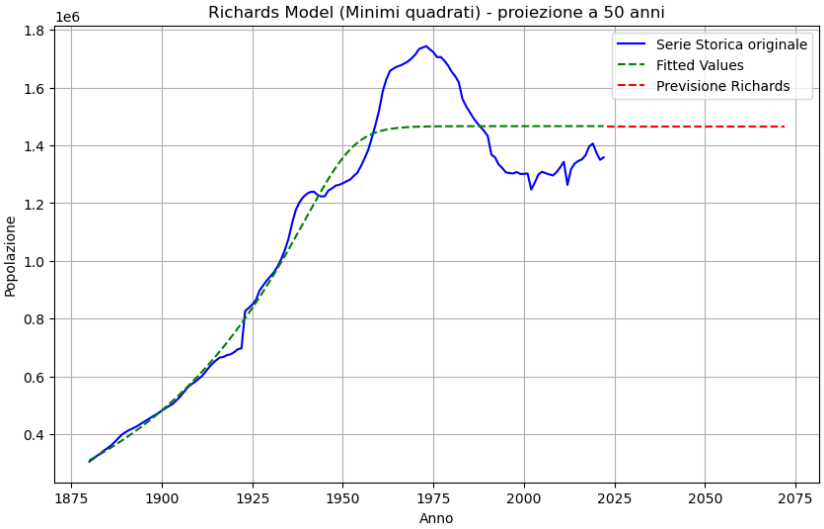

Figura 6.4. Proiezione dell'andamento della popolazione fino al 2072, sulla base del Modello di Richards.

6.3 Dashboard implementate per la Data Visualization

In questa sezione si vanno ad esibire le dashboard costruite in Power BI, sulla base della progettazione descritta nella Sezione 5.4. In Figura 6.5 si mostra la dashboard 'overview', nella quale si riconoscono le informazioni principali relative alle serie storiche dei dataset analizzati. Cliccando nelle opzioni relative al filtro sugli anni, si ottengono i valori corrispondenti nei callout presenti sulla dashboard, permettendo di avere una immediata idea sulla situazione residenti-nascite-decessi relativa a quell'anno (Figura 6.6). Per quanto riguarda il secondo livello di profondità del report, si osservano nelle Figure successive le dashboard create per ciascun dataset, mantenendo la coerenza grafica su ciascuna sezione, così da aiutare e orientare nella navigazione chi consulta, come spiegato già in precedenza.

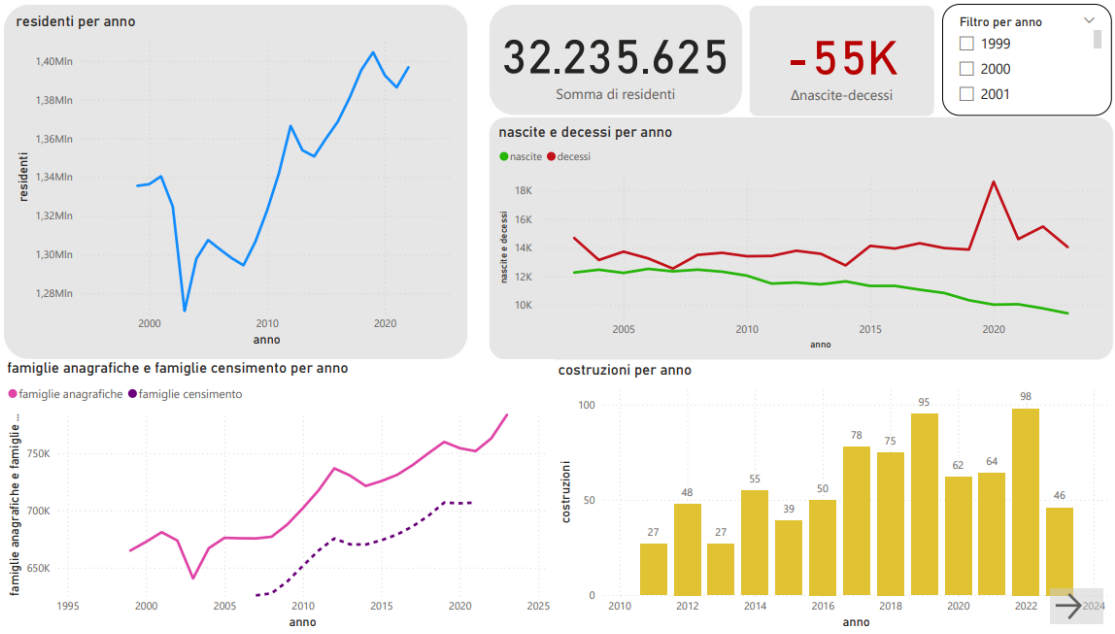

Figura 6.5. Dashboard 'overview', contenente le informazioni principali dei dataset analizzati.

Figura 6.6. Dettaglio della schermata 'overview' dove è stato selezionato l'anno 2022 nel filtro.

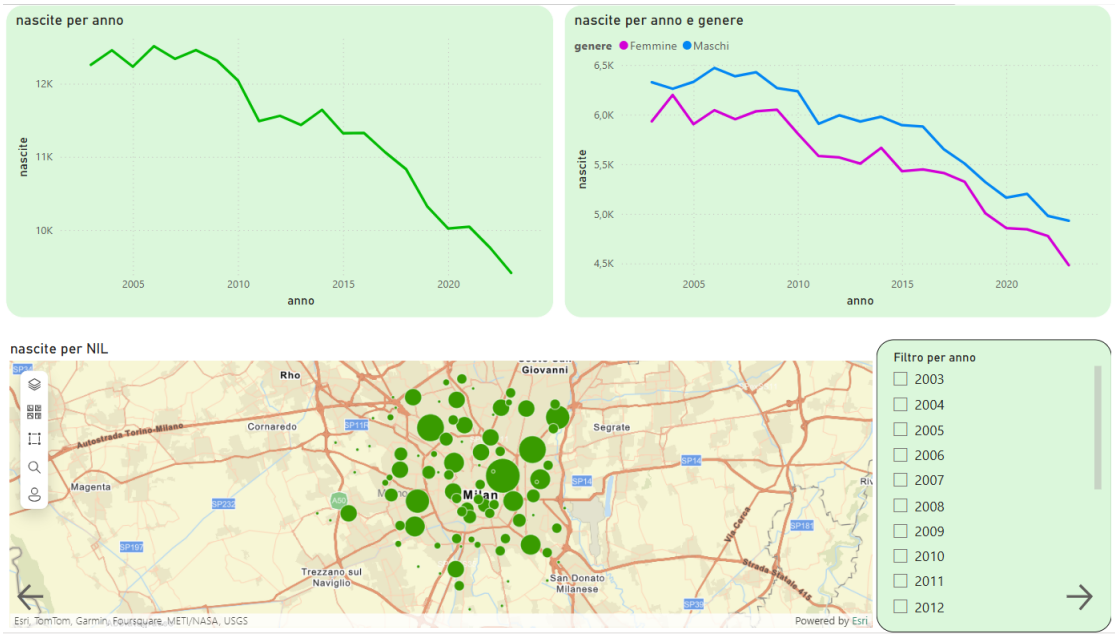

Figura 6.7. Dashboard relativa al dataset nascite.

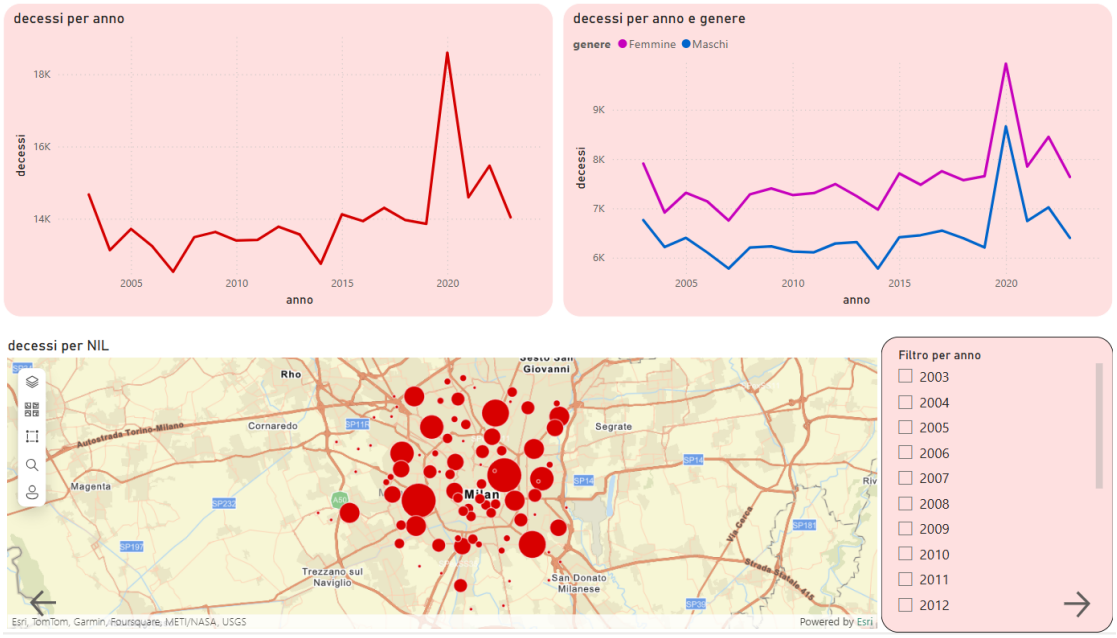

Figura 6.8. Dashboard relativa al dataset decessi.

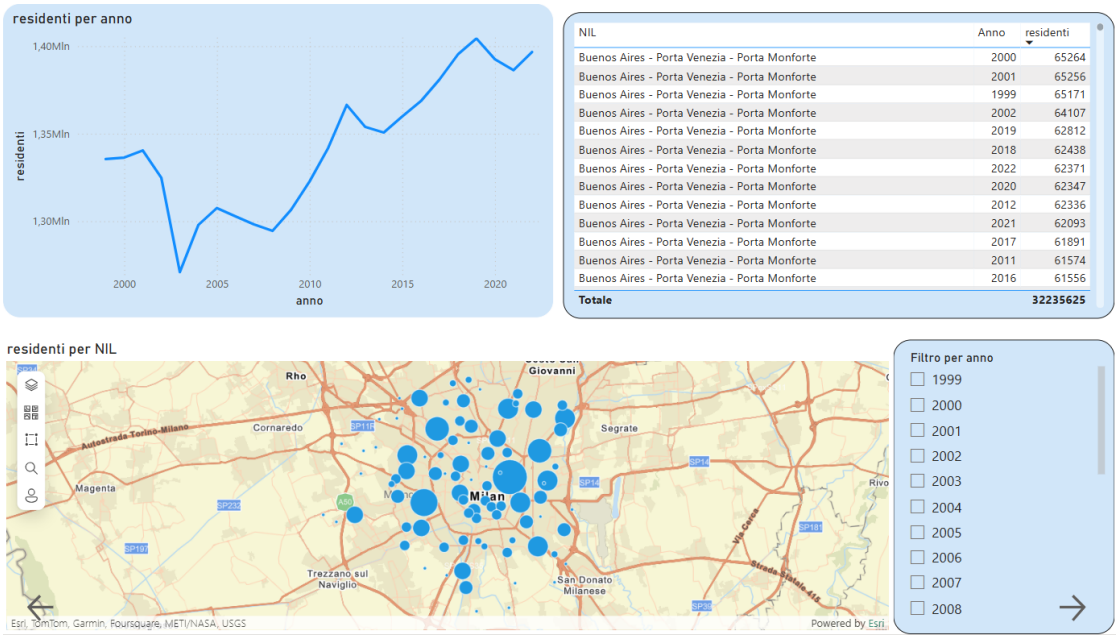

Figura 6.9. Dashboard relativa al dataset residenti.

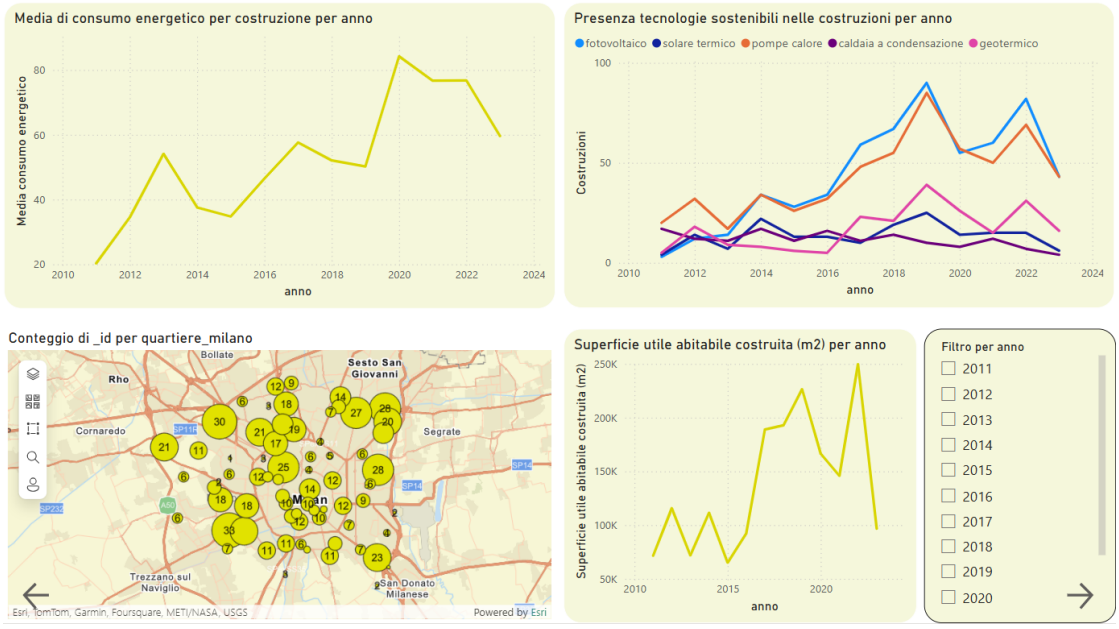

Figura 6.10. Dashboard relativa al dataset costruzioni.

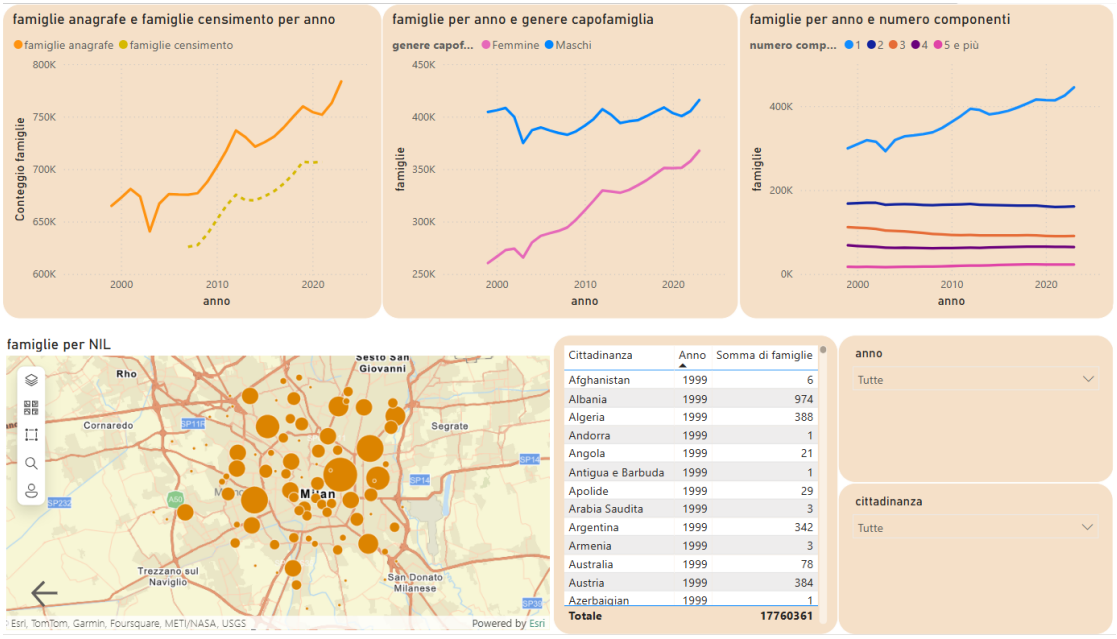

Figura 6.11. Dashboard relativa ai dataset delle tipologie di famiglie secondo i dati anagrafici e censuari.

Capitolo 7

Conclusioni

L'approccio full stack illustrato in questo elaborato ha permesso di tracciare l'intero processo di elaborazione dei dati demografici, partendo dalla loro estrazione fino alla presentazione delle informazioni ricavate. L'analisi si è focalizzata sulla demografia del Comune di Milano e ha integrato diverse fonti di dati, includendo nascite, decessi, residenti, costruzioni, tipologie di famiglie provenienti dall'anagrafe e dai censimenti, insieme a dati relativi al movimento naturale e migratorio della popolazione. Nella prima fase di Data Integration, sono stati descritti i flussi implementati tramite Talend, che hanno permesso l'aggiornamento continuo delle informazioni attraverso il metodo di Ingestion Delta. Si è evidenziato, in particolare, che per ottimizzare questo processo, sarebbe utile avere un campo di tipo timestamp contenente l'ultima modifica del record nella sorgente. Questo campo consentirebbe di filtrare i dati da integrare, includendo solo quelli con data di modifica o inserimento più recente rispetto all'ultimo aggiornamento effettuato. In questa fase sono stati esposti anche i criteri adottati per garantire la qualità dei dati e ottimizzare i processi di acquisizione, migliorando la consistenza e l'affidabilità delle informazioni nel tempo. La successiva fase di Data Analysis ha visto l'impiego di Python per condurre analisi descrittive e predittive. In primo luogo, l'analisi descrittiva ha permesso di individuare le caratteristiche principali e più rilevanti relative ai dati, fornendo un quadro chiaro delle variabili più significative, utili per la composizione delle dashboard riassuntive finali per descrivere la situazione generale del contesto milanese. A questo è seguita un'analisi predittiva sulla base della serie storica dei residenti a Milano dal 1880 al 2022, tramite l'impiego di modelli statistici comunemente utilizzati nell'ambito di serie storiche demografiche, per ottenere proiezioni sulle tendenze future. Sono stati selezionati due modelli, sulla base di un confronto tra le performance ottenute dalle diverse metodologie descritte, per poter consegnare due proiezioni. La prima proveniente dal modello Double Exponential Smoothing, il quale consegna un trend lievemente decrescente che prevede una popolazione residente nel 2072 pari a 1.305.758 persone, con un decremento rispetto alla popolazione attuale di 1.358.420 pari al -3,88%. La seconda proiezione proviene dal modello di Richards, il quale appartiene alla famiglia dei modelli logistici generalizzati, che consegna un trend crescente che nel 2072 consegna una popolazione pari a 1.466.534, con un incremento rispetto alla popolazione attuale del +7,96%. Infine, nella fase di Data Visualization è stato progettato un insieme di dashboard, sviluppato secondo principi di

chiarezza e usabilità, e realizzato su Power BI, comprendente le informazioni principali estratte dall'analisi. Questo processo ha permesso di rendere accessibili e comprensibili i risultati delle analisi, fornendo strumenti visivi che facilitano la comunicazione dei dati demografici ad un pubblico esterno. La limitazione del numero di osservazioni disponibili a livello di NIL, ha portato a concentrare le previsioni sulla serie storica complessiva dei residenti, offrendo un'indicazione del trend generale della popolazione di Milano. Tuttavia, uno sviluppo futuro dell'analisi demografica, in uno scenario in cui si abbia a disposizione un quantitativo adeguato di dati associati a ciascun NIL, potrebbe integrare modelli statistici e di machine learning più complessi su scala locale, per una rappresentazione più granulare e mirata delle dinamiche urbane. Questo permetterebbe di tenere conto di una molteplicità di fattori socio-economici, demografici e urbanistici attualmente esclusi, fornendo proiezioni più precise e utili per il policy-making e la pianificazione urbana. L'analisi demografica, come quella condotta in questo studio, rappresenta uno strumento utile per comprendere le trasformazioni profonde che interessano le città contemporanee. La capacità di anticipare i cambiamenti demografici è cruciale per affrontare le sfide legate all'invecchiamento della popolazione, all'urbanizzazione e alla sostenibilità. L'analisi demografica si configura come un potente strumento di conoscenza e di intervento, in grado di fornire informazioni preziose e utili ai decisori politici, agli operatori economici e alla collettività nel suo complesso, per dare forma a considerazioni oggettive sulle dinamiche che interessano un contesto urbano.

Bibliografia

[1] Muluken Admasu, Ashager Adane, and Mulugeta Tesfa. Estimation of Growth model for Population of Ethiopia using least square method. *IOSR Journal of Mathematics*, 16:1–11, 08 2020. doi: 10.9790/5728-1604050111.

[2] Ahmad Aimran. A comparison between single exponential smoothing (ses), double exponential smoothing (des), holt's (brown) and adaptive response rate exponential smoothing (arres) techniques in forecasting malaysia population. *Global Journal of Mathematical Analysis*, 2:276–280, 09 2014. doi: 10.14419/gjma.v2i4.3253.

[3] H.T. Banks and Michele L. Joyner. Aic under the framework of least squares estimation. *Applied Mathematics Letters*, 74:33–45, 2017. ISSN 0893-9659. doi: https://doi.org/10.1016/j.aml.2017.05.005.

[4] Anthony Bracken and Henry Tuckwell. Simple mathematical models for urban growth. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 438:171–181, 07 1992. doi: 10.1098/rspa.1992.0100.

[5] Jorge M. Bravo and Edviges Coelho. Forecasting subnational demographic data using seasonal time series methods. 2019.

[6] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, 1991. ISBN 9781441903198.

[7] AnHai Doan. *Principles of Data Integration*. Morgan Kaufmann, Waltham, Mass, 1st edition edition, 2012. ISBN 0-12-416044-1.

[8] A. Ferrari and M. Russo. *Introducing Microsoft Power BI*. Microsoft Press, 2016. ISBN 9781509302284.

[9] Stephen Few. *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media, Inc., 2006. ISBN 0596100167.

[10] Idemudia Esosa G. and Ojo Oluwadare O. A comparative analysis of growth models on nigeria population. *FUDMA JOURNAL OF SCIENCES*, 7(6):373 – 381, Feb. 2024.

[11] Lili He and Zhao Jin. The population predicting based on the curve fitting least square method. 01 2015. doi: 10.2991/amcce-15.2015.258.

[12] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014. ISBN 9780987507105.

[13] Saeed Imani, Yagob Dinpashoh, Esmail Asadi, and Ahmad fakheri fard. A mathematical population growth model for urban water security management and sustainability in cities (case study: Tabriz city). 7 2023. doi: 10.20944/preprints202307.1669.v1.

[14] Indicatori demografici. URL https://demo.istat.it/tavole/?t=indicatori.

[15] Ufficio Stampa Istat. Migrazioni interne e internazionali della popolazione residente anni 2022-2023. Technical report, ISTAT, 2024.

[16] Kamal Kishore and Vidushi Jaswal. Statistics corner: Chi-squared test. *Journal of Postgraduate Medicine, Education and Research*, 57:40–44, 04 2023. doi: 10.5005/jp-journals-10028-1618.

[17] Maurizio Lenzerini. Data integration: A theoretical perspective. *In Proceedings of the twentyfirst ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002.

[18] L. Lipkin, David Smith, J.-C Chuang, and J. Michel. Logistic growth model. 1, 12 2001.

[19] Thomas Robert Malthus. *An Essay on the Principle of Population*. History of Economic Thought Books. McMaster University Archive for the History of Economic Thought, 1798.

[20] Samson Linus Manu and Samuel Shikaa. Mathematical modeling of taraba state population growth using exponential and logistic models. *Results in Control and Optimization*, 12:100265, 2023. ISSN 2666-7207. doi: https://doi.org/10.1016/j.rico.2023.100265.

[21] Nuclei di Identità Locale (NIL). URL https://www.pgt.comune.milano.it/psschede-dei-nil-nuclei-di-identita-locale/nuclei-di-identita-locale-nil.

[22] Sofi Parvez and Nur Hosain. Analyzing bangladesh's present patterns in population growth and prediction by arima and exponential smoothing model. 10:41–48, 06 2023.

[23] Popolazione calcolata Istat: movimento naturale (nascite e decessi) e migratorio (iscrizioni e cancellazioni dall'anagrafe) (1880-2022). URL https://dati.comune.milano.it/dataset/ds71-popolazione-movimento-naturale-migratorio.

[24] Yang W. Lee Richard Y. Wang, Mostapha Ziad. *Data Quality*. Springer Science & Business Media, 2006.

[25] F. J. Richards. A flexible growth function for empirical use. *Journal of Experimental Botany*, 10(2):290–301, 06 1959. ISSN 0022-0957. doi: 10.1093/jxb/10.2.290.

[26] A. Tsoularis and J. Wallace. Analysis of logistic growth models. *Mathematical Biosciences*, 179(1):21–55, 2002. ISSN 0025-5564. doi: https://doi.org/10.1016/S0025-5564(02)00096-2.

[27] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3 edition, 2012. ISBN 9780123814647.