Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Tesi di Laurea

Analisi Predittiva per la determinazione del Full Time Equivalent necessario per le attività di backend di un Gruppo Bancario Nazionale

Predictive Analytics for the computation of the Full Time Equivalent required by the backend activities of a National Banking Group

Relatore

Prof. Domenico Ursino

Candidato

José Junior Paricagua Siñani

Anno Accademico 2020-2021

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Le analisi temporali hanno rappresentato, e continuano a rappresentare, uno degli aspetti principali che vengono considerati in qualsiasi campagna di Data Analytics. Esse, infatti, giocano un ruolo fondamentale in qualsiasi aspetto della Data Analytics, da quello descrittivo a quello diagnostico, da quello predittivo a quello prescrittivo. Per tale ragione, nel tempo, sono stati definiti degli approcci ben collaudati per effettuare questo tipo di attività. Tra questi approcci un ruolo chiave viene certamente giocato dalle serie temporali. In questo settore, infatti, oramai sono ben note tecniche per effettuare previsioni molto accurate, capaci di tenere conto delle interazioni tra più fenomeni e degli effetti della stagionalità. I ricercatori si sono anche sforzati di individuare una serie di parametri per calcolare la potenziale accuratezza delle previsioni e nel quantificare il possibile errore.

Sfruttando questo bagaglio culturale, nella presente tesi abbiamo condotto una campagna di Data Analytics per conto di un importante gruppo bancario. Per ragioni di riservatezza commerciale non possiamo indicare il nome del gruppo, per cui nel seguito, per indicarlo useremo lo pseudonimo di National Bank Group. Tale gruppo è in continua crescita e prevede, nel futuro, di acquisire varie filiali che attualmente non controlla. Ogni acquisizione comporta un incremento delle attività da compiere e un parallelo incremento delle risorse umane a disposizione. Si pone, pertanto, il problema di una allocazione ottimale delle risorse nelle varie attività e nelle varie filiali complessive del gruppo. Supportare i decision maker in tale attività di allocazione rappresenta l'obiettivo principale della presente tesi.

In particolare, ci si avvarrà di un opportuno parametri, denominato Full Time Equivalent (nel seguito, FTE). Questo parametro esprime il numero di risorse a tempo pieno per svolgere una determinata attività nonché il numero di risorse a tempo pieno presenti in un'azienda, in relazione al totale dei soggetti. A tal fine viene calcolato l'equivalente delle ore anche in presenza di part time e altre forme contrattuali con meno ore giornaliere rispetto al Full Time. Avendo a disposizione una quantificazione unica per le necessità e le risorse, è possibile procedere con l'utilizzo delle serie temporali sui dati passati in modo da effettuare previsioni per il futuro. Tali previsioni saranno di supporto per il management della banca nel determinare l'allocazione ottima delle risorse sui carichi ogniqualvolta verrà acquisita una nuova filiale.

La presente tesi illustrerà in dettaglio tutto il lavoro svolto e i risultati ottenuti.

Essa è strutturata come di seguito specificato:

La presente tesi verrà strutturata come di seguito specificato:

- Il Capitolo 1 illustrerà il contesto di riferimento relativo alla presente tesi, e quindi descriverà l'organizzazione e il modo di procedere del National Bank Group.
- Il Capitolo 2 presenterà i dati di partenza e descriverà tutte le attività di Extraction, Transformation and Loading (ETL) che abbiamo dovuto svolgere per rendere i dati in grado di essere utilizzati per l'estrazione di informazione e conoscenza affidabili.
- Il Capitolo 3 presenterà il processo di modellazione delle serie temporali e le metodologie utilizzate in tale contesto per stimare i corrispettivi parametri e per valutare l'accuratezza dei risultati ottenuti.
- Il Capitolo 4 illustrerà l'attività da noi condotta per scegliere il modello più adeguato e, successivamente, effettuare con il suo supporto le previsioni.
- Il Capitolo 5 presenterà l'attività di validazione dei risultati ottenuti dal modello e il loro utilizzo nell'ambito delle attività del National Bank Group.
- Infine, nel Capitolo 6, saranno tratte le conclusioni relative alla presente tesi e verranno presentati alcuni possibili sviluppi futuri.

1

Ambito di riferimento

In questo primo capitolo verranno descritte, a livello generale le entità presenti nel progetto, le loro attività nel settore bancario/finanziario, prodotti e servizi offerti, nonchè le modalità con cui esse interagiscono nel processo di una transazione bancaria. Si parlerà infine dell'obbiettivo generale con una breve sintesi delle linee guida del presente progetto.

1.1 Gruppo Bancario di riferimento: National Bank Group

La National Bank Group (N.B.C. nome di fantasia dietro cui si cela il nome effettivo del gruppo bancario, che non può essere citato per raggioni di riservatezza) è uno dei principali gruppi bancari, impegnato a sostenere l'economia nel Paese in cui opera, attraverso un approccio al business sostenibile e responsabile.

Si propone come partner globale al servizio di clienti Corporate, Public Finance e Financial Institutions, su basi nazionali e internazionali, distinguendoli per una storica presenza sui mercati finanziari e un'offerta estesa e innovativa.

Essa opera a fianco dei clienti con un approccio dedicato, una consolidata esperienza e una presenza internazionale, per offrire loro la migliore consulenza nella gestione dei rischi e per accompagnarli nel loro sviluppo, attraverso il cambiamento, individuando le opportunità offerte dal mercato.

I prodotti i servizi finanziari mirati all'eccellenza vengono offerti, in Italia, attraverso un Network costituito da Sedi Corporate coordinate da Aree Territoriali e da strutture di relazione dedicate alle Financial Institutions. A livello internazionale N.B.C. offre un supporto all'attività cross-border della clientela sia italiana che internazionale, con una rete estera specializzata costituita da Hub, Filiali, Uffici di Rappresentanza e controllate che svolgono attività di corporate e investment banking.

N.B.C crea in maniera durevole valore per i propri clienti, e per l'intero Gruppo, attraverso l'offerta di prodotti e servizi di:

- *Commercial Banking*;
- *Transaction Banking*;

- *Finanza Strutturata*;
- *Investment Banking*;
- *Capital Markets*.

Commercial Banking è nata grazie alla conoscenza approfondita delle necessità dei clienti e alla capacità di sviluppare soluzioni specifiche per il settore di appartenenza.

Nel Transaction Banking si offre alla clientela di riferimento un'ampia gamma di servizi transazionali, come il cash management, la trade & export finance, e securities services, tramite Inbiz, una piattaforma di Corporate Internet Banking.

N.B.C mostra una leadership storica nel mercato della Finanza Strutturata, unita a un solido track record internazionale e alla capacità di portare a termine operazioni complesse in numerosi ambiti come ad esempio il Project Finance.

N.B.C rappresenta un punto di riferimento nell'Investment Banking per l'attività di M&A Advisory, vantando inoltre una forte presenza nell'ambito dei collocamenti azionari e obbligazionari, con clientela italiana e internazionale.

Il gruppo si posiziona come riferimento nell'attività d'intermediazione sui Capital Markets, fornendo consulenza specialistica nella gestione dei rischi finanziari, operando sui mercati equity e fixed income, e ricoprendo il ruolo di market maker su una vasta gamma di strumenti finanziari: Titoli di Stato, corporate e financial bonds, cambi e derivati.

Infine, NBC emette strumenti finanziari (ad esempio obbligazioni e certificati) per investitori e risparmiatori, e fornisce ai clienti il servizio di best execution dinamica degli ordini MiFID compliant attraverso la piattaforma proprietaria Market Hub.

1.2 Provider di Servizi di National Bank Group: Energia S.P.A

Energia S.p.A. è una società nel perimetro diretto del National Bank Group, che si occupa di tutte le attività di operazioni e di back office bancario e supporta la banca in tutti processi organizzativi e di change management legati alle scelte di esternalizzazione, con l'obiettivo di garantire la riduzione dei costi mediante un costante efficientamento processuale, operativo e tecnologico.

Aderendo a Energia le Banche uniformano i processi di lavoro, la percezione del rischio, i profili operativi e il piano dei conti, riuscendo pertanto a:

- realizzare economie di utilizzo, riducendo i rischi connessi alla sostituzione di risorse e ai picchi di lavorazione;
- avere un maggior controllo di tutto il processo produttivo;
- disporre di un monitoraggio efficace sul controllo di gestione;
- adottare iter operativi certificati.

1.2.1 Modello Operativo

Il modello operativo con cui Energia S.p.A. si interfaccia è totalmente integrato: la banca che decide l'esternalizzazione di una attività/processo, di fatto, demanda ad Energia tutto il processo.

La filiale raccoglie la documentazione inoltrata dal cliente e la invia a Energia S.p.A., che contabilizza, gestisce la rete interbancaria in entrata ed in uscita, gestisce tutte le anomalie, quadra i conti di contabilità ed i conti reciproci su N.B.C, archivia in ottico e fisicamente la documentazione cartacea, gestisce i vari caveau dei valori, provvede alla postalizzazione se richiesta dalla banca, etc.

La Figura 1.1 mostra, a livello astratto, le relazioni e gli interscambi che avvengono tra N.B.C ed Energia S.p.A.

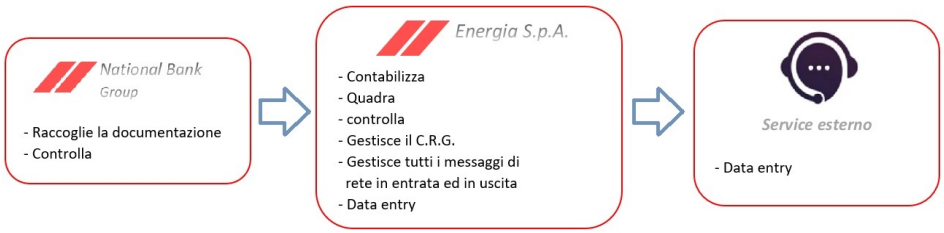

Figura 1.1. Relazioni e interscambi che avvengono tra N.B.c. ed Energia S.p.A.

In questo modello operativo, Energia esternalizza tutta l'operatività di data entry, di archiviazione e di imbustamento/spedizione a dei service esterni, anche se, nei confronti della banca l'unico, responsabile contrattuale rimane Energia.

Il modello di business è basato sulla massima flessibilità lasciata alle banche aderenti, sia per quanto riguarda i processi esternalizzati che per quanto attiene le modalità di gestione del personale in esubero.

Le banche possono aderire a tutti i processi o anche ad uno soltanto, secondo le proprie valutazioni strategico/economiche, così come hanno la possibilità - ma non l'obbligo - di distaccare proprio personale al consorzio; questo elemento è distintivo rispetto a molte altre esperienze analoghe, in cui il distacco del personale non è previsto.

In questo modo, è facoltà della banca cedente decidere se utilizzare la leva in una logica di puro re-engineering, riconvertendo le risorse rese disponibili su altre aree, o sfruttare l'occasione per ridurre in modo netto i propri costi operativi.

1.3 Obbiettivo del Progetto

Nel presente progetto l'obbiettivo principale è lo studio delle dipendenze che intercorrono tra le risorse presenti nell'organico di Energia S.p.A. e tutte le variabili inerenti le loro attività finalizzate ad effettuare i diversi servizi che l'azienda offre alle banche facenti parte della N.B.C.

Un primo passo in questa direzione è stato effettuato realizzando il cruscotto BLA (*Business Line Analysis*) che, a fronte dei dati relativi alle ore di lavoro dei dipendenti, derivati dal sistema Zucchetti (*Software per la gestione delle risorse umane*), e di quelli relativi ai diversi servizi offerti, fornisse una prima macroscopica stima dei livelli ottimali di gestione del personale, attraverso un'analisi ed una navigazione dinamiche ed interattive, per poter produrre indicatori a supporto del Management volti ad anticipare carichi di lavoro futuri derivanti da nuove acquisizioni o cambiamenti.

Questo progetto permette uno studio descrittivo e diagnostico del fenomeno, consentendo la navigazione dei dati storici e recenti, nonché una serie di indagini successive. Esso, però, da solo, non riesce a consentire uno studio predittivo e prescrittivo.

É proprio in questo scenario che si contestualizza l'attività di ricerca e di sviluppo di una soluzione avanzata che, in fasi successive, possa fornire anche delle indicazioni sui possibili andamenti di tutti i fenomeni in gioco, consentendo al Management di Energia S.p.A. un intervento più preciso, mirato e tempestivo nel gestire/incentivare i potenziali nuovi flussi nelle banche presenti nel gruppo, e pure arrivo di nuove Banche e/o diversa modulazione dei servizi offerti.

Il primo passo di questo percorso sarà, dunque, quello di, a partire dai dati storici già a disposizione del BLA, uno o più modelli che, governando secondo i principi del Machine Learning le molte variabili in gioco, sappiano fornire informazioni previsionali sui carichi di lavoro.

Sintetizzando il progetto, l'obiettivo che qui ci si pone di centrare è un sistema o modello che:

- Riceva in ingresso i dati relativi alle variabili legate ai servizi offerti ed ai clienti serviti (ovvero, le banche); tali variabili verranno individuate congiuntamente dal team di sviluppo e dagli esperti di Energia S.p.A..
- Apprenda comportamenti inerenti all'ambiente a partire dai dati storici.
- Suggerisca, per ogni possibile macro-scenario (ovvero, tipologia di banca che viene acquisito) quali siano le variabili più importanti capaci di condizionare i carichi di lavoro.
- Proponga un peso per ciascuna di queste variabili, consentendo comunque all'esperto umano di poter modificare i pesi e/o le variabili selezionate dal sistema stesso.

2

Analisi dei Dati

Nel capitolo corrente si fornirà una panoramica dei dati a disposizione per realizzare la rispettiva analisi del comportamento di ognuna delle variabili in gioco e per individuare le relazioni possibili tra di loro ci aiuterà a capire come verrà fatto lo sviluppo del modello in base alle variabili scelte. Successivamente ci si focalizzerà nella standarizzazione e analisi delle variabili nel tempo; i corrispondenti risultati saranno usati come futuro input del modello o dei modelli scelti.

2.1 Base di dati di partenza

La base di dati preliminare, dove verrà fatta tutta l'analisi, tanto delle variabili presenti come delle attività di ETL, si riferisce ad una serie di tabelle preliminari con cui procedere con i rispettivi lavori. Si tratta di una base di dati con informazione relativa ai processi/movimenti distribuiti nel tempo, tipologia di servizi, aree, settori, attività, banche, localizzazione, etc. La base di dati è composta da informazioni che vanno da gennaio 2014 a dicembre 2020, disaggregate a livello mensile e giornaliero.

2.1.1 Dizionario dei Dati

Nella presente sezione verrà proposta una descrizione delle variabili presenti dentro il dataset di partenza, attraverso lo strumento di un dizionario dei dati, fornendo, in questo modo, un dettaglio dei campi presenti nella rispettiva tabella/entità di dati. Come tabelle di partenza considereremo le tabelle delle anagrafiche(tabelle 2.1, 2.2 a 2.3) presenti nella base di dati fornita.

Altre tabelle che prenderemo in considerazione riguardano informazioni delle operazioni/effettuate presso la National Bank Group. La loro struttura viene mostrata nelle tabelle 2.4, 2.5 e 2.6.

Come ultima tabella, consideriamo la tabella Ore personale, dove sono presenti informazioni come il campo QtaOre (Quantità d'ore lavorate), che risulteranno estremamente utiliti per i nostri obiettivi.

CAMPI	TIPO DI DATO
CodIstituto	Intero
CodABI	Intero
DesBanca	Stringa
DataOrigineBanca	Date
NumDipendentiAttuali	Intero
IdProvincia	Stringa
IdRegione	Intero
IdIstatProvincia	Intero
DataRecesso	Date

Tabella 2.1. Tabella Anagrafica Banche

CAMPI	TIPO DI DATO
CodIstituto	Intero
CodFiliale	Intero
TipoFiliale	Stringa
DesFiliale	Stringa
ViaFiliale	Stringa
LocalitaFiliale	Stringa
ProvinciaFiliale	Stringa

Tabella 2.2. Tabella Anagrafica Filiali

CAMPI	TIPO DI DATO
idArea	Stringa
idSettore	Stringa
idScheda	Stringa
idAttivita	Stringa

Tabella 2.3. Tabella Anagrafica Schede Tecniche

CAMPI	TIPO DI DATO
CodIstituto	Intero
IdArea	Stringa
IdSettore	Stringa
IdAttivita	Stringa
IdScheda	Stringa
Anno	Stringa
Mese	Intero
NumeroComunicazioni	Intero
TempoMedioRisposta	Decimal
TipoComunicazione	Stringa

Tabella 2.4. Tabella Comunicazioni Tempi Medi

CAMPI	TIPO DI DATO
CodIstituto	Intero
IdArea	Stringa
IdSettore	Stringa
IdAttivita	Stringa
IdScheda	Stringa
Anno	Intero
mMse	Intero
NumeroPending	Intero

Tabella 2.5. Tabella Pending

CAMPI	TIPO DI DATO
IdArea	Stringa
IdSettore	Stringa
IdAttivita	Stringa
IdScheda	Stringa
CodIstituto	Intero
Anno	Intero
Mese	Intero
NumeroOperazioni	Intero
ImportoFattura	Decimal

Tabella 2.6. Tabella Volumi

CAMPI	TIPO DI DATO
keyZucchetti	Stringa
anno	Intero
mese	Intero
idDipendente	Intero
idArea	Stringa
idSettore	Stringa
idAttivita	Stringa
idScheda	Stringa
tipoOre	Stringa
contaGiorniLav	Intero
qtaOre	Decimal
FTEs	Decimal

Tabella 2.7. Tabella Ore Personale

2.2 Attività di ETL condotte

In questa sezione, dopo una breve descrizione del processo ETL, verranno analizzate nel dettaglio le singole fasi.

2.2.1 Processo di ETL

Il processo di estrazione e compilazione dei dati grezzi, la loro trasformazione per renderli comprensibili e il loro caricamento in un sistema di destinazione, come un database o un Data Warehouse, per un facile accesso e analisi posteriore, è noto come processo di Extract - Transform - Load (ETL). Essa risulta una procedura fondamentale per la gestione dell'ecosistema dei dati (figura 2.1).

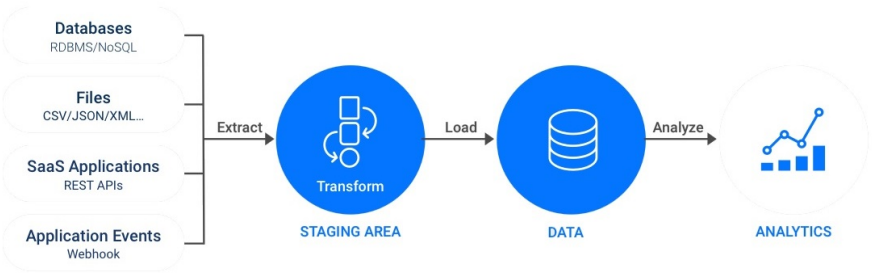

Figura 2.1. Processo di ETL

Poiché i dati provenienti da più fonti hanno un modello diverso, ogni set di dati deve essere trasformato in modo diverso prima di poter essere utilizzato per la Business Intelligence e l'analisi.

Per il processo di estrazione si farà utilizo di una macchina virtuale con la presenza di un ambiente di lavoro costituito dal tool Jupyter, basato sul linguaggio di programmazione Python.

2.2.2 Strumenti di estrazione

L'estrazione dei dati sarà realizzata con aiuto degli strumenti web come Jupyter Notebook.

Jupyter Notebook è un'applicazione web, open source che permette di creare e condividere documenti con codice dal vivo, equazioni, visualizzazioni e testo esplicativo.

Insieme a questo strumento, nel presente progetto verrà utilizzato il linguaggio Python, adottato in diversi settori, Python è diventato il linguaggio più utilizzato per la programmazione scientifica.

In particolare, faremo uso delle seguenti librerie di Python:

- *Numpy*: per elaborazione di dati con matrici.
- *Matplotlib*: per la visualizzazione.

- *Pandas*: per l'analisi statistica dei dati.

Un'esempio del processo di estrazione dei dati con i tool menzionati precedentemente, viene mostrato nella figura 2.2.

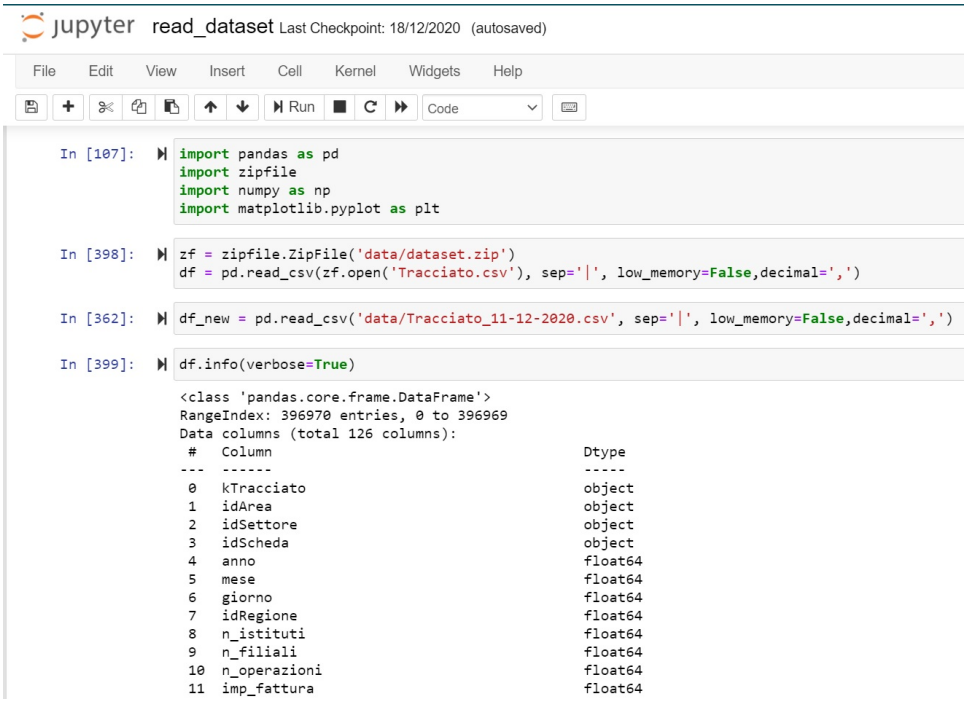

Figura 2.2. Estratto del processo di estrazione dei dati con Jupyter e Python

Attraverso questi strumenti, si procede nel processo di analisi e revisione dei dati relativi a National Bank Group.

Ad esempio, partendo dei dati presenti da due sorgenti diverse, si realizzeranno l'individuazione e pulizia dei dati attraverso la comparazione delle due sorgenti, operando in tal modo si osserva una differenza costante nei dati che, probabilmente, è il risultato di un'estrazione errata dei dati. Nelle figure 2.3 e 2.4 viene mostrato un riepilogo delle differenze trovate tra le due sorgenti.

In queste due figure, non si osservano differenze nei dati. Procediamo, quindi, con il confronto delle due sorgenti per ciò che concerne gli altri attributi. Tale confronto viene mostrato nelle figure 2.5, 2.6 e 2.7.

In queste figure, esiste una differenza notevole nei dati, per cui si procede alla scelta della seconda sorgente di dati per il presente gruppo di attributi. Continuiamo allo stesso modo con i confronti mostrati nelle figure 2.8, 2.9 e 2.10.

Dall'analisi di queste figure emerge che esiste una differenza nel comportamento temporale tra le due sorgenti.

Anche in questo caso viene scelta la seconda sorgente. In quest'ultima sorgente, infatti, il comportamento dei dati appare più realistico.

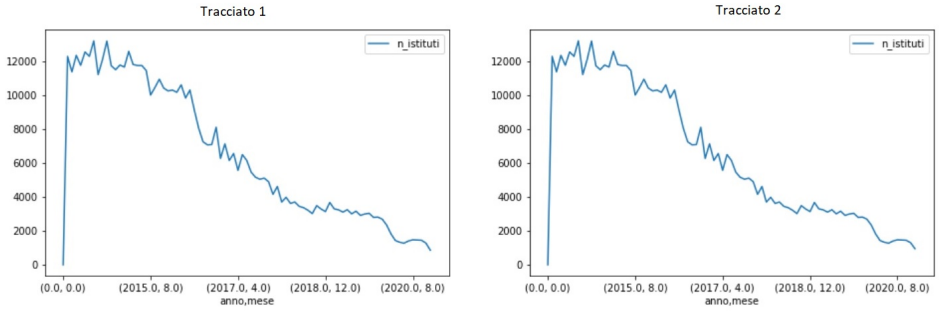

Figura 2.3. Confronto dei dati dell'attributo N_istituti

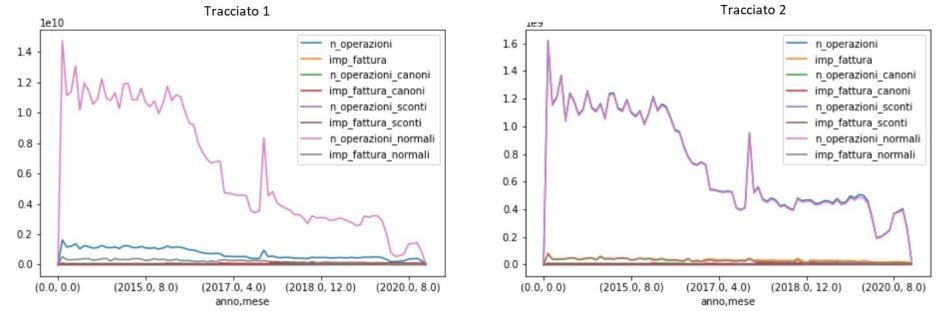

Figura 2.4. Confronto dei dati degli attributi Nro di operazioni e Importo Fatture

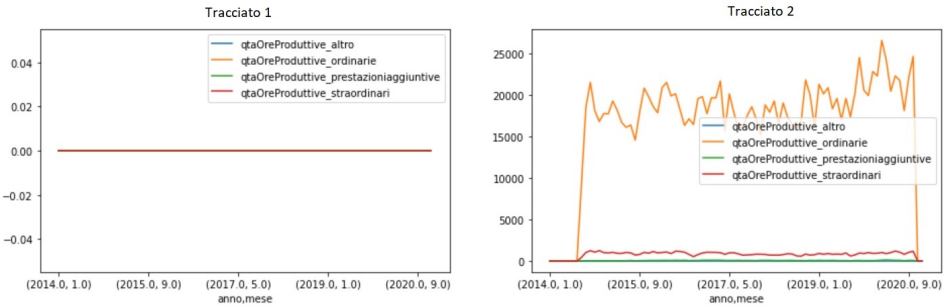

Figura 2.5. Confronto dei dati del gruppo di attributi Qta. Ore Produttive ordinarie e straodinari

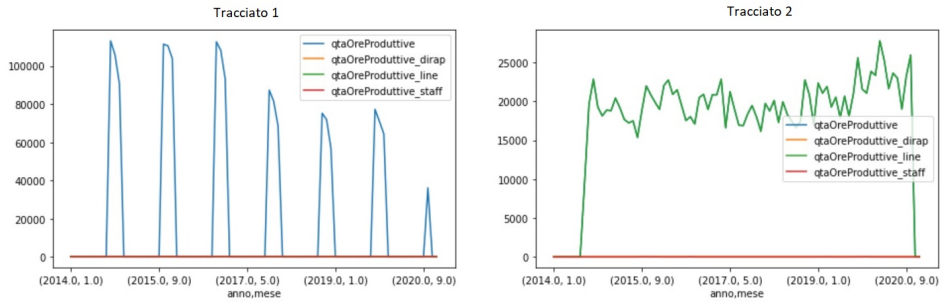

Figura 2.6. Confronto dei dati degli attributi Qta. Ore Produttive dirap, line e staff

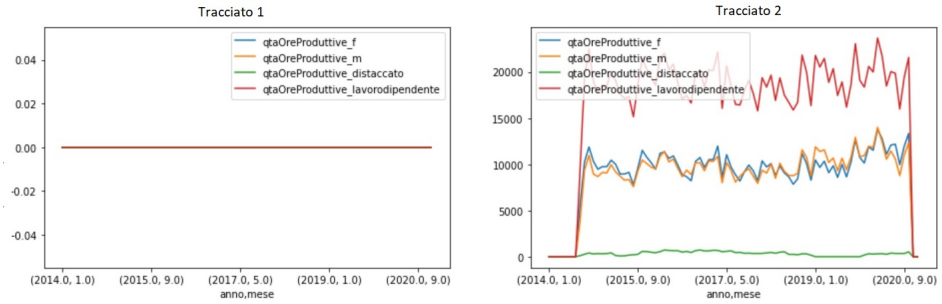

Figura 2.7. Confronto dei dati degli attributi Qta. Ore Produttive per genere

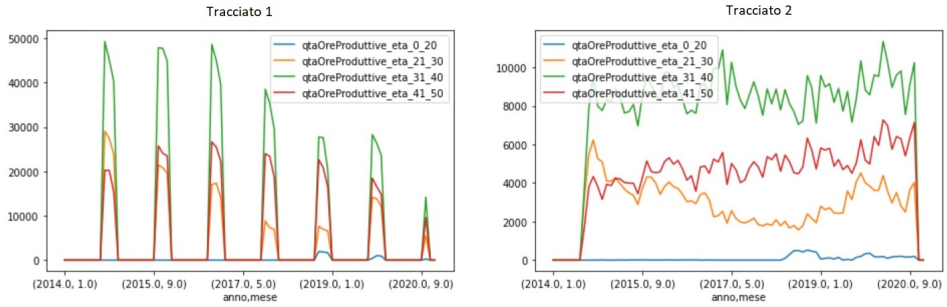

Figura 2.8. Confronto dei dati degli attributi Qta. Ore Produttive per fasce di età (gruppo 1)

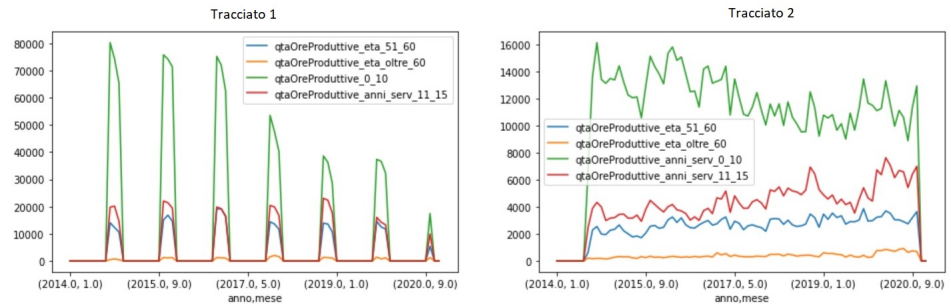

Figura 2.9. Confronto dei dati degli attributi Qta. Ore Produttive per fasce di età (gruppo 2)

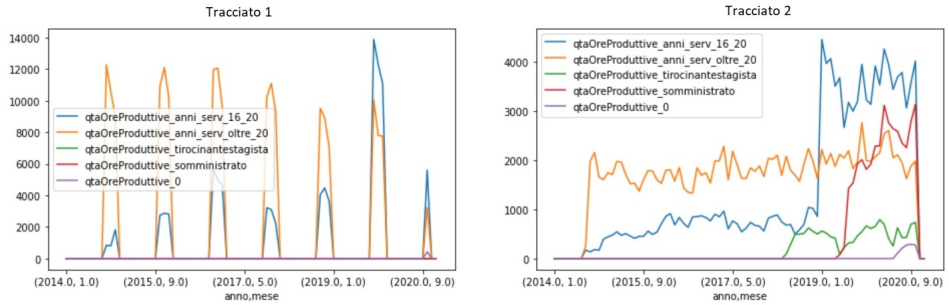

Figura 2.10. Confronto dei dati degli attributi Qta. Ore Produttive per fasce di età (gruppo 3)

2.3 Schema Finale dei Dati

Succesivamente ai lavori realizzati nella sezione precedente, si ottiene uno schema della base di dati con le rispettive relazione tra le tabelle. A partire da queste si può ricostruire uno schema E-R, come mostrato in figura 2.11.

Nel seguito, farremo particolare enfasi nella tabella Ore personale e gli attributi come “qtaOre” (quantità di ore lavorate) o l’attributo diretto FTE.

A partire dallo schema mostrato nella figura 2.11, insieme al National Bank Group, si approfondisce l’acquisizione di nuovi campi relazzionati a qtaOre, come quelli visti dal confronto delle sorgenti. Al termine di questa attività si ottiene come risultato l’ultima sorgente, vista come una estensione della tabella Ore personale, come si vede di seguito.

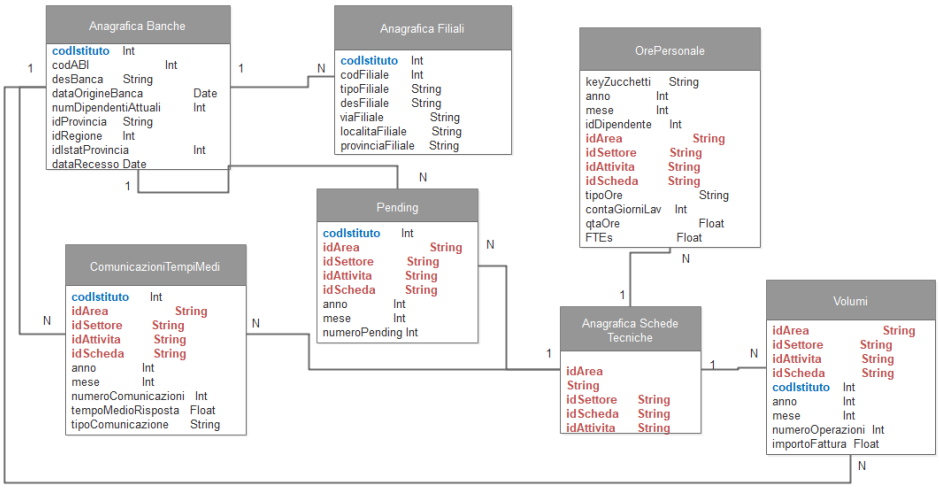

Figura 2.11. Schema finale della base di dati

CAMPI	TIPO DI DATO
kTracciato	Stringa
idArea	Stringa
idSettore	Stringa
idScheda	Stringa
anno	Intero
mese	Intero
giorno	Intero
n_istituti	Intero
n_filiali	Intero
n_operazioni	Intero
imp_fattura	Intero
qtaOreProduttive	Decimale
qtaOreProduttive_dirap	Decimale
qtaOreProduttive_line	Decimale
qtaOreProduttive_staff	Decimale
qtaOreProduttive_altro	Decimale
qtaOreProduttive_ordinarie	Decimale
qtaOreProduttive_prestazioniaggiuntive	Decimale
qtaOreProduttive_straordinari	Decimale
qtaOreProduttive_f	Decimale
qtaOreProduttive_m	Decimale
qtaOreProduttive_distaccato	Decimale
qtaOreProduttive_lavorodipendente	Decimale
qtaOreProduttive_eta_0_20	Decimale
qtaOreProduttive_eta_21_30	Decimale
qtaOreProduttive_eta_31_40	Decimale
qtaOreProduttive_eta_41_50	Decimale
qtaOreProduttive_eta_51_60	Decimale
qtaOreProduttive_eta_oltre_60	Decimale
qtaOreProduttive_anni_serv_0_10	Decimale
qtaOreProduttive_anni_serv_11_15	Decimale
qtaOreProduttive_anni_serv_16_20	Decimale
qtaOreProduttive_anni_serv_oltre_20	Decimale
qtaOreProduttive_tirocinantestagista	Decimale
qtaOreProduttive_somministrato	Decimale
qtaOreProduttive_0	Decimale

Tabella 2.8. Tabella Ore Personale Estesa

In base all'ultima estensione della tabella Ore Personale della National Group Bank, l'analisi succesiva sarà svolta prendendo gruppi di campi relazionati, come, ad esempio:

Per fasce di età:

- QtaOreProduttive_eta_0_20: quantità di ore produttive di dipendenti con età minore di 20 anni.
- QtaOreProduttive_eta_21_30: quantità di ore produttive di dipendenti con età tra 21 e 30 anni.
- QtaOreProduttive_eta_31_40: quantità di ore produttive di dipendenti con età tra 31 e 40 anni.
- QtaOreProduttive_eta_41_50: quantità di ore produttive di dipendenti con età tra 41 e 50 anni.
- QtaOreProduttive_eta_51_60: quantità di ore produttive di dipendenti con età tra 51 e 60 anni.
- QtaOreProduttive_eta_oltre_60: quantità di ore produttive di dipendenti con età maggiore di 60 anni.

Per genero:

- QtaOreProduttive_f: quantità di ore produttive di dipendenti di sesso femminile.
- QtaOreProduttive_m: quantità di ore produttive di dipendenti di sesso maschile.

Per anni di servizio:

- QtaOreProduttive_anni_serv_0_10: quantità di ore produttive di dipendenti con un numero di anni di servizo compreso tra 0 e 10.
- QtaOreProduttive_anni_serv_11_15: quantità di ore produttive di dipendenti con un numero di anni di servizo compreso tra 11 e 15.
- QtaOreProduttive_anni_serv_16_20: quantità di ore produttive di dipendenti con un numero di anni di servizo compreso tra 16 e 20.
- QtaOreProduttive_anni_serv_oltre_20: quantità di ore produttive di dipendenti con oltre 20 anni di servizio.

Per tipo dello svolgimento del lavoro:

- QtaOreProduttive_anni_serv_0_10: quantità di ore produttive di lavoro distaccato.
- QtaOreProduttive_anni_serv_oltre_20: quantità di ore produttive di lavoro dipendente.

Dopo l'individuazione dei gruppi, è stato realizzato un elenco dell'attributo "Scheda Tecnica" presente nella tabella finale "Ore Personale", tale tabella è quella di riferimento per il servizio svolto in quella transazione. L'elenco che ha un totale di 96 Schede Tecniche presenti in un'intervallo temporale compreso tra gennaio 2014 e dicembre 2020. Esso è mostrato nella figura 2.12.

2.4 Analisi di Serie Temporali

In questa sezione, si procede con i lavori di analisi delle serie temporali estratte dal schema finale definito nella sezione precedente.

```
In [465]: len(df_new['idScheda'].unique())

Out[465]: 96

In [466]: df_new['idScheda'].unique()

Out[466]: array(['SASSCC', 'SASSSI', 'SBECBI', 'SCOBAM', 'SCOBIL', 'SCOCRG',
                 'SCOMAN', 'SESGAR', 'SESASS', 'SIPMPP', 'SIPCPD', 'SIPIMM',
                 'SUMARK', 'SGTATM', 'SCAIAS', 'SASSO', 'SASSCI', 'SBONIF',
                 'SCARTE', 'SRID', 'SPTFDI', 'SPENSI', 'SGEPOS', 'SPTEFF', 'SACCAR',
                 'SACCBA', 'SACCRE', 'SARCFI', 'SCOCSR', 'SCOTRA', 'SDOCOT',
                 'SFORMA', 'SMONET', 'SPOSTA', 'SPOSMA', 'SSISBT', 'SSISIN',
                 'SWEBEX', 'SSTIPR', 'SSTIBP', 'SSTIHR', 'SPCO', 'SSVPCF', 'SSVPSE',
                 'STEINC', 'STESOR', 'STEWEB', 'STIANA', 'STITES', 'STIASI',
                 'STITBA', 'STITCO', 'SCOFOR', 'SESTES', 'SASSE', 'SDELFI',
                 'SBOLUT', nan, 'SASSAC', 'SSIPRE', 'NOATT', 'SSWCFD', 'SSISVI',
                 'SCREBO', 'SCOABB', 'SSISVB', 'SLOGAS', 'SSWEXC', 'SCREBI',
                 'SCREIN', 'SIPPEC', 'SCREIM', 'SANATO', 'SCOMAS', 'SANPEF',
                 'SREPSI', 'SCRECM', 'SBOTIF', 'SCABA', 'SCRECR', 'SCRECI',
                 'SOUTBD', 'SSUPCE', 'SISIRA', 'SMDPRO', 'SCRERE', 'SGAMMA',
                 'SCOFIS', 'SDICHI', 'SARCPR', 'SGTPOS', 'SCREPD', 'SGTCAR',
                 'SCRELI', 'SCRESO', 'SIPIPD'], dtype=object)
```

Figura 2.12. Elenco Schede Tecniche

2.4.1 Stazionarietà

Nell'analisi delle serie storiche è possibile che si manifestino dei trend che potrebbero rendere le regressioni spurie.

Questi trend possono essere stocastici, nel caso ci sia non stazionarietà in varianza, o deterministici, nel caso la non stazionarietà sia in media.

Il test di Dickey-Fuller permette di valutare se esiste un trend nelle variabili che renda la regressione spuria. Nel caso sussista tale trend è possibile creare la differenza tra le variabili al tempo t e quelle $t-1$ e lavorare su queste.

Partendo dell'ultimo argomento, il test procede nel seguente modo.

Partiamo da un modello autorregressivo di primo ordine, AR(1):

$$Y_t = a + \varphi Y_{t-1} + \varepsilon \tag{2.1}$$

Sottraiamo il termine Y_{t-1} in entrambi membri.

$$Y_t - Y_{t-1} = a - Y_{t-1} + \varphi Y_{t-1} - Y_{t-1} + \varepsilon - Y_{t-1} \tag{2.2}$$

Mettendo in evidenza il fattore comune avremo come risultato:

$$\varphi Y_{t-1} - Y_{t-1} = Y_{t-1}(\varphi - 1) = Y_{t-1}(\delta - 1) \tag{2.3}$$

$$Y_t - Y_{t-1} = \Delta Y \tag{2.4}$$

Con quest'ultima considerazione, riformuliamo l'equazione nel seguente modo:

$$\Delta Y_t = a + \Delta Y_{t-1} + \varepsilon \tag{2.5}$$

Partendo da questa equazione si possono verificare due ipotesi:

- *H0*: $\Delta = 0$: Presenza di trend stocastico nella serie temporale.

- *H1*: $\Delta < 0$: Assenza di trend stocastico nella serie temporale.

In base al risultato di tale test si procede all'analisi delle varibili presenti nella sorgente/tabella Ore Personale Estesa, facendo la seguente stima dei parametri presenti nella equazione.

Dato il modello:

$$\Delta Y_t = \Delta Y_{t-1} + \varepsilon \tag{2.6}$$

facendo l'ugualianza con il modello di una serie temporale:

$$Y_t = \rho x + e_t \tag{2.7}$$

con $e_t = Error$:

$$e_i = Y_t - \hat{Y}_t \tag{2.8}$$

Allora:

$$E_T = \sum (Y_t - \hat{Y}_t)^2 \tag{2.9}$$

con $\hat{Y}_t = \rho x$, quindi:

$$E_T = \sum (Y_t - \rho x)^2 \tag{2.10}$$

Procedendo con la derivata:

$$\frac{\partial E_T}{\partial \rho} = \sum (Y_t - \rho x)^2 \tag{2.11}$$

$$\frac{\partial E_T}{\partial \rho} = (-2) \sum (Y_t - \rho x) x \tag{2.12}$$

Infine. la stima del parametro del modello viene effettuata ottenendo la seguente formula:

$$\rho = \frac{\sum (Y_t x)}{\sum x_t^2} \tag{2.13}$$

In alternativa, si può procedere con il calcolo utilizzando il metodo matriciale. Se Y e X rappresentano delle matrici, il modello può essere definito nel seguente modo:

$$Y = \alpha X \tag{2.14}$$

Facendo la molteplicazione per X^T. e $(X^T X)^{-1}$in entrambi membri, avremo:

$$(X^T X)^{-1} X^T Y = X^T X (X^T X)^{-1} \alpha \tag{2.15}$$

A questo punto, il risultato finale della stima sarà:

$$\alpha = (X^T X)^{-1} X^T Y \tag{2.16}$$

Il valore del parametro α ci consentirà di verificare la stagionalità o meno della serie temporale.

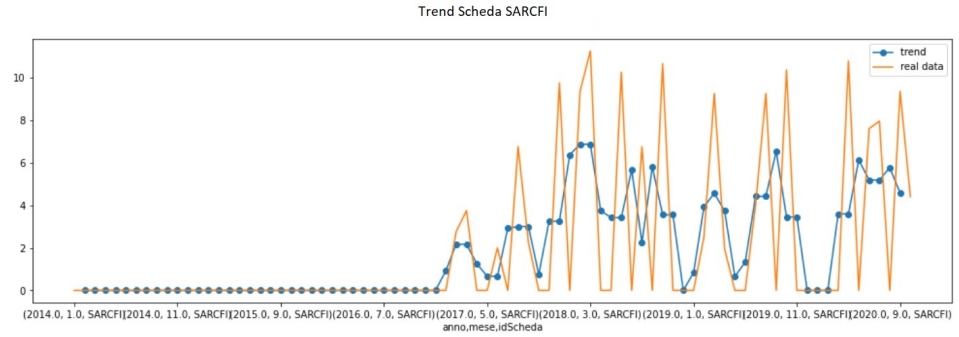

Figura 2.13. Trend Scheda Tecnica SARCFI

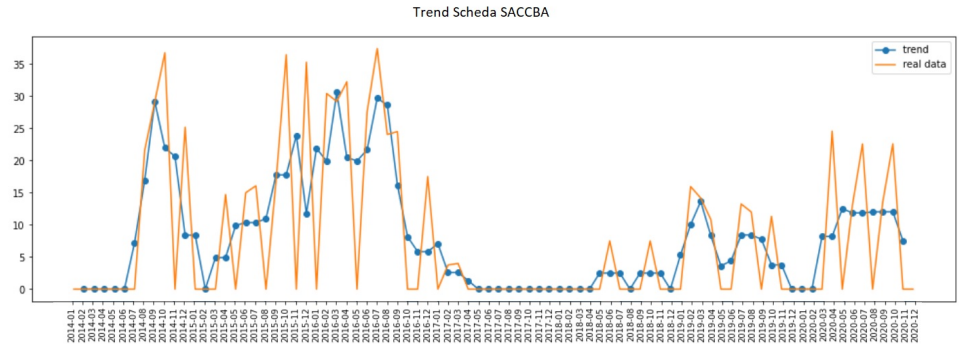

Figura 2.14. Trend Scheda Tecnica SACCBA

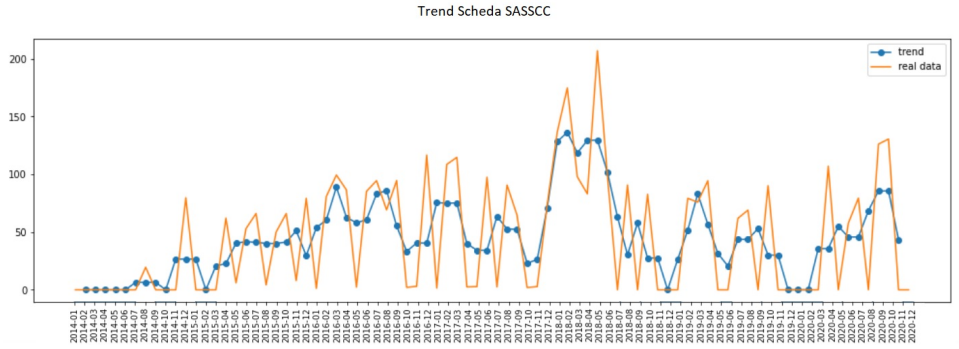

Figura 2.15. Trend Scheda Tecnica SASSCC

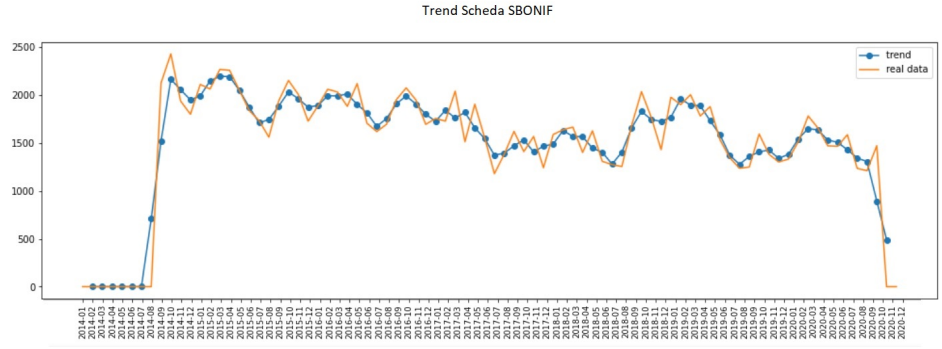

Figura 2.16. Trend Scheda Tecnica SBONIF

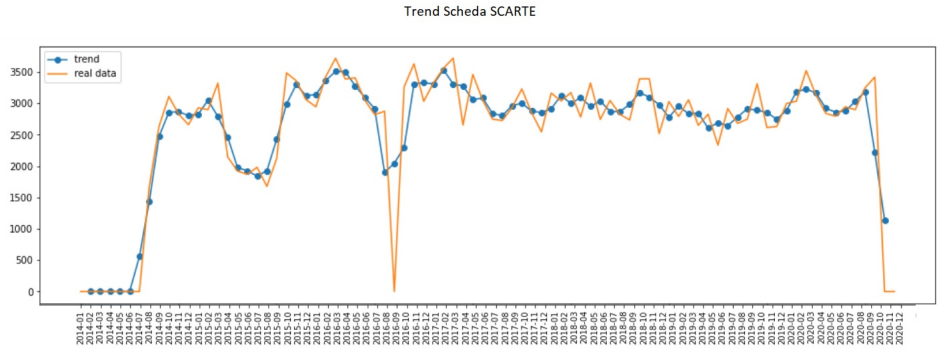

Figura 2.17. Trend Scheda Tecnica SCARTE

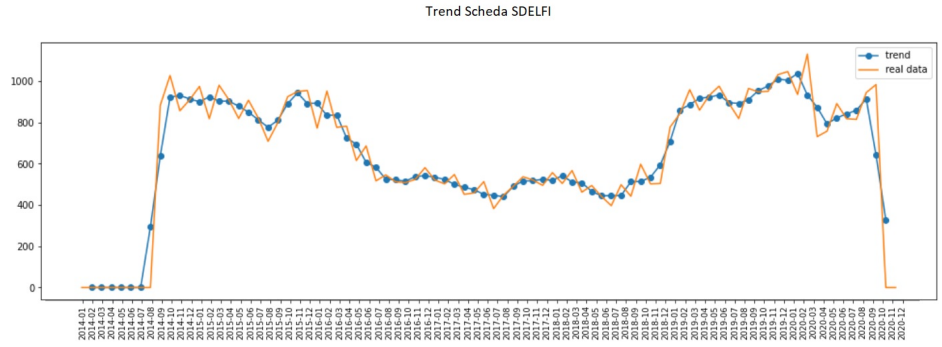

Figura 2.18. Trend Scheda Tecnica SDELFI

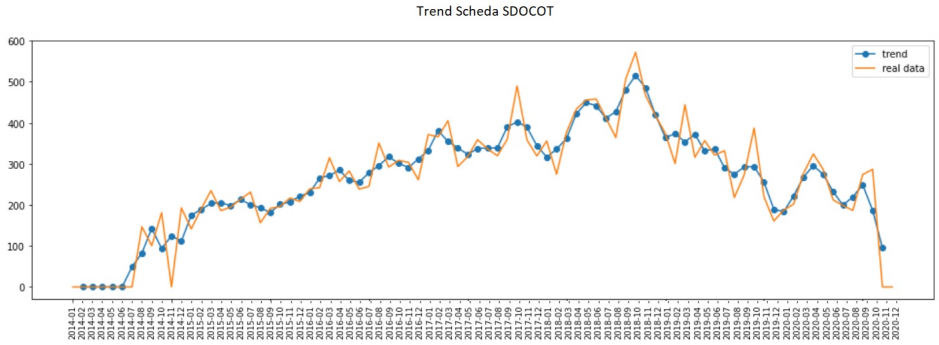

Figura 2.19. Trend Scheda Tecnica SDOCOT

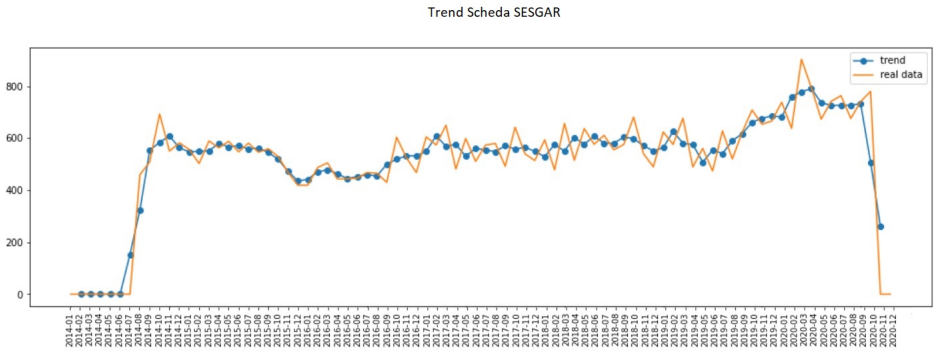

Figura 2.20. Trend Scheda Tecnica SESGAR

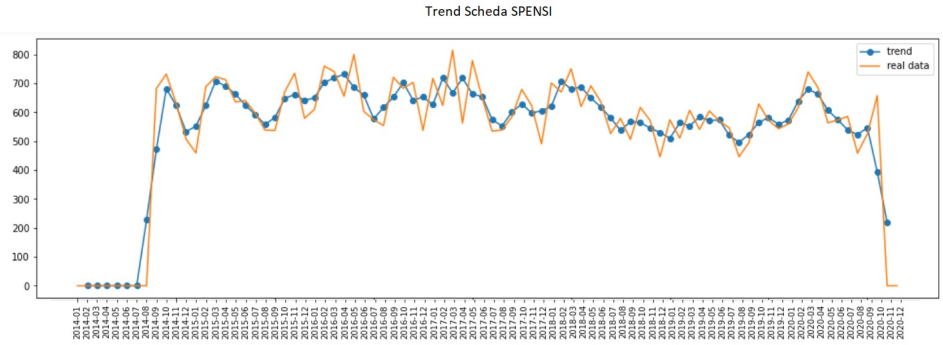

Figura 2.21. Trend Scheda Tecnica SPENSI

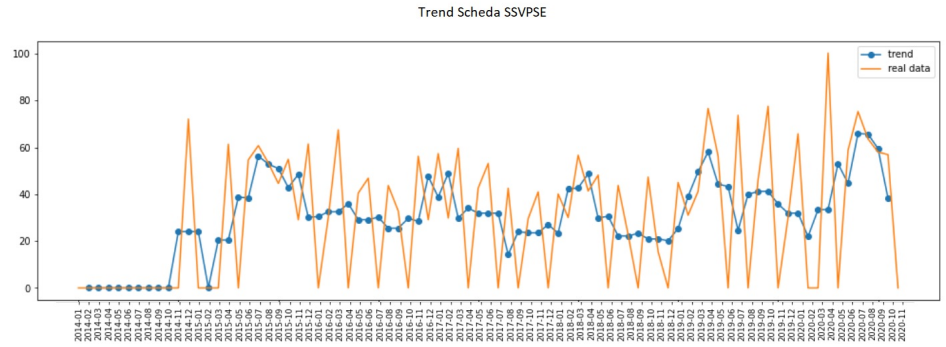

Figura 2.22. Trend Scheda Tecnica SSVPSE

Di seguito si procede con l'applicazione della verifica di Dickey Fuller a 10 schede tecniche riprese dal totale individuato di 96; per la verifica utilizzeremo l'attributo Qta.Ore produttive in ciascuna scheda.

Come risultato dell'analisi del trend del test di Dickey Fuller (radice unitaria) effettuata precedentemente, otteniamo i risultati riportati nella tabella 2.9.

SCHEDA TECNICA	**DICKEY FULLER**
SARCFI	stazionaria
SACCBA	non stazionaria
SASSCC	stazionaria
SBONIF	stazionaria
SCARTE	stazionaria
SDELFI	non stazionaria
SDOCOT	stazionaria
SESGAR	non stazionaria
SPENSI	stazionaria

Tabella 2.9. Risultati Test Dickey Fuller

3

Modelli di Predizione

Data la descrizione delle serie temporali ai fini di una migliore comprensione, nel presente capotolo si procede con il processo di modellazione, che collabora con le serie temporali; ciò servirà a costruire i modelli di predizione che si intendono formulare per quanto riguarda l'analisi dell' FTE e la sua composizione.

3.1 Modellazione

Per capire la modellazione bisogna prima avere chiari alcuni concetti che sono legati a questo tema. Essi saranno l'oggetto della presente sezione.

3.1.1 Definizione di Modello

Si chiama modello una rappresentazione astratta, concettuale, grafica o visiva (ad esempio, una mappa concettuale), fisica, di fenomeni, sistemi o processi per analizzare, descrivere, spiegare, simulare (in generale, esplorare, controllare e prevedere) gli stessi. Un modello permette di determinare un risultato finale a partire da dati di input. La creazione di un modello è considerata una parte essenziale di ogni attività scientifica.

3.1.2 Metodologia di Modellazione

La metodologia di modellizzazione si riferisce all'insieme delle procedure sistematiche basate sulle conoscenze acquisite che mirano ad affrontare e risolvere i problemi.

La modellizzazione è il processo attraverso il quale si instaura un rapporto tra i principali enti di un sistema che si esprime in termini di obiettivi, criteri di attuazione e restrizioni; questi nel complesso, costituiscono il modello. Ogni modello ha un modello di base, che costituisce la visione o l'immagine particolare che ha sul sistema a partire da cui sarà costruito un modello semplificato. Attraverso la sperimentazione di questo modello semplificato, si spera di migliorare la comprensione del modello di base e del sistema reale da esso caratterizzato.

La modellazione è un processo iterativo. Inoltre, il passaggio dal modello di base al modello semplificato è generalmente accompagnato da una trasformazione delle informazioni qualitative in informazioni quantitative. Il processo di modellazione è di natura evolutiva, passando da speculazioni iniziali a ipotesi, a modelli generali e, infine, a un modello specifico semplificato La partecipazione del computer non è solo per la simulazione o predizione, ma anche in varie fasi del processo di modellazione, che includono la raccolta e il trattamento dei dati, la progettazione di esperimenti, confronti, verifiche e convalidazioni.

3.2 Modelli di Serie Temporali

Il modello delle serie temporali tiene conto dello schema dei movimenti pregressi di una determinata variabile e utilizza queste informazioni per prevedere i suoi movimenti futuri. È solo un metodo di estrapolazione sofisticato. Può fornire uno strumento molto efficace per la previsione.

Un modello di serie temporale per la quantità di ore lavorate collegherebbe tale variabile ai suoi valori storici e ad altre variabili che descrivono la natura "aleatoria" del suo comportamento nel passato. Come la maggior parte dei modelli di regressione, essa consiste in un'equazione che contiene un insieme di coefficienti da stimare. Tuttavia, a differenza di questi ultimi, l'equazione è non lineare nei coefficienti, per cui per la stima è necessario utilizzare una versione non lineare degli MQO (Minimi Quadrati Ordinari).

3.2.1 Modello Autorregressivo AR(p)

Un modello AR auto-regressivo descrive una particolare classe di processi in cui le osservazioni in un momento dato sono prevedibili dalle osservazioni preliminari del processo, più un termine di errore. Il caso più semplice è ARIMA(1,0,0) o AR(1), o di primo ordine, la cui espressione matematica è:

$$AR(1) = x_t = \phi_1 x_{t-1} + a_t \tag{3.1}$$

Il processo di auto-regressione d'ordine p, rappresentato da ARIMA(p,0,0), o semplicemente da AR(p), viene formulato come:

$$AR(p) = x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + a_t \tag{3.2}$$

che, per il tramite dell'operatore di cambio retroattivo B, può assumere la forma:

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) x_t = a_t \tag{3.3}$$

$$B^k(x_t) = x_{t-k} \tag{3.4}$$

Un processo auto regressivo AR(p) è stazionario se le radici del polinomio in B dato da: $(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)$ cadono al di fuori del cerchio unità. Questa condizione è equivalente al fatto che le radici della formula: $(x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - ... - \phi_{p-1} x - \phi_p = 0)$ siano tutte inferiori a 1 in modulo. Un processo autoregressivo è sempre invertibile.

3.2.2 Modello di Media Mobile MA(q)

Un modello di medie mobili MA descrive una serie temporale stazionaria. In questo modello il valore attuale può essere previsto a partire dalla componente casuale di questo momento e, in misura minore, dagli impulsi casuali di cui sopra. Il modello ARIMA(0,0,1), anche indicato da MA(1), è dato dall'espressione:

$$x_t = a_t - v_1 a_{t-1} \tag{3.5}$$

Il processo di medie mobili d'ordine q, rappresentato da ARIMA(0,0,q), o anche da Ma(q), viene dato dall'espressione:

$$x_t = a_t - v_1 a_{t-1} - v_2 a_{t-2} - ... - v_q a_{t-q} \tag{3.6}$$

che, per il tramite dell'operatore di cambio retroattivo B, può assumere la forma:

$$x_t = (1 - v_1 B - v_2 B^2 - ... - v_q B^q) a_t \tag{3.7}$$

Un processo di medie mobili è sempre stazionario. Un processo di medie mobili MA(q) è invertibile se le radici del polinomio in B definito da $(1 - v_1 B - v_2 B^2 - ... - v_q B^q)$ cadono al di fuori del cerchio unità. Questa condizione è equivalente al fatto che le radici dell'equazione $(x^q - \phi_1 x^{q-1} - \phi_2 x^{q-2} - ... - \phi_{q-1} x - \phi_q = 0)$ siano tutte inferiori a 1 in modulo.

3.2.3 Modello ARIMA

Box e Jenkins hanno sviluppato modelli statistici per serie temporali che tengono conto della dipendenza esistente tra i dati, vale a dire che ogni osservazione in un momento dato è modellata in funzione dei valori di cui sopra. Le analisi sono basate su un modello specifico. I modelli sono conosciuti con il nome generico di ARIMA (Autoregresive Integrated Moving Average), che deriva dai suoi tre componenti, ovvero AR (Autoregressivo), I (Integrato) e MA (Media Mobile).

Il modello ARIMA permette di descrivere un valore come una funzione lineare di dati precedenti ed errori dovuti al caso; inoltre, può includere un componente ciclico o stagionale. In altre parole, esso contiene tutti gli elementi necessari per descrivere il fenomeno. Box e Jenkins consigliano almeno 50 osservazioni nelle serie temporali.

La metodologia di Box e Jenkins si articola in quattro fasi:

- *La prima fase consiste nell'identificare il possibile modello ARIMA che segue la serie, il che richiede* decidere che informazioni applicare per convertire la serie osservata in una serie stazionaria e determinare un modello ARMA per la serie stazionaria, cioè gli ordini di p e q della sua struttura autoregressiva e media mobile.
- Selezionando provvisoriamente un modello per la serie stazionaria, si passa alla *seconda fase di stima*, dove i parametri AR e MA del modello sono stimati per la massima plausibilità e si ottengono i loro errori standard e i residui del modello.

- *La terza fase è la diagnosi*, dove si verifica che i descarti non abbiano una struttura di dipendenza e seguono un processo di rumore bianco. Se i descarti mostrano una struttura, il modello viene modificato per incorporarlo e ripetere le fasi precedenti fino a ottenere un modello adeguato.
- *La quarta fase è la previsione*; una volta ottenuto un modello adeguato, si fanno previsioni con esso.

Identificare un modello significa utilizzare i dati raccolti e tutte le informazioni su come genera la serie temporale oggetto di studio, per suggerire una serie ristretta di possibili modelli che hanno molte possibilità di adattarsi ai dati. In presenza di una serie temporale empirica, è necessario trovare i valori (p, d, q) più appropriati.

- Se la serie temporale presenta una tendenza, la prima cosa da fare è trasformarla in stazionaria mediante una differenziazione di ordine d. Una volta differenziata la serie, una buona strategia consiste nel confrontare le correlazioni della funzione di autocorrelazione (ACF) e della funzione di autocorrelazione parziale (ACFP), un processo che, di solito, fornisce una guida per la formulazione del modello indicativo.
- I processi di auto-regressione hanno una funzione di autocorrelazione parziale (ACFP) con un numero finito di valori diverso da zero. Un processo AR(p) ha i primi p termini della funzione di autocorrelazione parziale diversi da zero e gli altri sono nulli. Questa affermazione è molto forte e in pratica si ritiene che un dato campione provenga da un processo di auto-regressione di ordine p se i termini della funzione di autocorrelazione parziale sono quasi zero a partire da quello che occupa il posto p. Un valore è considerato quasi zero quando il modulo è inferiore a $\frac{2}{\sqrt{T}}$. Esistono programmi per computer che costruiscono l'intervallo $(-\frac{2}{\sqrt{T}}, \frac{2}{\sqrt{T}})$ e rilevano i valori dell'ACFP che cadono al di fuori di esso.
- I processi di medie mobili hanno funzione di autocorrelazione con un numero finito di valori diversi da zero. Un processo MA(q) ha i primi q termini della funzione di autocorrelazione diversi da zero, mentre gli altri sono nulli.

Un modello ARIMA(0, d, 0) è una serie temporale che si trasforma in rumore bianco (processo puramente casuale) dopo essere stata differenziata d volte. Il modello (0, d, 0) è espresso da:

$$(1 - B)^d X_t = a_t \tag{3.8}$$

Il modello generale ARIMA(p, d, q), denominato processo autoregressivo integrato delle medie mobili di ordine p, d, q, assume l'espressione:

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)(1 - B)^d X_t = (1 - v_1 B - v_2 B^2 - ... - v_q B^q) \tag{3.9}$$

Un modello ARIMA(p, d, q) permette di descrivere una serie di osservazioni dopo diverse volte, al fine di estrarre le possibili fonti di non stagionalità. Questa formula può essere applicata a qualsiasi modello. Se c'è un componente p, d e q, uguale a zero, il termine viene dal eliminato corrispondente della formula generale. I modelli ciclici o stagionali sono quelli caratterizzati anche da oscillazioni cicliche,

dette variazioni stagionali. Le variazioni cicliche si sovrappongono, talvolta, a una tendenza secolare.

Le serie con tendenza secolare e variazioni cicliche possono essere rappresentate dai modelli ARIMA(p, d, q)(P, D, Q). La prima parentesi (p, d, q) si riferisce alla tendenza secolare, o parte regolare della serie, mentre la seconda parentesi (P, D, Q) si riferisce alle variazioni stagionali, o alla parte ciclica della serie temporale.

3.2.4 Modello SARIMA

Il modello SARIMA (Seasonal, Autoregressive, Integrated, Moving Average) è una estensione dei modelli ARIMA, proposti degli studiosi Box-Jenkins (1976), che tiene in considerazione: un parte stagionale autoregressiva (P), una parte di integrazione stagionale (D), una parte a media mobile stagionale (Q) e, infine, una quarta parte che descrive il tipo di stagionalità (S), permettendo, in questo modo, di considerare movimenti periodici di tipo stagionale. La parte stagionale può essere sia deterministica e indipendente dalle altre componenti, o stocastica e correlata con la componente non stagionale. Grazie a questo tipo di modellazione potrà venire trattata una stazionarietà anche di tipo non periodico. Il modello SARIMA $(p, d, q)(P, D, Q)_s$ è definito nella seguente forma:

$$\alpha(B)A(B^s)\nabla^d\nabla_s^D X_t = \theta(B)\Theta(B^s)_{\varepsilon_t} \tag{3.10}$$

dove:

- x_t è la variabile causale osservata.
- S è la periodicità stagionale.
- $\alpha(B) = (1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)$ è l'operatore autoregressivo non stagionale di ordine p stazionario. .
- $A(B^s) = (1-\phi_1 B-\phi_2 B^2 s-...-\phi_p B^p s)$ è l'operatore autoregressivo stagionale di ordine P stazionario.
- $\nabla^d = (1-B)^d$ è l'operatore differenze di ordine d non stagionale.
- $\nabla^D = (1-B^s)^D$ è l'operatore differenze di ordine D stagionale.
- $\theta(B) = (1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q)$ è l'operatore a media mobile non stagionale di ordine q invertibile.
- $\Theta(B^S) = (1-\Theta_1 B^S-\Theta_2 B^2 S-...-\Theta_Q B^Q S)$ è l'operatore a media mobile stagionale di ordine Q invertibile.

L'idea che sta alla base di questo modello è che le osservazioni lontane, ovvero che distano tra loro S periodi dovrebbero essere simili o fortemente correlate tra loro. Nel caso la serie abbia frequenza mensile avremo una periodicità pari a 12, nel caso di trimestrale la stagionalità è pari a 4, è così via.

La presenza di una eventuale non stazionarietà viene presa modellata con la stessa logica dei processi ARIMA ed eliminata grazie all'operatore "differenza stagionale" ∇_S^D, mentre gli operatori $\Phi(B^S)$ e $\Theta(B^S)$ permettono di cogliere e modellare la dipendenza stagionale tra le osservazioni lontane $S, 2S, ..., QS$ periodi tra loro.

3.2.5 Modello SARIMAX

I modelli fino a qui presentati mettono in relazione esclusivamente le osservazioni passate della sola variabile in esame e non tengono in considerazione eventuali informazioni contenute nella realizzazione di altre serie storiche attinenti. Le realizzazioni infatti possono essere influenzate anche da realizzazioni presenti e passate di altre serie storiche oltre che dal suo stesso passato.

I modelli SARIMAX sono, come nel caso dei modelli SARIMA, una estensione dei modelli ARIMA (p, d, q) tradizionali. In questo caso sono incluse le variabili esogeni che possono influenzare positivamente il processo predittivo ottenendo un modello del tipo $(p, d, q)x(P, D, Q)s$ con la seguente costruzione matematica:

$$(1 - B)^d(1 - b^S)^d z_t = \mu + \psi_i(B)X_{i,t} + \frac{\theta(B)\theta_S(B^S)}{\phi(B)\phi S(B^S)} + a_t \tag{3.11}$$

dove:

- z_t è la serie temporale.
- $X_{i,t}$ sono le serie temporali di predittori esterni.
- a_t è l'errore casuale.
- μ_t è la media della serie temporale.
- B è l'operatore di regressione.
- $\phi(B)$ è l'operatore autoreregressivo, un polinomio d'ordine p nell'operatore di regressione:

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p \tag{3.12}$$

- $\theta(B)$ è l'operatore di media mobile, un polinomio d'ordine q nell'operatore di regressione:

$$\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q \tag{3.13}$$

- $\psi_i(B)$ è l'operatore di trasferimento per l'effetto di $X_{i,t}$

3.2.6 Metodi di valutazione e calcolo dei parametri

Quando si costruisce un modello ci si imbatte nella stessa difficoltà che si presenta quando si tratta di costruire un modello di regressione. Uno dei criteri per valutare un modello è la "regolazione" delle singole variabili in un contesto di simulazione. Ci si aspetterebbe che i risultati di una simulazione storica si avvicinassero abbastanza accuratamente al comportamento del mondo reale. Un modo per testare un modello consiste nell'eseguire una simulazione storica esaminando fino a che punto ciascuna delle variabili endogene traccia la corrispondente serie storica. È, quindi, necessario disporre di una qualche misura quantitativa di questo fenomeno. La misura più frequentemente utilizzata è la "radice dell'errore quadratico medio" (RMSE).

La RMSE per la variabile Y_t è definita come segue:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_t^s - Y_t^a)^2}{N}} \tag{3.14}$$

dove:

- Y_t^s è il valore predetto di Y_t.
- Y_t^a è il valore reale di Y_t.
- N è il numero dei periodi della serie.

La RMSE costituisce, pertanto, una misura della deviazione della variabile simulata rispetto alla sua traiettoria temporale effettiva. È evidente che l'entità di tale errore può essere valutata solo rispetto al valore medio della variabile in questione.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{Y_t^s - Y_t^a}{Y_t^a})^2} \tag{3.15}$$

Si tratta, anche, di una misura della deviazione della variabile simulata rispetto alla sua traiettoria temporale reale, ma espressa in percentuale.

L'errore percentuale medio è definito come:

$$ME = \frac{\sum_{i=1}^{N}(Y_t^s - Y_t^a)}{N} \tag{3.16}$$

E l'errore medio percentuale è definito come:

$$MEP = \frac{1}{N}\sum_{i=1}^{N}(\frac{Y_t^s - Y_t^a}{Y_t^a})^2 \tag{3.17}$$

Lo svantaggio degli errori medi è che possono assumere valori prossimi allo zero quando i grandi errori positivi annullano i grandi errori negativi.

Nel caso in cui l'errore medio di simulazione assuma un valore prossimo a zero e il valore della REMC di simulazione abbia un valore più elevato, la REMC rappresenterebbe una migliore misura del funzionamento della simulazione. È molto probabile che un'equazione la cui regolazione statistica è molto buona abbia una regolazione di simulazione molto carente.

Nella presente tesi verrano effettuate una serie di metodi di validazioni, come:

Errore Percentuale Assoluto Medio

Errore percentuale Assoluto Medio (MAPE o Mean Absolute Percentage Error) è un indicatore delle prestazioni della previsione della domanda che misura la dimensione dell'errore (assoluto) in termini percentuali.

$$MAPE = \frac{\sum_{i=1}^{N}(\frac{|Y_t^s - Y_t^a|}{|Y_t^a|})}{N} \tag{3.18}$$

Errore Percentuale Mediano Assoluto

Il Median Absolute Percentage Error (Mdape) si trova ordinando la percentuale assoluta di errore (APE) dal più piccolo al più grande, e usando il suo valore medio (o la media dei due valori medi se N è un numero pari) come la mediana.

$$MDAPE = Median(\frac{|Y_t^s - Y_t^a|}{|Y_t^a|}) \tag{3.19}$$

Errore Percentuale Simmetrico Medio Assoluto

Symmetric mean absolute percentage error, è una misura di accuratezza basata su errori percentuali (o relativi). Di solito è definita come segue:

$$SMAPE = \frac{\sum_{i=1}^{N}(\frac{|Y_t^s - Y_t^a|}{(|Y_t^s| + |Y_t^a|)/2})}{N} * 100 \tag{3.20}$$

Errore Percentuale Assoluto Mediano Simmetrico

La symmetric median absolute percentage error, è una misura di accuratezza basata su errori percentuali, come nel caso precedente, però facendo uso della mediana. Viene definita come segue:

$$SMDAPE = Median(\frac{|Y_t^s - Y_t^a|}{(|Y_t^s| + |Y_t^a|)/2}) * 100 \tag{3.21}$$

Errore percentuale assoluto medio di Arcotangente

MAAPE è una nuova metrica di errore percentuale assoluto, ed è stato sviluppato guardando MAPE da un angolo diverso. In sostanza, MAAPE è una pendenza come angolo, mentre MAPE è una pendenza come rapporto. Viene definita come segue:

$$MAAPE = \frac{\sum_{i=1}^{N} arctan(\frac{|Y_t^s - Y_t^a|}{|Y_t^a|})}{N} \tag{3.22}$$

Stima dei parametri

Le correzioni auto regressive migliorano i risultati del modello aggiungendovi informazioni. Anche se è possibile che non si sappia perché i termini di errore additivo si comportano nel modo in cui lo fanno, si può osservare che sono fortemente correlati e includere tali informazioni nel modello.

Finora sono state menzionate due possibili alternative al metodo dei minimi quadrati ordinari, ma esistono altri metodi di stima.

Ciascuno di questi metodi è stato utilizzato per stimare i modelli delle schede tecniche nel presente progetto. I metodi utilizzati sono i seguenti:

- Minimi quadrati ordinari (MCO).
- Minimi quadrati in due fasi (MC2).
- MCO più una correzione auto-regressiva di primo ordine (MCOAUTO1).
- MC2E più una correzione auto-regressiva di primo ordine (MC2EAUTO1).
- MCO più una correzione auto-regressiva di secondo ordine (MCOAUTO2).

4

Metodologia ed Implementazione del modello

Nel capitolo corrente verrà trattata in dettaglio la metodologia con cui sono state effettuate la fase di modellazione e la succesiva implementazione del modello scelto sui dati storici delle schede tecniche con i rispettivi risultati della predizione delle ore produttive.

4.1 Preparazione dei Dati

Per la preparazione del dataset è stato effettuato un lavoro di pulizia e standardizzazione dei dati.

Come prima attività abbiamo aggregato il dataset "Tracciato_09-03-2021.csv", per i campi anno, mese e giorno, ottenendo 149214 righe. Come seconda attività, abbiamo raggruppato questi tre campi in un unico campo chiamato "data", con il formato "anno/mese", come si può osservare nella Figura 4.2.

```
In [569]:
# Rimuovo tutte le righe che non hanno: anno, mese e giorno
df = df[df['anno'].notna() & df['mese'].notna() & df['giorno'].notna()]
# Dimensioni del dataset senza NaN nelle date
df.shape

Out[569]: (149214, 123)

In [570]:
# Transform la data in una stringa
def date_parse(val):
    return str(int(val)).zfill(2).strip()

In [571]:
# Aggiungo il campo data
df['data'] = df.apply(lambda row: date_parse(row['anno']) +'-'+ date_parse(row['mese']) +'-'+ date_parse(row['giorno']), axis

In [572]:
# Vengono rimosse date non valide
#df = df[~df['data'].isin(['2014-11-31'])] # non esiste
#df = df[~df['giorno'].isin([50, 51, 70, 71, 80, 81, 90, 98, 99])] # errore nel campo giorni
# Vengono rimossi tutti i dati del 2021
#df = df[~df['anno'].isin([2021])]

In [573]:
df['data'] = pd.to_datetime(df['data'], format='%Y-%m-%d')
```

Figura 4.1. Lettura e standardizzazione del dataset

4.1.1 Training e Test set

Per la fase successiva, si procede alla scelta delle variabili esogene relative alla scheda da modellare e predire. Per questo compito sono state scelte le seguenti variabili: “n_istituti”, “n_filiali”, “n_operazioni”, “imp_fattura”, “qtaOreProduttive”. Le attività effettuate sono dettagliate di seguito:

- Individuazione dei giorni mancanti dentro la serie temporale e resampling.
- Individuazione dei dati di tipo NaN, con la rivalorizzazzione al valore zero.
- Raggrupamento della data per anno e mese (Figura 4.2).
- Rimozione dei primi 8 mesi del 2014, per mancanza di dati relativi a quel periodo (Figura 4.3).

Figura 4.2. Attività di creazione del dataframe raggrupato per anno e mese

Con la selezione delle variabili esogene, viene realizzato un calcolo della correlazione tra queste variabili, e la variabile “qtaOreProduttive”, come si può osservare nella Figura 4.4.

4.2 Selezione dei parametri ottimali

Per la selezione dei parametri ottimali, generalmente, nel modello di base ARIMA, bisogna fornire i valori p,d e q, che sono essenziali nella sua struttura. Si utilizzano tecniche statistiche per individuare questi valori, facendo la differenza per eliminare la non stagionalità. Nel modulo Auto ARIMA, utilizzato in questo progetto per la ricerca dei parametri, il modello stesso genera i valori ottimali di p, d e q sulla base della migliore previsione ottenibile con il set di dati di training.

Dopo aver preparato i dati per il modulo Auto ARIMA, si procede con il training del modello (Figura 4.6). A tal fine vengono impostati i seguienti parametri:

- Come variabile dipendente, la serie temporale “QtaOre”.
- Come variabile indipendente tutte le altre variabili esogene.
- Un parametro m = 12 mesi, con m pari al periodo di differenza stagionale.
- Il campo “Seasonal” pari a true, che indica ad ARIMA la presenza di una stagionalità.

```
In [1041]:
if df_single[:1].index[0][0] == 2014 and df_single[:1].index[0][1] == 1:
    # Remove first 8 months of 2014 without qtaOre
    df_single = df_single[8:]
```

```
In [1042]:
df_single
```

Out[1042]:

anno	mese	n_istituti	n_filiali	n_operazioni	imp_fattura	qtaOreProduttive
2014	9	0.0	0.0	0.0	0.0	1007.87200
	10	0.0	0.0	0.0	0.0	1151.59040
	11	0.0	0.0	0.0	0.0	1028.12160
	12	0.0	0.0	0.0	0.0	935.28832
2015	1	0.0	0.0	0.0	0.0	962.39040
...	...	...	...	...	...	...
2020	8	20.0	260.0	2857.0	43628.0	1472.03488
	9	22.0	286.0	6546.0	97240.0	1598.73056
	10	22.0	286.0	5567.0	83384.0	1650.53056
	11	21.0	273.0	5202.0	79901.0	1897.18208
	12	21.0	273.0	4412.0	67681.0	1751.45664

Figura 4.3. Attività di rimozione dei primi 8 mesi del 2014

```
In [1043]:
df_single.corr().style.background_gradient(cmap='viridis').set_precision(2)
```

Out[1043]:

	n_istituti	n_filiali	n_operazioni	imp_fattura	qtaOreProduttive
n_istituti	1.00	0.92	0.94	0.93	-0.16
n_filiali	0.92	1.00	0.87	0.88	-0.15
n_operazioni	0.94	0.87	1.00	0.99	-0.08
imp_fattura	0.93	0.88	0.99	1.00	-0.04
qtaOreProduttive	-0.16	-0.15	-0.08	-0.04	1.00

Figura 4.4. Calcolo della correlazione tra le variabili di analisi

4.3 Predizione del modello

In questa sezione, si inizia con la fase di predizione dei dati, utilizzando i parametri ottimali trovati dal modulo Auto ARIMA, che sono, rispettivamente, (1,0,0), e 12 mesi come periodo di differenza stagionale.

A partire dal modello definito, viene eseguita l'attività predittiva del modello SARIMAX, a 6 mesi, con un intervallo di confidenza al 95% per la scheda tecnica SCREBO. I risultati saranno analizzati nel prossimo capitolo.

Addestramento del modello

In [1045]:
```
otherFields = fields.copy()   # variabili diverse da qtaOre utilizzate per fare prediction
varQtaOre = otherFields.pop() # variabile qtaOre

MONTHS = 6
```

In [1046]:
```
# Creo una lista con tutti i valori delle varibili esogene
exogen_list = []
for otherField in otherFields:
    exogen_list.append(df_single[otherField].values)
```

In [1047]:
```
# Mi salvo le date che andrò a predire
date = df_single.index[-MONTHS:]
# Resetto gli index del dataframe
df_single = df_single.reset_index()

# Prendo la qtaOre lavorate
qtaOre  = df_single[varQtaOre]
# Faccio lo stack delle variabili esogene
exogVar = np.column_stack(tuple(exogen_list))

qtaOre_train = qtaOre[:-MONTHS]
qtaOre_test = qtaOre[-MONTHS:]
exogVar_train = exogVar[:-MONTHS]
exogVar_test = exogVar[-MONTHS:]
```

Figura 4.5. Preparazione dati per Modulo Auto ARIMA

In [1048]:
```
# Ricerca dei miglior parametri per il modello SARIMAX
par = auto_arima(y=qtaOre_train, X=exogVar_train, m=12, seasonal=True, trace=True)
```

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(1,0,1)[12] intercept   : AIC=794.809, Time=1.01 sec
 ARIMA(0,0,0)(0,0,0)[12] intercept   : AIC=868.960, Time=0.18 sec
 ARIMA(1,0,0)(1,0,0)[12] intercept   : AIC=785.448, Time=0.62 sec
 ARIMA(0,0,1)(0,0,1)[12] intercept   : AIC=843.338, Time=0.69 sec
 ARIMA(0,0,0)(0,0,0)[12]             : AIC=1072.363, Time=0.08 sec
 ARIMA(1,0,0)(0,0,0)[12] intercept   : AIC=782.955, Time=0.18 sec
 ARIMA(1,0,0)(0,0,1)[12] intercept   : AIC=784.524, Time=0.71 sec
 ARIMA(1,0,0)(1,0,1)[12] intercept   : AIC=786.629, Time=0.86 sec
 ARIMA(2,0,0)(0,0,0)[12] intercept   : AIC=785.317, Time=0.32 sec
 ARIMA(1,0,1)(0,0,0)[12] intercept   : AIC=783.699, Time=0.28 sec
 ARIMA(0,0,1)(0,0,0)[12] intercept   : AIC=843.825, Time=0.16 sec
 ARIMA(2,0,1)(0,0,0)[12] intercept   : AIC=785.574, Time=0.37 sec
 ARIMA(1,0,0)(0,0,0)[12]             : AIC=783.259, Time=0.14 sec

Best model:  ARIMA(1,0,0)(0,0,0)[12] intercept
Total fit time: 5.638 seconds
```

Figura 4.6. Tracing del modello Auto ARIMA

In [1049]:
```
# Faccio il fit del modello SARIMAX
mod = sm.tsa.statespace.SARIMAX(endog=qtaOre_train, exog=exogVar_train, order=par.order, seasonal_order=par.seasonal_order)
res = mod.fit()
```

Figura 4.7. Fit del modello SARIMA con i parametri ottimali

Plot dei risultati

In [1051]:
```
final_date = []
for i in range(0, len(date.get_level_values(0))):
    final_date.append(str(date.get_level_values(0)[i]) + '-' + str(date.get_level_values(1)[i]).zfill(2))
```

In [1052]:
```
final_results = pd.concat([qtaOre_test, predictions.predicted_mean, interval.iloc[:,0], interval.iloc[:,1]], axis=1).clip(low
final_results.columns = ['actual', 'predicted', 'lower', 'upper']
final_results['data'] = final_date
```

In [1053]:
```
# grafico dei risultati
fig, ax = plt.subplots(figsize=(20,8))
ax.plot(final_results['data'], final_results['predicted'], '--', color='blue', label='Prediction')
ax.plot(final_results['data'], final_results['actual'], lw=1, color='black', alpha=0.5, label='Real Data')

ax.fill_between(final_results['data'], final_results['lower'], final_results['upper'], alpha=0.05, color='red')

ax.yaxis.grid(color='gray', linestyle='dashed', alpha=0.3)
plt.xticks(rotation='45')
plt.legend()
plt.title('Zoom sul forecast')
plt.show()
```

Figura 4.8. Attività di stampa dei risultati della predizione del modello SARIMAX

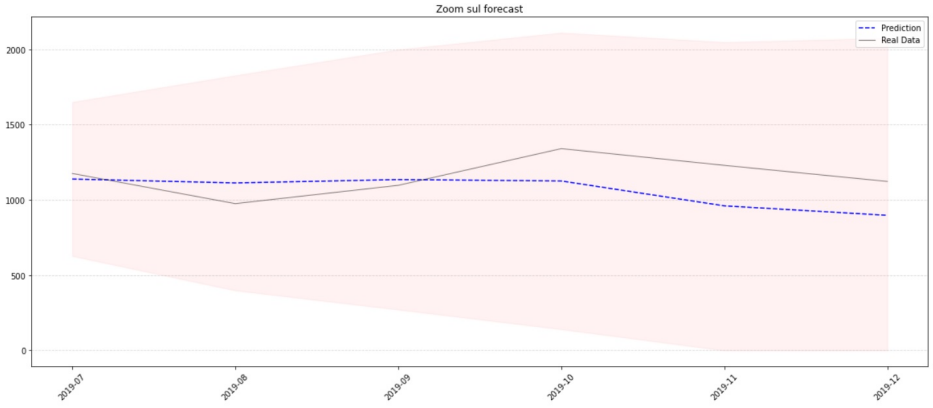

Figura 4.9. Risultato della predizione del modello SARIMAX a 6 mesi

5

Risultati

In questo capitolo si discuterà l'ultima fase del lavoro finora presentato, ovvero la validazione dei risultati della predizione del modello SARIMAX.

5.1 Validazione dei risultati

In accordo con quanto definito durante la fase di progettazione, sono state implementate varie metodologie di calcolo per valutare l'accuratezza del modello SARIMAX.

5.1.1 Indicatori di performance

Gli indicatori di performance, applicati ai risultati finali della predizione, sono riportati nel seguente elenco:

- Mean Absolute Percentage Error (MAPE).
- Median Absolute Percentage Error (MDAPE).
- Symmetric Mean Absolute Percentage Error (SMAPE).
- Symmetric Median Absolute Percentage Error (SMDAPE).
- Mean Arctangent Absolute Percentage Error (MAAPE).
- Root Mean Square Percentage Error (RMSPE).
- Root Median Square Percentage Error (RMDSPE).

Con gli indicatori di performance definiti, si procede alla loro applicazione nella fase di validazione dei dati di predizione in confronto con i dati reali, come si vede nella Figura 5.1.

In seconda fase, si procede con un elenco di schede tecniche preselezionate in base alla quantità di dati presenti delle variabili, raggruppate per mese e anno; viene anche, considerata la presenza di dati in variabili come "n_istituti", "n_filiali", etc. Questo elenco, viene visualizzato nella Figura 5.1.

In base ad un elenco di schede tecniche selezionate, si procede all'attività di predizione del modello SARIMAX, per 6, 9 e 12 mesi futuri. Questo ci ha permesso osservare il comportamento del modello a lungo termine, con la valutazione di ogni indicatore (Figure 5.2, 5.3 e 5.4;).

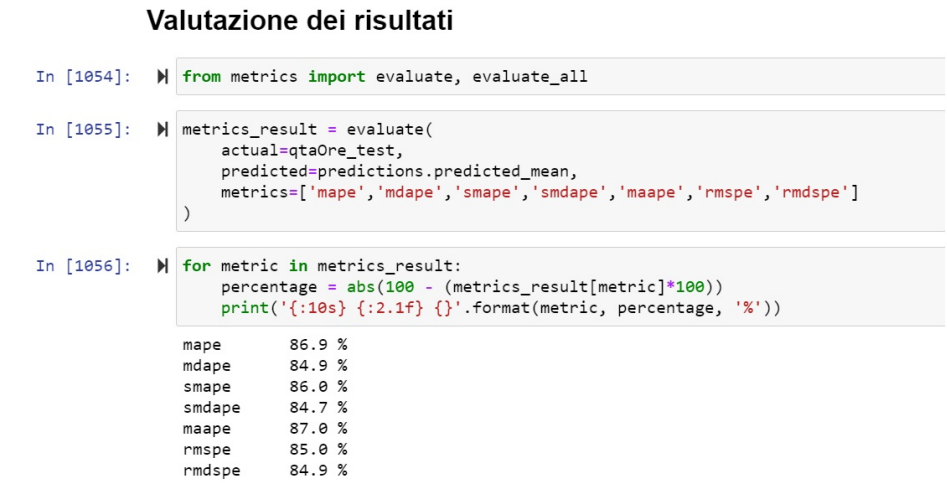

Figura 5.1. Validazione dei risultati della predizione

SCHEDE TECNICHE
SCARTE
STITBA
SBONIF
SPTFDI
SASSE
STEINC
SPTEFF
STITES
SCREBO
SESASS
SRID
SESTES
SDELFI
SPENSI
SESGAR
SBECBI
SDOCOT
SCAIAS
SCOTRA
SCRECM
SCREIN
SCOABB
SGTATM
SSISVI
SCREIM

Tabella 5.1. Elenco Schede Tecniche per la Predizione

	Forecast a 6 MESI						
SCHEDA	mape	mdape	smape	smdape	maape	rmspe	rmdspe
SCARTE	80.70	76.70	78.20	73.70	81.10	79.00	76.70
STITBA	88.90	91.40	88.10	91.40	89.00	86.80	91.40
SBONIF	82.30	86.30	81.90	85.30	82.60	80.60	86.20
SPTFDI	83.00	83.10	82.80	82.10	83.30	80.80	82.50
SASSE	91.20	90.40	91.30	89.90	91.20	90.30	90.40
STEINC	90.70	90.30	90.00	89.70	90.80	88.70	89.70
SPTEFF	85.50	87.50	85.80	86.70	85.80	81.90	87.50
STITES	68.20	72.50	59.00	67.40	70.10	62.90	71.00
SCREBO	90.90	91.80	91.20	91.40	90.90	89.70	91.60
SESASS	93.90	95.00	94.00	94.90	93.90	92.10	94.50
SRID	68.10	69.50	58.60	66.40	70.40	61.00	68.70
SESTES	89.60	92.00	90.10	91.70	89.80	85.70	92.00
SDELFI	81.80	85.30	78.80	84.00	82.40	77.30	84.80
SPENSI	80.90	76.90	81.00	76.60	81.40	77.20	76.90
SESGAR	92.40	92.20	92.10	92.20	92.50	91.40	92.20
SBECBI	23.30	19.60	22.31	31.10	37.70	16.80	19.60
SDOCOT	60.40	69.30	48.90	66.70	64.70	51.20	68.80
SCAIAS	87.80	89.70	89.10	90.20	88.10	83.30	89.50
SCOTRA	53.10	52.90	36.70	38.30	56.70	51.20	52.90
SCRECM	90.10	90.70	90.00	91.10	90.10	88.70	90.30
SCREIN	89.10	89.10	89.00	89.20	89.20	86.60	89.10
SCOABB	55.20	47.90	37.40	29.40	59.50	49.20	47.80
SGTATM	56.30	61.90	59.00	52.80	63.20	41.40	61.60
SSISVI	53.40	51.90	38.00	41.00	57.60	49.30	51.50
SCREIM	53.90	51.30	44.30	46.80	58.20	49.00	51.20

Figura 5.2. Forecast SARIMAX a 6 mesi

	Forecast a 9 MESI						
SCHEDA	mape	mdape	smape	smdape	maape	rmspe	rmdspe
SCARTE	80.10	78.30	77.30	75.70	80.50	77.80	78.30
STITBA	75.70	76.40	71.70	73.30	76.40	73.90	76.40
SBONIF	71.30	67.90	65.00	61.80	72.60	67.30	67.90
SPTFDI	89.80	90.70	89.70	90.20	89.90	87.60	90.70
SASSE	74.40	74.00	77.90	77.00	75.40	71.10	74.00
STEINC	92.20	91.10	91.90	91.00	92.20	91.10	91.10
SPTEFF	88.30	89.90	88.10	89.40	88.40	87.40	89.90
STITES	70.10	72.90	63.10	68.70	71.50	66.70	72.90
SCREBO	86.80	89.40	85.10	88.80	87.00	83.30	89.40
SESASS	86.10	85.00	87.30	86.00	86.30	83.80	85.00
SRID	58.70	65.80	42.10	58.80	62.40	52.80	65.80
SESTES	91.40	92.70	91.90	92.90	91.50	87.10	92.70
SDELFI	20.40	24.50	17.10	21.20	35.50	13.80	24.50
SPENSI	81.20	84.00	80.00	83.30	81.70	78.30	84.00
SESGAR	92.00	93.50	92.30	93.80	92.10	89.60	93.50
SBECBI	16.70	19.10	33.10	35.70	33.90	9.80	19.10
SDOCOT	65.10	67.40	49.20	71.90	68.90	53.90	67.40
SCAIAS	15.30	10.20	53.50	11.00	30.80	13.40	10.20
SCOTRA	7.30	62.90	47.80	54.50	57.50	45.40	62.90
SCRECM	90.30	91.30	89.70	91.60	90.40	88.60	91.30
SCREIN	40.00	63.00	58.80	68.80	53.10	20.90	63.00
SCOABB	52.20	63.80	27.00	55.80	58.00	43.80	63.80
SGTATM	38.20	46.60	28.10	27.20	44.70	32.60	46.60
SSISVI	19.70	48.50	28.40	30.60	47.20	30.70	48.50
SCREIM	83.00	83.00	82.80	81.40	83.50	77.30	83.00

Figura 5.3. Forecast SARIMAX a 9 mesi

	Forecast a 12 MESI						
SCHEDA	mape	mdape	smape	smdape	maape	rmspe	rmdspe
SCARTE	38.40	32.10	6.80	2.80	45.80	35.80	32.10
STITBA	13.90	9.10	15.20	16.70	29.50	12.90	9.10
SBONIF	63.80	59.40	52.00	48.70	66.40	58.10	59.00
SPTFDI	71.50	71.60	65.50	66.90	72.60	68.80	71.60
SASSE	39.90	36.90	3.60	7.00	48.40	33.40	36.50
STEINC	38.90	30.00	4.00	27.70	47.10	33.70	30.00
SPTEFF	73.30	71.70	68.00	66.90	74.20	70.70	71.60
STITES	23.50	16.50	31.60	33.40	36.10	20.40	16.50
SCREBO	67.80	64.50	59.30	56.50	69.60	63.40	64.00
SESASS	76.80	74.50	73.30	70.80	77.30	75.40	74.50
SRID	43.20	38.10	14.80	9.50	49.80	39.40	37.70
SESTES	53.50	49.70	38.10	32.80	56.80	52.20	49.70
SDELFI	72.50	74.10	64.60	70.20	74.20	65.80	73.80
SPENSI	75.60	72.40	70.60	67.90	76.60	71.30	72.20
SESGAR	78.10	75.90	74.80	72.60	78.60	76.10	75.90
SBECBI	1.10	8.70	6.20	1.20	24.60	3.80	8.70
SDOCOT	56.80	61.50	36.80	51.70	61.90	47.20	60.70
SCAIAS	31.40	12.90	22.10	24.50	43.40	23.30	12.90
SCOTRA	46.40	53.00	40.40	38.10	55.10	32.80	52.50
SCRECM	88.30	91.10	88.80	90.70	88.50	84.50	91.10
SCREIN	38.70	54.90	58.00	63.20	52.10	19.60	54.80
SCOABB	35.20	31.30	22.10	14.70	43.70	32.20	31.30
SGTATM	39.60	51.20	23.50	35.40	40.10	26.00	51.10
SSISVI	31.00	33.80	9.50	2.10	19.40	12.30	33.80
SCREIM	21.40	22.90	32.10	21.50	33.40	28.30	11.90

Figura 5.4. Forecast SARIMAX a 12 mesi

5.1.2 Analisi dei risultati

Secondo i risultati ottenuti nella fase di predizione, si procede all'analisi del modello nelle varie tipologie presentate, ovvero SARIMAX a 6, 9 e 12 mesi.

- Nel primo caso, con un SARIMAX a 6 mesi, sono presenti buoni risultati a seconda degli indicatori calcolati, con una precisione alta presente in più della metà degli indicatori e delle schede tecniche selezionate. Si raggiunge un 93.90% di precisione in casi come la scheda SESASS, mentre, in casi come la scheda SBECBI, otteniamo solo un 16.80%.
- Nella seconda impostazione del modello SARIMAX a 9 mesi, gli indicatori di precisione tendono a scendere in più della metà delle schede tecniche, ottenendo diminuzioni in schede tecniche come SASSE, che con il modello a 6 mesi raggiungeva 91.3%, ma nel presente modello presenta solo un 77.9% di precisione.
- Nell'ultimo modello SARIMAX impostato per 12 mesi di predizione, il comportamento degli indicatori tende a diminuire per ogni scheda, come già osservato nel modello a 9 mesi.

6

Conclusioni

In questo ultimo capitolo saranno trattate le conclusioni riguardanti il lavoro svolto e verranno esposte le possibili migliorie da apportare in futuro.

6.1 Conclusioni

L'analisi dei dati di partenza, insieme a tutte le attività di ETL, sono state un punto fondamentale della presente tesi, ottenendo come risultato un modello del database, insieme alle variabili principali identificate per la costruzione del modello.

Nell'elaborato corrente è stato presentato lo sviluppo di un modello di predizione come prodotto finale, basato sui dati storici concessi dalla N.B.C, per dare supporto al calcolo della quantità di ore produttive future, effettuate dentro una specifica scheda tecnica.

Il modello SARIMAX si è dimostrato un buon strumento per la predizione a breve termine della variabile "QuantitaOreProduttive". Dato che il modello si basa principalmente sull'informazione a disposizione, esso ha raggiunto un accettabile livello di precisione per la maggioranza delle schede tecniche.

La funzionalità dei modelli di serie temporali è ottimale a breve termine, ma mostra una certa difficoltà nei periodi di predizione a lungo termine.

L'utilizzo di Python si è dimostrato ottimale per la realizzazione di processi ETL, come per l'applicazione di modelli di serie temporali e analisi di dati, a partire da librerie sviluppate a questo scopo.

Infine, dopo aver finalizato tutta la fase di modellazione e predizione dei dati, è stata discussa la loro applicazione a predizioni a lungo termine, se il modello scelto fosse adatto a questo lavoro.

6.2 Sviluppi Futuri

Il lavoro realizzato finora, tanto in fase di costruzione del modello come nella scelta dei parametri ottimali, per la predizione successiva, rappresenta una parte importante del progetto in generale, visto che l'obiettivo futuro è avere uno strumento software che contenga il modello di predizione.

Per quanto concerne il modello finale SARIMAX, si propone di attuare nuovi meccanismi di stima dei parametri, visto che quelli forniti dal modulo auto-arima, funzionano automaticamente e sono maggiormente basati su metodologie di minimi quadrati ordinari.

Si pensa di implementare un'altra tipologia di modelli che potrebbero dare più precisione alla predizione sia a breve che a lungo termine. Un esempio molto promettente sono i modelli di equazioni simultanee, che mostrano ad un livello più profondo le relazioni tra tutte le variabili presenti, dove ogni variabile viene rappresentata con un modello.

Per quanto riguarda lo strumento utilizzato, Python, e le sue librerie di modellazione, si pensa allo sviluppo di un software sulla base di questo linguaggio, in cui i dati possano essere selezionati attraverso un'interfaccia e il modello agisca da solo nella ricerca dei parametri ottimali per la variabile scelta.

Riferimenti bibliografici

1. R. Adhikari and R.K. Agrawal. An Introductory Study on Time Series Modeling and Forecasting: Time Series Forecasting Using Stochastic Models. *Lambert Academic Publishing, Germany*, 2013.
2. J. S. Armstrong. Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting 22*, 2006.
3. G.E.P Box., M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2008.
4. P.P. Dabral and M.Z. Murry. Modelling and Forecasting of Rainfall Time Series Using SARIMA. *Environ. Process*, 2017.
5. S. de la Fuente Fernandez. Series temporales. *Departamento de Economia y Empresariales- Universidad Autonoma de Madrid*, 2009.
6. N. D'Ercole. Progettazione e realizzazione di un data warehouse per il supporto al controllo contabile e l'analisi delle vendite. Master's thesis, Università Degli Studi di Pisa, 2007.
7. S. Feuerstein. *Oracle PLSQL Best Practices*. O'Reilly Media, 2007.
8. J. De Gregorio. *Macroeconomía, Teoría y Políticas*. Pearson-Educación, 2012.
9. J.E. Hernandez. Aplicacion del modelo Box-Jenkins. Master's thesis, Fundacion Universitaria Los Libertadores, 2019.
10. Y. Hilpisch. *Python for Finance 2e: Mastering Data-Driven Finance*. O'reilly, 2019.
11. L. Igual and S. Segui. *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer, 2017.
12. P. Kumar, K. Sinha, K. Nere, Y. Shin, R. Ho, and B. Mlinar. A machine learning framework for computationally expensive transient models. *Springer*, 2020.
13. E. Manzato. Combinazione di modelli stagionali per la previsione dei prezzi dell'elettricità nel mercato Nord Pool Spot. Master's thesis, Università Ca'Foscari Venezia, 2013.
14. H. Mombeni, S. Rezaei, and S. Nadarajah. Estimation of Water Demand in Iran Based on SARIMA Models. *Environ Model Assess*, 2013.
15. S.S. Namini and A.S. Namin. Forecasting Economics and Financial Time Series: ARIMA vs. LSTM. *Statistical Finance*, 2018.
16. R. Pindick and D. Rubinfeld. *Econometría: Modelos y pronósticos*. Mc Graw Hill Education, 2001.
17. N. Chavez Quisbert. Modelos ARIMA. *Universidad Catolica Boliviana*, 2009.
18. D. Sartore. Introduzione all'econometria: Processi stocastici a media mobile e auto regressivi. Master's thesis, Università Ca'Foscari Venezia, 2011.
19. D. Stoffer and R. Shumway. *Time series analysis and its applications*. Springer, 2011.

20. J. VanderPlas. *Python Data Science Handbook: Tools and Techniques for Developers: Essential Tools for working with Data.* O'reilly, 2016.
21. R. Weron. Modeling and Forecasting Electricity Loads and Prices. *NewYork: John Wiley & Sons*, 2006.
22. M.E. Williams. *The Ultimate Beginners' Guide to Learning Python Data Science Step by Step.* Springer, 2019.

Ringraziamenti

Un enorme ringraziamento alla mia bella famiglia, per avermi dato tutto il vostro sostegno e supporto per poter realizzare questo incredibile sogno. Per avermi dato la forza di continuare nei momenti di difficoltà, nonostante le distanze, non mi resta che ringraziarvi.

I miei genitori, Arminda e Gabriel, per essere sempre con me, dandome consigli, incoraggiandomi a seguire i miei studi, e continuare a superarmi giorno per giorno.

A mio fratello Nick, per avermi dato queste gioie nei giorni più difficili, questo sogno è anche grazie a te.

Vorrei ringraziare al mio tutor prof. Domenico Ursino, per accompagnarmi in tutto il processo, fino al culmine, offrendomi il suo costante sostegno, insegnamenti e consigli per la presente tesi.

Un grande ringraziamento a Gianluca Bonifazi, per guidarmi nello sviluppo delle pratiche, ottenendo nuove conoscenze nell'area, e per la sua collaborazione nella scrittura della tesi.

Ringraziare ai miei colleghi Elia, Luca e Giacomo, per il supporto che mi forniscono, consigli e la sua amicizia dal primo momento che ci siamo conosciuti, diventando la mia seconda famiglia in Italia.

All'università UNIVPM, per avermi dato l'opportunità di continuare i miei studi in Italia, e quindi di continuare la mia formazione professionale con la acquisizione di nuove conoscenze.

Infine, vorrei ringraziare tutti i miei belli amici del cuore Carlo, Mattia, Fabio, Giovanna, Sara e Morris per essere sempre al mio fianco, sempre presenti al tavolo di 160, aiutandomi a migliorare il mio italiano e mostrarmi la bella cultura italiana.