Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Matematica

A.a. 2025/2026

Sessione di laurea Marzo 2026

Modelli di previsione su serie temporali per la gestione delle Spare Parts

Relatore:

Gianluca Mastrantonio

Candidata:

Sabrina Bucelli

Alla mia famiglia

Alle volte uno si crede incompleto
ed è soltanto giovane.
[I. Calvino]

Sommario

Il presente elaborato affronta il tema del forecasting della domanda applicato alla gestione delle spare parts in ambito industriale, con particolare attenzione ai contesti caratterizzati da elevata variabilità e presenza di domanda intermittente. L'obiettivo del lavoro è sviluppare un framework metodologico e applicativo in grado di analizzare, modellizzare e confrontare differenti approcci previsionali, integrandoli all'interno di uno strumento strutturato a supporto dei processi decisionali.

Nel primo capitolo viene introdotto il contesto di riferimento e vengono esplicitate le motivazioni del lavoro. Si analizzano le peculiarità della domanda delle parti di ricambio, la sua rappresentazione tramite serie temporali e l'importanza di una preliminare caratterizzazione statistica degli articoli. In questa sezione viene inoltre descritto il tool di inventory management sviluppato e l'impianto infrastrutturale che ne consente il funzionamento, delineando in modo sintetico l'architettura tecnologica a supporto delle attività di forecasting e simulazione.

Il secondo capitolo presenta i principali modelli per serie temporali analizzati nel lavoro. Vengono approfondite le tecniche di decomposizione, i modelli di smoothing esponenziale, i modelli ARIMA e gli approcci specificamente progettati per la domanda intermittente, come il metodo di Croston e le sue varianti, evidenziandone ipotesi, caratteristiche e ambiti di applicazione.

Il terzo capitolo descrive il contesto applicativo e la metodologia adottata. Vengono illustrati il dataset utilizzato, la struttura dei dati, il processo di valutazione del demand pattern e la clusterizzazione della domanda, nonché il processo di filtering e la metodologia di costruzione delle previsioni.

Il quarto capitolo è dedicato ai risultati sperimentali e all'analisi comparativa tra i modelli considerati. Vengono definiti gli obiettivi dell'analisi, l'impostazione delle simulazioni e i criteri di confronto, per poi discutere in modo sistematico le performance ottenute sui diversi cluster di articoli.

Il quinto capitolo esplora infine nuovi approcci per la gestione della domanda intermittente, introducendo metodi basati su Markov chain e bootstrap e proponendo un'estensione metodologica attraverso il Modified Markov Chain Model (MMCM). Il lavoro si conclude con una riflessione sulle prospettive di sviluppo future nel forecasting della domanda intermittente e non intermittente.

Nel suo complesso, l'elaborato integra analisi teorica, implementazione metodologica e sviluppo applicativo, con l'obiettivo di costruire un sistema di forecasting strutturato e coerente con le specificità operative del contesto industriale analizzato.

Ringraziamenti

Desidero dedicare queste righe a tutte le persone che, in modi diversi, hanno accompagnato e sostenuto il mio percorso fino a questo traguardo.

Ringrazio innanzitutto il mio Relatore per la guida, la disponibilità e i preziosi suggerimenti che hanno orientato lo sviluppo di questo lavoro.

Il ringraziamento più importante va alla mia famiglia, per il sostegno costante, la fiducia e la presenza fondamentale in ogni fase del mio percorso accademico. Un pensiero speciale va alle mie nipoti, che con la loro spontaneità e il loro affetto hanno saputo rendere più leggeri anche i momenti più impegnativi.

Alle mie amiche va la mia riconoscenza per la vicinanza, l'ascolto e la condivisione dei momenti più impegnativi, ma anche di quelli più leggeri; per avermi sempre spronata, sostenuta e per aver creduto in me, anche quando io facevo più fatica a farlo.

Infine, ringrazio i miei colleghi, per il confronto quotidiano, la collaborazione e il supporto che hanno contribuito alla mia crescita personale e professionale.

Indice

Elenco delle figure

Capitolo 1

Forecasting e gestione delle spare parts: contesto e motivazioni

1.1 Contesto e obiettivi

Nel panorama industriale contemporaneo, e in particolare nel settore automotive, il forecasting, o previsione, della domanda rappresenta una leva strategica all'interno dei processi di gestione e organizzazione della supply chain, in quanto rappresenta uno strumento di supporto alle decisioni di pianificazione, approvvigionamento e distribuzione lungo l'intero flusso logistico dei materiali. In termini generali, il forecasting può essere definito come il processo analitico di stima di eventi futuri basato su dati storici, informazioni sul contesto e modelli quantitativi o qualitativi, con l'obiettivo di ridurre l'incertezza decisionale e migliorare l'allocazione delle risorse. Tale attività assume una rilevanza ancora maggiore nella gestione dei componenti di ricambio, la quale rappresenta uno degli ambiti più complessi e strategicamente critici dei sistemi logistici, poiché richiede di garantire elevati livelli di servizio in presenza di una domanda spesso intrinsecamente incerta. A differenza dei prodotti finiti, le spare parts sono caratterizzate da lunghi cicli di vita, bassi volumi medi di consumo e una domanda spesso irregolare, intermittente e sporadica, influenzata da politiche di manutenzione e dinamiche del parco installato circolante. La disponibilità del ricambio al momento della necessità diviene pertanto un fattore chiave per la continuità operativa dei beni strumentali, per la qualità del servizio percepita e, in definitiva, per la soddisfazione overall del cliente. In questo contesto, il demand forecasting si configura come un elemento strutturale dei processi di inventory management e inventory optimization, poiché i modelli di

gestione delle scorte utilizzano le previsioni della domanda come input fondamentale per la definizione delle politiche di riordino, incluse le scorte di sicurezza e i parametri di riapprovvigionamento. Un miglioramento dell'accuratezza del forecast consente, a parità di livello di servizio, di ridurre i livelli di inventario necessari, aumentando l'efficienza complessiva del sistema logistico. Nel caso delle spare parts, come anticipato, tale relazione risulta ulteriormente amplificata dall'elevata eterogeneità dei profili di domanda: infatti, articoli a domanda regolare e volumi elevati reagiscono in modo differente alle politiche di gestione dell'inventario rispetto ai ricambi a bassa rotazione o a domanda intermittente, che richiedono approcci specificamente progettati per rappresentare fenomeni discreti e irregolari. Ne consegue che l'ottimizzazione delle scorte non può prescindere da una comprensione approfondita delle caratteristiche della domanda, derivante da un'analisi sistematica e strutturata, e dall'adozione di modelli di previsione coerenti con tali specificità e caratteristiche statistiche, al fine di ridurre il trade-off tra costi e livello di servizio e migliorare l'affidabilità complessiva della filiera distributiva.

1.2 Caratterizzazione della domanda e rappresentazione tramite serie temporali

Per poter affrontare in modo strutturato il problema del forecasting, la domanda delle spare parts viene comunemente rappresentata attraverso serie temporali che descrivono l'evoluzione nel tempo delle quantità richieste o consumate, costituendo la base informativa su cui si fondano i modelli di previsione. Tuttavia, la natura intrinsecamente variabile di questi dati di domanda ne complicano l'analisi: in tali condizioni, le componenti classiche delle serie temporali — livello, trend e stagionalità — risultano spesso debolmente identificabili o statisticamente instabili, riducendo l'efficacia di approcci previsivi generalisti. In presenza di tale eterogeneità, un'analisi aggregata della domanda risulta poco informativa e potenzialmente fuorviante, rendendo necessaria una segmentazione preliminare degli articoli. Per questo motivo, la comprensione delle proprietà statistiche della domanda diviene un passaggio preliminare essenziale al forecasting. Un approccio consolidato consiste nella segmentazione degli articoli in cluster omogenei, costruiti sulla base di dimensioni quali la presenza di trend, la stagionalità, il grado di regolarità del flusso di domanda e la profondità della storicità disponibile. Tale clusterizzazione consente di riconoscere l'eterogeneità tipica della popolazione di spare parts e di superare la logica del modello unico, che tende a penalizzare le performance complessive del sistema. Ne consegue che la scelta del modello di previsione statistica deve essere adattata al profilo di domanda del cluster di appartenenza: modelli capaci di catturare dinamiche strutturate risultano adeguati per articoli con domanda regolare, mentre ricambi a bassa rotazione o a domanda intermittente richiedono

approcci specificamente progettati per rappresentare fenomeni discreti e irregolari. In questo senso, la clusterizzazione della domanda non assume un ruolo meramente descrittivo, ma diventa un elemento abilitante per la costruzione di un sistema di forecasting coerente, flessibile e allineato alle esigenze dell'inventory management. Questo approccio costituisce la base metodologica per la selezione e l'applicazione dei modelli di forecast analizzati nei capitoli successivi.

1.3 Inventory Management tool

Alla luce delle considerazioni precedenti, si rende necessario disporre di uno strumento in grado di tradurre il framework concettuale di forecasting e inventory management in un processo operativo strutturato. Per rispondere a tali esigenze, per l'azienda presa in esame è stato sviluppato un tool integrato di inventory management, concepito come piattaforma analitica avanzata a supporto dei processi di forecasting e ottimizzazione delle scorte di parti di ricambio. Il tool si inserisce in un percorso di evoluzione verso soluzioni di analytics strutturate e scalabili, con l'obiettivo di costruire un sistema di previsione e gestione dell'inventario in grado di rispondere in modo efficace alle specificità della domanda delle spare parts e di supportare decisioni complesse in contesti caratterizzati da elevata incertezza. Un requisito fondamentale alla base dello sviluppo del tool è la necessità di disporre di una soluzione agile e performante, capace di supportare un rapido ciclo di analisi e sperimentazione e di garantire tempi di risposta adeguati anche in presenza di elevata complessità computazionale. In questo senso, il sistema è pensato per abilitare l'esecuzione di simulazioni su larga scala, necessarie per valutare l'impatto di differenti ipotesi di forecast e di politiche di inventario attraverso algoritmi complessi, mantenendo al contempo un'elevata efficienza operativa. Dal punto di vista architetturale, il tool è concepito come soluzione integrata, in grado di valorizzare in modo coerente le diverse fonti informative disponibili e di interfacciarsi agevolmente con i sistemi a valle del processo decisionale. Questa integrazione consente di garantire coerenza tra le analisi di forecasting, le logiche di inventory optimization e le successive fasi di pianificazione ed esecuzione, favorendo una convergenza tra ambito informatico e ambito business nei processi di gestione della domanda e delle scorte. Inoltre, il sistema è progettato con una prospettiva globale e di lungo periodo, così da poter essere esteso e arricchito progressivamente con nuove funzionalità, mantenendo un'unica piattaforma di riferimento per le attività di pianificazione, quali, ad esempio, ulteriori moduli di distribuzione o riapprovvigionamento. Un elemento distintivo del tool è rappresentato dalla componente di simulazione, che consente all'utente di creare e analizzare differenti scenari what-if a supporto delle decisioni strategiche e operative. Attraverso il sistema è possibile definire nuove simulazioni in funzione di diverse tipologie di analisi, visualizzare in

modo strutturato i parametri utilizzati e analizzare i risultati mediante indicatori di performance e dashboard dedicate. Questo approccio permette di valutare in modo trasparente le conseguenze delle scelte di business, riducendo l'impatto del cambiamento organizzativo e favorendo un utilizzo consapevole degli strumenti analitici. All'interno di questo framework, la componente di demand forecasting assume un ruolo centrale, in quanto costituisce il punto di partenza per l'intero processo decisionale e influenza direttamente l'efficacia delle politiche di inventario.

1.4 Architettura del Sistema: Integrazione e Componenti Tecnologiche

La traduzione del framework operativo in una soluzione tangibile ha richiesto lo sviluppo di un'architettura complessa, capace di armonizzare diverse tecnologie per garantire l'efficacia del processo di *inventory management*. Sviluppare un sistema di questa portata non significa soltanto implementare modelli di calcolo, ma orchestrare una serie di componenti che spaziano dall'infrastruttura *cloud* alla gestione dei flussi informatici, fino alla definizione di interfacce applicative e logiche di sviluppo avanzate.

Una delle componenti fondamentali dell'intero ecosistema è rappresentata da Microsoft Azure, una piattaforma di **Cloud Computing** che offre un insieme vasto e integrato di servizi gestiti a supporto di applicazioni enterprise e soluzioni di big data analytics. A differenza di un approccio tradizionale **On-Premises** — dove l'infrastruttura fisica (server, storage e networking) risiede all'interno dei perimetri aziendali richiedendo ingenti investimenti iniziali e una manutenzione costante della componente hardware — il *cloud* di Azure permette di astrarre completamente la gestione dell'infrastruttura sottostante. Questo paradigma abilita una scalabilità dinamica e on-demand, permettendo al sistema di allocare risorse computazionali in base al carico di lavoro istantaneo, garantendo al contempo standard di sicurezza, disponibilità e resilienza di livello enterprise che sarebbero difficilmente replicabili su server locali.

All'interno di questo scenario, l'orchestrazione dei flussi di dati è affidata ad **Azure Data Factory (ADF)**, un servizio di data integration serverless progettato per gestire processi complessi di ETL (Extract–Transform–Load) ed ELT (Extract–Load–Transform) in contesti sia cloud-native sia ibridi. ADF si configura come il vero e proprio regista dei dati dell'intero sistema, fornendo una piattaforma end-to-end per i data engineer finalizzata alla trasformazione di grandi volumi di dati grezzi in informazioni strutturate e utilizzabili a fini decisionali. Il ruolo principale di Azure Data Factory è la data ingestion, ovvero l'acquisizione coordinata di dati provenienti da fonti eterogenee tramite l'uso di Linked Services, che definiscono le connessioni verso le risorse esterne, e di Datasets, che

descrivono la struttura logica dei dati in ingresso e in uscita. L'organizzazione dei processi avviene attraverso le Pipelines, che rappresentano unità logiche di lavoro composte da una o più Activities. Ogni activity implementa uno specifico passo di elaborazione, come la copia dei dati (data movement), la loro trasformazione (data transformation) o la gestione del flusso di controllo (control activities). Un aspetto cruciale gestito da Data Factory è la data preparation, spesso pianificata in finestre temporali notturne tramite Triggers, al fine di ottimizzare l'utilizzo delle risorse computazionali. Durante queste fasi, ADF automatizza processi di pulizia, filtraggio, arricchimento e trasformazione dei dati; nel complesso, questa architettura di orchestrazione garantisce che, all'inizio di ogni giornata operativa, i decisori aziendali abbiano a disposizione basi dati aggiornate, consistenti e prive di ridondanze, pronte per alimentare modelli analitici avanzati e attività di forecasting affidabili e riproducibili.

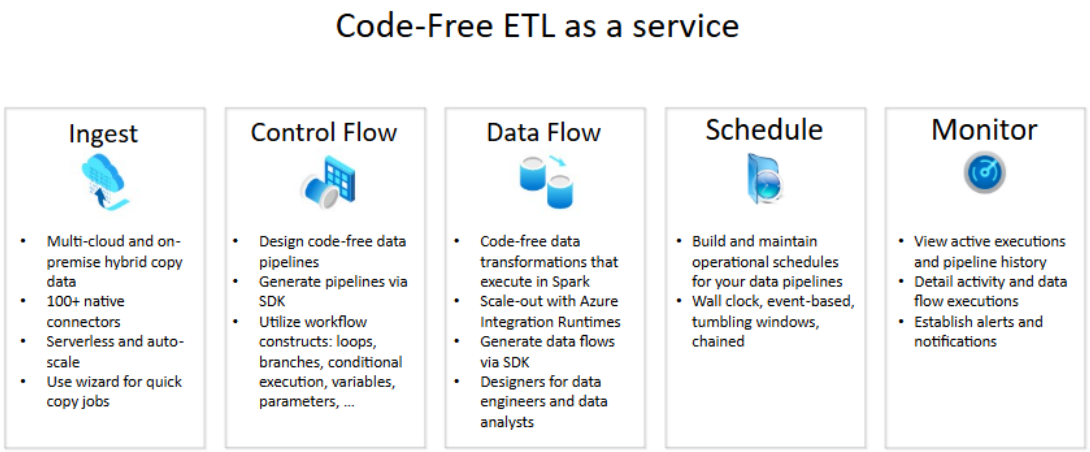

Figura 1.1: Struttura ETL di Azure Data Factory

Per quanto riguarda la persistenza delle informazioni, il sistema si appoggia a **PostgreSQL**, un database relazionale di tipo oggetto-relazionale noto per la sua robustezza e flessibilità. PostgreSQL agisce come il custode dei dati strutturati, gestendo con estrema precisione le tabelle relative all'inventario, alle anagrafiche dei ricambi e ai risultati delle simulazioni. La sua capacità di garantire l'integrità dei dati tramite transazioni sicure lo rende il componente ideale per supportare interrogazioni complesse e per archiviare le configurazioni degli scenari what-if generati dagli utenti.

L'intelligenza applicativa e la comunicazione tra i moduli sono invece assicurate da una struttura a più livelli:

- **Sviluppo Logico in Python:** Il linguaggio **Python** costituisce il motore computazionale del tool e rappresenta il livello in cui vengono implementate le componenti analitiche a maggiore intensità matematica. In questo strato risiedono gli algoritmi di calcolo statistico, le logiche di previsione della domanda, i modelli di ottimizzazione delle scorte e i moduli di simulazione what-if. Python è stato scelto non solo per la sua sintassi chiara e per la rapidità di sviluppo, ma anche per l'ampio ecosistema di librerie dedicate al data processing e al calcolo numerico, che consentono di gestire dataset ad alta dimensionalità e di implementare in modo efficiente procedure computazionalmente complesse. Tale flessibilità permette di tradurre modelli teorici in procedure operative strutturate e automatizzate, garantendo al contempo replicabilità dei risultati, controllo delle performance e possibilità di estensione futura verso algoritmi più avanzati.

- **Backend e API:** Lo strato di Backend rappresenta il livello logico–applicativo del sistema, responsabile del coordinamento e del governo del funzionamento complessivo del tool. Esso agisce come punto centrale di orchestrazione, occupandosi di ricevere le richieste provenienti dal Frontend, interpretarle secondo le regole di business e instradarle verso i componenti appropriati, come il database PostgreSQL o i moduli di calcolo sviluppati in Python. In questo senso, il Backend svolge un ruolo di mediazione tra l'interazione dell'utente e i processi interni del sistema, assicurando coerenza, controllo e affidabilità delle operazioni eseguite. Le funzionalità del sistema sono esposte tramite **API (Application Programming Interface)**, che costituiscono interfacce standardizzate di comunicazione tra componenti software differenti. Le API permettono di definire in modo chiaro quali operazioni possono essere richieste e come tali richieste devono essere formulate, senza esporre i dettagli interni dell'implementazione. Attraverso le API, le azioni dell'utente — come l'avvio di una simulazione, la modifica di parametri di inventario o la consultazione dei risultati — vengono tradotte in operazioni strutturate, elaborate dal Backend e restituite sotto forma di output coerente e interpretabile. Questo approccio consente di disaccoppiare l'interfaccia utente dalla logica applicativa, rendendo il sistema più flessibile e facilmente estendibile. In particolare, l'uso delle API favorisce l'integrazione con altri servizi o strumenti esterni e permette di trattare le simulazioni come servizi applicativi autonomi, invocabili in modo controllato e riproducibile. Nel complesso, tale architettura contribuisce a migliorare modularità, scalabilità e manutenibilità del sistema nel tempo, riducendo l'impatto delle modifiche e facilitando l'evoluzione futura della piattaforma.

- **Frontend:** Il Frontend è l'interfaccia utente finale, il punto di contatto dove la complessità tecnologica sottostante viene semplificata in visualizzazioni

grafiche. Attraverso componenti visuali, il frontend permette di esplorare i risultati delle simulazioni e di confrontare diversi scenari operativi, traducendo l'output dei modelli matematici in informazioni fruibili per il business.

L'integrazione coordinata di queste componenti tecnologiche definisce un'architettura multilivello coerente, in cui l'infrastruttura cloud garantisce scalabilità e affidabilità, i servizi di data integration assicurano la qualità e l'aggiornamento continuo delle basi informative, il database relazionale presidia la consistenza dei dati e lo strato applicativo – costituito da Python, Backend e API – governa la logica analitica e decisionale del sistema.

Nel loro insieme, tali elementi concorrono alla costruzione di una piattaforma solida, estendibile e orientata alla performance, in grado di supportare in modo strutturato i processi di forecasting e di inventory optimization in contesti caratterizzati da elevata complessità operativa e incertezza decisionale.

Capitolo 2

Modelli per Serie Temporali

2.1 Modelli analizzati

Dal momento che la modellazione corretta della domanda rappresenta un passaggio cruciale nel processo di previsione per le serie storiche, è necessario conoscere in che modo questa è strutturata e da quale componenti è rappresentata. In questo capitolo vengono pertanto introdotti e discussi i modelli di previsione utilizzati nel presente lavoro, selezionati in funzione delle caratteristiche statistiche delle serie temporali analizzate.

La trattazione che segue ha l'obiettivo di fornire il quadro teorico necessario per interpretare le scelte metodologiche del capitolo applicativo. Tale analisi si fonda sui contributi classici e consolidati della letteratura sul forecasting, con particolare riferimento al quadro metodologico generale proposto da Hyndman e Athanasopoulos [1] per la modellazione delle serie temporali e ai lavori pionieristici di Croston [2] sulla domanda intermittente e ai successivi sviluppi di Syntetos e Boylan [3].

2.1.1 Decomposizione STL e MSTL

La decomposizione di una serie temporale rappresenta uno strumento fondamentale nell'analisi esplorativa dei dati e nella modellazione delle dinamiche sottostanti, in quanto consente di separare e interpretare i diversi pattern che caratterizzano l'evoluzione della serie nel tempo. In questo contesto si inseriscono i metodi STL (Seasonal and Trend decomposition using Loess) e la sua estensione MSTL, particolarmente adatti alla gestione di strutture stagionali anche complesse. I dati delle serie temporali possono esibire una varietà di pattern, ed è spesso utile suddividere una serie temporale in più componenti, ciascuna rappresentativa di una specifica categoria di pattern sottostante. Una serie temporale può essere vista come composta da tre elementi: una componente di trend, una componente

stagionale e una componente di resto (che contiene tutto ciò che non rientra nelle altre componenti). Per alcune serie temporali, ad esempio quelle osservate con frequenza almeno giornaliera, può essere presente più di una componente stagionale, corrispondente ai diversi periodi di stagionalità.

Assumendo una decomposizione additiva, si può scrivere

$$y_t = S_t + T_t + R_t, \tag{2.1}$$

dove y_t rappresenta il dato osservato, S_t la componente stagionale, T_t la componente di trend e R_t la componente di residuale, tutte riferite al periodo t. In alternativa, una decomposizione moltiplicativa può essere espressa come

$$y_t = S_t \times T_t \times R_t.$$

La decomposizione additiva risulta appropriata quando l'ampiezza delle fluttuazioni stagionali, così come la variabilità attorno al trend, rimane approssimativamente costante al variare del livello della serie temporale, inteso come il livello medio della serie al netto delle componenti stagionali e irregolari. Quando invece la variabilità del pattern stagionale o la dispersione attorno al trend appaiono proporzionali al livello della serie, una decomposizione di tipo moltiplicativo risulta più adeguata.

La decomposizione delle serie storiche può essere utilizzata per misurare quanto le componenti del trend e della stagionalità influiscono su di esse, quindi, una volta specificata la forma della decomposizione, è possibile isolare le singole componenti della serie. Rimuovendo la componente stagionale dai dati originali si ottengono i dati "destagionalizzati". Nel caso additivo, i dati destagionalizzati sono dati da $y_t - S_T$, mentre nel caso moltiplicativo i valori destagionalizzati si ottengono come y_t/S_t; le serie destagionalizzate contengono sia la componente di resto sia la componente di trend. Di conseguenza, l'analisi della variabilità dei dati destagionalizzati consente di valutare la rilevanza della componente di trend rispetto alla componente di resto.

Una decomposizione di una serie temporale può essere utilizzata per misurare l'intensità del trend e della stagionalità in una serie. Per dati con un forte trend, i dati destagionalizzati dovrebbero presentare una variabilità significativamente maggiore rispetto alla componente di resto, poiché la variabilità aggiuntiva è attribuibile principalmente alla presenza della componente di trend. Di conseguenza, il rapporto $\mathrm{Var}(R_t)/\mathrm{Var}(R_t + T_t)$ dovrebbe risultare relativamente piccolo. Al contrario, per serie con poco o nessun trend, le due varianze dovrebbero essere approssimativamente uguali. Definiamo quindi la forza del trend come:

$$\mathrm{F}_T = \max\left(0{,}1 - \frac{\mathrm{Var}(R_t)}{\mathrm{Var}(R_t + T_t)}\right)$$

Questa quantità fornisce una misura della forza del trend compresa tra 0 e 1. Poiché la varianza della componente di resto può occasionalmente risultare maggiore della

varianza dei dati destagionalizzati, si impone che il valore minimo possibile di F_T sia pari a zero.

La forza della stagionalità è definita in modo analogo, ma con riferimento ai dati detrendizzati anziché a quelli destagionalizzati:

$$\mathrm{F}_S = \max\left(0{,}1 - \frac{\mathrm{Var}(R_t)}{\mathrm{Var}(R_t + S_t)}\right)$$

Una serie con una forza stagionale F_S prossima a 0 presenta una stagionalità quasi assente, mentre una serie con forte stagionalità avrà un valore di F_S vicino a 1, poiché $\mathrm{Var}(R_t)$ risulterà molto più piccola rispetto a $\mathrm{Var}(R_t + S_t)$.

Tuttavia, le definizioni precedenti assumono implicitamente la presenza di una singola componente stagionale, ipotesi che risulta limitante nel caso di serie temporali ad alta frequenza, nelle quali possono coesistere più cicli stagionali associati a differenti scale temporali.

A tal fine, introduciamo la decomposizione STL multipla (Multiple STL Decomposition, MSTL), un algoritmo di decomposizione additiva delle serie temporali progettato per gestire la presenza di più componenti stagionali. MSTL rappresenta un'estensione dell'algoritmo STL, in cui la procedura di decomposizione viene applicata in modo iterativo al fine di stimare separatamente le diverse componenti stagionali presenti nella serie.

Questo approccio consente di isolare in modo più efficace le variazioni stagionali associate a ciascun ciclo, migliorando la qualità della decomposizione rispetto all'applicazione di un singolo modello STL. Nel caso di serie temporali prive di stagionalità, MSTL individua esclusivamente le componenti di trend e di resto.

Analogamente a STL, MSTL fornisce una decomposizione additiva della serie temporale, permettendo di ottenere una rappresentazione strutturata dei diversi pattern che caratterizzano l'evoluzione dei dati nel tempo. Indicando con y_t l'osservazione al tempo t, MSTL estende l'equazione (2.1) per includere più pattern stagionali in una serie temporale nel modo seguente:

$$y_t = S_t^1 + S_t^2 + \cdots + S_t^n + T_t + R_t,$$

dove n rappresenta il numero di cicli stagionali presenti in y_t.

2.1.2 Exponential Smoothing

Lo smoothing esponenziale è stato proposto alla fine degli anni Cinquanta e ha dato origine ad alcuni dei metodi di previsione di maggior successo. Nel presente lavoro, i modelli di exponential smoothing vengono utilizzati come baseline per la previsione della domanda regolare, al fine di fornire un termine di confronto con gli approcci specificamente progettati per la domanda intermittente analizzati

nelle sezioni successive. Le previsioni prodotte mediante metodi di smoothing esponenziale sono medie ponderate delle osservazioni passate, in cui i pesi decadono esponenzialmente man mano che le osservazioni sono più remote nel tempo. In altre parole, quanto più recente è un'osservazione, tanto maggiore è il peso ad essa associato.

Il più semplice tra i metodi di smoothing esponenziale è comunemente definito smoothing esponenziale semplice (**Simple Exponential Smoothing, SES**). Questo metodo è adatto alla previsione di dati che non presentano un chiaro trend né un pattern stagionale. Le previsioni sono calcolate come medie ponderate, in cui i pesi diminuiscono esponenzialmente man mano che le osservazioni appartengono a un passato più remoto — i pesi più piccoli sono associati alle osservazioni più vecchie:

$$\hat{y}_{T+h|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \ldots, \tag{2.2}$$

dove $0 \leq \alpha \leq 1$ è il parametro di smoothing. La previsione per il tempo $T+1$ è una media ponderata di tutte le osservazioni della serie $y_1, \ldots, y_T$. La velocità con cui i pesi decrescono è controllata dal parametro α. Se α è piccolo (cioè vicino a 0), viene assegnato un peso maggiore alle osservazioni più lontane nel passato. Se α è grande (cioè vicino a 1), viene assegnato un peso maggiore alle osservazioni più recenti. Nel caso estremo in cui $\alpha = 1$, si ha $\hat{y}_{T+h|T} = y_T$.

L'equazione di previsione (2.2) può essere scritta anche nella **forma a componenti** nel seguente modo:

Equazione di previsione

$$\hat{y}_{t+h|t} = \ell_t$$

Equazione di smoothing

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1},$$

dove ℓ_t rappresenta il livello (o valore smussato) della serie al tempo t.

Ottimizzazione

L'applicazione di qualsiasi metodo di smoothing esponenziale richiede la scelta dei parametri di smoothing e dei valori iniziali. In particolare, per lo smoothing esponenziale semplice è necessario selezionare i valori di α e ℓ_0. Una volta noti questi valori, tutte le previsioni possono essere calcolate a partire dai dati. Nei metodi successivi, in genere, vi è più di un parametro di smoothing e più di una componente iniziale da determinare.

I parametri ignoti e i valori iniziali per qualsiasi metodo di smoothing esponenziale possono essere stimati minimizzando la somma dei quadrati degli errori (SSE). I residui sono definiti come

$$e_t = y_t - \hat{y}_{t|t-1}$$

per $t = 1, \ldots, T$. Pertanto, si ricercano i valori dei parametri ignoti e dei valori iniziali che minimizzano

$$SSE = \sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1})^2 = \sum_{t=1}^{T} e_t^2.$$

Metodo del trend lineare di Holt

Lo smoothing esponenziale semplice risulta inadeguato quando la serie presenta una dinamica di trend persistente. In questi casi è necessario introdurre una componente che catturi esplicitamente l'evoluzione del livello nel tempo. Holt (1957) ha esteso lo smoothing esponenziale semplice per consentire la previsione di dati con un trend. Questo metodo prevede un'equazione di previsione e due equazioni di smoothing (una per il livello e una per il trend):

Equazione di previsione

$$\hat{y}_{t+h|t} = \ell_t + hb_t$$

Equazione del livello

$$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$$

Equazione del trend

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$$

dove ℓ_t indica una stima del livello della serie al tempo t, b_t indica una stima del trend (pendenza) della serie al tempo t, α è il parametro di smoothing per il livello, $0 \leq \alpha \leq 1$, e β^* è il parametro di smoothing per il trend, $0 \leq \beta^* \leq 1$.

Come nello smoothing esponenziale semplice, l'equazione del livello mostra che ℓ_t è una media ponderata tra l'osservazione y_t e la previsione a un passo in avanti sul campione di addestramento per il tempo t, qui data da $\ell_{t-1} + b_{t-1}$. L'equazione del trend mostra che b_t è una media ponderata tra il trend stimato al tempo t basato su $\ell_t - \ell_{t-1}$ e b_{t-1}, la stima precedente del trend.

Metodi con trend smorzato

Le previsioni generate dal metodo lineare di Holt mostrano un trend costante (crescente o decrescente) che si estende indefinitamente nel futuro. Evidenze empiriche indicano che questi metodi tendono a sovrastimare le previsioni, soprattutto per orizzonti previsivi lunghi. Motivati da questa osservazione, Gardner & McKenzie (1985) hanno introdotto un parametro che "smorza" il trend verso una linea piatta dopo un certo periodo nel futuro.

In aggiunta ai parametri di smoothing α e β^* (con valori compresi tra 0 e 1 come nel metodo di Holt), questo metodo include anche un parametro di smorzamento $0 < \phi < 1$:

$$\begin{aligned}\hat{y}_{t+h|t} &= \ell_t + (\phi + \phi^2 + \cdots + \phi^h)b_t \\ \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}.\end{aligned}$$

Se $\phi = 1$, il metodo coincide con il metodo lineare di Holt. Per valori compresi tra 0 e 1, ϕ smorza il trend in modo tale che le previsioni di breve periodo presentino un trend, mentre quelle di lungo periodo siano costanti.

Metodi che utilizzano la stagionalità

Holt (1957) e Winters (1960) hanno esteso il metodo di Holt per catturare la stagionalità. Il metodo stagionale di Holt-Winters comprende un'equazione di previsione e tre equazioni di smoothing — una per il livello ℓ_t, una per il trend b_t e una per la componente stagionale s_t, con i corrispondenti parametri di smoothing α, β^* e γ. Si utilizza m per indicare il periodo della stagionalità, ossia il numero di stagioni in un anno. Ad esempio, per dati trimestrali $m = 4$, mentre per dati mensili $m = 12$.

Esistono due varianti di questo metodo che differiscono per la natura della componente stagionale. Il metodo additivo è preferibile quando le variazioni stagionali sono approssimativamente costanti nel corso della serie, mentre il metodo moltiplicativo è preferibile quando le variazioni stagionali cambiano in modo proporzionale al livello della serie.

Tabella 2.1: Combinazioni di componenti di trend e stagionalità

Componente di trend	**Componente stagionale**		
	N (Nessuna)	**A (Additiva)**	**M (Moltiplicativa)**
N (Nessuna)	(N,N)	(N,A)	(N,M)
A (Additiva)	(A,N)	(A,A)	(A,M)
A_d **(Additiva smorzata)**	(A_d,N)	(A_d,A)	(A_d,M)

Tabella 2.2: Metodi di smoothing esponenziale e notazione abbreviata

Notazione abbreviata	**Metodo**
(N,N)	Smoothing esponenziale semplice
(A,N)	Metodo lineare di Holt
(A_d,N)	Metodo additivo con trend smorzato
(A,A)	Metodo additivo di Holt-Winters
(A,M)	Metodo moltiplicativo di Holt-Winters
(A_d,M)	Metodo smorzato di Holt-Winters

Di seguito vengono fornite le formule ricorsive per applicare i nove metodi di smoothing esponenziale della tabella 2.1.

Trend	**Seasonal**		
	N	**A**	**M**
N	$\hat{y}_{t+h\|t} = \ell_t$ $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = \ell_t s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
A	$\hat{y}_{t+h\|t} = \ell_t + hb_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$
$\mathbf{A_d}$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$

Figura 2.1: Formule per il calcolo ricorsivo e le previsioni puntuali. Per ciascun caso, l_t indica il livello della serie al tempo t, b_t indica la pendenza al tempo t, s_t indica la componente stagionale della serie al tempo t, m indica il numero di stagioni in un anno; $\alpha, \beta^*, \gamma, \phi$ sono i parametri di smoothing, $\phi_h = \phi + \phi^2 + \cdots + \phi^h$ e k è la parte intera di $(h-1)/m$

In questo testo non vengono considerati i metodi con trend moltiplicativo, poiché tendono a produrre previsioni di scarsa qualità.

Modelli a spazio degli stati con innovazioni per lo smoothing esponenziale

Per ciascun metodo esistono due modelli: uno con errori additivi e uno con errori moltiplicativi. Le previsioni puntuali prodotte dai modelli sono identiche se utilizzano gli stessi valori dei parametri di smoothing. Tuttavia, essi generano intervalli di previsione differenti.

Per distinguere tra un modello con errori additivi e uno con errori moltiplicativi (e anche per distinguere i modelli dai metodi), estendiamo la classificazione usata finora. Indichiamo ciascun modello a spazio degli stati come ETS$(\cdot,\cdot,\cdot)$ per (Error, Trend, Seasonal). Le possibilità per ciascuna componente (o stato) sono: Error $= \{A, M\}$, Trend $= \{N, A, A_d\}$ e Seasonal $= \{N, A, M\}$. Viene riportato per scopo esemplificativo il seguente caso:

ETS(A,N,N): smoothing esponenziale semplice con errori additivi

Richiamiamo la forma a componenti dello smoothing esponenziale semplice:

Equazione di previsione

$$\hat{y}_{t+1|t} = \ell_t$$

Equazione di smoothing

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}.$$

Se riarrangiamo l'equazione di smoothing per il livello, otteniamo la forma di "correzione dell'errore",

$$\ell_t = \ell_{t-1} + \alpha(y_t - \ell_{t-1}) = \ell_{t-1} + \alpha e_t,$$

dove

$$e_t = y_t - \ell_{t-1} = y_t - \hat{y}_{t|t-1}$$

è il residuo al tempo t.

Possiamo anche scrivere

$$y_t = \ell_{t-1} + e_t,$$

così che ciascuna osservazione possa essere rappresentata come il livello precedente più un errore. Per trasformare questo in un modello a spazio degli stati con innovazioni, è sufficiente specificare la distribuzione di probabilità di e_t. Per un modello con errori additivi, assumiamo che i residui e_t siano rumore bianco normalmente distribuito con media 0 e varianza σ^2, scritto come

$$e_t = \varepsilon_t \sim \text{NID}(0, \sigma^2).$$

ADDITIVE ERROR MODELS

Trend	Seasonal		
	N	**A**	**M**
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal		
	N	**A**	**M**
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathbf{A_d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Figura 2.2: Equazioni a spazio degli stati per ciascuno dei modelli nel framework ETS.

Le equazioni del modello possono quindi essere scritte come

$$\begin{aligned} y_t &= \ell_{t-1} + \varepsilon_t \\ \ell_t &= \ell_{t-1} + \alpha\varepsilon_t. \end{aligned}$$

Stima dei modelli ETS

Un'alternativa alla stima dei parametri mediante la minimizzazione della somma dei quadrati degli errori consiste nel massimizzare la *verosimiglianza* (*likelihood*). La verosimiglianza è la probabilità che i dati osservati siano generati dal modello specificato. Di conseguenza, un'elevata verosimiglianza è associata a un buon

modello. Per un modello con errori additivi, la massimizzazione della verosimiglianza (assumendo errori normalmente distribuiti) fornisce gli stessi risultati della minimizzazione della somma dei quadrati degli errori. Tuttavia, per i modelli con errori moltiplicativi si ottengono risultati differenti.

I possibili valori che i parametri di smoothing possono assumere sono soggetti a restrizioni. Tradizionalmente, i parametri sono stati vincolati a giacere tra 0 e 1 in modo che le equazioni possano essere interpretate come medie pesate, cioè

$$0 < \alpha, \beta^*, \gamma^*, \phi < 1.$$

Per i modelli a spazio degli stati, abbiamo posto $\beta = \alpha\beta^*$ e $\gamma = (1 - \alpha)\gamma^*$. Pertanto, i vincoli tradizionali si traducono in

$$0 < \alpha < 1, \qquad 0 < \beta < \alpha, \qquad 0 < \gamma < 1 - \alpha.$$

In pratica, il parametro di smorzamento ϕ è solitamente vincolato ulteriormente per prevenire difficoltà numeriche nella stima del modello. Di solito esso è ristretto in modo che

$$0.8 < \phi < 0.98.$$

Selezione del modello

L'AIC e l'AIC$_c$ possono essere impiegati per determinare quale modello ETS sia il più appropriato per una data serie temporale.

Il criterio di informazione di Akaike (AIC) è definito come

$$\text{AIC} = -2\log(L) + 2k, \tag{2.3}$$

dove L è la verosimiglianza del modello e k è il numero totale di parametri e stati iniziali stimati (inclusa la varianza dei residui).

L'AIC corretto per la distorsione in piccoli campioni (AICc) è definito come

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{T-k-1}$$

I modelli con errori moltiplicativi sono utili quando i dati sono strettamente positivi, ma non sono numericamente stabili quando i dati contengono zeri o valori negativi.

2.1.3 ARIMA

I modelli ARIMA rappresentano, insieme ai modelli di smoothing esponenziale, uno dei due approcci più utilizzati per la previsione delle serie temporali, offrendo una prospettiva complementare sul problema. Mentre i modelli di smoothing

esponenziale descrivono esplicitamente trend e stagionalità, i modelli ARIMA si basano sulla struttura di autocorrelazione dei dati. Prima di introdurre i modelli ARIMA, è necessario discutere il concetto di stazionarietà e la tecnica della differenziazione delle serie temporali.

In generale, una serie temporale è stazionaria se le sue statistiche non dipendono dal tempo in cui la serie è osservata e si può concludere, pertanto, che le serie temporali con trend o con stagionalità non sono stazionarie, poiché il trend e la stagionalità influenzano il valore della serie in istanti diversi. Un modo per rendere stazionaria una serie temporale non stazionaria consiste nel calcolare le differenze tra osservazioni consecutive. Questa procedura è nota come differenziazione.

La differenziazione può aiutare a stabilizzare la media di una serie temporale rimuovendo le variazioni nel livello della serie e, di conseguenza, eliminando (o riducendo) il trend e la stagionalità.

La serie differenziata rappresenta la variazione tra osservazioni consecutive della serie originale e può essere scritta come

$$y'_t = y_t - y_{t-1}.$$

In alcuni casi, la prima differenziazione non è sufficiente a rendere la serie stazionaria e può essere necessario applicare una seconda differenza. In pratica, tuttavia, è raro dover ricorrere a differenziazioni di ordine superiore al secondo.

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}.$$

È possibile considerare anche una differenziazione stagionale, in cui si calcola la differenza tra un'osservazione e l'osservazione precedente appartenente alla stessa stagione. Pertanto,

$$y'_t = y_t - y_{t-m},$$

dove m indica il numero di stagioni. Queste sono anche chiamate *differenze a lag, o ritardo, m*, poiché si sottrae l'osservazione con un ritardo di m periodi.

In questo caso, le previsioni tendono a coincidere con l'ultima osservazione disponibile della stessa stagione.

Talvolta è necessario applicare sia una differenza stagionale sia una prima differenza per ottenere dati stazionari.

Per descrivere il processo di differenziazione, risulta preferibile utilizzare l'operatore di backshift B, uno strumento notazionale utile quando si lavora con i ritardi nelle serie temporali:

$$By_t = y_{t-1}.$$

Per cui, una prima differenza può essere riscritta come

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t.$$

Una volta ottenuta una serie temporale approssimativamente stazionaria tramite differenziazione, è possibile modellarne la dinamica attraverso componenti autoregressive e a media mobile, come avviene nei modelli ARIMA.

Modelli autoregressivi

In un modello autoregressivo, la variabile di interesse viene prevista utilizzando una combinazione lineare dei valori passati della variabile stessa. Il termine *autoregressione* indica che si tratta di una regressione della variabile su sé stessa.

Un modello autoregressivo di ordine p può quindi essere scritto come

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

dove ε_t è rumore bianco. Ci si riferisce a questo modello come AR(p), ovvero un modello autoregressivo di ordine p.

Modelli a media mobile

Anziché utilizzare i valori passati della variabile da prevedere in una regressione, un modello a media mobile utilizza gli errori di previsione passati in un modello di tipo regressivo:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

dove ε_t è rumore bianco. Ci si riferisce a questo modello come MA(q), ovvero un modello a media mobile di ordine q. Naturalmente, i valori di ε_t non sono osservabili, quindi non si tratta di una regressione nel senso tradizionale.

Si noti che ciascun valore di y_t può essere interpretato come una media mobile pesata dei più recenti errori di previsione (anche se i coefficienti non sommano generalmente a uno). La variazione dei parametri $\theta_1, \ldots, \theta_q$ produce differenti pattern nella serie temporale.

Modelli ARIMA non stagionali

Combinando la differenziazione con un modello autoregressivo e un modello a media mobile, si ottiene un modello ARIMA non stagionale. ARIMA è un acronimo di *AutoRegressive Integrated Moving Average* (in questo contesto, il termine "integrated" indica l'operazione inversa della differenziazione). Il modello completo può essere scritto come

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

dove y'_t rappresenta la serie differenziata (che può essere stata differenziata più di una volta). I "regressori" sul lato destro includono sia valori ritardati di y_t sia errori ritardati. Questo modello è indicato come ARIMA(p, d, q), dove

- p è l'ordine della componente autoregressiva;
- d è il grado di differenziazione applicato;
- q è l'ordine della componente a media mobile.

Quando si combinano più componenti per costruire modelli più complessi, risulta più agevole utilizzare la notazione di backshift. Ad esempio, il modello ARIMA sopra descritto può essere riscritto come

$$(1 - \phi_1 B - \cdots - \phi_p B^p)\,(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\,\varepsilon_t.$$

AIC e AICc

Una volta definita la struttura analitica del modello, il problema principale risiede nell'identificazione degli ordini ottimali. In questo contesto, per determinare l'ordine di un modello ARIMA risulta utile utilizzare il criterio di informazione di Akaike (Akaike's Information Criterion, AIC). Esso può essere scritto come

$$\text{AIC} = -2\log(L) + 2k = -2\log(L) + 2(p + q + l + 1),$$

dove L è la verosimiglianza dei dati, k rappresenta il numero di parametri del modello e $l = 1$ se $c \neq 0$ e $l = 0$ se $c = 0$.

Può essere utile utilizzare l'AICc (Corrected AIC), che per i modelli ARIMA può essere scritta come

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{T-k-1} = \text{AIC} + \frac{2(p+q+l+1)(p+q+l+2)}{T-p-q-l-2},$$

con T rappresenta il numero totale di osservazioni nella serie storica.

I modelli migliori sono quelli che minimizzano l'AIC o l'AICc. È importante notare che questi criteri di informazione non costituiscono una buona guida per la selezione dell'ordine di differenziazione (d) del modello, ma risultano utili solo per la scelta dei valori di p e q. Questo perché la differenziazione modifica i dati sui quali viene calcolata la verosimiglianza, rendendo i valori di AIC non confrontabili tra modelli con diversi ordini di differenziazione. Di conseguenza, è necessario adottare un approccio alternativo per scegliere d, dopo di che l'AICc può essere utilizzato per selezionare p e q.

Modelli ARIMA stagionali

Finora l'attenzione è stata limitata a dati non stagionali e a modelli ARIMA non stagionali. Tuttavia, i modelli ARIMA sono in grado di modellare anche un'ampia gamma di serie temporali stagionali.

Un modello ARIMA stagionale si ottiene includendo termini stagionali aggiuntivi nei modelli ARIMA introdotti in precedenza. Esso viene indicato come

$$\text{ARIMA}(p,d,q)(P,D,Q)_m,$$

dove la prima terna (p,d,q) rappresenta la parte non stagionale del modello, la seconda terna (P,D,Q) rappresenta la parte stagionale e m è il periodo stagionale (ad esempio, il numero di osservazioni per anno). La notazione maiuscola viene utilizzata per le componenti stagionali, mentre quella minuscola per le componenti non stagionali.

La parte stagionale del modello è costituita da termini analoghi alle componenti non stagionali, ma che coinvolgono operatori di backshift di ampiezza m. Ad esempio, un modello ARIMA$(1,1,1)(1,1,1)_4$ (senza termine costante) per dati trimestrali $(m=4)$ può essere scritto in notazione di backshift come

$$(1-\phi_1 B)(1-\Phi_1 B^4)(1-B)(1-B^4)y_t = (1+\theta_1 B)(1+\Theta_1 B^4)\varepsilon_t.$$

La selezione degli ordini del modello (p,q,P,Q) può essere determinata minimizzando il valore di AIC o AICc, come abbiamo visto per i modelli non stagionali (utilizzando in questo caso $k = p+q+P+Q+l+1$).

ARIMA vs ETS

Tabella 2.3: Relazioni di equivalenza tra modelli ETS e ARIMA

Modello ETS	**Modello ARIMA**	**Parametri**
ETS(A,N,N)	ARIMA(0,1,1)	$\theta_1 = \alpha - 1$
ETS(A,A,N)	ARIMA(0,2,2)	$\theta_1 = \alpha + \beta - 2$ $\theta_2 = 1 - \alpha$
ETS(A,A_d,N)	ARIMA(1,1,2)	$\phi_1 = \phi$ $\theta_1 = \alpha + \phi\beta - 1 - \phi$ $\theta_2 = (1-\alpha)\phi$
ETS(A,N,A)	ARIMA(0,1,m)(0,1,0)$_m$	
ETS(A,A,A)	ARIMA(0,1,m+1)(0,1,0)$_m$	
ETS(A,A_d,A)	ARIMA(1,0,m+1)(0,1,0)$_m$	

Si riportano le relazioni di equivalenza tra alcuni modelli ETS e ARIMA. Notiamo che AIC e AICc sono utili per la selezione tra modelli appartenenti alla stessa classe. Ad esempio, può essere utilizzata per scegliere un modello ARIMA tra diversi modelli ARIMA candidati o un modello ETS tra diversi modelli ETS candidati.

Tuttavia, non può essere utilizzata per confrontare modelli ETS e ARIMA tra loro, poiché appartengono a classi di modelli differenti e la verosimiglianza viene calcolata in modi diversi.

2.1.4 Croston e sue varianti

Dopo aver discusso i modelli più adatti a serie con struttura regolare, è ora necessario considerare approcci progettati specificamente per domande intermittenti. Infatti, gran parte dei metodi di previsione convenzionali assume che i dati appartengano a uno spazio campionario continuo; tuttavia, in molti contesti applicativi, i dati si presentano sotto forma di conteggi discreti, proprio come nel caso del numero di componenti di ricambio richiesti. Sebbene per volumi elevati la distinzione tra uno spazio campionario continuo e uno discreto sia trascurabile, quando i conteggi sono piccoli e caratterizzati da numerosi valori nulli — una condizione nota come domanda intermittente — è necessario adottare approcci specifici. In questo ambito, il riferimento principale è il metodo di Croston, introdotto da John Croston nel 1972 in [2]. Il metodo si distacca dallo smoothing esponenziale semplice (SES) tradizionale, il quale tende a produrre previsioni distorte, poiché interpreta erroneamente gli "zeri" come cali strutturali del livello medio piuttosto che come assenza temporanea di eventi.
La giustificazione teorica del metodo risiede nella scomposizione della domanda osservata y_t in un processo composto da due componenti indipendenti:

$$y_t = x_t \cdot z_t$$

dove:

- x_t **(Occorrenza):** è una variabile casuale binaria che segue una distribuzione di Bernoulli. In ogni intervallo, la domanda si verifica ($x_t = 1$) con probabilità $1/p$ e non si verifica ($x_t = 0$) con probabilità $1 - 1/p$.
- z_t **(Ampiezza):** rappresenta la dimensione della domanda quando essa è positiva ($y_t > 0$), con media μ e varianza σ^2.

Sotto queste ipotesi, l'intervallo di tempo tra due domande consecutive segue una distribuzione geometrica con media p. Il valore atteso della domanda in ogni istante t è dunque definito dal rapporto tra i momenti del processo: $E(y_t) = \mu/p$.
Per ottenere una stima robusta del tasso di domanda, Croston ha proposto di costruire due nuove serie derivate dalla serie storica originale: la serie delle quantità non nulle z_t (spesso chiamata demand) e la serie degli intervalli di tempo p_t trascorsi tra di esse (inter-arrival time).Il metodo applica due procedure di smoothing esponenziale separate che vengono aggiornate solo quando si verifica una domanda non nulla ($y_t > 0$). Siano $\hat{z}_t$ e $\hat{p}_t$ le stime aggiornate, rispettivamente, dell'ampiezza

media e dell'intervallo medio; la procedura di aggiornamento al tempo t segue il seguente schema:

- **Se** $y_t > 0$**:**
$$\begin{cases} \hat{z}_t = (1-\alpha)\hat{z}_{t-1} + \alpha z_t \\ \hat{p}_t = (1-\alpha)\hat{p}_{t-1} + \alpha p_t \end{cases}$$
- **Se** $y_t = 0$ **(domanda assente):** Le stime strutturali non subiscono variazioni, garantendo che l'assenza di osservazioni non influenzi la stima del livello medio:
$$\begin{cases} \hat{z}_t = \hat{z}_{t-1} \\ \hat{p}_t = \hat{p}_{t-1} \end{cases}$$
In questa fase, il sistema incrementa esclusivamente il contatore temporale.

Si noti che le quantità z_t e p_t sono osservate solo negli istanti in cui $y_t > 0$; di conseguenza, le stime $\hat{z}_t$ e $\hat{p}_t$ vengono aggiornate esclusivamente in tali periodi e restano costanti negli intervalli senza domanda.

La previsione della domanda è data dal rapporto tra le due componenti:

$$\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}$$

Questo approccio garantisce che la previsione rifletta il reale tasso di domanda μ/p senza le distorsioni tipiche dei metodi che trattano la serie come continua.

Il limite di Croston: il bias di Syntetos e Boylan

Nonostante l'efficacia nel gestire la domanda intermittente, Syntetos e Boylan [3] hanno dimostrato che la previsione ottenuta con il metodo di Croston è affetta da bias positivo. Il problema risiede nella natura matematica del rapporto tra le due componenti: anche assumendo l'indipendenza statistica tra l'ampiezza delle domande z_t e l'intervallo tra di esse p_t, il valore atteso del loro rapporto non coincide con il rapporto dei loro valori attesi. In termini formali:

$$E\left[\frac{\hat{z}_t}{\hat{p}_t}\right] \neq \frac{E[\hat{z}_t]}{E[\hat{p}_t]}$$

Nello specifico, poiché $E[1/\hat{p}_t] > 1/E[\hat{p}_t]$ (per la disuguaglianza di Jensen), il metodo di Croston tende a sovrastimare sistematicamente il tasso di domanda medio. L'entità di questa distorsione dipende linearmente dal valore della costante di smoothing α: maggiore è α, più elevato sarà l'errore di sovrastima.
Per correggere tale distorsione, Syntetos e Boylan hanno introdotto una variante

nota come SBA (Syntetos-Boylan Approximation), proponendo un nuovo stimatore della domanda media:

$$\hat{y}_t^{SBA} = \left(1 - \frac{\alpha}{2}\right) \frac{\hat{z}_t}{\hat{p}_t}$$

Questa modifica mantiene invariata la procedura di aggiornamento ricorsivo di $\hat{z}_t$ e $\hat{p}_t$ descritta in precedenza, ma applica il coefficiente $(1 - \alpha/2)$ al momento della sintesi finale. In questo modo, il metodo SBA risulta approssimativamente corretto e tende a fornire prestazioni superiori rispetto al metodo di Croston originale, riducendo l'errore medio di previsione specialmente in presenza di elevata intermittenza.

Capitolo 3

Contesto applicativo e metodologia

Nell'ambito industriale attuale, all'interno del quale si colloca questo lavoro, le aziende hanno la necessità di dotarsi di sistemi avanzati a supporto della pianificazione, in grado di trasformare grandi quantità di dati in informazioni utili per il processo decisionale. In particolare, la gestione efficace della domanda e delle scorte richiede strumenti capaci non solo di elaborare dati storici, ma anche di supportare il decisore nella valutazione di scenari alternativi e nell'analisi delle conseguenze delle diverse scelte operative. In questo contesto, i sistemi di Demand Management e Forecasting assumono un ruolo centrale come componenti di un processo decisionale più ampio, che include anche la pianificazione delle scorte e l'ottimizzazione dei livelli di servizio. Il flusso implementato per la gestione e previsione dei dati storici di domanda a disposizione riflette l'operatività reale del tool sviluppato, in cui Demand Management e Forecast rappresentano fasi consecutive ma strettamente interconnesse. Le previsioni generate costituiranno un input diretto per i processi di Inventory Optimization, consentendo una visione unificata dell'impatto della domanda prevista lungo l'intera catena decisionale.

Nel presente capitolo verrà descritto lo scenario applicativo in cui si colloca il lavoro di tesi, il dataset utilizzato e la metodologia adottata per la gestione dei dati storici e la loro previsione. L'obiettivo è fornire una descrizione strutturata dell'intero processo, a partire dai dati grezzi fino alla selezione del modello di forecasting più appropriato per ciascun articolo.

3.1 Dataset e struttura dei dati

Il dataset di riferimento è costituito da serie storiche di domanda mensile associate a coppie *item–warehouse*, che rappresentano l'unità minima di analisi del sistema.

Nel contesto applicativo reale, il volume complessivo dei dati è pari a diversi milioni di record, a testimonianza dell'elevata complessità e varietà dei profili di domanda gestiti dal tool.

Ai fini dello sviluppo e della validazione della metodologia descritta in questa tesi, le analisi e i risultati sono stati ottenuti su un sottoinsieme rappresentativo del dataset complessivo, composto da circa 30 000 serie *item–warehouse*. Tale sottoinsieme è stato selezionato in modo da preservare l'eterogeneità dei comportamenti di domanda, includendo articoli con profili regolari, stagionali, con trend e con domanda intermittente.

Per ciascuna coppia articolo–magazzino sono disponibili fino a 60 osservazioni mensili di domanda storica; questo formato consente di operare su un orizzonte temporale sufficientemente esteso per l'identificazione di pattern strutturali della domanda.

Oltre alle serie storiche, il dataset include informazioni descrittive e parametri gestionali necessari per il processo di simulazione e previsione, tra cui l'identificativo dell'articolo, il codice del magazzino e la classe ABC. In particolare, la classe ABC viene utilizzata come proxy della movimentazione dell'articolo e consente di associare ogni *item* a una categoria di *mover* (FAST, MEDIUM, SLOW). Questa classificazione influenza sia le soglie utilizzate per l'identificazione di trend e stagionalità, sia le logiche di segmentazione applicate nelle fasi successive.

Il processo seguito per la tesi non parte direttamente dalla previsione, ma da una fase preliminare di pulizia e preparazione dei dati. La domanda storica viene infatti sottoposta a operazioni di *filtering* volte a individuare e gestire valori anomali o non rappresentativi, secondo un insieme di regole definite a livello applicativo e descritte nella sezione successiva. Il dataset così filtrato rappresenta l'input per le fasi di clusterizzazione degli articoli e di selezione dei modelli previsionali.

3.2 Valutazione del demand pattern e clusterizzazione della domanda

Nel processo di Demand Management, una fase preliminare fondamentale è l'analisi del comportamento della domanda nel tempo, con l'obiettivo di comprenderne la regolarità e la variabilità prima di procedere alle fasi di previsione. Questa analisi, definita come valutazione del *demand pattern*, consente di distinguere tra fluttuazioni fisiologiche della domanda e anomalie potenzialmente riconducibili a errori o eventi eccezionali.

Un primo aspetto analizzato è la *smoothness* della domanda, ovvero il grado di regolarità con cui essa si manifesta nel tempo. Tale classificazione viene introdotta principalmente per supportare le operazioni di filtering, permettendo di calibrare in modo differenziato il trattamento degli outlier. In particolare, per gli articoli

che presentano un numero sufficiente di osservazioni informative (almeno due mesi con domanda positiva e una profondità storica adeguata), la smoothness viene valutata attraverso due indicatori statistici di riferimento: l'*Average Demand Interval* (ADI), che misura la frequenza media degli eventi di domanda, calcolato come l'intervallo medio tra i mesi in cui l'articolo presenta una domanda diversa da zero, e il Coefficiente di Variazione al quadrato (CV^2), definito come il rapporto tra la varianza e il quadrato della media, e viene utilizzato per quantificare il grado di dispersione dei dati rispetto al loro valore medio.

Sulla base di tali indicatori e delle soglie definite nei parametri di simulazione, la domanda viene classificata come *smooth* o *non smooth*. Gli articoli caratterizzati da una domanda troppo rara o eccessivamente discontinua vengono invece esclusi da questa valutazione e assegnati a una classe per la quale non viene applicato un filtering statistico strutturato. Questa distinzione consente di evitare un trattamento eccessivamente aggressivo nei confronti di articoli per i quali l'irregolarità rappresenta un comportamento atteso.

Le informazioni ottenute dalla valutazione della smoothness, pur essendo introdotte per guidare il filtering della domanda, vengono successivamente riutilizzate anche nella fase di clusterizzazione, che è invece finalizzata al forecasting. In particolare, la clusterizzazione degli articoli ha lo scopo di segmentare la popolazione in gruppi omogenei, ai quali associare strategie previsionali differenti.

Oltre alla smoothness, la clusterizzazione tiene conto della presenza di componenti strutturali nella serie storica, quali trend e stagionalità. Tali componenti vengono valutate tramite indicatori di *trend strength* e *seasonal strength*, calcolati confrontando la variabilità spiegata dalle rispettive componenti rispetto alla variabilità residua. Anche in questo caso, l'identificazione di un trend o di una stagionalità significativa avviene in base al superamento di soglie parametriche definite a livello di simulazione e viene effettuata solo in presenza di una profondità storica sufficiente, al fine di garantire l'affidabilità delle valutazioni.

Combinando le informazioni relative a smoothness, trend, stagionalità e disponibilità di dati storici, ciascun articolo viene assegnato a uno specifico cluster di domanda, tra cui:

- **NNN**: in assenza di trend e stagionalità e nel caso di domanda non smooth;
- **NNS**: in assenza di trend e stagionalità e nel caso di domanda smooth;
- **NSZ**: in assenza di trend, ma presenza di stagionalità;
- **ZNZ**: in presenza di trend, ma assenza di stagionalità;
- **ZSZ**: in presenza sia di trend sia di stagionalità.

Questa clusterizzazione consente di trattare in modo differenziato articoli con comportamenti di domanda strutturalmente diversi e rappresenta un elemento

chiave per la selezione dei modelli previsionali, evitando l'adozione di un approccio uniforme che risulterebbe inefficace in presenza di una popolazione eterogenea come quella tipica delle spare parts.

3.3 Processo di Filtering

La fase di pulizia della domanda ha l'obiettivo di individuare e trattare valori anomali che potrebbero compromettere la qualità delle previsioni, senza alterare comportamenti che sono fisiologici per specifiche tipologie di articoli. Il filtering non viene applicato in modo uniforme, ma è modulato in funzione del comportamento della domanda, precedentemente classificato attraverso la valutazione del demand pattern.

Gli articoli caratterizzati da domanda *smooth*, ovvero frequente e con variazioni contenute nel tempo, sono sottoposti a criteri di filtering più restrittivi, in quanto eventuali picchi improvvisi sono più facilmente riconducibili a eventi anomali. Al contrario, per gli articoli con domanda *non smooth*, in cui la domanda si manifesta con una certa frequenza ma presenta un'elevata variabilità mese su mese, vengono adottate soglie più tolleranti, poiché ampie oscillazioni rappresentano un comportamento naturale. Gli articoli caratterizzati da una domanda troppo scarsa o instabile vengono invece assegnati a una classe di *no filtering*, nella quale non viene applicato alcun trattamento statistico strutturato: in questi casi, l'informazione disponibile non è considerata sufficiente per distinguere in modo affidabile tra valori normali e outlier.

Per gli articoli idonei al filtering, l'individuazione degli outlier avviene mediante un approccio basato sull'Interquartile Range (IQR), calcolato sui soli valori di domanda positivi. A partire dal primo e dal terzo quartile della distribuzione storica, viene definita una soglia massima accettabile di domanda, pari al terzo quartile incrementato di un termine proporzionale all'IQR. Il coefficiente moltiplicativo utilizzato in tale calcolo è parametrico e dipende dalla classificazione della domanda, risultando più contenuto per le domande smooth e più elevato per le domande non smooth.

Un valore di domanda viene segnalato come outlier solo qualora superi tale soglia e presenti contemporaneamente un rapporto anomalo tra quantità domandata e numero di righe d'ordine, in modo da evitare l'identificazione di falsi positivi dovuti a normali fluttuazioni operative. Inoltre, il sistema introduce ulteriori salvaguardie per preservare pattern ricorrenti: picchi che si ripresentano in modo sistematico nello stesso periodo dell'anno vengono interpretati come comportamento legittimo e non come anomalie.

Una volta identificati, gli outlier vengono corretti limitando il valore osservato a una soglia massima ritenuta realistica, garantendo in ogni caso che il valore corretto

non superi mai la domanda originale. Successivamente, la serie di domanda filtrata può essere sottoposta a una fase di smoothing, che combina ciascun valore mensile con il contesto temporale circostante tramite pesi parametrizzabili. Questo passaggio consente di ridurre discontinuità residue e transizioni troppo brusche, ottenendo una curva di domanda più coerente con il comportamento sottostante e più adatta alle successive fasi di forecasting.

3.4 Metodologia di previsione

La metodologia di previsione adottata si basa su un approccio che combina la clusterizzazione della domanda con la selezione automatica del modello previsionale più appropriato per ciascuna coppia *item–warehouse*. In questo contesto, la clusterizzazione non ha un ruolo puramente descrittivo, ma viene utilizzata per restringere l'insieme degli algoritmi candidati a quelli più coerenti con le caratteristiche statistiche della domanda osservata.

A ciascun cluster viene associato un insieme di famiglie modellistiche, comprendenti modelli per domanda intermittente, modelli di smoothing esponenziale, modelli ARIMA/SARIMA e approcci basati su decomposizione STL. In Tabella 3.1 è riportata la corrispondenza tra cluster di domanda e famiglie di algoritmi considerate.

Per ciascun articolo, l'insieme degli algoritmi associato al cluster di appartenenza viene utilizzato per addestrare più modelli in parallelo. La serie storica di domanda viene suddivisa in un insieme di training e uno di validazione, utilizzando una quota prefissata della finestra storica più recente come orizzonte di controllo. Ogni modello viene addestrato sul training set e utilizzato per generare una previsione sul periodo di validazione.

La selezione del modello migliore avviene tramite il confronto delle prestazioni previsionali misurate attraverso metriche di errore. In particolare, vengono calcolati il Mean Absolute Error (MAE) e il Mean Squared Error (MSE), definiti rispettivamente come:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} |y_t - \hat{y}_t| \tag{3.1}$$

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \tag{3.2}$$

dove y_t rappresenta la domanda osservata al tempo t, $\hat{y}_t$ la previsione fornita dal modello e T il numero di osservazioni nel periodo di validazione.

La metrica utilizzata per la selezione finale dipende dal cluster di appartenenza dell'articolo. Per articoli caratterizzati da domanda intermittente o fortemente irregolare (cluster NNN e NNS), viene privilegiato il MAE, in quanto più robusto

Tabella 3.1: Associazione tra cluster di domanda e famiglie di modelli previsionali

Cluster	Famiglie di modelli previsionali considerate
In tutti i cluster sono inclusi come baseline comune i modelli per domanda intermittente di tipo Croston (e relative varianti) e il modello di smoothing esponenziale semplice (SES).	
NNN	Modelli baseline
NNS	Modelli baseline e modelli ARIMA non stagionali, adatti alla rappresentazione di serie regolari in assenza di componenti strutturali
NSZ	Modelli baseline e modelli di smoothing esponenziale con componente stagionale, modelli SARIMA e approcci basati su decomposizione STL
ZNZ	Modelli baseline e modelli di smoothing esponenziale con componente di trend e modelli ARIMA non stagionali
ZSZ	Modelli baseline e modelli di smoothing esponenziale con componenti di trend e stagionalità, modelli SARIMA e approcci basati su decomposizione STL

rispetto a errori estremi. Per articoli con domanda più strutturata e regolare, viene invece adottato il MSE, che penalizza maggiormente deviazioni rilevanti e consente una migliore identificazione di modelli capaci di catturare la dinamica complessiva della serie.

Il modello che minimizza la metrica selezionata viene infine riaddestrato utilizzando l'intera serie storica disponibile e impiegato per generare la previsione finale su un orizzonte futuro di medio termine.

3.5 Sintesi del processo

Il processo complessivo può essere riassunto nei seguenti passi:

1. acquisizione e pre–processing dei dati di domanda;
2. calcolo delle metriche di trend, stagionalità e smoothness;
3. clusterizzazione degli articoli;

4. selezione automatica del modello di previsione;

5. generazione delle previsioni future.

L'approccio adottato è modulare e fortemente parametrizzato: ciascuna fase del processo può essere configurata tramite parametri applicativi, consentendo all'utente di scegliere se applicare o meno il filtering, di definire le soglie utilizzate per la classificazione della domanda e di impostare i parametri dei modelli previsionali. Questa flessibilità rende il tool adatto non solo alla produzione del forecast operativo, ma anche all'esecuzione di simulazioni di tipo *what–if*, permettendo di valutare l'impatto di diverse configurazioni sulla qualità delle previsioni e sui processi decisionali a valle.

Capitolo 4

Risultati sperimentali e analisi comparativa

4.1 Obiettivi dell'analisi

Il presente capitolo è dedicato alla presentazione e all'analisi dei risultati sperimentali ottenuti dall'applicazione della metodologia di forecasting descritta nei capitoli precedenti. L'obiettivo principale del lavoro non è limitato alla valutazione dell'accuratezza dei singoli modelli previsionali, ma riguarda in modo più ampio la verifica dell'efficacia dell'intero framework di previsione implementato all'interno del tool di inventory management, con particolare riferimento al ruolo della clusterizzazione della domanda.

In particolare, il lavoro di tesi si propone di rispondere alle seguenti domande di ricerca:

- l'introduzione di una fase di clusterizzazione degli articoli consente di ottenere risultati previsionali comparabili o migliori rispetto a un approccio che non prevede alcuna segmentazione della domanda?
- in che misura la clusterizzazione influenza la selezione dell'algoritmo di forecast, riducendo o meno la necessità di ricorrere a modelli non coerenti con il profilo di domanda dell'articolo?
- il beneficio computazionale derivante dalla riduzione dello spazio di ricerca degli algoritmi giustifica l'introduzione della clusterizzazione, in termini di tempi di calcolo e risorse impiegate?

Per rispondere a tali quesiti, è stata condotta un'analisi comparativa basata sull'esecuzione di simulazioni alternative, distinguendo tra:

- simulazioni in cui la selezione del modello di forecast avviene all'interno di un insieme di algoritmi predefinito sulla base del cluster di domanda;
- simulazioni in cui la clusterizzazione non viene utilizzata e tutti gli algoritmi disponibili vengono valutati indistintamente per ciascun articolo.

Il confronto tra questi due approcci consente di valutare se la clusterizzazione rappresenti esclusivamente uno strumento di efficientamento computazionale o se, al contrario, introduca un vincolo che incide in modo significativo sulla qualità delle previsioni e sulle decisioni a valle del processo di inventory management.

4.2 Impostazione delle simulazioni

I risultati delle simulazioni, presentati in questo capitolo, sono stati ottenuti utilizzando il tool di inventory management descritto nel Capitolo 3, mantenendo invariati il dataset di riferimento, i parametri e le metriche di valutazione delle performance previsionali. L'unica differenza tra gli scenari analizzati riguarda la modalità di selezione degli algoritmi di forecast.

Nel caso con clusterizzazione, per ciascun articolo l'insieme dei modelli candidati è limitato alle famiglie previsionali associate al cluster di appartenenza, come definito nella Sezione 3.4. Nel caso senza clusterizzazione, invece, tutti gli algoritmi disponibili vengono testati per ogni serie storica, indipendentemente dalle caratteristiche del demand pattern.

Questo secondo approccio, pur rappresentando una soluzione teoricamente più generale, potrebbe comportare un significativo incremento della complessità computazionale, rendendo necessario valutare se il miglioramento potenziale delle performance giustifichi l'aumento dei tempi di calcolo e delle risorse richieste.

4.3 Criteri di confronto

Il confronto tra le diverse configurazioni di simulazione viene effettuato lungo tre dimensioni principali:

- **performance previsionali**, misurate tramite le metriche di errore definite nel Capitolo 3;
- **coerenza nella selezione degli algoritmi**, analizzando la frequenza con cui il modello ottimale individuato in assenza di clusterizzazione appartiene a una famiglia diversa da quella prevista per il cluster di domanda;
- **efficienza computazionale**, valutata in termini di numero di modelli addestrati e tempi di esecuzione delle simulazioni.

In particolare, l'analisi della frequenza con cui l'algoritmo selezionato "esce" dal cluster di riferimento consente di quantificare il rischio introdotto dalla clusterizzazione, ovvero la possibilità che il vincolo sullo spazio degli algoritmi escluda modelli potenzialmente migliori. Allo stesso tempo, tale analisi permette di verificare se, nella maggior parte dei casi, la selezione effettuata con clusterizzazione coincida con quella ottenuta da un approccio completamente non vincolato.

4.4 Discussione dei risultati ed analisi comparativa

Impostazione metodologica del confronto

L'analisi comparativa tra l'approccio con clusterizzazione della domanda e quello senza clusterizzazione è stata condotta adottando, in entrambi i casi, la metrica Mean Squared Error (MSE) come unico criterio di selezione del modello ottimale. Tale scelta metodologica risponde all'esigenza di garantire piena omogeneità nel confronto: l'utilizzo della stessa funzione di errore consente infatti di isolare l'effetto della clusterizzazione, evitando che eventuali differenze nei risultati siano attribuibili a metriche di valutazione differenti o a criteri di selezione eterogenei.

È importante sottolineare che l'intera analisi è stata volutamente svolta su dati grezzi, non sottoposti né a filtering né a smoothing. Questa scelta rappresenta uno scenario conservativo, assimilabile a un "caso peggiore", ed è stata adottata con l'obiettivo di valutare il comportamento strutturale del framework in condizioni caratterizzate da maggiore rumore, presenza di outlier e discontinuità nella domanda. In un contesto simile, la rilevazione di trend e stagionalità risulta inevitabilmente più instabile e soggetta a variabilità, rendendo il confronto particolarmente sfidante. L'analisi su dati grezzi non è quindi finalizzata a massimizzare le performance assolute, bensì a stressare il sistema e a verificarne la robustezza in condizioni meno favorevoli rispetto a quelle operative.

ADI	CV^2	Trend strength slow	Trend strength medium	Trend strength fast	Seasonal strength
1.32	0.49	0.1	0.4	0.6	0.6

Tabella 4.1: Soglie per le metriche della serie temporale utilizzate per l'analisi: la domanda viene considerata smooth se ADI < 1.32 e $CV^2 < 0.49$; si attesta la presenza di trend se la forza del trend supera la soglia associata alla classe di movimentazione dell'item; si attesta la presenza di stagionalità se la forza della stagionalità supera la soglia decisa.

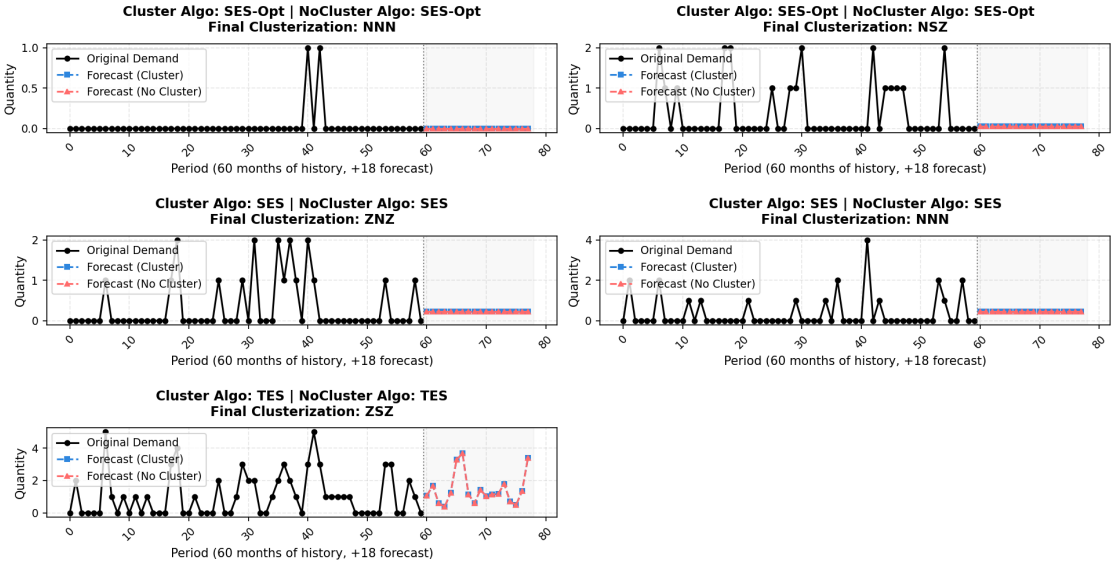

Figura 4.1: Confronto delle previsioni per lo stesso item su warehouse differenti: dall'analisi dei grafici si osserva che, nella maggior parte dei magazzini, l'algoritmo selezionato sia nel caso con cluster sia nel caso senza cluster è SES (simple exponential smoothing), coerentemente con la classificazione delle serie come prive di trend e/o stagionalità.
Nel caso del warehouse in cui la serie è classificata come ZSZ, indicando la presenza sia di trend sia di stagionalità, viene selezionato l'algoritmo TES, corrispondente allo smoothing esponenziale con trend e stagionalità, che genera una previsione più variabile e dinamica rispetto agli altri casi.

Coerenza tra approccio con e senza clusterizzazione

Nel complesso, i risultati mostrano un elevato grado di coerenza tra i due approcci. In oltre il 70% dei casi, il modello selezionato nello scenario con clusterizzazione coincide con quello individuato nello scenario senza cluster. Tale percentuale cresce ulteriormente per gli articoli caratterizzati da pattern di domanda strutturati e informativi, in particolare in presenza di stagionalità evidente, per i quali la corrispondenza tra i due approcci risulta pressoché totale. [4.1]

Questo risultato è particolarmente significativo: indica che, quando la struttura della serie è chiaramente identificabile, la restrizione dello spazio di ricerca agli algoritmi coerenti con il demand pattern non comporta una perdita di qualità previsionale. In altre parole, la clusterizzazione non agisce come un vincolo penalizzante, ma come un meccanismo di razionalizzazione che, nella maggior parte dei casi, conduce alla stessa scelta che verrebbe effettuata in uno scenario completamente non vincolato.

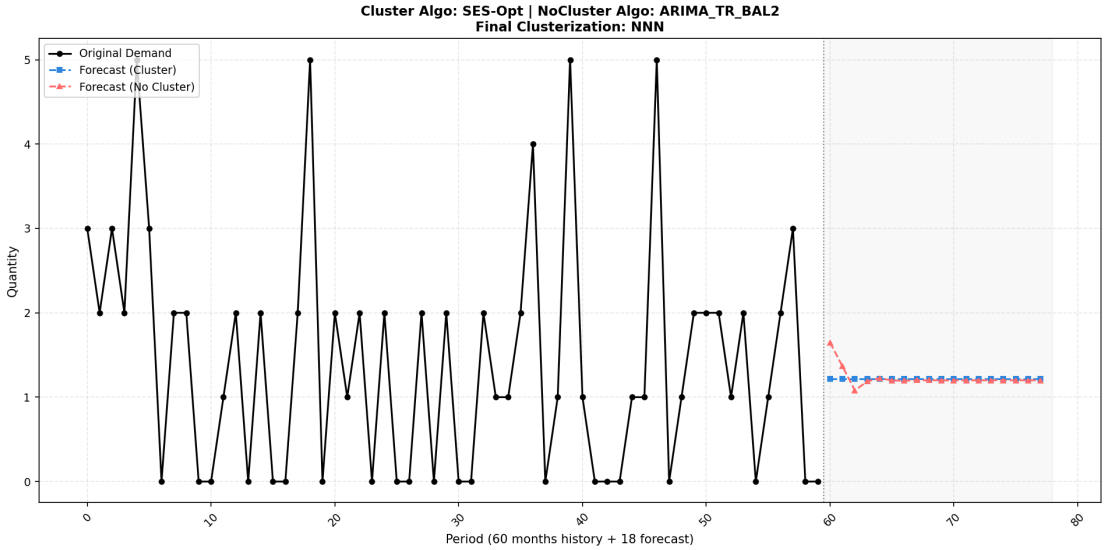

Figura 4.2: Serie storica di item classificato con NNN: dal modello con cluster viene scelto l'algoritmo di smoothing esponenziale semplice, metre nel caso in cui si analizzano tutti i cluster viene scelto un modello ARIMA. E' chiaro, però, che, nonostante la scelta ricada su modelli diversi, le previsioni siano pressocchè identiche, non giustificando la maggior complessità computazionale del valutare tutti gli algoritmi.

Analisi delle divergenze e ruolo delle soglie parametriche

Le differenze osservate nei casi residuali sono riconducibili principalmente alla maggiore flessibilità dell'approccio senza clusterizzazione. In assenza di vincoli, vengono talvolta selezionati modelli stagionali o modelli basati su decomposizione STL anche per serie che, secondo gli indicatori adottati, non superano le soglie parametriche di stagionalità o di trend. Anche se, spesso, in questi casi la scelta di modelli con stagionalità o trend non porta realmente un beneficio alla previsione. [4.2]

Questo comportamento mette in evidenza la sensibilità della clusterizzazione ai valori soglia utilizzati per la valutazione delle componenti strutturali. Soglie troppo restrittive possono escludere modelli in grado di catturare pattern ricorrenti deboli ma comunque rilevanti dal punto di vista predittivo. Al contrario, soglie troppo permissive rischierebbero di ampliare eccessivamente lo spazio di ricerca, riducendo il beneficio computazionale della segmentazione. [4.3]

In questo senso, la clusterizzazione non emerge come un limite concettuale del framework, bensì come un elemento la cui efficacia dipende in modo cruciale dalla qualità della fase di identificazione del demand pattern. La corretta misurazione di trend e stagionalità diventa quindi un nodo metodologico centrale.

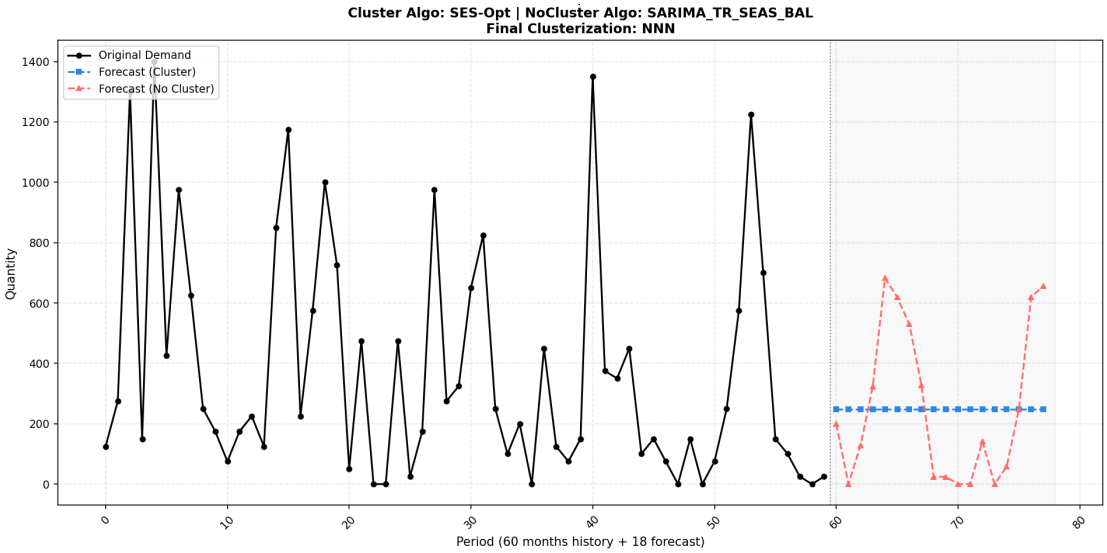

Figura 4.3: In questo caso di item classificato con NNN, dal modello con cluster viene scelto l'algoritmo di smoothing esponenziale semplice, metre nel caso in cui si analizzano tutti i cluster viene scelto un modello SARIMA. In questo caso, utilizzare trend e seasonal strength parametrici per valutare l'appartenenza al cluster risulta limitante.

Selezione di modelli stagionali in assenza di stagionalità "evidente"

Un risultato ricorrente riguarda la frequente selezione di modelli stagionali anche nei casi in cui la seasonal strength non segnala una stagionalità marcata. Questo fenomeno può essere interpretato sotto diverse prospettive.

In primo luogo, la misura di seasonal strength rappresenta un indicatore sintetico, che può risultare sensibile alla presenza di rumore o di outlier. Su dati grezzi, una stagionalità irregolare o non perfettamente stabile può essere sottostimata dall'indicatore, pur mantenendo una rilevanza predittiva.

In secondo luogo, l'introduzione di una componente stagionale in modelli come SARIMA o nelle decomposizioni STL aumenta la flessibilità della struttura dinamica. Anche in assenza di una stagionalità "classica" ben visibile, la componente stagionale può contribuire ad assorbire autocorrelazioni residue o dipendenze a distanza multipla di 12 mesi, migliorando la performance sul periodo di validazione.

Infine, occorre considerare che la selezione è guidata dal minimo MSE su un orizzonte finito. In presenza di rumore, un modello più complesso può adattarsi meglio alla specifica finestra di validazione, ottenendo un vantaggio locale che non necessariamente riflette una struttura sistematica della serie.

Complessità modellistica e similarità delle previsioni

Un ulteriore elemento emerso dall'analisi è che modelli formalmente differenti possono produrre previsioni molto simili. In presenza di trend o stagionalità deboli, l'introduzione esplicita di una componente strutturale aggiuntiva non sempre si traduce in un miglioramento significativo delle performance, ma può generare risultati sostanzialmente sovrapponibili in termini di MSE.

Questo suggerisce che, in numerosi casi, la maggiore complessità del modello non rappresenta un vantaggio sostanziale dal punto di vista predittivo. Ne consegue che limitare lo spazio di ricerca a un insieme coerente di modelli non comporta necessariamente una perdita di informazione, ma può anzi evitare l'adozione di strutture eccessivamente complesse rispetto al contenuto informativo effettivo della serie.

Possibili sviluppi metodologici

Alla luce di queste evidenze, una possibile evoluzione del framework riguarda la modalità di identificazione di trend e stagionalità. In particolare, la valutazione potrebbe essere affiancata o in parte sostituita da un approccio basato su regressione lineare multivariata, in cui una componente di trend deterministica e una o più componenti stagionali vengano stimate simultaneamente, assumendo ad esempio una struttura moltiplicativa della domanda

$$y_t = Trend_t \times Seasonal_t \times \varepsilon_t$$

dove y_t rappresenta la domanda osservata nel periodo t, $Trend_t$ la componente di trend deterministica, $Seasonal_t$ la componente stagionale e ε_t il termine di errore.

Assumendo una struttura moltiplicativa, il modello può essere linearizzato applicando il logaritmo alla variabile dipendente, ottenendo una formulazione di regressione lineare del tipo

$$\log(y_t) = \beta_0 + \beta_1 T_t + \sum_{m=1}^{12} \gamma_m S_{m,t} + \varepsilon_t$$

dove T_t rappresenta la variabile di trend (ad esempio il tempo o il numero di anni trascorsi) e $S_{m,t}$ sono variabili dummy che rappresentano le componenti stagionali mensili. I coefficienti β_1 e γ_m consentono rispettivamente di stimare l'intensità del trend e gli indici stagionali.

All'interno di questo framework possono essere considerati diversi modelli annidati, tra cui:

$$\log(y_t) = \beta_0 + \varepsilon_t$$

(modello costante),

$$\log(y_t) = \beta_0 + \beta_1 T_t + \varepsilon_t$$

(modello con sola componente di trend),

$$\log(y_t) = \sum_{m=1}^{12} \gamma_m S_{m,t} + \varepsilon_t$$

(modello con sola stagionalità) e

$$\log(y_t) = \beta_1 T_t + \sum_{m=1}^{12} \gamma_m S_{m,t} + \varepsilon_t$$

(modello con trend e stagionalità).

La stima dei parametri può essere effettuata tramite il metodo dei minimi quadrati ordinari (Ordinary Least Squares). Una volta stimati i diversi modelli candidati, la selezione del modello più appropriato può essere effettuata confrontando misure di errore predittivo, come il Mean Squared Error (MSE) scegliendo il modello che minimizza tale misura e che quindi fornisce la migliore approssimazione dei dati osservati.

Un simile approccio consentirebbe di testare la significatività statistica delle componenti strutturali in modo congiunto, riducendo la dipendenza da soglie parametriche fissate a priori. Questo migliorerebbe la gestione dei casi borderline e rafforzerebbe la coerenza tra cluster assegnato e modello effettivamente ottimale.

È tuttavia fondamentale evidenziare che molte delle discrepanze osservate sono amplificate dal fatto che l'analisi è stata condotta su dati non filtrati e non smoothed. In un contesto applicativo reale, il filtering elimina outlier non rappresentativi e lo smoothing riduce discontinuità spurie e rumore ad alta frequenza, rendendo più stabile la stima delle componenti strutturali. È quindi ragionevole attendersi che, operando su domanda pulita, le differenze tra approccio con e senza clusterizzazione si riducano ulteriormente.

Efficienza computazionale e scalabilità

Dal punto di vista aziendale e operativo, il risultato più rilevante riguarda l'efficienza computazionale del processo. La simulazione su circa 30.000 serie storiche ha evidenziato una differenza di tempo di esecuzione superiore a un fattore sette tra i due approcci. Nel caso senza clusterizzazione vengono valutati circa 30 algoritmi per ciascuna serie, mentre nello scenario con clusterizzazione il numero di modelli testati rimane significativamente inferiore, anche nei casi più completi.

A fronte di questa drastica riduzione dello spazio di ricerca, le performance previsionali risultano nella grande maggioranza dei casi sovrapponibili. Questo

risultato assume particolare rilevanza in un contesto industriale caratterizzato da milioni di record e da esigenze di simulazione what–if ripetute nel tempo: la clusterizzazione si configura quindi come un elemento essenziale per garantire scalabilità e sostenibilità computazionale del sistema.

Considerazioni conclusive

Per sintetizzare, anche in uno scenario conservativo basato su dati grezzi, la clusterizzazione non introduce una perdita significativa di qualità previsionale. Le differenze osservate sono riconducibili principalmente alla sensibilità delle soglie parametriche e alla maggiore flessibilità dell'approccio non vincolato in presenza di rumore.

Nel complesso, i risultati suggeriscono che la clusterizzazione rappresenti non solo uno strumento di efficientamento computazionale, ma anche un meccanismo di coerenza modellistica. In un contesto operativo reale, basato su domanda filtrata e smoothed, i benefici dell'approccio cluster-based risultano verosimilmente ancora più evidenti, configurandolo come un compromesso efficace tra accuratezza, robustezza ed efficienza computazionale.

Alla luce dell'analisi condotta sulle serie storiche di domanda, emerge inoltre la possibilità di rivedere in modo mirato l'associazione tra cluster di domanda e famiglie di modelli previsionali. In particolare, i risultati suggeriscono che alcune famiglie modellistiche, come i modelli ARIMA non stagionali, possano essere considerate come baseline trasversali anche per cluster in cui non sono attualmente incluse. In diversi casi, tali modelli risultano infatti competitivi in termini di performance, pur in assenza di componenti strutturali fortemente marcate.

Questo tipo di evidenza non deve essere interpretato come una criticità del framework, ma come un'indicazione utile per un suo affinamento progressivo. La struttura modulare del tool consente infatti di estendere o modificare l'insieme degli algoritmi associati a ciascun cluster in modo semplice e controllato, mantenendo invariata l'architettura complessiva del processo di forecasting.

In un contesto applicativo reale, tale flessibilità assume un ruolo centrale. La selezione degli algoritmi di previsione può essere guidata non solo da criteri puramente statistici, ma anche dalla conoscenza di dominio relativa ai prodotti, ai mercati e ai comportamenti di domanda attesi. Ad esempio, per specifiche famiglie di articoli o categorie merceologiche, l'esperienza di business può suggerire l'utilizzo sistematico di determinati modelli, indipendentemente dal cluster statistico di appartenenza.

La clusterizzazione si configura quindi come uno strumento di supporto alla decisione, e non come un meccanismo prescrittivo rigido: essa fornisce una segmentazione coerente della popolazione, all'interno della quale è possibile operare scelte modellistiche informate e adattabili. Questo approccio consente di integrare in

modo naturale analisi quantitativa e conoscenza esperta, rafforzando ulteriormente l'efficacia del framework in un contesto industriale reale.

Un ulteriore aspetto rilevante riguarda la natura fortemente configurabile del tool sviluppato. La scelta di introdurre una clusterizzazione della domanda non rappresenta infatti un vincolo rigido o una limitazione strutturale del framework, ma si inserisce all'interno di un impianto progettuale pensato per essere adattabile alle specifiche esigenze di business. Per ciascun cluster di domanda è possibile associare in modo flessibile le famiglie di algoritmi ritenute più appropriate, sulla base sia delle evidenze statistiche sia della conoscenza esperta del comportamento della domanda.

In questo senso, la clusterizzazione non impone una segmentazione "chiusa", ma costituisce un livello intermedio di astrazione che consente di tradurre la conoscenza di dominio in scelte modellistiche coerenti. Qualora il contesto applicativo suggerisca, ad esempio, l'utilizzo di modelli stagionali anche in presenza di stagionalità debole, o l'inclusione di specifiche famiglie algoritmiche per particolari categorie di articoli, il framework consente di integrare tali indicazioni attraverso una semplice configurazione dei parametri applicativi, senza modifiche strutturali al processo. Di conseguenza, la clusterizzazione non deve essere interpretata come una semplificazione eccessiva del problema, ma come uno strumento di governo della complessità. Essa permette di bilanciare in modo esplicito accuratezza previsionale, efficienza computazionale e coerenza con le logiche di business, rendendo il sistema non solo tecnicamente valido, ma anche utilizzabile e sostenibile in un contesto industriale reale.

Capitolo 5

Nuovi approcci per la domanda intermittente

Questa tesi nasce dal bisogno di rispondere ad uno dei bisogni fondamentali nella gestione della supply chain, ovvero quello di generare previsioni quanto più accurate possibile dai dati di domanda che si hanno a disposizione. Tale processo risulta particolarmente complesso nel caso di domanda intermittente, o irregolare, caratterizzata da una prevalenza di osservazioni nulle e da valori positivi sporadici e altamente variabili che, come abbiamo visto, rappresenta una componente molto impattante nell'intero sistema dei componenti di ricambio.

Come è noto, nella letteratura sulle serie temporali sono stati sviluppati metodi specifici per gestire tali peculiarità. Tra questi, il metodo di Croston e la sua variante Syntetos–Boylan Approximation (SBA) — analizzati nel capitolo 2 — rappresentano gli approcci di riferimento sia in ambito accademico, sia nella pratica industriale.

Accanto a questi standard, però, la letteratura propone metodi in grado di superare alcuni limiti strutturali di Croston e variante, così da modellare esplicitamente la struttura stocastica dell'intermittenza. In particolare, i modelli basati su catene di Markov permettono di descrivere il processo di alternanza tra periodi di stasi e periodi di attività, superando il limite teorico dell'ipotesi di indipendenza tra osservazioni successive.

Questo capitolo esamina due contributi particolarmente rilevanti in tale ambito. Il primo è il lavoro di Willemain, Smart e Schwarz [4], che combina un modello di Markov a due stati con una procedura di bootstrapping per prevedere la distribuzione della domanda su un lead time fisso. Il secondo, sviluppato da Uzunoğlu Köçer [5], introduce un modello di catena di Markov 'modificato' (MMC) che segmenta la domanda in diversi stati di intensità per migliorarne la precisione quantitativa.

Nel seguito verranno descritti i principi fondamentali di tali approcci, evidenziandone la struttura modellistica e i vantaggi competitivi rispetto ai metodi di previsione tradizionali.

5.1 Markov chain e Bootstrap

Il lavoro di Willemain, Smart e Schwarz [4] si pone l'obiettivo di prevedere la distribuzione cumulativa della domanda su un lead time fisso, utilizzando un nuovo tipo di bootstrap per serie temporali. Questo compito è reso difficile da un'alta proporzione di valori pari a zero nella storia della domanda.

E' bene ricordare che il bootstrap è una tecnica statistica di ricampionamento con reimmissione, il cui scopo è quello di fare inferenza statistica sul modello identificato senza effettuare ipotesi a priori, quando, in particolare, non si conosce la distribuzione della statistica di interesse. Nello specifico, le tecniche bootstrap permettono di stimare la distribuzione di un determinato parametro della popolazione sulla base dei dati osservati.

Nel contesto della gestione delle scorte, che è quello studiato dagli autori, è fondamentale lo studio della teoria delle quantità economiche d'ordine (EOQ), che determina due quantità per ogni articolo: un punto di riordino e una quantità d'ordine. Quando il livello di inventario raggiunge il punto di riordino viene effettuato un ordine pari alla quantità d'ordine prefissata. Il calcolo di tale quantità richiede generalmente soltanto una previsione della domanda media per periodo. Al contrario, la determinazione corretta del punto di riordino richiede una stima dell'intera distribuzione della domanda sull'intervallo di tempo, detto *lead time*, che intercorre tra l'emissione dell'ordine e il suo arrivo in magazzino. La letteratura tradizionale semplifica questo problema assumendo che la domanda nei singoli periodi sia indipendente e normalmente distribuita; tuttavia, nessuna di queste ipotesi risulta valida nel caso di domanda intermittente. Il metodo bootstrap proposto dagli autori non richiede invece alcuna di tali assunzioni. Anche nello scenario preso in esame dagli autori, i dati analizzati consistono in serie storiche di spedizioni mensili.

All'interno dell'articolo, viene confrontano il metodo proposto con alcuni approcci di previsione concorrenti, in particolare lo smoothing esponenziale e il metodo di Croston. Di seguito, per riportare l'approccio seguito dagli autori, sarà utilizzato X_t per riferirsi alla domanda osservata nel periodo t, con $t = 1, \ldots, T$; tale domanda è spesso nulla e, quando diversa da zero, presenta un'elevata variabilità. L invece rappresenterà il lead time fisso sul quale si desidera effettuare le previsioni. L'obiettivo è stimare l'intera distribuzione della somma delle domande sul lead time.

5.1.1 Metodo bootstrap proposto

Il metodo bootstrap modificato sviluppato dagli autori tiene conto di tre caratteristiche critiche della domanda intermittente: la presenza di autocorrelazione, la frequente ripetizione di valori e la ridotta lunghezza delle serie storiche. L'analisi dei dati industriali mostra infatti che la domanda tende spesso a presentarsi in sequenze, con serie più lunghe di valori zero o non zero rispetto a quanto atteso da un semplice processo di Bernoulli, evidenziando autocorrelazione positiva. In altri casi, si osserva un'alternanza più frequente tra valori zero e non zero, indicativa di autocorrelazione negativa. In entrambi gli scenari, risulta fondamentale che le previsioni di breve periodo siano in grado di catturare e sfruttare tali strutture di dipendenza temporale. Per modellare tale dipendenza temporale nella sequenza di periodi con e senza domanda, gli autori introducono un modello probabilistico discreto basato su un processo di Markov.

L'autocorrelazione viene infatti modellata mediante un processo di Markov a due stati di primo ordine, che descrive la sequenza di periodi con domanda nulla o non nulla.

$$\mathbb{P}(X_{t+1} = j \mid X_t = i, X_{t-1}, \ldots, X_0) = \mathbb{P}(X_{t+1} = j \mid X_t = i), \qquad i, j \in \{0{,}1\}.$$

La previsione della sequenza di zeri e non zeri sugli L periodi del lead time è condizionata allo stato dell'ultima osservazione. Le probabilità di transizione tra gli stati vengono stimate a partire dalla serie storica.

$$Q = \begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix}$$

Dove ogni elemento q_{ij} rappresenta la probabilità condizionata:

$$q_{ij} = P(X_{t+1} = j \mid X_t = i)$$

$$q_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$$

Una volta generata la sequenza prevista di valori zero e non zero sul lead time, è necessario assegnare valori numerici specifici ai periodi con domanda positiva. Un approccio diretto consisterebbe nel campionare dai valori storici non nulli; tuttavia, si fa notare, tale procedura impedirebbe la comparsa di nuovi valori in futuro, producendo campioni bootstrap poco realistici e stime scadenti delle code della distribuzione della domanda sul lead time LTD, soprattutto nel caso di serie storiche brevi. Questa limitazione è nota in letteratura come una delle principali debolezze del bootstrap tradizionale.

Per ovviare a tale problema, dopo aver selezionato casualmente un valore storico di domanda non nulla, si introduce una perturbazione casuale (*jittering*) al fine

di consentire la generazione di valori vicini. Il processo di jittering, infatti, è una tecnica statistica utilizzata principalmente nell'analisi dei dati, introducendo una piccola quantità di rumore casuale, per permettere una maggiore variabilità intorno a valori di domanda più elevati e migliorarene la chiarezza e l'interpretabilità. Nel caso studiato, dato X^* valore storico di domanda selezionato casualmente e Z una variabile casuale normale standard; il valore perturbato è ottenuto come:

$$X_{jittered} = \begin{cases} 1 + \lfloor X^* + Z\sqrt{X^*} \rfloor & \text{se } 1 + \lfloor X^* + Z\sqrt{X^*} \rfloor > 0 \\ X^* & \text{altrimenti} \end{cases}$$

La somma delle previsioni sui singoli periodi del lead time fornisce una realizzazione della LTD. Il processo viene ripetuto fino a ottenere 1000 repliche bootstrap, che consentono di stimare l'intera distribuzione della domanda sul lead time.

Il processo seguito da Willemain, Smart e Schwarz può essere schematizzato nel modo seguente:

1. Acquisizione di dati storici della domanda in intervalli temporali scelti.
2. Stima delle probabilità di transizione per un modello di Markov a due stati in cui la domanda viene ricondotta a una variabile di stato binaria che indica la presenza o l'assenza di domanda in ciascun periodo.
3. Condizionatamente all'ultima domanda osservata, si utilizza il modello di Markov per generare una sequenza di valori zero/non zero sull'orizzonte di previsione.
4. Sostituzione di ogni indicatore di stato non zero con un valore numerico campionato casualmente con reintroduzione dall'insieme dei valori osservati non zero.
5. Applicazione del jittering ai valori di domanda non zero.
6. Si sommano i valori previsti sull'orizzonte per ottenere un valore previsto di LTD.
7. I passi 3–6 vanno ripetuti molte volte, in modo da ottenere la distribuzione risultante dei valori di LTD.

Gli autori hanno valutato l'accuratezza dei diversi metodi di previsione applicandoli a oltre 28.000 articoli provenienti da nove aziende industriali. I risultati mostrano che il metodo di Croston, pur fornendo stime più accurate del livello medio di domanda nei periodi in cui questa si manifesta, non determina un miglioramento complessivo rispetto allo smoothing esponenziale quando l'obiettivo è la previsione dell'intera distribuzione della domanda sul lead time (LTD). Al contrario,

l'approccio bootstrap risulta nettamente superiore allo smoothing esponenziale, in particolare nel caso di lead time brevi.

Oltre al miglioramento dell'accuratezza previsiva, il metodo bootstrap presenta ulteriori vantaggi. In primo luogo, esso può essere facilmente esteso al caso di lead time variabili, semplicemente campionando da distribuzioni empiriche di lead time specifiche per ciascun articolo. In secondo luogo, il metodo bootstrap fa affidamento sulla storia empirica dettagliata di ciascun articolo di inventario piuttosto che su ipotesi matematiche semplificative. Inoltre, l'asimmetria delle distribuzioni ottenute mediante bootstrap è ritenuta più realistica rispetto a quella di una distribuzione normale convenzionale.

Nonostante l'approccio bootstrap alla previsione della domanda intermittente rappresenti una nuova e potente alternativa, gli autori riconoscono l'esistenza di alcune limitazioni e problematiche ancora irrisolte. Una prima criticità riguarda il processo di jittering, che in determinate circostanze può generare previsioni prive di significato fisico. Ad esempio, nel caso di prodotti venduti esclusivamente in confezioni di quantità fissa, la domanda sul lead time dovrebbe assumere soltanto valori compatibili con tali vincoli, mentre il jittering potrebbe produrre valori non ammissibili. In situazioni di questo tipo, può risultare opportuno omettere la fase di jittering. Una seconda limitazione riguarda il problema della non stazionarietà: come per tutti gli algoritmi bootstrap sviluppati fino ad oggi, il metodo assume che la serie storica sia stazionaria o possa essere facilmente trasformata in tale forma. La presenza di tendenze complesse o di stagionalità marcata costituisce pertanto una potenziale fonte di criticità.

5.2 Modified Markov chain model (MMCM)

Kocer riprende ed estende nel suo articolo [5] i risultati presentati da Willemain [4], sfruttando però un modello Markov chain di ordine superiore al primo, in modo da non prevedere semplicemente le sequenze di domanda nulla e non nulla, ma anche per prevedere i valori effettivi della domanda; in questo modo, un ordine superiore viene utilizzato per modellare la struttura dei pattern di domanda. Con questo approccio, ogni valore della domanda corrisponde ad uno stato della catena e, di conseguenza, la previsione dello stato successivo corrisponde alla previsione del valore successivo di domanda.

Quando si applica un modello di catena di Markov di ordine superiore a dati intermittenti, è necessaria una modifica per la procedura di ottenimento dei valori di previsione. La modifica proposta da Kocer può essere spiegata nel modo seguente: l'idea di base, come detto prima, è che gli stati siano le entità della domanda piuttosto che le categorie; tuttavia, in alcuni casi è necessario categorizzare le entità della domanda. Si ricordi, infatti, che la struttura della domanda intermittente

comporta valori di domanda come 0, 1 o 2 con frequenze elevate, mentre i valori di domanda più alti si riscontrano con frequenze più basse. Qui sono possibili due casi. Il primo caso riguarda entità di domanda basse con frequenze elevate. In questo caso, poiché ogni stato corrisponde a un valore di domanda, la stima dello stato successivo fornisce semplicemente la stima della successiva entità della domanda. Le entità di domanda più elevate con frequenze basse corrispondono al secondo caso. Frequenze basse o nulle influenzano la matrice delle probabilità di transizione e talvolta portano a trovare probabilità di transizione pari a zero. Ciò è indesiderabile perché influisce sul calcolo delle probabilità della distribuzione stazionaria della relativa catena di Markov. Pertanto, queste domande vengono raggruppate in un unico stato e quando si stima lo stato successivo, le occorrenze delle domande raggruppate nello stesso stato sono trattate con uguale probabilità.

5.2.1 Stima della domanda intermittente

In un modello di catena di Markov del primo ordine, la stima dello stato successivo si ottiene dipendendo solo dalla matrice delle probabilità di transizione a un passo, mentre per un modello di catena di Markov di ordine k è necessaria una matrice di transizione a k passi.

$$\hat{X}^{(n)} = \sum_{i=1}^{k} \lambda_i \hat{Q}_i X^{(n-i)}$$

$\hat{X}^{(n)}$ è il vettore di stato che rappresenta la previsione dello stato successivo al tempo n. $\hat{Q}_i$ è la matrice delle probabilità di transizione a i passi e λ_i sono i pesi reali non negativi tali che

$$\sum_{i=1}^{k} \lambda_i = 1$$

λ_i $(i = 1, 2, \ldots, k)$ può essere stimato tramite la stima di massima verosimiglianza o ottenuto risolvendo un modello di programmazione lineare (LP).

Le fasi dell'algoritmo proposto da Kocer sono qui descritte:

1. Si ottiene la frequenza di ogni valore di domanda per il set di dati j. Mentre alcuni valori di domanda sono osservati frequentemente, le frequenze di alcuni valori di domanda sono basse. Ogni valore di domanda con alta frequenza corrisponde a uno stato di una catena di Markov. Le domande con frequenze basse sono raccolte in un unico gruppo per essere riconsiderate nella procedura di previsione.

2. Vengono determinati gli stati della catena di Markov.

3. Vengono calcolate le matrici di probabilità di transizione a uno e due passi. Per m stati, le matrici generiche $\hat{Q}_1$ e $\hat{Q}_2$ sono definite come:

$$\hat{Q}_1 = \begin{pmatrix} q_{00}^{(1)} & q_{01}^{(1)} & \cdots & q_{0m}^{(1)} \\ q_{10}^{(1)} & q_{11}^{(1)} & \cdots & q_{1m}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m0}^{(1)} & q_{m1}^{(1)} & \cdots & q_{mm}^{(1)} \end{pmatrix}, \quad \hat{Q}_2 = \begin{pmatrix} q_{00}^{(2)} & q_{01}^{(2)} & \cdots & q_{0m}^{(2)} \\ q_{10}^{(2)} & q_{11}^{(2)} & \cdots & q_{1m}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m0}^{(2)} & q_{m1}^{(2)} & \cdots & q_{mm}^{(2)} \end{pmatrix}$$

Dove $q_{ij}^{(k)} = P(X_n = j | X_{n-k} = i)$.

4. Viene calcolata la distribuzione stazionaria, rappresentata dal vettore $\pi = [\pi_1, \pi_2, \ldots, \pi_m]^T$.

5. Viene costruito il modello di programmazione lineare (LP). Le probabilità della distribuzione stazionaria e le probabilità di transizione sono i parametri del modello LP e i λ_i sono le variabili decisionali. Il sistema LP risolto è:

$$\min_{\lambda} \quad \left\| \sum_{i=1}^{k} \lambda_i \hat{Q}_i \pi - \pi \right\|$$

Soggetto ai vincoli:

$$\sum_{i=1}^{k} \lambda_i = 1, \quad \lambda_i \geq 0$$

6. I valori λ_i vengono calcolati come soluzione del modello LP. Il numero di valori determina l'ordine del modello della catena di Markov.

7. Utilizzando i valori e le matrici delle probabilità di transizione, viene costruito il modello della catena di Markov per il set di dati j secondo la formula:

$$\hat{X}^{(n)} = \sum_{i=1}^{k} \lambda_i \hat{Q}_i X^{(n-i)}$$

8. In base al modello di catena di Markov risultante, si ottengono le previsioni della domanda. Secondo la distribuzione di probabilità dello stato di ordine k, la previsione dello stato successivo ($\hat{X}^{(n)}$) deve essere presa come lo stato con la massima probabilità; inoltre, se più di una domanda con bassa frequenza corrisponde ad uno stato, le domande per quello stato si presenteranno con probabilità eguali.

9. Viene calcolata la misura di accuratezza r per i valori previsionali.

10. I passaggi 8 e 9 sono ripetuti fino ad ottenere il valore di r più alto.

11. Le domande nel lead time LTD sono calcolate coerentemente con le informazioni sul tempo di consegna per il set di dati j; la somma delle previsioni di domanda mensili per i mesi di LT compongono la LTD.

12. Come visto anche nell'approccio precedente, i passaggi 8-11 vengono ripetuti molte volte per lo stesso set di dati in modo da ottenere la distribuzione LTD.

Nell'articolo di Kocer, per valutare le prestazioni e l'efficacia del modello di catena di Markov di ordine superiore, l'accuratezza della previsione r viene definita come

$$r = \frac{1}{T} \sum_{t=n+1}^{T} \delta_t$$

dove T è la lunghezza della sequenza di dati e

$$\delta_t = \begin{cases} 1, & \text{se } \hat{X}_{(t)} = X_{(t)} \\ 0, & \text{altrimenti} \end{cases}$$

Poiché il delta δ misura la coerenza tra i valori previsti e quelli reali, sono auspicabili valori di delta più elevati.

Nel paper, si conclude che il metodo MMCM risulta avere le migliori prestazioni in termini di accuratezza previsiva, rivelandosi superiore sia ai metodi classici per domanda intermittente (Croston e SBA) sia alla variante Markov Chain Bootstrap.

5.3 Prospettive e sviluppi futuri nel forecasting della domanda intermittente e non intermittente

L'analisi dei modelli Markoviani e bootstrap evidenzia come l'evoluzione del forecasting per la domanda intermittente stia progressivamente superando la logica deterministica e puntuale dei metodi tradizionali, orientandosi verso approcci probabilistici in grado di modellare in modo esplicito la struttura stocastica del processo generativo della domanda. In particolare, la rappresentazione dell'alternanza tra periodi di domanda nulla e periodi di attività mediante catene di Markov consente di incorporare la dipendenza temporale, superando una delle principali limitazioni strutturali dei metodi di Croston e delle sue varianti.

Nel contesto del framework sviluppato in questa tesi, tali approcci non rappresentano semplicemente un'alternativa modellistica, ma un possibile ampliamento coerente dell'architettura metodologica proposta. La clusterizzazione della domanda, infatti, costituisce un punto di ingresso naturale per l'introduzione di modelli Markoviani: gli articoli classificati nei cluster caratterizzati da elevata intermittenza (ad esempio NNN o NNS) potrebbero beneficiare di una modellazione esplicita delle probabilità di transizione tra stati, integrando la selezione automatica già implementata nel tool con una nuova famiglia di algoritmi dedicata.

Un ulteriore elemento di interesse riguarda l'allineamento tra forecasting e inventory optimization. I modelli bootstrap-Markov, in particolare, consentono di stimare l'intera distribuzione della domanda sul lead time, anziché fornire una previsione puntuale. Tale caratteristica risulta pienamente coerente con le esigenze dei modelli di gestione delle scorte, in cui la determinazione del punto di riordino richiede una stima affidabile delle code della distribuzione. L'integrazione di previsioni distribuzionali nel tool di inventory management potrebbe quindi migliorare la coerenza tra fase previsionale e decisioni di safety stock e EOQ.

Dal punto di vista applicativo, l'introduzione di modelli Markoviani comporta tuttavia alcune sfide rilevanti. In primo luogo, l'aumento della complessità computazionale richiede un'attenta valutazione in termini di scalabilità, soprattutto in presenza di milioni di serie item–warehouse. In secondo luogo, la stima robusta delle matrici di transizione necessita di una profondità storica adeguata, non sempre disponibile nel caso di articoli a bassissima rotazione. Tali criticità suggeriscono la possibilità di sviluppare approcci ibridi, nei quali i modelli Markoviani vengano attivati selettivamente solo per gli articoli che presentano determinate caratteristiche statistiche, preservando l'efficienza complessiva del sistema.

Sebbene i modelli presentati non siano stati implementati nell'ambito sperimentale di questo lavoro, essi rappresentano una direzione di ricerca promettente e coerente con l'evoluzione del sistema sviluppato. L'integrazione di approcci probabilistici strutturati all'interno del tool di inventory management potrebbe costituire un passo ulteriore verso la costruzione di una piattaforma decisionale capace non solo di prevedere la domanda, ma di modellarne esplicitamente l'incertezza, rafforzando l'efficienza della gestione delle spare parts.

In una prospettiva più ampia, l'evoluzione del time series forecasting non si limita al dominio della domanda intermittente, ma coinvolge l'intero spettro delle serie temporali aziendali. In tale contesto, un ruolo crescente è assunto dalle reti neurali artificiali, capaci di modellare relazioni non lineari complesse tra la variabile risposta e i suoi predittori. Le architetture feed-forward multistrato, costituite da nodi organizzati in livelli di input, strati nascosti e output, permettono di apprendere dai dati attraverso la stima iterativa dei pesi mediante algoritmi di ottimizzazione che minimizzano una funzione di costo (ad esempio l'errore quadratico medio), introducendo al contempo meccanismi di regolarizzazione volti a contenere l'ampiezza dei parametri e a ridurre il rischio di overfitting.

Particolarmente rilevante, nell'ambito delle serie storiche, è l'approccio di autoregressione a rete neurale (NNAR), nel quale i valori ritardati lag della serie vengono utilizzati come input del modello, in analogia con le strutture AR tradizionali ma senza imporre vincoli di linearità o di stazionarietà. L'estensione a componenti stagionali consente inoltre di incorporare ritardi multipli legati alla periodicità del fenomeno, ampliando la capacità del modello di catturare pattern ricorrenti

complessi.

Tali configurazioni risultano particolarmente promettenti in presenza di dinamiche non lineari, effetti soglia o interazioni tra trend e stagionalità difficilmente rappresentabili attraverso modelli lineari classici. In questa direzione, una frontiera avanzata è rappresentata dall'adattamento delle architetture 1D U-Net al dominio temporale. Originariamente concepite per la segmentazione d'immagine da Olaf Ronneberger et al., queste reti operano su serie storiche trattandole come segnali monodimensionali, sfruttando una struttura simmetrica a "U" composta da un percorso di contrazione (encoder) per l'estrazione del contesto macroscopico e uno di espansione (decoder) per la ricostruzione granulare del segnale. L'elemento distintivo risiede nelle skip connections, che permettono di preservare i dettagli ad alta frequenza dei dati storici più recenti integrandoli con le componenti di lungo periodo, rendendo il modello estremamente efficace sia nel denoising dei segnali sporchi che nella previsione multi-step di serie temporali ad alta dimensionalità.

Nel quadro evolutivo delineato in questa tesi, l'integrazione di reti neurali, tecniche ensemble e metodologie di simulazione rappresenta quindi un possibile sviluppo verso sistemi previsionali ibridi, nei quali modelli lineari, approcci probabilistici e architetture non lineari cooperano in modo complementare. Una tale impostazione consentirebbe di superare la contrapposizione tra metodi statistici tradizionali e algoritmi di machine learning, orientando il forecasting verso piattaforme modulari e adattive, capaci di selezionare o combinare automaticamente la struttura modellistica più adeguata in funzione delle caratteristiche empiriche della serie analizzata.

Bibliografia

[1] R. J. Hyndman e G. Athanasopoulos. *Forecasting: Principles and Practice*. 3rd. Melbourne, Australia: OTexts, 2021 (cit. a p. 8).

[2] J. D. Croston. «Forecasting and stock control for intermittent demands». In: *Operational Research Quarterly* 23.3 (1972), pp. 289–303 (cit. alle pp. 8, 22).

[3] A. A. Syntetos e J. E. Boylan. «The accuracy of intermittent demand estimates». In: *International Journal of Forecasting* 21.2 (2005), pp. 303–314 (cit. alle pp. 8, 23).

[4] Thomas R. Willemain, Charles N. Smart e Henry F. Schwarz. «A new approach to forecasting intermittent demand for service parts inventories». In: *International Journal of Forecasting* 20.3 (2004), pp. 375–387. DOI: 10.1016/S0169-2070(03)00013-X (cit. alle pp. 42, 43, 46).

[5] Umay Uzunoğlu Kocer. «Forecasting intermittent demand by Markov chain model». In: *International Journal of Innovative Computing, Information and Control* 9.8 (2013), pp. 3307–3318 (cit. alle pp. 42, 46).

[6] R. B. Cleveland, W. S. Cleveland, J. E. McRae e I. Terpenning. «STL: A Seasonal-Trend Decomposition Procedure Based on Loess». In: *Journal of Official Statistics* 6.1 (1990), pp. 3–73.

[7] K. Bandara, R. J. Hyndman e C. Bergmeir. «MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns». In: *International Journal of Forecasting* 37.2 (2021), pp. 488–503.