Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

COMBINAZIONI DI PREVISIONI PROBABILISTICHE PER LA DOMANDA INTERMITTENTE

Relatrice Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Correlatore Dott. Daniele Girolimetto
Dipartimento di Scienze Statistiche

Laureanda: Aurora Ciandri
Matricola N. 2027845

Anno accademico 2024/2025

Indice

Introduzione

La previsione della domanda intermittente è una sfida di primaria importanza in molti settori lavorativi, poiché i costi elevati associati, ad esempio, alla gestione dell'inventario ne rendono cruciale una stima accurata. Spesso è associata a dati di conteggio caratterizzati da movimenti molto irregolari con un'elevata presenza di valori pari a zero. Di conseguenza, prevedere con precisione tali fluttuazioni diventa essenziale per ottimizzare le decisioni aziendali legate alla gestione delle scorte, come la pianificazione dello stoccaggio e il rifornimento degli articoli. Una previsione accurata della domanda intermittente può infatti contribuire ad evitare costi superflui legati a scorte eccessive o insufficienti, migliorando l'efficienza complessiva dell'intera catena di approvvigionamento.

Nonostante la natura irregolare della domanda intermittente tenda ad avere un impatto limitato a livello di ricavi, una sua previsione adeguata può portare ad un significativo contenimento dei costi. Questa è una delle ragioni per cui l'importanza di tali dati è cresciuta negli ultimi anni: dal 2010, infatti, si è assistito allo spostamento dell'interesse aziendale dalla massimizzazione dei ricavi alla minimizzazione dei costi (Boylan & Syntetos, 2021). Tuttavia, la letteratura ha prestato limitata attenzione allo studio della domanda intermittente, a causa, almeno in parte, della sua complessità previsiva. Questa, infatti, è caratterizzata da una duplice fonte di incertezza, legata all'irregolarità dell'arrivo della domanda e alla possibile alta variabilità della sua dimensione, che rende i metodi previsivi tradizionali poco idonei (Nikolopoulos, 2021).

L'obiettivo di questa tesi è l'analisi e il confronto di diversi modelli e metodi di previsione probabilistica applicati alla domanda intermittente. L'adozione di un approccio probabilistico (Gneiting & Katzfuss, 2014), in contrasto con la tradizionale previsione puntuale, si rende necessaria per fornire informazioni più ricche a supporto delle decisioni aziendali. In particolare, si intendono confrontare i risultati ottenuti dall'applicazione di combinazioni di previsioni con quelli derivanti

dai singoli metodi che le compongono, sia in termini di qualità previsiva (tramite calibrazione e *sharpness*, Gneiting et al. (2007)), sia in relazione alle implicazioni pratiche sulla gestione dell'inventario. Questo approccio è ampiamente utilizzato per migliorare l'accuratezza delle previsioni, integrando informazioni provenienti da diverse fonti. Questo è particolarmente utile nell'ambito della domanda intermittente, che è spesso caratterizzata da un processo generatore dei dati complesso, con trend che varia nel tempo, cambiamenti stagionali e rotture strutturali. In tali situazioni, la combinazione di previsioni ottenute da modelli con diversi gradi di non corretta specificazione e adattabilità (Wang et al., 2023), può mitigare il problema. Seguendo l'approccio proposto da Wang et al. (2024), si prendono in considerazione combinazioni lineari di previsioni, utilizzando diverse tipologie di pesi associati ai singoli metodi univariati, come pesi semplici (ad esempio, la media aritmetica), pesi basati su funzioni di punteggio (ottimizzati, ad esempio, tramite il punteggio logaritmico), e pesi basati sui costi di inventario.

Questa tesi si suddivide in cinque capitoli. Il primo capitolo introduce il contesto della domanda intermittente e propone un metodo per la classificazione di tali serie storiche. Il secondo capitolo esplora i principali approcci e le criticità nella previsione della domanda intermittente, nonché vari metodi e modelli probabilistici, facendo riferimento principalmente a Wang et al. (2024) e Svetunkov & Boylan (2023). Il terzo capitolo esamina il metodo del *linear opinion pool*, focalizzandosi su tre differenti tipologie di pesi. Il quarto capitolo valuta la qualità delle previsioni probabilistiche in termini di calibrazione, *sharpness* e *performance* nell'ambito della gestione dell'inventario. Infine, nel quinto capitolo viene svolta un'analisi empirica basata sulle serie storiche tratte dalla "M5 Competition" (Makridakis et al., 2022).

Capitolo 1

Introduzione e contesto della domanda intermittente

La previsione della domanda di un certo bene o servizio guida alcune delle principali decisioni aziendali, come la pianificazione della produzione e degli ordini di rifornimento dell'inventario. Spesso la domanda si presenta regolarmente nel tempo e viene definita "non intermittente".

Ci sono dei casi, però, in cui possono essere presenti dei periodi, più o meno lunghi, caratterizzati da completa assenza di domanda. In queste situazioni la domanda si presenta in maniera sporadica nel tempo e viene definita "intermittente" (Boylan & Syntetos, 2021). Tale andamento è prevalente nell'ambito della gestione dell'inventario e influenza varie industrie come quella automobilistica, della vendita al dettaglio e aerospaziale (Wang et al., 2024). In particolare, Nikolopoulos (2021) afferma che la domanda intermittente è molto frequente nell'ambito delle parti di ricambio, nel quale caratterizza approssimativamente il 50% degli inventari. Un altro ambito particolarmente rilevante per lo studio della domanda intermittente è l'industria post-vendita[1] (Wang et al., 2024).

Nonostante i beni caratterizzati da domanda intermittente, noti anche come "*slow-moving*", non generino ricavi particolarmente alti, una loro gestione corretta può comportare un significativo contenimento dei costi. Questo spiega, almeno parzialmente, l'aumento di interesse sull'argomento da parte delle aziende in anni recenti. A partire dal 2010, infatti, si è assistito ad uno spostamento dell'interesse aziendale dalla massimizzazione del guadagno alla minimizzazione dei costi, au-

[1] L'applicabilità delle metodologie studiate per la domanda intermittente non si limitano all'ambito economico a cui è prevalentemente associata (vedi Appendice A.1).

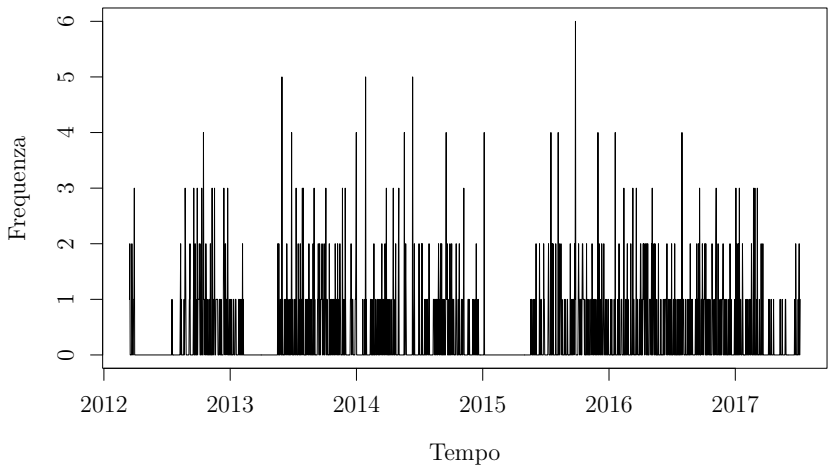

Figura 1.1: Esempio di una serie intermittente (dati della competizione M5).

mentando quindi la rilevanza della gestione dei prodotti a domanda intermittente (Boylan & Syntetos, 2021).

Un altro fattore che ha contribuito a sottolineare l'importanza della corretta gestione dei beni *slow moving* è la crescente attenzione verso le questioni ambientali. Poiché i movimenti di questi articoli sono intrinsecamente poco frequenti, avendo una richiesta sporadica nel tempo, essi sono particolarmente soggetti ad obsolescenza, problema ulteriormente aggravato dalla riduzione del ciclo di vita del prodotto che caratterizza l'industria moderna. Ne consegue che una migliore previsione e gestione di questi *item* può contribuire alla riduzione di rifiuti e scarti prodotti, oltre che dei costi (Boylan & Syntetos, 2021).

Tutti questi fattori, ambientali ed economici, hanno reso i beni intermittenti, e la loro previsione, una delle aree più importanti nelle organizzazioni moderne (Boylan & Syntetos, 2021).

In questa tesi ci si concentrerà quindi sull'analisi della domanda intermittente, il cui studio ha ricevuto limitata attenzione in ambito accademico (si veda Nikolopoulos (2021) e il capitolo 2). In termini più formali, con il termine "domanda intermittente" si fa riferimento a dati di conteggio con una forte componente temporale contenenti un elevato numero di osservazioni pari a zero. In Figura 1.1 se ne può osservare un esempio.

1.1 Il contesto economico

Dal punto di vista operativo, la domanda intermittente è solitamente osservata a livelli gerarchici granulari, ossia i livelli più dettagliati della serie, sia considerando un'aggregazione temporale (ad esempio, le vendite giornaliere, piuttosto che quelle mensili) che contemporanea (ad esempio, le vendite di un singolo prodotto, piuttosto che una categoria di beni). Questa definizione comprende quindi la domanda giornaliera delle *Stock Keeping Units* (SKU), ossia le unità operative di base per la pianificazione dei rifornimenti e la distribuzione di stock giornaliero caratterizzate da un codice univoco (Fildes et al., 2022). Ad esempio, se si considerano i dati relativi ad un supermercato, si osserva intermittenza a livello di vendite giornaliere per ogni SKU di un punto vendita. Ne consegue che ottenerne previsioni accurate è essenziale per il miglioramento del processo decisionale relativo all'inventario e contenere i costi, costituiti principalmente da quelli relativi al magazzino e alle vendite perse o al *backordering*[2] (Wang et al., 2024).

Infatti, se le previsioni della domanda sono troppo alte, l'azienda produrrà più beni del necessario, che non riusciranno ad essere venduti, implicando una produzione non necessaria e un aumento dei costi delle risorse (materiali, forza lavoro e spazio di magazzino). D'altra parte, se le previsioni sono troppo basse, si potrebbe avere una carenza di prodotti e perdita di opportunità di vendita (Dmitry et al., 2019).

1.1.1 La domanda intermittente e le decisioni di inventario

Le principali decisioni basate sullo studio della domanda intermittente riguardano (CMAF, 2023) lo stoccaggio, l'organizzazione del rifornimento dell'inventario, i rendimenti della domanda (*returns*) e l'ultimo ordine da effettuare (*last-time buy*):

- Lo stoccaggio dipende dalla decisione dell'azienda di mantenere un certo oggetto nell'inventario o meno. Questa scelta ha implicazioni dirette sui costi di inventario: mantenere un certo *item* in stock comporta dei costi di mantenimento (*inventory holding charge*), che includono i costi opportunità (ossia l'impossibilità di investire i soldi legati allo stock in altri ambiti), i costi dello spazio occupato, potenziali furti, deterioramento e obsolescenza.

[2]Per coerenza con *newsvendor problem* (vedi sezione 3.4) non si considera la possibilità di *backordering*.

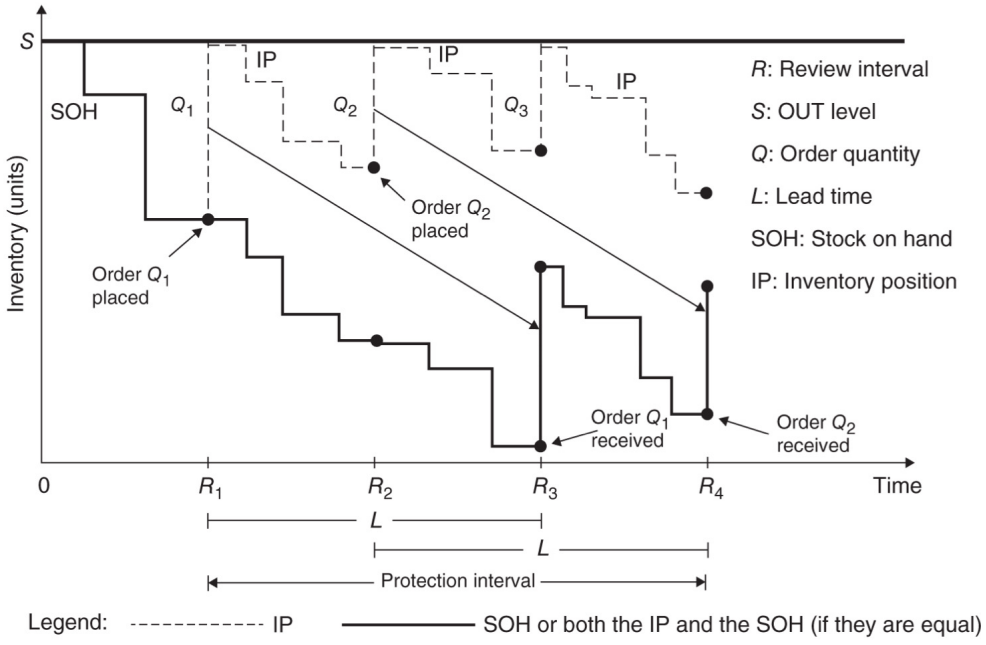

Figura 1.2: Rappresentazione della politica di inventario per le ordinazioni *order-up-to level* con periodi di revisione periodici, (Boylan & Syntetos, 2021).

Allo stesso tempo, il non stoccaggio dell'*item* può determinare la perdita di opportunità di vendita e clienti, che solitamente comportano costi più gravosi rispetto a quelli di mantenimento, ma meno frequenti e onerosi nel lungo termine (Boylan & Syntetos, 2021). È quindi importante avere buone previsioni della domanda media per bilanciare adeguatamente questo *trade-off* e ottenere decisioni adeguate.

- Le decisioni riguardanti il rifornimento sono particolarmente rilevanti nell'ambito delle serie intermittenti. L'andamento di questi *item*, infatti, implica la presenza di periodi con domanda insufficiente, in cui può essere conveniente aspettare di finire lo stock prima di riordinarlo o produrlo.

  Seguendo Boylan & Syntetos (2021), si considera una politica di inventario per le ordinazioni *order-up-to level* (OUT) con periodi di revisione periodici, indicata come maggiormente idonea nello studio della domanda intermittente. Tale approccio prevede l'ordinazione di una certa quantità di beni tale da portare l'inventario ad un certo livello S, studiato per coprire l'intervallo di protezione e il *lead time*[3], dopo ogni periodo di revisione R. La Figura 1.2

[3]Il *lead time* è il tempo necessario per stoccare l'*item*.

riporta graficamente tale politica, ipotizzando per semplicità l'assenza di *stockout* e un intervallo di previsione più corto del *lead time*.

- I *returns* sono definiti come la differenza tra la domanda lorda e quella netta. In questo caso si può procedere con la previsione della domanda lorda e netta oppure calcolando direttamente i *returns* futuri.

- Il *last-time buy* è la decisione aziendale di acquistare per l'ultima volta un certo bene ed è notoriamente difficile. Questa, infatti, richiede la previsione accurata della diminuzione della domanda nel tempo, ossia del modo in cui il numero di richieste dell'*item* diminuirà nei periodi di tempo successivi all'ultimo ordine effettuato.

  La sua rilevanza è esacerbata nel caso di *item* intermittenti, in quanto particolarmente suscettibili ad obsolescenza, fenomeno indicato da una riduzione nella probabilità di domanda o da un aumento delle osservazioni pari a zero (Sanguri et al., 2024). Questa caratteristica comporta che tali beni perdano di valore ed efficienza economica nel tempo, a causa della comparsa sul mercato di beni migliori e più competitivi, oppure in caso di usura dell'*item* stesso. Ne consegue, che la decisione di *last-time buy* è importante nell'ambito intermittente, in quanto ne è particolarmente soggetto.

La corretta previsione della domanda intermittente, quindi, porta a grandi benefici in termini economici, sia a livello di ricavi sia di costi, ma non solo. La complessità della *supply chain* e il ridursi del ciclo di vita dei prodotti osservato negli ultimi anni ha, infatti, portato all'aumento dell'obsolescenza, incrementando la creazione di rifiuti ambientali legati alla produzione di beni inutili e i conseguenti costi di trasporto e smaltimento. Una previsione accurata delle serie intermittenti, quindi, può determinare una significativa riduzione di sprechi e scarti, con importanti benefici ambientali (Boylan & Syntetos, 2021).

In questa tesi ci si concentra sullo studio della domanda indipendente (*top level*), ossia quella relativa al prodotto finito, ipotizzando una politica di inventario *Make To Stock* (MTS), in cui si ha disponibilità e spedizione immediata degli *item* richiesti (Boylan & Syntetos, 2021).

1.2 Classificazione della domanda intermittente

In molti ambiti applicativi, le previsioni devono essere svolte su un ampio insieme di prodotti, servizi e locazioni. In queste situazioni, conviene introdurre delle regole di classificazione che permettano di individuare i metodi o modelli più idonei per un'insieme di dati.

In alcuni contesti, come nel caso dei modelli ARIMA, la classificazione delle serie è ben consolidata tramite l'utilizzo di criteri di informazione, come, per esempio, l'AIC (Hyndman & Athanasopoulos, 2018). Tuttavia nell'ambito delle serie intermittenti la questione è più complessa, poiché spesso si utilizzano metodi privi di una base teorica pienamente soddisfacente, come il metodo di Croston (Petropoulos et al., 2022).

1.2.1 Classificazione ABC

A livello operativo, il metodo più comunemente usato per la categorizzazione della domanda intermittente è la classificazione ABC o di Pareto, che si basa sul volume delle vendite e i prezzi delle SKU.

Il metodo si fonda sulla regola di Pareto, detta anche regola 80:20[4], la cui validità nell'ambito dell'inventario è confermata da diversi studi, e divide gli *items* in tre categorie, in base alla domanda annuale espressa in termini di valore monetario (Boylan & Syntetos, 2021):

- *item* A: sono gli articoli che portano il massimo valore all'azienda, costituendo circa l'80% del valore dell'inventario con un numero di beni presenti che si aggira attorno al 20%. Per tale ragione, sono considerati a livello operativo quelli più critici, richiedendo una gestione accurata e frequente.

- *item* B: sono beni meno cruciali rispetto ai primi ma comunque importanti, determinando circa il 15% del valore dell'inventario con un numero di articoli presenti pari circa al 30%.

- *item* C: sono gli articoli che hanno un impatto economico minore sui ritorni dell'azienda, costituendo solo il 5% del valore dell'inventario pur determinandone il 50% in termini di volume.

[4]La regola di Pareto afferma che il 20% degli oggetti contenuti nell'inventario generano circa l'80% dei volumi di vendita, mentre il restante 80% determina il 20% rimanente.

Si vede, quindi, che la classe A rappresenta i beni che producono maggiori ricavi e sono considerati, spesso, dal punto di vista operativo più importanti. Questa classificazione è usata per stabilire, quindi, politiche di inventario che si concentrano su poche parti critiche, basandosi appunto sul principio di Pareto: "*there are few critical and many trivial*" (Dmitry et al., 2019).

Secondo questa classificazione, è opportuno usare i metodi classici per gli *item* A e B, e quelli per la domanda intermittente per la classe C. Tuttavia, Boylan & Syntetos (2021) sottolineano che il valore e il volume della domanda non sono dei criteri sufficienti per caratterizzare gli andamenti delle serie e risultano quindi inadatti per la classificazione in ambito previsivo:

- La dimensione della domanda (quando avviene) può falsare la percezione dell'intermittenza. Ad esempio, si considerino due *item* che si muovono rispettivamente tre volte all'anno, con 50 vendite l'una, e 10 volte all'anno, con due unità vendute alla volta. In base al criterio ABC, il secondo *item* verrebbe classificato come C, mentre il primo come A, anche se in realtà è più intermittente del secondo.

- Il prezzo del bene può distorgere la rappresentazione della realtà: nell'esempio sopra riportato, se si ipotizza che il primo *item* sia venduto a 2€ per pezzo e il secondo a 50€, si incorre nello stesso problema esposto nel primo punto.

1.2.2 Classificazione SBC

Una strategia alternativa è la classificazione attraverso la previsione della domanda (*forecasting-based classification*). Questa consiste nella comparazione diretta di diversi metodi previsivi tramite una misura teorica per la quantificazione dell'errore, come l'errore quadratico medio o MSE (*Mean Square Error*), al fine di identificare le caratteristiche delle serie che portano alla preferenza di un metodo rispetto ad un altro. In altre parole, si cercano di definire regioni di *performance* superiore dei metodi e i relativi *pattern* della domanda.

Un metodo ampiamente riconosciuto per distinguere tra diverse tipologie di domanda è la categorizzazione SBC (Syntetos Boylan Croston), introdotta da Syntetos et al. (2005), che si basa sull'intermittenza della serie e sulla variabilità del volume della domanda. Originariamente concepita per confrontare il metodo

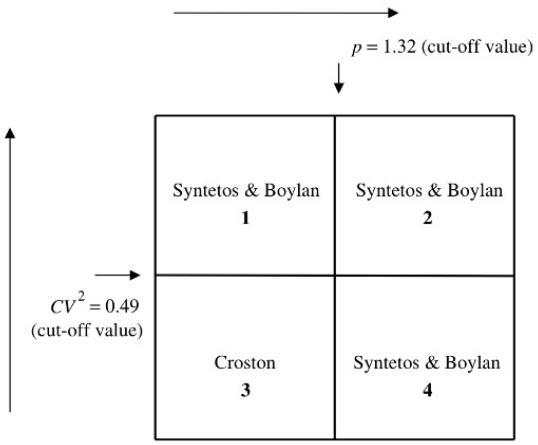

Figura 1.3: Schema di classificazione SBC: Croston vs. SBA, (Syntetos et al., 2005).

di Croston (Croston, 1972), la Syntetos-Boylan Approximation (SBA, Syntetos & Boylan (2001)) e il Simple Exponential Smoothing (SES, Brown (1956)), la classificazione SBC ha assunto una rilevanza più ampia, divenendo uno strumento generale per distinguere diverse tipologie di domanda (Makridakis et al., 2022).

Le categorie sono espresse in termini di intervallo inter-domanda medio, ossia l'*Average Demand Interval* (ADI), e di coefficiente di variazione della dimensione della domanda al quadrato (CV^2), definiti rispettivamente come

$$ADI = \frac{\sum_{t=1}^{T} \Delta_t}{T}$$

$$CV^2 = \frac{\sqrt{\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T}}}{\bar{y}}$$

dove Δ_t è il periodo di tempo tra due domande positive consecutive, T rappresenta il numero totale di periodi, y_t è la domanda osservata al tempo t, e $\bar{y}$ è la domanda positiva media (Kaya et al., 2020).

I risultati delle operazioni di confronto sono espressi come rapporto tra gli MSE dei due metodi e definiscono quattro aree, come illustrato in Figura 1.3:(1) erratiche, (2) *lumpy*, (3) *smooth* e (4) "strettamente" intermittenti.

In base a tali risultati, le serie sono considerate "strettamente" intermittenti quando sono caratterizzate da ampi periodi senza domanda e poca variazione nella dimensione quando presente. Si definiscono inoltre altre tre categorie per la domanda: erratiche, caratterizzate da domande molto variabili in dimensione

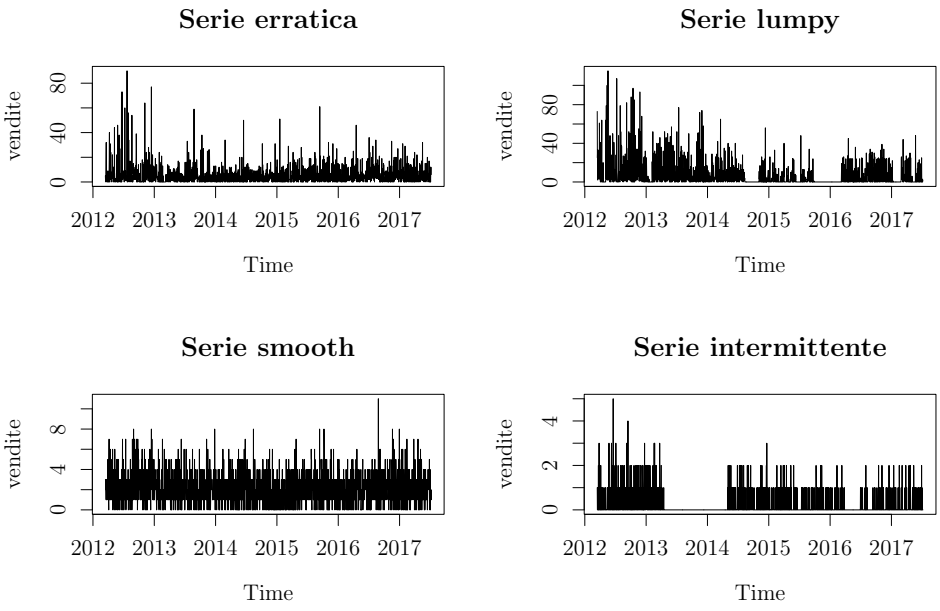

Figura 1.4: Esempi di serie in base alla classificazione SBC (dati della competizione M5). La categorizzazione è stata applicata utilizzando il pacchetto R `tsintermittent` (Kourentzes, 2022).

e intervalli inter-domanda corti; *lumpy*, con variazioni nella dimensione e intervalli senza domanda ampi; e *smooth*, caratterizzate da domande frequenti e poco variabili. Degli esempi di tali categorie possono essere osservati in Figura 1.4.

Capitolo 2

Previsioni probabilistiche della domanda intermittente

Come argomentato nel capitolo 1, la previsione della domanda intermittente è essenziale per la maggior parte delle attività di controllo e pianificazione di qualsiasi azienda. Tuttavia, la sua previsione non è un'operazione banale, a causa della complessità dei dati e del suo ingente peso nelle decisioni aziendali. Questi fattori determinano l'inadeguatezza dei metodi previsivi standard e la necessità di estrarre informazioni che vanno oltre alla semplice previsione puntuale.

In questo capitolo, si introdurranno le principali problematiche e approcci alla previsione della domanda intermittente. Si sottolineerà poi la necessità di considerare previsioni probabilistiche, di cui si esploreranno vari modelli e metodi che forniscono una base eterogenea per il calcolo di combinazioni di previsioni, esaminate nel capitolo 3. Per chiarire la terminologia, un modello di previsione è una rappresentazione matematica di un fenomeno reale con una specificazione completa della distribuzione e dei parametri, mentre un metodo previsivo è una procedura matematica che genera previsioni con o senza un modello previsivo. Un esempio classico è il metodo SES, che utilizza il modello di lisciamento esponenziale con la sola componente di errore additiva, ETS(A, N, N) nella formulazione proposta da Svetunkov & Boylan (2023).

Nonostante alcuni dei metodi descritti non siano nati per produrre previsioni probabilistiche, come il metodo di Croston, i risultati puntuali forniti vengono usati in fase di analisi per calcolare la densità della domanda, ipotizzando un'adeguata distribuzione di probabilità.

2.1 Problematiche e approcci alla previsione della domanda intermittente

Nonostante la rilevanza di questo tipo di dato, la letteratura si è principalmente focalizzata su serie *fast moving*, dando limitata attenzione alla domanda intermittente e ai relativi metodi di previsione. Tale mancanza di interesse può essere attribuita alla percezione della domanda intermittente come utile solo nell'ambito delle parti di ricambio e alla complessità della sua previsione. L'intermittenza, infatti, implica una duplice fonte di incertezza, derivante dalla natura sporadica del volume della domanda osservata e della tempistica di arrivo delle richieste del prodotto (ossia dei suoi elementi costituenti), che ne rende la previsione particolarmente ardua (Nikolopoulos, 2021). Gli intervalli di tempo senza domanda possono essere altamente variabili e lunghi (intermittenza), e la dimensione della domanda, quando presente, può assumere valori piccoli o costanti, oppure, in alternativa, essere altamente variabile (domanda erratica).

Inoltre, l'alta presenza di zeri può precludere l'identificazione delle componenti delle serie, come trend e stagionalità, e le informazioni sulla domanda storica sono spesso limitate e con valori assunti difficili da prevedere, in quanto spaziano da costanti ad altamente variabili.

Questa mancanza di informazione, sommata alla duplice fonte di variabilità che la caratterizza, implica la necessità di ipotesi semplificatrici per la modellazione di questo tipo di dati. Un'assunzione comune è quella di assenza di stagionalità. Nonostante tale ipotesi impedisca lo sviluppo di soluzioni ottimali[1] in senso statistico, permette di sviluppare metodi molto robusti[2] e di facile implementazione (Boylan & Syntetos, 2021).

Uno dei metodi più utilizzati per la previsione della domanda intermittente è il Simple Exponential Smoothing (SES), il quale tuttavia non riesce a tenere conto della duplice variabilità delle serie, portando ad una distorsione positiva, detta "*decision point bias*" (vedi Appendice A.2). Per affrontare questo problema, Croston (1972) propone di modellare separatamente la dimensione della domanda e la lunghezza degli intervalli inter-domanda tramite il metodo SES, riuscendo ad eliminare, almeno a livello teorico, la distorsione suddetta. Nel tempo sono

[1] L'ottimalità è definita come "migliore" performance sotto certe condizioni.

[2] Con robustezza si intende una performance sufficientemente buona per una vasta gamma di situazioni.

stati sviluppati altri metodi che si basano su questa idea, sia in ambito parametrico (come SBA e Teunter-Syntetos-Babai (TSB, Teunter et al. (2011)), varianti del metodo di Croston) che non parametrico (come il Willemain-Smart-Schwarz (WSS, Willemain et al. (2004)), un'estensione del bootstrap semplice ideato per trattare la domanda intermittente).

Un approccio alternativo consiste nell'aggregazione temporale dei dati. Dato che l'intermittenza è osservabile a livelli granulari, diminuire la frequenza delle osservazioni, ad esempio trasformando serie giornaliere in settimanali o mensili, riduce o addirittura elimina l'intermittenza che li caratterizza.

Se i dati sono sufficientemente numerosi, questo procedimento consente quindi l'uso di metodi validi per variabili continue. In questo caso, infatti, l'errore relativo tra i dati di conteggio sottostanti e la loro approssimazione continua diventa trascurabile, permettendo l'applicazione di molti teoremi statistici basati sulla legge dei grandi numeri (Kolassa, 2016). Un esempio di questo approccio è l'ADIDA (*Aggregate Disaggregate Intermittent Demand Approach*), il quale aggrega le serie in periodi di tempo *non-overlapping*, vi applica un qualche metodo previsivo per serie *fast moving* e poi le disaggrega nuovamente (Boylan & Syntetos, 2021).

La dottrina corrente, seguita anche in questo testo, si discosta da tale *modus operandi* e prevede l'utilizzo di modelli probabilistici avanzati sulle serie originali, cioè intermittenti (Petropoulos et al., 2022). Kolassa (2016) sottolinea, infatti, la propensione della ricerca verso i *big data* e l'utilizzo di dati a livelli granulari molto fini per la previsione operativa. Seguendo tali considerazioni, ci si concentra sull'analisi delle SKU e quindi su dati che mantengono intermittenza. L'utilizzo di modelli probabilistici avanzati è, invece, giustificato dai buoni risultati forniti, che superano la plethora di limiti che li caratterizzano, legati all'alto numero di dati necessari, la complessità matematica e la poca accettazione pratica dei metodi.

2.2 Perché usare previsioni probabilistiche?

Con il termine "previsione probabilistica" si fa riferimento a tutto ciò che va oltre alla previsione puntuale, ad esempio densità, quantili e intervalli di previsione. Nonostante la maggiore complessità di elicitazione, valutazione e aggregazione dei metodi probabilistici, le previsioni fornite contengono un'informazione più ricca, andando ad esaminarne anche l'incertezza (Petropoulos et al., 2022).

Questo approccio risulta particolarmente utile nell'ambito della domanda intermittente principalmente per due ragioni. La prima è intrinseca nella natura dei dati, che rende le metriche standard per la valutazione delle previsioni puntuali, come MAPE[3] e RMSE[4], poco adatte alla trattazione del problema. Questo è dovuto in primo luogo all'elevato numero di osservazioni nulle, che potrebbero comportare divisioni per zero. In secondo luogo, esse sono misure di errore relative che presuppongono una distribuzione normale dei dati, caratteristiche poco adatte per lo studio della domanda intermittente. La seconda ragione riguarda invece l'eccessiva sinteticità dei risultati prodotti dalle previsioni puntuali, che determina un contenuto informativo insufficiente per supportare adeguatamente le decisioni inerenti alla gestione dell'inventario (vedi sezione 1.1).

Considerate le ragioni sopra esposte, è opportuno concentrarsi sulle previsioni probabilistiche, allontanandosi dal *framework* puntuale, maggiormente studiato in letteratura (Wang et al., 2024). In questa tesi, saranno esaminati alcuni metodi per la previsione di densità, considerando l'informazione più completa riguardo ai dati come suggerito da Kolassa (2016) e Snyder et al. (2012), che ne sottolineano l'importanza.

2.3 ARIMA e ETS

Si considerano due metodi classici per l'analisi di serie storiche come *benchmark*: gli ARIMA e gli ETS (Error Trend Seasonality). Dato che i modelli ARIMA sono ben consolidati e ampiamente riconosciuti, in questa sezione si fornirà solo una breve introduzione ai modelli ETS. Questi sono delle "generalizzazioni" dei metodi di lisciamento esponenziale che si basano sulla scomposizione delle serie nelle sue componenti di trend, stagionalità e errore. Grazie alla sua formulazione *state space*, le informazioni relative a questi elementi possono essere incorporati all'interno del modello in diversi modi.

[3]Il *Mean Absolute Percentage Error* (MAPE) è definito come $MAPE = \frac{100}{N}\sum_{i=1}^{N}|\frac{y_i-\hat{y}_i}{y_i}|$, dove N è il numero di osservazioni, y_i sono i valori osservati e $\hat{y}_i$ sono i valori previsti.

[4]Il *Root Mean Squared Error*(RMSE) è definito come $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$, dove N è il numero di osservazioni, y_i sono i valori osservati e $\hat{y}_i$ sono i valori previsti.

Rappresentazione *state space*

Il principale vantaggio di questo modello è la sua scrittura in forma *single source of error* (SSOE) *state space* (vedi Appendice A.3), che gli conferisce ampia flessibilità. Questa formulazione cattura le dinamiche delle serie osservate tramite un vettore latente $\mathbf{v}_t$, noto come vettore degli stati (Hamilton, 1994). La sua rappresentazione è quindi composta da un'equazione di misura, che mostra il legame tra le variabili osservate e quelle latenti, e una di transizione, che modella $\mathbf{v}_t$ in base alle sue componenti, descrivendone l'evoluzione temporale. In termini analitici, l'equazione di misura è espressa nel seguente modo:

$$y_t = w(\mathbf{v}_{t-1}) + r(\mathbf{v}_{t-1})\epsilon_t \ , \tag{2.1}$$

dove $r(\cdot)$ è la funzione legata al termine di errore e $w(\cdot)$ è la funzione di misura. L'equazione di transizione è invece scritta come:

$$\mathbf{v}_t = f(\mathbf{v}_{t-1}) + g(\mathbf{v}_{t-1})\epsilon_t \tag{2.2}$$

dove $f(\cdot)$ è la funzione di transizione e $g(\cdot)$ indica la funzione di persistenza, che descrive l'estensione degli effetti dell'innovazione (ossia dell'errore) sullo stato (Svetunkov, 2023; Hyndman et al., 2008).

Modello ETS

Nell'ambito degli ETS, l'equazione (2.1) descrive il legame esistente tra la serie e le sue componenti (contenute nel vettore $\mathbf{v}_t$), mentre l'equazione (2.2) mostra la modellazione dei singoli elementi. In particolare, il modello ETS consente di incorporare le informazioni riguardanti le componenti in vari modi:

- l'errore può essere inserito come effetto additivo (A) o moltiplicativo (M);
- il trend può essere incorporato in maniera additiva, moltiplicativa, *additive damped* (Ad), *multiplicative damped* (Md) oppure non essere inserito (N);
- la stagionalità può essere inserita in maniera additiva, moltiplicativa oppure non essere considerata.

Ogni combinazione di queste alternative porta ad un modello distinto e definisce, in linea teorica, 30 possibili ETS.

Denotando $\mathbf{v}_t = (l_t, b_t, s_t)$, dove l_t è il livello della serie, b_t è il trend e s_t è la stagionalità, si esamina la formulazione *state space* scritta nelle equazioni 2.1 e 2.2 nell'ambito degli ETS, usando come esempio illustrativo il modello ETS(M, M, M). Si parte considerando l'equazione di misura (2.1), i cui elementi variano dipendentemente dalla natura delle componenti (Svetunkov, 2023):

- La funzione di misura $w(\mathbf{v}_{t-1})$ corrisponde all'addizione o moltiplicazione delle componenti, dipendentemente dal tipo di trend e stagionalità introdotti nel modello.

  Nel caso degli ETS(M, M, M), si ha che $w(\mathbf{v}_{t-1}) = l_{t-1}b_{t-1}s_{t-1}$.

- La funzione associata al termine di errore $r(\mathbf{v}_{t-1})$ varia in base all'errore, assumendo valore unitario e costante nel caso additivo e pari a $w(\mathbf{v}_{t-1})$ se l'errore è moltiplicativo.

  Ne consegue che nel caso di errore additivo l'equazione (2.1) può essere scritta come $y_t = w(\mathbf{v}_{t-1}) + \epsilon_t$, mentre nel caso di errore moltiplicativo come $y_t = w(\mathbf{v}_{t-1})(1 + \epsilon_t)$.

Per quanto riguarda invece l'equazione di transizione (2.2):

- La funzione di transizione $f(\mathbf{v}_{t-1})$ descrive il modo in cui le componenti interagiscono tra loro e variano nel tempo, e dipende da trend e stagionalità.

  Nel caso ETS(M, M, M), non considerando la seconda parte dell'equazione (2.2), si può scrivere:

$$\begin{aligned} l_t &= l_{t-1}b_{t-1} \\ b_t &= b_{t-1} \\ s_t &= s_{t-m} \ . \end{aligned} \tag{2.3}$$

- La funzione di persistenza $g(\mathbf{v}_{t-1})$ differisce in base alle specifiche del modello, ma si possono distinguere due casi speciali: ETS(A, A, A), in cui $g(\mathbf{v}_{t-1}) = \mathbf{g}$, e ETS(M, M, M), in cui $g(\mathbf{v}_{t-1}) = f(\mathbf{v}_{t-1})\mathbf{g}$. In entrambi i casi la funzione dipende dal vettore dei parametri di lisciamento $\mathbf{g}$, che nel contesto degli ETS è chiamato "vettore di persistenza".

L'equazione di transizione del modello ETS(M, M, M), quindi, si può scrivere come:

$$\begin{aligned} l_t &= l_{t-1}b_{t-1} + l_{t-1}b_{t-1}\alpha\epsilon_t \\ b_t &= b_{t-1} + b_{t-1}\beta\epsilon_t \\ s_t &= s_{t-m} + s_{t-m}\gamma\epsilon_t \ , \end{aligned} \tag{2.4}$$

dove α, β e γ sono i parametri di lisciamento e può essere semplificata nel seguente modo:

$$l_t = l_{t-1}b_{t-1}(1 + \alpha\epsilon_t) \tag{2.5}$$

$$b_t = b_{t-1}(1 + \beta\epsilon_t) \tag{2.6}$$

$$s_t = s_{t-m}(1 + \gamma\epsilon_t) \ . \tag{2.7}$$

Ognuna delle equazioni esposte descrive un aspetto della serie al tempo t: il modello rappresentato dall'equazione (2.5) descrive il livello e mostra il valore medio della serie per periodo di tempo, l'equazione (2.6) ne indica l'inclinazione, ossia ne descrive le variazioni nei valori, e l'equazione (2.7) esprime la componente stagionale.

Per garantire il corretto funzionamento del modello, gli ETS richiedono l'assunzione di diverse ipotesi standard di corretta specificazione del modello, incorrelazione tra le variabili esplicative e l'errore e tra le variabili esogene, e di normalità degli errori, che devono essere i.i.d. e omoschedastici. Inoltre, per far sì che le previsioni puntuali e le medie condizionate h passi in avanti coincidano, è necessario che $E(\epsilon_t) = 0$ nel caso di errore additivo e $E(1 + \epsilon_t) = 1$ nel caso moltiplicativo (vedi Appendice A.4).

Il calcolo delle previsioni del modello avviene tramite un algoritmo basato sul procedimento ideato da Hyndman et al. (2002). Questo prevede di ottimizzare i parametri per ogni serie e selezionare il miglior modello tramite AIC. Successivamente, si calcolano le previsioni puntuali e si usa un metodo bootstrap per simulare 5000 andamenti campionari futuri per la variabile risposta. Questo consente di individuare i percentili $1-\alpha/2$ e $\alpha/2$ per i dati simulati per ogni orizzonte previsivo, corrispondenti agli estremi dell'intervallo di confidenza. Sotto l'ipotesi di normalità degli errori è possibile utilizzare un bootstrap parametrico, il quale sfrutta tale assunzione per il calcolo dei residui. In mancanza dell'ipotesi di normalità, è necessario utilizzare gli errori ricampionati, cioè al bootstrap ordinario.

Dato che i modelli ETS con errore moltiplicativo sono numericamente instabili quando i dati contengono dei valori pari a zero o negativi (Hyndman & Athanaso-

poulos, 2018), per dati di domanda intermittente sono preferibili gli ETS con errori additivi, le cui possibili specificazioni sono riportate in Tabella 2.1 e raffigurate in Figura 2.1.

	Nonseasonal	Additive	Multiplicative
No trend	$y_t = l_{t-1} + \epsilon_t$ $l_t = l_{t-1} + \alpha\epsilon_t$	$y_t = l_{t-1} + s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + \alpha\epsilon_t$ $s_t = s_{t-m} + \gamma\epsilon_t$	$y_t = l_{t-1}s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + \alpha\frac{\epsilon_t}{s_{t-m}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}}$
Additive	$y_t = l_{t-1} + b_{t-1} + \epsilon_t$ $l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t$ $b_t = b_{t-1} + \beta\epsilon_t$	$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t$ $b_t = b_{t-1} + \beta\epsilon_t$ $s_t = s_{t-m} + \gamma\epsilon_t$	$y_t = (l_{t-1} + b_{t-1})s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + b_{t-1} + \alpha\frac{\epsilon_t}{s_{t-m}}$ $b_t = b_{t-1} + \beta\frac{\epsilon_t}{l_{t-1}s_{t-m}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}+b_{t-1}}$
Additive damped	$y_t = l_{t-1} + \phi b_{t-1} + \epsilon_t$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha\epsilon_t$ $b_t = \phi b_{t-1} + \beta\epsilon_t$	$y_t = l_{t-1} + \phi b_{t-1} + s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha\epsilon_t$ $b_t = \phi b_{t-1} + \beta\epsilon_t$ $s_t = s_{t-m} + \gamma\epsilon_t$	$y_t = (l_{t-1} + \phi b_{t-1})s_{t-m} + \epsilon_t$ $l_t = l_{t-1} + \phi b_{t-1} + \alpha\frac{\epsilon_t}{s_{t-m}}$ $b_t = \phi b_{t-1} + \beta\frac{\epsilon_t}{l_{t-1}s_{t-m}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}+\phi b_{t-1}}$
Multiplicative	$y_t = l_{t-1}b_{t-1} + \epsilon_t$ $l_t = l_{t-1}b_{t-1} + \alpha\epsilon_t$ $b_t = b_{t-1} + \beta\frac{\epsilon_t}{l_{t-1}}$ $s_t = s_{t-m} + \gamma\epsilon_t$	$y_t = l_{t-1}b_{t-1} + s_{t-m} + \epsilon_t$ $l_t = l_{t-1}b_{t-1} + \alpha\epsilon_t$ $b_t = b_{t-1} + \beta\frac{\epsilon_t}{l_{t-1}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}}$	$y_t = l_{t-1}b_{t-1}s_{t-m} + \epsilon_t$ $l_t = l_{t-1}b_{t-1} + \alpha\frac{\epsilon_t}{s_{t-m}}$ $b_t = b_{t-1} + \beta\frac{\epsilon_t}{l_{t-1}s_{t-m}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}b_{t-1}}$
Multiplicative damped	$y_t = l_{t-1}b_{t-1}^{\phi} + \epsilon_t$ $l_t = l_{t-1}b_{t-1}^{\phi} + \alpha\epsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\frac{\epsilon_t}{l_{t-1}}$ $s_t = s_{t-m} + \gamma\epsilon_t$	$y_t = l_{t-1}b_{t-1}^{\phi} + s_{t-m} + \epsilon_t$ $l_t = l_{t-1}b_{t-1}^{\phi} + \alpha\epsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\frac{\epsilon_t}{l_{t-1}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}b_{t-1}^{\phi}}$	$y_t = l_{t-1}b_{t-1}^{\phi}s_{t-m} + \epsilon_t$ $l_t = l_{t-1}b_{t-1}^{\phi} + \alpha\frac{\epsilon_t}{s_{t-m}}$ $b_t = b_{t-1}^{\phi} + \beta\frac{\epsilon_t}{l_{t-1}s_{t-m}}$ $s_t = s_{t-m} + \gamma\frac{\epsilon_t}{l_{t-1}b_{t-1}^{\phi}}$

Tabella 2.1: Equazioni degli ETS con errori additivi, (Svetunkov, 2023).

2.4 Modelli standard per la domanda intermittente

Il metodo standard per fare previsioni puntuali sulla domanda intermittente è il metodo di Croston (Croston, 1972), introdotto al fine di "correggere" la distorsione nei periodi successivi a domanda positiva (*issue point forecast*) che caratterizza i metodi di previsione standard, detta *"decision-point bias"*.

Come introdotto all'inizio del capitolo, una delle principali difficoltà nella previsione della domanda intermittente è la duplice incertezza che la caratterizza, legata alle sue componenti: la dimensione della domanda e la lunghezza degli intervalli di tempo che intercorrono tra domande positive (intervalli inter-domanda). Per ovviare tale problema, il metodo di Croston utilizza l'*exponential smoothing* per aggiornare separatamente le stime delle componenti della domanda intermittente, quando si verifica una domanda positiva.

Nonostante la superiorità teorica della metodologia proposta rispetto a quelle standard, l'evidenza empirica suggerisce che il metodo non apporti il migliora-

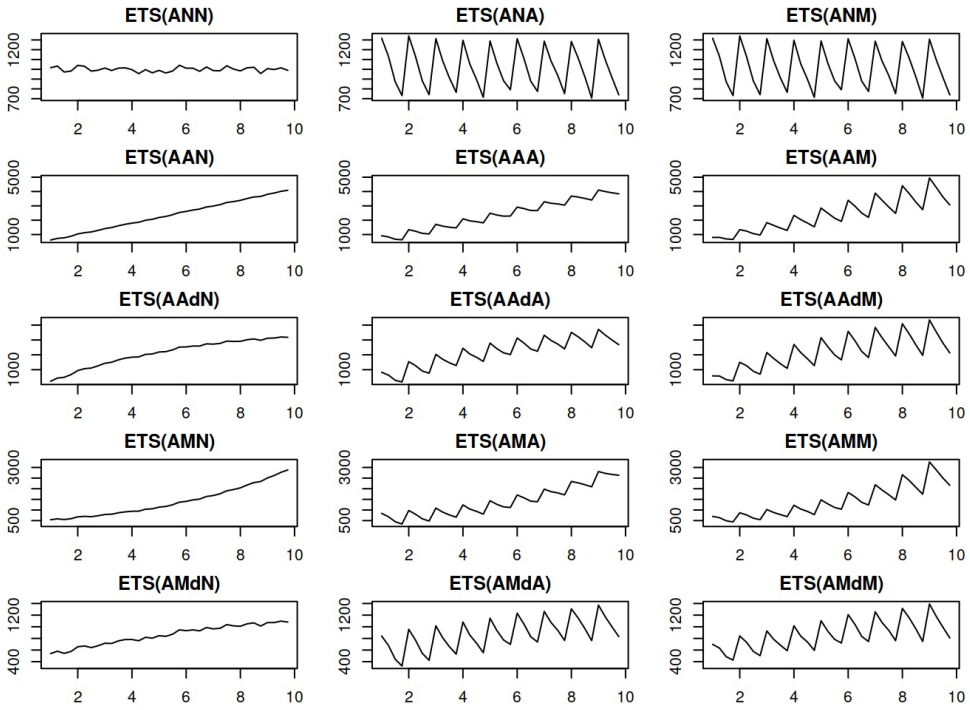

Figura 2.1: Serie storiche corrispondenti a modelli ETS con errori additivi, (Svetunkov, 2023).

mento atteso a livello di accuratezza delle previsioni (Syntetos & Boylan, 2001). Tale risultato è imputabile ad un errore matematico nella derivazione del valore atteso, che porta all'introduzione di distorsione nei risultati. È stata quindi introdotta una modifica per la correzione della distorsione: la Syntetos-Boylan Approximation, nota come SBA (Syntetos & Boylan, 2001).

Un altro difetto del metodo di Croston è dato dall'aggiornamento delle stime solo nel momento in cui si presenta una domanda positiva. Data la natura intermittente dei dati, questo può portare all'assenza di aggiornamento per molteplici periodi di tempo, che lo rendono quindi poco adatto per la gestione di problemi inerenti all'obsolescenza, come la gestione della rimozione di *stock* obsoleto (*dead stock*[5]) o in eccesso e la stima del rischio di obsolescenza. Per risolvere tale problema, è stata sviluppata un'altra variante del metodo di Croston: la Teunter-Syntetos-Babai, nota come TSB (Teunter et al., 2011), la quale aggiorna la probabilità di domanda positiva regolarmente.

[5]Per *dead stock* si intendono i prodotti che non si riescono a vendere e che hanno una bassa probabilità di essere venduti in futuro.

2.4.1 Il metodo di Croston

Il metodo di Croston è considerato il metodo principale per ottenere previsioni puntuali per la domanda intermittente e costituisce un metodo robusto, ma non ottimale (Boylan & Syntetos, 2021).

La metodologia, introdotta da Croston (1972), utilizza un *modus operandi* diverso dai metodi di previsione convenzionali, stimando e aggiornando separatamente le componenti della domanda, ovvero la sua dimensione e la lunghezza degli intervalli inter-domanda. Ciò gli consente di tenere conto della duplice fonte di incertezza della domanda intermittente, derivante dalla variabilità contemporanea dei due elementi costituenti, eliminando così la distorsione che caratterizza i metodi standard. Si consideri, ad esempio, il Simple Exponential Smoothing (SES) (vedi Appendice A.2), ampiamente usato nelle applicazioni relative all'inventario. La sua formulazione attribuisce pesi più alti alle osservazioni recenti, calcolando quindi previsioni più alte in periodi preceduti da valori positivi. Tuttavia, dato l'alto numero di zeri contenuti nelle serie, periodi successivi ad una domanda positiva sono spesso seguiti da assenza di domanda, determinando quindi una sovrastima delle previsioni del metodo. In termini più formali, Croston (1972) dimostra che il valore atteso della domanda al tempo *t+1* condizionatamente alla presenza di domanda positiva al tempo t previsto tramite il metodo SES è pari a:

$$E(\hat{y}_{t+1}|y_t > 0) = \frac{\mu}{p}(1 + \alpha(p-1)) \tag{2.8}$$

dove $\hat{y}_{t+1}$ è la previsione della domanda al tempo *t+1*, μ e p sono rispettivamente i valori attesi del volume e della lunghezza degli intervalli, e α è la costante di lisciamento. Si può quindi osservare una distorsione attesa percentuale del metodo pari a $100\alpha(p-1)$ (Boylan & Syntetos, 2021).

Il metodo di Croston, quindi, definisce la domanda in funzione delle sue componenti:

$$y_t = o_t z_t = \frac{z_t}{q_t} \tag{2.9}$$

dove y_t è la domanda, z_t è la sua dimensione, e o_t ne indica l'occorrenza e coincide con l'inversa dell'intervallo inter-domanda q_t. La modellazione delle due componenti si basa sulle ipotesi che:

i) $o_t \sim Be(1/p)$, dove $1/p$ è la probabilità di avere domanda diversa da zero che è assunta costante nel tempo;

ii) z_t abbia una qualche distribuzione di probabilità.

Il valore atteso della domanda intermittente è quindi calcolato come segue:

$$E(y_t) = \frac{\mu}{p} \ . \tag{2.10}$$

Ne consegue che la previsione della domanda al tempo *t+1* calcolata tramite il metodo di Croston è data dalla seguente equazione:

$$\hat{y}_{t+1} = \frac{\mu_{t+1}}{p_{t+1}} \ . \tag{2.11}$$

dove i valori attesi delle due componenti sono calcolati tramite il metodo SES, come:

$$\begin{aligned} \mu_{t+1} &= \alpha z_t + (1-\alpha)\mu_t \\ p_{t+1} &= \alpha q_t + (1-\alpha)p_t \end{aligned} \tag{2.12}$$

dove p_t e μ_t sono le medie condizionate stimate al tempo t. La previsione della domanda ottenuta per il tempo *t+1*, $\hat{y}_{t+1}$, è utilizzata per ogni periodo di tempo futuro fino a che non si osserva una domanda positiva, in cui si ipotizza che il valore atteso rimanga costante.

Si può vedere che l'aggiornamento avviene separatamente per le due componenti e solo dopo un periodo con domanda non zero. Ne consegue che, nel caso in cui sia presente domanda in ogni periodo, i risultati dell'equazione (2.10) coincidono con quelli del metodo del lisciamento esponenziale e può essere quindi applicato anche a serie storiche non intermittenti.

Esempio numerico

Per comprendere meglio il funzionamento del metodo descritto, si supponga di avere a disposizione i dati riguardanti la domanda per otto periodi di tempo, riportati in Tabella 2.2 (Boylan & Syntetos, 2021). Per ottenere le previsioni tramite il metodo di Croston, si eseguono i seguenti passaggi:

1. Si scompone la serie in Tabella 2.2 in altre due serie inerenti la dimensione della domanda e la lunghezza degli intervalli, riportate in Tabella 2.3.

2. Si ottengono le previsioni iniziali delle due componenti usando i valori attesi dei dati storici. Nell'esempio proposto si ha quindi che $\mu_9 = 3$ e $p_9 = 2$.

3. Si utilizzano le previsioni ottenute al punto precedente per calcolare la previsione della domanda in base alla formula riportata nell'equazione (2.11). Nell'esempio, quindi, si ottiene una previsione iniziale $\hat{y}_9 = 1.5$ di unità vendute per periodo, che corrisponde alla media delle osservazioni in Tabella 2.2.

4. Si applica il metodo SES separatamente alle serie delle componenti della domanda e si aggiornano ogni volta che si osserva una domanda positiva.

   Nell'esempio, si ipotizza una costante di lisciamento pari a 0.2 e che la domanda osservata nei periodi 9 e 10 siano rispettivamente pari a 6 e 0. Il metodo SES per il periodo 10 è calcolato quindi come[6]

$$\begin{aligned} \mu_{10} &= 0.2z_9 + 0.8\mu_9 = 3.6 \\ p_{10} &= 0.2o_9 + 0.8p_9 = 1.8 \ . \end{aligned} \tag{2.13}$$

   Quindi, la previsione della domanda per il periodo di tempo 10 calcolata tramite il metodo di Croston è pari a $\hat{y}_{10} = 2$.

   Alla fine del periodo 10, dato che è privo di domanda, le previsioni non verranno aggiornate. Ne consegue che $p_{11} = p_{10},\ \mu_{11} = \mu_{10}$ e $\hat{y}_{11} = \hat{y}_{10}$.

Tabella 2.2: Esempio di serie di domanda intermittenti (primi otto periodi).

Periodo di tempo	1	2	3	4	5	6	7	8
Domanda	4	0	2	5	0	0	0	1

Tabella 2.3: Esempio di serie del volume della domanda e degli intervalli inter-domanda (primi otto periodi).

Dimensione della domanda	4	2	5	1
Intervalli inter-domanda	1	2	1	4

Il metodo di Croston in forma *state space*

Il modello *state space* del metodo di Croston può essere quindi esplicitato come segue (Svetunkov & Boylan, 2023):

[6] o_9 utilizzato per il calcolo viene posto pari ad 1 per riflettere le occorrenze di domanda consecutive (Boylan & Syntetos, 2021).

$$\begin{aligned} \hat{y}_t &= \hat{y}_{j_t} = \frac{1}{\hat{q}_{j_t}} \hat{z}_{j_t} \\ \hat{z}_{j_t} &= \alpha z_{j_t-1} + (1-\alpha)\hat{z}_{j_t-1} \ , \\ \hat{q}_{j_t} &= \alpha q_{j_t-1} + (1-\alpha)\hat{q}_{j_t-1} \\ j_t &= j_{t-1} + o_t \end{aligned} \tag{2.14}$$

dove $\hat{y}_{j_t}$ è la domanda media prevista, $\hat{z}_{j_t}$ è la sua dimensione prevista, $\hat{q}_{j_t}$ è l'intervallo di domanda previsto, α è il parametro di lisciamento, e $j_t = 1, ..., N$ è un insieme di numeri sequenziali con N domande non zero.

2.4.2 Approssimazione SBA

La Syntetos-Boylan Approximation, nota come SBA, è stata introdotta da Syntetos & Boylan (2001) al fine di correggere la distorsione presente nel metodo di Croston, derivante da un errore nella derivazione numerica del valore atteso, detto *inversion bias*:

$$E\left(\frac{1}{p}\right) \neq \frac{1}{E(p)} \ . \tag{2.15}$$

L'approssimazione mantiene lo stesso concetto di Croston, costruendo le previsioni della domanda usando i suoi eventi costituenti. Al fine di eliminare la distorsione suddetta, il metodo utilizza una formulazione per l'aggiornamento alternativa, in grado di ridurre, in linea teorica, la distorsione del metodo originale tramite l'introduzione di un fattore di correzione. La SBA, quindi, formula le previsioni e gli aggiornamenti nel seguente modo:

$$\hat{y}_{t+1} = \left(1 - \frac{\alpha}{2}\right)\frac{\mu}{p} \tag{2.16}$$

dove la notazione è la stessa usata nella sezione 2.4.1.

Quantificazione dell'*inversion bias*

Sebbene la quantificazione dell'*inversion bias* sia stata introdotta da Syntetos & Boylan (2001), in questa tesi si illustra la procedura esposta da Syntetos & Boylan (2005) in quanto più diretta. Syntetos & Boylan (2005) considerano le previsioni SES non distorte delle componenti della domanda $\hat{z}_{t+1}$ e $\hat{q}_{t+1}$ tali che $E(\hat{z}_{t+1}) = \mu$ e $E(\hat{q}_{t+1}) = p$. Applicando il teorema di Taylor alla funzione $g(\hat{z}_{t+1}, \hat{q}_{t+1}) = \frac{\hat{z}_{t+1}}{\hat{q}_{t+1}}$, sotto l'ipotesi di indipendenza delle stime lisciate, si può scrivere:

$$g(\hat{z}_{t+1}, \hat{q}_{t+1}) = g(\mu, p) + \frac{\partial g(\mu,p)}{\partial \hat{z}_{t+1}}(\hat{z}_{t+1} - \mu) + \frac{\partial g(\mu,p)}{\partial \hat{q}_{t+1}}(\hat{q}_{t+1} - \mu)$$
$$+ \frac{1}{2}\frac{\partial^2 g(\mu,p)}{\partial \hat{z}_{t+1}^2}(z_{t+1} - \mu)^2 + \frac{\partial g(\mu,p)}{\partial \hat{z}_{t+1}\partial \hat{q}_{t+1}}(z_{t+1} - \mu)(\hat{q}_{t+1} - p)$$
$$+ \frac{1}{2}\frac{\partial^2 g(\mu,p)}{\partial \hat{q}_{t+1}^2}(\hat{q}_{t+1} - \mu) + ...$$

Ne consegue che il valore atteso della funzione $g(\hat{z}_{t+1}, \hat{q}_{t+1}) = \frac{\hat{z}_{t+1}}{\hat{q}_{t+1}}$ può essere espresso come segue:

$$E\left(g(\hat{z}_{t+1}, \hat{q}_{t+1})\right) = \frac{\mu}{p} + \frac{\mu}{p}\text{Var}(q_{t+1}) + ... \tag{2.17}$$

Questo risultato deriva principalmente dalla non distorsione delle previsioni (che rende nulli i valori attesi degli elementi $(\hat{z}_{t+1} - \mu)$ e $(\hat{q}_{t+1} - p)$) e dal fatto che la derivata seconda parziale di $g(\hat{z}_{t+1}, \hat{q}_{t+1})$ rispetto al volume della domanda è pari a zero.

Dato che gli intervalli sono indipendenti tra di loro e geometricamente distribuiti[7] con varianza $\sigma_q^2 = p(p+1)$, l'equazione (2.17) può essere espressa approssimativamente come

$$E\left(\frac{\hat{z}_{t+1}}{\hat{q}_{t+1}}\right) \approx \frac{\mu}{p} + \frac{\alpha}{1-\alpha}\frac{p-1}{p^2}\mu \ . \tag{2.18}$$

Ne consegue che l'*inversion bias* del metodo di Croston può essere approssimato dal secondo termine dell'equazione (2.18).

Correzione della distorsione del metodo di Croston

Per eliminare l'*inversion bias*, Syntetos & Boylan (2001) propongono il seguente stimatore

$$\hat{y}_{t+1} = \left(1 - \frac{\alpha}{2}\right)\frac{\mu}{p} \tag{2.19}$$

La SBA corregge il valore atteso della domanda nell'equazione (2.10) come segue

[7]Dato che l'occorrenza della domanda si distribuisce come una Bernoulli e gli intervalli corrispondono alla sua inversa, q_{t+1} segue una distribuzione geometrica.

$$E(\hat{y}_t) = E(\hat{z}_t)E\left(\frac{1}{\hat{p}_t\alpha^{\hat{p}_t-1}}\right) = \frac{\mu}{p} \tag{2.20}$$

dove α è una costante introdotta per eliminare la distorsione. In teoria, per far sì che questa equazione sia corretta, α dovrebbe assumere un valore infinitamente alto. Tuttavia, Syntetos et al. (2005) mostrano che l'equazione (2.20) è una buona approssimazione per α sufficientemente grande, portando a dei miglioramenti a livello di accuratezza delle previsioni rispetto al metodo di Croston. Il valore atteso della SBA è ricavabile direttamente dall'equazione (2.18):

$$E\left(\left(1-\frac{\alpha}{2}\right)\frac{\mu}{p}\right) \approx \frac{\mu}{p} - \frac{\alpha}{2}\frac{\mu}{p} + \frac{\alpha}{2}\frac{\mu-1}{p^2}\mu = \frac{\mu}{p} - \frac{\alpha}{2}\frac{\mu}{p^2}\ . \tag{2.21}$$

Il metodo presenta quindi una distorsione negativa proporzionale a $1/p^2$, invece che a $1/p$. Dato che $p > 1$, la distorsione della SBA diminuisce più rapidamente rispetto a quella di Croston all'aumentare della lunghezza degli intervalli inter-domanda.

Nel caso di assenza di domanda, l'approssimazione coincide con il metodo di Croston.

2.4.3 La variante TSB

Il TSB è una variante del metodo di Croston introdotto da Teunter et al. (2011) per migliorare la gestione dell'obsolescenza. Infatti, sia l'approccio originale che la SBA aggiornano le stime solo dopo l'arrivo di domanda positiva. Di conseguenza, in caso di improvvisa obsolescenza, questi metodi rimangono ancorati alle stime pre-esistenti e reagiscono lentamente ad aumenti del rischio di obsolescenza, poiché l'aggiustamento non avverrà fino ad una domanda positiva. Come introdotto nel capitolo 1, queste considerazioni hanno particolare valenza nell'ambito della domanda intermittente, data la rilevanza nell'ambiente dell'inventario, specialmente per beni *slow moving*.

Al contrario dei metodi precedentemente descritti, il metodo TSB aggiorna la probabilità di domanda regolarmente, aumentandola in periodi con domanda positiva e riducendola in sua assenza. Grazie a questa caratteristica, il metodo è in grado di reagire velocemente a situazioni di obsolescenza improvvisa, come l'eliminazione di alcune parti di ricambio nel processo produttivo, o di aumento del rischio di obsolescenza, ad esempio all'entrata nel mercato di un competitor.

Il metodo TSB, perciò, aggiorna le stime delle componenti della domanda con frequenze diverse, seguendo l'approccio di Croston per l'aggiornamento di z_t e calcolando o_t in ogni periodo. Le operazioni sono svolte usando il SES con costanti di lisciamento diverse per le componenti della domanda, data la differenza di tempistiche nell'aggiornamento dei due elementi.

Si sottolinea che nessun metodo può prevenire completamente l'obsolescenza, poiché potrebbe dipendere da una cattiva gestione e, anche senza questo fattore, spesso è difficile determinare quando un articolo dovrebbe essere dismesso. Tuttavia, il TSB costituisce un metodo valido per cercare di tenerne conto.

2.5 ETS intermittenti

Unendo le idee espresse nelle sezioni 2.3 e 2.4 , si ottengono i modelli ETS intermittenti (iETS), dei modelli *single source of error* (SSOE) *state space* generali introdotti da Svetunkov & Boylan (2023) al fine di incorporare l'intermittenza dei dati nel modello ETS.

L'idea parte dalla formulazione del metodo di Croston espressa nell'equazione (2.9) e sviluppa un modello *state space* che inserisce una interazione diretta tra il termine di errore e la dimensione della domanda, e parametri di lisciamento diversi per la dimensione della domanda e gli intervalli. Il modello segue l'approccio di Croston, modellando separatamente la dimensione della domanda e la sua occorrenza, sfruttando una struttura *state space* simile a quella degli ETS. In particolare, Svetunkov & Boylan (2023) si concentrano sul caso ETS(M, N, N), ossia con errore moltiplicativo e senza trend e stagionalità. Questa scelta è dettata dalla semplicità e familiarità del metodo, usato come metodo principale nella stima del metodo di Croston e TSB (descritti nell'sezione 2.4).

Tuttavia, a differenza degli ETS, il modello iETS non assume la normalità degli errori, prediligendo distribuzioni Log-Normali, Gamma o Gaussiana Inversa (vedi Appendice A.5). Queste possono prendere solo valori positivi e regolano l'asimmetria e la curtosi tramite i parametri, consentendo la flessibilità necessaria per l'analisi della domanda intermittente. Infatti, la non negatività e i valori potenzialmente bassi che caratterizzano questo tipo di dati rendono poco adatte l'ipotesi di simmetria e la possibilità di valori negativi della distribuzione normale in questo contesto.

In termini formali, il modello è definito come segue:

$$\begin{aligned} y_t &= o_t z_t \\ z_t &= w(\mathbf{v}_{t-\mathbf{l}}) + r(\mathbf{v}_{t-\mathbf{l}})\epsilon_{z,t} \\ \mathbf{v}_t &= f(\mathbf{v}_{t-\mathbf{l}}) + g(\mathbf{v}_{t-\mathbf{l}})\epsilon_{z,t} \end{aligned} \tag{2.22}$$

dove o_t è una variabile casuale Bernoulli, z_t è la dimensione della domanda potenziale caratterizzata da una qualche distribuzione condizionale, $\mathbf{l}$ è il vettore dei ritardi delle componenti (denotando la possibilità di considerare *lags* differenti del vettore $\mathbf{v}_t$) e le restanti componenti sono le stesse descritte per le equazioni 2.1 e 2.2. Si ha quindi che la prima equazione corrisponde alla formulazione originale del metodo di Croston, la seconda riflette l'evoluzione temporale della domanda potenziale (equazione di misura) e la terza descrive il cambiamento delle componenti del modello nel tempo (ossia l'equazione di transizione standard per modelli SSOE). Questa struttura può essere vista come un modello mistura, composto da un modello sottostante la domanda potenziale e uno riguardante la realizzazione della domanda.

Il modello descritto nell'equazione (2.22) richiede quattro assunzioni:

A1. z_t è continua nei suoi valori. Nonostante questa ipotesi possa risultare restrittiva, Svetunkov & Boylan (2023) mostrano che i modelli forniscono buone prestazioni anche nel caso di dimensione della domanda discreta.

A2. La domanda potenziale potrebbe fluttuare nel tempo anche in assenza di osservazioni dirette. Questa ipotesi riflette i cambiamenti nei bisogni dei consumatori, anche se non vi è acquisto effettivo di prodotto.

A3. z_t è indipendente da o_t.

A4. $o_t \sim Be(p_t)$, dove p_t è la probabilità di avvenimento di domanda che, nel caso più generale, varia nel tempo.

Alcune di queste ipotesi possono essere rilassate e conducono a modelli differenti.

Modello iETS generale

Sotto le assunzioni (A1), (A2), (A3) e (A4), il modello prende la forma iETS_G, che corrisponde al modello *state space* intermittente continuo generale presentato

nell'equazione (2.22). Sotto l'ipotesi di modellazione di z_t tramite ETS(M, N, N), si ottiene la seguente forma:

$$\begin{aligned} y_t &= o_t z_t \\ z_t &= l_{z,t-1}(1+\epsilon_{z,t}) \ , \\ l_{z,t} &= l_{z,t-1}(1+\alpha_z \epsilon_{z,t}) \end{aligned} \tag{2.23}$$

Per far sì che il modello fornisca previsioni puntuali adeguate, si richiede la condizione $E(1+\epsilon_{z,t}) = 1$, che permette di derivare le formule di media e varianza condizionate per valori h passi in avanti come segue:

$$\mu_{z,t+h|t} = l_{z,t}$$

$$\sigma^2_{z,t+h|t} = l^2_{z,t}\left((1+\alpha_z^2\sigma_\epsilon^2)^{h-1}(1+\sigma_\epsilon^2) - 1\right)$$

dove σ^2_ϵ è la varianza del termine di errore.

In merito alla parte riguardante l'occorrenza della domanda, invece, è naturale ipotizzare una distribuzione Bernoulli con parametro p_t che varia nel tempo, stimato tramite le medie condizionate di due variabili latenti: l'occorrenza della domanda, $\mu_{a,t}$, e la sua assenza, $\mu_{b,t}$.

Tali considerazioni implicano la seguente struttura del modello:

$$\begin{aligned} y_t &= o_t l_{z,t-1}(1+\epsilon_{z,t}) \\ l_{z,t} &= l_{z,t-1}(1+\alpha_z\epsilon_{z,t}) \\ o_t &\sim \text{Be}(p_t) \\ p_t &= \frac{\mu_{a,t}}{\mu_{a,t}+\mu_{b,t}} \\ \mu_{a,t} &= w(\mathbf{v}_{t-\mathbf{l}_a}) \\ \mathbf{v}_{a,t} &= f(\mathbf{v}_{a,t-\mathbf{l}_a}) + g(\mathbf{v}_{a,t-\mathbf{l}_a})\epsilon_{a,t} \\ \mu_{b,t} &= w(\mathbf{v}_{t-\mathbf{l}_b}) \\ \mathbf{v}_{b,t} &= f(\mathbf{v}_{b,t-\mathbf{l}_b}) + g(\mathbf{v}_{b,t-\mathbf{l}_b})\epsilon_{a,t} \end{aligned} \tag{2.24}$$

dove gli indici a e b denotano le parti del modello relative rispettivamente all'occorrenza e all'assenza della domanda, $1+\epsilon_{a,t}$ e $1+\epsilon_{b,t}$ sono termini di errore mutualmente e serialmente indipendenti, e α_a e α_b sono i parametri di lisciamento.

La presenza di due parametri variabili nel tempo, $\mu_{a,t}$ e $\mu_{b,t}$, consente al modello di coprire tutti i possibili casi di variazione temporale della probabilità. Il modello generale, infatti, è in grado di catturare i casi di domanda fissa, crescente

e decrescente (dipendentemente dai valori dei parametri), insieme ad una situazione in cui la domanda evolve da uno stato all'altro e non converge nè a 0 nè a 1.

Focalizzandosi sull'analisi degli ETS(M, N, N), si definisce il seguente modello:

$$\begin{aligned} y_t &= o_t l_{z,t-1}(1+\epsilon_{z,t}) \\ l_{z,t} &= l_{z,t-1}(1+\alpha_z \epsilon_{z,t}) \\ o_t &\sim \text{Be}(p_t) \\ p_t &= \frac{\mu_{a,t}}{\mu_{a,t}+\mu_{b,t}} \\ \mu_{a,t} &= l_{a,t-1} \\ l_{a,t} &= l_{a,t-1}(1+\alpha_a \epsilon_{a,t}) \\ \mu_{b,t} &= l_{b,t-1} \\ l_{b,t} &= l_{b,t-1}(1+\alpha_b \epsilon_{b,t}) \end{aligned} \tag{2.25}$$

dove $l_{a,t}$ e $l_{b,t}$ sono i livelli delle variabili.

2.5.1 Specificazioni derivanti dal modello iETS generale

In questa sottosezione, si presentano alcuni modelli derivati dal modello iETS_G, i quali differiscono per proprietà, formulazione e convergenza stocastica degli ETS(M, N, N). Tutte le specificazioni che seguono, derivano dal rilassamento di diverse ipotesi del modello presentate sopra.

iETS_F

iETS_F è il più semplice dei modelli presentati ed è appropriato nel caso di probabilità di occorrenza della domanda costante nel tempo, ossia l'intervallo di domanda medio è fisso. Questo modello, infatti, ipotizza che $\mu_{a,t}$ e $\mu_{b,t}$ rimangano costanti nel tempo, eliminando di fatto le equazioni delle dimensioni della domanda:

$$\begin{aligned} o_t &\sim Be(p) \\ p &= \frac{\mu_a}{\mu_a+\mu_b} \end{aligned} \tag{2.26}$$

In questo modello, quindi, si assume che la probabilità di occorrenza della domanda non vari nel tempo, riducendo il numero di parametri necessari e di con-

seguenza la flessibilità. Per tali ragioni, il modello non verrà utilizzato nell'analisi descritta nel capitolo 5.

$\mathbf{iETS_O}$

iETS_O, detto modello *odds-ratio* (OR), risulta appropriato nel caso in cui i dati mostrino obsolescenza, ossia quando la domanda cala lentamente e i clienti smettono di acquistare il prodotto. In altre parole, fornisce buoni risultati in situazioni in cui la probabilità di occorrenza della domanda converge a zero. Il modello è ottenuto vincolando $\mu_{b,t} = 1$ nell'equazione (2.25):

$$\begin{aligned} o_t &\sim \text{Be}(p_t) \\ p_t &= \frac{\mu_{a,t}}{\mu_{a,t} + 1} \\ \mu_{a,t} &= l_{a,t-1} \\ l_{a,t} &= l_{a,t-1}(1 + \alpha_a \epsilon_{a,t}) \end{aligned} \tag{2.27}$$

Da questa definizione, si può osservare che $\mu_{a,t} = \frac{p_t}{1-p_t}$. Si ha quindi che quando $\mu_{a,t}$ decresce, si riduce anche l'OR e perciò anche la probabilità di occorrenza della domanda.

$\mathbf{iETS_I}$

iETS_I, detto modello OR inverso, cattura correttamente le dinamiche dei dati quando la domanda è in aumento, ossia situazioni in cui la probabilità eventualmente converge ad uno. Questo modello ipotizza $\mu_{a,t} = 1$, imponendo una dinamica specifica dell'aumento della probabilità di occorrenza:

$$\begin{aligned} o_t &\sim \text{Be}(p_t) \\ p_t &= \frac{1}{1 + \mu_{b,t}} \\ \mu_{b,t} &= l_{b,t-1} \\ l_{b,t} &= l_{b,t-1}(1 + \alpha_b \epsilon_{b,t}) \ . \end{aligned} \tag{2.28}$$

Si può osservare che $\mu_{b,t} = \frac{1-p_t}{p_t}$, ossia la parte di occorrenza modella l'OR inverso, da cui deriva il nome del modello.

$\mathbf{iETS_D}$

iETS_D funziona bene su dati con obsolescenza, in quanto cattura le variazioni nella probabilità di occorrenza della domanda più rapidamente rispetto agli iETS_O. Il modello pone i vincoli

$$\mu_{a,t} + \mu_{b,t} = 1 \ , \ \mu_{a,t} \leq 1$$

da cui si può vedere che, una volta inseriti nell'equazione (2.25), $p_t = \mu_{a,t}$. Il modello diventa quindi:

$$\begin{aligned} o_t &\sim \text{Be}(\mu_{a,t}) \\ \mu_{a,t} &= \min(l_{a,t-1}, 1) \\ l_{a,t} &= l_{a,t-1}(1 + \alpha_a \epsilon_{a,t}) \end{aligned} \tag{2.29}$$

2.5.2 Stima degli iETS

Seguendo l'idea di Croston, la stima degli iETS richiede il calcolo di stime distinte per la dimensione della domanda e la sua occorrenza. Le procedure descritte di seguito sono adattate sulla base del modello generale iETS_G e, se opportuno, per i casi specifici esaminate nella sottosezione precedente.

Stima della dimensione della domanda

La parte del modello riguardante z_t appare nell'equazione (2.25) come:

$$l_{z,t} = l_{z,t-1}(1 + \alpha_z \epsilon_{z,t}) \ .$$

La procedura di stima distingue i casi di presenza e di assenza di domanda, ossia rispettivamente $o_t = 1$ e $o_t = 0$.

Nel primo caso, la costruzione è semplice, in quanto il termine di errore è stimabile direttamente tramite l'errore di previsione $e_{z,t} = \frac{z_t - \hat{l}_{z,t-1}}{\hat{l}_{z,t-1}}$, in modo tale che lo stato si aggiorni ad ogni osservazione.

Nel secondo, invece, si riscontrano delle difficoltà nella stima del livello della dimensione della domanda, dato che tale variabile non è osservabile. Al suo posto, quindi, si può utilizzare il valore atteso condizionato, che, seguendo direttamente le proprietà degli ETS(M, N, N), si ottiene come $\hat{l}_{z,t+h|t} = \hat{l}_{z,t}$.

Stima dell'occorrenza della domanda

Per quanto riguarda l'occorrenza della domanda, si è interessati alla stima della seconda parte del modello nell'equazione (2.25):

$$\begin{aligned} o_t &\sim \text{Be}(p_t) \\ p_t &= \frac{\mu_{a,t}}{\mu_{a,t} + \mu_{b,t}} \\ \mu_{a,t} &= l_{a,t-1} \\ l_{a,t} &= l_{a,t-1}(1 + \alpha_a \epsilon_{a,t}) \\ \mu_{b,t} &= l_{b,t-1} \\ l_{b,t} &= l_{b,t-1}(1 + \alpha_b \epsilon_{b,t}) \end{aligned} \tag{2.30}$$

Tali equazioni possono essere utilizzate direttamente per ottenere previsioni puntuali della probabilità di occorrenza della domanda h passi in avanti:

$$\hat{p}_{t+h|t} = \frac{\hat{l}_{a,t+h|t}}{\hat{l}_{a,t+h|t} + \hat{l}_{b,t+h|t}} \tag{2.31}$$

dove $\hat{l}_{a,t+h|t} = \hat{l}_{a,t}$ e $\hat{l}_{b,t+h|t} = \hat{l}_{b,t}$ sono le parti del modello corrispondenti. Questo si traduce in diverse formulazioni dipendentemente dal caso speciale di iETS$_\text{G}$ in analisi:

- iETS$_\text{F}$: $\hat{p}_{t+h|t} = \hat{p} = \frac{T_1}{T}$, dove T_1 è il numero di osservazioni diverse da zero;
- iETS$_\text{O}$: $\hat{p}_{t+h|t} = \frac{\hat{l}_{a,t}}{\hat{l}_{a,t}+1}$;
- iETS$_\text{I}$: $\hat{p}_{t+h|t} = \frac{1}{\hat{l}_{b,t}+1}$;
- iETS$_\text{D}$: $\hat{p}_{t+h|t} = \min(\hat{l}_{a,t-1}, 1)$.

Si osserva quindi che il calcolo di $\hat{p}_{t+h|t}$ richiede la stima degli errori di previsione un passo in avanti dell'occorrenza della domanda, i quali però non sono osservabili. Per tale ragione, si utilizzano delle proxy che consentono di trasformare le distanze tra il risultato e la probabilità nella scala dei livelli $l_{a,t}$ e $l_{b,t}$ (vedi Appendice A.5.1):

$$\begin{aligned} e_{a,t} &= \frac{u_t}{1 - u_t} - 1 \\ e_{b,t} &= \frac{1 - u_t}{u_t} - 1 \end{aligned}$$

dove $u_t = \frac{1+o_t+\hat{p}_{t|t-1}}{2}$ e $\hat{p}_{t|t-1}$ è il valore atteso un passo in avanti della probabilità di occorrenza.

Stima del modello

Infine, data l'ipotesi di indipendenza tra l'occorrenza e la dimensione della domanda, la media condizionata un passo in avanti di y_t può essere calcolata come:

$$\hat{y}_{t|t-1} = \hat{p}_{t|t-1}\hat{z}_{t|t-1}$$

In sintesi, quindi, la stima del modello iETS$_{\text{G}}$ può essere rappresentata come segue:

$$\begin{aligned}
\hat{y}_{t|t-1} &= \hat{p}_{t|t-1}\hat{z}_{t|t-1} \\
e_{z,t} &= o_t \frac{y_t - \hat{z}_{t|t-1}}{\hat{z}_{t|t-1}} \\
\hat{z}_t &= \hat{l}_{z,t-l} \\
\hat{l}_{z,t} &= \hat{l}_{z,t-1}(1 + \hat{\alpha}_z e_{z,t}) \\
\hat{p}_{t|t-1} &= \frac{\hat{l}_{a,t}}{\hat{l}_{a,t} + \hat{l}_{b,t}} \\
u_t &= \frac{1 + o_t + \hat{p}_{t|t-1}}{2} \\
e_{a,t} &= \frac{u_t}{1 - u_t} - 1 \\
\hat{l}_{a,t} &= \hat{l}_{a,t-1}(1 + \hat{\alpha}_a e_{a,t}) \\
e_{b,t} &= \frac{1 - u_t}{u_t} - 1 \\
\hat{l}_{b,t} &= \hat{l}_{b,t-1}(1 + \hat{\alpha}_b e_{b,t})
\end{aligned} \tag{2.32}$$

Tali equazioni richiedono l'inizializzazione di alcuni dei valori di $\hat{l}_{z,0}$, $\hat{l}_{a,0}$ e $\hat{l}_{b,0}$. L'approccio convenzionale usato per gli ETS prevede di stimarli insieme ai parametri di lisciamentro tramite massimizzazione della funzione di verosimiglianza, che nel nostro caso corrisponde a:

$$\ell(\theta, \sigma_\epsilon^2 \mid \mathbf{Y}) = \sum_{o_t=1} \log f_z\left(z_t \mid l_{z,t-1}\right) + \sum_{o_t=0} \mathcal{H}_z(z_t) + \sum_{o_t=1} \log(\hat{p}_t) + \sum_{o_t=0} \log(1 - \hat{p}_t),$$

dove $\mathbf{Y}$ è il vettore di tutte le osservazioni *in-sample*, θ è il vettore di parametri da stimare (valori iniziali e parametri di lisciamento), $f_z\left(z_t \mid l_{z,t-1}\right)$ è la funzione

di densità della distribuzione ipotizzata per le dimensioni della domanda e $\mathcal{H}_z(z_t)$ è l'entropia differenziale della distribuzione[8], mentre $\hat{p}_t$ è la probabilità stimata dal modello iETS. Per la derivazione della funzione di log-verosimiglianza vedi l'Appendice B di Svetunkov & Boylan (2023).

Le stime prodotte da questi metodi sono consistenti e efficienti, ma le stime $\hat{\alpha}_z$ mostrano una distorsione positiva derivante dall'ipotesi di variazione del livello tra domande non-zero e aumenta con il ridursi della probabilità di occorrenza.

2.5.3 Intervalli di previsione per i modelli iETS

Il calcolo degli intervalli di previsione per il modello iETS utilizza la funzione di ripartizione (CDF, *Cumulative Distribution Function*) della domanda. Ad esempio, se si è interessati a costruire l'intervallo di previsione al 95%, si utilizza la funzione di ripartizione per calcolare i quantili corrispondenti al 97.5% e al 2.5% tramite la CDF.

La funzione di ripartizione della domanda è esplicitata come:

$$F_y(y_{t+h} \leq Q) = \hat{p}_{t+h|t} F_z(z_{t+h} \leq Q) + (1 - \hat{p}_{t+h|t}) \tag{2.33}$$

dove $F_z(z_{t+h} \leq Q)$ è la CDF della dimensione della domanda, z_t, h passi in avanti e Q è il valore del quantile desiderato della distribuzione.

L'unica parte ignota della funzione è $F_z(z_{t+h} \leq Q)$, che può essere calcolata come:

$$F_z(z_{t+h} \leq Q) = \frac{F_y(y_{t+h} \leq Q) - (1 - \hat{p}_{t+h|t})}{\hat{p}_{t+h|t}} \ . \tag{2.34}$$

Nonostante la funzione di distribuzione condizionata h passi in avanti di z_t sia ignota, una soluzione relativamente semplice è quella di ottenere i valori tramite simulazioni delle possibili traiettorie della domanda in base al modello applicato.

Per ottenere dei valori che abbiano senso nell'ambito dei dati di conteggio, si arrotondano per eccesso i quantili risultanti: $y_t = o_t \lceil z_t \rceil$.

[8] Analogamente alla sua controparte discreta, l'entropia differenziale misura la casualità di una variabile, ma, al contrario del caso discreto, non fornisce una descrizione esatta, ma cerca piuttosto di descrivere la variabile in un intervallo di ampiezza unitaria (Orlitsky, 2003). In termini più formali, è definita nel seguente modo (Blahut, 2002): sia X una variabile casuale continua con funzione di densità di probabilità $p(x)$, la sua entropia differenziale è definita come

$$H(X) = -\int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx \ .$$

2.6 Quantile GAM per dati di conteggio

Seguendo Wang et al. (2024), in questa sezione si esamina il modello di regressione quantilica per dati di conteggio con modello additivo generalizzato (GAM), detto modello GAM-QR. L'idea di sfruttare un approccio misto con GAM (Hastie et al., 2009) e regressione quantilica (Koenker & Bassett Jr, 1978) ha mostrato risultati promettenti, risultando vincente nella competizione "Global Energy Forecasting Competition 2014".

La regressione quantilica è ampiamente utilizzata e costituisce un approccio valido per ottenere le stime dei quantili, con l'incorporazione di informazioni esogene. In questa tesi, in particolare, si applica una regressione quantilica lineare per dati di conteggio (Machado & Silva, 2005), incorporando la non linearità delle relazioni utilizzando gli effetti stimati dal GAM (Gaillard et al., 2016) come covariate.

2.6.1 Modello di regressione quantilica per dati di conteggio

La regressione quantilica (QR) è uno strumento inizialmente introdotto da Koenker & Bassett Jr (1978) al fine di migliorare l'efficienza dei modelli lineari nel caso di errori non normali. La stima dei quantili condizionali, inoltre, permette di ottenere la maggior parte dei risultati solitamente raggiungibili solo tramite modelli più strutturati, pur richiedendo ipotesi deboli. In particolare, la QR consente di studiare l'impatto dei regressori su ogni quantile della distribuzione e di produrre degli *statements* probabilistici sui conteggi.

L'applicazione della metodologia a dati di conteggio, tuttavia, presenta un ostacolo significativo: la distribuzione discreta della variabile di risposta Y comporta che la funzione quantile condizionato $Q_Y(\tau|\mathbf{x})$ non possa essere continua nei parametri di interesse. Infatti, la congiunzione di una funzione obiettivo campionaria non differenziabile, quale la *pinball*, con una variabile dipendente discreta, implica che la mancanza di lisciamento della funzione non sia necessariamente compensabile tramite media. Di conseguenza, le usuali strategie basate sull'espansione di Taylor per ottenere la distribuzione asintotica dei quantili condizionali non sono applicabili in questo caso.

Regressione quantilica con dati perturbati

Per evitare l'insorgenza di questo problema, è necessario applicare un certo grado di lisciamento artificiale. In particolare, il metodo proposto da Machado & Silva (2005) prevede l'applicazione di una specifica forma di *jittering*, che somma un rumore $U \sim U[0,1)$ indipendente da Y alla variabile di risposta originaria[9].

La procedura si basa, quindi, sulla costruzione di una variabile continua $Z = Y + U,\ U \sim U[0,1)$, la quale condivide una relazione biunivoca con i quantili della variabile di conteggio ed è usata per fare inferenza[10]. $Q_Z(\tau|\mathbf{x})$, infatti, è una funzione continua che interpola ogni salto della funzione della variabile originale $Q_Y(\tau|\mathbf{x})$ usando un kernel integrato.

Nonostante la misura presentata possieda una funzione di distribuzione continua, questa non è liscia sull'intero supporto: si nota infatti che non ha derivate continue per valori interi di Z. Per evitare di ricadere nel problema presentato sopra, si sfruttano le ipotesi usate per derivare la distribuzione asintotica dello stimatore QR, cioè:

A1. Y è una variabile casuale discreta con supporto in $\mathbb{N}_0$ e $\mathbf{X}$ è un vettore casuale in $\mathbb{R}^k$; la funzione di probabilità condizionata di Y dati $\mathbf{X}$ in $Q_Y(\tau|\mathbf{X})$ è vincolata uniformemente lontana da 0 per quasi ogni realizzazione di $\mathbf{X}$.

A2. I regressori $\mathbf{X}$ sono tali che:

- $E(\mathbf{X}\mathbf{X}')$ è finita e non singolare.
- $\mathbf{X}' = (X_1, ..., X_k)$ può essere partizionato in $(\mathbf{X}^{(d)'}\mathbf{X}^{(c)'})$ con $X_1^{(d)} = 1$ e $X_1^{(c)} \in \mathbb{R}^{k_c},\ 1 \leq k_c \leq k-1$, soddisfacendo $P(\mathbf{X}^{(c)} \in C) = 0$ per qualsiasi sottoinsieme numerabile C di $\mathbb{R}^{k_c}$.

A3. Sia $Z = Y + U$ e $U \sim U[0,1)$ indipendente da $\mathbf{X}$ e Y. Per qualche trasformazione monotona nota, $T(\,\cdot\,;\tau)$, possibilmente dipendente da τ, vale il seguente vincolo sul processo quantile di Z dato $\mathbf{X}$:

$$Q_{T(Z;\tau)}(\tau|\mathbf{x}) = \mathbf{x}'\beta(\tau) \tag{2.35}$$

[9]Il rumore di lisciamento non deve essere necessariamente distribuito come una Uniforme, bensì potrebbe essere generato da una qualsiasi distribuzione continua con supporto in $[0,1)$ e una densità vincolata lontana da 0. La scelta di utilizzare una Uniforme deriva da importanti precedenti storici e dal fatto che consente semplificazioni algebriche e computazionali.

[10]Il Teorema 1 proposto in Machado & Silva (2005) mostra che è possibile usare metodi convenzionali per fare inferenza su $Q_Z(\tau|\mathbf{x})$.

dove $\boldsymbol{\beta} \in$ B, sottospazio compatto di $\mathbb{R}^k$.

La maggior parte di queste ipotesi sono standard nella letteratura relativa alla QR, la sola assunzione mancante è quella di continuità della densità condizionale.

Per costruzione, i punti di discontinuità della densità di Z dato $\mathbf{x}$ si trovano in $\mathbb{N}_0$. Defininendo, quindi, $T^{-1}(\cdot)$ come l'inversa della funzione di trasformazione $T(Z;\tau)$, le ipotesi (A2) e (A3) assicurano che $P(T^{-1}(\mathbf{x}'\boldsymbol{\beta}(\tau)) \in \mathbb{N}_0) = 0$. Di conseguenza, per quasi ogni realizzazione di $\mathbf{X}$, la densità condizionata della variabile al quantile di interesse sarà continua.

Average jittering

Un altro aspetto degno di nota è la dipendenza delle stime dei coefficienti QR con il campione specifico estratto dalla Uniforme [0,1). Poiché tale rumore ha una funzione meramente tecnica, è naturale cercare stime che siano il meno dipendenti possibile dalla specifica realizzazione del campione casuale di U.

Per tale ragione, si considera lo stimatore *average-jittering*, che utilizza la media delle stime applicate per m campioni perturbati $\{y_i + u_i^{(l)}, \mathbf{x}_i\}_{i=1}^n$, $l = 1, ..., m$, costruiti a partire da m campioni casuali indipendenti di dimensione n estratti da una distribuzione Uniforme. Lo stimatore *average jittering* si può definire come

$$\hat{\boldsymbol{\beta}}_m^A(\tau) = \frac{1}{m} \sum_{l=1}^{m} \hat{\boldsymbol{\beta}}_m^{(l)}(\tau)$$

dove $\hat{\boldsymbol{\beta}}_m^{(l)}(\tau)$ è lo stimatore QR basato su $\{y_i + u_i^{(l)}, \mathbf{x}_i\}_{i=1}^n$.

Implementazione

Dato che $Q_Z(\tau|\mathbf{x})$ è limitato inferiormente da τ, rimanendo in linea con le tradizionali ipotesi dei dati di conteggio, si può specificare la seguente rappresentazione parametrica:

$$Q_Z(\tau|\mathbf{x}) = \tau + \exp\left(\mathbf{x}'\boldsymbol{\beta}(\tau)\right) \tag{2.36}$$

Questa specificazione permette importanti semplificazioni matematiche e fornisce un'approssimazione per le funzioni quantile condizionali non note. Inoltre, le simulazioni eseguite in Machado & Silva (2005) suggeriscono la ragionevolezza della specificazione per una vasta classe di modelli.

Usando la funzione quantile nell'equazione (2.36), è possibile stimare $\boldsymbol{\beta}(\tau)$ eseguendo la regressione quantile lineare della trasformazione

$$T(Z;\tau) = \begin{cases} \log(Z-\tau) & Z > \tau \\ \log(\zeta) & Z \leq \tau \end{cases}$$

dove ζ è un valore piccolo e positivo. Questa operazione è realizzabile grazie all'equivarianza rispetto a trasformazioni monotone dei quantili e all'invarianza rispetto alla censura dal basso fino al quantile di interesse.

A seguito della stima di $Q_Z(\tau|\mathbf{x})$ tramite la specificazione esposta nell'equazione (2.36), si sfrutta la Teorema 2 esposto in Machado & Silva (2005) per ottenere $Q_Y(\tau|\mathbf{x})$, il quale afferma $Q_Y(\tau|\mathbf{x}) = \lceil Q_Z(\tau|\mathbf{x}) - 1 \rceil$.

2.6.2 Quantile GAM standard e per dati di conteggio

Il *quantile generalized additive model* (quantGAM) è una procedura multi-step basata sull'utilizzo delle stime delle funzioni di lisciamento del GAM come covariate nel modello di regressione quantile. Si tratta quindi di un approccio misto introdotto durante la "Global Energy Forecasting Competition 2014", in cui è stato usato come base del metodo vincitore (Gaillard et al., 2016).

In questa tesi, si unisce tale approccio con la regressione quantile per dati di conteggio, come descritto sotto, in modo da ottenere un metodo di stima dei quantili per dati di conteggio in grado di incorporare la non linearità delle relazioni.

QuantGAM

Nella presente sezione, si introduce una procedura generica per il calcolo della regressione quantile usando il GAM. La scelta di utilizzare l'approccio multi-step descritto sotto, piuttosto che sostituire semplicemente la funzione obiettivo *pinball* con quella utilizzata dalle splines di lisciamento, deriva dall'aumento della difficoltà di ottimizzazione e dall'insorgenza di problemi numerici.

1. Linearizzazione del problema tramite GAM: si adatta un modello GAM tramite la minimizzazione della funzione obiettivo delle splines cubiche di lisciamento e si stimano gli effetti che catturano le relazioni non lineari tra la media della variabile di interesse al tempo t, Y_t, e le covariate $\mathbf{X}_t = (X_{t,1}, ..., X_{t,p})$: $\mathbf{Z}_t = \left(\hat{f}_1(X_{t,1}), ..., \hat{f}_p(X_{t,p})\right)$.

2. Regressione quantilica: si applica una regressione quantilica lineare per i quantili $\tau \in \{0.01, 0.02, ..., 0.99\}$ usando gli effetti stimati dal GAM al punto

1come covariate per stimare la funzione quantile $\hat{q}_\tau$. A tale scopo, il vettore $\mathbf{Z}_t$ sostituisce il vettore delle covariate $\mathbf{X}_t$ nel problema di minimizzazione convesso

$$\hat{\boldsymbol{\beta}}_\tau \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \sum_{t=1}^{n} \rho_\tau(Y - g(\mathbf{X}))$$

ottenendo così la stima di q_τ.

GAM-QR

La procedura seguita in questa tesi unisce i due metodi descritti sopra e consiste, quindi, in una procedura a 2 passi:

1. **Stima del GAM.** Si stima un modello GAM con la seguente forma:

$$g(E(Y_t|X_{t,1},...,X_{t,P})) = f_1(X_{t,1}) + ... + f_P(X_{t,P}) \tag{2.37}$$

dove Y_t denota la domanda al tempo t, mentre $X_{t,1},...,X_{t,P}$ denotano le covariate al tempo t. Si utilizza una funzione legame logaritmica e come funzioni di lisciamento si usano rispettivamente le splines cubiche per le variabili continue e la funzione identità per le variabili categoriali.

2. **Stima della regressione quantilica per dati di conteggio.** Si applica la regressione quantilica per dati di conteggio con le componenti $\hat{f}_1(\cdot),...,\hat{f}_P(\cdot)$ stimate al punto (1) per i livelli dei quantili $\tau \in \{0.01, 0.02, ..., 0.99\}$. Si ottiene così la seguente formulazione della previsione dei quantili:

$$Q_{T(Z,\tau)}(\tau, \hat{\boldsymbol{f}}) = \hat{\boldsymbol{f}}'\boldsymbol{\beta}(\boldsymbol{\tau}) \tag{2.38}$$

dove $\hat{\boldsymbol{f}} = (\hat{f}_1(\cdot),...,\hat{f}_P(\cdot))$ e $T(Z,\tau)$ è la funzione di trasformazione

$$T(Z,\tau) = \begin{cases} \log(Z-\tau) & Z > \tau \\ \log(10^{-5}) & Z \leq \tau \end{cases}$$

dove $Z = Y + U,\ U \sim U[0,1)$, come specificato nella sezione 2.6.1.

2.7 Modelli di distribuzione con media *damped*

Seguendo il lavoro di Wang et al. (2024), si considerano due distribuzioni con media *damped*, in modo da ridurre l'effetto di fluttuazioni improvvise o eccessive

nelle previsioni e stabilizzare il modello nel lungo termine. Tale specificazione della media, infatti, integra nei modelli un meccanismo che riduce gradualmente l'impatto delle deviazioni dalla media storica nel tempo e che, in termini formali, può essere visto come un modello autoregressivo stazionario per la media:

$$\mu_t = (1 - \phi - \alpha)\mu + \phi\mu_{t-1} + \alpha y_{t-1} \ , \tag{2.39}$$

sotto i vincoli $\mu > 0$, $\phi > 0$, $\alpha > 0$ e $\alpha + \phi < 1$, dove μ e μ_{t-1} indicano rispettivamente la media di lungo e di breve termine al tempo $t-1$[11]. Avvalersi di una media variabile temporalmente, consente di ottenere previsioni probabilistiche adeguate durante molteplici periodi di tempo per serie non stazionarie. In questo modo, infatti, si consente alla media della distribuzione di variare casualmente nel tempo, riflettendo l'effetto di possibili cambiamenti strutturali (Snyder et al., 2012).

In particolare, data la natura dei dati, si considerano le distribuzioni Poisson, $Y_t \sim Poi(\lambda_t)$:

$$\frac{\lambda_t^{y_t}}{y_t!} \exp(-\lambda_t) \ , \ \lambda_t > 0 \ , \tag{2.40}$$

e Binomiale Negativa, $Y_t \sim NB(a_t, b)$:

$$\frac{\Gamma(a_t + y_t)}{\Gamma(a_t) y_t!} \left(\frac{b}{1+b}\right)^{a_t} \left(\frac{1}{1+b}\right)^{y_t} \ , \ a > 0, \ b > 0 \tag{2.41}$$

dove i parametri λ_t e a_t variano nel tempo e sono legati alla media *damped* tramite le seguenti relazioni:

$$\lambda_t = \mu_t \tag{2.42}$$

$$a_t = b\mu_t \ . \tag{2.43}$$

Tutti i parametri ignoti specificati sopra sono stimati usando il metodo di massima verosimiglianza. Per ogni serie, si considera quindi la distribuzione congiunta $f(y_1, ..., y_T|\mu_1, \theta)$, dove θ contiene tutti i parametri non noti oltre a μ_1. Usando il principio di induzione in congiunzione con la legge di probabilità condizionata $P(A, B) = P(B|A)P(A)$, si ottiene:

$$f(y_1, ..., y_T|I_{T-1}) = \prod_{t=1}^{T} f(y_t|I_{t-1}) \tag{2.44}$$

[11]La scelta di considerare un solo *lag* è guidata dal fatto che la complessità derivante dall'inserimento di ritardi aggiuntivi probabilmente sovrasterebbe il guadagno ottenuto dall'operazione (Snyder et al., 2012).

dove $I_{t-1} = \{\mu_t, \theta\}$, $t = 1, ..., T$, e le distribuzioni univariate sono successioni di distribuzioni di previsione un passo in avanti. Le medie delle distribuzioni univariate sono ricavate usando la relazione dinamica appropriata, espressa nell'equazione (2.39).

Le distribuzioni previste sono poi ottenute tramite simulazioni, sfruttando le relazioni di ricorsione di primo ordine:

$$f(y_{T+1}, ..., y_{T+h}|I_T) = \prod_{t=T+1}^{T+h} f(y_t|I_{t-1}) \ . \tag{2.45}$$

2.8 Bootstrap WSS

Il bootstrap è un approccio non parametrico utilizzato per prevedere la domanda totale durante un periodo di interesse. Questo metodo genera un'elevata quantità di dati estraendo ripetutamente campioni casuali dai dati storici della domanda, permettendo così di costruire una distribuzione della domanda futura o di determinare la posizione dell'inventario necessaria per garantire un certo *service rate* (Zhou & Viswanathan, 2011). L'obiettivo di questo metodo è quindi quello di stimare l'intera distribuzione della domanda nel periodo di interesse, ossia: $\sum_{t=T+1}^{T+h} y_t$, dove y_t è la domanda osservata in t e h è il periodo di interesse fissato.

In questo lavoro, si considera un'estensione del bootstrap ideata per gestire correttamente la domanda intermittente, ovvero il metodo Willemain-Smart-Schwarz (WSS, Willemain et al. (2004)). L'utilizzo di questa estensione mira a risolvere alcuni dei problemi riscontrati nell'utilizzo del metodo di boostrap semplice (Efron, 1979) per la previsione della domanda intermittente. Quest'ultimo, infatti, crea pseudo-dati tramite il campionamento con reinserimento di singole osservazioni, ignorando la presenza di autocorrelazione e riproducendo valori storici. La procedura bootstrap WSS risolve queste limitazioni rispettivamente tramite l'applicazione di un modello Markoviano a 2 stati e di *jittering*.

Di seguito si presentano i passi della procedura proposta da Willemain et al. (2004).

1. Stima della probabilità di transizione della domanda tra due stati (presenza e assenza di domanda) della serie, usando le probabilità condizionate dei due stati della domanda storica (ad esempio, la probabilità di osservare una domanda positiva al tempo *t*, dato che in *t-1* non vi era).

2. Applicazione di un modello Markoviano per generare una sequenza di 0 e 1 durante l'orizzonte previsivo, condizionatamente all'ultima domanda osservata, dove 0 indica l'assenza di domanda e 1 la presenza. In altre parole, in questo passaggio si modella l'autocorrelazione tramite un modello Markoviano di 1° ordine con 2 stati.

3. Sostituzione dei *marker* di stato 1 con un valore numerico campionato casualmente con reinserimento dall'insieme di domande positive osservate.

4. *Jittering* dei valori della domanda positiva: si aggiunge una variazione casuale ai valori positivi della domanda, rendendoli quindi vicini ma non identici a quelli osservati. In termini più formali, si svolge la seguente operazione:

$$jittered_t = 1 + INT\{y^* + Z\sqrt{y^*}\}$$

$$\text{if } jittered_t \leq 0, \text{ then } jittered = y^*$$

   dove y^* sono i valori della domanda storica selezionati casualmente e Z è una normale standard.

   In alcune circostanze il *jittering* può creare previsioni che non sono fisicamente significative, ad esempio, se le birre sono vendute in lotti da 6, 12 o 24 lattine, usando il *jittering* si potrebbe ottenere un valore pari a 13. In questi casi, può essere ragionevole saltare questo step.

5. Somma dei valori previsti sull'orizzonte temporale per ottenere un valore predetto della domanda.

6. Iterazione dei passi da 2 a 5 molteplici volte.

7. Ordinare e ricavare la distribuzione risultante.

Dato che in questo lavoro si è interessati a generare previsioni probabilistiche si saltano i passi 5 e 7 e si salvano i risultati di ogni iterazione della procedura. Per ogni osservazione futura, si calcola poi la distribuzione cumulativa empirica in base ai dati generati dalla procedura.

Capitolo 3

Combinazione di previsioni probabilistiche

Combinare i risultati di diversi processi previsivi è una pratica ampiamente utilizzata per migliorare l'accuratezza delle previsioni. Questa procedura permette di integrare informazioni raccolte da diverse fonti, risultando particolarmente utile quando i singoli modelli e metodi in analisi contengono informazioni parziali non completamente sovrapponibili. Di conseguenza, elimina la necessità di identificare un unico modello "migliore", una strategia alternativa spesso sub-ottimale. È, infatti, poco verosimile ipotizzare che le serie storiche osservate siano generate da un processo semplice, con una forma funzionale specifica. Queste serie sono spesso caratterizzate da un processo generatore dei dati complesso, con trend che variano nel tempo, cambiamenti stagionali e rotture strutturali. Quest'ultimo aspetto costituisce una motivazione comune per l'utilizzo di combinazioni di previsioni: in presenza di rotture strutturali o altre instabilità, la combinazione di previsioni ottenute da modelli con diversi gradi di non corretta specificazione e adattabilità può mitigare il problema (Wang et al., 2023).

Nell'ambito delle previsioni probabilistiche, i risultati individuali possono essere elicitati in modi diversi, ad esempio come densità, quantili e intervalli di previsione. Seguendo Kolassa (2016), si considera l'utilizzo di distribuzioni previsive come più appropriato per il supporto alle decisioni complesse riguardanti la gestione dell'inventario. L'uso dei quantili, infatti, riduce la quantità di informazione contenuta e potrebbe introdurre supporto non intero, violando le proprietà della domanda intermittente.

3.1 *Linear opinion pool*

Un approccio comune per la combinazione di previsioni è il "*linear opinion pool*" (Hall & Mitchell, 2007; Petropoulos et al., 2022). Si considerano N previsioni individuali ottenute da modelli diversi e caratterizzate da funzioni di ripartizione condizionate al tempo $T+h$, denotate come $F_i^{(T+h)}(\, y \mid I_T \,)\ ,\ i = 1, ..., N$, dove y è la serie di interesse e I_T è l'informazione disponibile al tempo T. Le previsioni combinate possono essere ottenute tramite il calcolo di una mistura finita:

$$F_{\text{comb}}^{(T+h)}(\, y \mid I_T \,) = \sum_{i=1}^{N} w_{\,T+h|T,i}\, F_i^{(T+h)}(y|I_T) \tag{3.1}$$

dove $w_{\,T+h|T,i}$ è il peso associato all'i-esima previsione probabilistica individuale $F_i^{(T+h)}(y|I_T)$. Per garantire che le previsioni combinate mantengano le proprietà delle distribuzioni di probabilità, i pesi $w_{\,T+h|T,i}$ sono vincolati ad essere non negativi e sommare ad uno. Per semplicità, si considerano pesi costanti nel tempo. Data la formula nell'equazione (3.1), si possono ricavare le seguenti equazioni dei momenti:

$$E[f_{\text{comb}}^{(t)}(y_t)] = m_t* = \sum_{i=1}^{N} w_{it} m_{it} \tag{3.2}$$

$$Var[f_{\text{comb}}^{(t)}(y_t)] = \sum_{i=1}^{N} w_{it} v_{it} + \sum_{i=1}^{N} w_{it}\{m_{it} - m_t*\} \tag{3.3}$$

dove $m_{it} = \int_{-\infty}^{\infty} y_t f_{it} dy_t$ è la media della previsione del metodo i al tempo t e $v_{it} = \int_{-\infty}^{\infty} (y_t - m_{it})^2 f_{it}(y_t) dy_t$ è la relativa varianza. Si ha quindi che la varianza della combinazione è la somma dell'incertezza media inviduale ("*within*" al modello) e di quella tra i modelli ("*between*" i modelli). Questo risultato è opposto a quello che si ottiene nel caso di previsioni puntuali, in cui si osserva che la varianza della combinazione è minore rispetto a quella dei metodi individuali. L'incremento della varianza, però, non è necessariamente deleterio, in quanto potrebbe comunque portare a *performance* migliori rispetto alle sue componenti (Hall & Mitchell, 2007). Queste considerazioni spiegano la popolarità del metodo nel caso di previsioni caratterizzate da una varianza minore di quella osservata (Gneiting & Katzfuss, 2014).

È rilevante notare che il guadagno derivante dall'utilizzo di combinazioni di previsioni dipende non solo dalla bontà dei singoli elementi, ma anche dai pesi di combinazione assegnati a ciascuno di essi. Di conseguenza, seguendo Wang et al.

(2024), verranno esaminati tre approcci per il calcolo dei pesi: semplici, basati sull'ottimizzazione di funzioni di punteggio e basati sui costi di inventario.

3.2 Pesi semplici

Il modo più semplice per calcolare i pesi è il Simple Average (SA), il quale pone peso identico alle previsioni di ciascun modello, pari a $1/N$. Nonostante la semplicità del metodo, questo porta a risultati robusti e spesso domina empiricamente schemi di pesi più sofisticati, che dovrebbero essere asintoticamente superiori (Wang et al., 2023).

Un altro schema di pesi semplici utilizzato è il punteggio logaritmico, il quale si basa sulla *performance* storica dei metodi ed è calcolato come:

$$\begin{aligned} \text{logscore}_i &= \frac{1}{h} \sum_{t=T-h+1}^{T} \log(f_{\text{comb}}^{(t)}(y_t|I_{T-h})) \\ w_{T+h|T,i} &= \frac{1}{\text{logscore}_i} / \sum_{j=1}^{N} \left(\frac{1}{\text{logscore}_j} \right) \end{aligned} \tag{3.4}$$

dove logscore_i è il punteggio logaritmico medio per la previsione individuale i durante gli h periodi precedenti e $f_{\text{comb}}^{(t)}(y_t|I_{T-h})$ è la funzione di probabilità (PMF) prevista della combinazione.

Visto che i dati in esame sono costituiti da valori interi con distribuzione asimmetrica, si arrotondano i valori prodotti e si considera anche la mediana come metodo di combinazione.

3.3 Pesi ottimali basati funzioni di punteggio

Il calcolo dei pesi ottimali in base a funzioni di punteggio è un approccio *data-driven*, utile per la combinazione diretta di previsioni di densità utilizzando una *linear pooling* pesata (Hall & Mitchell, 2007). Tale metodo valuta la qualità delle densità tramite l'assegnazione di un punteggio numerico, basato sulla previsione e la successiva realizzazione della variabile.

In questa tesi, si esamina l'ottimizzazione di quattro funzioni di punteggio: il punteggio logaritmico (log *scoring*, Hall & Mitchell (2007)) e la sua variante *Censored Likelihood score* (CL *scoring*), il punteggio di Brier (Kolassa, 2016) e il punteggio Discrete Ranked Probability Score (DRPS, Snyder et al. (2012)).

3.3.1 Punteggio logaritmico e CL

Log *scoring*

La funzione di punteggio logaritmico (Hall & Mitchell, 2007) mira ad identificare l'insieme di pesi che forniscono la previsione più accurata in senso statistico, ossia tali da minimizzare il criterio di informazione di Kullback-Leibler (KLIC) del *linear opinion pooling*

$$f^{(t)}_{\text{comb}}(y_t|I_{T-h+1}) = \sum_{i=1}^{n} w_i f^{(t)}_i(y_t|I_{T-h+1}) \ . \tag{3.5}$$

Il criterio di informazione KL misura la distanza tra la vera densità $f^{(t)}(y_t)$ e la densità previsiva combinata $f^{(t)}_{\text{comb}}(y_t|I_{T-h+1})$, $i = 1, ..., n$, ed è espresso come:

$$KLIC_t = E[\ln f^{(t)}(y_t) - \ln f^{(t)}_{\text{comb}}(y_t|I_{T-h+1})] \ . \tag{3.6}$$

Questo criterio risulta particolarmente conveniente, in quanto, al contrario di altre misure, non richiede la specificazione o la stima né della vera densità (non nota) della variabile di interesse, né della densità della trasformata integrale di probabilità (PIT)[1].

In base alla Definizione 1 data da Hall & Mitchell (2007), la densità previsiva combinata ottimale è definita come

$$f^{(t)}_{\text{comb}}(y_t|I_{T-h+1}) = \sum_{i=1}^{n} w_i f^{(t)}_i(y_t|I_{T-h+1})$$

dove il vettore di pesi ottimali $\boldsymbol{w} = (w_1, ..., w_N)$ minimizza il criterio nell'equazione (3.6).

Sotto alcune condizioni di regolarità, il KLIC può essere stimato in maniera consistente come la media dell'informazione campionaria sulla vera densità e quella predetta, la cui minimizzazione è ottenibile come segue:

$$\boldsymbol{w}_{T+h|T} = \arg\max_{\boldsymbol{w}} \frac{1}{h} \sum_{t=T-h+1}^{T} \log(f^{(t)}_{\text{comb}}(y_t|I_{T-h+1})) \tag{3.7}$$

dove $\boldsymbol{w}_{T+h|T} = (w_{T+h|T,1}, ..., w_{T+h|T,n})'$ sono i pesi per la combinazione delle previsioni. In altre parole, i pesi ottimali sono individuati via ricerca numerica sull'insieme di valori che massimizzano il punteggio logaritmico medio della densità previsiva combinata. Il problema di individuazione dei pesi ottimali a partire

[1] Trattata in dettaglio nella sezione 4.1.1.

da dati empirici si riduce, quindi, alla massimizzazione della funzione di costo concava definita nell'equazione (3.7), nota come punteggio logaritmico predittivo.

CL *scoring*

Data l'importanza dei quantili alti nella gestione di inventario, si considera una variante del punteggio logaritmico, utile nel caso in cui ci si voglia concentrare su un'area specifica della distribuzione probabilistica: la funzione di punteggio di verosimiglianza censurata.

Si definisca A come l'area di interesse. Allora il punteggio CL è denotato come:

$$\begin{aligned}\mathrm{CL}(f_{\mathrm{comb}}^{(t+h)})(y|T_t) =& \mathbf{1}[y_{t+h} \in A_{t+h}] \log(f_{\mathrm{comb}}^{(t+h)}(y_{t+h}|I_t)) \\ &+ \mathbf{1}[y_{t+h} \in A_{t+h}^C] \log\left(\int_{A_{t+h}^C} f_{\mathrm{comb}}^{(t+h)}(y|I_t)dy\right)\end{aligned} \tag{3.8}$$

dove A_{t+h}^C è il complemento di A_{t+h} e $\mathbf{1}(\cdot)$ è la funzione indicatrice. Si può quindi vedere che la formula nell'equazione (3.8) cattura la forma della distribuzione nell'area A_{t+h}, che si considera essere definita come i valori di y_{t+h} superiori al 90% quantile delle distribuzioni. I pesi ottimali sono individuati nel seguente modo:

$$\boldsymbol{w}_{T+h|T} = \arg\max_{\boldsymbol{w}} \frac{1}{h} \sum_{t=T-h+1}^{T} \mathrm{CL}(f_{\mathrm{comb}}^{(t)}(y_t|I_{T-h+1})) \ . \tag{3.9}$$

Ottimizzazione

Computazionalmente, ci si avvale di un semplice algoritmo iterativo basato sulla strategia MM (*"minimization-maximizazion"*), che considera massimizzazioni successive di funzioni di costo ausiliarie "surrogate" per ogni iterazione (Conflitti et al., 2015). Le operazioni sono descritte in funzione del punteggio logaritmico, ma possono essere facilmente riadattate per essere utilizzate per il CL *score*.

In termini più formali, si definisce una matrice $\hat{P}_{h\times N}$ con elementi non negativi $\hat{P}_{ti} = f_i^{(t)}(y_t|I_{T-h+1})$, dove h è l'orizzonte temporale di interesse e N è il numero di previsioni individuali. Allora, la funzione obiettivo da massimizzare nell'equazione (3.7) può essere riscritta come:

$$\Phi(\mathbf{w}) = \frac{1}{h} \sum_{t=T-h+1}^{T} \ln(\hat{P}\mathbf{w})_t \ . \tag{3.10}$$

Al fine di introdurre i vincoli di non negatività e di somma unitaria dei pesi, si introduce un moltiplicatore di Lagrange λ:

$$\Phi(\mathbf{w}) = \frac{1}{h} \sum_{t=T-h+1}^{T} \ln(\hat{P}\mathbf{w})_t - \lambda \sum_{i=1}^{N} w_i \tag{3.11}$$

Si considera la seguente funzione di costo ausiliaria:

$$\Psi_\lambda(\mathbf{w}; \mathbf{a}) = \frac{1}{h} \sum_{t=T-h+1}^{T} \sum_{i=1}^{N} \frac{\hat{P}_{ti} a_i}{\sum_{l=1}^{N} \hat{P}_{tl} a_l} \ln\left(\frac{w_i}{a_i} \sum_{l=1}^{N} \hat{P}_{tl} a_l \right) - \lambda \sum_{i=1}^{N} w_i \tag{3.12}$$

dove $\mathbf{a}$ è un vettore di pesi arbitrari (vedi Appendice A.6). Allora l'algoritmo iterativo MM si può definire come:

$$\mathbf{w}_\lambda^{(k)} = \arg\max_{\mathbf{w}} \Psi_\lambda(\mathbf{w}; \mathbf{w}_\lambda^{(k)}) \tag{3.13}$$

che produce un incremento monotonico di Φ_λ, ossia $\Phi_\lambda(\mathbf{w}_\lambda^{(k+1)}) \geq \Phi_\lambda(\mathbf{w}_\lambda^{(k)})$. La funzione surrogata definita nell'equazione (3.12) è più semplice da massimizzare rispetto a Φ_λ, in quanto separabile: la sua formula consiste, infatti, di una somma di N termini, ognuno dei quali dipende da un solo peso.

Ponendo le derivate di tale funzione rispetto ad ogni peso pari a zero, si ottiene che la sua massimizzazione è data da $w_{\lambda,i} = \frac{1}{\lambda h} \sum_{t=1}^{h} b_{ti}$, $b_{ti} = \frac{\hat{P}_{ti} a_i}{\sum_{l=1}^{N} \hat{P}_{tl} a_l}$. In base al vincolo di somma unitaria dei pesi, si ha quindi che $\lambda = 1$. Di conseguenza, l'iterazione *k+1*-esima del peso associato all'i-esimo metodo di previsione individuale dell'algoritmo in equazione (3.13) può essere scritto come segue:

$$w_{T+h|T,i}^{(k+1)} = w_{T+h|T,i}^{(k)} \; \frac{1}{h} \sum_{t=T-h+1}^{T} \frac{\hat{f}_i(y_t|I_{T-h})}{\sum_{j=1}^{N} \hat{f}_j(y_t|I_{T-h}) w_{T+h|T,j}^{(k)}}, \tag{3.14}$$

in cui i vincoli di non negatività dei pesi sono soddisfatti automaticamente ad ogni iterazione. Per il punteggio CL, si sostituisce $\hat{f}_i(y_t|I_{T-h})$ con $\mathbf{1}[y_t \in A_t]\hat{f}_i(y_t|I_{T-h}))+$ $\mathbf{1}[y_t \in A_t^C] \int_{A_t^C} \hat{f}_i(y|I_t)dy$.

L'algoritmo può essere terminato tramite un appropriato criterio di stop, ad esempio quando la differenza tra le componenti di due iterazioni successive non supera un certo livello di tolleranza predefinito.

3.3.2 Punteggio di Brier e DRPS

Le funzioni riguardanti il punteggio di Brier e il DRPS sono due metriche per la valutazione delle previsioni probabilistiche che si basano sulla distanza tra gli

esiti reali e quelli previsti. Entrambi gli *score* vengono usati per il calcolo dei pesi tramite la loro minimizzazione.

Brier *score*

Il punteggio di Brier è stato ideato da Brier (1950) come metodo di verifica per le previsioni meteorologiche, ed è stato introdotto da Kolassa (2016) come metrica di valutazione delle previsioni per la domanda intermittente. La funzione di *scoring* prende la seguente forma:

$$\text{Brier}(f_{\text{comb}}^{(t+h)}(y|I_t)) = -2f_{\text{comb}}^{(t+h)}(y_{t+h}|I_t) + \sum_{k=0}^{\infty} f_{\text{comb}}^{(t+h)}(k|I_t)^2 \tag{3.15}$$

dove $f_{\text{comb}}^{(t+h)}(k|I_t)$ è il valore della PMF $f_{\text{comb}}^{(t+h)}(y|I_t)$ al valore k. L'individuazione dei pesi tramite minimizzazione del punteggio di Brier consiste, quindi, in un problema di ottimizzazione di una funzione obiettivo quadratica:

$$\boldsymbol{w}_{T+h|T} = \arg\min_{\boldsymbol{w}} \frac{1}{h} \sum_{t=T-h+1}^{T} \text{Brier}(f_{\text{comb}}^{(t)}(y_t|I_{T-h+1})) \ . \tag{3.16}$$

DRPS

Il DRPS è una delle metriche utilizzate da Snyder et al. (2012) per la valutazione della distribuzione della domanda intermittente e si basa sulla norma-L_2 per misurare la distanza tra due distribuzioni:

$$L_2(y, F) = \sum_{k=0}^{\infty} (\hat{F}_y(k) - F_y(k))^2$$

dove $\hat{F}(k)$ è un'approssimazione campionaria della funzione di distribuzione F, che, nel caso in cui $F(y)$ è discreta e si ha una sola osservazione, si definisce come:

$$\hat{F}_y(k) = \begin{cases} 0 & \text{se } y > k \\ 1 & \text{se } y \leq k \end{cases} .$$

Si denota, quindi, la funzione da minimizzare come segue:

$$\text{DRPS}(F_{\text{comb}}^{(t+h)}(y|I_t), y_{t+h}) = \sum_{k=0}^{\infty} (F_{\text{comb}}^{(t+h)}(k|I_t) - \mathbf{1}[y_{t+h} \leq k])^2 \tag{3.17}$$

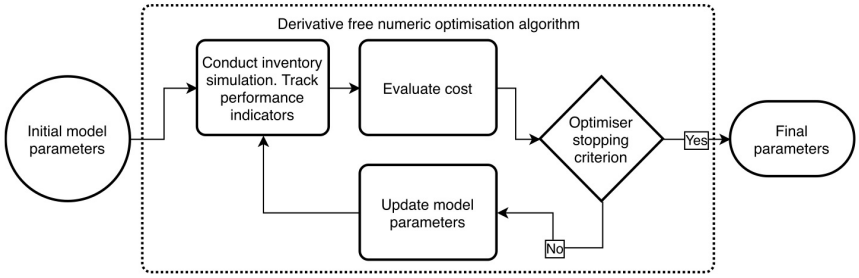

Figura 3.1: *Flowchart* dell'approccio di stima dei parametri proposta da Kourentzes et al. (2020), (Kourentzes et al., 2020).

dove $F_{\text{comb}}^{(t+h)}(k|I_{T-h+1})$ è il valore della probabilità cumulativa $F_{\text{comb}}^{(t+h)}(y_t|I_{T-h+1})$ in k. I pesi che minimizzano tale funzione sono:

$$\boldsymbol{w}_{T+h|T} = \arg\min_{\boldsymbol{w}} \frac{1}{h} \sum_{t=T-h+1}^{T} \text{DRPS}(f_{\text{comb}}^{(t)}(y_t|I_{T-h+1})) \ . \tag{3.18}$$

3.4 Pesi ottimali basati sui costi di inventario

L'ultimo metodo utilizzato per il calcolo dei pesi si basa sulla simulazione dei costi di inventario, sottolineando l'importanza di considerare le implicazioni pratiche delle previsioni. Infatti, la previsione finalizzata all'ottimizzazione dei modelli in termini di accuratezza previsiva, non implica necessariamente dei miglioramenti nelle *performance* di inventario. Tale *modus operandi* si concentra solo sull'ottimizzazione della previsione a livello statistico, senza considerare l'influenza della posizione di inventario[2] e delle politiche di ordinamento sulle decisioni aziendali. Per risolvere tale problema, si producono delle simulazioni di inventario e si calcolano i pesi in modo da ottimizzare delle metriche di *performance* dell'inventario (Kourentzes et al., 2020), seguendo il procedimento illustrato in Figura 3.1.

A tale scopo, si fissa un *customer service level* (CSL) target[3], τ, come la probabilità designata della distribuzione e si calcolano i corrispondenti quantili come *holdings* di inventario. I costi imputabili agli *holdings* e alle vendite perse sono valutati in base alla differenza tra le previsioni e la domanda reale, e determinano la funzione di costo obiettivo:

[2]Per posizione di inventario si intende il livello di scorte totale, comprendendo quelle presenti in magazzino e in transito.

[3]vedi Appendice A.7.

$$\boldsymbol{w}_{T+h|T} = \arg\min_{\boldsymbol{w}} \text{ cost} = \arg\min_{\boldsymbol{w}} [c_1 * \text{stock} + c_2 * \text{lostsales}] \tag{3.19}$$

dove

$$\text{stock} = \frac{1}{h} \sum_{t=T-h+1}^{T} [\hat{q}_t(\tau) - x_t]_+ \text{ e } \text{lostsales} = \frac{1}{h} \sum_{t=T-h+1}^{T} [x_t - \hat{q}_t(\tau)]_+$$

in cui x_t sono le vendite reali al tempo t, c_1 e c_2 sono i costi marginali, e $\hat{q}_t(\tau)$ è il quantile stimato al CSL τ, tale che $F_{\text{comb}}(\hat{q}_t(\tau)) \geq \tau$ e $F_{\text{comb}}(\hat{q}_t(\tau) - 1) < \tau$.

Si considera un classico *newsvendor problem* con distribuzione della domanda non nota. In altre parole, (Huber et al., 2019) si ipotizza il caso di un'azienda che vende articoli deperibili e si considera il problema di scegliere la quantità di prodotto da acquistare (prima dell'inizio della stagione di vendita), in modo tale da minimizzare i costi di *overage*, ossia di eccedenza (c_2), e di *underage*, derivanti dalla mancanza di stock (c_1). In questo contesto, il CSL target può essere definito come $\tau = \frac{c_2}{c_1+c_2}$ e permette di simulare la variazione delle proporzioni dei costi marginali corrispondenti a diversi livelli obiettivo. Seguendo Wang et al. (2024), si considerano tre combinazioni dei costi marginali (c_1, c_2): (1, 4), (1, 9) e (1, 19), che corrispondono rispettivamente a CSL target pari a 80%, 90% e 95%.

Il problema di minimizzazione presentato nell'equazione (3.19) potrebbe non essere convesso, a causa della complessità della funzione di inventario presentata: dato che i pesi sono contenuti nella distribuzione combinata (discreta), la funzione di costo è discontinua. Al fine di risolvere il problema di ottimizzazione, si applica, quindi, il *Particle Swarm Optimization* (PSO, Kennedy & Eberhart (1995)).

3.4.1 Breve *overview* del PSO

Il PSO è un algoritmo di ottimizzazione per funzioni non lineari che sfrutta la metodologia *swarm* introdotto da Kennedy & Eberhart (1995).

Dalla sua ideazione, questo ottimizzatore ha attirato l'attenzione degli studiosi grazie ai vantaggi che apporta, quali una struttura semplice, alta velocità di convergenza e ampio raggio di ricerca. In particolare, è risultato utile per problemi di ottimizzazione globale, specialmente per funzioni ad alta dimesionalità, multi-picco e discontinue, oltre ad avere una forte propensione per problemi di ottimizzazione combinatoria (Zhang et al., 2019).

L'idea del PSO parte dallo studio dei comportamenti sociali degli animali, i quali sfruttano la formazione di *swarms*, appunto, in cui ogni membro dello sciame cambia continuamente il suo percorso di ricerca in base alle esperienze fatte da lui o da altri membri per trovare cibo. L'algoritmo, quindi, simula un gruppo di particelle che si muovono indipendentemente nello spazio di ricerca, condividendo tra loro le informazioni necessarie per l'individuazione della soluzione ottimale. L'idea principale dell'algoritmo è basata su due metodologie: l'*evolutionary algorithm* e l'*artificial life*. Il suo legame con il primo deriva dalla ricerca simultanea in grandi regioni nello spazio delle soluzioni della funzione da ottimizzare, mentre il secondo deriva dallo studio di sistemi artificiali con caratteristiche "di vita" del PSO (Wang et al., 2018).

Sinteticamente, l'algoritmo inizializza i parametri della funzione (dette particelle), la direzione dei loro spostamenti e la relativa velocità con valori casuali. Iterativamente, ogni parametro si sposta nello spazio di ricerca a lui designato, il quale è influenzato dall'esperienza personale (ossia, la migliore posizione individuata dalla particella stessa) e da quella collettiva (ovvero, la posizione ottimale individuata dall'intero sciame), aggiornando la velocità e la direzione. Ad ogni passo, il PSO registra e memorizza il minimo storico. L'algoritmo si ferma in base ad un criterio di arresto, come una certa soglia di miglioramento del minimo o un numero massimo di passi.

Capitolo 4

Metodi per la valutazione delle previsioni probabilistiche

Il processo di valutazione delle previsioni nell'ambito probabilistico differisce da quello classico puntuale. Mentre quest'ultimo si concentra sull'accuratezza di un singolo valore, per valutare le densità previsive è necessario valutare l'intera funzione predittiva. Per tale ragione, solitamente la valutazione di questi risultati è misurata in termini di calibrazione, la quale misura la coerenza tra la densità predetta e quella osservata, tramite l'analisi grafica della trasformata integrale di probabilità (PIT, Willemain et al. (2004)). Considerando la natura discreta della domanda intermittente, si utilizza una variante normalizzata della trasformata, detta rPIT (Kolassa, 2016). Seguendo il lavoro di Wang et al. (2024), si affianca la rappresentazione grafica con il test Kolmogorov-Smirnov (KS, Massey Jr (1951)), che fornisce una misura sintetica della bontà di adattamento della densità ai dati.

Nonostante la trasformata integrale di probabilità sia un buon indicatore per la valutazione delle previsioni di probabilità, non è sufficiente per identificare il metodo previsivo "migliore" (Boylan & Syntetos, 2021). Per tale ragione, si considerano anche altri due aspetti: la *sharpness* e la *performance* di inventario. Il primo indica la dispersione della densità predetta attorno ai valori centrali della previsione, ed è misurata tramite regole di punteggio corrette (*proper scoring rules*), ossia di punteggi numerici ottimizzati quando la funzione prevista coincide con quella reale. Seguendo Wang et al. (2024), si utilizzano il punteggio logaritmico, quello di Brier e il DRPS descritti nella sezione 3.3. Il secondo, invece, analizza i costi simulati come in sezione 3.4 derivanti dalle previsioni.

4.1 Calibrazione

La calibrazione è uno strumento che valuta l'accuratezza di una previsione probabilistica, confrontandola con le realizzazioni delle serie. In altre parole, una previsione di densità è calibrata se è coerente con i dati osservati (Kolassa, 2016).

La nozione di calibrazione più comunemente usata nelle applicazioni è quella probabilistica[1], in base alla quale un sistema di previsioni è calibrato se la trasformata integrale di probabilità[2] (PIT) segue una distribuzione Uniforme in [0,1] (Petropoulos et al., 2022). Nel caso in cui la distribuzione calcolata sia corretta, infatti, la rispettiva funzione di ripartizione assume valori tra zero ed uno nei valori osservati e ci si aspetta un'allocazione approssimativamente uguale per ogni *item*.

In termini più formali, si considera temporaneamente il contesto delle variabili casuali continue e si definiscono $(\hat{F}_t)_{t=1,2,...}$ e $(F_t)_{t=1,2,...}$ come sequenze di funzioni di ripartizione (CDF) continue e strettamente crescenti, relative rispettivamente le previsioni probabilistiche e il vero processo generatore dei dati. La sequenza $(\hat{F}_t)_{t=1,2...}$ è probabilisticamente calibrata relativamente a $(F_t)_{t=1,2,...}$ se

$$\frac{1}{T}\sum_{t=1} F_t^T \circ \hat{F}_t^{-1}(p) \to p \ , \qquad p \in (0,1)$$

dove p indica un certo livello di probabilità e $F_t^T \circ \hat{F}_t^{-1}(p)$ denota la probabilità reale associata al valore previsto[3] (Gneiting et al., 2007).

La valutazione della calibrazione delle previsioni si avvale di due strumenti: la rPIT (Kolassa, 2016) e la statistica D del test KS (Massey Jr, 1951).

4.1.1 Randomized Probability Integral Transform (rPIT)

La rPIT è un metodo per valutare la calibrazione di previsioni probabilistiche per variabili discrete, suggerito da Kolassa (2016) come approccio alternativo a quello proposto da Willemain et al. (2004) per variabili continue.

[1] Esistono quattro modalità di calibrazione: probabilistica, di eccedenza, marginale e forte (Gneiting et al., 2007).

[2] La PIT converte un valore osservato di una variabile casuale nel suo corrispondente frattile utilizzando la CDF ed è esaminata in dettaglio nella sezione 4.1.1.

[3] $\circ$ indica l'operazione di composizione di funzioni. In questo caso quindi $\hat{F}_t^{-1}(p)$ viene usata come *input* da F_t^T.

PIT

In generale, la PIT converte un valore osservato di una variabile casuale nel suo corrispondente frattile utilizzando la CDF: ad esempio, se il valore mediano è 2.35, allora la trasformazione lo converte in 0.5. Se la vera funzione di ripartizione fosse esattamente nota, la trasformazione produrrebbe una distribuzione Uniforme. Tuttavia, in casi più realistici, la CDF non è nota e quindi la variabilità campionaria provoca un allontanamento dalla distribuzione desiderata. Nonostante ciò, il grado di conformità dei frattili stimati all'Uniforme può essere usato come un indicatore della qualità relativa della funzione di ripartizione stimata (Boylan & Syntetos, 2021).

In termini più formali, assumiamo che ad un tempo futuro t si preveda una densità $\hat{f}_t$ con funzione di ripartizione prevista $\hat{F}_t$ e denotiamo la vera distribuzione cumulativa come F_t. Si assuma inoltre di osservare y_t. Allora, la PIT è definita come:

$$p_t = \hat{F}_t(y_t) = \int_{-\infty}^{y_t} \hat{f}_t dy \ .$$

Si ha quindi che se la previsione è corretta, ossia $\hat{F}_t = F_t$, allora $p_t \overset{IID}{\sim} U[0,1]$ (Kolassa, 2016).

rPIT

Nell'ambito delle variabili casuali discrete non è possibile utilizzare direttamente il concetto di PIT definito nel paragrafo precedente. In questo caso, infatti, la PIT è discreta e assumerebbe una forma diversa da quella di una Uniforme in [0,1] anche se la densità prevista fosse corretta.

Per tale ragione, Kolassa (2016) propone di campionare casualmente p_t da una distribuzione $\tilde{p_t} \sim U[\hat{F}_t(y_t - 1), \hat{F}_t(y_t)]$. Così facendo, la distribuzione di $\tilde{p_t}$ corrisponde ad una mistura di Uniformi su sotto intervalli di [0,1], con pesi corrispondenti alle lunghezze dei sotto intervalli. L'approccio fornito corrisponde, perciò, con quello introdotto da Brockwell (2007) ed è possibile sfruttarne il Lemma 2[4]. Ne consegue che se $\hat{F}_t$ è specificata correttamente, allora $\tilde{p}_t$ si distribuisce

[4]Lemma 2 di Brockwell (2007): sia X una variabile casuale con funzione di distribuzione cumulativa F. Sia U una variabile Uniforme sull'intervallo $[0,1]$, indipendente da X. Allora

$$Z(X) = (1-U)F(X-) + UF(X)$$

è anch'essa distribuita in maniera Uniforme sull'intervallo $[0,1]$.

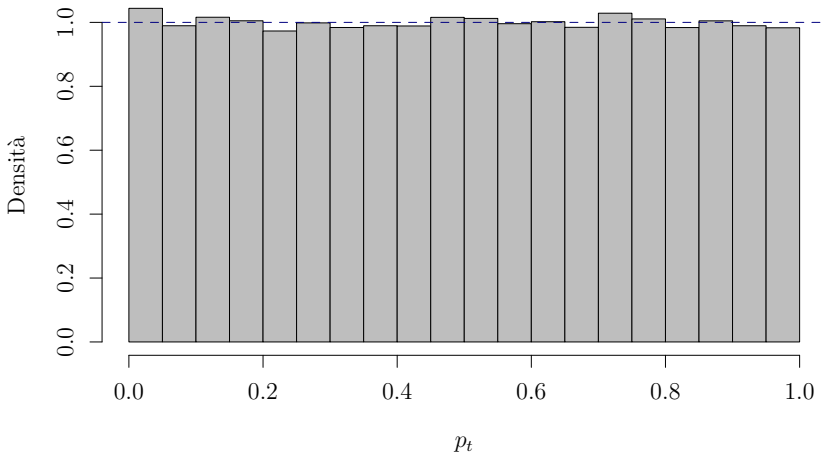

Figura 4.1: rPIT di simulazioni di dati Poisson e previsioni corrette.

nuovamente in maniera i.i.d. come $U[0,1]$, consentendo di sfruttare il *modus operandi* sviluppato per il caso continuo. In Figura 4.1 si mostra un esempio di rPIT nel caso di previsione corretta della densità, ricavato tramite dati simulati da una Poisson.

La valutazione della PIT e della rPIT avviene tramite l'analisi grafica della distribuzione di p_t. La Figura 4.2 mostra quattro casi in cui la trasformata non assume la forma di una Uniforme. In particolare, i grafici in alto presentano rispettivamente sotto-previsione, ossia media prevista troppo bassa, e sovra-previsione, mentre quelli in basso rispettivamente sotto-dispersione, ossia una varianza troppo bassa, e sovra-dispersione.

4.1.2 Test Kolmogorov-Smirnov (test KS)

Il test KS è un test non parametrico alternativo al test χ^2 per valutare la bontà di adattamento di un modello. Inizialmente proposto da Kolmogorov per l'analisi di un solo campione, ed esteso poi al caso di due campioni da Smirnov, il test è stato dettagliato da Massey Jr (1951). Esso considera l'ipotesi nulla di corretta previsione, ovvero $H_0 : \hat{F}_t = F_t$, e utilizza la distanza tra la funzione di ripartizione ipotetica, $\hat{F}_t$, e la funzione a gradini cumulativa di un campione casuale estratto dalla popolazione, F_t, per la verifica. La lontananza delle due funzioni, infatti,

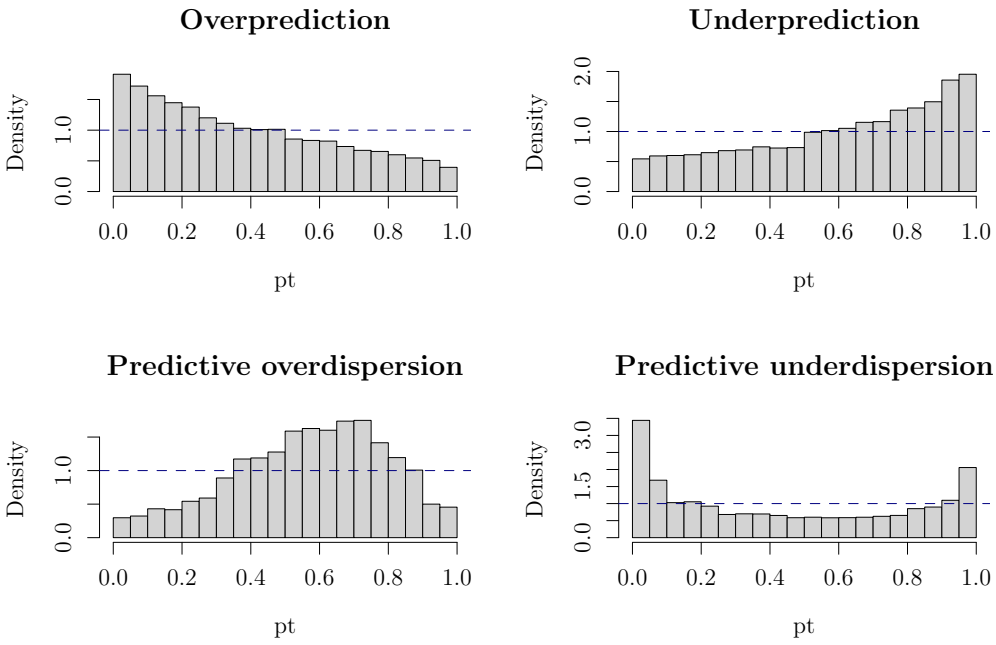

Figura 4.2: rPIT di simulazioni di dati Binomiale Negativa e previsioni non corrette.

costituisce evidenza della mancanza di calibrazione della previsione.

La valutazione si basa sulla statistica $D = \max |\hat{F}_t - F_t|$, la cui distribuzione è nota e indipendente da $\hat{F}_t$ se $\hat{F}_t$ è continua. Questo valore viene confrontato con i valori critici della Differenza Assoluta Massima tra le Distribuzioni Cumulative della Popolazione e del Campione, $d_\alpha(N)$, tabulati da Massey Jr (1951). In Figura 4.3 è riportata la procedura grafica per la valutazione del test.

Seguendo Wang et al. (2024), il test viene utilizzato per calcolare la distanza massima tra la distribuzione della rPIT e quella desiderata Uniforme in [0,1], in accompagnamento alla rappresentazione grafica della trasformata integrale.

4.2 *Sharpness*

La *sharpness* è una misura della concentrazione delle previsioni probabilistiche ed è valutata tramite funzioni di punteggio proprie (*proper scoring rules*, Gneiting et al. (2007)). Come introdotto all'inizio del capitolo, infatti, la calibrazione misurata tramite PIT o rPIT è una condizione necessaria, ma non sufficiente affinché la densità previsiva sia ideale. Le funzioni di punteggio sono quindi introdotte

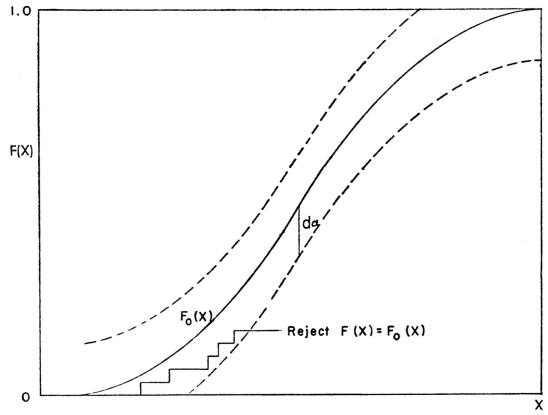

Figura 4.3: procedura grafica del test KS, (Massey Jr, 1951).

come strumenti per valutare sia la calibrazione sia la *sharpness* delle distribuzioni predittive.

Le funzioni di punteggio, denotate come $s(\hat{f}_t, y_t)$, sono funzioni di penalizzazione applicabili a previsioni probabilistiche e sono dette proprie se sono ottimizzate dalla migliore distribuzione predetta possibile, ossia quella reale (Gneiting & Katzfuss, 2014). In altre parole, una *scoring rule* è una funzione s che mappa una distribuzione predittiva $\hat{f}$ e una singola realizzazione y ad un valore di penalità $s(\hat{f}, y)$, di cui nella pratica viene solitamente riportato il valore atteso:

$$S := \frac{1}{t} \sum_{t=1}^{T} s(\hat{f}_t, y_t) \ .$$

Quindi, $s(\hat{f}, y)$ è detta propria se il suo valore atteso è minimo quando $\hat{f}$ corrisponde alla vera distribuzione futura di y, ossia $E_{y\sim f}\left(s(f, y)\right) \leq E_{y\sim f}\left(s(\hat{f}, y)\right)$.

Seguendo Wang et al. (2024), per la valutazione della *sharpness* delle previsioni, si utilizzano le tre funzioni introdotte nella sezione 3.3:

- **Punteggio logaritmico.** Si tratta di una metrica sensibile alle osservazioni con probabilità vicina a zero, sottolineando la significatività della copertura previsiva in situazioni estreme.

  Al fine di rendere i risultati comprarabili, si moltiplica il punteggio logaritmico per -1, in modo tale che valori più piccoli dello *score* indichino una maggiore *sharpness* della previsione.

- **Punteggio di Brier.** Si tratta di una misura legata direttamente alla funzione di massa di probabilità (PFM).

- **Punteggio DRPS.** Si tratta di una misura associata con la CDF.

4.3 *Performance* di inventario

Come ultimo criterio per la valutazione delle previsioni probabilistiche, si utilizzano delle misure della *performance* di inventario basate sui costi simulati durante il periodo previsivo (vedi sezione 3.4).

L'analisi della *performance* si basa sul rapporto tra i costi simulati totali per tutte le SKU e il totale di vendite nell'orizzonte previsivo, e su tre metriche cruciali di inventario:

- *Customer service level* (CSL) deviation: è il divario tra il CSL raggiunto, calcolato come la probabilità $\hat{F}_{\text{comb}}(\hat{q}_t(\tau))$ definita nella sezione 3.4, e il target. Questa misura rappresenta quindi la probabilità di non esaurire lo stock ad uno specifico livello *target* .

- Vendite perse: sono le parti positive della differenza tra la domanda reale e l'inventario previsto ad un certo livello durante il periodo di tempo considerato. Se ne considera la media per tutti gli SKU e rappresenta le vendite perse medie.

- Investimento di inventario: corrisponde all'inventario medio previsto ai livelli di servizio target tra i periodi e le SKU, denotando il livello di *safety stock* medio mantenuto.

La valutazione di queste tre metriche viene fatta tramite delle *trade-off curves*, strumenti grafici che misurano gli effetti diretti sul controllo dell'inventario e che consentono di comparare le diverse metodologie in maniera realistica, dato che sono le più significative dal punto di vista del *practitioner* (Trapero et al., 2019).

Capitolo 5

Analisi di dati reali

In questo capitolo si applicano i metodi individuali e le combinazioni descritte nel capitoli precedenti tramite l'analisi dei dati forniti da Kaggle[1] per la "M5 Competition"(Makridakis et al., 2022), una delle competizioni ideate dal Prof. Makridakis per la valutazione empirica della *performance* di metodi previsivi esistenti o nuovi.

In particolare, in questo testo si trattano solo i dati a livello di SKU, concentrandosi quindi sulla previsione di dati registrati a livelli granulari. Questa scelta è motivata dal *focus* della ricerca degli ultimi anni sui *big data*, ossia sull'analisi di dati dettagliati che consentono previsioni più precise e adattive, essenziali per gestire la natura intermittente delle vendite al dettaglio (Kolassa, 2016).

5.1 Descrizione e preparazione dei dati

I dati utilizzati nell'analisi, sono stati forniti da Kaggle per la "M5 Competition", il cui scopo era la previsione di 42840 serie storiche relative alle vendite unitarie di prodotti commercializzati da Walmart tra il 29/01/2011 e il 23/05/2016 (1941 giorni). I dati presentano una struttura gerarchica complessa, di cui si considera solo il livello più basso (30490 serie storiche).

Oltre allo storico delle vendite unitarie per prodotto, i *dataset* forniscono anche alcune informazioni esogene riguardo al bene stesso e agli eventi di calendario che potrebbero influenzare l'andamento delle vendite, come ad esempio l'attività a scopo promozionale SNAP (*Supplemental Nutrition Assistance Program*), e alcune festività.

[1]Fonte dei dati: https://www.kaggle.com/c/m5-forecasting-accuracy/data

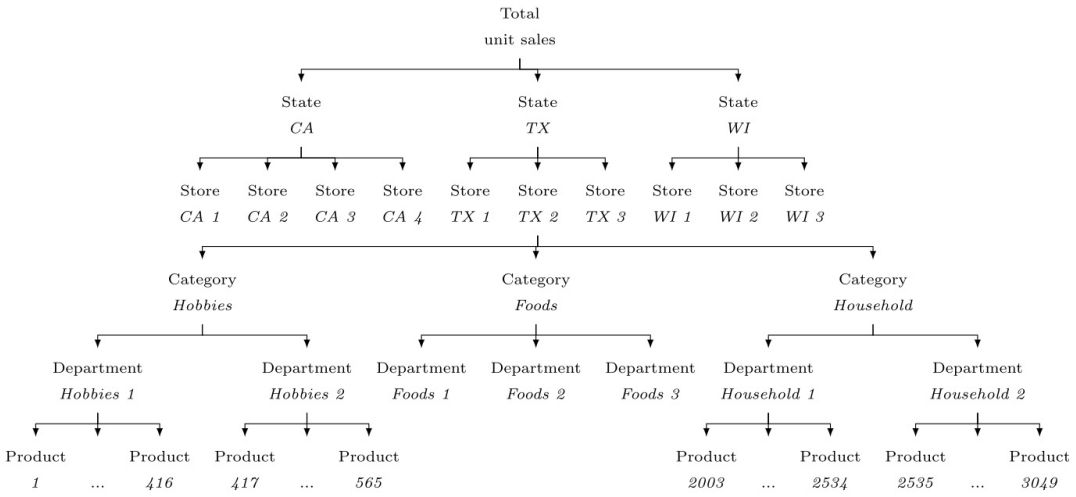

Figura 5.1: Struttura dei dati, (Makridakis et al., 2022).

Nello specifico, i dati contengono 3049 prodotti venduti in 10 negozi (locati in California, Texas e Wisconsin) e classificati in 7 dipartimenti, a loro volta disaggragati in 3 categorie (*Hobbies, Foods, Household*). I diversi livelli di aggregazione dei dati sono osservabili in Figura 5.1.

I dati originali forniti da Kaggle sono suddivisi in 3 *datasets*, relativi rispettivamente allo storico delle vendite del prodotto (ogni riga corrisponde allo storico delle vendite di una SKU), alle informazioni di calendario e ai prezzi di vendita e alle caratteristiche della SKU specifica (e.g., dipartimento, categoria, ...).

I dati disponibili (1969 giorni, ossia circa 5.4 anni) sono stati suddivisi in tre parti, in base alla divisione adottata durante la competizione:

- Il periodo 29/01/2011 - 27/03/2016 (circa 1885 osservazioni) è stato utilizzato come *training set* dei metodi individuali introdotti nel capitolo 2. Per ogni serie, si è selezionato solo il periodo a partire dalla prima vendita positiva del prodotto.

- Il periodo 28/03/2016 - 24/04/2016 (28 osservazioni) è stato usato come *validation set* per il calcolo dei pesi per la combinazione dei metodi individuali, introdotti nel capitolo 3, e per il calcolo delle previsioni dei metodi individuali.

- Infine, il periodo 25/04/2016 - 22/05/2016 (28 osservazioni) è stato usato come *test set* per il calcolo delle previsioni delle combinazioni dei metodi di previsione probabilistica.

Si è poi eseguita una scrematura delle serie disponibili, eliminando quelle che contengono vendite costanti pari a zero nel *validation set*, al fine di evitare l'introduzione di una potenziale distorsione nelle combinazioni. In caso contrario, infatti, si potrebbe incorrere in pesi sproporzionatamente alti per metodi che tendono a prevedere zero, determinando un errore considerabile quando si incorre in valori positivi nel periodo di valutazione (Wang et al., 2024). Un altro fattore considerato nella riduzione del *dataset* è stato il numero di zeri contenuti nella serie: una volta eliminati i valori nulli iniziali, alcune delle serie contengono un basso numero di osservazioni pari a zero o non ne contengono. Si decide, quindi, di usare una soglia arbitraria del 20% di osservazioni pari a zero per determinare l'intermittenza delle serie, eliminandone 1934. A seguito di queste operazioni restano 26969 serie, contenenti in media 62.57% di valori pari a zero.

È stata poi applicata una categorizzazione SBC, da cui sono risultate 437 serie *smooth* (21.59% di zeri), 20832 serie "strettamente" intermittenti (66.45%), 272 erratiche(22.26%) e 5428 *lumpy* (53.01%). La Figura 5.2 mostra una rappresentazione grafica della classificazione delle serie.

Da questi dati, si è ricavato un *dataset* per ognuna delle serie analizzate in cui ogni riga rappresenta una vendita e si sono poi trasformate alcune delle covariate presenti nei dati da inserire nel modello GAM-QR (vedi Appendice B.1.1).

Per limitare i costi computazionali, si è analizzato un campione casuale contenente circa il 50% delle serie, utilizzando 12902 *datasets* per l'analisi. In base alla classificazione SBC, si elaborano 250 serie *smooth* (21.63%), 10448 intermittenti (66.44%), 140 erratiche (22.27%) e 2064 *lumpy* (53.34%). Tra queste, si escludono altre 12 serie dall'analisi, in quanto contenenti informazioni insufficienti per ottenere delle previsioni adeguate nella valutazione in termini di *performance* di inventario. Le serie escluse, appartenenti alle categorie intermittenti e *lumpy*, contengono un'alta variabilità della domanda e un ingente numero di osservazioni pari a zero nel *training set*. Questo ha determinato l'esplosione dei valori dei quantili alti della distribuzione previsiva, determinando dei costi di inventario simulati estremamente elevati.

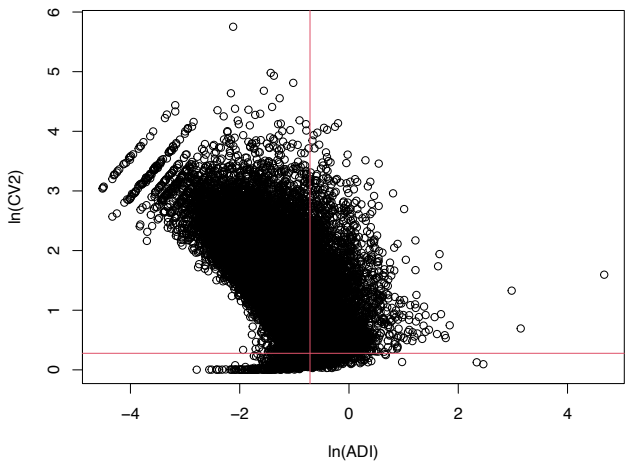

Figura 5.2: Rappresentazione grafica della classificazione delle serie della competizione M5 in base all'intermittenza(ADI) e l'erraticità (CV^2). Le linee rosse indicano le soglie dei due indici che dividono le classi. Partendo dal quadrante in alto a sinistra e procedendo in senso orario, i quadranti indicano le classi di serie "strettamente" intermittenti, *lumpy*, erratiche e *smooth*.

5.2 Implementazione

L'esperimento di previsione è stato eseguito utilizzando il linguaggio R (R Core Team, 2024), sfruttando i pacchetti disponibili per i vari modelli laddove possibile. Nei casi in cui tali pacchetti non erano disponibili, le funzioni necessarie sono state implementate manualmente. L'intero progetto è consultabile in un repository GitHub al seguente URL: https://github.com/aciandri/domanda-intermittente. Inoltre, si rimanda all'Appendice B.1 per ulteriori chiarimenti relativi ai procedimenti eseguiti.

Dato che ognuno dei metodi individuali descritti nel capitolo 2 fornisce un *output* differente, al fine di uniformare i risultati, si applica la seguente procedura (Wang et al., 2024):

1. **Applicazione dei metodi individuali.** Si utilizzano i metodi descritti nel capitolo 2 per fare previsioni sui dati 28 passi in avanti.

2. **Generazione dei quantili.** Si calcolano i quantili $\tau \in \{0.01, 0.02, ..., 0.99\}$ per ogni orizzonte temporale.

3. **Calcolo della funzione di massa di probabilità (PFM).** Si arrotondano tutte le previsioni dei quantili e si pone il 100-esimo percentile pari al 99-esimo. Si calcolano le frequenze di ogni valore come la sua probabilità. Infine, le previsioni dei quantili vengono trasformate nella relativa PFM.

4. **Stima dei pesi di combinazione.** Si calcolano i pesi per combinare i risultati dei vari metodi, come descritto nel capitolo 3.

5.3 Risultati

In questa sezione si riportano i risultati delle analisi svolte in termini di calibrazione, *sharpness* e *performance* dell'inventario, come descritto nel capitolo 4.

La calibrazione viene valutata tramite l'analisi grafica della rPIT e la statistica D del test KS, la quale indica la distanza massima tra la distribuzione campionaria dei $\tilde{p}_t$ e quella desiderata, ossia, in questo caso, Uniforme [0,1]. Per la rappresentazione grafica si sono estratti 10 campioni $\tilde{p}_t$ per ogni istante temporale, seguendo la procedura suggerita da Kolassa (2016), ottenendo 280 valori in totale per le previsioni 28 passi in avanti di una singola serie.

I grafici delle rPIT dei metodi individuali sono riportati nella Figura 5.3, insieme alla rispettiva statistica D del test KS. In particolare, si può subito notare come i modelli GAM-QR, iETS_G e iETS_D risultino ben calibrati rispetto agli altri. Anche i metodi classici per la domanda intermittente e il modello Poisson appaiono abbastanza calibrati, anche se tendono a formare un tenue andamento ad "u". Questo comportamento indica una leggera sottostima della distribuzione degli scenari estremi, quali vendite molto alte o pari a zero. I metodi tradizionali per serie storiche, usati come *benchmark*, e l'iETS_O mostrano una forte sottoprevisione, sovrastimando la parte alta della distribuzione e sottostimando quella bassa. Al contrario, il modello Binomiale Negativa, il WSS e l'iETS_I tendono a sottostimare la parte alta della distribuzione e sovrastimare quella alta.

La Figura 5.4 mostra gli rPIT delle varie combinazioni. Si osserva che il metodo SA, la mediana e log-opt mostrano una buona calibrazione e, complessivamente, i risultati delle combinazioni appaiono più calibrati rispetto a quelli dei metodi individuali, superando i problemi di sottostima della domanda osservati in alcuni di essi.

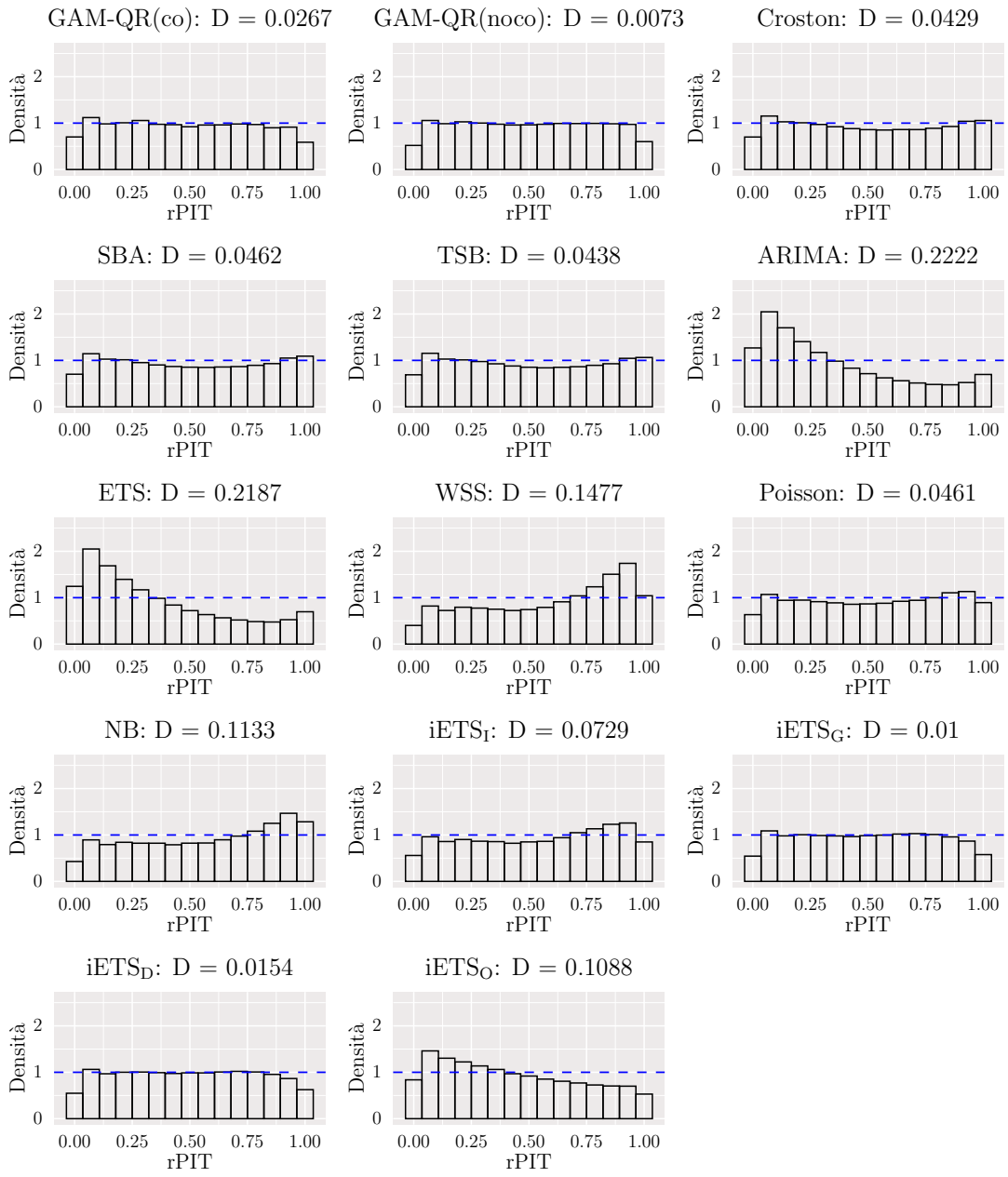

Figura 5.3: rPIT di tutti i metodi individuali. La linea tratteggiata indica la distribuzione uniforme ideale e le barre rappresentano le densità dei p_t.

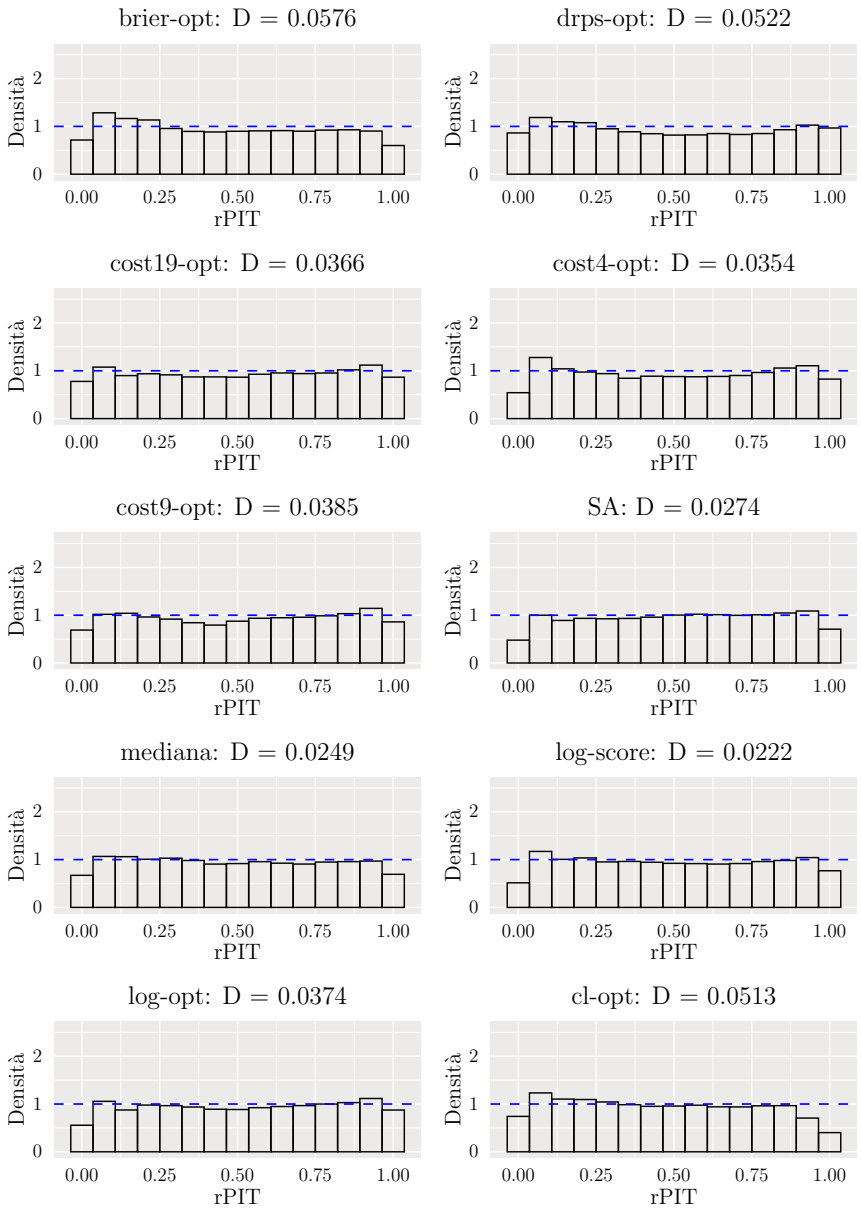

Figura 5.4: rPIT dei metodi di combinazione. La linea tratteggiata indica la distribuzione uniforme ideale e le barre rappresentano le densità dei p_t.

I risultati della *sharpness* sono riportati nella Tabella 5.1. Tra i metodi di combinazione, il punteggio di Brier fornisce i risultati migliori in termini di punteggio logaritmico e di Brier, mentre il *Simple Average* in termini di DRPS. Complessivamente, le combinazioni forniscono dei risultati migliori in termini di *sharpness* per tutte le metriche in esame.

Per la valutazione della *performance* di inventario, si esaminano i costi simulati sotto diverse condizioni, ipotizzando le seguenti combinazioni dei costi marginali (Wang et al., 2024): cost(1,4), cost(1,9) e cost(1,19). I risultati sono riportati nella Tabella 5.2. Per quanto riguarda i metodi individuali, si osserva che il modello ARIMA tende a determinare costi più bassi nelle situazioni in cui il costo marginale delle vendite perse è molto superiore rispetto a quelli di stoccaggio, ovvero cost(1,9) e cost(1,19). Per quanto riguarda le combinazioni, invece, la media sembra fornire i risultati migliori nel caso cost(1,4), mentre nelle altre due casistiche l'ottimizzazione delle funzioni logaritmica e DRPS sembrano essere i metodi più idonei. Complessivamente, i costi derivanti dalle combinazioni sono più alti rispetto ai metodi individuali. In generale, le tabelle relative ai costi e alla *sharpness*, indicano la presenza di un *trade-off* tra le misure previsive e quelle di inventario, la cui esistenza è confermata anche da Wang et al. (2024).

Seguendo Wang et al. (2024), come ulteriore analisi delle implicazioni delle previsioni a livello di inventario, si presenta un'analisi grafica delle relazioni esistenti tra le varie componenti dei costi simulati. La Figura 5.5 mostra, separatamente per i metodi individuali sopra e per le combinazioni sotto, le relazioni esistenti tra la deviazione del CSL da quello desiderato, le vendite perse cumulate su orizzonti temporali di 28 giorni e i costi di stoccaggio giornalieri. Idealmente, si vorrebbe una deviazione del CSL e vendite perse pari a zero, e costi di stoccaggio bassi. Ci si concentra sulle casistiche studiate anche nella Tabella 5.2, ossia su CSL target pari a 0.8, 0.9 e 0.95, indicate dai punti presenti sulle linee.

Nei grafici, si è scelto di non includere i risultati del modello GAM-QR con covariate, poiché i costi di stoccaggio particolarmente elevati associati a tale modello compromettono la chiarezza visiva delle illustrazioni. Inoltre, il modello non offre miglioramenti significativi negli altri due fattori considerati, rendendo la sua inclusione superflua ai fini dell'analisi. Osservando i primi due grafici in alto, si vede che il modello iETS_I spicca sia in funzione della deviazione dal CSL target che dei costi di stoccaggio per tutti i livelli target. Tuttavia, il *trade-off* tra le vendite perse e gli investimenti è evidente in tutti i metodi in esame, determinando

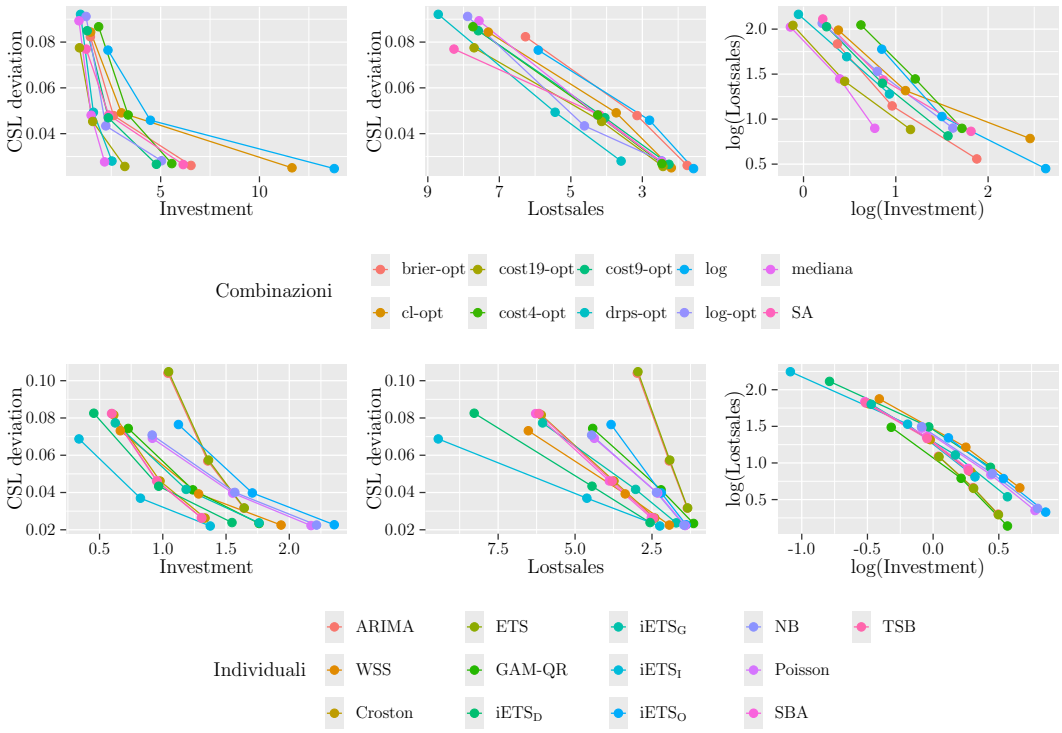

Figura 5.5: Relazione tra il CSL, gli investimenti e le vendite perse. I tre punti sulle linee rappresentano i livelli target dell'80%, 90% e 95%. La deviazione dal CSL è il divario tra il livello di servizio raggiunto e quello target. L'investimento è la media delle *stock* giornaliere su tutte le SKU e tutti gli orizzonti. Le vendite perse in figura indicano la somma delle vendite perse su tutti gli orizzonti temporali, calcolata in media per SKU. Seguendo Wang et al. (2024), i grafici sulla destra presentano degli assi logaritmici per semplicità di visualizzazione, rendendo le forme della curva concave anziché convesse. Inoltre, per consentire una rappresentazione leggibile delle relazioni, si è escluso dal grafico il modello GAM-QR con covariate, in quanto presenta dei costi di investimento particolarmente elevati.

vendite perse più elevate per $iETS_I$ rispetto agli altri. Tale relazione è chiaramente visibile dai grafici sulla destra della Figura 5.5. Confrontando tali risultati con i rispettivi grafici nella seconda riga, si nota che complessivamente le combinazioni tendono a mostrare dei costi relativi agli investimenti più alti rispetto ai metodi individuali.

I risultati in termini di *sharpness* e misure di inventario sono state esaminate anche separatamente per ogni categoria di domanda in base alla classificazione SBC, riportati nell'Appendice B.2. Dall'analisi di queste tabelle, sorge che la classe "strettamente" intermittente mostra dei livelli di *sharpness* migliori rispetto alle altre categorie, oltre che un maggiore peso del *trade-off* sui risultati. Si nota inoltre che i metodi migliori in base al criterio logaritmico rimangono costanti tra le classi, ma cambiano se si considerano il DRPS e punteggio di Brier. Utilizzando tali misure, si notano dei *pattern* comuni tra le diverse categorie: nelle serie con bassa intermittenza, ossia *smooth* e erratiche, la DRPS identifica la combinazione con pesi ottimizzati per la funzione di punteggio logaritmico come migliore, mentre per le serie con alta variabilità nella domanda, il punteggio di Brier seleziona il DRPS come ottimale.

Tipo	Metodo	Logaritmico	DRPS	Brier
Individuali	GAM-QR(co)	1.332	0.593	**-0.457**
	GAM-QR(noco)	**1.313**	**0.579**	-0.456
	Croston	1.702	0.639	-0.432
	SBA	1.719	0.639	-0.432
	TSB	1.699	0.636	-0.434
	ARIMA	1.735	0.706	-0.319
	ETS	1.732	0.710	-0.323
	WSS	1.555	0.665	-0.403
	Poisson	1.571	0.625	-0.431
	NB	1.788	0.648	-0.411
	$iETS_I$	1.624	0.682	-0.392
	$iETS_G$	1.630	0.651	-0.416
	$iETS_D$	1.519	0.628	-0.436
	$iETS_O$	1.592	0.631	-0.434
Combinazioni	brier-opt	**1.353**	0.596	**-0.457**
	drps-opt	1.646	0.627	-0.455
	cost19-opt	1.493	0.623	-0.430
	cost4-opt	1.506	0.626	-0.429
	cost9-opt	1.483	0.619	-0.430
	SA	1.475	**0.585**	-0.432
	mediana	1.416	0.672	-0.400
	log-score	1.407	0.613	-0.437
	log-opt	1.598	0.599	-0.434
	cl-opt	1.490	0.634	-0.430

Tabella 5.1: Risultati della sharpness in termini di punteggi logaritmico, di Brier e DRPS per tutti i metodi individuali e di combinazione. I numeri in grassetto indicano il metodo migliore per le metriche del relativo gruppo (individuali e combinazioni).

Tipo	Metodo	Cost(1,4)	Cost(1,9)	Cost(1,19)
Individuali	GAM-QR(co)	7.873	13.650	43.324
	GAM-QR(noco)	**1.965**	2.967	4.119
	Croston	2.080	3.133	4.451
	SBA	2.097	3.176	4.544
	TSB	2.081	3.140	4.467
	ARIMA	2.072	**2.819**	**3.668**
	ETS	2.089	2.837	3.680
	WSS	2.324	3.308	4.356
	Poisson	2.097	3.111	4.294
	NB	2.242	3.344	4.659
	$iETS_I$	2.295	3.386	4.695
	$iETS_G$	2.285	3.392	4.657
	$iETS_D$	2.138	3.211	4.469
	$iETS_O$	2.210	3.321	4.590
Combinazioni	brier-opt	9.257	7.175	7.388
	drps-opt	5.872	**3.965**	4.084
	cost-opt	9.173	6.556	6.960
	SA	**5.732**	5.278	5.139
	mediana	7.265	9.087	15.473
	log-score	9.512	6.971	7.326
	log-opt	7.647	5.918	**3.5**
	cl-opt	11.122	8.699	13.364

Tabella 5.2: Costi di inventario simulati per tutti i metodi individuali e di combinazione. I risultati dei pesi ottimali per i costi di inventario sono riportati per la corrispondente condizione in esame, quindi cost4-opt per cost(1,4), cost9-opt per cost(1,9) e cost19-opt per cost(1,19).

Conclusioni

In questa tesi sono state esplorate varie procedure per la previsione della domanda intermittente, con particolare attenzione all'utilizzo di combinazioni probabilistiche. L'analisi svolta mira ad esaminare e confrontate le diverse procedure sia a livello di accuratezza previsiva che di implicazioni pratiche sui costi aziendali.

A tale scopo, si è condotta un'analisi su un ampio campione di 12902 serie storiche di domanda intermittente, selezionate casualmente dai dati forniti per la "M5 Competition" (Makridakis et al., 2022). I dati sono stati inizialmente analizzati tramite l'applicazione di dodici metodi individuali, parametrici e non, ideati per la domanda intermittente e due modelli tradizionali per l'analisi di serie storiche usate come *benchmark*. Le previsioni ottenute sono state poi utilizzate per il calcolo dei pesi delle combinazioni, prodotti tramite dieci diverse procedure, riassumibili in tre categorie: semplici, ottimali per tre diverse funzioni di punteggio e ottimali basati sui costi d'inventario. I risultati dei metodi individuali e di combinazione sono stati valutati e confrontati sia in termini di accuratezza previsiva tramite calibrazione e *sharpness*, che delle implicazioni economiche, attraverso i costi d'inventario simulati.

Dai risultati ottenuti emerge che l'utilizzo delle combinazioni probabilistiche offre un miglioramento significativo dell'accuratezza previsiva, sia in termini di calibrazione che di *sharpness*. La procedura porta a risultati calibrati, riuscendo a mitigare i problemi di sovra e sotto-previsione che caratterizzano alcuni dei metodi individuali. Anche in termini di *sharpness*, misurata tramite delle funzioni di punteggio proprie, le combinazioni mostrano risultati interessanti, portando a previsioni complessivamente migliori rispetto ai singoli modelli.

Rispetto ai costi d'inventario, invece , l'applicazione di combinazioni non porta a risultati altrettanto soddisfacenti. In questo ambito infatti i risultati in termini di costi simulati risultano maggiori rispetto a quelli dei metodi individuali. Questa relazione inversa tra i risultati delle due misure di valutazione, è riscontrata anche

negli altri metodi in esame, indicando la presenza di un *trade-off* tra le misure previsive e quelle d'inventario. Tali risultati sono coerenti con quanto riscontrato anche da Wang et al. (2024), che affermano che i modelli migliori in termini di accuratezza delle previsioni, non necessariamente implicano una riduzione dei costi, e viceversa.

In conclusione, l'analisi condotta ha dimostrato che l'adozione di combinazioni probabilistiche porta a un miglioramento significativo dell'accuratezza previsiva rispetto ai singoli modelli. Tuttavia, le combinazioni non hanno mostrato vantaggi altrettanto evidenti quando si considerano i costi d'inventario simulati. La presenza di questo *trade-off* sottolinea la necessità di sviluppare strategie che bilancino adeguatamente la qualità dei risultati con l'efficienza dei costi, per supportare decisioni aziendali che non solo migliorino la precisione delle stime, ma anche l'efficacia complessiva delle operazioni aziendali.

Appendice A

Approfondimenti sulla metodologia

A.1 La domanda intermittente al di fuori del contesto economico

L'applicabilità delle metodologie studiate per la domanda intermittente non si limita all'ambito economico a cui è prevalentemente associata. Ad esempio, attraverso la decomposizione della serie in "*peak over threshold*" (POT), è possibile trasformare vari tipi di serie in intermittenti, facilitando lo studio di fenomeni altrimenti complessi. Questo approccio divide i dati disponibili in tre serie, dette *white*, *grey* e *black swans*. La prima costituisce la baseline, ossia gli eventi di base, mentre le altre due denotano i fenomeni estremi attesi e non (Nikolopoulos, 2021). Isolando i valori che eccedono una soglia fissata (POT), ossia gli eventi estremi, si è in grado di estrarre i momenti di interesse, di bassa numerosità e molto intermittenti per natura. Questo rappresenta uno dei due approcci per l'applicazione della Extreme Value Theory (EVT), che analizza le deviazioni estreme da una misura statistica di locazione centrale al fine di stimare la probabilità degli eventi più estremi della serie storica (Petropoulos et al., 2022).

A.2 Il metodo SES

Il SES, in quanto metodo di lisciamento esponenziale, fa una media pesata delle osservazioni passate, dove il peso decade esponenzialmente all'allontanamento

dall'osservazione presente. Nell'ambito della domanda intermittente, quindi, associa pesi più alti subito dopo la domanda e più bassi nel periodo che la precede. (Syntetos & Boylan, 2001)

Il lisciamento esponenziale è un metodo disegnato per fornire una risposta appropriata a cambiamenti nei *pattern* della domanda e la sua forma più semplice, il SES, risulta adeguata quando non sono presenti chiari pattern stagionali o trend. E' un metodo robusto che ha trovato successo in diverse applicazioni riguardanti l'inventario: nonostante non sia stato costruito per la domanda intermittente, infatti, fornisce buoni risultati (Boylan & Syntetos, 2021).

In termini formali, il modello può essere espresso come segue

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

dove $\hat{y}_t$ sono le stime e $\alpha \in [0,1]$ è un parametro di lisciamento esponenziale.

Il modello può essere interpretato sia come un meccanismo di correzione dell'errore che come una "media pesata".

Nel primo caso, α denota la responsività del sistema a cambiamenti sottostanti alla domanda media, in cui un valore più alto indica una maggiore responsività. Quindi, dipendentemente dal valore scelto, si possono avere *over reaction* o *under reaction*.

Nel secondo, invece, il parametro determina il peso delle osservazioni: un suo valore basso implica poca differenza nei pesi assegnati a osservazioni vicine e lontane nel tempo ed è utile in situazioni in cui i *pattern* della domanda sono abbastanza stabili. L'assunzione di valori estremi del parametro, 0 e 1, implicano rispettivamente un stima pari al valore iniziale (quindi un modello con media stazionaria) e uno stimatore naive.

A.2.1 Distorsione del metodo SES

Seguendo Boylan & Syntetos (2021), si considera il metodo SES descritto nell'equazione (A.1) per un periodo di revisione di lunghezza R:

$$\hat{y}_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + ... + \alpha(1-\alpha)^R y_{t-R+1} + (1-\alpha)^R \hat{y}_{t-R+1} \tag{A.2}$$

dove il termine $\hat{y}_{t-R+1}$ rappresenta l'ultima previsione calcolata prima dell'inizio dell'intervallo di previsione più recente. Dato che questa previsione non è condizionata all'occorrenza di domanda durante il periodo precedente, il suo valore atteso è pari a $E(\hat{y}_{t-R+1}) = \frac{\mu}{p}$.

Le domande $y_t, ..., y_{t-R+1}$ sono invece condizionate al fatto che almeno una di esse sia positiva. Questo implica che il valore atteso condizionato per una generica domanda $y_{t-j},\ j = 0, ..., R-1$, può essere espresso come:

$$E(y_{t-j}) = \frac{\mu}{p}\left(\frac{1}{1-(1-1/p)^R}\right) \ . \tag{A.3}$$

Sostituendo il valore atteso di $\hat{y}_{t-R+1}$ e y_{t-j} per $j = 0, ..., R-1$ nell'equazione (A.2), si ottiene l'espressione generale del valore atteso di previsioni SES alla fine degli intervalli di revisione (di lunghezza R) contenenti domanda positiva:

$$E(\hat{y}_{t+1}|Y_R > 0) = \frac{\mu}{p}\left(\frac{1-(1-\alpha)^R}{1-(1-1/p)^R} + (1-\alpha)^R\right) \tag{A.4}$$

per $0 < \alpha \leq 1$, $p \geq 1$ e $Y_R = y_t + ... + y_{t-R+1}$ è la domanda totale dell'intervallo di revisione più recente. Tale risultato vale generalmente per qualsiasi intervallo di revisione di lunghezza strettamente positiva. Il caso speciale di $R = 1$ è riportato nell'equazione (2.8).

A.3 I modelli *state space* SSOE

L'approccio *state space* alla modellazione di serie storiche univariate è largamente usato, sia nella teoria che nella pratica. La ricchezza del *framework* implica che sia possibile ottenere formulazioni differenti del modello anche quando si cerca di descrivere lo stesso fenomeno. A questo proposito, si possono distinguere due schemi che utilizzano specificazioni diverse degli errori del modello: il *single source of error* (SSOE) e il *multiple source of error* (MSOE). Il primo è caratterizzato da errori perfettamente correlati, mentre il secondo prevede un termine di errore diverso per ogni equazione del modello.

Per comprendere meglio la differenza tra i due, si considera il metodo SES descritto nell'equazione (A.1).

Lo schema MSOE scrive il modello in forma *state space* nel seguente modo:

$$\begin{aligned} y_t &= l_t + u_t \\ l_t &= l_{t-1}\alpha_1 w_t, \end{aligned} \tag{A.5}$$

dove l_t è una variabile di stato non osservata che denota il livello o media della serie. Si può osservare che il modello contiene due fonti di errori: u_t e w_t, assunte rispettivamente $u_t \sim N(0, \sigma_u^2)$ e $w_t \sim N(0, \sigma_w^2)$, e mutualmente indipendenti.

Il modello *state space* SSOE, invece, scrive il metodo SES come:

$$\begin{aligned} y_t &= l_t + \epsilon_t \\ l_t &= l_{t-1}\alpha_1\epsilon_t \ , \end{aligned} \tag{A.6}$$

dove l_t è analogo alla variabile descritta per l'equazione (A.5) e α è un parametro. In questo caso, si può osservare un'unica fonte di errore ϵ_t, assunta $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ e serialmente indipendenti (Ord et al., 2005).

A.4 Condizione di uguaglianza tra previsione e media condizionata negli ETS

Le condizioni $E(\epsilon_t) = 0$ per i modelli additivi e $E(1 + \epsilon_t) = 1$ sono necessarie per l'uguaglianza tra la previsione puntuale e la media condizionata h passi in avanti, ma non sono sufficienti. In situazioni in cui il trend e/o stagionalità sono moltiplicativi, infatti, i valori degli stati molteplici passi in avanti ($h > 1$) introducono moltiplicazioni dei termini di errore, dato che il modello ETS ha un'unica fonte di errore (Svetunkov, 2023). Ne consegue che i valori attesi condizionati potrebbero non avere forma chiusa e, quindi, l'utilizzo di questi modelli potrebbe richiedere lo svolgimento di simulazioni. Questo problema chiaramente non si pone nel caso in cui $h = 1$.

Nonostante queste considerazioni, l'importanza delle ipotesi $E(\epsilon_t) = 0$ per i modelli additivi e $E(1 + \epsilon_t) = 1$ per quelli moltiplicativi permane. In caso contrario, infatti, le equazioni di livello si comporterebbero come un *drift*, richiedendo dei cambiamenti nella struttura del modello e rendendo quindi impossibile l'applicazione efficiente dei modelli ETS. Un valore atteso dell'errore diverso da quello specificato porterebbe inoltre a delle difficoltà nella stima del modello: gli ETS mostrano un effetto "*pull to centre*" tale per cui la vicinanza delle previsioni ai valori reali dipende dagli errori, implicando una difficoltà nella stima di errori con media diversa da 0.

In termini più analitici, considerando il caso semplice ETS(M, N, N), la previsione puntuale è definita come

$$\hat{y}_{t+h} = \exp\left(\left(\mathbf{w}_1'\mathbf{F}_1^{h-1} + \mathbf{w}_m'\mathbf{F}_m^{\lceil \frac{h}{m} \rceil - 1}\right)\log \mathbf{v}_t\right) \tag{A.7}$$

dove $\mathbf{w}$ è il vettore di misura, $\mathbf{v}_t$ è il vettore degli stati e $\mathbf{F}$ è la matrice di transizione. Si può vedere che $\hat{y}_{t+h}$ coincide con le media aritmetica e geometrica, seguendo la ricorsione per gli ETS(M, N, N):

$$y_{t+h} = l_t \prod_{j=1}^{h-1} (a + \alpha\epsilon_{t+j})(1 + \epsilon_{t+h}) \ , \tag{A.8}$$

il cui valore atteso coincide con l_t purché

$$E(1 + \epsilon_t) = 1$$

e il termine di errore è i.i.d. (Svetunkov & Boylan, 2024)

Questa ipotesi, quindi, garantisce che le previsioni puntuali h passi in avanti corrispondano ai valori attesi dei modelli, condizione necessaria per la costruzione del modello e per iniziare la ricorsione. Inoltre, se questa non valesse, l'equazione di transizione differirebbe da quella tipicamente assunta, implicando la necessità di calcolare il valore atteso dei prodotti delle variabili casuali.

A.5 Approfondimenti sugli iETS

Breve introduzione agli ETS(M, N, N) con diverse distribuzioni

Svetunkov & Boylan (2024) sottolineano la rilevanza degli ETS con errore moltiplicativo, nonostante siano meno trattati in letteratura, a favore della controparte additiva, in termini di problemi teorici e implementazioni pratiche. Questi modelli, infatti, risultano particolarmente importanti quando si è in presenza di asimmetria nei dati, specialmente in dati *low volume*, come si possono trovare nell'ambito della *supply chain* e vendita al dettaglio.

In particolare, qua siamo interessati ad un caso particolare: gli ETS(M, N, N), ossia senza trend e stagionalità, formulato come segue:

$$\begin{aligned} y_t &= l_{t-1}(1 + \epsilon_t) \\ l_t &= l_{t-1}(1 + \alpha\epsilon_t) \end{aligned} \tag{A.9}$$

dove $\epsilon_t \sim N(0, \sigma^2)$. Questo risulta particolarmente interessante, in quanto ha forme chiuse per i momenti condizionati.

La principale limitazione del modello descritto nell'equazione (A.9) applicato al caso di *low volume data* è l'ipotesi di normalità. Questa non risulta più ragionevole, dato che il modello potrebbe produrre valori negativi anche per dati

strettamente positivi. Per tale ragione, Svetunkov & Boylan (2024) propongono di utilizzare distribuzioni Gamma, Log-Normali o Normale Inversa, implicando una delle sequenti situazioni:

$$\begin{aligned} 1+\epsilon_t &\sim IG(1,\sigma^2) \\ 1+\epsilon_t &\sim \log N\left(-\frac{\sigma^2}{2},\sigma^2\right) \\ 1+\epsilon_t &\sim \Gamma(\sigma^{-2},\sigma^2)\ . \end{aligned} \tag{A.10}$$

Le restrizioni imposte nell'equazione (A.10) sono state inserite per garantire il rispetto della condizione $E(1+\epsilon_t)=1$ e fanno sì che la media e la varianza h passi in avanti del modello ETS(M, N, N) possa essere calcolato usando le stesse formula degli ETS(M, N, N) convenzionali:

$$E(l_{t+h} \mid t) = E\left(l_t \prod_{j=1}^{h}(1+\alpha\epsilon_{t+j}) \mid l_t\right) = l_t, \tag{A.11}$$

$$V(l_{t+h} \mid t) = l_t^2\left(\left(1+\alpha^2\sigma^2\right)^h - 1\right). \tag{A.12}$$

Inoltre, si possono ottenere i valori reali un passo in avanti utilizzando la proprietà di scalabilità:

$$\begin{aligned} IG : y_{t+1} &\sim IG\left(l_t, \frac{\sigma^2}{l_t}\right) \\ \log N : y_{t+1} &\sim \log N\left(\log(l_t) - \frac{\sigma^2}{2}, \sigma^2\right) \\ \Gamma : y_{t+1} &\sim \Gamma(\sigma^{-2}, l_t\sigma^2) \end{aligned} \tag{A.13}$$

A.5.1 Proxy dei termini di errore nel modello iETS

La probabilità di occorrenza della domanda in equazione (2.25) al tempo t è definita come:

$$p_t = \frac{\mu_{a,t}}{\mu_{a,t}+\mu_{b,t}} \tag{A.14}$$

Quindi, se la probabilità fosse nota, sarebbe possibile calcolare $\mu_{a,t}$ e $\mu_{b,t}$ tramite le seguenti formule

$$\mu_{a,t} = \mu_{b,t}\frac{p_t}{1-p_t} \tag{A.15}$$

$$\mu_{b,t} = \mu_{a,t}\frac{1-p_t}{p_t}\ . \tag{A.16}$$

Tuttavia, p_t non è mai nota. Al suo posto si può utilizzare la sua stima al tempo t , $\hat{p}_t$, per il calcolo dell'errore, ipotizzando che quando $o_t = 1$ la probabilità dovrebbe essere il più vicina possibile ad 1 e, in caso contrario, a 0. In base a questa idea, l'errore può essere calcolato come:

$$\upsilon_t = o_t - \hat{p}_t \tag{A.17}$$

quindi $\upsilon_t \in (0, 1)$.

Per far sì che l'errore stia nell'intervallo (0, 1), vi si applica la seguente trasformazione:

$$u_t = \frac{1 + \upsilon_t}{2} \tag{A.18}$$

in modo tale che $u_t = 0.5$ corrisponda alla situazione ideale di eguaglianza tra l'outcome effettivo e quello previsto. Nel caso limite $\hat{p}_t = 1$ e $o_t = 0$, la proxy dell'errore equazione (A.18) è pari a 0 e, viceversa, pari ad 1.

Inserendo la nuova variabile u_t nelle equazioni A.14 e A.16 al posto di p_t e ponendo $\mu_{a,t} = \mu_{b,t} = 1$ (i.e., ipotizzando l'indipendenza dei due modelli), si ottengono le seguenti proxy del termine di errore:

$$1 + e_{a,t} = \frac{u_t}{1 - u_t} \tag{A.19}$$

$$1 + e_{b,t} = \frac{1 - u_t}{u_t} \ . \tag{A.20}$$

A.5.2 Metodi alternativi di stima per iETS$_\mathbf{I}$ e iETS$_\mathbf{D}$

iETS$_\mathbf{I}$

Ci sono due modi per stimare la probabilità di occorrenza nei modelli iETS$_\mathrm{I}$. Il primo consiste nella procedura illustrata nella sezione 2.5.2, mentre il secondo deriva dal fatto che p_t è inversamente proporzionale al valore di $\mu_{b,t}$. Questa proprietà implica che, ipotizzando che $\mu_{b,t}$ sia arrotondato per difetto, questo coincida con gli intervalli di domanda osservati. Ne consegue che è possibile applicare la sostituzione di $\hat{q}_{j_t} = \lfloor 1+\hat{\mu}_{b,t} \rfloor$, in modo tale da ottenere il meccanismo di aggiornamento delle probabilità del metodo di Croston:

$$\hat{p}_{j_t} = \frac{1}{\lfloor 1 + \hat{\mu}_{b,t} \rfloor} = \frac{1}{\hat{q}_{j_t}} \ . \tag{A.21}$$

Questo dimostra che il metodo di Croston è solo uno dei modi in cui si può stimare il modello iETS$_\text{I}$.

È importante osservare che, anche se $\hat{\mu}_{b,t}$ potrebbe variare nel tempo ad ogni osservazione (influenzando quindi $\hat{p}_t$), il modello iETS$_\text{I}$ che utilizza l'equazione (A.21) non è stimabile quando la domanda è pari a zero. Ne consegue che in questo modello è necessario ipotizzare che gli stati di $\mu_{b,t}$ non cambino nel periodo inter-domanda, che corrisponde all'ipotesi originale del metodo di Croston. Questo dimostra che l'iETS$_\text{I}$ è un modello che sottostà al metodo di Croston.

iETS$_\text{D}$

Se tutti i valori di $\mu_{a,t} \leq 1$, allora tutti i valori condizionati di iETS$_\text{D}$ corrispondono a quelli del modello ETS(M, N, N). Questa situazione si potrebbe verificare quando il livello è basso e converge a zero, i.e., la domanda è sparsa e diventa obsoleta. In questo caso, il modello iETS$_\text{D}$ può essere stimato usando il metodo TSB.

Usando la procedura descritta in sezione 2.5.2 per il modello iETS$_\text{D}$, il termine di errore è calcolato come

$$1 + e_{a,t} = \frac{o_t}{\hat{\mu}_{a,t}} \ .$$

Questa formula, tuttavia, è irrealistica, in quanto nel caso in cui $o_t = 0$, l'errore diventa pari a 0, rendendo il modello non stimabile. Per tale ragione, si introduce un'approssimazione del termine di arrore, in modo da garantire errori di previsioni diversi da zero nei casi limite:

$$e_{a,t} = \frac{o_t(1-2k) + k - \hat{\mu}_{a,t}}{\hat{\mu}_{a,t}} \tag{A.22}$$

dove k è un valore molto piccolo, la cui unica funzione è quella di rendere il modello stimabile.

Alternativamente, è possibile utilizzare il metodo TSB, il quale ha una connessione diretta col modello ETS(M, N, N) che sottostà il metodo SES. Questo significa che iETS$_\text{D}$ è un modello sottostante al TSB.

A.6 Funzione ausiliaria nell'algoritmo iterativo MM

La funzione definita in equazione (3.12) soddisfa le proprietà necessarie perché sia una "surrogata" appropriata di Φ_λ, i.e.:

(i) $\Psi_\lambda(\mathbf{a};\mathbf{a}) = \Phi_\lambda(\mathbf{a})$ per qualsiasi $\mathbf{a}$;

(ii) $\Psi_\lambda(\mathbf{w};\mathbf{a}) \leq \Phi_\lambda(\mathbf{w})$ per qualsiasi $\mathbf{a}$ e qualsiasi $\mathbf{w}$.

La proprietà (i) è immediata date le formule equazione (3.10) e equazione (3.12), mentre la (ii) deriva dalla disequazione che esprime la concavità della funzione logaritmica:

$$\sum_{i=1}^{N} b_{ti} \ln(x_{ti}) \leq \ln\left(\sum_{i=1}^{N} b_{ti} x_{ti}\right)$$

con $b_{ti} = \frac{\hat{P_{ti}} a_i}{\sum_{l=1}^{N} \hat{P}_{tl} a_l}$ e $x_{ti} = \frac{w_i}{a_i} \sum_{l=1}^{N} \hat{P}_{tl} a_l$, notando che $\sum_{i=1}^{N} b_{ti} = 1,\ \forall t$.

A.7 CSL e politica di inventario OUT

Come scritto nella sezione 1.1.1, si considera una politica di inventario OUT, la quale richiede di determinare l'intervallo di revisione e il livello OUT per ogni SKU. Il primo nella pratica solitamente è impostato come uguale per tutte le SKU, o per una classe, e varia in base al settore di interesse. Il secondo invece dovrebbe essere impostato separatamente per ogni SKU per tenere conto dell'incertezza della domanda. La criticità di questa scelta dipende dal contesto considerato, ma in ogni caso ha effetti rilevanti sugli inventari aggregati, e dipende dalle misure di servizio, dalla distribuzione della domanda e dal metodo di previsione.

Come misura del livello di servizio si considera il *customer service level* (CSL), ossia la proporzione di domanda soddisfatta. Nonostante il suo utilizzo non sia appropriato nell'ambito della domanda intermittente (Boylan & Syntetos, 2021), nel contesto del *newsvendor problem* in cui ci poniamo in questo testo, la misura coincide con la più appropriata Revised Customer Service Level (CSL^+). Quest'ultima richiede la valutazione separata delle probabilità della presenza di domanda sull'intervallo di revisione (condizionata) e durante il *lead time* (non condizionata). Nel *newsvendor problem* gli intervalli considerati coincidono e sono unitari, determinando la sovrapposizione di CSL e CSL^+.

Appendice B

Approfondimenti sull'analisi

B.1 Implementazione dei metodi

In questa sezione si esamineranno alcune delle implementazioni dei metodi introdotti nel capitolo 2 e nel capitolo 3.

Alcuni dei metodi sono già presenti su R, quali:

- ARIMA e ETS (pacchetto R `forecast`) ;
- Metodi standard per la domanda intermittente descritti nella sezione 2.4 (pacchetto R `tsintermittent` (Kourentzes, 2022)). Per questi metodi si ipotizza una distribuzione Binomiale Negativa con varianza pari a 1.1 volte il valore atteso (Syntetos et al., 2015);
- iETS (pacchetto R `smooth`).

Per quanto riguarda il calcolo dei pesi, si è utilizzata la funzione R `constrOptim` per l'ottimizzazione dei punteggi di Brier e DRPS e il pacchetto R `pso` (Bendtsen., 2022) per il PSO.

I metodi restanti, quali GAM-QR, modelli di distribuzione con media *damped* e WSS sono stati implementati seguendo le rispettive procedure descritte nel capitolo 2.

B.1.1 GAM-QR

GAM

Seguendo Wang et al. (2024), si adattano due modelli GAM, GAM-QR(noco) e GAM-QR(co), che utilizzano diversi insiemi di covariate per fare le previsioni. In particolare, GAM-QR(noco) usa solamente le informazioni riguardo alle vendite storiche (media delle vendite riguardanti i 7 e i 28 giorni precedenti), mentre GAM-QR(co) si avvale di un insieme di covariate ulteriori generate durante il processo di creazione dei *dataset*:

- Caratteristiche di calendario: la proporzione del giorno dell'anno, la proporzione del giorno della settimana, il tipo e il nome dell'evento, e l'esistenza di SNAP nel giorno in analisi.

- Caratteristiche del prezzo: prezzo originale, il prezzo relativo al prezzo massimo per categoria e dipartimento per un certo giorno.

Nella stima dei modelli, si utilizza il comando `gam` del pacchetto R `mgcv` (Wood, 2015), impostando una famiglia Binomiale Negativa con funzione legame logaritmica.

Nel modello con covariate si è inoltre deciso di applicare una regolarizzazione al processo di stima, impostando `irls.reg` di `gam.control` pari a 0.5, al fine di evitare problemi di convergenza. Il metodo IRLS alla base della stima dei GAM, infatti, può non riuscire a convergere in alcune circostanze, ad esempio nel caso di dati con alte frequenze di valori 0 in congiunzione con l'utilizzo di una funzione legame logaritmica: in questo caso, il problema di convergenza è causato dalla mancanza di identificabilità del modello dovuta al fatto che una media pari a 0 corrisponde ad un *range* infinito di valori del predittore lineare. Per evitare tale problema, si applica una penalizzazione regressione ridge[1] per imporre l'identificabilità del modello (Wood, 2015).

Inoltre, per garantire la convergenza dell'algoritmo anche in presenza della variabile esogena contenente il nome dell'evento (come fatto da Wang et al. (2024)), si imposta l'ottimizzatore "`bfgs`", un metodo *quasi-Newton* iterativo utile per l'ottimizzazione di problemi non lineari non vincolati. Essendo un metodo *quasi-*

[1] Aggiunge un termine costante alla diagonale della matrice di pesi per stabilizzare la procedura di stima dei parametri.

Newton, l'algoritmo utilizza un'approssimazione dell'inversa dell'Hessiana al posto di quella reale, spesso poco pratica e costosa da usare (Luenberger et al., 1984).

Regressione quantile

Si procede poi creando una funzione iterativa che per ogni quantile τ ($\tau \in \{0.01, 0.02, ..., 0.99\}$) stima il modello di regressione quantile per la trasformata $T(Z, \tau)$ come definito in equazione (2.38), tramite il comando `qr` del pacchetto R `quantreg` (Koenker, 2024).

Si applica la funzione usando gli effetti stimati dei due modelli GAM descritti sopra come covariate e si creano previsioni a 28 giorni per ogni quantile stimato. Si ricava poi $Q_Z(\tau|\hat{\boldsymbol{f}})$ in base all'equazione (2.36) e si utilizza per calcolare i quantili della variabile di interesse: $Q_Y(\tau|\mathbf{x}) = \lceil Q_Z(\tau|\mathbf{x}) - 1 \rceil$.

B.1.2 Distribuzioni Poisson e Binomiale Negativa con media *damped*

Per entrambe le distribuzioni, si sono inizializzati i parametri $\phi = \alpha = 0.1$ e μ come la media di tutte le vendite contenute nei dati di stima e si è usata la funzione `maxLik` con `method = "bfgs"` del pacchetto R `maxLik` (Henningsen & Toomet, 2011) per calcolare le stime di massima verosimiglianza dei parametri usando le rispettive funzioni distributive. È stata poi applicata la funzione dell'equazione (2.39) per calcolare iterativamente i parametri tempo dipendenti della distribuzione, il cui valore finale è stato usato per iniziare la procedura iterativa di previsione. Per ogni orizzonte temporale nei dati di validazione e di verifica, sono state calcolati 1000 valori simulati dalle due distribuzioni, che sono stati poi usati per calcolare i quantili di interesse.

Nel caso della Binomiale Negativa, si sono usati $a_t = b\mu_t$ e $\frac{b}{1+b}$ come *input* delle funzioni `dnbinom` e `rnbinom`, che corrispondono rispettivamente al numero di successi per periodo desiderati e la probabilità di avere un successo. Nei casi limite in cui b prende valori iniziali o stimati non positivi o maggiori di 99, indicativi della presenza di sottodispersione, la Binomiale Negativa è stata sostituita con una distribuzione Poisson.

B.2 Metriche delle diverse categorie di domanda intermittente

Nelle Tabella B.1, Tabella B.2, Tabella B.3 e Tabella B.4 sono riportati i risultati in termini di *sharpness* per le diverse categorie di domanda intermittente. Nelle Tabella B.5, Tabella B.6, Tabella B.7 e Tabella B.8 sono riportati i risultati in termini di costi simulati per le diverse categorie di domanda intermittente.

Tipo	**Metodo**	**Logaritmico**	**DRPS**	**Brier**
Individuali	GAM-QR(co)	2.441	**1.453**	**-0.163**
	GAM-QR(noco)	**2.406**	1.514	-0.160
	Croston	3.406	1.745	-0.128
	SBA	3.372	1.772	-0.127
	TSB	3.413	1.742	-0.129
	ARIMA	2.695	1.695	-0.126
	ETS	2.684	1.761	-0.126
	WSS	3.124	2.134	-0.086
	Poisson	2.843	1.713	-0.134
	NB	2.682	1.834	-0.124
	$iETS_I$	2.605	1.647	-0.105
	$iETS_G$	2.515	1.607	-0.135
	$iETS_D$	2.549	1.625	-0.133
	$iETS_O$	2.510	1.633	-0.130
Combinazioni	brier-opt	2.531	1.547	-0.143
	drps-opt	10.306	4.840	0.069
	cost19-opt	2.687	1.574	-0.136
	cost4-opt	2.661	1.543	-0.133
	cost9-opt	2.697	1.583	-0.134
	SA	2.526	1.538	**-0.144**
	mediana	3.031	1.531	-0.133
	log-score	2.609	1.571	-0.142
	log-opt	**2.367**	**1.442**	-0.163
	cl-opt	2.372	1.646	-0.134

Tabella B.1: Risultati in termini di *sharpness* per la domanda *smooth*.

Tipo	Metodo	Logaritmico	DRPS	Brier
Individuali	GAM-QR(co)	2.420	**1.618**	**-0.212**
	GAM-QR(noco)	**2.327**	1.632	-0.208
	Croston	6.721	3.002	-0.057
	SBA	6.584	2.881	-0.059
	TSB	6.440	2.852	-0.060
	ARIMA	3.048	2.704	-0.085
	ETS	3.143	2.881	-0.082
	WSS	3.066	2.215	-0.116
	Poisson	5.700	2.665	-0.078
	NB	2.715	2.360	-0.112
	$iETS_I$	3.226	2.156	-0.065
	$iETS_G$	2.872	2.045	-0.116
	$iETS_D$	2.999	1.994	-0.131
	$iETS_O$	3.057	2.088	-0.104
Combinazioni	brier-opt	4.598	2.403	-0.092
	drps-opt	2.528	**0.583**	-0.318
	cost9-opt	4.053	2.185	-0.101
	cost19-opt	4.488	2.273	-0.094
	cost4-opt	3.687	2.003	-0.107
	log-opt	**2.540**	1.586	**-0.220**
	cl-opt	2.698	2.313	-0.109
	SA	4.743	2.234	-0.092
	mediana	3.773	2.084	-0.097
	log-score	5.384	2.515	-0.087

Tabella B.2: Risultati in termini di *sharpness* per la domanda erratica.

Tipo	Metodo	Logaritmico	DRPS	Brier
Individuali	GAM-QR(co)	1.887	1.041	-0.345
	GAM-QR(noco)	**1.837**	**1.022**	**-0.344**
	Croston	2.820	1.186	-0.283
	SBA	2.878	1.191	-0.282
	TSB	2.856	1.188	-0.281
	ARIMA	2.375	1.323	-0.168
	ETS	2.396	1.355	-0.169
	$iETS_I$	2.423	1.143	-0.266
	$iETS_G$	2.192	1.102	-0.300
	$iETS_D$	2.299	1.099	-0.300
	$iETS_O$	2.288	1.150	-0.265
	WSS	2.296	1.233	-0.277
	Poisson	2.361	1.156	-0.293
	NB	2.269	1.181	-0.282
Combinazioni	brier-opt	2.020	1.091	-0.289
	drps-opt	**1.663**	**0.846**	**-0.464**
	cost9-opt	2.197	1.085	-0.286
	cost19-opt	2.195	1.098	-0.285
	cost4-opt	2.186	1.072	-0.289
	log-opt	2.059	1.016	-0.310
	cl-opt	1.950	1.206	-0.260
	SA	2.008	1.074	-0.301
	mediana	2.470	1.061	-0.299
	log-score	2.175	1.120	-0.285

Tabella B.3: Risultati in termini di *sharpness* per la domanda *lumpy*.

Tipo	Metodo	Logaritmico	DRPS	Brier
Individuali	GAM-QR(co)	1.166	**0.464**	**-0.492**
	GAM-QR(noco)	**1.153**	0.450	-0.491
	Croston	1.391	0.479	-0.474
	SBA	1.402	0.479	-0.474
	TSB	1.384	0.477	-0.476
	ARIMA	1.572	0.540	-0.356
	ETS	1.562	0.536	-0.361
	WSS	1.340	0.530	-0.440
	Poisson	1.368	0.484	-0.468
	NB	1.654	0.516	-0.443
	$iETS_O$	1.447	0.551	-0.427
	$iETS_I$	1.459	0.497	-0.454
	$iETS_G$	1.308	0.484	-0.475
	$iETS_D$	1.413	0.483	-0.475
Combinazioni	brier-opt	**1.150**	**0.451**	**-0.503**
	drps-opt	1.424	0.484	-0.467
	cost9-opt	1.291	0.488	-0.470
	cost19-opt	1.303	0.489	-0.469
	cost4-opt	1.286	0.488	-0.469
	log-opt	1.321	0.464	-0.466
	cl-opt	1.255	0.524	-0.438
	SA	1.215	0.478	-0.475
	mediana	1.379	0.468	-0.472
	log-score	1.282	0.489	-0.470

Tabella B.4: Risultati in termini di *sharpness* per la domanda "strettamente" intermittente.

Metodo	Cost(1,4)	Cost(1,9)	Cost(1,19)
GAM-QR(co)	**4.109**	5.695	7.433
GAM-QR(noco)	4.312	5.918	7.470
Croston	4.878	7.209	10.165
SBA	5.039	7.548	10.879
TSB	4.928	7.281	10.331
ARIMA	4.811	6.337	7.851
ETS	4.978	6.489	8.008
WSS	6.091	7.771	9.523
Poisson	4.824	6.699	8.690
NB	5.107	7.319	9.818
iETS_{I}	4.726	6.570	8.683
iETS_{I}	4.602	6.539	8.726
iETS_{D}	4.585	6.473	8.635
iETS_{O}	4.643	6.571	8.743
brier-opt	4.398	6.087	4.940
drps-opt	12.113	21.418	35.785
cost19-opt	4.437	6.092	5.276
cost4-opt	4.463	6.103	5.076
cost9-opt	4.590	**4.174**	5.299
SA	4.381	6.071	4.910
mediana	4.253	5.946	7.589
log-score	4.443	6.137	4.975
log-opt	4.111	5.663	7.271
cl-opt	4.713	6.893	**1.744**

Tabella B.5: Costi relativi dei metodi per le serie *smooth*.

Metodo	Cost(1,4)	Cost(1,9)	Cost(1,19)
GAM-QR(co)	5.008	7.183	9.267
GAM-QR(noco)	5.006	7.149	9.279
Croston	7.074	10.055	14.325
SBA	6.901	9.786	13.747
TSB	6.822	9.756	13.814
ARIMA	7.610	9.965	12.178
ETS	8.095	10.556	12.893
$iETS_I$	6.711	9.778	12.991
$iETS_G$	6.275	9.161	12.433
$iETS_D$	6.098	8.914	12.138
$iETS_O$	6.455	9.411	12.778
WSS	7.335	10.498	13.641
Poisson	7.011	9.851	13.366
NB	7.969	11.981	16.055
brier-opt	6.790	9.535	12.629
drps-opt	4.937	7.478	10.236
cost9-opt	6.083	8.513	11.111
cost19-opt	6.179	8.518	11.033
cost4-opt	5.811	8.246	11.152
log-opt	**4.823**	**6.955**	**9.265**
cl-opt	7.343	10.990	15.037
SA	6.360	8.994	12.024
mediana	6.283	8.868	11.732
log-score	6.672	9.262	12.117

Tabella B.6: Costi relativi dei metodi per le serie erratiche.

Metodo	Cost(1,4)	Cost(1,9)	Cost(1,19)
GAM-QR(co)	3.842	8.911	104.257
GAM-QR(noco)	3.798	6.397	9.888
Croston	3.741	5.982	9.278
SBA	3.793	6.112	9.573
TSB	3.770	6.048	9.388
ARIMA	3.797	5.268	6.955
ETS	3.889	5.375	**7.062**
$iETS_I$	3.993	6.274	9.125
$iETS_G$	3.805	5.969	8.593
$iETS_D$	3.786	5.926	8.599
$iETS_O$	3.915	6.092	8.755
WSS	4.322	6.428	8.832
Poisson	3.732	5.735	8.357
NB	4.117	6.234	8.663
brier-opt	3.568	5.872	25.362
drps-opt	**2.831**	4.308	7.456
cost9-opt	3.551	5.734	22.142
cost19-opt	3.584	5.563	19.799
cost4-opt	3.489	5.346	17.933
log-opt	3.332	5.859	18.575
cl-opt	4.131	8.868	58.285
SA	3.554	5.856	24.639
mediana	3.495	**5.267**	7.287
log-score	3.732	6.938	52.319

Tabella B.7: Costi relativi dei metodi per le serie *lumpy*.

Metodo	Cost(1,4)	Cost(1,9)	Cost(1,19)
GAM-QR(co)	8.795	14.861	32.638
GAM-QR(noco)	1.507	2.165	2.833
Croston	1.619	2.382	3.231
SBA	1.628	2.404	3.279
TSB	1.617	2.379	3.232
ARIMA	1.593	2.157	2.806
ETS	1.585	2.146	2.787
WSS	1.846	2.530	3.177
Poisson	1.639	2.381	3.233
NB	1.817	2.665	3.659
$iETS_O$	1.862	2.696	3.692
$iETS_I$	1.760	2.623	3.617
$iETS_G$	1.700	2.547	3.484
$iETS_D$	1.699	2.553	3.534
brier-opt	1.989	3.044	4.177
drps-opt	1.656	2.439	3.312
cost9-opt	2.058	3.165	4.293
cost19-opt	2.086	3.199	4.399
cost4-opt	2.091	3.179	4.371
log-opt	**1.55**	2.239	2.980
cl-opt	2.989	4.985	7.641
SA	2.032	3.112	4.243
mediana	1.571	**2.220**	**2.892**
log-score	2.775	4.486	6.311

Tabella B.8: Costi relativi dei metodi per le serie strettamente intermittente.

Bibliografia

Bendtsen., C. (2022). *pso: Particle Swarm Optimization*. R package version 1.0.4, https://CRAN.R-project.org/package=pso.

Blahut, R. E. (2002). Information theory and coding. In Middleton, W. M. & Van Valkenburg, M. E., editors, *Reference Data for Engineers*, pages 25(1–31). Newnes, Woburn, 9th edition edition.

Boylan, J. & Syntetos, A. (2021). *Intermittent Demand Forecasting: Context, Methods and Applications*. Wiley.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statistics & probability letters*, 77(14):1473–1478.

Brown, R. G. (1956). *Exponential smoothing for predicting demand*. Little.

CMAF, L. (2023). CMAF FFT: Intermittent Demand Forecasting. YouTube. https://www.youtube.com/watch?v=PRidAjQRtCs.

Conflitti, C., De Mol, C., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096–1103.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3):289–303.

Dmitry, I., Alexander, T., & Jörn, S. (2019). *Global Supply Chain and Operations Management A Decision-Oriented Introduction to the Creation of Value*. Spinger, 3rd edition.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 24(1):1–26.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4):1283–1318.

Gaillard, P., Goude, Y., & Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32(3):1038–1050.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268.

Gneiting, T. & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Hall, S. G. & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13.

Hamilton, J. D. (1994). State-space models. *Handbook of econometrics*, 4:3039–3080.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Henningsen, A. & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458.

Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, 278(3):904–915.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts, 2nd edition. https://otexts.com/fpp2/.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–454.

Kaya, G. O., Sahin, M., & Demirel, O. F. (2020). Intermittent demand forecasting: a guideline for method selection. *Sādhanā*, 45:1–7.

Kennedy, J. & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948.

Koenker, R. (2024). *quantreg: Quantile Regression*. R package version 5.98. https://CRAN.R-project.org/package=quantreg.

Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32(3):788–803.

Kourentzes, N. (2022). *tsintermittent: Intermittent Time Series Forecasting*. R package version 1.10. https://CRAN.R-project.org/package=tsintermittent.

Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, 225:107597.

Luenberger, D. G., Ye, Y., et al. (1984). *Linear and nonlinear programming*, volume 2. Springer.

Machado, J. A. F. & Silva, J. M. C. S. (2005). Quantiles for Counts. *Journal of the American Statistical Association*, 100(472):1226–1237.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Nikolopoulos, K. (2021). We need to talk about intermittent demand forecasting. *European Journal of Operational Research*, 291(2):549–559.

Ord, J. K., Snyder, R. D., Koehler, A. B., Hyndman, R. J., & Leeds, M. (2005). Time series forecasting: the case for the single source of error state space approach. *Unpublished manuscript, Monash University*.

Orlitsky, A. (2003). Information Theory. In Meyers, R. A., editor, *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 751–769. Academic Press, New York, 3rd edition edition.

Petropoulos, F. et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sanguri, K., Patra, S., Nikolopoulos, K., & Punia, S. (2024). Intermittent demand, inventory obsolescence, and temporal aggregation forecasts. *International Journal of Production Research*, 62(5):1663–1685.

Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). Forecasting the intermittent dem& for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2):485–496.

Svetunkov, I. (2023). *Forecasting and analytics with the augmented dynamic adaptive model (ADAM)*. CRC Press.

Svetunkov, I. & Boylan, J. E. (2023). iETS: State space model for intermittent demand forecasting. *International Journal of Production Economics*, 265:109013.

Svetunkov, I. & Boylan, J. E. (2024). Staying positive: challenges and solutions in using pure multiplicative ETS models. *IMA Journal of Management Mathematics*, 35(3):403–425.

Syntetos, A. A., Babai, M. Z., & Gardner Jr, E. S. (2015). Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. *Journal of Business Research*, 68(8):1746–1752.

Syntetos, A. A. & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71(1-3):457–466.

Syntetos, A. A. & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of forecasting*, 21(2):303–314.

Syntetos, A. A., Boylan, J. E., & Croston, J. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society*, 56:495–503.

Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615.

Trapero, J. R., Cardós, M., & Kourentzes, N. (2019). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, 35(1):239–250.

Wang, D., Tan, D., & Liu, L. (2018). Particle swarm optimization algorithm: an overview. *Soft computing*, 22(2):387–408.

Wang, S., Kang, Y., & Petropoulos, F. (2024). Combining probabilistic forecasts of intermittent demand. *European Journal of Operational Research*, 315(3):1038–1048.

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547.

Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of forecasting*, 20(3):375–387.

Wood, S. (2015). *Package 'mgcv'*. R package version 1.9-1, https://CRAN.R-project.org/package=mgcv.

Zhang, W., Zhang, R., Shang, R., Li, J., & Jiao, L. (2019). Application of natural computation inspired method in community detection. *Physica A: Statistical Mechanics and its Applications*, 515:130–150.

Zhou, C. & Viswanathan, S. (2011). Comparison of a new bootstrapping method with parametric approaches for safety stock determination in service parts inventory systems. *International Journal of Production Economics*, 133(1):481–485.