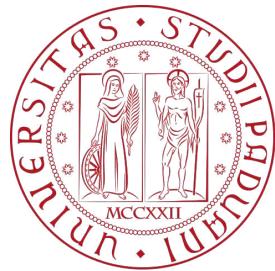


Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche



**STUDIO DELLA DEGRADAZIONE TEMPORALE
NEI MODELLI DI MACHINE LEARNING**

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Laureando: Marco Pellizzari
Matricola N 2054751

Anno Accademico 2023/2024

Desidero ringraziare il mio relatore, il Professor Bruno Scarpa, per il suo prezioso supporto e la sua guida, che hanno reso possibile questo lavoro.

Ringrazio in particolare la mia famiglia, che mi ha sempre sostenuto
durante questo difficile percorso.

Infine ringrazio i miei amici, per essere sempre stati al mio fianco.

Indice

Introduzione	1
1 Metodologia e definizioni	7
1.1 “Test” di <i>degradazione temporale</i>	8
1.2 Approccio simulativo	16
1.2.1 <i>Concept drift</i>	23
1.2.2 Possibili cause di <i>degradazione temporale</i>	26
1.2.3 Dataset simulati	27
1.2.4 Applicazione del “test” di <i>degradazione temporale</i>	31
2 Analisi su dati reali	37
2.1 Dati sugli accessi alle cliniche	38
2.2 Preparazione dei dataset	43
2.3 Analisi della <i>degradazione temporale</i>	47
3 Simulazioni: assenza di <i>concept drift</i>	61
Riassunto dei risultati	81
4 Simulazioni: <i>concept drift</i>	87
Riassunto dei risultati	116
5 Simulazioni: stagionalità	125
Riassunto dei risultati	141
Conclusioni	147
Bibliografia	151

A Materiale capitolo 3	153
B Materiale capitolo 4	173
C Materiale capitolo 5	199
D Materiale aggiuntivo	219

Introduzione

L'utilizzo di *intelligenza artificiale* (AI) e di *machine learning* (ML) continua ad aumentare nelle applicazioni pratiche. In un contesto come questo, la robustezza e l'affidabilità di tali strumenti, intese come la capacità di produrre buoni risultati in modo consistente, con supervisione minima, diventa sempre più importante.

L'approccio corrente all'impiego di metodologie di questo tipo non sembra però dare sufficiente importanza a tali aspetti. Spesso, nella realtà d'uso, i modelli vengono costruiti sulla base dei dati più recenti a disposizione e, una volta perfezionati per raggiungere le prestazioni desiderate, vengono impiegati senza ulteriori aggiornamenti.

I processi che generano dati però spesso evolvono (Žliobaitė et al., 2016), e le conseguenze di ciò sono chiare: i modelli, addestrati su uno specifico insieme di dati, diventano, con il passare del tempo, meno affidabili nello svolgere la funzione per la quale sono stati costruiti. Questo fenomeno prende il nome di *concept drift*, termine utilizzato per descrivere le situazioni in cui le proprietà statistiche di un processo cambiano nel tempo (Gama et al., 2014; Lu et al., 2018). Questi cambiamenti sono noti per ridurre le prestazioni dei modelli: più il vero processo generatore si discosta da quello su cui i modelli sono stati adattati, più la qualità della previsione peggiora.

Nella pratica esistono metodi per limitare l'impatto di questi *drift* sulle prestazioni dei modelli, i più noti sono il *continuous learning* e l'*online learning*. Tramite l'aggiornamento automatico dei modelli all'arrivo di nuovi dati, questi possono rimanere al passo con il contesto che evolve e mantenere una qualità adeguata (Gama et al., 2014).

Questi approcci hanno però dei limiti. Il primo è la necessità di sviluppare metodi efficienti e robusti per aggiornare in modo automatico il modello. Gli aggiornamenti sono infatti spesso onerosi, in termini di tempo, e la necessità di automatizzare completamente il processo ne rende l'impiego più complesso. Per fare questo, è indispensabile non solo individuare delle condizioni, sui dati in arrivo, che innescino l'aggiornamento (minimizzando così il numero di volte in cui eseguirlo), ma anche assicurarsi che le anomalie vengano gestite correttamente. Il secondo limite è la necessità di avere sempre accesso ai nuovi dati e ai veri valori della variabile risposta; tutto questo limita la reale applicabilità di queste soluzioni.

Viste le difficoltà nell'aggiornare un modello, si presenta il problema delle prestazioni a medio/lungo termine: la qualità iniziale del modello si mantiene nel tempo, quando questo non viene aggiornato? E se no, di quanto peggiora?

Un primo studio del fenomeno è contenuto nell'articolo “*Temporal quality degradation in AI models*” di Vela et al. (2022), che introduce il termine *degradazione temporale* per definire i casi in cui le prestazioni di un modello, stimato utilizzando un certo insieme di dati e mai aggiornato, si riducono all'allontanarsi dal momento della stima. Un modello che presenta un decadimento della qualità è un modello che invecchia.

Gli autori cercano di studiare la *degradazione temporale* di alcuni modelli sfruttando un vasto numero di dataset reali, provenienti da ambiti diversi, in cui i valori della variabile risposta non presentano cambiamenti improvvisi. Questi cambiamenti improvvisi nei comportamenti osservati sono infatti segno di evoluzioni altrettanto rapide della *distribuzione sottostante*, e in quei casi la riduzione delle prestazioni è attesa. Questi sono gli stessi segni che suggeriscono l'impiego delle soluzioni indicate in precedenza, per l'aggiornamento dei modelli.

L'idea dello studio è di valutare la stabilità temporale delle prestazioni quando i metodi menzionati non vengono applicati; quando non ci sono i segnali di potenziali problemi.

Nell'articolo sono stati studiati quattro diversi algoritmi di *machine learning* di comune utilizzo nell'ambito della regressione: la *Regressione Lineare Penalizzata Ridge* (Hoerl et al., 1970), la *Foresta Casuale* (Breiman 2001), il *Gradient Boosting* (Friedman, 2001) e la *Rete Neurale* (LeCun et al., 2015), con il sospetto iniziale che diverse logiche matematiche rispondessero in modo diverso alla prova del tempo. Nello studio della *degradazione temporale* è stata utilizzata una procedura proposta dagli autori, il “test” di *degradazione temporale*, che è stata riadattata ed utilizzata durante questo lavoro di tesi.

I risultati rilevanti ai fini di questo lavoro sono riassunti di seguito:

1. Le prestazioni dei modelli considerati possono presentare segni di *degradazione temporale* (peggioramenti della qualità nel tempo) anche in contesti in cui il *drift* è *minimo o assente*; in cui il comportamento della variabile risposta rimane regolare. Quindi anche in situazioni in cui non c’è motivo di sospettare una riduzione della qualità del modello, le prestazioni possono ridursi con il passare del tempo.
2. La *degradazione temporale* può presentarsi in modo differente per le diverse categorie di modello, sugli stessi dati: *l’evoluzione temporale della qualità dipende in parte dai dati, e in parte dal modello utilizzato*.

Lo studio suggerisce quindi che la stabilità temporale delle prestazioni sia determinata dalle caratteristiche dello specifico modello utilizzato e dal contesto di impiego; in particolare, gli autori affermano che si sappia ben poco delle proprietà di stabilità numerica di questi metodi (di ML), e che gli sforzi per studiarli in questo senso siano stati scarsi.

La componente temporale dei dati deve perciò essere approfondita nelle applicazioni pratiche. Il modello migliore, individuato utilizzando i dati disponibili, non rimane necessariamente il migliore a medio/lungo termine: nella scelta va considerata la stabilità temporale delle prestazioni.

Con l’intenzione di approfondire questo aspetto legato all’utilizzo dei modelli di *machine learning*, lo studio presentato in queste pagine cerca di individuare le caratteristiche dei dati che influiscono sulla stabilità a medio/lungo

termine delle stesse quattro logiche matematiche, sempre nell'ambito della regressione.

Partendo dai risultati e dalle proposte degli autori, questo lavoro cerca di fare luce su come le diverse famiglie di modelli rispondono a diverse situazioni, provando ad evidenziarne le differenze.

Il raggiungimento dello scopo permetterebbe di scegliere il modello migliore per svolgere un compito non solo sulla base della sua qualità iniziale, ma considerando anche la stabilità delle prestazioni a medio/lungo termine. Potrebbe inoltre essere possibile individuare quali casi, e per quali modelli, *online learning* e *continuous learning* sono veramente necessari.

Per raggiungere questi obiettivi è stato condotto uno studio di simulazione: per ciascuna situazione di interesse sono stati simulati dei dati su cui applicare la metodologia descritta in Vela et al. (2022), e i risultati combinati. Ciò permette di capire come uno specifico modello risponde ad una data situazione, e il confronto tra le diverse categorie considerate permette di individuare i più stabili.

Il lavoro si articola nei seguenti capitoli:

1. Nel capitolo 1 viene presentata la metodologia utilizzata per raggiungere gli obiettivi, quindi il “test” di *degradazione temporale* e l’approccio di simulazione.
2. Nel capitolo 2 il “test” viene applicato su alcuni dati reali, provenienti da Vela et al. (2022), per capire come utilizzare l’informazione sulla stabilità per selezionare il modello (un aspetto tralasciato nell’articolo), combinandola con le prestazioni iniziali; ciò ha fornito alcuni spunti per sviluppare lo studio di simulazione.
3. Nei capitoli 3, 4 e 5 sono invece presentati i risultati delle simulazioni condotte.

Questo approccio non ha permesso di evidenziare delle differenze “significative” tra le diverse logiche matematiche considerate, ma ha invece permesso

di evidenziare alcune criticità dello strumento impiegato e dell'approccio al problema, che li rendono inadatti a studiare il fenomeno della stabilità delle prestazioni a medio/lungo termine. Al contrario, sono state messe in evidenza delle similitudini nel modo in cui i modelli invecchiano sugli stessi dati, di fronte alle stesse problematiche.

Per produrre i risultati presentati sono stati utilizzati R e Python. Il codice utilizzato e gli ambienti, che permettono di riprodurre le analisi, sono disponibili al seguente link: “<https://github.com/marco-pll/MarcoPellizzariLM.git>”.

Capitolo 1

Metodologia e definizioni

Questo capitolo presenta la metodologia impiegata per raggiungere l’obiettivo introdotto. Questo può essere definito, formalmente, come “evidenziare delle differenze nel modo in cui le prestazioni dei modelli evolvono nel tempo (e quindi nella *degradazione temporale*) quando questi non vengono aggiornati”. Ciò permetterebbe di ottenere delle informazioni utili per selezionare un modello da utilizzare, in una data circostanza, non solo sulla base della qualità iniziale raggiunta, ma anche considerando la stabilità della qualità nel tempo, anche in contesti in cui la quantità di dati disponibile è limitata e quindi il “test” di *degradazione temporale* non applicabile.

In questo lavoro sono state considerate le stesse quattro logiche matematiche studiate in Vela et al. (2022), quindi la *Regressione Lineare Penalizzata Ridge*, la *Foresta Casuale (RF)*, il *Gradient Boosting (GB)* e la *Rete Neurale (NN)*. Nella prima sezione (1.1) viene presentato il fenomeno della *degradazione temporale* e la procedura proposta nell’articolo per studiarla. Nella seconda sezione (1.2) viene invece presentato il nuovo approccio al problema, lo studio di simulazione, specifico di questo lavoro, entrando nel dettaglio della metodologia utilizzata per simulare i dataset, dell’aspetto computazionale, del fenomeno del *concept drift* e delle situazioni in cui la stabilità dei modelli viene valutata.

1.1 “Test” di *degradazione temporale*

La *degradazione temporale* della qualità può presentarsi in una tipica procedura di sviluppo e impiego di soluzioni di ML: scelto un dataset su cui adattare un modello (tipicamente l’insieme di dati più recente a disposizione) questo viene regolato, perfezionato, e poi impiegato in un periodo futuro, senza essere aggiornato. Questa procedura viene rappresentata nella figura 1.1.



Figura 1.1: Procedura di sviluppo e impiego di un modello di ML. L’insieme di stima viene utilizzato per adattare il modello, con l’obiettivo di approssimare la relazione tra variabili esplicative (X) e variabile risposta (Y).

La *degradazione temporale* si ha in quei casi in cui la qualità previsiva del modello, impiegato sui dati futuri, tende a ridursi all’allontanarsi dal momento della stima. Il verificarsi o meno di questo fenomeno è strettamente legato allo specifico modello utilizzato e al contesto applicativo.

La conoscenza pregressa delle proprietà di stabilità temporale dei diversi modelli di *machine learning* permetterebbe ad un loro utilizzatore di scegliere la metodologia più adatta a garantire buone prestazioni a medio/lungo termine, dato il contesto di impiego.

Il “test” (tra virgolette perché non è un test statistico, ma solo un termine utilizzato dagli autori dell’articolo) di *degradazione temporale*, proposto in Vela et al. (2022), permette di studiare e visualizzare come le prestazioni di uno specifico modello, utilizzato nel contesto della regressione, variano su uno specifico insieme di dati all’allontanarsi dal momento della sua stima;

questo viene fatto tramite l’emulazione dell’impiego pratico precedentemente descritto.

Per poter condurre l’analisi è necessario disporre di diversi anni di rilevazione, con osservazioni marcate temporalmente (con l’indicazione della posizione temporale nel dataset, come una data).

Questo “test” consiste in una procedura iterativa, descritta tramite i seguenti passi:

1. Scegliere un sottoinsieme di dati su cui regolare e adattare il modello, con l’obiettivo di prevedere la variabile risposta (Y) in funzione delle esplicative (X). Nella prima iterazione viene utilizzato il sottoinsieme meno recente a disposizione.
Questo insieme di stima viene costruito utilizzando la marcatura temporale delle osservazioni, che dovranno essere contigue.
2. Calcolare l’*MSE* (*errore quadratico medio*) nel periodo immediatamente successivo a quello di stima (indicato con $MSE(t_0)$, l’errore iniziale), che verrà utilizzato come punto di riferimento per misurare le variazioni nell’errore di previsione. L’insieme di dati utilizzato per il calcolo prende il nome di insieme di verifica.
3. Allontanarsi dall’insieme di stima e calcolare l’*MSE* tramite una finestra mobile, in modo da visualizzare l’errore come funzione del tempo passato dal momento della stima ($MSE(t_0 + dT)$, dove dT è il tempo intercorso, chiamato “età del modello”). Possiamo quindi misurare la variazione nelle prestazioni tramite il calcolo dell’errore relativo:

$$e_{rel}(dT) = \frac{MSE(t_0 + dT)}{MSE(t_0)}$$

dove $e_{rel}(dT)$ indica l’errore commesso dal modello a distanza dT , *relativamente* al periodo iniziale t_0 .

4. Ripartire da 1. con un nuovo sottoinsieme di dati, mantenendo la stessa ampiezza per l’insieme ma spostandosi verso osservazioni più recenti.

La scelta del numero di iterazioni (n) è arbitraria. La sequenza di passi viene rappresentata nell'immagine 1.2.

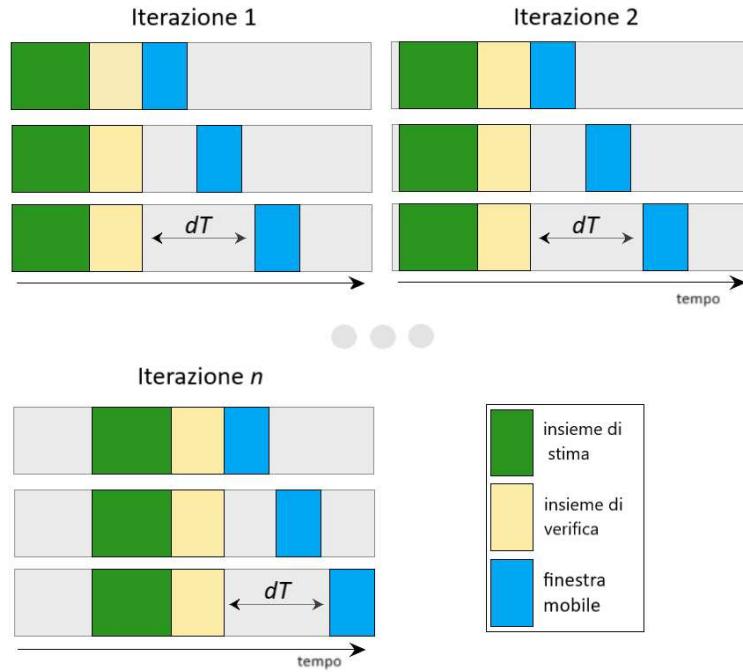


Figura 1.2: Rappresentazione della procedura per lo studio della *degradazione temporale*. L'insieme di stima viene usato per l'adattamento del modello, l'insieme di verifica per il calcolo dell'errore iniziale, l' $MSE(t_0)$, mentre la finestra mobile è usata per misurare l' $MSE(t_0 + dT)$, al variare della distanza dal momento della stima.

Una volta conclusa la procedura abbiamo a disposizione n tracciati, misure dell'*errore relativo* del modello al variare dell'insieme di stima utilizzato. Questi tracciati vengono quindi combinati, in modo intuitivo, per studiare l'evoluzione temporale della qualità del modello: volendo avere un'idea di quanto variano le prestazioni a dT giorni dal momento della stima possiamo prendere, dagli n tracciati, gli *errori relativi* alla distanza corrispondente ($e_{rel}(dT)$) e calcolarne *primo, secondo e terzo quartile*. Questo permette di visualizzare una distribuzione dell'errore, condizionatamente all'età del modello (dT).

Possiamo quindi rappresentare i risultati tramite il grafico di *AI aging*, da cui possiamo analizzare la *degradazione temporale* (alcuni esempi, tratti da Vela et al. (2022), sono riportati in figura 1.3). Il tracciato del 75esimo percentile rappresenta l’andamento dell’errore nel caso peggiore, quello del 25esimo percentile l’andamento nel caso migliore, mentre il tracciato della mediana l’andamento nel caso mediano.

Si è in presenza di *degradazione temporale* della qualità del modello in uno di due casi:

1. Il tracciato dell’errore relativo mediano presenta un trend crescente (grafici 2, 3 e 4 in figura 1.3): ciò indica che, allontanandosi dal momento della stima, le prestazioni del modello tendono a ridursi. In assenza di *degradazione temporale* il livello dovrebbe essere attorno ad 1, indicando che le prestazioni iniziali si mantengono.
2. La variabilità dell’errore relativo (la distanza tra i tracciati dei quartili, che rappresentano il caso migliore e peggiore), aumenta con il tempo trascorso: ciò indica che le prestazioni del modello diventano più imprevedibili (grafico 4 in figura 1.3).

Un modello che presenta *degradazione temporale* è un modello che invecchia. Nel caso i 3 quantili rimangano ad un livello costante, come nel grafico 1 di figura 1.3, possiamo dire che non c’è *degradazione* nella qualità.

I risultati ottenuti per ciascun modello possono quindi essere confrontati: i tracciati dei percentili possono essere utilizzati per valutare come e quanto variano le prestazioni nel tempo, permettendo di identificare i modelli più e meno stabili sullo specifico set di dati.

Alcuni limiti dell’approccio sono i seguenti:

1. Il prodotto di questa procedura è il grafico di *AI Aging*, che fornisce un’indicazione approssimativa dell’invecchiamento di un modello. È adeguato quando le differenze tra i modelli sono evidenti, in quanto

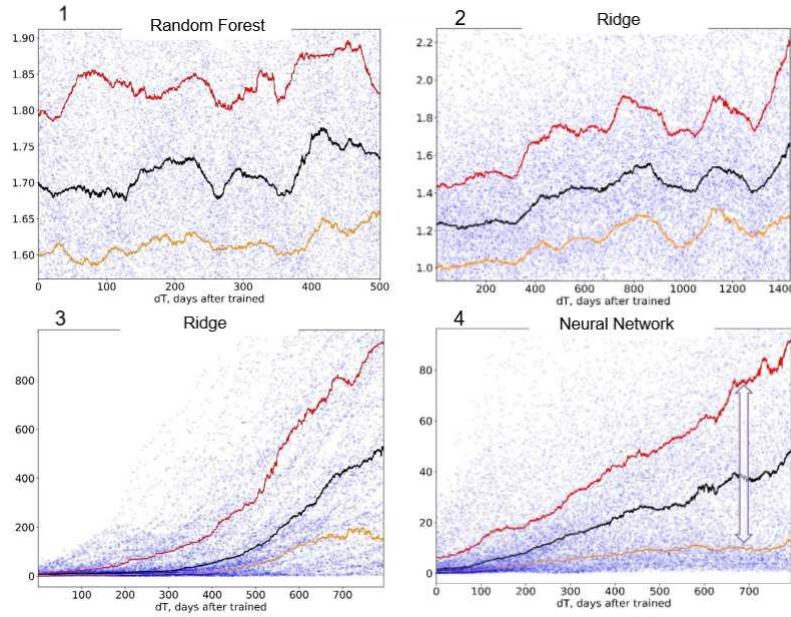


Figura 1.3: Grafici di *AI Aging*. Ciascun grafico fa riferimento ad un dataset diverso e ad uno specifico modello. Il singolo punto rappresenta un errore relativo ad una certa distanza (asse delle ascisse) dal momento di stima. La linea gialla è il tracciato del primo quartile, quella rossa del terzo, mentre la nera della mediana. Fonte: Vela et al. (2022).

possono essere confrontate le forme complessive, ma non permette considerazioni più precise. Nel caso in cui le differenze siano più ridotte, come accadrà nel capitolo 2, l'utilità del grafico è minima.

2. Nei casi in cui le differenze siano minime il solo grafico è di utilità limitata, non conservando l'informazione sulla qualità dei modelli. Supponendo di voler confrontare due modelli possiamo valutare se un modello è più stabile di un altro, ma senza la qualità iniziale non è semplice sapere quale sia il migliore.

Tutto ciò indica che il “test” vada utilizzato per studiare il comportamento complessivo, le tendenze, di un modello alla *degradazione temporale*, e che sia più importante la forma osservata nel grafico rispetto ai livelli.

Detto ciò, è possibile ricavare delle statistiche aggiuntive che possano aiutare

a misurare il decadimento della qualità all'aumento dell'età e la prevedibilità delle prestazioni. Le misure utilizzate dipendono però da ciò che osserviamo nel grafico di *AI Aging*:

1. Se è presente *degradazione temporale*, quindi un aumento dell'errore relativo nel tempo, l'inclinazione della mediana può essere una misura adeguata per quantificare l'invecchiamento del modello. Questa può essere misurata tramite un modello di regressione lineare.
2. Se le prestazioni rimangono stabili è più importante valutare la prevedibilità (variabilità) delle prestazioni, che può essere misurata tramite il livello medio del 75esimo percentile (più informativo dello spazio tra i quartili), che indica quanto le prestazioni del modello si riducono nel caso peggiore. Livelli più alti indicano prestazioni che possono ridursi maggiormente durante il periodo di utilizzo del modello, rendendo la sua qualità meno consistente (meno prevedibile, o più variabile). Il livello del terzo quartile è comunque fortemente legato a quello del primo quartile: quando il livello del caso peggiore aumenta, quello del caso migliore si abbassa, rendendo la misura analoga all'ampiezza della banda.

In aggiunta, il livello medio del caso mediano indica quanto bene la qualità iniziale si mantiene (per costruzione dovrebbe essere attorno ad 1).

Queste misure non sono le stesse che vengono utilizzate in Vela et al. (2022), ma la logica di fondo è la stessa e si adattano bene allo studio di simulazione; possono inoltre essere combinate con la qualità iniziale dei modelli, rendendo più semplice identificare il migliore da utilizzare. La procedura comporta infatti la stima del modello diverse volte, ad ogni iterazione, e in ciascun caso la sua qualità iniziale viene misurata sull'insieme di verifica tramite *errore quadratrico medio* ($MSE(t_0)$). La sua media può costituire quindi una misura della qualità dell'adattamento raggiunta sui dati, e può essere combinata con i risultati relativi alla stabilità in modo intuitivo per selezionare il modello migliore; ciò verrà mostrato nel capitolo 2.

Alcuni note finali relative alla procedura:

1. E' necessario fissare una distanza dT massima a cui valutare l'errore in modo che, a prescindere dall'insieme di stima utilizzato, possa essere calcolato l'errore relativo ($e_{rel}(dT)$) per ogni $dT = 1, \dots, \text{massimo}$. Questo si traduce nell'utilizzare sempre n osservazioni per il calcolo delle statistiche di sintesi (i percentili). Il dT massimo scelto limita però il numero di dataset di stima che possiamo utilizzare, visto che la finestra per l'addestramento si sposta verso destra ad ogni iterazione (escludiamo osservazioni vecchie e ne includiamo di più recenti). È necessario quindi individuare un equilibrio tra dT massimo e numero di insiemi di stima.
2. Ad ogni iterazione della procedura il modello studiato deve essere addestrato sul nuovo insieme di stima, e i suoi parametri devono essere ottimali, in modo da emulare un contesto applicativo reale. Regolare il modello ogni volta è tuttavia computazionalmente oneroso, soprattutto se si utilizzano procedure di convalida incrociata. La soluzione è evitare di regolare i parametri ad ogni iterazione, ma ad intervalli più ampi. La giustificazione sta nel fatto che i dataset di iterazioni vicine sono composti, per la maggior parte, dalle stesse osservazioni, e quindi la parametrizzazione ottimale sarà simile.
3. Ad ogni iterazione, le variabili esplicative nell'insieme di stima vengono standardizzate, e le stesse medie e deviazioni standard vengono utilizzate per la standardizzazione nel periodo successivo, in cui vengono misurate le prestazioni.
4. La procedura descritta non è identica a quella presentata nell'articolo. Nella versione originale, ad ogni iterazione viene scelto casualmente sia l'insieme di stima che la distanza dT a cui valutare l'errore (viene valutato una sola volta per insieme di stima). Non c'è però motivo di limitarsi alla misurazione dell'errore su una singola finestra ad una distanza casuale, e misurarlo al variare di questa non aumenta in modo significativo il tempo impiegato dalla procedura. Anche la scelta di estrarre

casualmente l’insieme di stima non è necessaria: nella procedura adottata dagli autori vengono estratti casualmente 20000 sottoinsiemi (uno per ogni iterazione) contenenti un anno di osservazioni, tramite la scelta casuale dell’ultimo giorno di dati da utilizzare. Il numero è talmente elevato che lo stesso insieme viene selezionato più volte, e quindi le prestazioni dei modelli vengono misurate molteplici volte, per ciascun insieme, a distanze casuali. Non ci sono quindi differenze nei risultati delle due procedure.

1.2 Approccio simulativo

In questa sezione viene descritto lo studio di simulazione condotto. In Vela et al. (2022) viene affermato che:

1. Sugli stessi dati, i modelli possono presentare distinti comportamenti a medio/lungo termine, presentando differenze nel modo in cui invecchiano.
2. Non esiste uno studio esaustivo su come i modelli rispondano alle diverse possibili insidie che si presentano durante il periodo di utilizzo, e che le proprietà di stabilità di questi strumenti dovrebbero quindi essere studiate a fondo.

Gli autori affermano inoltre che spesso non è chiaro, osservando i dati, perché i modelli possano presentare queste differenze, e ciò è stato confermato nel capitolo 2. Per questi motivi è stato condotto uno studio sistematico in cui la stabilità dei modelli è stata testata di fronte a diverse situazioni, con l'intenzione di ottenere delle informazioni che possano essere utilizzate in fase di selezione.

Il punto di partenza sono i risultati presentati nell'articolo, dove la *degradazione temporale* è stata studiata in contesti applicativi reali (figura 1.4). Un limite nell'utilizzo di dati reali, nell'affrontare il problema, è la mancata conoscenza del vero processo generatore dei dati. Non è possibile separare una causa di *degradazione* da un'altra, e determinare l'effetto dei singoli problemi sui comportamenti osservati; ciò rende l'approccio scelto in questo lavoro particolarmente adatto.

Sono stati perciò creati dei dataset che incorporano uno alla volta i problemi di interesse, su cui è stata applicata la metodologia descritta nella sezione 1.1, per ciascuno dei modelli considerati. Per limitare l'effetto della casualità è stato condotto un vasto numero di esperimenti simulativi: per ogni problema di interesse sono stati generati 100 dataset su cui applicare la procedura, e i risultati combinati; quest'ultimo passaggio avviene in modo intuitivo. Data una categoria di modello, da ognuna delle replicazioni è stato ricavato un

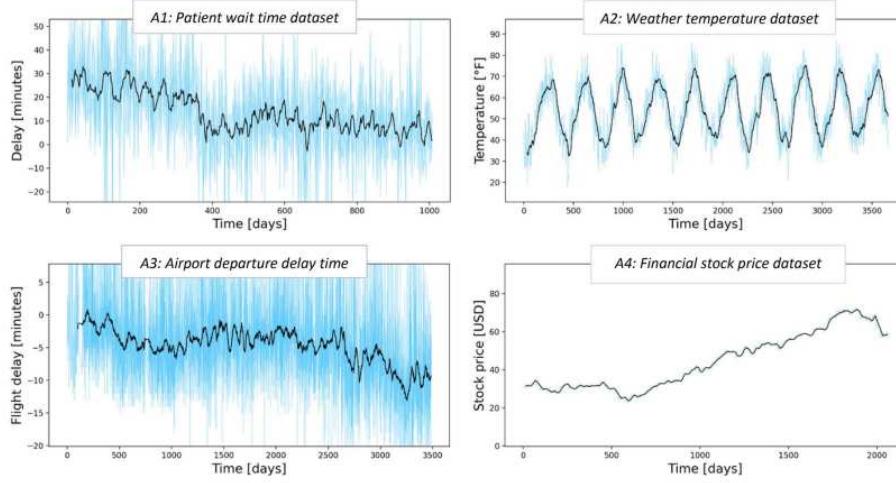


Figura 1.4: Quattro dataset utilizzati in Vela et al. (2022). Sull'asse verticale è rappresentata la media giornaliera della variabile risposta (in azzurro) e una media mobile (in nero), mentre su quello orizzontale il giorno. Nessuno dei dataset impiegati presenta evoluzioni improvvise nei valori della variabile risposta, in corrispondenza dei quali ci si aspetta una riduzione della qualità dei modelli. Fonte: Vela et al. (2022).

grafico di *AI Aging*. Di questi sono stati tenuti solo mediana, primo e terzo quartile, che sono stati combinati, separatamente, tramite mediana, condizionatamente alla distanza (dT). Questa procedura è rappresentata in figura 1.5, mentre la combinazione dei grafici di *AI Aging* in figura 1.6.

Mentre questo metodo permette di evidenziare il comportamento complessivo dei modelli, non permette di conservare tutta l'informazione disponibile nelle singole replicazioni o di fare considerazioni più precise. Per questo motivo vengono monitorate delle quantità aggiuntive:

1. Il livello medio del terzo quartile, che indica quanto aumenta l'errore nel caso peggiore e quindi la variabilità dell'errore relativo nei casi in cui non vi è *degradazione temporale*. La distribuzione di questa misura nelle replicazioni, per ciascun modello, viene rappresentata tramite istogramma (esempio in figura 1.7) e la media dei livelli utilizzata come sintesi.
2. L'inclinazione del tracciato della mediana, una misura del grado di *de-*

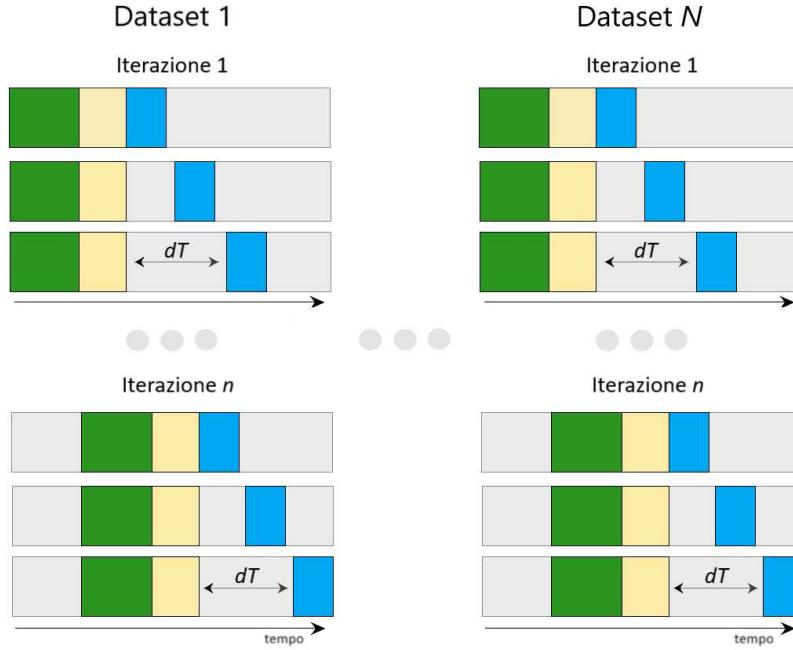


Figura 1.5: Procedura simulativa, estensione naturale del “test” di *degradazione temporale*, rappresentato in figura 1.2.

gradazione temporale, la cui distribuzione viene nuovamente rappresentata tramite istogramma (esempio in figura 1.8) e sintetizzata tramite media. L’inclinazione è riportata come differenza tra i valori iniziali (a $dT = 0$) e finali (a dT massimo) della retta interpolata (per stimare l’inclinazione), in quanto è un valore più indicativo della pendenza.

Dal capitolo 4 in avanti queste misure si sono rivelate insufficienti per comprendere il comportamento dei modelli nel tempo, a causa di alcune criticità del “test” di *degradazione temporale*. In quei casi sono stati studiati direttamente gli errori non-relativi dei modelli ($\text{l}'MSE$), all’aumento della distanza (dT), combinati in modo identico a quanto fatto per produrre il grafico di *AI Aging* (i percentili sono calcolati direttamente con gli $MSE(dT)$). I tracciati dei quartili possono dunque essere combinati con la procedura descritta in figura 1.9 e confrontati.

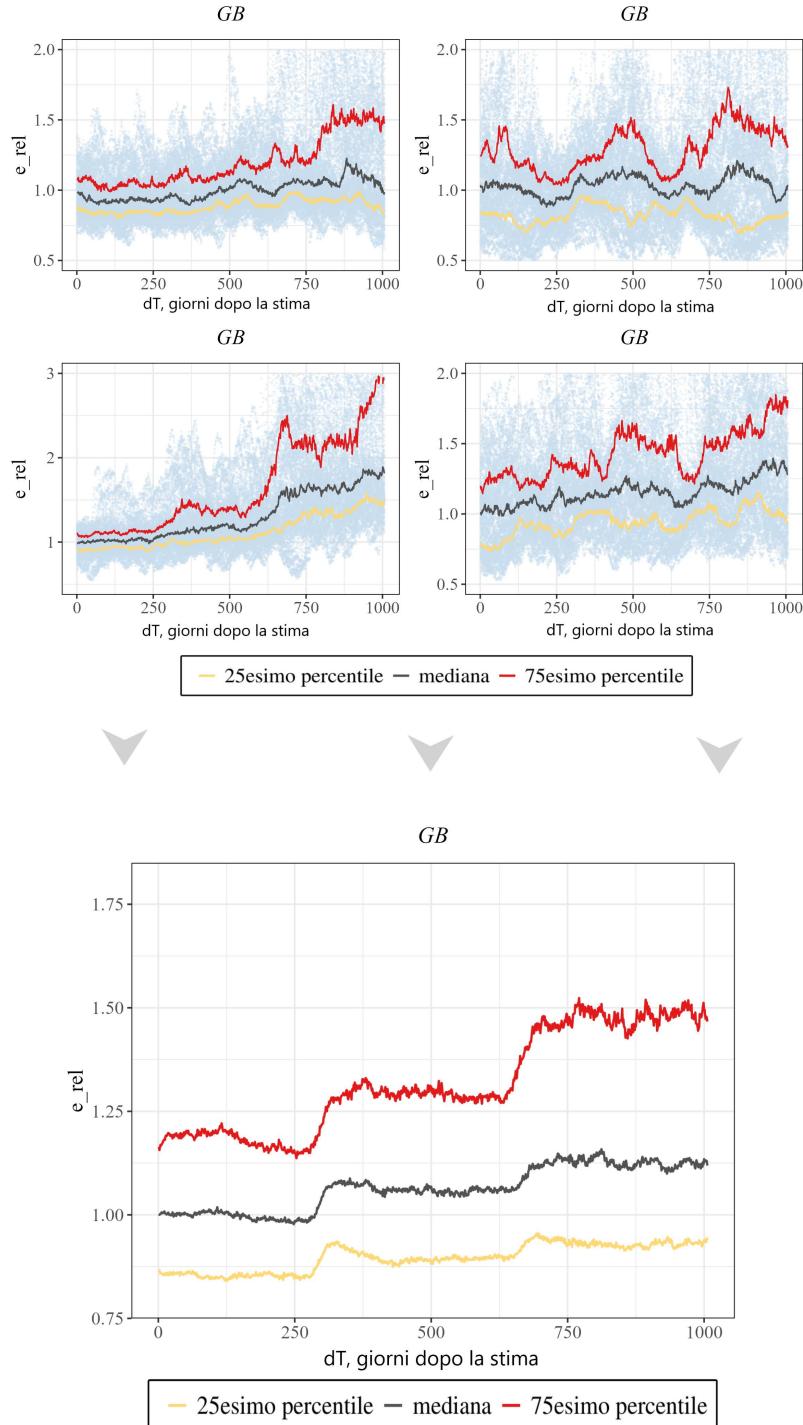


Figura 1.6: Data la categoria di modello, in questo caso il *gradient boosting*, e una specifica causa di *degradazione temporale*, l'output prodotto in ciascuna delle replicazioni viene combinato. Qui vengono rappresentati solo 4 dei 100 grafici che verrebbero prodotti.

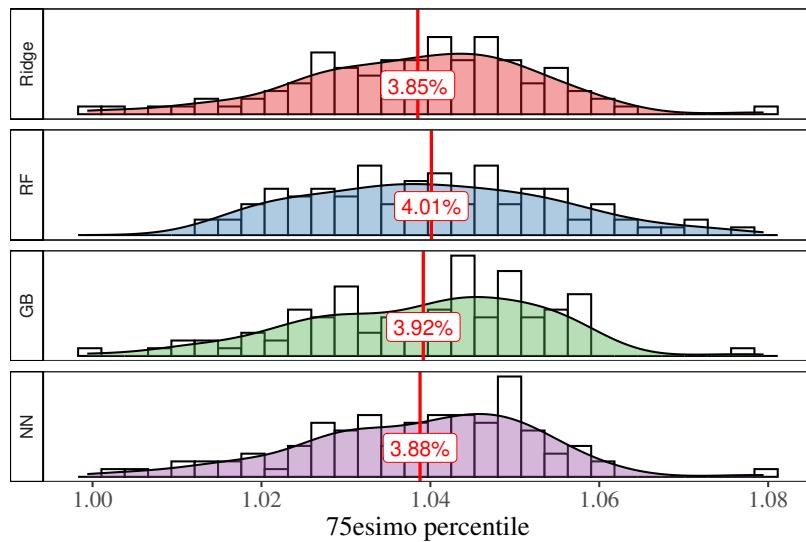


Figura 1.7: Distribuzione del livello medio del 75esimo percentile nelle replicazioni dell'esperimento simulativo, per ciascuno dei quattro modelli. Dalla distribuzione è possibile confrontare il campo di variazione dei valori, in modo da contestualizzare le differenze nei livelli medi. I livelli medi del terzo quartile, che nei valori originali superano 1, sono rappresentati nel grafico in termini di percentuale: un valore medio di 1.04 indica che, nel caso peggiore, in media, l'errore del modello aumenta del 4%, e ciò è rappresentato nelle etichette dei grafici.

I valori medi ottenuti possono essere confrontati per identificare il modello le cui prestazioni sono meno variabili (prevedibili), ma ha poco significato al variare della situazione simulata: ciascun caso ha dei livelli differenti. Non solo, i valori assoluti osservati cambiano in base alle scelte arbitrarie effettuate (capitolo 3), come il numero di osservazioni, quindi qualsiasi quantificazione delle differenze non ha significato. Il grafico è pensato quindi esclusivamente per osservare l'ordinamento dei modelli rispetto alla variabilità delle prestazioni.

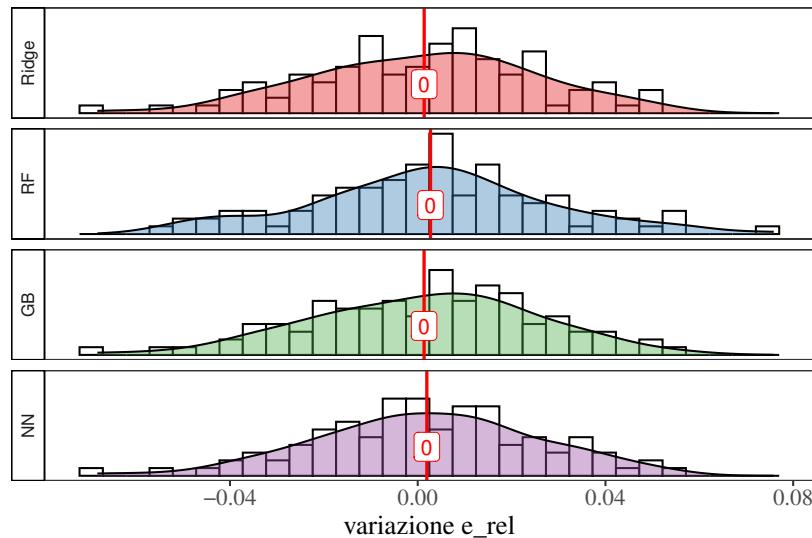


Figura 1.8: Distribuzione delle pendenze dell'errore relativo mediano nelle replicazioni della simulazione. Per ottenerle è stata stimata l'inclinazione della mediana (dal grafico di *AI Aging*) sulle singole replicazioni dell'esperimento di simulazione, tramite modello di regressione lineare. La differenza poi tra il valore iniziale (a $dT = 0$) e finale (a dT massimo) della retta, che indica la variazione dell'errore relativo nel periodo, è stata utilizzata per realizzare il grafico, in quanto più informativa rispetto ad una pendenza. Come nel caso precedente i valori medi possono essere utilizzati per confrontare i modelli, in quanto valori maggiori indicano maggiore *degradazione*.

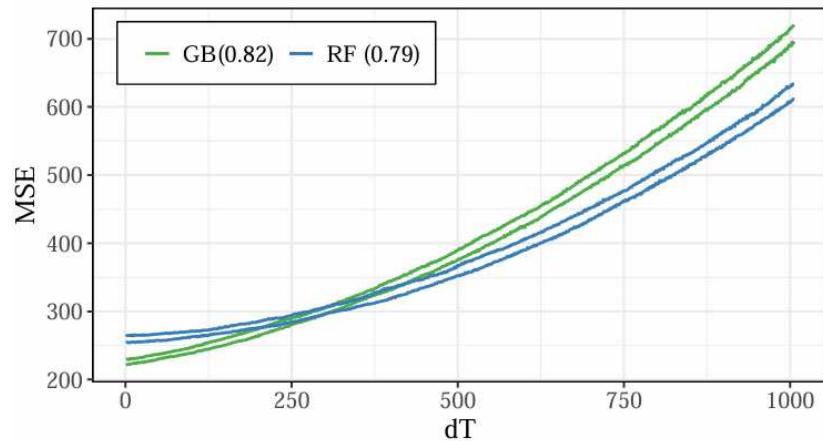


Figura 1.9: Dopo aver riprodotto i grafici di *AI Aging* utilizzando gli $MSE(dT)$ è possibile combinare i tracciati delle mediane (o di altri percentili), in modo da valutare come l'errore effettivo del modello evolve nel tempo. In questo caso sono confrontati *gradient boosting* e *foresta casuale*: dato un modello, i tracciati delle mediane sono combinati tramite il calcolo del 25esimo e del 75esimo percentile, punto per punto (dT), in modo da mostrare una banda di variabilità.

I valori riportati nella legenda indicano l' R^2 *predittivo* iniziale medio dei modelli nelle replicazioni della procedura, una quantità ulteriore che viene monitorata per valutare l'adeguatezza dei dataset simulati (introdotta nella sezione 1.2.3).

1.2.1 *Concept drift*

Prima di presentare i casi che vengono studiati tramite approccio simulativo è necessario introdurre il fenomeno del *concept drift*.

In un flusso di dati la generica osservazione nella sequenza può essere indicata con d_i e si compone di due parti, (X_i, Y_i) , dove X_i è l'insieme di variabili esplicative, mentre Y_i è la variabile di interesse, o risposta. L'obiettivo della modellazione è quello di prevedere Y_i in funzione di X_i .

Come indicato in Lu et al. (2018), il *concept drift* si presenta al tempo $t + 1$ se $P_t(X, Y) \neq P_{t+1}(X, Y)$, dove $P_t(X, Y)$ è la *distribuzione congiunta* di X e Y al tempo t . Si tratta quindi di un'evoluzione nella relazione dopo un certo istante temporale t .

La distribuzione congiunta può essere scomposta nel prodotto di due parti:

$$P(X, Y) = P(Y|X) \cdot P(X)$$

e ciò permette di evidenziare 2 possibili fonti di cambiamento:

1. Nel primo caso $P_t(X) \neq P_{t+1}(X)$ mentre $P_t(Y|X) = P_{t+1}(Y|X)$. Questa è la situazione in cui la sola distribuzione delle variabili esplicative evolve, ma la dipendenza di Y da queste rimane invariata. Questo tipo di *concept drift* prende il nome di *virtual shift* o *data drift*.
2. La seconda è dovuta all'evolversi di $P(Y|X)$, per cui $P_t(Y|X) \neq P_{t+1}(Y|X)$ ma $P_t(X) = P_{t+1}(X)$. La distribuzione delle variabili X rimane invariata, ma la dipendenza di Y da queste evolve nel tempo. Questa seconda tipologia è chiamata *real concept drift*.

Il *real concept drift* è sicuramente il più problematico dei due nell'ambito della previsione. Un modello può infatti essere stimato e perfezionato su un sottoinsieme finito dei dati, fino a riuscire ad approssimare bene la relazione tra le variabili X ed Y . All'evolversi della relazione, l'errore commesso dal

modello aumenterà, in quanto adattato ad un processo che non corrisponde più a quello vero.

Anche il *data drift* può causare problemi nei contesti pratici. L’evoluzione nella distribuzione delle X può portare alla presenza, nel flusso di dati, di osservazioni sempre più lontane da quelle che compaiono nell’insieme di stima, con la conseguente riduzione della prestazione dei modelli.

Comunemente i *concept drift* possono essere distinti in 4 tipi, rappresentati nella figura 1.10. Questi sono:

1. *Drift improvviso*, quando la distribuzione congiunta cambia in modo repentino, da un istante al successivo, passando da un concetto vecchio ad uno nuovo.
2. *Drift incrementale*, caratterizzato da un periodo di transizione graduale da un concetto precedente al successivo. Nel corso del periodo possono apparire concetti “intermedi”.
3. *Drift graduale*, caratterizzato da una progressiva sostituzione del concetto precedente con il successivo.
4. *Drift ricorrente*, dove 2 o più concetti si alternano.

La distinzione delle possibili manifestazioni di *concept drift* è rilevante ai fini delle simulazioni, in quanto le diverse tipologie possono portare a diverse manifestazioni della *degradazione temporale*.

In Vela et al. (2022) sono state spese alcune parole sull’effetto di questo fenomeno sui risultati presentati, in quanto la causa più nota per peggiorare le prestazioni dei modelli, e che motiva l’adozione di metodi per l’aggiornamento automatico. Nel loro lavoro hanno quindi cercato di utilizzare dati che non presentassero segni di *drift*: il comportamento della variabile risposta è infatti regolare nel tempo. Il comportamento di questa variabile è però poco indicativo per individuare *real concept drift*: nella sezione dedicata ai risultati delle simulazioni viene presentato come, anche in presenza di importanti

drift di questo tipo, il comportamento di Y rimanga regolare, rendendolo più insidioso da identificare.

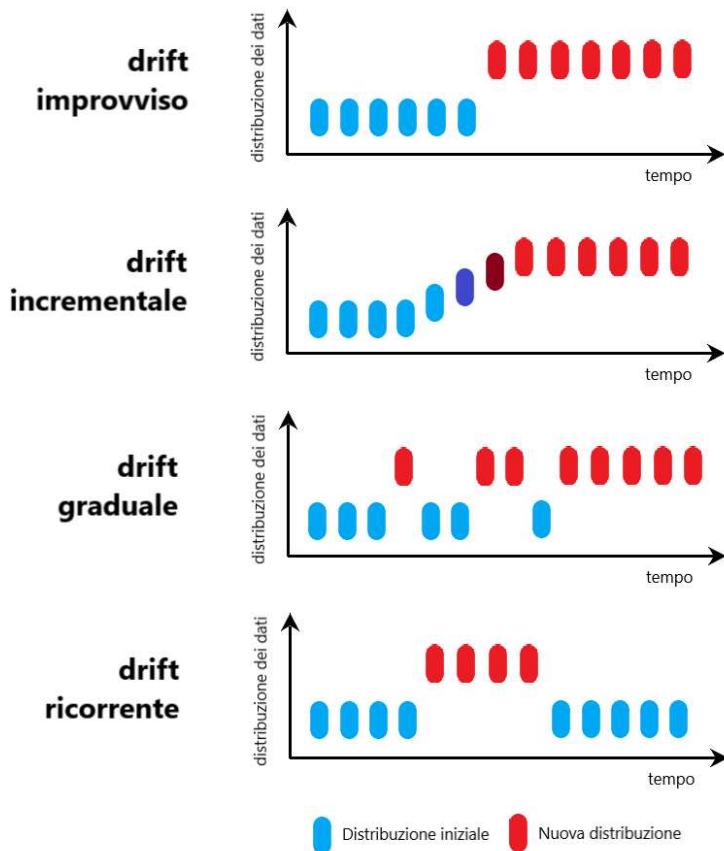


Figura 1.10: Tipologie di *concept drift*

L'impatto dei *drift* sulla qualità dei modelli è comunque appurato, però potrebbe essere fonte di differenze nel comportamento a medio/lungo termine. Gli autori affermano infatti che nonostante molta ricerca sia stata fatta su come individuare questi *drift* (Lu et al., 2018; contenente i principali metodi e test), non è noto come i singoli modelli vi rispondano.

Una parte delle simulazioni è quindi dedicata allo studio del problema, separando *data drift* da *real concept drift*.

1.2.2 Possibili cause di *degradazione temporale*

Un punto di partenza nella scelta delle situazioni da studiare è costituito dai suggerimenti degli autori dell'articolo che propongono, come cause dei comportamenti osservati, l'evoluzione del processo (*concept drift*) e la presenza di effetti latenti ciclici o stagionali. Non è però chiaro come ciascuno di questi problemi si rifletta sui singoli modelli, o perché questi presentino le differenze dichiarate.

Nell'articolo inoltre il “test” viene applicato su dataset che non sembrano presentare segni di *drift*, in quanto la variabile risposta presenta un comportamento regolare. Il comportamento dei modelli, in queste simulazioni, è quindi stato studiato anche in situazioni in cui il processo che genera i dati non evolve: non è infatti scontato che le diverse logiche matematiche presentino la stessa stabilità, anche in situazioni regolari, in cui la *degradazione temporale* non è attesa. I risultati del capitolo 2 inoltre sono chiari: anche quando la *degradazione* è assente la variabilità dell'errore relativo può differire, e ciò influisce sulla scelta del modello.

Le diverse situazioni studiate tramite approccio simulativo sono quindi state riassunte in alcune categorie di interesse:

1. Nella prima categoria viene studiata la stabilità dei modelli in assenza di evoluzioni del processo, al variare di alcune caratteristiche dei dati. Viene quindi inizialmente studiato il comportamento a medio/lungo termine in presenza di relazioni più o meno complesse tra le variabili esplicative e la variabile risposta, l'effetto della quantità di errore e delle differenze iniziali nella qualità dei modelli, che il “test” di *degradazione temporale* non prende in considerazione.

Avere un processo che non evolve nel tempo non significa però che le osservazioni siano necessariamente indipendenti: spesso, in contesti

reali, si osserva dipendenza tra i valori passati e futuri della variabile risposta, catturati tramite l'utilizzo dei primi ritardi. Questa dipendenza può essere causata da una serie di fattori, tra cui la presenza di stagionalità, la presenza di processi latenti e fattori esterni, oppure da movimenti della distribuzione delle variabili esplicative. Mentre alcune di queste situazioni sono incluse nelle successive categorie, in questo caso viene simulata la presenza di fattori esterni tramite l'inclusione, nella componente non osservata dei dati, di processi autocorrelati.

2. Nella seconda categoria vengono simulati dei flussi di dati caratterizzati da un'evoluzione del processo generatore, e tratta quindi i casi di *data drift* (in modo limitato, in quanto il contributo delle simulazioni è piuttosto ridotto) e di *real concept drift*.
3. Nella terza categoria vengono simulati dei flussi di dati caratterizzati da stagionalità (annuale), evidenziata nell'articolo come uno dei fattori che influenzano l'evoluzione temporale delle prestazioni dei modelli. Questa problematica viene affrontata per ultima perché include elementi di *real concept drift* e *data drift*, in quanto le variazioni stagionali non si ripetono allo stesso modo da un anno al successivo (*real concept drift*), generando picchi di altezza variabile (potenzialmente *data drift*).

I risultati relativi alla prima categoria sono riportati nel capitolo 3, quelli della seconda nel capitolo 4, mentre quelli della terza nel capitolo 5.

1.2.3 Dataset simulati

In questa sezione viene presentata la struttura dei dataset simulati, scelta sulla base di una serie di fattori:

1. L'obiettivo della modellazione è la previsione di una variabile quantitativa. La struttura fondamentale è quindi costituita da un insieme di variabili esplicative, X , e da una variabile risposta, Y , che ne è funzione (non deterministica).
2. L'aspetto temporale è fondamentale nello studio del problema. Ogni dataset contiene 5 anni di osservazioni, una lunghezza adeguata a stu-

diare il comportamento del modello nel futuro, e 30 osservazioni al giorno, sufficienti a permettere ai modelli delle buone prestazioni iniziali.

3. Il numero di variabili esplicative simulate è pari a 10, una quantità tale da permettere ai modelli buone prestazioni iniziali limitando il numero di osservazioni. Questa configurazione permette di effettuare le simulazioni in un tempo ragionevole.
4. La relazione tra X ed Y è lineare, ad eccezione dei casi in cui è necessaria una forma più complessa. Questa relazione permette a tutti i modelli, soprattutto considerando lo stimatore *ridge*, di raggiungere prestazioni iniziali simili.

In Vela et al. (2022) tutte le logiche considerate raggiungono infatti sempre prestazioni iniziali elevate, e tre modelli su quattro (tutti tranne lo stimatore *ridge*) possono adattarsi bene a diverse relazioni tra X e Y . Per fare in modo che ciò avvenga (allineamento iniziale) è perciò stato necessario utilizzare relazioni di questo tipo.

La struttura fondamentale dei dataset simulati è quindi la seguente:

$$Y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \cdots + \beta_{10} X_{10,ij} + \varepsilon_{ij}$$

i = giorno; j = identificativo dell'osservazione nel giorno i .

dove ε_{ij} costituisce un termine d'errore, indipendente dalle covariate. Questa struttura sarà comunque soggetta a cambiamenti a seconda della situazione che viene studiata, che saranno indicati in modo appropriato di volta in volta.

Data una possibile situazione oggetto di studio vengono generati 100 dataset, di cui alcune componenti rimangono fisse, mentre altre vengono generate casualmente in ciascuna replicazione. In particolare, fissato il problema oggetto della simulazione:

1. Vengono fissati i parametri β , che vengono generati casualmente, una sola volta, da una distribuzione uniforme $U(-5, 5)$.

2. Viene generata una matrice di correlazione Σ , definita positiva, per le variabili esplicative, che rimane fissa. La metodologia seguita permette di simulare una correlazione leggera tra le variabili esplicative, che non solo permette di collocarci più vicino ad un caso reale, ma permette anche allo stimatore *ridge* di fare della regolazione (e differenziarsi da un modello lineare). I valori della matrice sono generati attraverso una distribuzione esponenziale, con parametro di *rate* pari a 7. Il 30% di questi valori vengono resi 0, e la metà cambiati di segno. La matrice così ottenuta è definita positiva e la correlazione tra le variabili esplicative è ragionevole (esempio in figura 1.11). Nell'evento che non sia definita positiva, questa viene corretta tramite un algoritmo iterativo (N. J. Higham, 2002), che la rende la matrice definita positiva più vicina. Negli esperimenti di simulazione è stato comunque osservato come la scelta della matrice non influenzi la stabilità dei modelli.
3. Per le osservazioni, i valori di X e il termine d'errore sono generati casualmente ad ogni replicazione, rispettivamente da una normale multivariata $N_{10}(0, \Sigma)$ e una normale univariata $N(0, \sigma_\varepsilon^2)$, indipendenti. Il valore di σ_ε , fisso, è determinato in funzione del resto del modello, in modo da ottenere delle prestazioni iniziali adeguate.

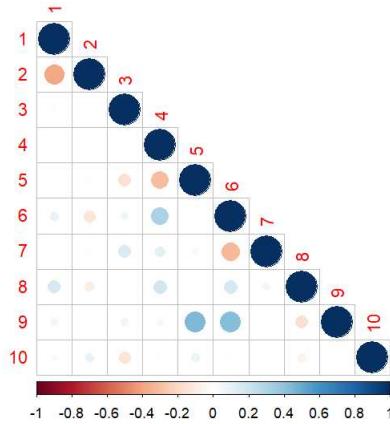


Figura 1.11: Esempio di matrice di correlazione generata tramite la metodologia descritta sopra, che mostra una correlazione ragionevole tra le variabili esplicative.

È poi necessario distinguere tra la struttura dei dataset generati e i dati utilizzati dai modelli. In genere i dataset su cui effettuare la stima sono costituiti dalle variabili X , quantitative, e dalla variabile Y . In aggiunta, a seconda del dataset simulato, l'insieme di variabili X può essere esteso per catturare strutture più complesse. Un esempio è l'inclusione dei primi ritardi di Y quando vi è una struttura autoregressiva nel processo generatore dei dati; questi casi verranno comunque indicati.

La valutazione dell'adeguatezza dei dataset generati si basa su una misura della qualità iniziale dei modelli. Nell'articolo originale (Vela et al., 2022) tutti i modelli presentavano qualità elevata, in quanto hanno raggiunto valori di R^2 *predittivo* (ovvero l'indice R^2 calcolato fuori dall'insieme di stima) di convalida incrociata compresi tra 0.7 e 0.9, stimati su sottoinsiemi dei dati di un anno. In modo analogo, in queste simulazioni sono stati utilizzati buoni modelli, ed è quindi necessario un indice che trascenda il singolo dataset e consideri complessivamente le 100 replicazioni; e che sia più semplice rispetto a ricorrere a convalide incrociate, che richiedono troppo tempo.

La soluzione è la seguente: concentrandoci su una sola replicazione della procedura di simulazione (un singolo dataset, figura 1.5) e un singolo modello, possiamo ricavare un numero di misure di errore iniziale ($MSE(t_0)$) pari al numero di dataset di stima utilizzati. Queste sono misure dell'errore nell'insieme di verifica, che possono essere usate come base per costruire l' R^2 *predittivo* iniziale. Questi valori possono quindi essere combinati tramite media, prima relativamente alla singola replicazione e poi tra replicazioni, fornendo un'indicazione del livello di qualità raggiunto dal modello sulla struttura utilizzata per generare i dataset. La qualità iniziale è quindi riportata tramite grafico (un esempio in figura 1.12) o in alcuni casi nei commenti di altri grafici.

Nelle simulazioni spesso lo studio di una problematica comporta la riduzione della qualità dei modelli, e in alcuni casi questa la impatta in modo differente, da un modello ad un altro, accentuando il divario di partenza. Nei casi in cui ciò è rilevante è comunque considerato nella lettura dei risultati.

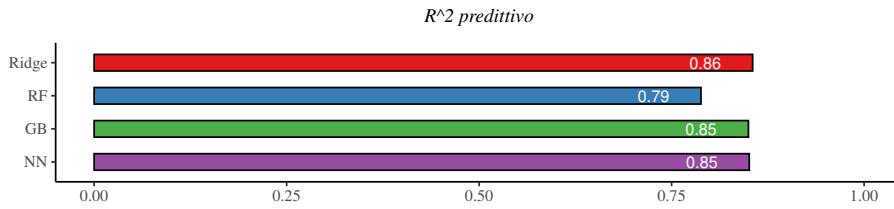


Figura 1.12: Media degli R^2 *predittivi* iniziali. Il grafico è tratto dal capitolo 3. È possibile notare che la *foresta casuale* tende ad avere una qualità iniziale più bassa degli altri modelli, ed è un divario che non è stato possibile chiudere nelle simulazioni condotte.

In alcuni, limitati, casi è stato di interesse approfondire alcune ramificazioni legate ai risultati di una simulazione. Visti i tempi richiesti non sempre è stato possibile approfondire con adeguato rigore: in questi casi, indicati, in cui le differenze nel comportamento dei modelli sono sufficientemente chiare, il numero di replicazioni effettuate è stato dimezzato (50 replicazioni invece di 100). Questo fornisce un’idea del comportamento dei modelli in queste specifiche situazioni, che altrimenti non sarebbero potute essere sviluppate. Se il numero di replicazioni è differente da 100 ciò viene indicato nei commenti dei grafici che riportano i risultati, oppure negli elenchi delle figure (all’inizio di ogni sezione dell’appendice, per le figure riportate nella sezione).

1.2.4 Applicazione del “test” di *degradazione temporale*

La procedura per lo studio della *degradazione temporale*, descritta nella sezione 1.1, è computazionalmente onerosa. Dato un singolo dataset vengono selezionati un elevato numero di sottoinsiemi di stima, su cui ciascuno dei 4 modelli deve essere stimato. Ogni modello deve inoltre utilizzare una parametrizzazione ottimale, e determinarla richiede tempi lunghi.

In seguito viene quindi descritta l’applicazione del “test” nelle singole re-

plicazioni della procedura di simulazione, con un'attenzione particolare alle soluzioni per ridurre il tempo richiesto. Questi espedienti sono stati impiegati anche nelle analisi condotte sui *dataset reali*, presentate nel capitolo 2.

La lunghezza temporale dei dataset simulati è di 5 anni. L'ampiezza dell'insieme di stima, il sottoinsieme di dati su cui stimare i modelli, è stata fissata ad un anno (la stessa utilizzata nell'articolo), e vengono costruiti utilizzando dati provenienti dai primi 2 anni.

Ad ogni iterazione della procedura un nuovo dataset viene selezionato partendo dal precedente: vengono escluse osservazioni meno recenti e incluse quelle più recenti, come una finestra mobile. Un nuovo sottoinsieme di stima viene costruito escludendo i 5 giorni di osservazioni più vecchie dal precedente e includendone 5 più recenti. Questo porta alla selezione, un'iterazione alla volta, di 74 sottoinsiemi di stima, su cui applicare i 4 modelli.

L'ampiezza dell'insieme di verifica è stata fissata a 60 giorni, mentre per la finestra mobile sono stati utilizzati 30 giorni. Queste scelte hanno definito in modo esatto la distanza massima, dal momento della stima, a cui calcolare l'errore di previsione (dT massimo), 1006 giorni.

L'individuazione della parametrizzazione ottimale per i modelli, ogni volta che questi vengono stimati, è l'elemento più dispendioso della procedura. Dovendo considerare i limiti di tempo e di risorse per il completamento del lavoro sono state limitate le possibili parametrizzazioni prese in considerazione.

Per lo stimatore *ridge* è stato regolato il valore del parametro di penalizzazione dei coefficienti, λ . Per la *foresta casuale* è stato regolato il numero di variabili considerate dall'algoritmo per determinare la suddivisione ottimale, mantenendo fisso il numero di alberi e lasciandoli crescere alla profondità massima (procedura descritta in Azzalini et al., 2012). Per il *gradient boosting* è stato regolato il numero di alberi utilizzati e la loro profondità, mantenendo costante il tasso di apprendimento. Nel caso della *rete neurale* è stata fissata la struttura, quindi il numero di strati latenti (3) e i nodi latenti in

ciascuno strato (due possibili configurazioni, che cambiano a seconda della complessità del problema: 50 o 200 nodi per strato), utilizzando come funzioni di attivazione la funzione *reLu*; mentre per la regolazione è stato usato il metodo *early stopping*.

Nonostante questo approccio alla regolazione possa sembrare piuttosto semplice, considerando le alternative (penalizzazioni *L1*, *L2* o *dropout*, tra le possibilità considerate), ha portato a buone prestazioni nelle simulazioni condotte e ad una grande riduzione dei tempi necessari per la regolazione, il collo di bottiglia principale nell'applicazione di questa metodologia.

Le funzioni obiettivo utilizzate nella stima di ciascuno di questi modelli sono di tipo quadratico.

La regolazione avviene in modo diverso da una categoria di modello all'altro. Nel caso dello stimatore *ridge* e della *foresta casuale* individuare la parametrizzazione ottimale su ogni sottoinsieme di stima è troppo oneroso. È possibile però sfruttare la similitudine tra insiemi di stima ad iterazioni vicine, che condividendo la quasi totalità delle osservazioni avranno una parametrizzazione ottimale pressoché identica. È quindi sufficiente effettuare la regolazione ogni volta che il nuovo insieme di stima si sposta troppo da quello su cui è avvenuta l'ultima regolazione. Questa avviene quindi ogni 60 giorni, tramite *convalida incrociata timewise* su una griglia di possibili configurazioni.

La procedura di *convalida incrociata timewise* è impiegata nell'ambito delle serie storiche e dei flussi di dati, ed è una modifica della *convalida incrociata* classica. Dato un insieme di stima, invece che fare una suddivisione casuale in k parti uguali più piccole, come nella procedura classica, l'insieme viene suddiviso in parti uguali seguendo l'ordinamento temporale delle osservazioni. Il primo sottoinsieme sarà costituito dalle prime osservazioni, l'ultimo dalle ultime. La procedura rimane iterativa. Nel caso classico, ad ogni iterazione, tutti i gruppi meno uno vengono utilizzati per la costruzione dell'insieme su cui stimare il modello, mentre il rimanente per misurare l'errore; nelle iterazioni successivi le singole parti vengono ruotate, finché tutte non sono state utilizzate una volta per la misurazione.

Nella versione *timewise* il criterio di assegnazione delle parti a insieme di stima o di misurazione cambia. Nella prima iterazione viene utilizzato il sottoinsieme meno recente per la stima, mentre quello subito successivo, in ordine di tempo, per la misurazione dell'errore. All'iterazione successiva, i primi due sottoinsiemi vengono utilizzati per la stima, e il successivo per la misurazione. Il procedimento è analogo per le altre iterazioni: all'ultima, sarà il sottoinsieme più recente ad essere utilizzato per la misurazione dell'errore, e tutti gli altri per la stima. Questo metodo permette di selezionare la parametrizzazione ottimale emulando l'impiego reale del modello nei contesti in cui questa metodologia viene applicata: la previsione nel futuro.

Nel caso della *rete neurale* il metodo di regolazione impiegato (*early stopping*) non necessita la determinazione di parametrizzazione ottimale, e quindi non viene seguita la procedura precedente. Invece, al momento dell'adattamento, l'insieme di stima viene separato casualmente in due parti: una prima parte (80%) da dedicare alla stima del modello, e la restante (20%) alla regolazione. La regolazione consiste nel monitorare le prestazioni del modello sulla parte di dati più piccola, durante la procedura iterativa di stima (*gradient descent*), interrompendola arrivati al picco, in modo da evitare sovradattamento.

Nel caso del *gradient boosting* viene regolata la profondità degli alberi e il loro numero, fissando il tasso di apprendimento. La profondità ottimale viene determinata tramite *convalida incrociata timewise* solamente sul primo sottoinsieme di stima (il primo anno), e mantenuta per l'intero dataset. Utilizzando lo stesso approccio di *ridge* e *foresta casuale* emergeva infatti la stabilità del parametro al variare dell'insieme di stima, e da qui la scelta per ridurre il tempo di esecuzione. Il numero di alberi ottimale viene invece individuato su ciascun sottoinsieme di stima, con un approccio analogo all'*early stopping* per la *rete neurale*.

Il tempo di esecuzione rimane comunque molto elevato. Scelta una situazione su cui valutare la stabilità dei modelli, lo studio di simulazione non può essere eseguito su una singola macchina. Nel lavoro è stato quindi impiegato il cluster del Dipartimento di Scienze Statistiche, per le sole simulazioni.

Queste sono state condotte utilizzando R per *ridge*, *foreste casuali* e *gradient boosting*, mentre Python per le *reti neurali*. Per le analisi del capitolo 2 è stato invece utilizzato esclusivamente R.

Capitolo 2

Analisi su dati reali

In questo capitolo sono descritte le analisi preliminari condotte per cercare di riprodurre i risultati e le conclusioni a cui sono giunti Vela et al. (2022); tramite l'applicazione del “test” per l'analisi della *degradazione temporale* ad alcuni dataset reali impiegati dagli autori nello studio.

I quattro dataset studiati provengono da altrettante strutture sanitarie e contengono informazioni riguardanti gli accessi dei pazienti alle strutture, e permettono di sviluppare modelli di previsione del tempo d'attesa per le visite mediche.

Lo studio degli altri dataset impiegati nell'articolo non è stato possibile. Gli autori hanno infatti indicato le fonti utilizzate, ma non hanno né messo a disposizione i dati (la maggior parte non era perciò reperibile), né indicato le variabili costruite ed utilizzate, rendendo difficile riprodurre i risultati.

Lo scopo di questa sezione è quello di capire come le prestazioni dei modelli possono evolvere in contesti reali e le differenze da attendersi, e come usare questa informazione per selezionare il modello, un aspetto non considerato in Vela et al. (2022).

2.1 Dati sugli accessi alle cliniche

I dati analizzati provengono dallo studio di Pianykh et al. (2020), il cui obiettivo era quello di migliorare la gestione e l'organizzazione degli appuntamenti medici in 4 strutture ambulatoriali. Per farlo sono stati raccolti un gran volume di dati provenienti dai 4 siti (indicati con le etichette F1 - F2 - F3 - F4), distribuiti su diversi anni e separati per struttura.

La struttura dei dataset è sostanzialmente identica. Ogni osservazione fa riferimento all'accesso di un paziente all'ambulatorio specifico, a cui il dataset fa riferimento, per una visita medica. Per ciascun ingresso alla struttura è presente un elevato numero di variabili, riportate nelle due tabelle 2.1 e 2.2 (la seconda contiene le esplicative).

Variabile	Descrizione
x_ArrivalDTTM	Data e ora di arrivo del paziente alla clinica.
x_ScheduledDTTM	Data e ora dell'appuntamento.
x_BeginDTTM	Data e ora dell'inizio della visita.
Wait	Tempo di attesa effettivo, dato dalla differenza tra x_BeginDTTM e x_ScheduledDTTM.

Tabella 2.1: Variabili che non sono incluse tra le esplicative.

L'obiettivo dello studio condotto in Pianykh et al. (2020) era quello, come già menzionato, di migliorare l'organizzazione e la gestione delle visite. I dati sono perciò stati impiegati per costruire dei modelli di *machine learning* che permetessero di prevedere l'attesa effettiva dei pazienti per la visita (variabile Wait, tabella 2.1), sulla base delle altre variabili raccolte (riportate nella tabella 2.2), e individuarne le determinanti. Il modello migliore sarebbe poi stato impiegato per fornire ai pazienti, nelle sale d'attesa di queste cliniche, una previsione in tempo reale del tempo di attesa per l'inizio della visita.

Variabile	Descrizione
SumHowEarlyWaiting	Somma dell'anticipo di arrivo dei pazienti in fila.
AvgHowEarlyWaiting	Media dell'anticipo dei pazienti in fila.
LineCount0Strict	Numero di pazienti in fila con appuntamento successivo all'orario corrente.
SumWaits	Tempo atteso totale, dal momento di arrivo, dei pazienti in fila.
LineCount0/1/2/3/4	Numero di pazienti in fila all'arrivo e 15, 30, 45 e 60 minuti prima.
FlowCount2/4	Numero di pazienti che hanno iniziato la visita nei 30 e 60 minuti precedenti all'arrivo.
SchFlowCount2/4	Numero di pazienti previsti nei 30 e 60 minuti precedenti all'arrivo.
FutFlowCount2/4	Numero di pazienti previsti nei 30 e 60 minuti successivi all'arrivo.
DelayCount	Numero di visite iniziate in ritardo fino al momento di arrivo.
DelayCountLastHour	Numero di visite iniziate in ritardo nell'ultima ora precedente all'arrivo.
mintime	Tempo di attesa minimo della giornata.
maxtime	Tempo di attesa massimo della giornata.
AheadCount	Numero di altri pazienti previsti, nella giornata, prima del paziente.
ThoracicCount	Numero di pazienti in attesa per visita toracica.
PediatricCount	Numero di pazienti in attesa per visita pediatrica.
NeuroCount	Numero di pazienti in attesa per visita neurologica.
AbdominalCount	Numero di pazienti in attesa per esame all'addome.
VascularCount	Numero di pazienti in attesa per esame vascolare.
CardiacCount	Numero di pazienti in attesa per visita cardiaca.

MSKCount	Numero di pazienti in attesa per esame muscolo-scheletrico.
NumScannersUsedToday	Numero di scanner utilizzati nella giornata nella clinica.
SumInProgress	Tempo totale impiegato per gli esami in corso, fino a questo momento.
BeforeSlot	Distanza dal precedente slot di appuntamento.
AfterSlot	Distanza dal successivo slot di appuntamento.
Median5	Tempo di attesa mediano per gli ultimi 5 pazienti.
MostRecent1/2/3/4/5	Tempo di attesa del paziente più recente, del secondo, fino al quinto più recente.
StartTime	Ora di arrivo e le prime potenze, per catturare relazioni non lineari.
IsLast	Indicatrice, primo paziente della giornata.
IsFirst	Indicatrice, ultimo paziente della giornata.
NoneInProgress	Indicatrice, nessun esame in corso.
NoneCompleted	Indicatrice, nessun esame condotto nella giornata.
NoneInLine	Indicatrice, nessun paziente in fila.
SumWaitByTaskTypeLine	Somma del tempo di attesa dei pazienti in fila, per tipologia di esame.
AvgWaitByTaskTypeLine	Media del tempo di attesa dei pazienti in fila, per tipologia di esame.
SumTimeToComplete-InProgress	Tempo atteso per il completamento degli esami in corso.
DelayedInLine	Numero di pazienti in fila la cui visita dovrebbe essere già iniziata.
SumDelayWaitingByExamCode	Ritardo totale per i pazienti in fila, per tipologia di esame.
SumDelayWaitingInLine	Ritardo totale dei pazienti in fila.
SumDelayInProgress	Ritardo totale fatto per gli esami in corso.
ExpectedDelayNextExam	Ritardo previsto per il prossimo esame in programma.

AvgAgePeopleWaiting	Età media dei pazienti in fila.
DayOfWeek	Giorno della settimana.
Month	Mese.
DayOfYear	Giorno dell'anno.
InProgressSize	Numero di esami in corso.
AvgWaitLastK1Customers	Tempo di attesa medio per gli ultimi 2, 4 e 8 pazienti.
NumCompletedToday	Esami completati nella giornata.
NumCompletedIn- LastW1/W2/W3	Numero di esami completati negli ultimi 30, 60 e 120 minuti.
NumCustomersIn- LastW1/W2/W3	Numero di pazienti arrivati negli ultimi 30, 60 e 120 minuti.
AvgWaitLastW1/W2/W3	Tempo di attesa medio negli ultimi 30, 60 e 120 minuti.
AvgDelayForDay	Ritardo medio della giornata.
OutpatientWaitingCount	Numero di pazienti ambulatoriali in fila.
MalesWaitingCount	Numero di uomini in fila.
NumAddOnsToday	Numero di pazienti che si sono aggiunti il giorno stesso.
NumAddOnsLastW2	Numero di pazienti che si sono aggiunti negli ultimi 60 minuti.
NumScheduledNextSlot	Numero di pazienti nel prossimo slot, relativamente al momento di arrivo.
NumScheduledNextW2	Numero di pazienti prenotati nei prossimi 60 minuti.
SumTimeToComplete- NextSlot	Tempo atteso per il completamento degli esami del prossimo slot.
SumTimeToComplete- NextW2	Tempo atteso per il completamento degli esami dei prossimi 60 minuti.
WithContrastCountWaiting	Numero di pazienti in attesa per visite con contrasto.
WithAndWithoutContrast- CountWaiting	Numero di pazienti in attesa per visite con e senza contrasto.

WithContrastCountIn- Progress	Numero di visite in corso con contrasto.
WithAndWithoutContrast- CountInProgress	Numero di visite in corso con e senza contrasto.

Tabella 2.2: Variabili raccolte e costruite dai ricercatori per spiegare il tempo di attesa dei pazienti.

Un concetto importante su cui si basa la modellazione, e che traspare dalle variabili presentate nella tabella 2.2, è la fila: quando un cliente arriva alla clinica si mette in fila, e sarà perciò preceduto da individui in attesa dell'inizio della loro visita; molte delle variabili disponibili si basano sui pazienti che vi appartengono. Le altre fanno invece riferimento agli esami in corso e a quelli già conclusi.

Tutti i dataset condividono la stessa struttura, presentata nelle tabelle, con alcune differenze:

1. Le cliniche F1 - F2 - F3 differiscono da F4, in quanto si basano su un sistema di prenotazione della visita (il paziente si presenta nella struttura dopo aver prenotato, in anticipo, una visita), mentre F4 è di tipo "walk-in". I pazienti si presentano quindi senza appuntamento, che viene fissato a pochi minuti dall'arrivo. In quest'ultimo caso, il tempo di attesa effettivo da prevedere non è la differenza tra l'appuntamento e il momento di inizio dalla visita, ma tra quest'ultimo e il momento di arrivo dal paziente in ambulatorio.
2. Gli appuntamenti vengono concessi ai pazienti a specifici slot temporali: in F1 l'appuntamento può essere dato in corrispondenza dell'inizio dell'ora o a metà, per F2 ed F3 ogni quarto di ora, mentre in F4 ogni 5 minuti. Oltre a questi slot, ci sono alcune eccezioni ad orari differenti. Le variabili che sono state costruite, e che andranno ad aggiungersi a quelle già presenti, ne hanno tenuto conto.

2.2 Preparazione dei dataset

I dataset necessitano di minime operazioni di pulizia: i pochi dati mancanti presenti possono essere imputati in modo esatto, e le variabili sono già pronte all'uso. Nella tabella 2.3 sono riportate alcune informazioni relative alla dimensione di questi.

Dataset	Numero di osservazioni	Giorni
F1	42765	1007
F2	15652	1007
F3	23583	666
F4	48430	598

Tabella 2.3: Dimensione dei dataset: numero di osservazioni e lunghezza in giorni del dataset. F1 - 1007 giorni indica che la distanza tra l'osservazione meno recente e quella più recente in F1 è di 1007 giorni, e in modo analogo gli altri dataset.

Gli interventi più importanti riguardano la creazione di nuove variabili, costruite da zero per spiegare il tempo di attesa. Queste sono presentate nella tabella 2.4.

Variabile	Descrizione
x_ArrivalDTTM_Quant	Arrivo alla clinica come minuti dall'orario di apertura.
x_ScheduledDTTM_Quant	Orario dell'appuntamento come minuti dall'orario di apertura.
how_Early	Quanto si è presentato in anticipo il paziente rispetto all'orario dell'appuntamento.
how_Late	Quanto si è presentato in ritardo il paziente rispetto all'orario dell'appuntamento.
NumScheduledSameSlot	Numero di pazienti con appuntamento nello stesso slot.

InLineNumberForScheduled	Numero di pazienti in fila con appuntamento nello stesso slot (+1); quindi la posizione nella fila, relativamente allo slot di appuntamento.
NumScheduledBefore-Slot5/10/15/30	Numero di visite in programma nei 5, 10, 15 e 30 minuti prima dello slot del paziente.
Schedule_Minute	Frazione di ora a cui è stato fissato l'appuntamento, categoriale. Gli orari eccezionali costituiscono categoria a sé.
perc_delay	Frazione di visite iniziata in ritardo nella giornata, fino al momento di arrivo.
perc_delay_high	Frazione di visite iniziata con elevato ritardo (>15 minuti) nella giornata, fino al momento di arrivo.
SumDelayInLine	Ritardo totale nell'inizio della visita dei pazienti in fila, calcolato al momento dell'arrivo.
LastPatient1/2	Tempo di attesa degli ultimi due pazienti.
MedianWaitTime	Tempo di attesa mediano della giornata.
WeekDelay	Tempo di attesa medio nell'ultima settimana.
previousDayDelay	Tempo di attesa medio dell'ultimo giorno di apertura.

Tabella 2.4: Nuove variabili costruite per spiegare il tempo atteso per l'inizio della visita.

L'inizio della visita può avvenire dal momento di arrivo del paziente, a prescindere dall'orario effettivo dell'appuntamento, e ciò limita l'informazione utilizzabile per la previsione a quel momento. Nel dataset originale è già disponibile un vasto insieme di variabili che permettono di valutare la situazione all'arrivo, appena precedente e nell'immediato futuro, che permettono di capire se un cliente può iniziare nell'immediato. In effetti, ci sono diversi casi di pazienti che si presentano con largo anticipo, e iniziano altrettanto presto; allo stesso modo alcuni si presentano in ritardo, iniziando, per costruzione, almeno altrettanto tardi. Per queste situazioni sono state costruite le variabili “how_Early” e “how_Late”.

Spesso però l'inizio della visita avviene attorno all'orario di appuntamen-

to, e l'informazione relativa a tale fascia oraria manca nel dataset iniziale. Essendo gli appuntamenti prenotati in anticipo, non vi è motivo di non includere l'informazione: è stato quindi aggiunto il numero di appuntamenti fissati nella stessa fascia oraria di quella concordata dal cliente (“NumScheduledSameSlot”), l'ordine di arrivo del cliente rispetto agli altri allo stesso orario (“InLineNumberForScheduled”), in quanto il paziente può essere superato da qualcuno arrivato dopo, ma generalmente avviene se l'appuntamento di quest'ultimo è prima (l'ordine in fila, relativamente ad uno slot orario, è più indicativo), e il numero di appuntamenti negli slot orari precedenti, per considerare l'intensità della fascia oraria (“NumScheduledBeforeSlot”).

Sono state inoltre aggiunte delle variabili relative ai giorni precedenti, in modo da catturare trend o situazioni di breve periodo (“previousDayDelay” e “WeekDelay”), possibilmente dovuti a cambiamenti organizzativi. Sono state inoltre incluse variabili per comprendere meglio la situazione del giorno stesso (“perc_delay” e “perc_delay_high”), tra cui “SumDelayInLine”, “LastPatient1/2” e “MedianWaitTime”, molto simili ad alcune già presenti, ma ricalcolate al momento di arrivo del paziente (non era chiaro il calcolo delle variabili “MostRecent”).

Queste variabili sono state utilizzate nei dataset F1, F2 ed F3, mentre nell'ultimo sono state apportate delle modifiche, vista la diversa definizione di tempo di attesa, calcolato come differenza tra l'arrivo e l'inizio. Le differenze principali sono le seguenti:

1. È stato distinto il tempo di attesa dal ritardo, ovvero la differenza dal momento di inizio e l'orario concordato (la variabile “Wait” negli altri dataset). Le variabili “LastPatient1/2”, “MedianWaitTime”, ”WeekDelay” e “previousDayDelay” sono presenti due volte: la prima fa riferimento ai tempi di attesa, la seconda all'effettivo ritardo rispetto a quanto concordato, due informazioni distinte.
2. La variabile “perc_delay_high” considera come ritardo elevato 5 minuti, vista la diversa scala della variabile risposta. È calcolata direttamente sul ritardo, e non sul tempo di attesa.

I quattro dataset finali sono quindi stati valutati per capire la qualità di previsione raggiungibile dai modelli: sono stati quindi presi singolarmente, separati in sottoinsiemi di stima e di verifica (85% - 15% rispettivamente), e sugli insiemi di stima sono stati regolati e stimati i 4 modelli oggetto di questo lavoro (la regolazione avviene come indicato nella sezione 1.2.4). La qualità di questi modelli è stata poi valutata sull'insieme di verifica, tramite il calcolo dell'indice R^2 predittivo (l'indice R^2 , il coefficiente di determinazione, calcolato fuori dall'insieme di stima). I risultati sono riportati in figura 2.1.

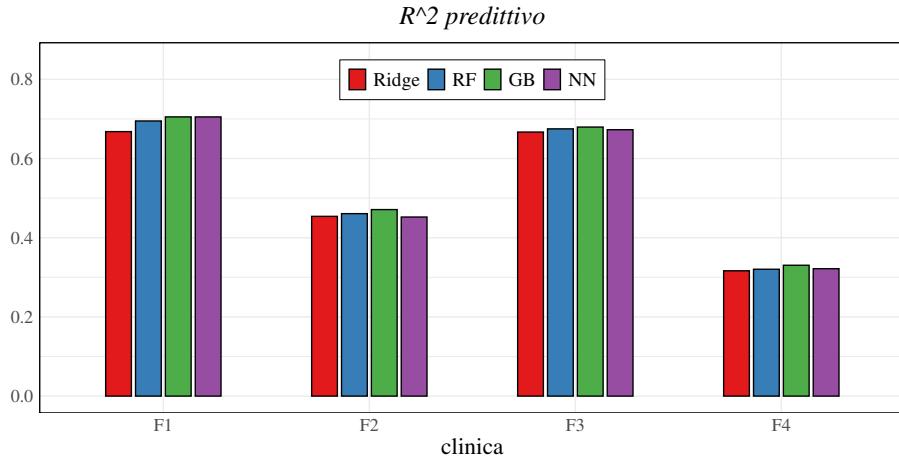


Figura 2.1: R^2 predittivo dei quattro modelli considerati, calcolati sugli insiemi di verifica, per ciascun dataset; i modelli presentano delle prestazioni simili su ciascuno.

I risultati migliori sono stati ottenuti nei dataset F1 ed F3, in cui l'indice di R^2 predittivo si ferma appena sotto a 0.7, seguiti dai risultati sul dataset F2 (R^2 predittivo appena sotto 0.5) e da quelli su F4, poco sopra lo 0.3. Le nuove variabili, costruite da zero, si sono rilevate essere tra le più rilevanti per la previsione, figurando tra le più importanti sulla base della *foresta casuale*.

Nel lavoro di Vela et al. (2022) viene indicato come i modelli, regolati usando un anno di osservazioni, raggiungano un R^2 predittivo iniziale di convalida incrociata superiore a 0.7; in questo lavoro è stato tuttavia impossibile ripro-

durre lo stesso livello di qualità.

Ottener livelli superiori a questi, limitando l'informazione utilizzata allo stesso modo, non è semplice. Nell'articolo gli autori hanno affrontato un problema differente, avendo previsto il "tempo di attesa del prossimo paziente". Ciò permette di utilizzare informazione molto più recente nella previsione, fino al paziente precedente, elevando le prestazioni dei modelli. Questo approccio al problema non è tuttavia corretto, o quantomeno realistico: perché la previsione sia significativa deve pervenire al paziente con un certo anticipo, e non quando il paziente precedente inizia la visita (ciò potrebbe addirittura avvenire una volta passato l'orario di appuntamento del paziente per cui stiamo facendo la previsione). Limitare l'informazione come fatto in questo lavoro è quindi l'approccio più corretto.

Detto questo, non vi è ragione di non studiare la *degradazione temporale* quando la qualità iniziale dei modelli è più bassa; nella realtà, non sempre questa è pienamente soddisfacente. Per queste ragioni il "test" di *degradazione temporale* è stato applicato su tutti e quattro i dataset.

2.3 Analisi della *degradazione temporale*

Le figure 2.2 - 2.5 presentano l'andamento nel tempo della variabile risposta, per ciascuna struttura. Dalle immagini si può vedere come il comportamento del tempo di attesa (Wait) non presenta cambiamenti improvvisi, che segnerebbero un processo generatore dei dati che evolve; non ci sono ragioni di pensare che la qualità dei modelli possa *degradare*.

La lunghezza dei dataset è molto diversa: per le strutture F1 ed F2 sono disponibili più di 1000 giorni di osservazioni, mentre per F3 ed F4 attorno a 600. Negli ultimi casi quindi la distanza massima, dal momento di stima, a cui è possibile misurare le prestazioni del modello (dT massimo) è piuttosto ridotta. Questo limita anche il tratto del flusso di dati da cui possiamo ri-

cavare degli insiemi di stima, che saranno in numero inferiore. In tutti i casi l'ampiezza dell'insieme di stima è stata fissata ad un anno di osservazioni, la stessa utilizzata nell'articolo.

I risultati per il dataset F1 sono riportati nelle figure 2.6 e 2.7, per il dataset F2 in 2.8 e 2.9, per il dataset F3 in 2.10 e 2.11, mentre per F4 in 2.12 e 2.13.

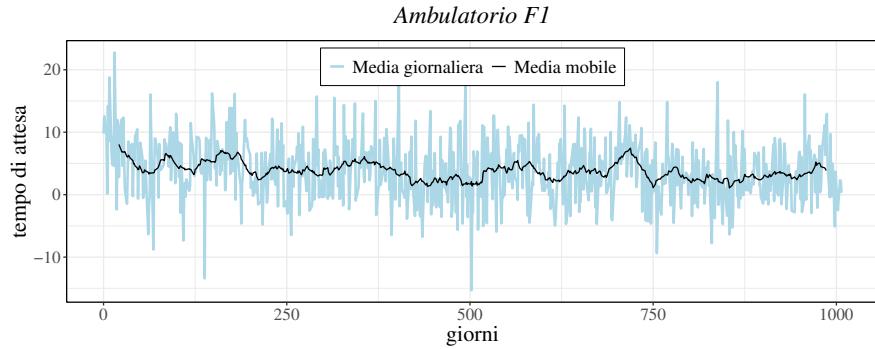


Figura 2.2: Variabile d'interesse, il tempo di attesa (Wait), nel dataset F1. Media giornaliera e media mobile.

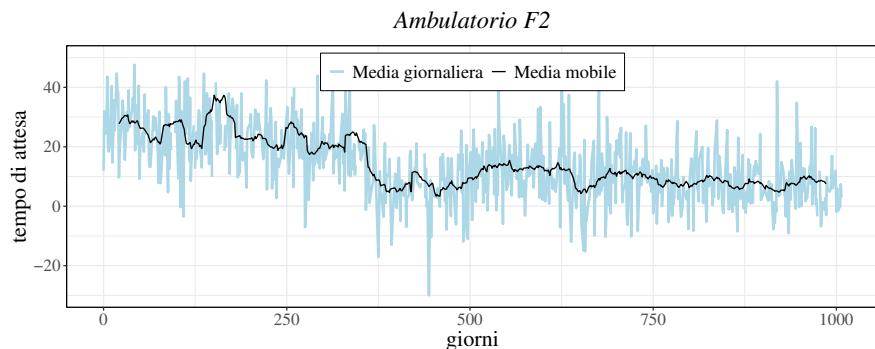


Figura 2.3: Variabile d'interesse, il tempo di attesa (Wait), nel dataset F2. Media giornaliera e media mobile.

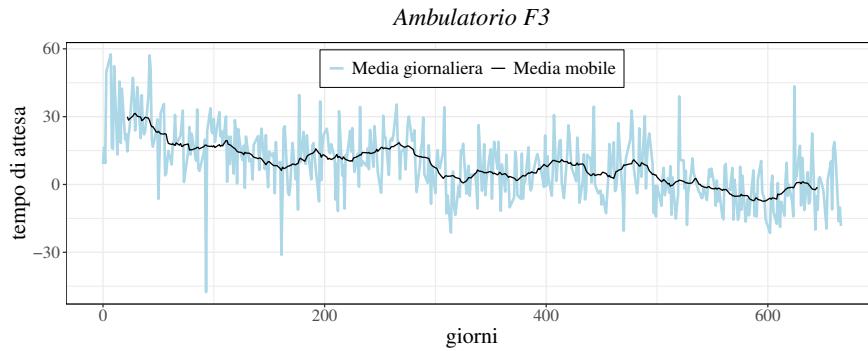


Figura 2.4: Variabile d'interesse, il tempo di attesa (Wait), nel dataset F3. Media giornaliera e media mobile.

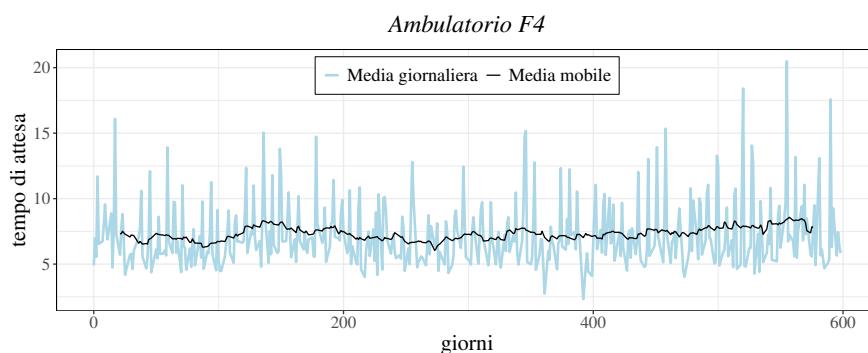


Figura 2.5: Variabile d'interesse, il tempo di attesa (Wait), nel dataset F4. Media giornaliera e media mobile.

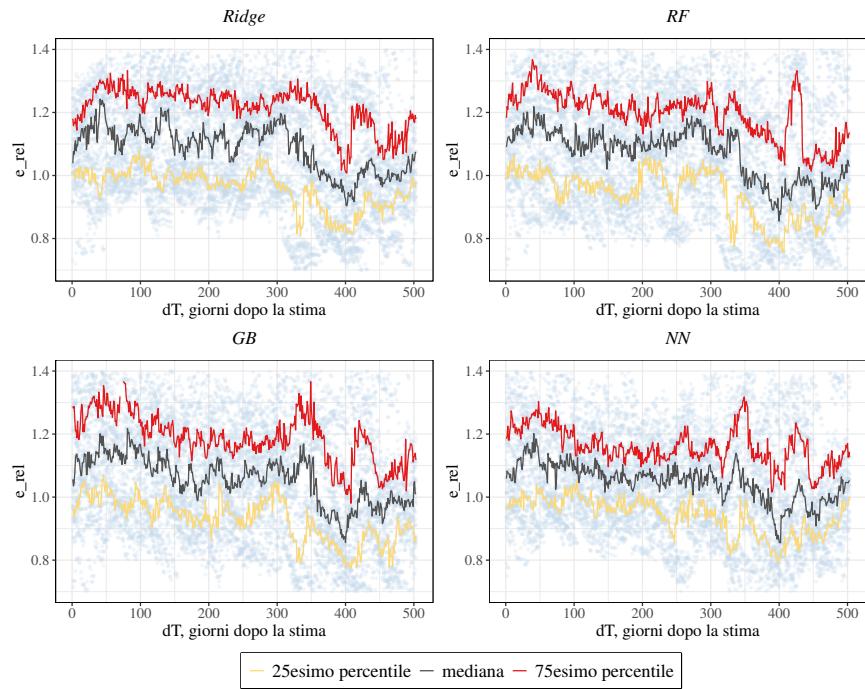
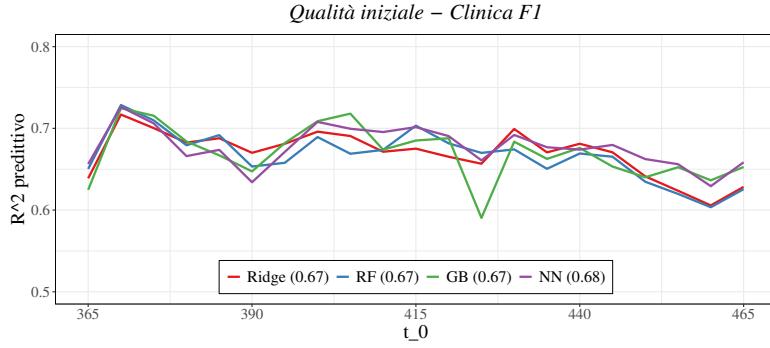


Figura 2.7: Risultati del test per la *degradazione temporale*, applicato al dataset F1. Gli insiemi di stima sono stati ricavati dal tratto iniziale, che va dal giorno 0 al giorno 465. L'insieme di convalida e la finestra mobile per valutare l'errore relativo (e_{rel} (dT)) sono entrambi di ampiezza 20 giorni.

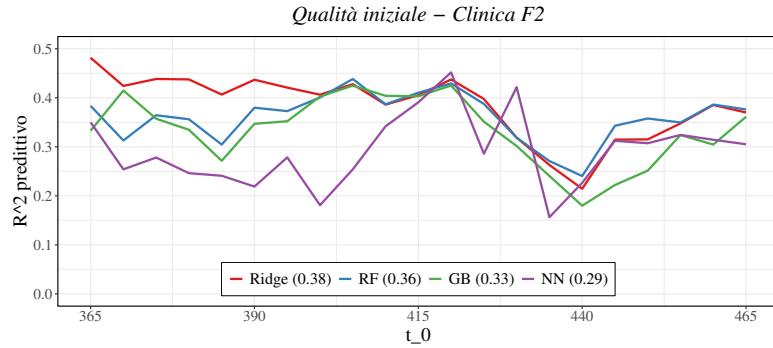


Figura 2.8: Qualità iniziale dei modelli nel dataset F2, ottenuta come R^2 predittivo negli insiemi di verifica, successivi rispetto ai dataset di stima, e la media per ciascun modello. Il valore sull'asse delle ascisse corrisponde all'ultimo giorno presente nell'insieme di stima (t_0) utilizzato per stimare i modelli. In questo caso i modelli presentano una differenza notevole nelle prestazioni, soprattutto nel tratto iniziale.

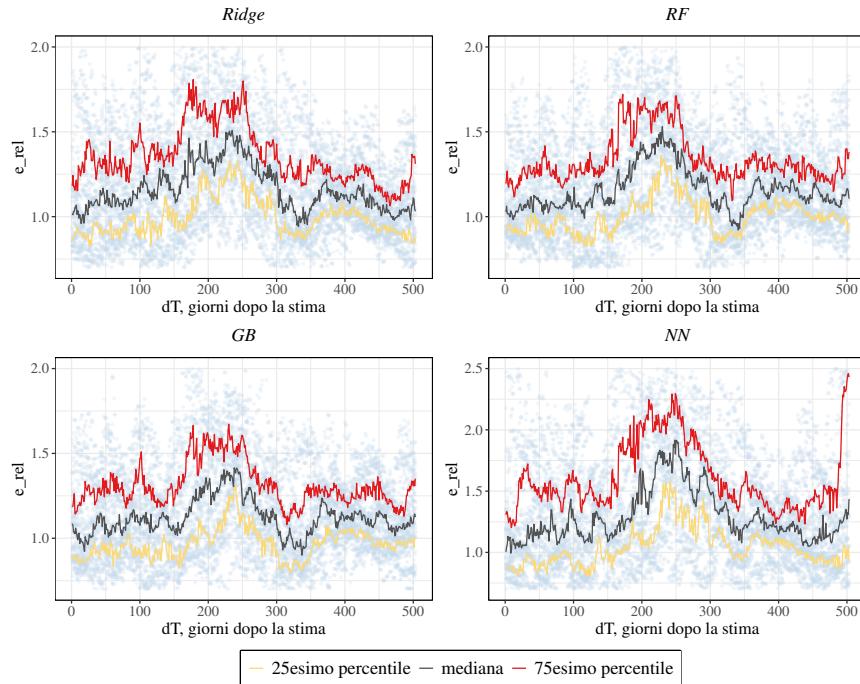


Figura 2.9: Risultati del test per la *degradazione temporale*, applicato al dataset F2. Gli insiemi di stima sono stati ricavati dal tratto iniziale, che va dal giorno 0 al giorno 465. L'insieme di convalida e la finestra mobile per valutare l'errore relativo (e_{rel} (dT)) sono entrambi di ampiezza 20 giorni.

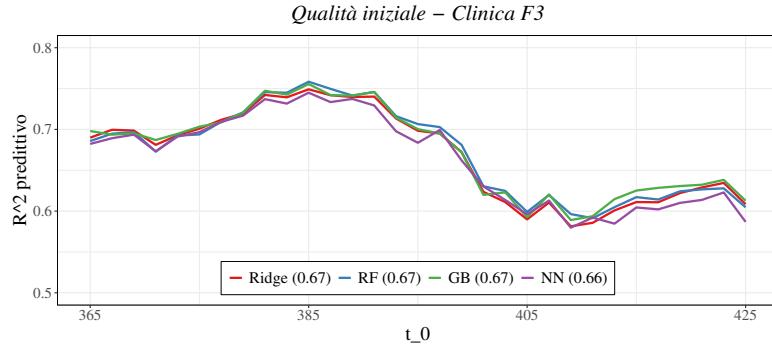


Figura 2.10: Qualità iniziale dei modelli nel dataset F3, ottenuta come R^2 predittivo negli insiemi di verifica, successivi rispetto ai dataset di stima, e la media per ciascun modello. Il valore sull'asse delle ascisse corrisponde all'ultimo giorno presente nell'insieme di stima (t_0) utilizzato per stimare i modelli.

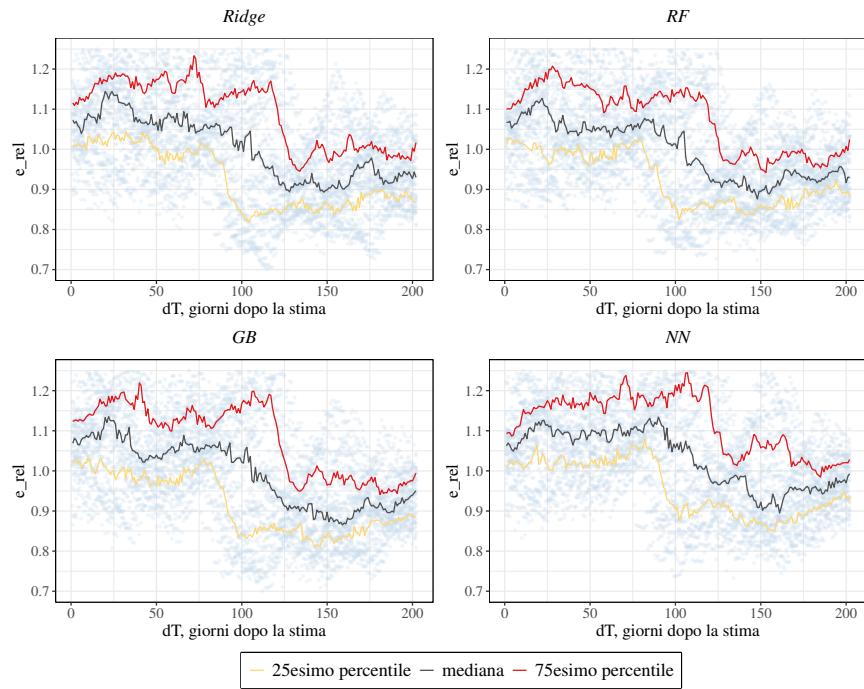


Figura 2.11: Risultati del test per la *degradazione temporale*, applicato al dataset F3. Gli insiemi di stima sono stati ricavati dal tratto iniziale, che va dal giorno 0 al giorno 425. L'insieme di convalida e la finestra mobile per valutare l'errore relativo (e_{rel} (dT)) sono entrambi ampiezza di ampiezza 20 giorni.

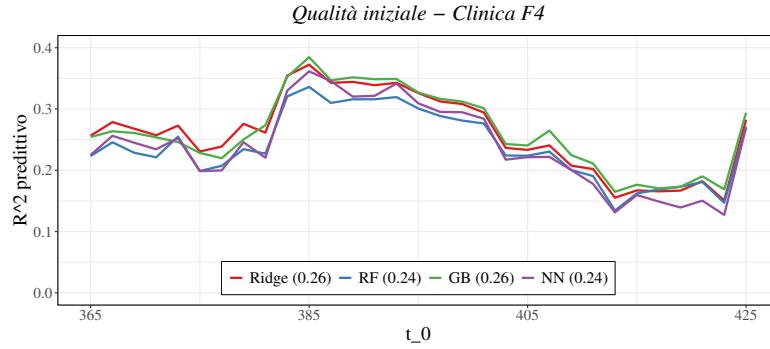


Figura 2.12: Qualità iniziale dei modelli nel dataset F4, ottenuta come R^2 predittivo negli insiemi di verifica, successivi rispetto ai dataset di stima, e la media per ciascun modello. Il valore sull'asse delle ascisse corrisponde all'ultimo giorno presente nell'insieme di stima (t_0) utilizzato per stimare i modelli.

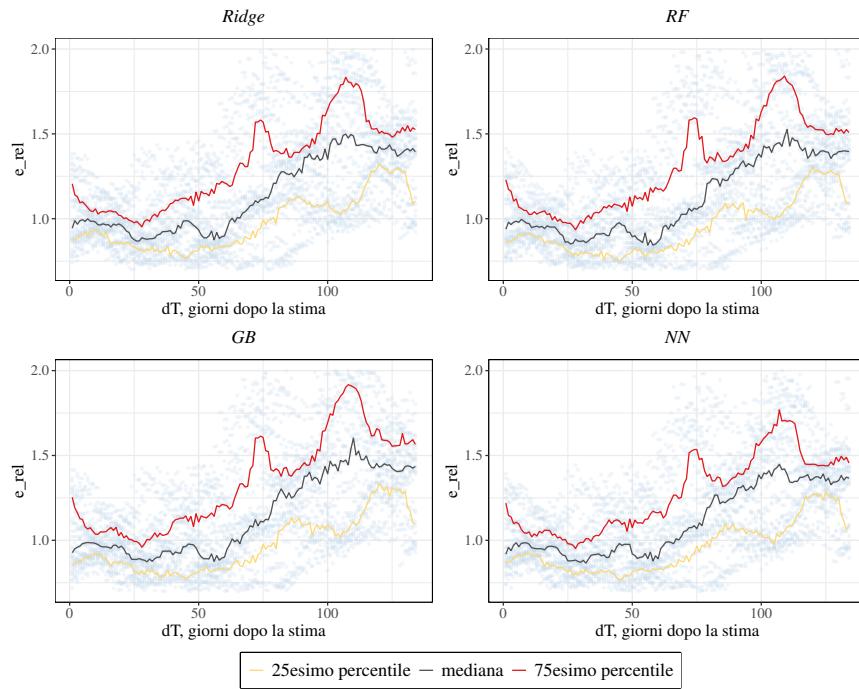


Figura 2.13: Risultati del test per la *degradazione temporale*, applicato al dataset F4. Gli insiemi di stima sono stati ricavati dal tratto iniziale, che va dal giorno 0 al giorno 425. L'insieme di convalida e la finestra mobile per valutare l'errore relativo (e_{rel} (dT)) sono entrambi ampiezza di ampiezza 20 giorni.

Dai risultati presentati è possibile affermare che:

1. Nei due dataset F1 ed F3 non sono presenti segni di *degradazione temporale*, in quanto vi è assenza di un trend positivo nell'andamento dei quantili e la variabilità dell'errore relativo non aumenta con la distanza. I modelli quindi non invecchiano, e anzi, sembra esserci il trend opposto: l'errore relativo si riduce con l'aumento dell'età (dT) del modello.
2. Nei due dataset F2 ed F4 possiamo notare invece dei segni di *degradazione temporale*. Nel caso del dataset F2, nonostante non ci sia un trend nei quartili, o un aumento della variabilità dell'errore, le prestazioni sono meno stabili: ci sono dei chiari picchi nell'errore relativo, che poi si riassorbono.

Nel caso di F4 è invece presente un chiaro andamento positivo, anche se la distanza massima dT è troppo piccola per sapere se si tratti di un picco isolato, come in F2, o di un peggioramento progressivo che proseguirebbe all'aumentare della distanza.

Questi sono casi in cui andrebbe considerata l'implementazione di metodi per l'aggiornamento dei modelli; e sarebbe una scelta giustificata, considerando che nessuno dei modelli impiegati si distingue con una buona stabilità e i cali di prestazioni sono condivisi negli stessi tratti.

Le differenze nel comportamento generale dei modelli, osservabili dalla rappresentazione grafica, sono minime. In tutti i dataset le prestazioni dei 4 modelli si riducono sui medesimi tratti e la forma complessiva è analoga, ed è dunque necessario utilizzare delle misure aggiuntive per determinare il miglior modello da utilizzare.

Nel caso di F1, F2 ed F3 non osserviamo un aumento progressivo dell'errore, perciò l'informazione più importante da ricavare è legata ai livelli dell'errore relativo: il livello medio del caso peggiore (terzo quartile) è la più importante, ma è possibile considerare anche il livello del caso mediano. Questi sono stati sintetizzati tramite media e riportati nelle tabelle 2.5 (per F1), 2.6 (per F2) e 2.7 (per F3), sotto forma di variazioni percentuali.

Dataset F1	R^2 <i>predittivo</i>	$MSE(t_0)$ medio	Mediana	Terzo quartile
Ridge	0.67	101.90	8.73%	21.38%
RF	0.67	102.68	6.84%	19.58%
GB	0.67	101.76	5.56%	18.82%
NN	0.68	99.36	4.48%	16.04%

Tabella 2.5: Dataset F1. MSE iniziale medio ed R^2 *predittivo* iniziale medio, livello medio della mediana e del terzo quartile, resi variazioni percentuali. Un livello medio della mediana pari a 1.0873 (*ridge*) viene convertito in variazione percentuale di 8.73%, e così per gli altri valori.

Dataset F2	R^2 <i>predittivo</i>	$MSE(t_0)$ medio	Mediana	Terzo quartile
Ridge	0.38	465.67	16.06%	35.91%
RF	0.36	485.91	15.62%	34.22%
GB	0.33	503.35	12.26%	31.31%
NN	0.29	540.42	29.04%	61.73%

Tabella 2.6: Dataset F1. MSE iniziale medio ed R^2 *predittivo* iniziale medio, livello medio della mediana e del terzo quartile, resi variazioni percentuali.

Dataset F3	R^2 <i>predittivo</i>	$MSE(t_0)$ medio	Mediana	Terzo quartile
Ridge	0.67	733.30	0.31%	9.01%
RF	0.67	726.19	-0.36%	7.59%
GB	0.67	724.06	-0.81%	7.95%
NN	0.66	746.90	3.12%	11.96%

Tabella 2.7: Dataset F3. MSE iniziale medio ed R^2 *predittivo* iniziale medio, livello medio della mediana e del terzo quartile, resi variazioni percentuali.

I modelli che hanno presentato la stabilità maggiore, ovvero i livelli più bassi per il terzo quartile (e successivamente la mediana) in questi tre casi sono

stati la *rete neurale* nel dataset F1, il *gradient boosting* nel dataset F2 e la *foresta casuale* nel dataset F3, riportati in grigio.

Differenze, seppur leggere, nel modo in cui i modelli invecchiano sono quindi state osservate, e questa informazione può essere combinata a quella relativa alla qualità iniziale per scegliere il modello migliore. Una misura dei livelli iniziali può essere ottenuta combinando i valori di $MSE(t_0)$ (tramite media), calcolati sui singoli insiemi di verifica e riportati nelle tabelle. La scelta di utilizzare l' MSE invece che l' R^2 *predittivo* è dovuta al fatto che permette considerazioni più precise, in quanto non è arrotondato quanto il secondo.

Possiamo quindi concludere che:

1. In F1 la scelta ottimale è la *rete neurale*, il modello inizialmente migliore e il più stabile.
2. In F2 i candidati sono lo stimatore *ridge*, con l'errore medio iniziale più basso, il *gradient boosting*, il più stabile, e la *foresta casuale*, una via di mezzo. Combinando l' MSE con gli aumenti nel caso peggiore, l'ordinamento basato sulla qualità iniziale si mantiene: $Ridge = 632.90$, $RF = 652.19$, $GB = 660.94$ (l' $MSE(t_0)$ medio è stato aumentato delle variazioni percentuali).
3. In F3 la stabilità dei modelli è molto simile, soprattutto per i modelli inizialmente migliori. In questo caso scegliere tra *foresta casuale* e *gradient boosting* non è semplice, in quanto nel caso mediano preferiamo di poco il secondo ($RF = 723.58$, $GB = 718.19$), in quello peggiore il primo ($RF = 781.31$, $GB = 781.62$). La differenza è minima, ma conduce alla scelta del *gradient boosting*, con migliore qualità iniziale, caso mediano e stesso caso peggiore.

Come possiamo osservare è quindi possibile utilizzare l'informazione ricavata dal “test” di *degradazione temporale* per selezionare il modello. Particolarmente interessante il caso F1, in cui tutte le categorie di modello raggiungono prestazioni iniziali analoghe ma mostrano diversi livelli di stabilità delle prestazioni. In modo analogo il caso F3, in cui la differenza tra le prestazioni iniziali di *rete neurale* e *foresta casuale* è molto ridotta, ma la differenza, in

media, tra i quartili è elevata.

Non solo, il modello migliore non è necessariamente quello le cui prestazioni sono più prevedibili: nel caso F2 lo stimatore *ridge* ha una qualità iniziale molto più elevata del *gradient boosting*, ma una maggiore variabilità delle prestazioni.

Nel caso F4 l'errore relativo presenta un trend positivo, e quindi le misure di stabilità basate sui livelli, da sole, sono meno informative sul comportamento del modello all'allontanarsi dal momento della stima. Queste sono comunque riportate nella tabella 2.8, in quanto l'aumento potrebbe rappresentare un picco isolato, come in F2. Ipotizzando però che la crescita continui questa può essere quantificata tramite l'inclinazione della mediana (stimata tramite modello di regressione lineare), riportata nella tabella 2.8. Questo approccio ha dei difetti, in quanto la mediana non ha una forma esattamente lineare, però fornisce un'indicazione del grado di *degradazione temporale*.

Dataset F4	R^2 predittivo	$MSE(t_0)$ medio	Mediana	Terzo quartile	Var. e_{rel}
Ridge	0.26	28.12	13.74%	31.58%	0.67
RF	0.24	28.97	11.52%	30.34%	0.67
GB	0.26	28.03	14.5%	34.48%	0.71
NN	0.24	28.88	11.36%	28.11%	0.61

Tabella 2.8: Dataset F3. MSE iniziale medio ed R^2 predittivo iniziale medio, livello medio della mediana e del terzo quartile, resi variazioni percentuali, e inclinazione della mediana, riportata come variazione dell'errore relativo (più informativa di una pendenza, come indicato nella sezione 1.2).

Il modello con la stabilità maggiore è la *rete neurale*. Confrontandola con i modelli inizialmente migliori, lo stimatore *ridge* e il *gradient boosting*, osserviamo che:

1. Nel caso peggiore *ridge* e *rete neurale* presentano il medesimo livello medio di MSE , pari a 37, mentre il *gradient boosting* pari a 37.7.

2. La variazione dell'errore relativo, secondo la retta stimata sulla media-
na, è molto minore per la *rete neurale*, indicando una minore *degrada-
zione temporale*.

Le differenze in questo caso sono piccole, tuttavia possono indicare una preferenza ad utilizzare la *rete neurale*, considerando che la sua qualità potrebbe decadere di meno nel lungo periodo.

Questi primi risultati permettono quindi di discutere quanto presentato in Vela et al. (2022), in particolare:

1. I modelli possono presentare segni di *degradazione temporale* anche in contesti in cui non abbiamo motivo di sospettare che ciò avvenga: nei casi F2 ed F4 in effetti l'andamento della variabile risposta non preannuncia la perdita di prestazioni osservata (nel primo caso il picco inizia a 200 giorni dal momento della stima). Tuttavia,
2. il fatto che la *degradazione temporale* possa presentarsi in modo diver-
so da un modello ad un altro, sugli stessi dati, non può essere con-
fermato. Tutti i modelli presentano il medesimo comportamento, in
quanto la forma complessiva del grafico di *AI Aging* corrisponde qua-
si perfettamente (le differenze osservate in F4 sono minime). Detto
questo,
3. i modelli possono, come è stato osservato, presentare diversi livelli di
stabilità delle prestazioni, sugli stessi dati, anche quando la qualità
iniziale è molto simile (tabelle 2.5, 2.7, 2.8). Sugli stessi dati, i modelli
hanno presentato differenze nel livello dei quartili, indicando che la
qualità di alcuni fosse meno prevedibile, e diverse inclinazioni per la
mediana (tabella 2.8), indicando che differenze, seppur molto leggere,
nel grado di *degradazione temporale* possono essere osservate.

Cosa determina quindi le differenze osservate nei modelli? In quali situazioni possiamo aspettarci che l'applicazione del “test” evidenzi delle differenze utili per la selezione del modello migliore? Le simulazioni condotte, presentate nei capitoli 3, 4 e 5, cercano di rispondere a queste domande.

Nonostante l'applicazione del "test" mostri del potenziale è decisamente oneroso, sia in termini di tempi richiesti che di quantità di dati necessari. Conoscere a priori quali modelli presenteranno maggiore stabilità futura può aiutare a considerare quest'informazione quando ricorrere a tale procedura non è possibile.

Capitolo 3

Simulazioni: assenza di *concept drift*

In questo capitolo vengono presentati i primi risultati delle simulazioni condotte, nelle quali la stabilità dei modelli è stata verificata in contesti in cui il processo generatore dei dati non evolve. Facendo riferimento alla sezione 1.2.2, in cui sono state introdotte le situazioni studiate, le simulazioni presentate di seguito contengono i casi appartenenti alla prima categoria.

In Vela et al. (2022) gli autori indicano come *degradazione temporale* sia stata osservata anche in casi in cui la variabile risposta mantiene un comportamento apparentemente regolare. Immaginare come un modello possa presentare un errore relativo che aumenta nel tempo, in assenza di cambiamenti, è però difficile. Dalle prime analisi, condotte su dataset reali, è stato però osservato come, anche in assenza delle principali manifestazioni di *degradazione*, nei casi in cui la qualità dei modelli si mantiene nel tempo, le categorie considerate possono presentare diversa variabilità dell'errore relativo (valutata usando il livello del terzo quartile, più indicativo della distanza tra quartili); e l'informazione ricavata può aiutare nel processo di selezione del modello da utilizzare.

Le prime simulazioni presentate in questo capitolo trattano casi in cui la componente temporale non ha alcuna rilevanza. Le osservazioni che costi-

tuiscono i dataset sono quindi indipendenti, e le regole che generano i dati rimangono invariate durante l'intero periodo, con l'idea di evidenziare differenze nella variabilità delle prestazioni.

In seguito vengono estese le situazioni considerate, cercando di simulare i casi reali in cui la variabile risposta presenta dipendenza con i suoi valori passati. Ciò può accadere per diverse ragioni: in questo caso viene simulata la presenza di fattori esterni tramite processi autocorrelati non osservati.

Situazioni di questo tipo non comportano un'evoluzione del processo, in quanto le regole attraverso le quali vengono generati i dati rimangono invariate nel tempo, tuttavia portano ad osservare dei cambi di livello nel tempo, nella variabile risposta, e ciò può influire sulla stabilità delle prestazioni.

Situazioni indipendenti dal tempo

In queste simulazioni la struttura del dataset e le sue caratteristiche sono state modificate, per valutare come i modelli rispondono di fronte a diverse situazioni. Partendo dal caso più semplice, in cui la struttura dei dataset simulati presenta una relazione lineare tra variabili esplicative e variabile risposta, questa è stata resa più complessa, tramite l'utilizzo di relazioni non lineari. In particolare è stato studiato il caso di relazioni cubiche e di effetti di interazione.

Per verificare l'impatto delle scelte arbitrarie sulla struttura del modello, vale a dire l'importanza della componente non osservata e la struttura di correlazione tra le X , la simulazione con relazioni lineari è stata ripetuta, al variare di queste componenti. Idealmente ciò andrebbe ripetuto in presenza di relazioni più complesse, ma ciò non può essere effettuato a causa dei lunghi tempi richiesti dalle simulazioni, rendendo impraticabile l'analisi. Detto questo, la struttura lineare è la più importante, in quanto la base per le simulazioni successive e perché permette di confrontare tutti e quattro i modelli in situazioni in cui le prestazioni iniziali sono simili (in particolare dello stimatore *ridge*, che altrimenti potrebbe essere escluso, in quanto non avrebbe prestazioni soddisfacenti).

Un altro aspetto esplorato in queste simulazioni è l'effetto, sulla stabilità dei modelli, delle diverse qualità iniziali che questi possono raggiungere, sugli stessi dati. Come già menzionato in precedenza il “test” di *degradazione temporale* non permette di considerare i diversi livelli iniziali dei modelli, un'informazione piuttosto rilevante, dovendo selezionare il modello migliore da utilizzare per svolgere un determinato compito.

Per verificare ciò alcune simulazioni sono state replicate con un numero di osservazioni ridotto, in modo da indurre una riduzione nella qualità iniziale dei modelli tramite una spazio delle esplicative più sparso. Nonostante questa possa non essere l'unica ragione che porta ad una riduzione della qualità, i modelli considerati (ad eccezione dello stimatore *ridge*) hanno il potenziale di adattarsi bene a qualsiasi struttura considerata, con un numero sufficiente di osservazioni. Individuare altri casi in cui, sugli stessi dati, i modelli presentassero qualità iniziali molto differenti (quando regolati e stimati correttamente) non è stato possibile.

Infine, la stabilità dei modelli è stata verificata in alcune situazioni note per mettere in difficoltà i modelli nei casi reali: quando la componente non osservata è maggiore (e quindi, complessivamente, la qualità iniziale è molto minore) e quando sono presenti variabili di disturbo (lo stimatore *ridge* e la *rete neurale* potrebbero essere particolarmente colpiti, in quanto non possono escludere le variabili). Dovendo individuare situazioni che possano minare la stabilità dei modelli è ragionevole partire dai casi noti per impattare negativamente la qualità di questi.

Complessivamente questi studi forniscono una panoramica sul comportamento dei modelli a medio/lungo termine, al variare di alcune caratteristiche dei dati e della qualità iniziale, sia come livello complessivo raggiunto dall'insieme dei modelli, sia come differenze osservate tra i modelli, sugli stessi dati.

L'informazione più importante da ricavare, in questi casi, riguarda il livello medio del 75esimo percentile, utilizzato per valutare la variabilità delle prestazioni. Non è infatti ragionevole attendersi le manifestazioni più im-

portanti di *degradazione temporale*, ovvero l'aumento progressivo dell'errore o della sua variabilità. L'errore, come funzione della distanza dal momento di stima, viene infatti misurato su finestre mobili, che in questi casi saranno uguali per processo generatore dei dati (PGD). Gli aumenti eventualmente osservati saranno dovuti alla sola generazione casuale delle osservazioni, e non alle caratteristiche del modello. Invece, le diverse situazioni considerate possono avere un impatto sulla qualità iniziale, misurata sulla prima finestra disponibile (l'insieme di verifica), identica per PGD alle finestre mobili; il livello raggiunto tenderà dunque a mantenersi nel tempo.

Dataset con relazioni lineari

Iniziando con una prima simulazione, caratterizzata da relazioni lineari tra variabili esplicative e variabile risposta, la struttura dei dataset simulati è descritta attraverso la seguente equazione:

$$Y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \cdots + \beta_{10} X_{10,ij} + \varepsilon_{ij}$$

i = giorno; j = identificativo dell'osservazione nel giorno i

dove ε_{ij} è un termine di errore gaussiano, con distribuzione $N(0, \sigma_\varepsilon^2)$.

Solo in questo caso sono state riportate tutte le metriche tenute sotto controllo, per fornire una panoramica dello stato iniziale delle simulazioni, ma alcune sono rilevanti solo in contesti in cui le prestazioni decadono nel tempo, e non sono i casi di questo capitolo. La qualità iniziale dei modelli è riportata nella figura 3.1, mentre i singoli grafici di *AI Aging* sono stati combinati nella figura 3.2. La distribuzione del livello medio del terzo quartile, per ogni modello, è riportata nella figura 3.3, mentre in figura 3.4 è riportata la distribuzione della pendenza della mediana, misurata come differenza tra il valore finale e iniziale della retta interpolata.

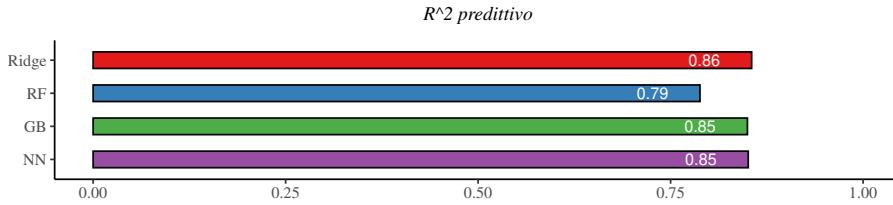


Figura 3.1: Prestazioni iniziali medie dei modelli nella simulazione in cui la struttura del dataset presenta relazioni lineari. I modelli hanno una simile qualità iniziale, anche se la *foresta casuale* presenta dei valori più bassi.

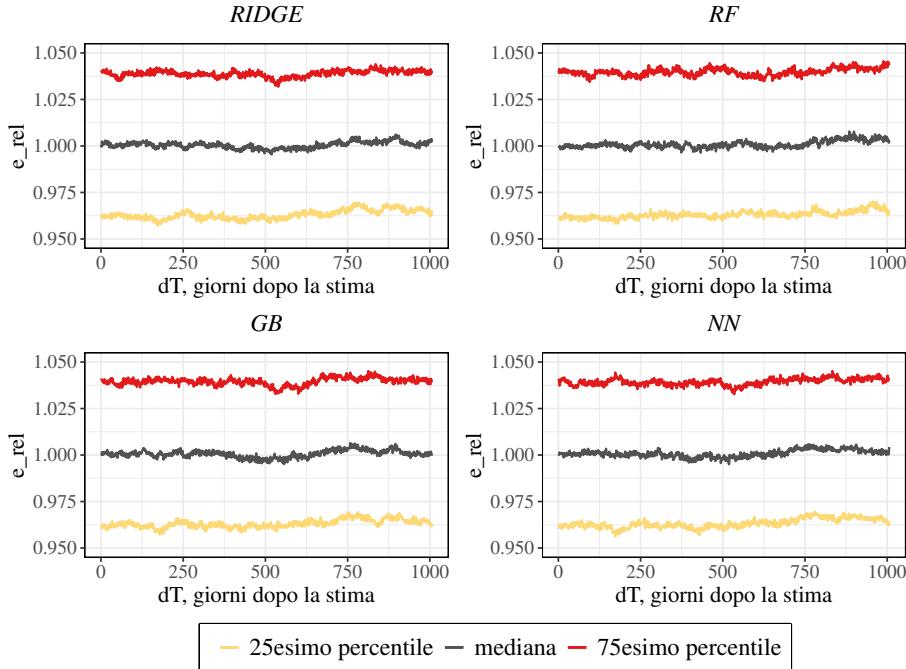


Figura 3.2: Combinazione dei grafici di *AI Aging*. La combinazione delle mediane mostra un'assenza di trend, mentre gli altri quartili mostrano come la qualità iniziale dei modelli si mantenga in modo simile per ogni modello. La tendenza complessiva, intesa come la forma dei grafici, coincide praticamente perfettamente.

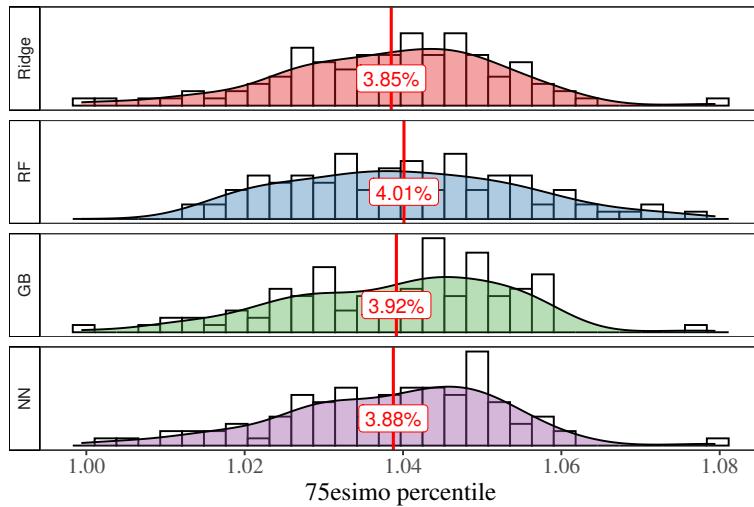


Figura 3.3: Distribuzione del livello medio del 75esimo percentile nelle replicazioni della simulazione. Tutti i modelli presentano una distribuzione simile, indicando una simile variabilità dell'errore relativo, e il livello medio, in modo analogo, è molto simile.

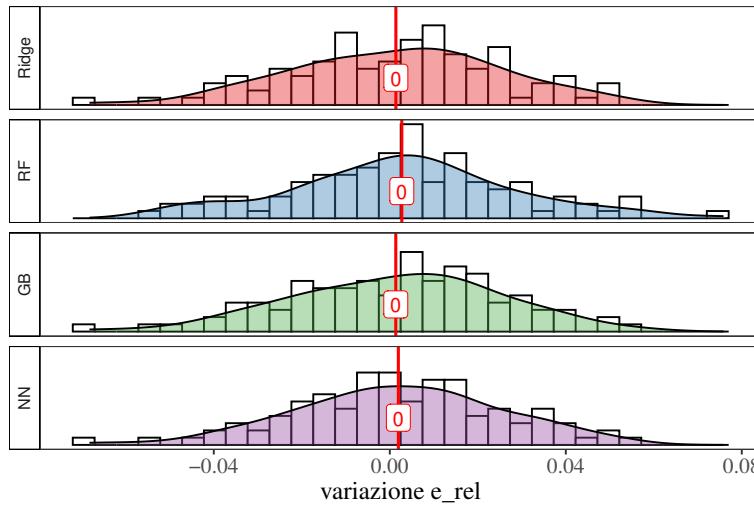


Figura 3.4: Distribuzione delle variazioni dell'errore relativo nelle replicazioni della simulazione. Tutti i modelli presentano la stessa distribuzione, con media attorno a 0, indicando assenza di *degradazione temporale*. I valori spaziano tra -0.08 e 0.08, indicando come nei casi con più decadimento della qualità questa si riduca, in 1006 giorni, dell'8%.

Come era atteso, nessun modello tende a presentare *degradazione temporale* (3.2 e 3.4), e le distribuzioni del terzo quartile sono molto simili (3.3), quasi totalmente sovrapposte.

Una leggera differenza può essere notata nella media del caso peggiore, tra replicazioni: tutti i modelli presentano un livello analogo, ad eccezione della *foresta casuale*, con un livello medio marginalmente più elevato. Detto questo la differenza è minima, considerando il campo di variazione osservato: nel caso peggiore, in media, mentre la qualità dello stimatore *ridge* aumenta del 3.85%, quella della *foresta* del 4.01%. L'entità di questa differenza è ancora più chiara dal grafico in figura 3.2, in cui osserviamo come i comportamenti di medio/lungo periodo dei modelli, combinati, siano praticamente identici dal punto di vista della variabilità dell'errore.

Come verrà chiarito tramite le successive simulazioni, quantificare la differenza non è possibile, in quanto i livelli cambiano da un caso all'altro, e dipendono da fattori indipendenti dai modelli, come il numero di osservazioni. È possibile però valutare l'ordinamento dei modelli e le distanze relative.

Quanto impatto hanno avuto, su questi risultati, le scelte arbitrarie relative alla struttura del flusso di dati? In particolare, quanto impatto ha la quantità di errore presente e la matrice di varianza/covarianza delle variabili esplicative?

Per rispondere a queste domande sono stati condotti altri esperimenti simulativi, in cui gli stessi coefficienti sono stati mantenuti, ma questi elementi modificati. In particolare, sono stati simulati dei flussi di dati in cui le variabili esplicative sono prima indipendenti, e poi altamente correlate (le matrici utilizzate sono riportate in appendice, figura A.1, così come la distribuzione dei valori medi del terzo quartile e le qualità iniziali, figure A.2 e A.3). Allo stesso modo è stata aumentata la varianza della componente non osservata, permettendo di valutare la stabilità in un contesto in cui la qualità iniziale complessiva dei modelli è minore (i risultati rilevanti sono riportati in appendice, figura A.4).

Al variare della dipendenza tra le variabili esplicative la qualità iniziale ha subito minime variazioni, ma non ci sono differenze sostanziali con il caso

precedente. Allo stesso modo, l'aumento del termine di errore ha comportato livelli iniziali minori (R^2 *predittivo* tra 0.4 e 0.5, e la *RF* rimane comunque il modello peggiore), ma non ha evidenziato alcuna differenza nella variabilità delle prestazioni tra i modelli.

L'aumento dell'importanza della componente non osservata ha quindi ridotto le differenze nella stabilità, anche se le distanze nei livelli iniziali sono rimaste di simile entità.

Qualità iniziale e numero di osservazioni

Un aspetto degno di approfondimento è l'effetto della diversa qualità iniziale dei modelli: la *foresta casuale* ha infatti mantenuto dei livelli di R^2 *predittivo* iniziali inferiori, e non è chiaro come ciò influenzi i risultati. Valutarne l'effetto si traduce nell'intervenire sulla struttura del dataset e sulle caratteristiche dei dati, in modo da indurne l'aumento o la riduzione, e valutare la stabilità in tali circostanze.

Questo tuttavia non si limita alla sola *foresta casuale*, ma può anche essere esteso agli altri modelli: in una situazione reale le differenze nella qualità iniziale sono attese, e ciò rende necessario il procedimento di scelta del modello. Al variare delle differenze iniziali però, come varia la stabilità dei modelli?

Agire sulla qualità iniziale non è semplice, in quanto tutte le categorie di modello considerate si adattano bene a relazioni lineari (e ad eccezione dello stimatore *ridge*, a molte situazioni più o meno complesse). È stato quindi ridotto il numero di osservazioni (più semplice rispetto ad aumentare il numero di variabili), aumentando di conseguenza la sparsità dello spazio (maledizione della dimensionalità) e riducendo la capacità dei modelli di adattarsi alla relazione.

Anche operando in questo modo, i modelli soffrono del problema in modo differente, in particolare la *foresta casuale*, la cui qualità si riduce eccessivamente prima di osservare differenze iniziali significative negli altri modelli.

Questo non è il solo problema: nelle figure 3.5 e 3.6 sono riportati i risultati di due simulazioni, condotte cambiando il numero di osservazioni ma

mantenendo altrimenti inalterata la struttura dei dataset, prima con 5 casi al giorno e poi con 50. È immediato osservare come la variazione del numero di osservazioni disponibili abbia un impatto considerevole sull'esito del "test" di *degradazione temporale*, ma non necessariamente attraverso la qualità dei modelli.

La riduzione del numero di osservazioni ha sicuramente comportato una riduzione della qualità iniziale della *foresta casuale*, l'unico modello veramente colpito, e il divario nei livelli del terzo quartile è aumentato. La riduzione del numero di osservazioni ha però comportato un aumento della variabilità dell'errore per tutti i modelli, seppur abbiano mantenuto le loro prestazioni iniziali. Allo stesso modo, l'aumento di osservazioni ha portato ad una riduzione della variabilità e delle differenze in stabilità tra i modelli, anche se non ha influito sulla qualità iniziale. Ciò è probabilmente dovuto a delle misurazioni degli errori più stabili, dovute alla maggior dimensione delle finestre mobili, la cui ampiezza in giorni è rimasta invariata.

Ridurre il numero di variabili considerate permette però di osservare come, a parità di qualità iniziale, la stabilità dei modelli sia la stessa (figura 3.7). È quindi ragionevole concludere che la maggiore instabilità della *foresta*, seppur minima, fosse attribuibile al livello iniziale più basso. Nelle simulazioni successive l'effetto della qualità iniziale sulla stabilità sarà approfondito.

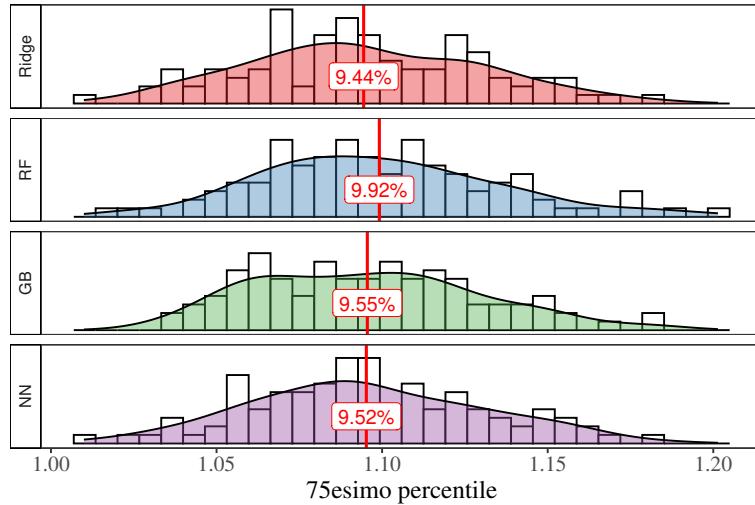


Figura 3.5: Distribuzione del livello medio del 75esimo percentile, con 5 osservazioni al giorno. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.85, RF = 0.74, GB = 0.83 e NN = 0.84. Le prestazioni iniziali della RF si sono ridotte maggiormente.

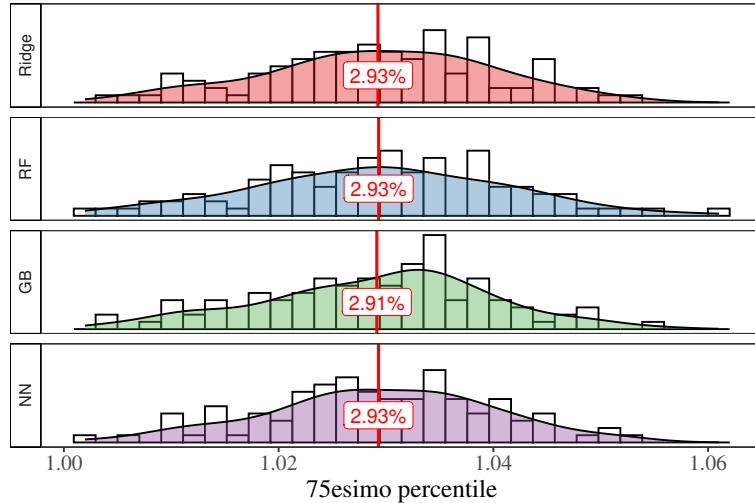


Figura 3.6: Distribuzione del livello medio del 75esimo percentile, con 50 osservazioni al giorno. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.86, RF = 0.8, GB = 0.85 e NN = 0.85.

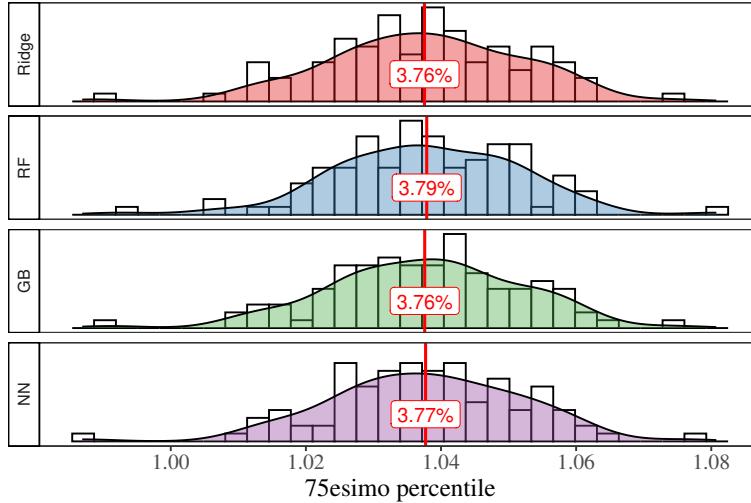


Figura 3.7: Distribuzione del livello medio del 75esimo percentile, con 30 osservazioni al giorno e 4 variabili esplicative. I livelli di R^2 predittivo iniziale raggiunti sono: $ridge = 0.87$, $RF = 0.85$, $GB = 0.86$ e $NN = 0.86$.

Variabili di disturbo e maggiore componente non osservata

In aggiunta alle situazioni appena studiate sono stati simulati dei flussi di dati caratterizzati da problematiche specifiche, note per rendere l'adattamento dei modelli più complesso e presenti spesso in contesti reali. In particolare, il termine di errore è stato notevolmente aumentato, più di quanto fatto in precedenza, portando la qualità iniziale a livelli più simili a quelli osservati nel capitolo 2 (nei casi F2 ed F4, attorno a 0.2). I risultati sono riportati in appendice, in figura A.5. Le differenze iniziali sono in questo caso molto ridotte (R^2 predittivo: $ridge = 0.18$, $RF = 0.15$, $GB = 0.17$ e $NN = 0.18$.), e le differenze nella stabilità sono minime/assenti (in modo analogo al caso in figura A.4, in appendice, in cui la componente non osservata è stata aumentata in misura minore).

È stata inoltre condotta una simulazione provando a valutare l'effetto di variabili di disturbo, presenti nel dataset utilizzato dai modelli ma che non contribuiscono alla determinazione della variabile risposta. La scelta è giustificata dal fatto che lo stimatore *ridge* e la *rete neurale* non possono escludere

le variabili, risentendo possibilmente in misura maggiore del problema; il primo, basato su una penalizzazione dei coefficienti quadratica, meno adatta a farlo, e il secondo perché non vengono impiegate penalizzazioni durante la stima. Per rendere più complesso il problema le variabili esplicative sono state simulate in modo da presentare un'elevata correlazione a coppie (± 0.91), rendendo più complesso distinguere il contributo della singola variabile. Queste sono state generate partendo da una distribuzione $N(0, 1)$, perturbata tramite altrettante distribuzioni normali; le variabili risultanti, che hanno una base comune, presentano elevata correlazione. L'80% di queste hanno un effetto pari a zero, andando a costituire variabili di disturbo. I risultati sono riportati in appendice, nelle figure A.6 (solo correlazione elevata) e A.7 (correlazione elevata e variabili di disturbo), riportati entrambi i casi per confrontare i risultati.

In presenza di variabili di disturbo la *rete neurale* presenta il livello medio del terzo quartile maggiore, pur avendo qualità iniziale maggiore, ma la differenza è talmente contenuta, rispetto al range di valori osservato, da non sembrare rilevante (livello medio $NN = 4.11\%$, contro quello del più stabile, $ridge = 4.02\%$; differenza molto più piccola di quella tra *RF* e *ridge* nella prima simulazione, in figura 3.3).

Nel caso di sola correlazione estremamente elevata (figura A.6) le differenze nella stabilità dei modelli sono maggiori, e il *gradient boosting* presenta livelli più elevati del terzo quartile rispetto alla *foresta casuale*, pur avendo un livello iniziale maggiore (dell'1% di R^2 predittivo). Qui le differenze nella stabilità dei modelli sono maggiori (livello medio del terzo quartile $GB = 3.69\%$, contro $NN = 3.56\%$, il più stabile; una differenza ancora molto contenuta, ma quasi a parità di livelli iniziali), ma non sembrano essere sistematiche, in quanto non sono state osservate negli altri casi più simili (A.3, elevata correlazione, e A.7, stessa correlazione e variabili di disturbo, entrambe in appendice). Essendo le differenze difficili da interpretare in termini assoluti il confronto tra casi è reso molto complicato.

Questi risultati indicano come i livelli iniziali non prevedano in modo perfetto la stabilità delle prestazioni: ci sono alcuni casi in cui il modello peggiore è

il meno stabile (figure 3.3 e 3.5, A.2 e A.3, in appendice), altri in cui sono presenti grosse differenze iniziali ma non in stabilità (3.6 e A.5, la seconda in appendice) e altri ancora in cui le differenze iniziali sono ridotte ma le differenze nella stabilità dei modelli sono più elevate (A.6 e A.7, in appendice). In questi casi lineari, però, anche quando sono state osservate differenze nella stabilità, queste sono molto ridotte. Non ci sono inoltre modelli che tendono a presentare instabilità maggiori, in modo consistente.

Dataset con relazioni non lineari

Il passo successivo è valutare la variabilità dell'errore relativo dei modelli in presenza di relazioni più complesse tra le variabili esplicative e la variabile risposta, iniziando con la presenza di interazioni. La struttura del dataset, in questo esperimento di simulazione, è definita dalla seguente equazione:

$$Y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \cdots + \beta_{10} X_{10,ij} + \\ + \beta_{11} X_{7,ij} X_{10,ij} + \cdots + \beta_{20} X_{9,ij} X_{1,ij} + \varepsilon_{ij}$$

i = giorno; j = identificativo dell'osservazione nel giorno i

identica ai casi precedenti, ma con l'aggiunta di 10 termini di interazione (di primo ordine). I valori dei coefficienti sono stati generati casualmente dalla stessa distribuzione uniforme, e la scelta è stata casuale anche per i termini che sono andati a formare l'interazione. I dataset utilizzati dai modelli contengono le sole X (e non i prodotti delle singole coppie). La qualità iniziale dei modelli è riportata in figura 3.8, mentre la distribuzione del caso peggiore in figura 3.9.

A differenza dei casi precedenti la specifica struttura simulata ha portato ad osservare grosse differenze nella stabilità dei modelli; in questo frangente, in ordine, i modelli più stabili sono il *gradient boosting*, seguito da *rete neurale* (con una stabilità simile), *foresta casuale* e *ridge*, che non cattura bene relazioni di questo tipo. I modelli peggiori presentano minore stabilità, ma non solo: il livello medio del terzo quartile tende ad essere più alto, ma anche più

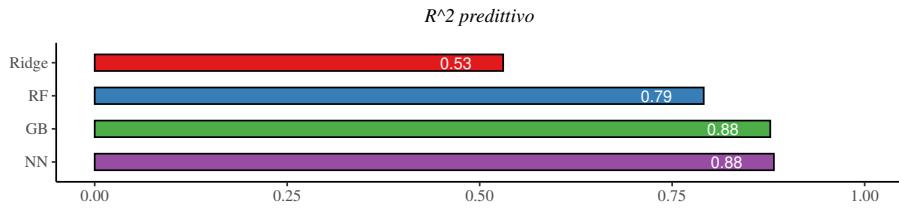


Figura 3.8: Prestazioni iniziali medie nelle diverse replicazioni, con effetti di interazione e 30 osservazioni al giorno. La qualità iniziale dello stimatore *ridge* è particolarmente bassa, in quanto non cattura bene relazioni di questo tipo.

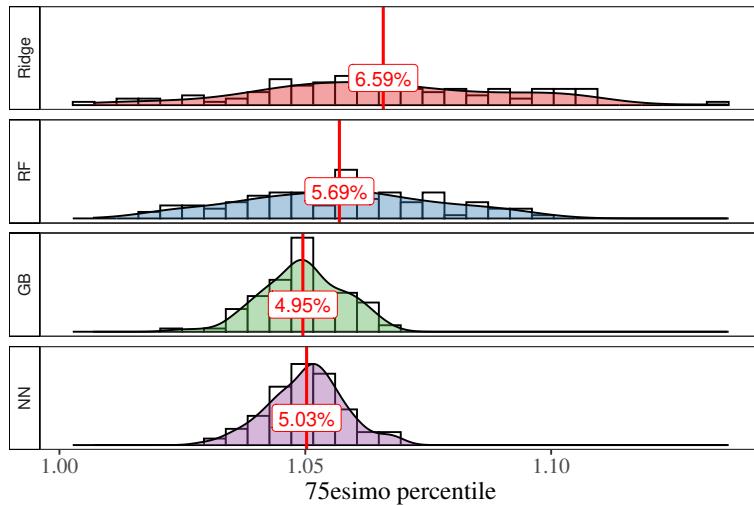


Figura 3.9: Distribuzione del livello medio del 75esimo percentile, con 30 osservazioni al giorno. Le differenze nei livelli medi sono maggiori rispetto ai casi precedenti, ma anche le differenze iniziali lo sono.

variabile. Dal grafico possiamo osservare che i modelli peggiori hanno presentato spesso livelli molto bassi di questa misura, indicando una stabilità molto inconsistente.

La questione delle prestazioni iniziali, non totalmente risolta in precedenza, deve essere approfondita. Lo stesso studio di simulazione è stato ripetuto, riducendo il numero di osservazioni, questa volta a 2 al giorno, per enfatizzare la perdita di prestazioni. I risultati, con annessa qualità iniziale, sono riportati nella Figura 3.10.

tati in figura 3.10. In questo caso, con differenze più elevate nelle prestazioni iniziali, le differenze nella stabilità sono maggiori: l'ordinamento dei modelli basato sulla stabilità coincide con quello basato sulle prestazioni iniziali.

Confermare che i modelli migliori tendono ad essere anche i più stabili sarebbe un risultato interessante, in quanto significa che la scelta del modello basata sulle prestazioni iniziali tende ad essere quella corretta.

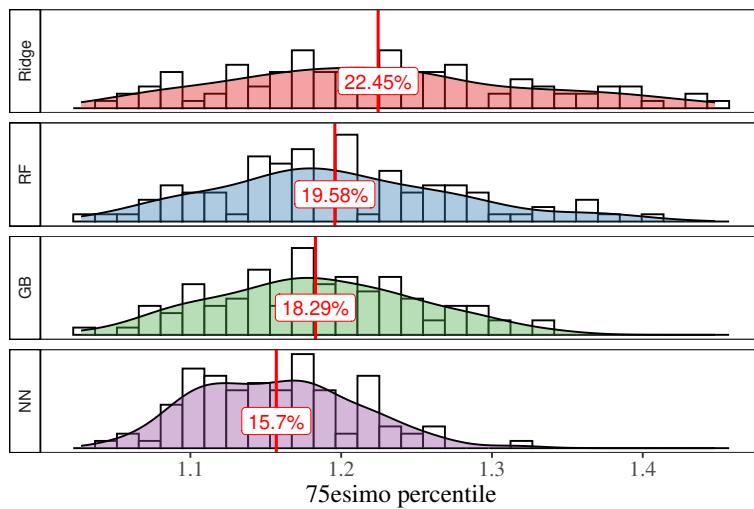


Figura 3.10: Distribuzione del livello medio del 75esimo percentile, con effetti di interazione e 2 osservazioni al giorno. Le differenze nei livelli medi sono maggiori rispetto ai casi precedenti, ma anche le differenze iniziali lo sono: *ridge* = 0.51, *RF* = 0.67, *GB* = 0.76 e *NN* = 0.82.

Nel caso di una relazione più complessa sono emerse maggiori differenze tra i modelli. Per questo motivo è stata esplorata una seconda situazione, caratterizzata da relazioni cubiche. I dataset simulati sono descritti tramite la seguente equazione:

$$Y_{ij} = \beta_0 + \beta_1 X_{1,ij}^3 + \cdots + \beta_{10} X_{10,ij}^3 + \varepsilon_{ij}$$

i = giorno; j = identificativo dell'osservazione nel giorno i

e costituiti da 30 osservazioni al giorno (nei dati utilizzati dai modelli compaiono le sole X , non i cubi). I risultati sono riportati nelle figure 3.11 e 3.12.

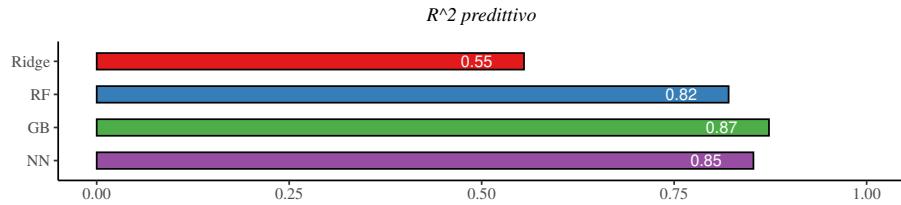


Figura 3.11: Prestazioni iniziali medie nelle diverse replicazioni, con relazioni cubiche. La qualità iniziale dello stimatore *ridge* è particolarmente bassa, in quanto non cattura bene relazioni di questo tipo. La qualità della rete neurale, in questo caso, è inferiore, ma può essere facilmente aumentata agli stessi livelli del *gradient boosting* utilizzando una struttura con più nodi latenti (appendice, figure A.8 e A.9). È stato mantenuto questo risultato in quanto mostra come migliori prestazioni iniziali non indichino necessariamente migliore stabilità, ma le conclusioni sono analoghe.

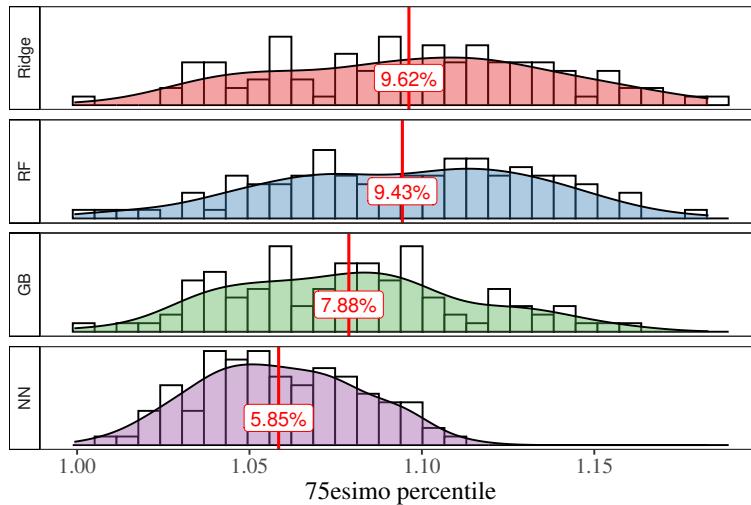


Figura 3.12: Distribuzione del livello medio del 75esimo percentile, nei dataset con relazioni cubiche.

Anche in questo caso le differenze sono più chiare, ma l'ordinamento basato

sulla stabilità non coincide con quello basato sulla qualità iniziale. La *rete neurale*, con livelli iniziali inferiori, presenta prestazioni meno variabili del *gradient boosting*, inizialmente migliore. Non solo, anche la *foresta casuale*, con una qualità molto superiore a quella del modello *ridge*, presenta una stabilità molto simile. Questa è una differenza importante rispetto a quanto osservato fino a questo punto, dovuta, stavolta, alle caratteristiche dei modelli.

Studiando i risultati delle singole simulazioni è stato possibile fare luce sulla causa. I risultati della replicazione 50, in cui il modello migliore (*GB*) non è il più stabile, sono i seguenti:

1. L' R^2 predittivo iniziale è pari a 0.82 per *RF*, 0.87 per *GB* e 0.85 per *NN*;
2. Il livello medio del terzo quartile è pari a 1.16 (16%) per *GB*, 1.15 (15%) per *RF* e 1.11 (11%) per *NN*.

Prendendo il tracciato dell'errore relativo dei 3 modelli, stimati su uno degli insiemi di stima, rappresentato in figura 3.13, possiamo osservare come la forma complessiva sia molto simile, ma la *rete neurale* non presenta gli stessi picchi, o li presenta in misura minore, condivisi invece da *RF* e *GB*.

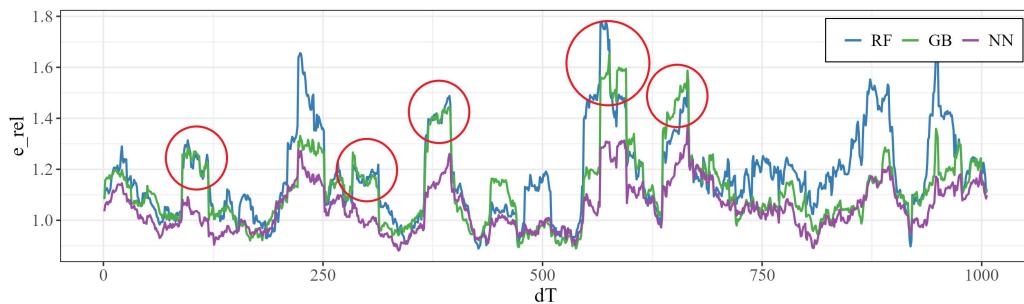


Figura 3.13: Tracciato dell'errore relativo di *foresta casuale*, *gradient boosting* e *rete neurale*, stimati sul 60esimo insieme di stima della 50esima replicazione (simili comportamenti anche per gli altri insiemi e simulazioni). In rosso sono indicati i picchi di errore relativo, condivisi dai primi due modelli, ma non dalla *rete*.

Sono questi picchi a causare la maggiore variabilità delle prestazioni, misurata tramite l'altezza del 75esimo percentile. La differenza è dovuta al fatto che i metodi basati sugli alberi estrapolano peggio, in questo caso, commettendo un errore molto più alto della *rete* sulle osservazioni collocate fuori dal range osservato durante l'adattamento, specialmente in relazione alle variabili con effetto elevato. In questo caso la differenza è chiara, a causa della relazione cubica: il valore estremo viene elevato al cubo e allontanato ulteriormente, andando a generare valori della variabile risposta troppo lontani da quelli osservati. La *rete*, che può generare valori non osservati, fa meglio, anche se non bene, su queste osservazioni, e ciò contiene l'errore relativo.

Situazioni dipendenti dal tempo

Nelle simulazioni precedenti le osservazioni sono tra loro indipendenti, ma ciò non è necessario per avere un processo che non evolve. In questa seconda sezione viene quindi aggiunta la componente temporale: è infatti comprensibile che, se questa non è presente, non è possibile osservare *degradazione temporale* in modo consistente.

Come già discusso, osservare dipendenza tra i valori passati e futuri della variabile risposta è normale, in un caso reale. Questa può essere causata dalla presenza di stagionalità, trattata nel capitolo 5, da variazioni nella distribuzione delle variabili esplicative, ovvero *data drift*, oppure dalla presenza di altri fattori latenti non osservati. Dalla simulazione con relazioni cubiche è inoltre emerso come una grossa differenza nella stabilità dei modelli si manifesta quando questi devono estrarre. In situazioni in cui la variabile risposta presenta cambi di livello ciò diventa più semplice, e la stabilità dei modelli ne risente.

L'intenzione quindi è quella di dare maggiore importanza alla componente temporale, e ciò viene fatto introducendo un processo autocorrelato nella componente non osservata dell'equazione che definisce la struttura del dataset, che diventa:

$$Y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \cdots + \beta_{10} X_{10,ij} + \varepsilon_{ij} + u_i$$

$i = \text{giorno}$; $j = \text{identificativo dell'osservazione nel giorno } i$

A differenza del caso precedente compare u_i , il processo autocorrelato, che varia assieme al giorno, ma non all'interno, portando ad osservare dipendenza tra le medie giornaliere.

Sono state condotte diverse simulazioni, al variare di questo termine. Le prime utilizzano dei processi stazionari, una scelta scontata, considerando che l'alternativa comporta senza dubbio *degradazione temporale*. Sono state fatte alcune prove, con relazioni lineari e non tra passato e futuro, mettendo a disposizione più o meno informazione per catturare la dipendenza (inserendo i primi ritardi della media giornaliera, della variabile risposta, nei dataset). I risultati e i dettagli delle prove condotte sono riportati in appendice A (da pagina 160).

I modelli non tendono a presentare *degradazione temporale*, quindi l'informazione più rilevante è legata alla variabilità delle prestazioni. Le differenze osservate sono totalmente prevedibili, considerando le caratteristiche dei metodi: le logiche basate sugli alberi presentano minore stabilità. *Rete neurale* e stimatore *ridge* presentano invece risultati simili: è stato infatti osservato, in Xu et al. (2020), che le previsioni del primo modello, con un'architettura uguale a quella utilizzata in questo lavoro (funzioni di attivazione *reLu*), convergono ad una funzione lineare fuori dal range osservato durante la stima. Delle differenze sono state osservate quando la dipendenza inserita nel flusso non è lineare (ma il processo rimane stazionario): in quei casi la *rete neurale* ha mostrato maggiore stabilità. La maggiore efficacia della *rete* è tuttavia legata allo specifico processo utilizzato (in questo caso, la convergenza delle previsioni fuori range ad una funzione lineare era particolarmente adatta) ma in casi differenti potrebbe non esserlo altrettanto. Considerando inoltre che dipende dalla specifica architettura utilizzata, sarebbe necessario studiare la stabilità della *rete* ogni volta che questa cambia, rendendo eccessivamente complicato lo studio esaustivo delle proprietà di stabilità di tale logica.

La logica matematica quindi, in questo caso, determina differenze, preve-

dibili però sulla base delle caratteristiche dei modelli: la logica matematica non ha delle proprietà di stabilità a sé che la rendono più adatta, ma è la qualità della sua estrapolazione sullo specifico dataset a determinare la differenza.

Anche in questi casi la qualità iniziale dei modelli ha giocato un ruolo importante: la similitudine nel comportamento a medio/lungo termine di *gradient boosting* e *foresta casuale* è stata parzialmente coperta dal diverso punto di partenza. Osservando gli errori relativi infatti la stabilità risulta differente, ma ciò non emerge dall'*MSE*, rendendo difficile distinguere differenze in stabilità legate ai metodi specifici senza considerarne la loro qualità (in appendice, da pagina 160, un approfondimento).

Se i processi inseriti nel flusso non sono stazionari (in media) si osserva, per forza, *degradazione temporale* nelle prestazioni di alcuni dei modelli considerati, considerando la presenza di trend. Considerando i risultati precedenti l'incognita più grande è la differenza tra *rete neurale* e *ridge*: quale dei due modelli presenta minore decadimento delle prestazioni?

Se il trend non è lineare allora la differenza è poco prevedibile (dipende dalla tipologia di trend e dall'entità), ed entrambi invecchieranno in misura importante. Se il trend fosse però lineare o catturabile tramite relazioni lineari con il passato, un caso in cui entrambi possono avere buone prestazioni, quanto può differire, e in che modo, l'evoluzione temporale delle prestazioni? È stata condotta una simulazione in cui il processo autocorrelato è di tipo ARIMA(0,1,0), che dimostra come, in modo abbastanza prevedibile, le prestazioni della *rete neurale* siano molto più imprevedibili nel medio/lungo termine (i risultati sono riportati in figura A.18, in appendice).

Riassunto dei risultati

La tabella 3.1 contiene un breve riassunto delle simulazioni condotte in questo capitolo.

Simulazione	Risultati	Figure
Effetti lineari	I modelli presentano una stabilità delle prestazioni simile, leggermente inferiore per la <i>foresta casuale</i> , con una qualità iniziale inferiore.	3.1-3.4
Effetti lineari - Differenti matrici Σ	Non ci sono differenze rilevanti nei risultati rispetto al caso iniziale.	A.2, A.3
Effetti lineari - Aumento varianza della componente non osservata	Le differenze nella stabilità dei modelli si sono notevolmente ridotte.	A.4, A.5
Effetti lineari - Diverso numero di osservazioni	La variazione del numero di osservazione ha comportato, complessivamente per tutti i modelli, diversi livelli del terzo quartile. Le differenze tra modelli, sugli stessi dati, sono però contenute, e in linea con quanto osservato nel caso iniziale.	3.5, 3.6
Effetti lineari - 4 variabili	Ridurre il numero di variabili esplicative ha portato ad un aumento della qualità iniziale della <i>foresta casuale</i> , con conseguente riduzione del divario in stabilità dagli altri modelli.	3.7
Effetti lineari - Variabili di disturbo	La <i>NN</i> presenta una stabilità leggermente più bassa pur avendo la qualità iniziale migliore, mentre con solo correlazione estremamente elevata il <i>GB</i> è risultato il meno stabile, pur avendo una buona qualità iniziale. Queste differenze non sembrano però essere sistematiche.	A.6, A.7
Effetti di interazione	Le differenze nella stabilità dei modelli sono maggiori, ma sono associate a diversi livelli di qualità iniziale.	3.8, 3.9

Effetti di interazione - Meno osservazioni	Con l'aumento della differenza iniziale tra i modelli è aumentata la differenza nella loro stabilità. I modelli peggiori sono anche i meno stabili.	3.10
Effetti cubici	I modelli basati sugli alberi hanno una stabilità inferiore, in quanto risentono maggiormente del dover prevedere su valori fuori dal range osservato durante la stima.	3.11-3.13
Processo AR(2)	I modelli basati sugli alberi hanno una stabilità inferiore, in quanto risentono maggiormente del dover prevedere su valori fuori dal range osservato durante la stima. La qualità iniziale inferiore della <i>RF</i> , in questo caso, ha nascosto la similitudine con il <i>GB</i> in termini di comportamento a medio/lungo termine. <i>Ridge</i> e <i>NN</i> hanno una stabilità analoga.	A.10- A.15
Processo ARMA(2,2)	Non ci sono differenze rilevanti nei risultati rispetto al caso dell'AR(2).	A.16
Processo SETAR	La <i>NN</i> ha una stabilità maggiore rispetto allo stimatore <i>ridge</i> .	A.17
Processo ARIMA(0,1,0)	La <i>degradazione temporale</i> è molto inferiore per lo stimatore <i>ridge</i> rispetto alla <i>NN</i> .	A.18

Tabella 3.1: Riassunto delle simulazioni condotte nel capitolo 3. Nella colonna “figure” sono indicate le figure che riportano i risultati. I numeri che iniziano con A indicano che le figure sono in appendice.

Dalle simulazioni effettuate in questo capitolo possiamo affermare quanto segue:

1. Nei casi più semplici, caratterizzati da una relazione lineare, è stato osservato come le differenze nella stabilità dei modelli siano, al più, molto ridotte, anche in presenza di differenze in qualità iniziali ampie (figura 3.5). Quantificare queste differenze è tuttavia impossibile, in quanto da una simulazione ad un'altra i livelli complessivi dei quartili dei modelli cambiano, rendendo i risultati poco confrontabili.
2. All'aumento della complessità della relazione sono state osservate maggiori differenze nella stabilità dei modelli, che sono però spesso accom-

pagnate da altrettanta distanza nei livelli di qualità iniziali. In questi casi i modelli inizialmente migliori si sono rivelati essere anche i più stabili (figure 3.9 e 3.10), e ciò è compatibile anche con i risultati osservati per la *foresta casuale* nei casi lineari iniziali (figure 3.3, 3.5 e A.3, in appendice). La dipendenza dei risultati dalla qualità iniziale è confermata dal fatto che ridurre la distanza iniziale porta a ridurre le differenze nella stabilità dei modelli (figura 3.7).

3. Il caso più importante in cui la precedente associazione, tra qualità iniziale e stabilità, non è stata osservata è in presenza di relazioni cubiche (figura 3.12), in cui la *rete neurale* ha presentato stabilità sensibilmente migliore del *gradient boosting*, pur avendo qualità iniziale minore. Ciò è dovuto all'estrapolazione: in questo caso i modelli basati sugli alberi commettono un errore maggiore sulle osservazioni fuori dal range osservato durante la stima, aumentando le fluttuazioni di errore relativo.
4. Oltre alla situazione appena citata ci sono stati dei casi in cui i modelli migliori hanno presentato maggiore instabilità. Nei casi riportati nelle figure 3.9 e A.7 (in appendice), la *rete neurale* ha presentato un livello medio del terzo quartile maggiore di quello di altri modelli, pur presentando una qualità iniziale migliore, oppure nel caso in figura A.6 (in appendice), dove è stato il *gradient boosting* a presentare minore stabilità (pur avendo buona qualità iniziale): le differenze sono però limitate a questi casi, e talmente contenute, rispetto al range osservato, da non sembrare rilevanti. Nella simulazione riportata in figura 3.2 è infatti stato osservato come il comportamento dei modelli sia praticamente identico, anche quando vengono osservate differenze di maggiore entità nel terzo quartile.
Essendo le differenze (in termini assoluti) poco confrontabili, da un caso all'altro, diventa molto difficile imputarle a qualcosa di strutturale.
5. Ci sono stati inoltre casi in cui le differenze iniziali sono state ampie, ma le differenze nella stabilità dei modelli minime o assenti (figure A.4 e A.5, in appendice, in cui la componente di errore ha importanza maggiore, e figura 3.6).

Le differenze più importanti nella stabilità sono state comunque osservate quando il divario iniziale era ampio, e quando la qualità iniziale dei modelli è simile la stabilità tende ad essere molto simile. Nei casi in cui sono state osservate delle differenze nella variabilità dell'errore, con livelli iniziali simili tra modelli, le differenze sono state minime, rispetto al range di valori possibili per il livello del terzo quartile.

In un flusso di dati in cui le regole non evolvono sembra dunque adeguato basare la scelta del modello sulla qualità iniziale: il miglior modello, che raggiunge prestazioni più elevate, spesso è anche il più stabile (ad eccezione dell'unico caso menzionato sopra, e di eccezioni minime). Le migliori prestazioni sono inoltre probabilmente sufficienti a compensare eventuali piccole maggiori instabilità (che sono differenze nei casi peggiori, non nei casi mediani).

Non solo, nei flussi di dati in cui è stata data importanza alla componente temporale è stato evidenziato come i modelli che estrapolano correttamente siano da preferire nei flussi di dati in cui il livello non rimane costante (figure A.12, A.16, A.17 e A.18, in appendice).

Anche considerando tutto ciò, non è garantito che il modello peggiore presenti prestazioni meno prevedibili. Le differenze osservate sono sui livelli medi, che tendono ad essere più alti per i modelli peggiori; tuttavia, sullo stesso dataset, non è raro osservare casi in cui i modelli più scarsi sono i più stabili. Prendendo ad esempio il caso degli effetti di interazione, con 2 osservazioni al giorno (figura 3.10), osserviamo, dalla figura 3.14, che il 31% delle volte la *foresta*, pur avendo qualità iniziale molto più bassa della *rete neurale*, presenta un livello del terzo quartile inferiore.

Non solo, osservando l'istogramma del caso con relazioni di interazione e un numero maggiore di osservazioni (figura 3.9), è possibile osservare che *ridge* e *foresta casuale*, nonostante presentino livelli iniziali molto inferiori rispetto agli altri modelli, spesso hanno prodotto livelli del terzo quartile inferiori, indicando come tendenzialmente siano più instabili, ma anche che la stabilità stessa sia più inconsistente, quando la qualità iniziale si riduce. La qualità

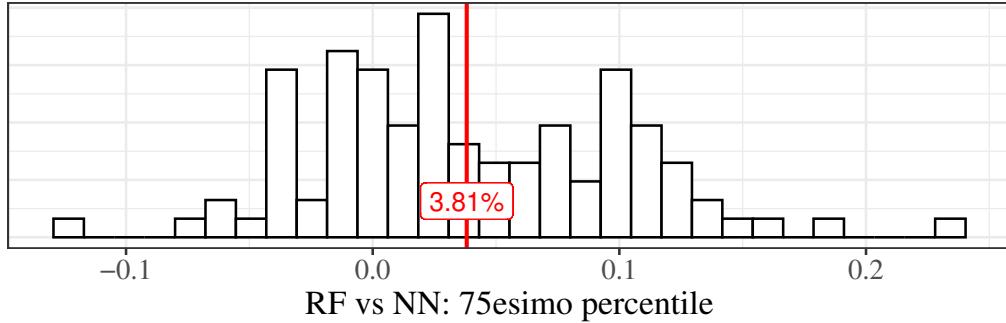


Figura 3.14: Distribuzione della differenza tra le medie del terzo quartile di *foresta casuale* e di *rete neurale*. Le qualità iniziali medie dei modelli sono pari a 0.67 per *RF* e 0.82 per *NN*. Nel 31% dei casi la differenza è inferiore a 0, indicando che la *foresta* è più stabile, in quei casi. Questo grafico fa riferimento alla simulazione in cui sono presenti effetti di interazione e le osservazioni giornaliere sono pari a 2.

iniziale non prevede quindi perfettamente la stabilità del modello, ma entra sempre in gioco nella lettura dei risultati (alcuni casi interessanti sono quelli riportati in appendice, da pagina 160: la diversa qualità iniziale ha nascosto similitudini a medio/lungo termine; figure A.10 - A.16).

Prevedere la stabilità delle prestazioni sulla base della qualità iniziale, delle caratteristiche dei dati o della tipologia di modello non porta a risultati esatti. Scegliere il modello sulla base delle prestazioni iniziali rimane quindi l'opzione più valida, con una preferenza per i modelli con migliore extrapolazione; criterio di scelta difficile da sfruttare, in quanto si tratta, per definizione, di scegliere il modello migliore su valori non osservati. È quindi necessaria una buona conoscenza del dominio di applicazione.

Nei casi in cui non è presente il problema dell'extrapolazione, non sono state evidenziate differenze nella stabilità dei modelli che possano essere attribuite direttamente alla loro logica matematica; nelle simulazioni condotte, i risultati legati alla variabilità delle prestazioni non sono abbastanza consistenti da indicare differenze di fondo.

Capitolo 4

Simulazioni: *concept drift*

In questo capitolo viene affrontato il problema della stabilità dei modelli in presenza di *concept drift*, ovvero nel caso in cui le regole del processo, attraverso il quale sono generati i dati, evolvono. Facendo riferimento alla sezione 1.2.2, in cui sono state introdotte le diverse situazioni oggetto di studio, le simulazioni presentate di seguito contengono i risultati appartenenti alla seconda categoria.

Come discusso nella sezione 1.2.1, il *concept drift* consiste nell’evoluzione del processo che interessa due componenti distinte della distribuzione congiunta di X ed Y , che può essere scomposta nella maniera seguente:

$$P(X, Y) = P(Y|X) \cdot P(X)$$

Cambiamenti che interessano la distribuzione delle variabili esplicative prendono il nome di *data drift*, mentre evoluzioni nella distribuzione condizionata di Y vengono chiamate *real concept drift*, o più spesso, *concept drift*. Il contributo che l’approccio simulativo può dare nel comprendere come il *data drift* influisca sulla stabilità dei modelli è molto ridotto, ma nel secondo caso il grado di approfondimento è maggiore.

Data drift

L’evoluzione della distribuzione delle variabili esplicative è un problema che influisce negativamente sulle prestazioni dei modelli. Ciò comporta infatti, con il passare del tempo, lo spostamento della regione delle esplicative su cui il modello deve operare, con effetti poco prevedibili sulle prestazioni.

La maggior parte del problema è associato all’estrapolazione: lo spostamento della distribuzione porta i modelli, con il tempo, ad operare su valori delle covariate mai osservati, con conseguenze poco prevedibili in un contesto reale (ma molto prevedibili in questo contesto simulativo). Qualsiasi valutazione possa essere fatta ha quindi poca utilità: in un caso reale, per definizione, il modello più stabile sarebbe quello con prestazioni migliori su casi non osservati, e quindi poco prevedibile.

Le conseguenze dell’estrapolazione, in questo contesto di simulazione, sono però prevedibili: tra i modelli considerati *ridge* e *rete neurale* presenterebbero prestazioni stabili nel tempo solo se la relazione fosse lineare (il primo più del secondo), ma non in altri; sarebbe possibile valutare quale risponde meglio in ciascuna situazione, ma ciò dipende dall’entità dello spostamento e dalla relazione, e l’informazione sarebbe poco utilizzabile in casi reali.

Il *data drift* non è un problema per la sola estrapolazione: questo fenomeno può infatti comportare la comparsa, nel flusso di dati, di osservazioni in regioni poco esplorate in fase di stima, con il conseguente aumento dell’errore del modello; oppure in regioni in cui l’adattamento è generalmente peggiore, con le stesse conseguenze. Diversi modelli, adattati allo stesso processo, possono inoltre esibire diversi livelli di qualità dell’adattamento nelle stesse regioni dello spazio, rendendo complesso prevedere l’evoluzione temporale delle prestazioni.

Certamente, tra i modelli considerati, tre (tutti tranne lo stimatore *ridge*) possono adattarsi adeguatamente ad un vasto insieme di strutture, e prevedere il perché uno di questi (regolato e stimato a dovere) dovrebbe presentare un adattamento minore in determinate aree è difficile.

Nel capitolo 3 è stato osservato come diversi modelli abbiano bisogno di un numero differente di osservazioni per produrre un adattamento soddisfacente: questi modelli soffriranno maggiormente se la distribuzione delle X si sposta verso regioni più sparse. In particolare, la *foresta casuale* ha sempre risentito maggiormente della scarsità di osservazioni, seguita dal *gradient boosting* e dalla *rete neurale*. Anche quest'ultimo modello è noto per aver bisogno di un elevato numero di osservazioni per produrre dei risultati soddisfacenti, quindi la sua stabilità non è garantita. Alcune simulazioni condotte, riportate in appendice B (da pagina 173), confermano proprio questo: i modelli con un adattamento peggiore in un area dello spazio soffrono maggiormente quando la distribuzione si sposta in tale zona, e i modelli che presentano maggiore *degradazione temporale* sono quelli indicati (in caso di effetti di interazione e relazioni cubiche).

Considerando tutto questo, in un contesto reale la scelta ottimale è sicuramente basata sulla qualità iniziale: il modello con un livello iniziale più alto tenderà a presentare l'adattamento complessivo migliore e a soffrire meno di *data drift* (questo non considera l'estrapolazione, problematica per la quale è necessaria una buona conoscenza del dominio di applicazione).

Le conseguenze del *data drift* dovrebbero quindi essere studiate caso per caso, nel contesto reale di impiego dei modelli, tramite l'utilizzo del “test” di *degradazione temporale* e conoscenza specifica del dominio.

Real concept drift

In questa sezione viene studiato l'effetto del *real concept drift* (da qui semplicemente *concept drift*) sulle prestazioni a medio/lungo termine dei modelli. Questo termine descrive i casi in cui la relazione tra le variabili esplicative e la variabile risposta evolve nel tempo, ed è un fenomeno noto per influire negativamente sulle prestazioni dei modelli, che adattati ad un processo differente non riescono a mantenere la loro qualità senza essere aggiornati.

L'effetto negativo di questo fenomeno è noto; di fatto, esistono diversi studi su tale problema. L'articolo di Lima et al. (2022) (che riassume, tra gli altri, Liu et al., 2019; Martínez-Rego et al., 2011; Budiman et al., 2016) ha costituito un valido punto di partenza per capire come i *concept drift* influiscono sulla qualità delle previsioni nell'ambito della regressione, e come simularli. Il focus di questi studi non è di fatto quello di valutare se l'errore di previsione aumenta o meno, ma confrontare diversi metodi per individuare il *drift* e aggiornare i modelli, confrontandone l'efficacia.

La riduzione delle prestazioni e della stabilità dei modelli in presenza di *concept drift* non viene quindi messa in discussione: un modello statico non può non risentirne; è tuttavia ragionevole pensare (e come suggerito, del resto, in Vela et al., 2022) che i modelli possano presentare differenze nel comportamento di medio/lungo periodo in presenza di tale problematica.

La *degradazione temporale* è stata quindi studiata in presenza di un flusso di dati che evolve in modo graduale e prevedibile nel corso dell'intero intervallo temporale simulato, in modo da esporre i modelli al cambiamento; evoluzioni di questo tipo hanno maggiori possibilità di produrre differenze nell'invecchiamento.

In quest'ottica, sono stati condotti degli esperimenti di simulazione in cui il flusso di dati viene caratterizzato da tre tipologie di *concept drift*:

1. *Concept drift graduale*, in cui un primo sistema di regole viene gradualmente sostituito da un secondo.

-
2. *Concept drift incrementale*, in cui un primo sistema di regole si trasforma progressivamente in un secondo.
 3. *Concept drift ricorrente*, un'evoluzione più irregolare, in questo caso, delle regole che legano variabili esplicative e variabile risposta.

Concept drift graduale

I primi flussi di dati simulati incorporano un'evoluzione progressiva della relazione tra le variabili esplicative, X , e la variabile risposta, Y . Per farlo, il valore dei coefficienti viene modificato nel tempo, un approccio tipico alla simulazione del problema (Lima et al., 2022; Liu et al., 2019; Martínez-Rego et al., 2011; Budiman et al., 2016). Nel caso del *drift graduale* ciò avviene tramite la progressiva sostituzione di un primo insieme di coefficienti con un secondo.

La struttura del flusso di dati, in questo primo caso, è descritta tramite la seguente equazione:

$$Y_{ij} = \beta_{i,0} + \beta_{i,1}X_{1,ij} + \cdots + \beta_{i,10}X_{10,ij} + \varepsilon_{ij}$$

i = giorno; j = identificativo dell'osservazione nel giorno i

A differenza dei casi precedenti il valore dei coefficienti cambia nel tempo. Il procedimento utilizzato per simulare un *drift graduale* è il seguente:

1. È stato scelto casualmente il primo insieme di coefficienti, che va a costituire la regola iniziale del flusso di dati.
2. È stato generato casualmente il secondo insieme di coefficienti, che gradualmente va a sostituirsi al primo. In questo esperimento solo metà dei coefficienti presentano un valore che evolve nel tempo, mentre gli altri rimangono costanti. Questo permette di avere una qualità iniziale dei modelli accettabile ma un impatto visibile della problematica.

3. La probabilità che un'osservazione venga generata utilizzando il primo insieme di coefficienti, all'inizio del flusso di dati, è pari a 0.8, nei restanti casi viene generata utilizzando il secondo. Questa probabilità si riduce in modo lineare nel corso del tempo, arrivando a 0 all'ultimo giorno simulato: ciò significa che, alla fine del dataset, tutte le osservazioni seguono la seconda relazione, che è andata a sostituire la prima.

Volendo indagare come i diversi approcci alla modellazione rispondono al problema, il valore dei coefficienti non dovrebbe essere rilevante.

La scelta di utilizzare inizialmente la seconda regola nel 20% dei casi permette ai modelli di osservarla durante la stima, ma come sarà osservato ciò non è particolarmente importante: la differenza tra un *drift* di questo tipo e uno *incrementale*, con gli stessi coefficienti, è minima.

Il valore dei coefficienti viene determinato una sola volta, e mantenuto in tutte le replicazioni dell'esperimento simulativo. Prendendo come riferimento però la singola osservazione (Y_{ij}, X_{ij}) , ad ogni replicazione viene deciso casualmente l'insieme di coefficienti da utilizzare per generarla, sulla base della probabilità nello specifico tratto di flusso (che varia linearmente nel corso del tempo). Un esempio di dataset simulato utilizzando questa metodologia è rappresentato nella figura 4.1.

Come possiamo vedere dal grafico, identificare il *concept drift* tramite l'osservazione del comportamento della risposta è difficile, in quanto si mantiene regolare per l'intero periodo. Questo può spiegare i casi in cui la qualità dei modelli decade nel tempo quando il comportamento della variabile risposta, almeno visivamente, non evolve nel tempo. Quindi spiega quei casi in cui non ci sono segni di un cambiamento nella distribuzione sottostante, ma le prestazioni dei modelli peggiorano (come osservato nei dataset reali studiati nel capitolo 2).

I dataset utilizzati dai modelli contengono le sole X .

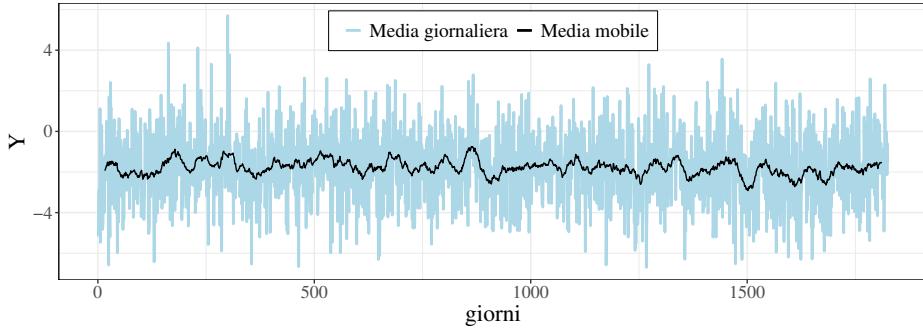


Figura 4.1: Primo dataset che include *concept drift graduale*. A differenza del *data drift*, è difficile capire da un’analisi grafica che il processo evolve nel tempo. Ciò può spiegare la *degradazione temporale* dei modelli nei casi in cui apparentemente i segni di un’evoluzione non ci sono.

I risultati rilevanti sono riportati nelle figure 4.2 (la combinazione dei grafici di *AI Aging*) e 4.3 (andamenti degli *MSE* mediani e qualità iniziale dei modelli).

Come era atteso, tutti i modelli soffrono di *degradazione temporale* in presenza di *drift graduale*, anche se la *foresta casuale* presenta un’inclinazione dell’errore relativo mediano molto inferiore rispetto agli altri (figura 4.2): invecchia quindi in modo differente. A distanza massima dal momento della stima, mentre l’*MSE* della *foresta casuale* aumenta del 20%, quello degli altri modelli aumenta del 30%. Non solo dal punto di vista relativo, ma anche assoluto: osservando i tracciati degli *MSE(dT)* mediani, combinati nuovamente per mostrare una banda di variabilità (figura 4.3, con un riepilogo della procedura; la combinazione risultante è chiamata “andamento dell’*MSE* mediano”, per mancanza di un termine più semplice), osserviamo che il modello, a distanza massima, batte gli altri, e lo preferiamo dunque a lungo andare.

Per valutare se questo comportamento è dovuto alle caratteristiche del modello, oppure al grado di adattamento iniziale (l’unica altra differenza osservata), sono state condotte ulteriori simulazioni:

1. Il numero di variabili esplicative è stato ridotto, in modo da permet-

tere alla *foresta casuale* un miglior adattamento iniziale (risultati in appendice, figura B.5); ridurre totalmente il divario non è però stato possibile.

2. Il numero di osservazioni è stato ridotto, in modo da valutare il comportamento degli altri modelli al ridursi della loro qualità iniziale (figura 4.4).
3. Il numero di variabili esplicative è stato aumentato (a 60, e il numero di osservazioni è stato ridotto a 10 al giorno), allo stesso scopo del caso precedente (figura 4.5).

L'utilizzo di questi approcci, per ridurre la qualità degli altri modelli, è dovuto al fatto che il “test” di *degradazione temporale* prevede di utilizzare dei modelli con una parametrizzazione ottimale: in questo contesto fare in modo che *rete neurale* e *gradient boosting* avessero prestazioni inferiori alla *foresta casuale* non è stato possibile. Le figure relative ai risultati delle simulazioni (4.4 e 4.5) riportano gli andamenti degli *MSE* mediani, in quanto gli errori relativi sono meno efficaci nel mostrare differenze sostanziali (nel modo in cui l'errore del modello cambia). Il problema viene chiarito nella sezione dedicata al riassunto dei risultati.

Da quanto è possibile osservare, l'effetto del *concept drift* non sembra essere necessariamente legato ad uno specifico modello, ma dipende dalla qualità iniziale raggiunta. Questa considerazione è basata sui risultati riportati nelle figure 4.3 e 4.4. In quest'ultimo caso possiamo osservare, in particolare, che all'aumento della distanza l'*MSE* dei modelli *aumenta maggiormente per i modelli con una qualità iniziale maggiore*. I modelli che perdono qualità in misura minore sono invece quelli con un adattamento peggiore.

Un comportamento analogo può essere osservato nel caso iniziale (figura 4.3), in cui il *gradient boosting* ha presentato una riduzione dell'errore leggermente minore dello stimatore *ridge*, con cui condivide la quasi totalità della banda a distanza massima.

Chiaramente questo metodo per combinare i tracciati degli $MSE(dT)$ mediani non permette di fare considerazioni precise, ma del resto non esiste una metodologia totalmente soddisfacente per farlo.

Aumentando il numero di variabili si osservano comportamenti simili: al più, due modelli con errore iniziale simile presentano aumenti simili (figura 4.5). Tutto considerato, un modello con qualità iniziale inferiore non ha presentato aumenti maggiori dell'errore.

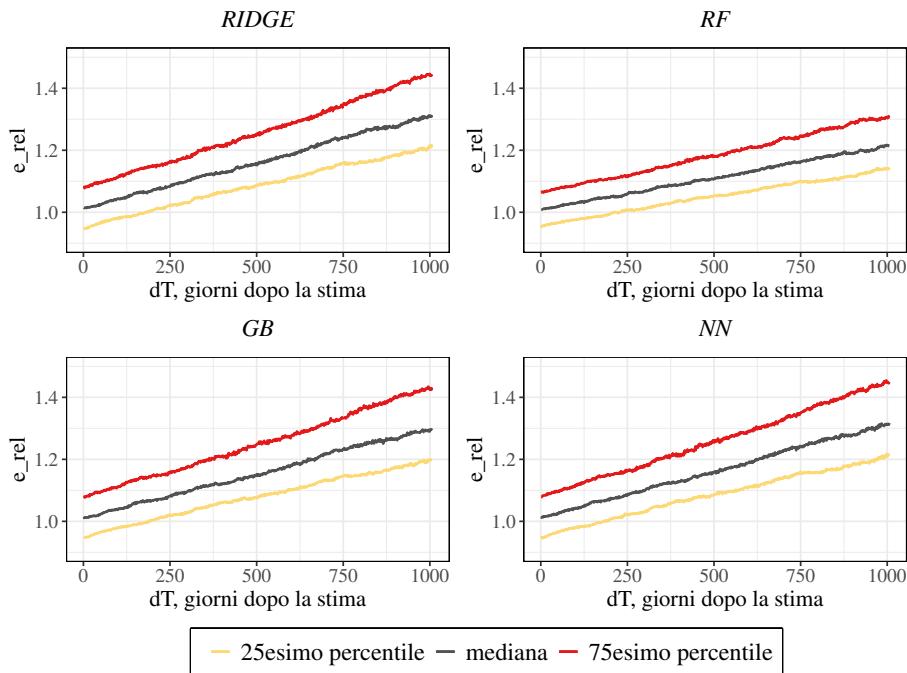


Figura 4.2: Combinazione dei grafici di *AI Aging*, che mostrano come tutti i modelli presentino *degradazione temporale*. L'inclinazione dell'errore relativo mediano, l'indicatore più importante di *degradazione temporale* in questo caso, è molto inferiore per la *foresta casuale*.

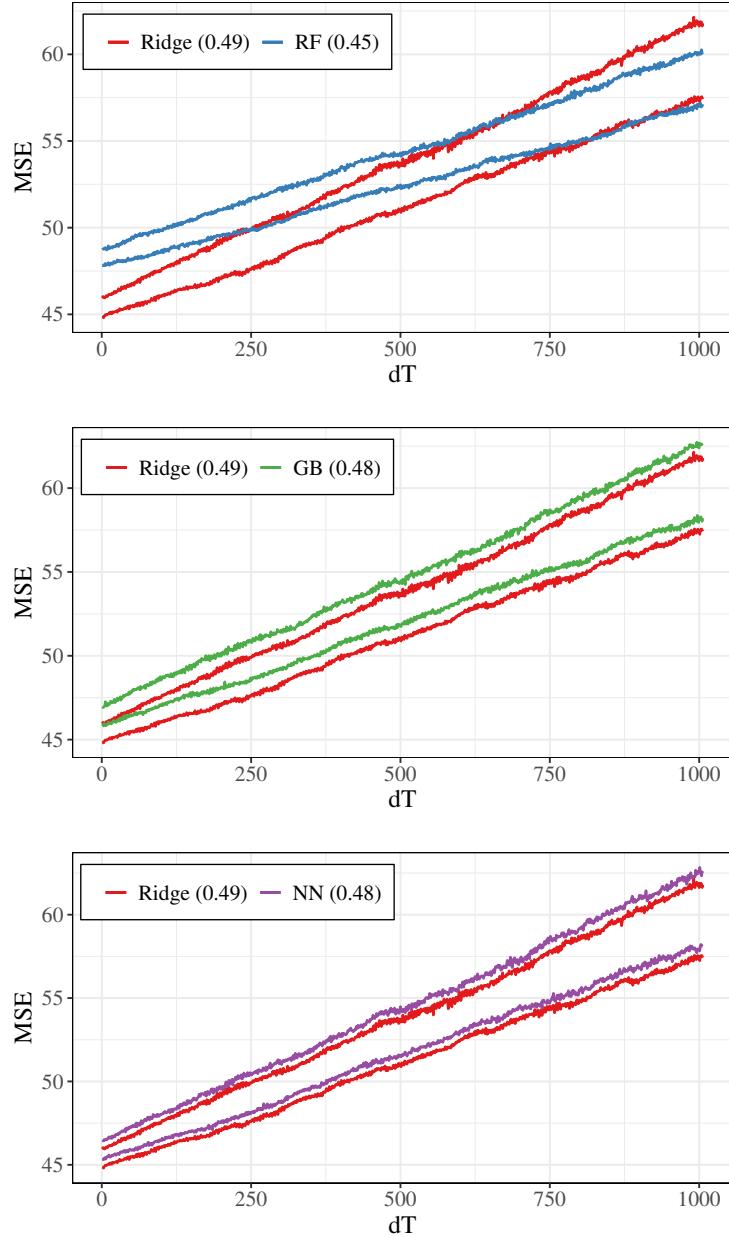


Figura 4.3: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, nel caso iniziale con 10 variabili esplicative e 30 osservazioni al giorno. Sulla singola replicazione è stato calcolato l' $MSE(dT)$ mediano, in modo analogo a quanto fatto per l'errore relativo mediano. I tracciati ottenuti nelle singole replicazioni sono poi stati combinati, nuovamente, calcolando il 25esimo e 75esimo percentile, ad ogni dT , in modo da ottenere una banda di variabilità. L'andamento dell'errore (della banda) di ciascun modello è quindi stato confrontato con quello del modello inizialmente migliore, in quanto presenta la maggiore *degradazione*. Nelle legende è riportato il valore di R^2 *predittivo* iniziale medio.

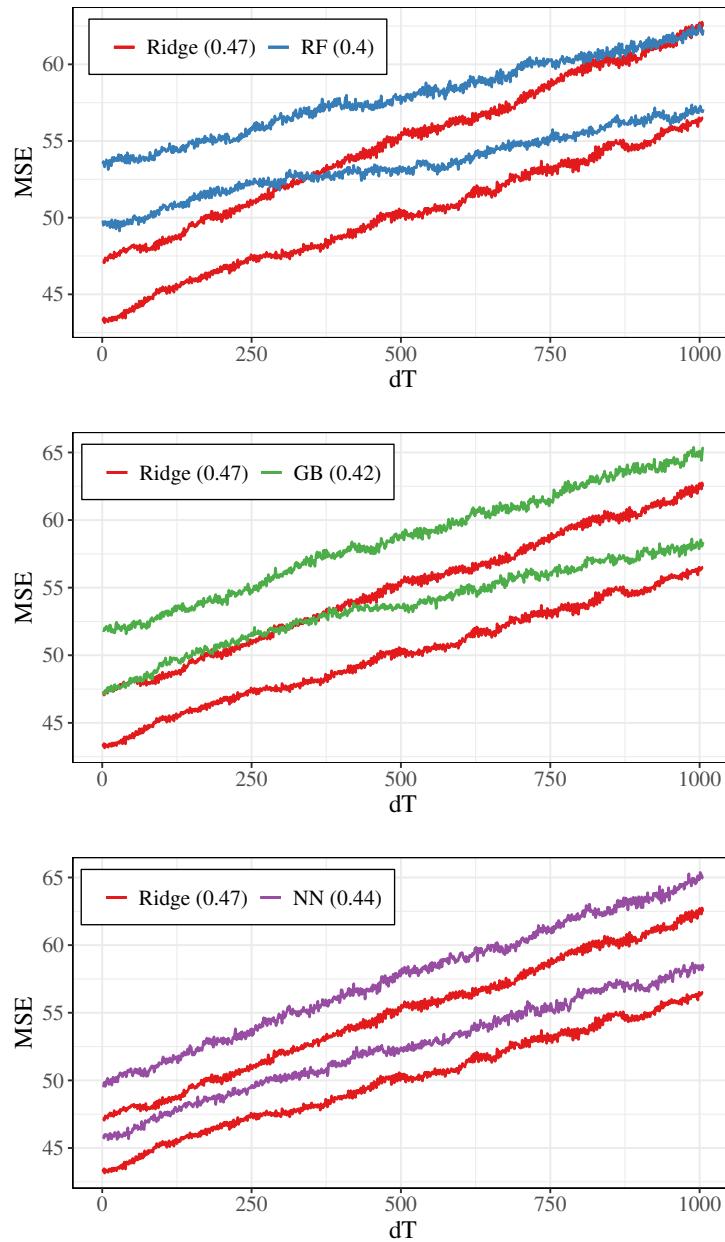


Figura 4.4: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, nel caso iniziale ma con 2 osservazioni al giorno (i coefficienti sono gli stessi). La *foresta casuale* presenta le stesse peculiarità, il minor decadimento della qualità, ma ora anche il *gradient boosting*, e in misura minore la *rete neurale*, la cui banda di variabilità della mediana si sovrappone quasi completamente a quella del *ridge*, a distanza massima.

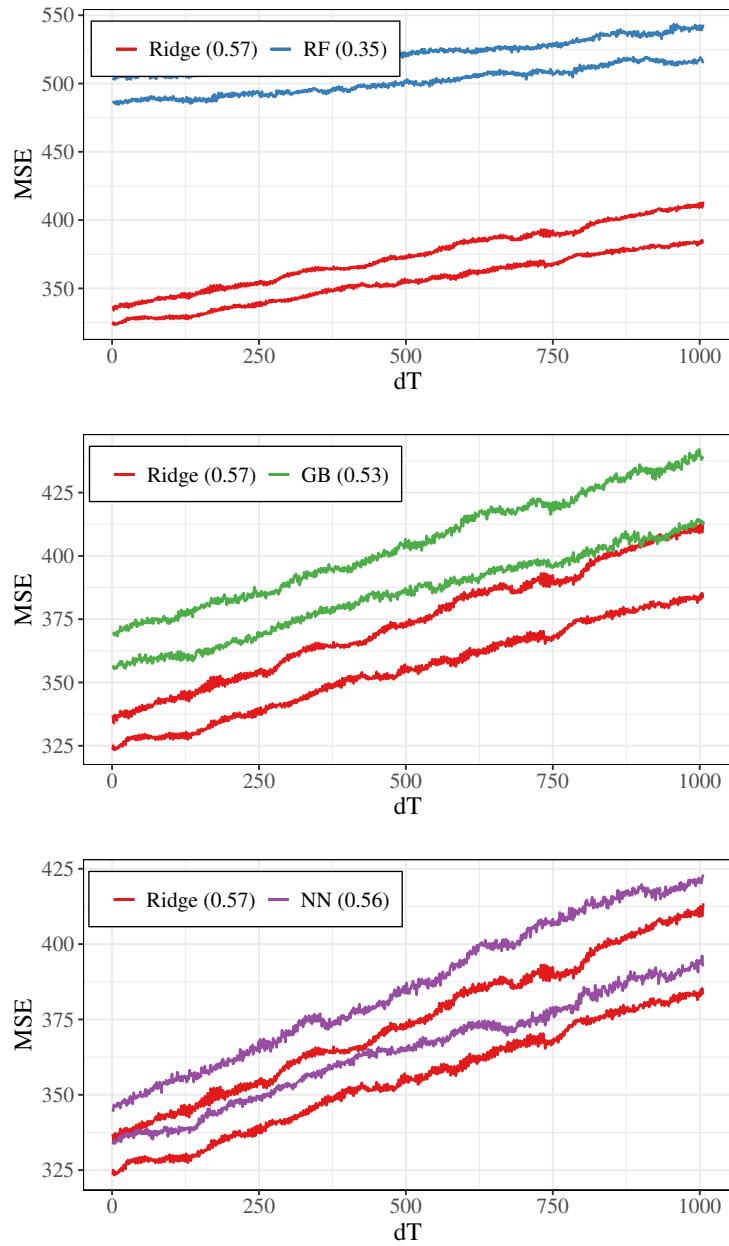


Figura 4.5: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, con 60 variabili e 10 osservazioni al giorno. Le conclusioni sono simili a quelle già riportate: il modello peggiore, la *foresta casuale*, presenta un aumento minore dell'errore, e un comportamento analogo può essere osservato per il *gradient boosting*. Al più, due modelli inizialmente simili presentano un aumento simile dell'errore. 50 repliche.

Considerato quanto osservato, il *concept drift* sembra influire maggiormente sui modelli con qualità iniziale maggiore. Questo è intuibile da come opera: un cambio di regole (in questo caso, del valore dei coefficienti) impatta maggiormente i modelli con un adattamento maggiore al primo sistema di regole.

In questo caso non è stato possibile ridurre totalmente la distanza che separa la *foresta casuale* dagli altri modelli, o invertire l'ordinamento dei modelli sulla base dei livelli iniziali: fare ciò avrebbe permesso di confermare come il minor decadimento della qualità non sia una peculiarità specifica del modello, ma che dipenda solo dal minor adattamento.

È stato condotto un ulteriore esperimento, per verificare ciò: la simulazione iniziale (in figura 4.3) è stata ripetuta per il *gradient boosting* e la *rete neurale*. Nel primo caso è stato limitato notevolmente il numero di alberi utilizzati, inducendo un peggiore adattamento (figura 4.6), mentre nel secondo è stato ridotto il numero di epoch, rispetto a quelle ottimali, per ottenere un risultato analogo (figura 4.7).

Queste riduzioni di qualità sono artificiali, e i risultati non sono necessariamente confrontabili con quelli ottenuti per gli altri modelli. Ciononostante questi evidenziano come l'effetto del *drift* dipenda dall'adattamento iniziale: in entrambi i casi la sua riduzione ha comportato un aumento minore dell'errore. Nel caso del *gradient boosting*, in particolare, sono stati prodotti dei modelli con adattamento inferiore a quello delle *foreste casuali* e che subiscono in misura minore l'effetto del *drift*.

La *rete neurale* ha invece complessivamente una qualità iniziale minore, ma risente in misura maggiore del *drift* rispetto agli altri modelli (con un valore di R^2 predittivo iniziale più alto, come il *GB* con meno alberi).

Dai risultati di queste simulazioni le diverse logiche matematiche non hanno esibito differenze specifiche (quando i modelli sono stati regolati a dovere): a prescindere dalla logica considerata, è stato osservato come un adattamento inizialmente peggiore è associato ad un decadimento minore della qualità (figure 4.3, 4.4 e 4.5). Similmente, modelli con qualità iniziale simile hanno presentato un decadimento simile. Ridurre l'adattamento dei modelli con

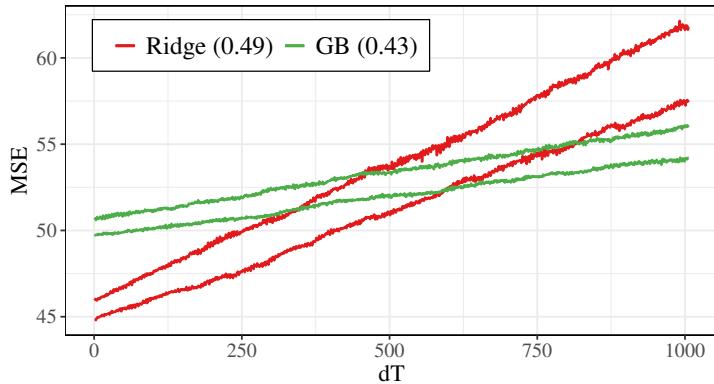


Figura 4.6: Andamento dell’ MSE mediano (GB), ottenuto riducendo il numero di alberi utilizzati. Questo è confrontato con l’andamento dell’ MSE mediano dello stimatore *ridge* (ricavato dalla simulazione iniziale). L’effetto del *concept drift* si è ridotto notevolmente, e il GB vince sul *ridge* a distanza massima.

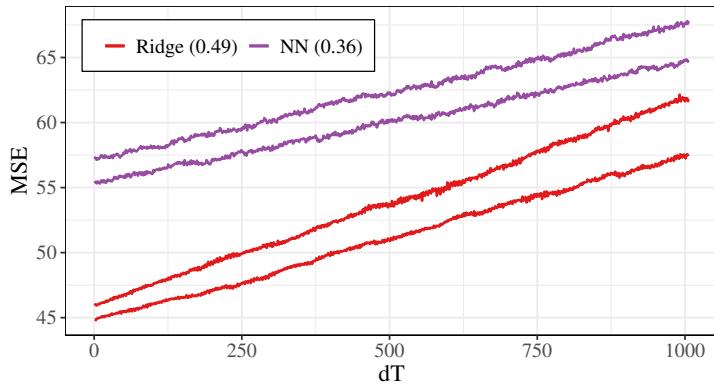


Figura 4.7: Andamento dell’ MSE mediano (NN), ottenuto usando un terzo delle epoche necessarie per l’adattamento ottimale; questo è confrontato con l’andamento dell’ MSE mediano dello stimatore *ridge* (ricavato dalla simulazione iniziale). L’effetto del *concept drift* sulle prestazioni si è ridotto, ma la NN non presenta risultati migliori di quelli ottenuti riducendo la qualità del GB , pur presentando un minor adattamento. I due approcci alla riduzione dell’adattamento non sono però necessariamente confrontabili.

qualità iniziale maggiore ha invece evidenziato come l'effetto del *drift* si sia ridotto, relativamente ad un adattamento ottimale. Certamente, a prescindere dalla logica matematica su cui il modello si fonda, la qualità iniziale è strettamente legata all'evoluzione temporale dell'errore.

Infine, relativamente al *concept drift graduale*, sono state condotte delle prove con relazioni più complesse della lineare, le stesse utilizzate nel capitolo 3 (cubiche ed effetti di interazione). Le simulazioni condotte, le cui figure sono riportate in appendice (le conclusioni sono analoghe), possono essere riassunte di seguito:

1. In un primo caso è stato simulato *concept drift graduale*, che coinvolge metà dei coefficienti, in una situazione caratterizzata dalla presenza di effetti di interazione e 30 osservazioni al giorno. Le specifiche del *drift* sono le stesse di quelle utilizzate nel caso lineare. Gli andamenti degli *MSE* mediani sono riportati nella figura B.6. In aggiunta, la simulazione è stata ripetuta riducendo il numero di osservazioni a 5 al giorno (figura B.7), per ridurre le qualità iniziali.
2. In un secondo caso è stato simulato *concept drift graduale*, con le stesse specifiche, in una situazione caratterizzata da relazioni cubiche (risultati in figura B.8) e, come già fatto, il numero di osservazioni è stato successivamente ridotto (risultati in figura B.9).

I risultati legati a queste simulazioni sono interessanti per diversi motivi:

1. È stato osservato un caso (figura B.8, in appendice, relazioni cubiche e 30 osservazioni al giorno) in cui il modello migliore è il *gradient boosting*, seguito dalla *rete neurale*, e le prestazioni del primo decadono in misura maggiore di quelle del secondo, confermando quanto osservato fino a questo punto (l'associazione tra qualità iniziale e peggioramento delle prestazioni).
2. È stato osservato un caso (figura B.7, in appendice, effetti di interazione e 5 osservazioni al giorno) in cui *foresta casuale* e *gradient boo-*

sting hanno una qualità iniziale identica (questa volta regolati e stimati correttamente), e il decadimento delle prestazioni è lo stesso.

3. In tutti i casi, eccetto con relazioni cubiche e un minor numero di osservazioni (figura B.9, in appendice), l'adattamento ha costituito un buon predittore del decadimento delle prestazioni. In questo caso però due modelli, *RF* e *GB* (figura 4.8), con la stessa qualità iniziale, hanno presentato due evoluzioni temporali delle prestazioni molto differenti: le prestazioni del primo si sono mantenute meglio di quelle del secondo. In questo caso anche le *reti* hanno presentato un comportamento differente, legato alla regolazione: pur presentando un adattamento minore le prestazioni si sono ridotte allo stesso modo che per modelli migliori (in particolare, come per il *GB*). Ciò è un caso simile a quanto sperimentato in figura 4.7: il modello ha presentato difficoltà a convergere a causa del minor numero di osservazioni, e l'effetto del *drift* è diventato meno prevedibile sulla base del livello iniziale.

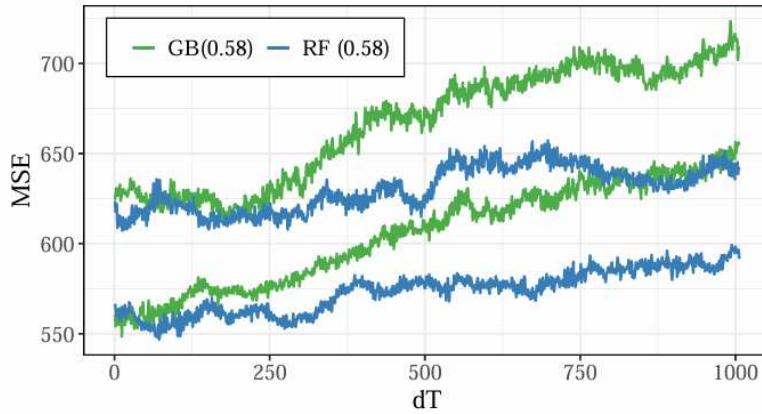


Figura 4.8: Andamenti degli *MSE* mediani di *foresta casuale* e *gradient boosting*, che mostrano come il decadimento delle prestazioni del primo sia minore, nonostante lo stesso livello iniziale.

In modo analogo a quanto osservato nel caso lineare i modelli che hanno sperimentato un aumento maggiore dell'*MSE* sono quelli con qualità maggiore, con la singola eccezione riportata, con l'aggiunta che sono stati osservati casi

in cui l'ordinamento basato sui livelli iniziali non è lo stesso di quanto visto nel caso lineare.

La determinante maggiore dell'effetto del *concept drift* sembra essere dunque la qualità iniziale, almeno per questa tipologia, ma non possono essere escluse differenze basate sulla logica matematica (vista l'eccezione osservata). Le considerazioni finali sono nella sezione dedicata al riassunto dei risultati.

Concept drift incrementale

Questa seconda tipologia costituisce un'evoluzione progressiva della relazione tra le X e la Y , ma stavolta non si tratta di una graduale sostituzione. I valori dei coefficienti invece si modificano nel tempo, spostandosi verso un secondo insieme di valori. Le simulazioni condotte con questa tipologia di *drift* utilizzano gli stessi insiemi di coefficienti delle precedenti (con *drift graduale*), per poter confrontare i risultati: la simulazione con effetti lineari, ad esempio, con quattro variabili esplicative e 30 osservazioni al giorno, utilizza gli stessi valori che nel caso di *drift graduale*, e ciò avviene per ogni simulazione per cui c'è una corrispondenza con quelle della sezione precedente. Le osservazioni stesse ($X_{i,j}$) però cambiano.

Per simulare un *drift incrementale*, dato un coefficiente che evolve, il valore iniziale del primo giorno viene modificato progressivamente, in modo lineare, verso il secondo valore, che raggiunge nell'ultimo giorno simulato. A differenza del *drift* precedente è quindi un'operazione deterministica. I risultati principali della prima simulazione (30 osservazioni al giorno, effetti lineari) sono riportati nelle figure 4.9 e 4.10. A differenza della tipologia (di *drift*) precedente, questa è più imprevedibile: un modello, stimato su un anno di dati, non ha modo di conoscere il valore finale dei coefficienti in corrispondenza del tratto finale del flusso; questo è molto simile, nella sostanza, a cambiare improvvisamente i valori.

I risultati ottenuti sono sostanzialmente identici ai precedenti: la *foresta casuale*, il modello che presenta un adattamento iniziale peggiore, presenta

un minor decadimento della qualità, sia in termini assoluti che relativi. Gli altri modelli, che presentano una qualità iniziale sostanzialmente identica, presentano anche un comportamento identico.

Di nuovo, le prestazioni sono state rivalutate variando come fatto in precedenza il numero di variabili esplicative (ridotto a 4, in figura 4.12, e portato a 60, in appendice, in figura B.11) e il numero di osservazioni (portato a 2 al giorno, in figura 4.11).

Due casi notevoli sono riportati nelle figure 4.11 (relativamente al caso con 2 osservazioni al giorno) e 4.12 (relativamente al caso con 4 variabili esplicative). Nel primo caso, è chiaro come il *concept drift* agisca maggiormente sui modelli con adattamento iniziale più scarso.

Nel secondo caso (figura 4.12), osserviamo come questa regola possa essere infranta: lo stimatore *ridge*, con una qualità iniziale maggiore, risente meno del *drift* rispetto a *gradient boosting* e *rete*, che presentano un adattamento minore. Questo non è necessariamente dovuto al caso: una seconda simulazione, condotta con coefficienti differenti, conferma quanto osservato (in appendice, figura B.10).

Non solo, dalla simulazione corrispondente con *concept drift graduale* (i coefficienti sono gli stessi), si intravede un comportamento simile, anche se è più complesso da distinguere (in appendice, figura B.5).

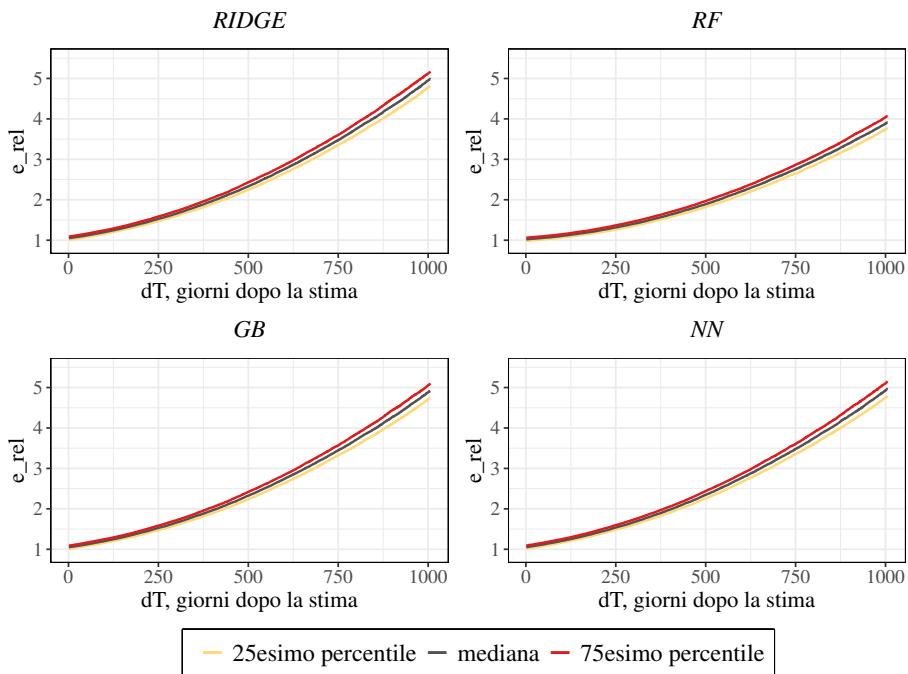


Figura 4.9: Combinazione dei grafici di *AI Aging* nel caso di *drift incrementale*, con effetti lineari e 30 osservazioni al giorno, da cui si può osservare come tutti i modelli presentino segni di *degradazione temporale*, anche se l'inclinazione dei quartili per la *foresta casuale* è inferiore. La qualità iniziale dei modelli è pari a 0.76 per *ridge*, *RF* e *GB*, e 0.72 per la *RF*.

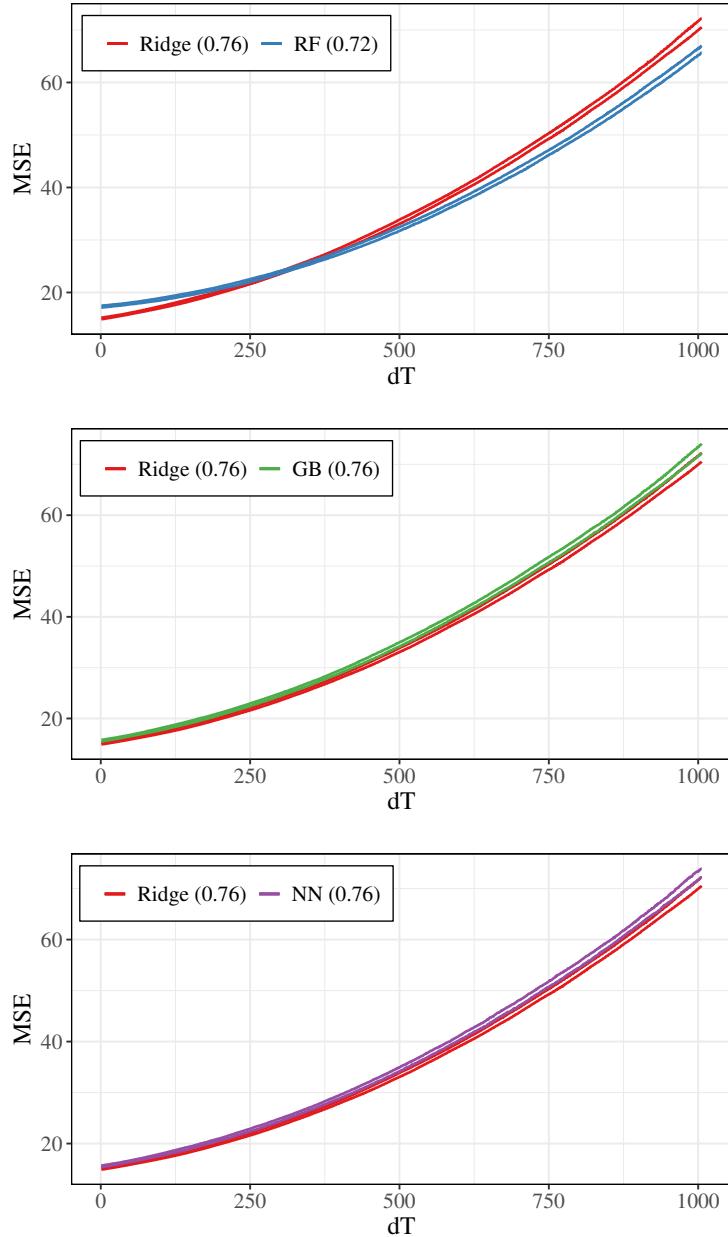


Figura 4.10: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con relazioni lineari e 30 osservazioni al giorno. La differenza negli andamenti è chiara, la *RF* mantiene meglio la qualità nel tempo. Negli altri casi gli andamenti sono invece appaiati.

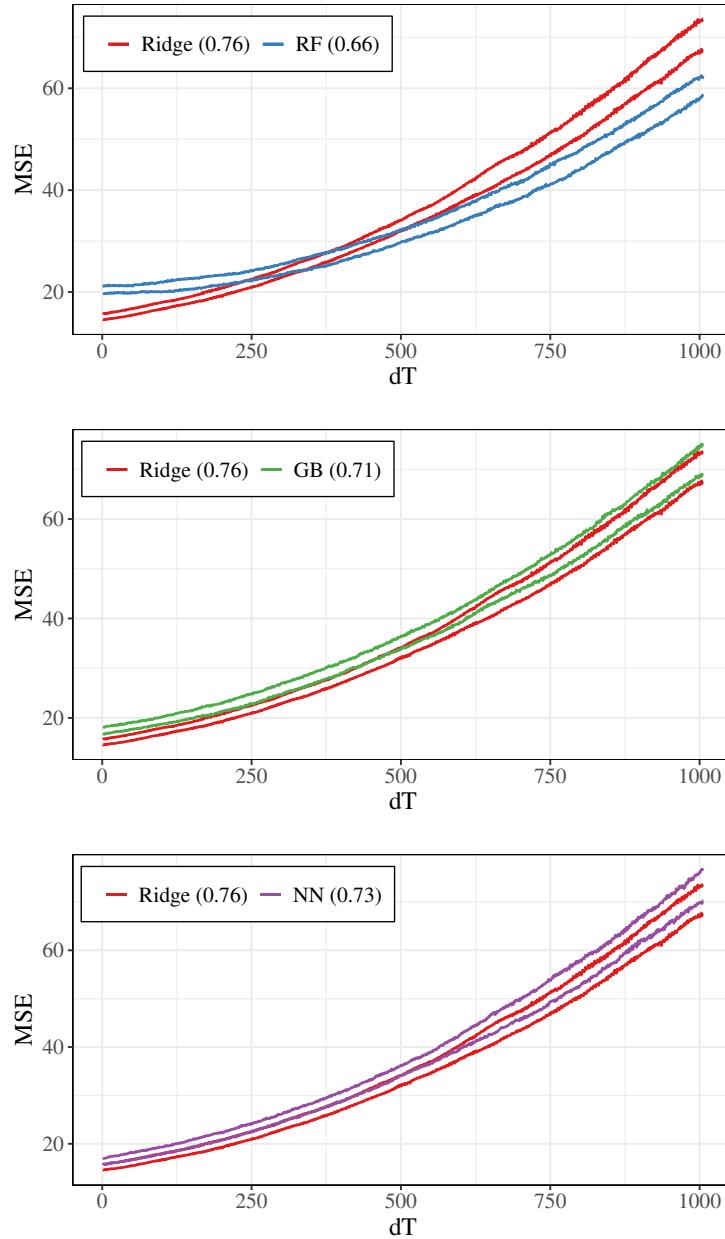


Figura 4.11: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, nel caso di 2 osservazioni al giorno (gli stessi coefficienti della prima simulazione, figura 4.10). Le considerazione fatte nel caso del *drift graduale* si presentano di nuovo: i modelli con una qualità iniziale inferiore presentano un'inclinazione minore.

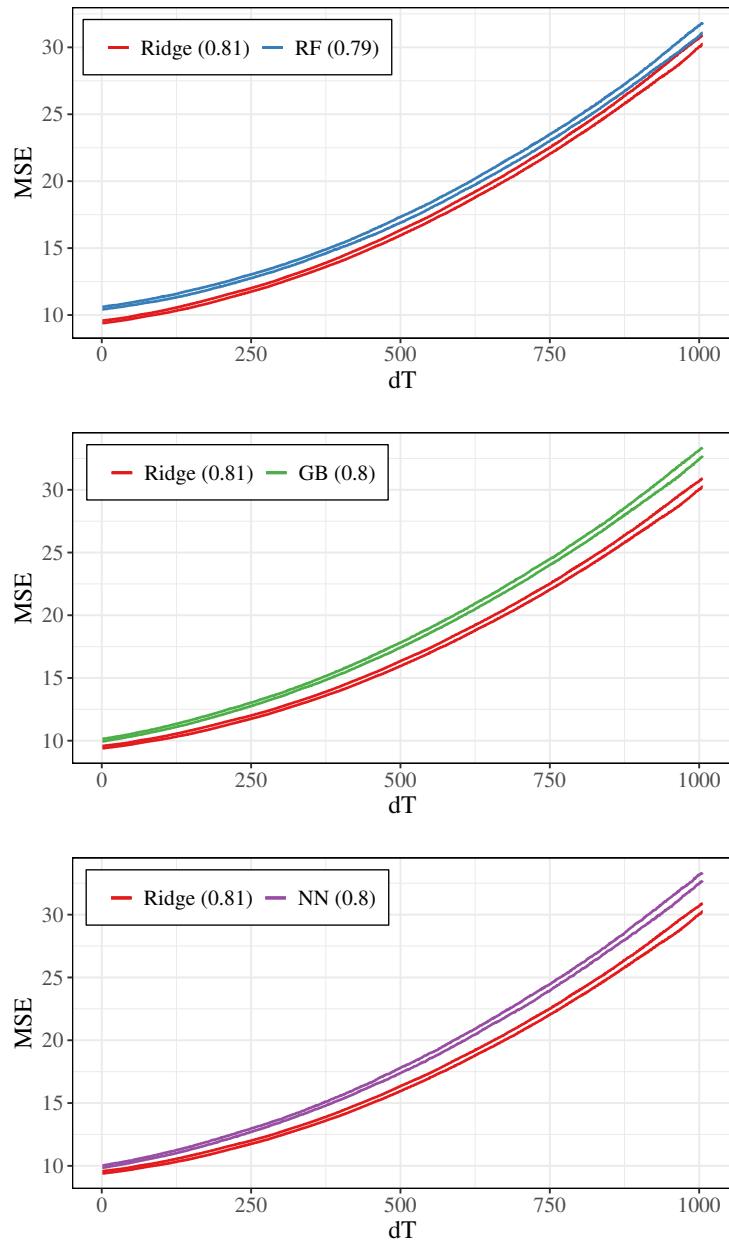


Figura 4.12: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale* e 4 variabili esplicative. In questo caso lo stimatore *ridge*, con una qualità iniziale maggiore, presenta un aumento minore dell'errore rispetto a *GB* e *NN*, con una qualità iniziale minore.

Come nel caso precedente sono state condotte alcune simulazioni aggiuntive:

1. La quantità di alberi del *gradient boosting* è stata ridotta, ottenendo un risultato analogo al caso del *drift graduale* (in appendice, figura B.12), così come sono state ridotte le epoche utilizzate per la stima delle *reti neurali* (gli stessi risultati ottenuti che per il *drift graduale*, in appendice, figura B.13).
2. I modelli sono stati applicati in flussi di dati caratterizzati da *drift incrementale* e relazioni non lineari, valutando gli stessi quattro casi. I risultati sono riportati in appendice (figure B.14 e B.15 per gli effetti di interazione, B.16 e B.17 per le relazioni cubiche).

I risultati ottenuti portano quindi a conclusioni molto simili a quelle già riportate:

1. La qualità iniziale ha un effetto che non può essere ignorato sull'evoluzione temporale della qualità dei modelli. Nei casi di effetti di interazione l'adattamento iniziale ha costituito un buon predittore del peggioramento delle prestazioni in presenza di *drift*.
2. Con relazioni cubiche la *foresta casuale* ha presentato un aumento dell'*MSE* minore rispetto a quanto atteso sulla base dell' R^2 . In particolare, con una qualità iniziale maggiore, ha risentito meno della problematica rispetto alle *reti neurali* (figura B.17, in appendice, 5 osservazioni al giorno). Ciò non è stato osservato, però, con un numero di osservazioni maggiore.

Sulla base di queste simulazioni è possibile dunque fare alcune considerazioni aggiuntive:

1. La qualità iniziale, nella maggior parte delle situazioni, ha fornito una buona indicazione del peggioramento delle prestazioni dei modelli.
2. Sono state osservate alcune eccezioni, legate allo stimatore *ridge*, alla *foresta casuale* e alla *rete neurale*. Come per la simulazione in figura 4.12 (effetti lineari e 4 variabili), in cui la prova è stata ripetuta per

confermare l’eccezione associata allo stimatore *ridge*, ciò è stato fatto per le simulazioni con effetti cubici. I risultati, riportati nelle figure B.18 e B.19, in appendice, non mostrano le stesse peculiarità nel comportamento della *foresta casuale*. Questo non è stato ripetuto nel caso di *drift graduale*, in quanto i risultati delle simulazioni con i due tipi di *drift* diversi hanno sempre concordato.

Infine, come ulteriore conferma della relazione tra qualità iniziale ed effetto del *drift*, il numero di alberi utilizzati dal *gradient boosting* è stato ridotto nei flussi di dati caratterizzati da relazioni non lineari. I risultati, riportati in appendice nelle figure B.20 e B.21, portano alle stesse conclusioni che nel caso lineare (per le relazioni cubiche sono stati utilizzati i dataset della seconda simulazione, in quanto non contengono l’eccezione della *foresta casuale*). Di nuovo, ciò è stato fatto solo nel caso di *drift incrementale*.

Concept drift ricorrente

In Vela et al. (2022), tra le possibili cause di *degradazione temporale*, viene indicata l’evoluzione dell’importanza delle variabili esplicative e del loro contributo allo spiegare la variabile risposta. Questo è stato mostrato tramite una rappresentazione grafica, che è stata riportata in figura 4.13.

In questo caso l’evoluzione del processo è molto irregolare, e avanzare nel flusso di dati non ci garantisce di trovare un processo più distante dalla condizione iniziale. Un caso di questo tipo è simile ad un *concept drift ricorrente*, presentato nella sezione 1.2.1: una regola viene temporaneamente sostituita da un’altra, per poi presentarsi nuovamente. La stabilità dei modelli è quindi stata messa alla prova in una situazione analoga.

Simulare degli effetti che si muovono nel tempo, in modo simile a quanto presentato dagli autori, può essere fatto in diversi modi. I dataset simulati, in questo caso, seguono una struttura analoga ai precedenti (effetti lineari e 10 variabili esplicative) con la differenza principale il modo in cui gli effetti evolvono. Per simulare un movimento analogo a quanto osservato in figura

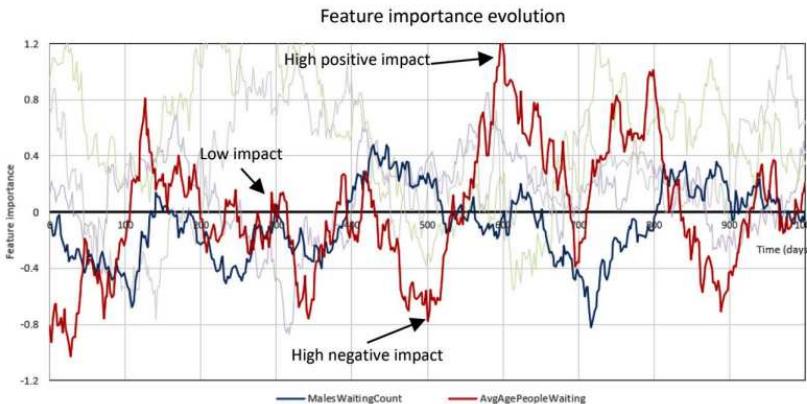


Figura 4.13: Grafico tratto da Vela et al. (2022). Gli autori cercano di mostrare come l'effetto di due variabili, stimato tramite stimatore *ridge*, cambia assieme all'insieme di stima, che si sposta poco a poco nel flusso di dati. Le due variabili presentano un effetto che cambia nel corso del tempo, avendo un diverso impatto a seconda del momento a cui il modello viene stimato. In questo esempio l'ampiezza degli insiemi di stima è di soli 3 mesi, differente dall'intero anno utilizzato negli altri casi.

4.13 sono stati utilizzati dei processi ARMA(2,2) (una scelta euristica) per i valori dei coefficienti, compatibili con i movimenti descritti sopra (esempio in appendice, pagina 197, dove sono brevemente discusse le scelte compiute).

Ad ogni replicazione dell'esperimento simulativo l'andamento degli effetti viene generato nuovamente. La prima simulazione è condotta con le solite specifiche: 30 osservazioni al giorno e ampiezza dell'insieme di stima pari ad un anno; i risultati sono riportati nelle figure 4.14 - 4.16.

Le differenze tra i modelli sono minime, complessivamente nessuno mostra una tendenza maggiore alla perdita di prestazioni, come dimostrano la combinazione dei grafici di *AI Aging* (figura 4.14) e la distribuzione delle pendenze dell'errore relativo mediano (figura 4.16); anche se, rispetto al caso con effetti lineari iniziale (del capitolo 3), è molto più semplice osservare pendenze estreme, come dimostra il range di valori osservati. Detto questo, vediamo come la distribuzione nella coda sia molto simile, indicando come la facilità con cui i modelli invecchiano lo sia.

Il livello medio del terzo quartile dell'errore è più basso per la *foresta casuale*: come osservato finora, ciò può essere associato alla minore qualità iniziale. In un contesto come questo è naturale che l'effetto del *concept drift* si manifesti sulla variabilità: i tracciati di $MSE(dT)$ dei singoli modelli (stimati su singoli sottoinsiemi di dati) non presentano tutti lo stesso andamento crescente, ma forme più irregolari. Detto questo, differenze in media del 3-4% nel caso peggiore non sono sufficienti a compensare le minori prestazioni iniziali, più basse rispetto agli altri modelli di circa il 7%.

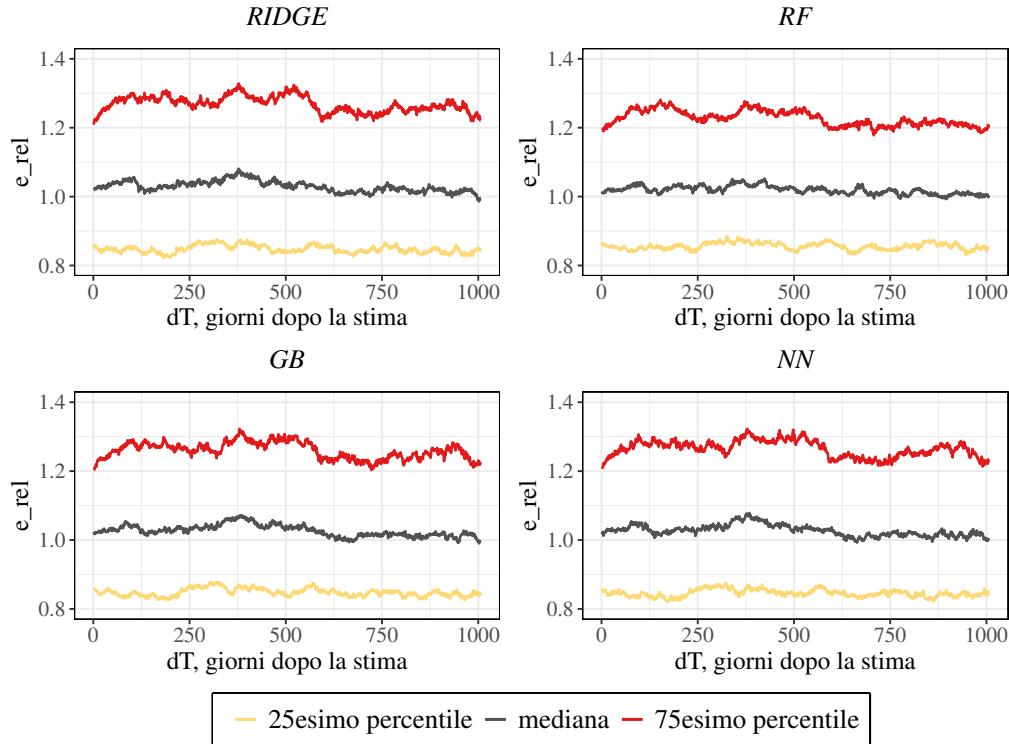


Figura 4.14: Combinazione dei grafici di *AI Aging*, che mostrano come le prestazioni dei modelli tendano a rimanere stabili. La qualità iniziale, raggiunta dai modelli, è pari a 0.75 per *ridge*, 0.69 per *RF*, 0.74 per *GB* e 0.75 per *NN*. In questo caso non ci attendiamo di osservare *degradazione temporale*, in quanto non vi è un allontanamento progressivo da una condizione iniziale: complessivamente i modelli dovrebbero mantenere la loro qualità, quando i tracciati dell'errore relativo ottenuti dalle singole stime vengono sovrapposti.

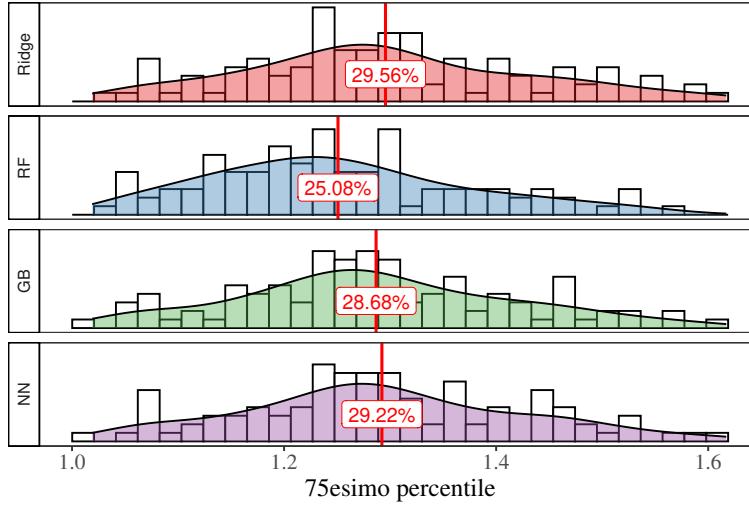


Figura 4.15: Distribuzione del livello medio del terzo quartile, nella simulazione con 30 osservazioni al giorno e un anno di ampiezza per gli insiemi di stima. La media complessiva, per la *RF*, è leggermente più bassa. Il range di valori osservato è molto ampio, indicando come questa problematica mini fortemente la stabilità dei modelli.

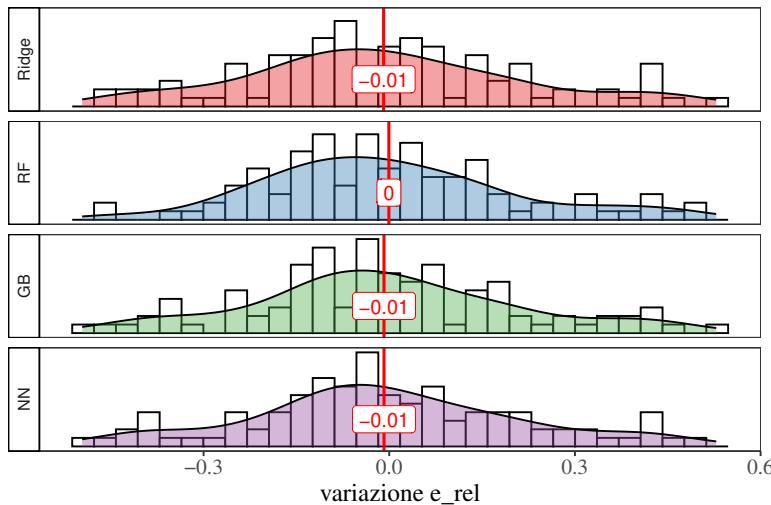


Figura 4.16: Distribuzione delle pendenze della mediana, nella simulazione con 30 osservazioni al giorno e un anno di ampiezza per gli insiemi di stima. Tutti i modelli presentano una distribuzione delle inclinazioni identica, e ci sono diversi casi in cui la *degradazione temporale* è elevata (ad esempio, variazioni maggiori di 0.3).

Questi risultati sono chiaramente ottenuti condizionatamente alla specifica ampiezza dell'insieme di stima, che al suo ridursi produce sicuramente un adattamento dei modelli più variabile nel tempo. Sperimentando con una dimensione inferiore, pari a 3 mesi (aumentando il numero di osservazioni a 60 al giorno per compensare la perdita di qualità) raggiungiamo però le stesse conclusioni (figura 4.17): i tre modelli simili mantengono una stabilità simile, mentre la *foresta* presenta una variabilità minore. Nonostante la differenza dagli altri modelli sia aumentata (9 - 10%, figura 4.17), e la qualità si sia mantenuta su livelli simili, valutando direttamente l'*MSE* (figura 4.18) questo modello non viene preferito. La stabilità è comunque maggiore, in linea con quanto osservato nei casi precedenti di *drift incrementale* e *graduale*.

Sulla base di quanto osservato fino a questo momento c'è sufficiente evidenza per credere che le differenze sarebbero molto ridotte, a parità di qualità iniziale, e che quanto osservato sia attribuibile alle differenze iniziali.

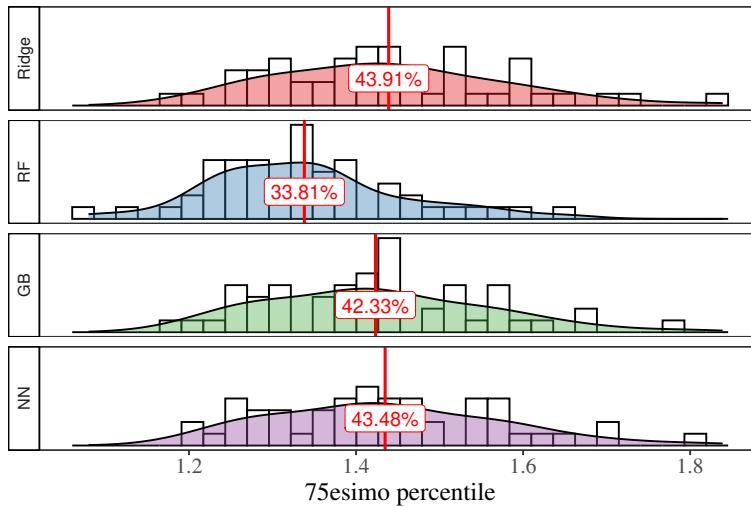


Figura 4.17: Distribuzione del livello medio del terzo quartile, nella simulazione con 60 osservazioni al giorno e tre mesi di ampiezza per gli insiemi di stima. La qualità iniziale dei modelli è pari a 0.74 per ridge, 0.67 per RF, 0.73 per GB e 0.73 per NN. Sono state condotte solo 50 replicazioni della procedura simulativa.

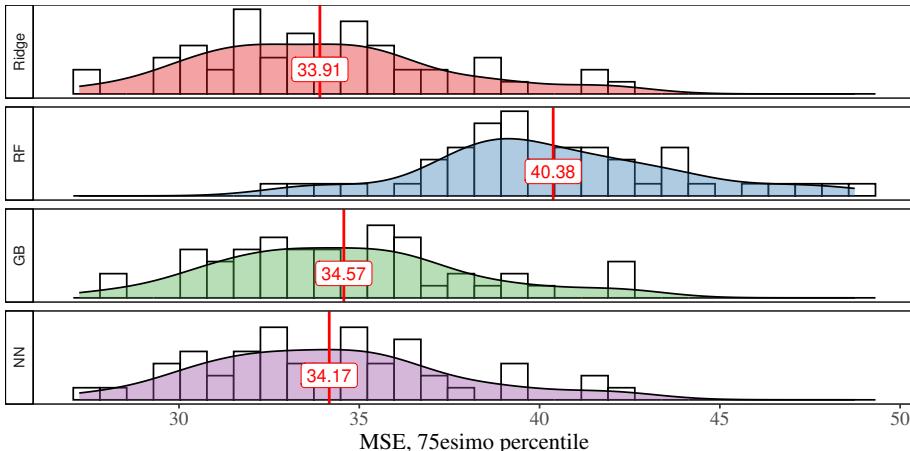


Figura 4.18: Distribuzione del livello medio del terzo quartile dell'MSE nella simulazione con 60 osservazioni al giorno e tre mesi di ampiezza per gli insiemi di stima. Nonostante la maggiore stabilità della foresta, comunque non viene preferita agli altri modelli, per via dell'errore maggiore. Sono state condotte solo 50 replicazioni della procedura simulativa.

Riassunto dei risultati

Le tabelle 4.1, 4.2 e 4.3 contengono dei brevi riassunti delle simulazioni condotte in questo capitolo (solo *real concept drift*, le uniche rilevanti). La prima fa riferimento ai casi di *concept drift graduale*, la seconda ai casi di *concept drift incrementale*, mentre la terza ai casi di *concept drift ricorrente*.

Simulazione	Risultati	Figure
Effetti lineari - 10 variabili	La <i>foresta casuale</i> soffre un minor calo delle prestazioni rispetto agli altri modelli, ma presenta un livello iniziale più basso.	4.2, 4.3
Effetti lineari - 4 variabili	Non ci sono differenze rilevanti nei risultati rispetto al caso iniziale.	B.5
Effetti lineari - Meno osservazioni	I modelli con una qualità iniziale minore presentano minori cali di prestazioni.	4.4
Effetti lineari - 60 variabili e meno osservazioni	Le prestazioni dei modelli si sono ridotte, nel tempo, in modo molto simile, con delle piccole differenze per i modelli peggiori.	4.5
Effetti lineari - 10 variabili e meno alberi per il <i>GB</i>	La qualità iniziale del modello si è ridotta, scendendo sotto i livelli iniziali degli altri modelli, e con essa l'effetto del <i>concept drift</i> sulle sue prestazioni, il cui peggioramento è minore rispetto a quello di tutti gli altri modelli.	4.6
Effetti lineari - 10 variabili e meno epoche per la <i>NN</i>	La qualità iniziale del modello si è ridotta, scendendo sotto i livelli degli altri modelli, e con essa l'effetto del <i>concept drift</i> sulle sue prestazioni, ma il peggioramento delle prestazioni è comunque maggiore rispetto a quello di altri modelli con qualità iniziale maggiore.	4.7
Effetti di interazione	Le prestazioni dei modelli migliori decadono di più.	B.6
Effetti di interazione - Meno osservazioni	Le prestazioni dei modelli migliori decadono di più.	B.7
Effetti cubici	Le prestazioni dei modelli migliori decadono di più.	B.8

Effetti cubici - Meno osservazioni	Le prestazioni della <i>foresta casuale</i> decadono in misura minore, pur presentando la migliore qualità iniziale.	B.9
------------------------------------	--	-----

Tabella 4.1: Riassunto delle simulazioni condotte nel capitolo 4, relativamente al *concept drift graduale*. Nella colonna “figure” sono indicate le figure che riportano i risultati. I numeri che iniziano con B indicano che le figure sono in appendice.

Simulazione	Risultati	Figure
Effetti lineari - 10 variabili	La <i>foresta casuale</i> soffre un minor calo delle prestazioni rispetto agli altri modelli, ma presenta un livello iniziale più basso.	4.9, 4.10
Effetti lineari - 4 variabili	Lo stimatore <i>ridge</i> , pur raggiungendo una qualità iniziale più alta, presenta un decadimento minore delle prestazioni rispetto a <i>RF</i> e <i>GB</i> . Una seconda simulazione, con differenti coefficienti, conferma lo stesso risultato.	4.12, B.10
Effetti lineari - Meno osservazioni	I modelli con una qualità iniziale minore presentano minori cali di prestazioni.	4.11
Effetti lineari - 60 variabili e meno osservazioni	I modelli con una qualità iniziale minore presentano minori cali di prestazioni. Le differenze, in questo caso, sono però minime.	B.11
Effetti lineari - 10 variabili e meno alberi per il <i>GB</i>	La qualità iniziale del modello si è ridotta, scendendo sotto i livelli degli altri modelli, e con essa l'effetto del <i>concept drift</i> sulle sue prestazioni, il cui peggioramento è minore rispetto a quello di tutti gli altri modelli.	B.12
Effetti lineari - 10 variabili e meno epoche per la <i>NN</i>	La qualità iniziale del modello si è ridotta, scendendo sotto i livelli degli altri modelli, e con essa l'effetto del <i>concept drift</i> sulle sue prestazioni, ma il peggioramento delle prestazioni è comunque maggiore rispetto a quello di altri modelli con qualità iniziale maggiore.	B.13
Effetti di interazione	Le prestazioni dei modelli migliori decadono di più.	B.14

Effetti di interazione - Meno osservazioni	Le prestazioni dei modelli migliori decadono di più.	B.15
Effetti cubici	Le prestazioni dei modelli migliori decadono di più, ma il peggioramento delle prestazioni della <i>foresta casuale</i> è minore di quanto atteso sulla base della qualità iniziale. Una seconda simulazione, con differenti coefficienti, non evidenzia la stessa peculiarità.	B.16, B.18
Effetti cubici - Meno osservazioni	Le prestazioni della <i>foresta casuale</i> decadono meno di quelle della <i>rete neurale</i> , pur presentando maggiore qualità iniziale. Una seconda simulazione, con differenti coefficienti, non evidenzia la stessa peculiarità.	B.17, B.19
Effetti di interazione - Meno alberi per il <i>GB</i>	I risultati sono analoghi a quelli ottenuti nel caso di effetti lineari e numero di alberi utilizzati ridotto.	B.20
Effetti cubici - Meno alberi per il <i>GB</i>	I risultati sono analoghi a quelli ottenuti nel caso di effetti lineari e numero di alberi utilizzati ridotto.	B.21

Tabella 4.2: Riassunto delle simulazioni condotte nel capitolo 4, relativamente al *concept drift incrementale*. Nella colonna “figure” sono indicate le figure che riportano i risultati. I numeri che iniziano con B indicano che le figure sono in appendice.

Simulazione	Risultati	Figure
Insieme di stima - 1 anno	L'effetto del <i>drift</i> si presenta in queste simulazioni tramite la variabilità dell'errore. La <i>foresta casuale</i> , in questi termini, presenta una stabilità maggiore degli altri modelli, ma è associata ad un livello iniziale minore.	4.14-4.16
Insieme di stima - 90 giorni	La <i>foresta casuale</i> presenta una stabilità molto maggiore degli altri modelli, a fronte di una riduzione contenuta della qualità iniziale. I risultati sono comunque in linea con quanto osservato nei casi di <i>concept drift graduale</i> e <i>incrementale</i> .	4.17, 4.18

Tabella 4.3: Riassunto delle simulazioni condotte nel capitolo 4, relativamente al *concept drift ricorrente*. Nella colonna “figure” sono indicate le figure che riportano i risultati. I numeri che iniziano con B indicano che le figure sono in appendice.

L’obiettivo delle simulazioni condotte in questa sezione era quello di evidenziare differenze nel modo in cui i modelli considerati rispondono al *real concept drift*; differenze potenzialmente attribuibili alla logica matematica su cui si basano. Ciò avrebbe permesso di scoprire perché diverse categorie di modello, sugli stessi dati, possono presentare differenze nella stabilità a medio/lungo termine delle prestazioni.

Per fare questo sono state simulate due principali tipologie di *drift*, *graduale* ed *incrementale*, le cui differenze in termini di risultati sono minime; ciò è confermato in quanto le simulazioni utilizzano gli stessi coefficienti, da una tipologia all’altra.

Questi *drift* sono stati inseriti in flussi di dati caratterizzati prima da relazioni lineari tra variabili esplicative e variabile risposta, in modo da poter confrontare i quattro modelli a parità di prestazioni iniziali, e successivamente non lineari.

I risultati delle due tipologie di *drift* concordano, perciò possono essere riassunti assieme di seguito (le figure il cui numero inizia per B sono riportate in appendice):

1. Nella maggior parte delle simulazioni condotte non sono emerse differenze nel modo in cui i modelli rispondono al *concept drift* che siano attribuibili, nello specifico, alle logiche matematiche impiegate (*CD graduale*: figure 4.3, 4.4, 4.5, B.6, B.7, B.8; *CD incrementale*: 4.10, 4.11, B.11, B.14, B.15, B.16).
2. Sono stati evidenziati invece dei punti in comune nel loro comportamento: sulla base delle stesse simulazioni riportate al punto precedente sono i modelli con un adattamento migliore a risentire maggiormente dei *drift*, a prescindere dalla logica su cui si basano. Questo è stato valutato tramite l’andamento degli *MSE* mediani, e non direttamente sulla base del “test” di *degradazione temporale* utilizzato, basato su errori relativi.
3. Allo stesso tempo i modelli con un adattamento iniziale peggiore sono risultati essere più resistenti al cambio di regole; e in diversi casi

modelli con qualità iniziale uguale/vicina hanno presentato un analogo peggioramento delle prestazioni.

Le differenze più importanti che sono state osservate nell'evoluzione delle prestazioni sono associate ad adattamenti iniziali differenti, ma non è possibile concludere che si tratti *solo* di una questione di adattamento. Differenze nel comportamento a medio/lungo termine dei modelli, dovute alla logica matematica, non possono infatti essere escluse.

In particolare, sono stati osservati dei casi in cui la qualità iniziale non è stata sufficiente a prevedere la risposta al *drift*:

1. Il primo in un caso di *drift incrementale*, in cui le prestazioni dello stimatore *ridge* peggiorano meno rispetto a quelle di *gradient boosting* e *rete neurale*, pur presentando un livello iniziale maggiore (figura 4.12). Non si tratta inoltre di un caso isolato, in quanto confermato tramite una seconda simulazione condotta con un insieme di coefficienti, iniziali e finali, differente (figura B.10, in appendice). Questa eccezione, relativamente a questo modello, è però stata osservata solo per quello specifico numero di variabili, e non in altri casi (10 o 60 variabili, figure 4.10 e B.11, in appendice).
2. Nel caso di relazioni cubiche e 5 osservazioni al giorno (figure B.9 e B.17, in appendice) la *foresta casuale*, pur con una qualità iniziale molto vicina a *gradient boosting* e *rete neurale*, ha presentato un aumento molto minore dell'*MSE*. Con 30 osservazioni al giorno e *drift incrementale* (figura B.16, in appendice), nonostante i livelli iniziali prevedessero correttamente i modelli con maggior decadimento, la *foresta* ha presentato un decadimento molto minore di quanto atteso sulla base dell'indice R^2 (*predittivo*). In una seconda simulazione, con un diverso insieme di coefficienti, iniziale e finale (e relazioni cubiche), non sono però state osservate tali irregolarità; potrebbero quindi trattarsi di casi isolati (figure B.18 e B.19, in appendice).
3. La *rete neurale* è il modello per il quale la previsione dell'effetto del *drift* sulla base della qualità iniziale è meno affidabile. Nei casi in

cui la procedura di stima ha avuto difficoltà a convergere, perché il numero di epoche è stato limitato di proposito (figure 4.7 e B.13, la seconda in appendice) oppure per altre ragioni (numero limitato di osservazioni, figure B.9 e B.17, in appendice) sono stati ottenuti dei modelli il cui livello iniziale è più variabile (nella stessa replicazione), complessivamente peggiore, ma per il quale l'effetto del *drift* è maggiore che per modelli con minore qualità. L'effetto del *drift* sulle *reti neurali* è quindi risultato essere meno prevedibile.

Questi risultati indicano che il valore di R^2 *predittivo* iniziale non è sufficiente a prevedere quanto un modello risentirà del *concept drift*, e ciò non permette di escludere l'idea che diverse logiche matematiche possano essere più adatte ad affrontare evoluzioni del processo, in determinati casi. Individuare tuttavia un modello che funzioni sempre meglio in determinate situazioni non è stato possibile.

Sulla base di questi risultati ritengo quindi che la risposta di un modello al *concept drift* sia prevalentemente un problema di adattamento al processo, e non di logica matematica. Detto questo, ciò non è in contrasto con quanto osservato in Vela et al. (2022), ma anzi, conferma che i modelli invecchiano in modo differente. L'invecchiamento di un modello è infatti considerato su una misura di errore relativo, e come indicano le simulazioni sono i modelli migliori che tendono a presentare gli aumenti maggiori di errore. Dal punto di vista relativo quindi, l'aumento di errore dovuto al cambio di regole è ulteriormente accentuato dal minor errore iniziale, e le differenze maggiori sono state osservate al variare della qualità iniziale, e non al pari.

Il “test” di *degradazione temporale* stesso non è perciò particolarmente adatto ad evidenziare differenze dovute alla logica matematica in quanto particolarmente sensibile alle differenze iniziali: applicarlo ad un flusso di dati in cui il processo evolve tende ad evidenziare come meno stabile il modello con un adattamento iniziale maggiore, senza che ci sia, necessariamente, una risposta sostanzialmente differente al *drift*. Un esempio è riportato in figura 4.19: possiamo osservare come il metodo evidensi come meno stabile lo stimatore *ridge*, seguito da *gradient boosting*, *rete neurale* e infine *foresta casuale*. Dal

punto di vista sostanziale però non c'è una vera differenza nel comportamento (figura 4.20). Non solo: lo stesso stimatore meno stabile è quello che presenta un errore minore a distanza massima, e quindi andrebbe comunque scelto.

Più in generale, il “test” *evidenzia sempre delle differenze nel comportamento a medio/lungo termine dei modelli* per il solo fatto che le qualità iniziali sono differenti. Per costruzione è quindi impossibile capire, dai soli risultati del “test”, se le differenze siano effettive oppure no. In aggiunta, come osservato in queste simulazioni, al variare dei livelli iniziali si osservano anche differenze sostanziali nella risposta del modello al *drift*, impossibili da distinguere rispetto al caso precedente. Più importante, in nessuna conclusione ottenuta fino a questo punto è stato possibile evitare di includere l'informazione relativa alle prestazioni iniziali.

Ritengo dunque che parlare di stabilità temporale del modello, di invecchiamento o di *degradazione temporale* della qualità, come qualcosa di separato dall'adattamento del modello, sia sbagliato; da ciò ne dipende l'evoluzione temporale delle prestazioni, molto più di quanto dipenda dalla logica matematica.

In conclusione quindi i modelli possono presentare differente *degradazione temporale*, sugli stessi dati, per il solo fatto di presentare una qualità iniziale non esattamente pareggiata. Dal punto di vista sostanziale, nella maggior parte dei casi, la logica matematica ha avuto ben poco a che fare con la risposta del modello al *drift*; anche se differenze, da quel punto di vista, non possono essere escluse, in quanto alcune eccezioni sono state osservate.

Tra i modelli considerati l'effetto del *drift* sulle prestazioni della *rete neurale* è più complesso da prevedere sulla base dell' R^2 *predittivo*, soprattutto nei casi in cui la procedura di stima fatica a raggiungere il minimo.

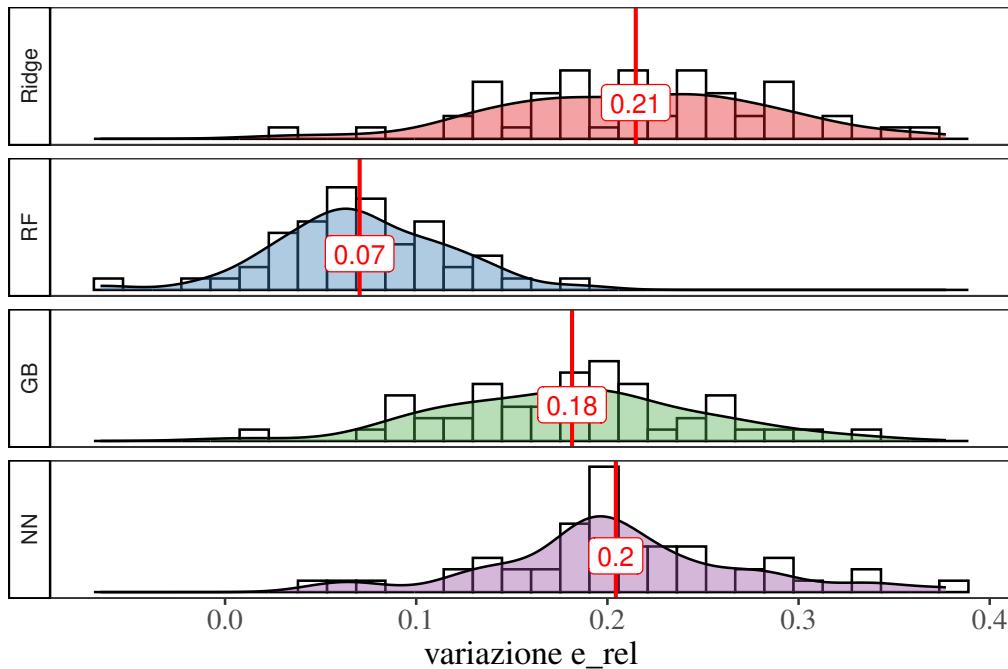


Figura 4.19: Distribuzione della pendenza delle mediane, calcolate sugli errori relativi. I modelli per i quali viene indicata maggiore *degradazione temporale* sono gli stessi che presentano un adattamento migliore (riportato in figura 4.20). Questi risultati sono relativi ad un caso in cui il numero di variabili esplicative è pari a 60, gli effetti sono lineari, il numero di osservazioni al giorno pari a 10 e in cui è presente *drift graduale*. La simulazione è stata condotta con 50 repliche.

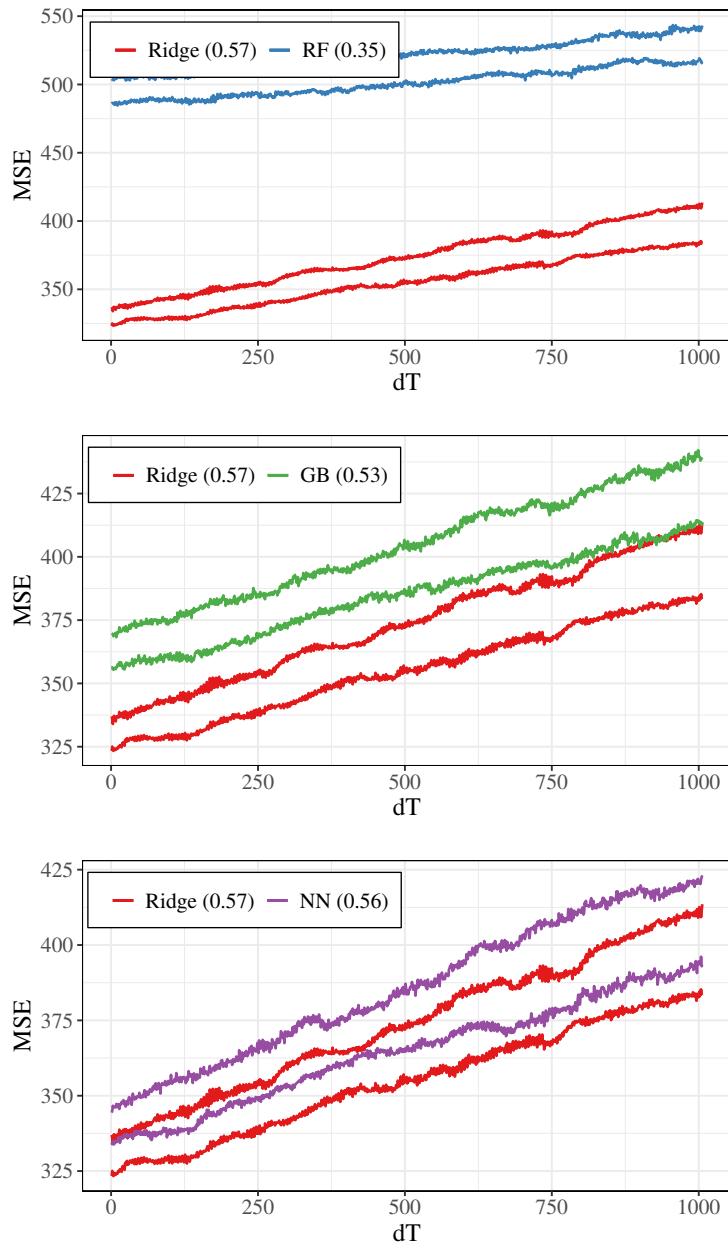


Figura 4.20: Andamenti degli MSE mediani, associati ai risultati riportati nella figura 4.19. Gli andamenti sono praticamente identici, con delle differenze molto ridotte per i modelli peggiori.

Capitolo 5

Simulazioni: stagionalità

I risultati presentati nel capitolo 4 hanno evidenziato come la *degradazione temporale* dei modelli sia strettamente legata alla qualità dell’adattamento iniziale, e queste conclusioni si ripresentano in questo capitolo.

Ciò rende impossibile distinguere la natura dei risultati ottenuti dal “test” di *degradazione temporale*, in quanto le differenze osservate tra i modelli sulla base dell’errore relativo possono:

1. Essere dovute esclusivamente alle differenze iniziali, e non indicare una reale differenza nell’evoluzione delle prestazioni a medio/lungo termine (valutate direttamente tramite *MSE*).
2. Indicare differenze reali nel modo in cui le prestazioni evolvono nel tempo (valutate tramite *MSE*) dovute alle diverse qualità iniziali.
3. Indicare differenze sostanziali (reali) nel modo in cui le prestazioni evolvono nel tempo, dovute alle diverse logiche dei modelli.

In quest’ultimo capitolo il “test” di *degradazione temporale*, nonostante parzialmente inadatto, viene utilizzato per esplorare un’altra possibile causa di *degradazione temporale* proposta in Vela et al. (2022): la presenza di fattori latenti ciclici o stagionali, che possono influire sull’evoluzione temporale delle prestazioni di un modello. Nel loro caso, in particolare, la valutazione deriva dai risultati ottenuti su un dataset caratterizzato da stagionalità annuale e

insiemi di stima di dimensione pari ad un anno: per questo motivo le stesse specifiche saranno utilizzate in queste simulazioni.

La stagionalità annuale viene generata utilizzando il modello strutturale di base per serie storiche proposto in Harvey (1989), che propone la seguente struttura per modellare tali dati:

$$y_t = \mu_t + \gamma_t + \varepsilon_t$$

La serie osservata (y_t) viene espressa in funzione di una componente di livello (μ_t), di una componente stagionale (γ_t) e di una componente non osservata (ε_t). L'evoluzione nel tempo delle prime due componenti può essere modellata in diversi modi; in questo caso la componente di trend non è stata utilizzata (una simulazione che la include è riportata alla fine del capitolo). La componente stagionale, scandita da s effetti stagionali, viene descritta dalla seguente equazione.

$$\gamma_t = \sum_{j=1}^{s-1} -\gamma_{t-j} + \omega_t \quad \omega_t \sim N(0, \sigma_\omega^2)$$

Secondo questa regola gli effetti stagionali, ogni anno, non sommano a zero: ciò fa sì che ogni stagione di un nuovo anno presenti un effetto differente dall'anno precedente, sporcato tramite un disturbo casuale.

Alla solita struttura lineare dei dataset simulati (con quattro variabili esplicative, per fare posto alle variabili necessarie a catturare la componente stagionale e ridurre il divario iniziale tra i modelli) viene quindi aggiunta una componente stagionale che evolve, generata tramite la seguente procedura:

1. È stato scelto un insieme di valori iniziali per gli effetti stagionali, pari a 12, disposti in una forma a campana, con somma a zero. Questi valori rimangono costanti per tutte le replicazioni condotte.
2. In modo iterativo, i valori per gli anni successivi vengono ottenuti secondo la formula descritta in precedenza, scegliendo un valore per la

varianza del termine di disturbo appropriato. Ad ogni iterazione i disturbi sono generati casualmente.

3. I valori così ottenuti vengono resi continui tramite lisciamento: questi vengono collocati ad equidistanza nell'arco dell'anno, vengono generati dei punti per riempire lo spazio e questi vengono infine lasciati, ottenendo un andamento continuo. Questi valori, uno per giorno, vengono sommati alla restante struttura lineare.

Ogni dataset prevede 2 anni di periodo di burn-in e ulteriori 10 anni di osservazioni, a differenza dei 5 usati finora. Questo permette di utilizzare un disturbo minore e osservare *degradazione temporale*, allungando la distanza massima a cui monitorare le prestazioni: 2831 giorni, quasi otto anni, determinata automaticamente una volta deciso di creare i sottoinsiemi di stima sui primi due anni (come fatto nelle precedenti simulazioni) e fissata l'ampiezza degli insiemi di verifica e delle finestre mobili (la stessa di sempre). Un esempio di dataset generato tramite questa metodologia è presentato in figura 5.1.

Per catturare la componente stagionale sono possibili differenti approcci, in questo lavoro ne sono stati considerati tre:

1. Utilizzare il giorno dell'anno e il mese, informazione disponibile quando limitiamo l'ampiezza dell'insieme di stima ad un anno. In una prima simulazione solo queste variabili sono state incluse nei dataset (in aggiunta alle X). Questo è anche il caso più significativo, in quanto in questo lavoro viene esplorata l'alternativa rispetto ai metodi di aggiornamento dei modelli, e uno dei motivi per cui la loro applicazione è resa ostica è l'impossibilità di accedere ai dati più recenti.
2. Utilizzare i ritardi stagionali (della media giornaliera): in questa seconda simulazione il primo è stato incluso nei dataset.
3. Utilizzare i primi ritardi (della media giornaliera): in questa terza simulazione sono stati inclusi nei dataset i primi due ritardi, che permettono di seguire le fasi di crescita e di decrescita.

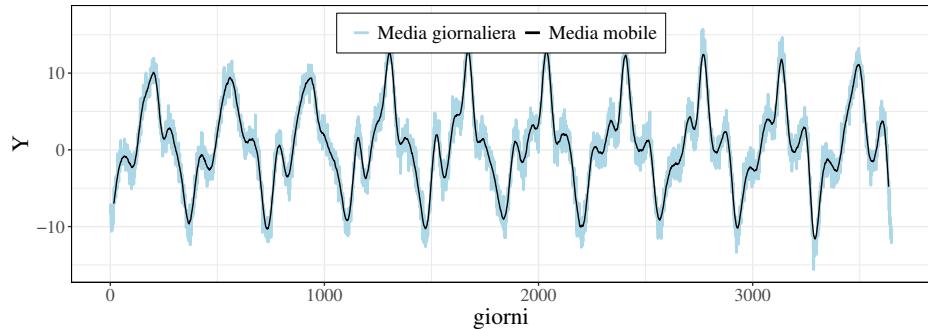


Figura 5.1: Primo dataset contenente una componente stagionale. L’importanza di questa componente all’interno del flusso è chiara, e il disturbo non è eccessivo: da un anno al successivo l’effetto stagionale rimane largamente invariato, ma è possibile osservare differenze nel lungo periodo.

L’utilizzo delle diverse possibilità comporta livelli di qualità iniziale complessivi differenti e invecchiamenti differenti. Le singole simulazioni condotte permettono di valutare le differenze osservate nel comportamento dei modelli legate a ciascun approccio per catturare la componente stagionale. Come in un caso reale, però, questi vengono raramente utilizzati da soli: sono quindi stati utilizzati assieme nelle successive simulazioni.

A differenza dei casi studiati finora i modelli hanno a disposizione diverse alternative per catturare le variazioni stagionali: ciascuna porta a diversi gradi di *degradazione temporale*, e quando usate assieme possono permettere di osservare differenze nel comportamento di medio/lungo termine dei modelli. Tutte le simulazioni di questo capitolo utilizzano non solo la stessa regola per generare la stagionalità, ma anche gli stessi dati, in modo da poter confrontare i risultati.

Risultati - Singoli approcci

Le prime simulazioni condotte permettono di valutare l’effetto di ciascun approccio (per catturare la componente stagionale) sulle prestazioni dei modelli. I valori di R^2 *predittivo* iniziali sono riportati nelle figure 5.2 - 5.4. I livelli iniziali raggiunti dai modelli sono molto differenti e dipendono dalle va-

riabili utilizzate, che influiscono sull’evoluzione temporale delle prestazioni: i grafici di *degradazione temporale* sono riportati in figura 5.5 per la prima simulazione, in figura 5.7 per la seconda, e in figura C.1, in appendice, per la terza (molto simile alla seconda).

Il risultato principale, tra queste prime simulazioni, è associato al caso in cui vengono utilizzate le sole variabili relative al giorno dell’anno e al mese (assieme alle X). Negli altri casi è atteso che la *degradazione temporale* sia molto ridotta: i modelli possono seguire il processo che evolve utilizzando informazione più recente nelle previsioni. Osservare *degradazione temporale* in quei casi implica aver utilizzato una componente di disturbo elevata, che comporta grandi differenze da un anno al successivo, e ciò non è necessariamente in linea con una manifestazione di una componente stagionale.

Nel primo caso (mese e giorno dell’anno) l’invecchiamento dei modelli è molto simile, come dimostra il grafico (figura 5.5), tuttavia il “test” utilizzato evidenzia delle differenze tra i modelli non imputabili, in questo caso, a qualcosa di sostanziale (figura 5.6, per le pendenze relative, figura 5.8, per gli andamenti degli *MSE* mediani, che mostrano un comportamento sostanzialmente identico): i modelli che invecchiano maggiormente sono anche quelli con una qualità iniziale maggiore, ma questa è un’indicazione del “test” che non riflette vere differenze nell’evoluzione temporale delle prestazioni, ed è dovuta solo alle differenti qualità iniziali.

I risultati relativi all’utilizzo degli altri approcci indicano una *degradazione temporale* molto minore (figura 5.7, per il ritardo stagionale, mentre figura C.1, in appendice, per il caso con i primi ritardi giornalieri), con una maggiore variabilità dell’errore relativo (livello medio del terzo quartile, in appendice, figure C.2 e C.3) per *gradient boosting* e *foresta casuale* (e in misura minore la *rete neurale*), in quanto soffrono maggiormente del *data drift*. Il disturbo infatti provoca, nel tempo, la comparsa di osservazioni esterne al range osservato durante la stima (picchi più alti e più bassi), i cui valori compaiono tra le variabili esplicative. Per lo stesso motivo questi modelli presentano anche maggiore *degradazione temporale*, in questi casi.

Queste sono differenze prevedibili in base alle caratteristiche dei modelli: è sempre un problema di estrapolazione.

Utilizzare i singoli approcci non evidenzia quindi differenze nel modo in cui le prestazioni dei modelli evolvono nel tempo, che non siano prevedibili a priori. In questi casi la diversa qualità iniziale dei modelli non ha avuto un effetto identificabile sull'evoluzione temporale effettiva delle prestazioni (figura 5.8 e, in appendice, C.4, C.5, C.6 e C.7, per gli andamenti degli *MSE* mediani nei casi non riportati).

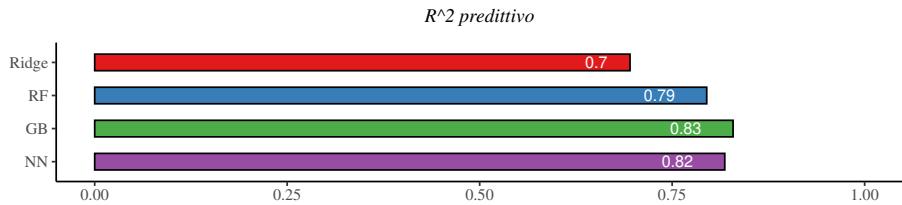


Figura 5.2: Qualità iniziale dei modelli, con il giorno dell'anno e il mese di appartenenza dell'osservazione. Le differenze sono minime, ad eccezione che per lo stimatore *ridge*, che presenta una qualità inferiore perché non fa un utilizzo appropriato della prima variabile.

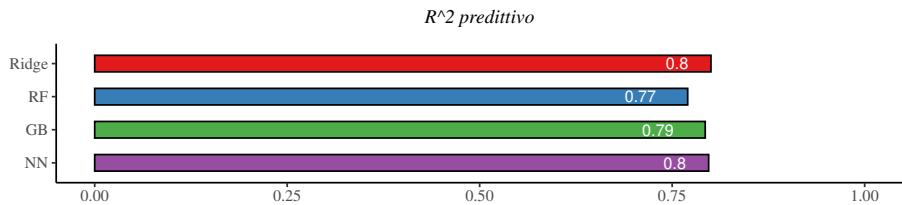


Figura 5.3: Qualità iniziale dei modelli, con il primo ritardo stagionale. La qualità iniziale è minore rispetto al caso precedente, e le differenze iniziali sono minime. Le replicazioni dell'esperimento sono 50.

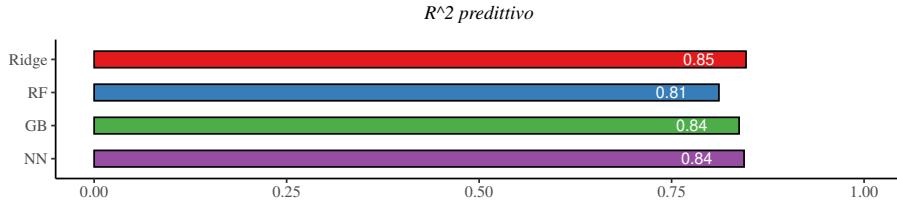


Figura 5.4: Qualità iniziale dei modelli, con i primi due ritardi. I livelli iniziali sono molto più alti dei casi precedenti, per tutti i modelli. Le replicazioni dell'esperimento sono 50.

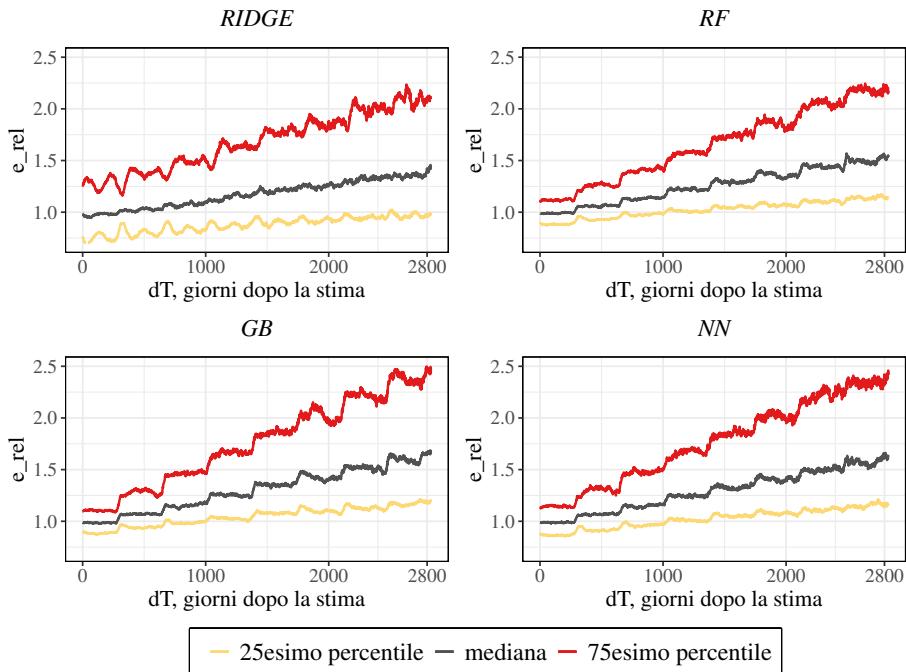


Figura 5.5: La combinazione dei grafici di *AI Aging* mostra degli andamenti simili per *RF*, *GB* e *NN*, con una forma a scalini, osservata in alcune simulazioni in cui è presente stagionalità. La differenza con lo stimatore *ridge* è dovuta al fatto che non fa utilizzo del giorno dell'anno. La variabilità dell'errore tra i tre modelli migliori è molto simile, con delle piccole differenze possibilmente imputabili ai livelli iniziali. Lo stimatore *ridge* presenta una variabilità simile alla *foresta casuale* nonostante le differenze iniziali, possibilmente attribuibile al diverso utilizzo delle variabili disponibili.

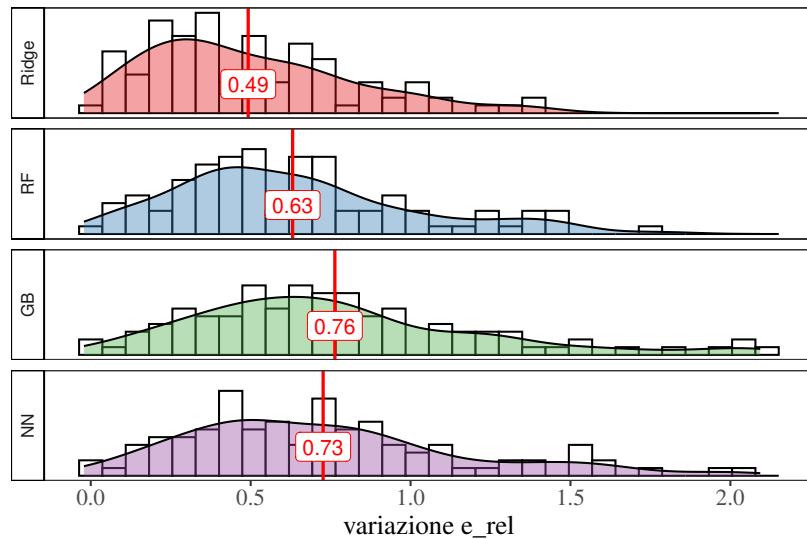


Figura 5.6: Il grafico mostra la distribuzione della pendenza delle mediane, calcolate sull'errore relativo, che, come nel capitolo precedente, mostra come l'ordinamento dei modelli sulla base della *degradazione temporale* coincida con quello basato sulla qualità iniziale. Le differenze non sono però sostanziali.

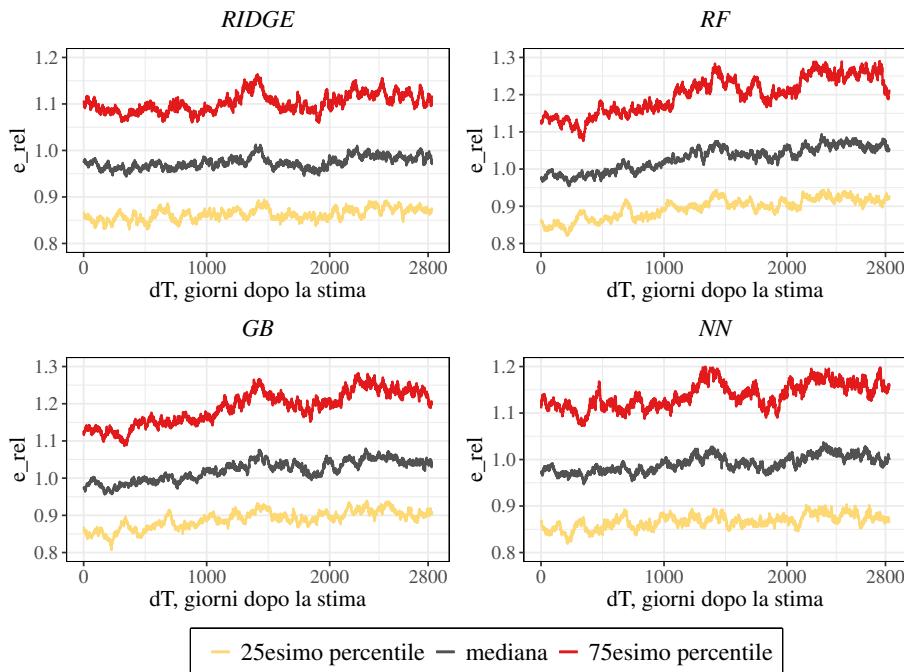


Figura 5.7: La combinazione dei grafici di *AI Aging*, nella simulazione in cui viene utilizzato solo il primo ritardo stagionale. Il grafico mostra come la *degradazione temporale* dei modelli sia molto ridotta, ma maggiore per *RF* e *GB*. Inoltre, *foresta casuale* e *gradient boosting*, seguiti dalla *rete neurale*, presentano un livello maggiore per il 75esimo percentile, probabilmente a causa della maggiore sensibilità al problema del *data drift* (che provoca la maggiore *degradazione*). Le replicazioni dell'esperimento sono 50.

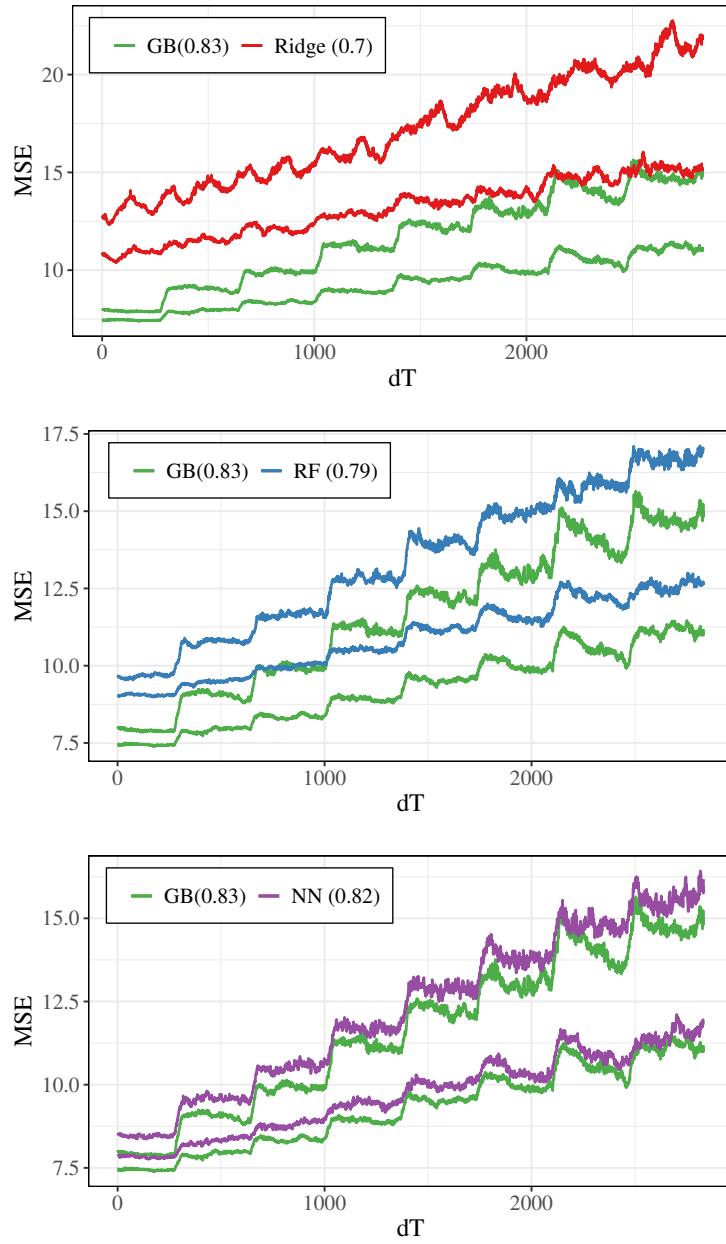


Figura 5.8: Andamenti degli MSE mediani, che mostrano un effetto più omogeneo del *concept drift* sui modelli. La forma è differente per lo stimatore *ridge*, in quanto utilizza il mese dell'anno, ma non ci sono differenze sostanziali evidenti.

Risultati - Più approcci in contemporanea

Le simulazioni precedenti sono state condotte per verificare come ogni variabile possa contribuire alla spiegazione delle variazioni stagionali e all’evoluzione temporale delle prestazioni di un modello. In un contesto reale però, spesso, non vengono utilizzate separatamente, e la presenza di alternative permette ai modelli di differenziarsi.

La prima simulazione condotta, in cui più approcci (più variabili, per catturare la componente stagionale) sono presenti, li include tutti, e mostra delle differenze sostanziali (e non dovute al “test”) nel comportamento a medio/lungo termine dei modelli (figura 5.11). Il *gradient boosting* presenta infatti un andamento complessivo dell’*MSE* mediano molto differente, ed è l’unico che si distingue. Osservando le singole replicazioni ciò è chiaro (un esempio riportato in figura 5.12): il modello presenta differenze pur partendo dallo stesso livello iniziale.

I risultati relativi alla qualità iniziale sono riportati in figura 5.9: essendo le differenze ridotte l’invecchiamento dei modelli dovrebbe essere simile. I risultati ottenuti contraddicono però ciò che è stato osservato fino ad ora: il *gradient boosting* presenta un invecchiamento maggiore (figura 5.10) di *ridge* e *rete neurale* pur avendo un livello iniziale identico, e maggiore della *foresta casuale* pur soffrendo in modo simile i problemi di *data drift*. Gli altri tre modelli presentano un comportamento di medio/lungo termine identico.

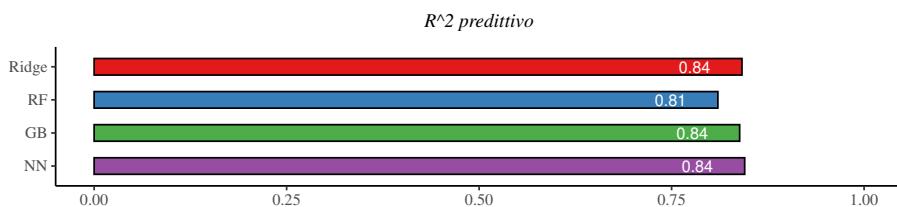


Figura 5.9: Qualità iniziale dei modelli, utilizzando tutte le variabili disponibili. Le differenze iniziali sono minime.

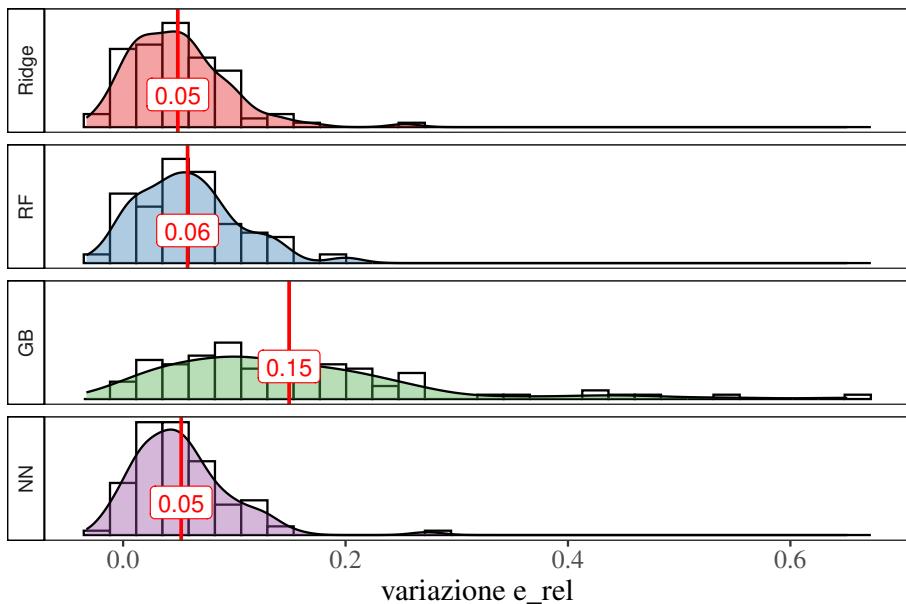


Figura 5.10: Variazioni dell'errore relativo, nella simulazione con variazioni stagionali e tutte le variabili disponibili. Il *gradient boosting* presenta un invecchiamento molto maggiore rispetto agli altri modelli, un risultato inatteso. Il maggior invecchiamento della *foresta casuale*, rispetto a *ridge* e *NN*, potrebbe essere dovuto alla maggiore sensibilità al *data drift*.

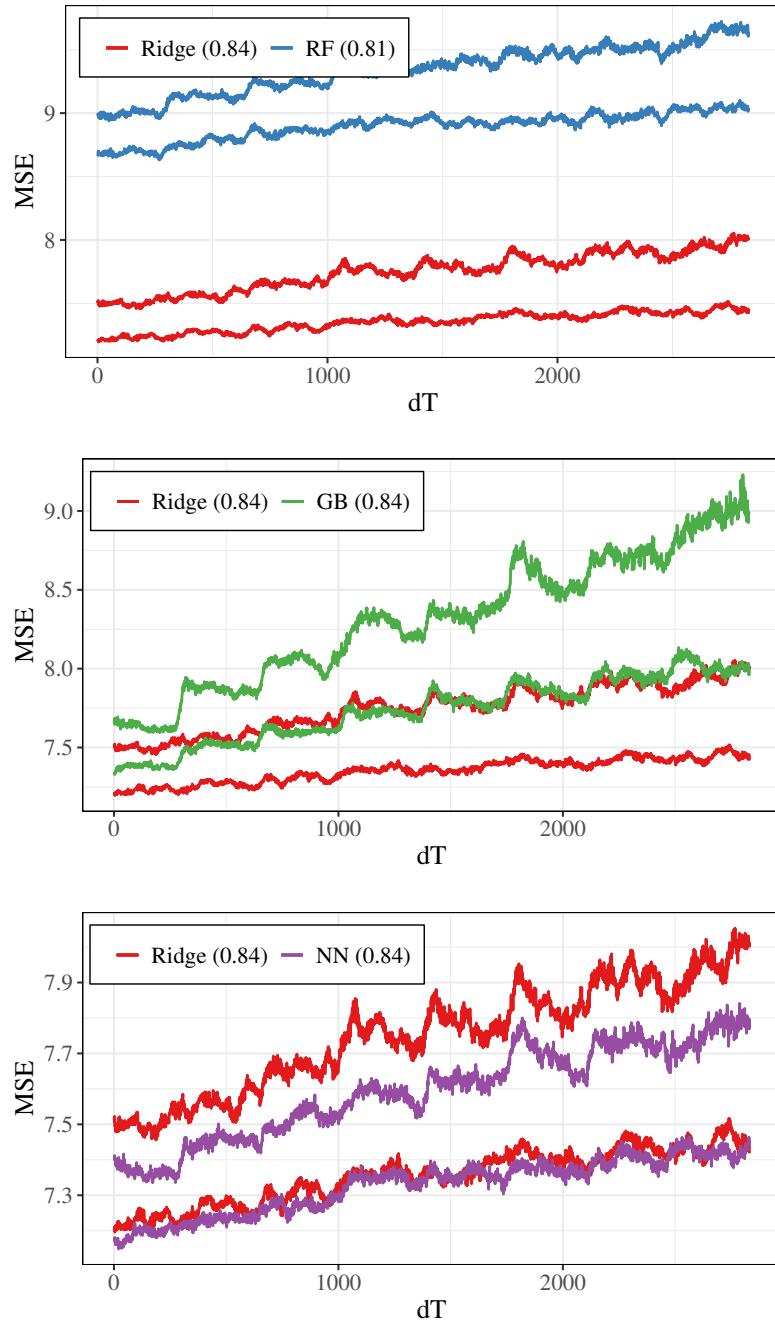


Figura 5.11: Andamenti degli MSE mediani, relativamente alla simulazione in cui tutte le alternative per catturare le variazioni stagionali sono utilizzate assieme. L'unico modello con un comportamento differente è il *gradient boosting*.

Il perché delle differenze non è chiaro, le singole variabili non ne sono direttamente la causa. La simulazione è stata ripetuta, cercando di evidenziare le cause, organizzando le variabili utilizzate nel modo seguente:

1. La prima simulazione utilizza mese, giorno dell'anno e primo ritardo stagionale (risultati in appendice, figura C.8).
2. La seconda simulazione utilizza mese, giorno dell'anno e primi due ritardi giornalieri (risultati in appendice, figura C.9).
3. Non è stata condotta una simulazione con i soli ritardi in quanto le uniche variabili per le quali le prestazioni si riducono in modo importante sono mese e giorno dell'anno.

In entrambi i casi il *gradient boosting* presenta un comportamento chiaramente distinto, nel secondo caso di più, differenziandosi decisamente dal modello più simile (*RF*).

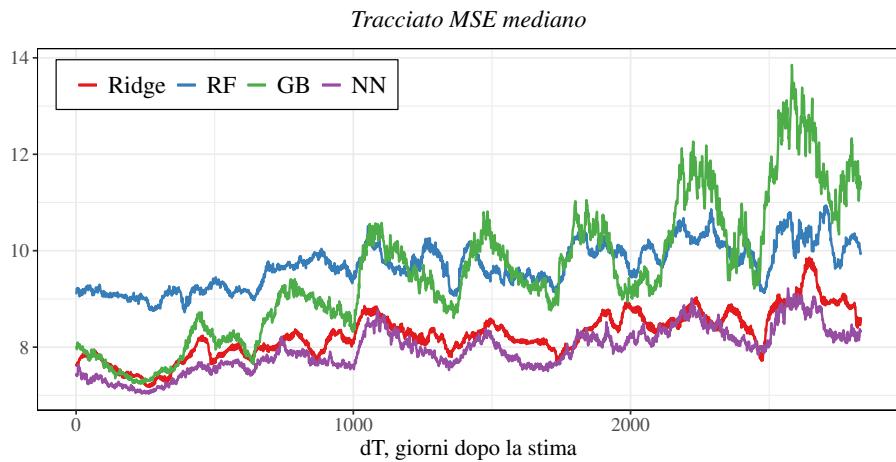


Figura 5.12: Tutti gli approcci considerati sono utilizzati contemporaneamente: i tracciati delle mediane degli $MSE(dT)$, di una replicazione (sono tutte analoghe), mostrano come le prestazioni del *gradient boosting* evolvano in modo diverso rispetto a quelle degli altri modelli.

Per cercare di approfondire l'origine delle differenze, e capire se queste possono essere sistematiche, sono state condotte ulteriori prove:

1. Per valutare quanto lo scostamento del *gradient boosting* dal modello più simile (*RF*) sia dovuto all’adattamento iniziale (ormai noto per influire sull’evoluzione temporale delle prestazioni) il numero di alberi è stato limitato. Nonostante le differenze si siano notevolmente ridotte, pareggiando i livelli iniziali dei due modelli, non sono del tutto scomparse (in appendice, figure C.10 e C.11), soprattutto nei tracciati dei casi peggiori.

Il livello iniziale gioca quindi un ruolo importante.

2. Sono state condotte due simulazioni in cui l’importanza della componente stagionale è stata prima aumentata (figura C.13, in appendice) assieme alla componente di disturbo, e poi ridotta (figure C.14 e C.16, in appendice). Mentre nel primo caso le differenze con la *foresta* sono scomparse, nel secondo anche le *reti neurali* hanno presentato un comportamento a medio/lungo termine differente: mentre le differenze negli andamenti degli *MSE* mediani sono minime (C.14), quelle nei casi peggiori sono maggiori (C.16, riportate anche in figura 5.13), a parità di qualità iniziale. Le combinazioni del terzo quartile dell’*MSE*, e i loro andamenti, non sono state riportate nei casi precedenti perché valevano le stesse considerazioni fatte per le mediane.

Il *gradient boosting* non è perciò l’unico modello che può presentare un comportamento differente, ma il verificarsi di ciò dipende dal dataset stesso.

3. In un’ultima simulazione, condotta utilizzando la stessa componente stagionale delle prime effettuate in questo capitolo, è stata inclusa una componente di trend; questo per rendere il problema affrontato più realistico. Sono state nuovamente impiegate tutte le variabili disponibili, e i risultati sono riportati in appendice (figure C.18, C.19, C.20 e C.21). In questo caso i risultati sono più difficili da leggere, e le differenze più complesse. Il *gradient boosting* rimane il più irregolare, e mostra differenze con la *foresta* solo rispetto al caso mediano (C.19), e non nel caso peggiore (C.20): la differenza principale è quindi nel peggioramento complessivo delle prestazioni, e non nella variabilità dell’errore; la stessa relazione vale tra *ridge* e *rete neurale* (figure C.18 e C.21) (nel

capitolo 3 i risultati erano molto peggiori per la *rete* in caso di trend, soprattutto nel terzo quartile), anche se le differenze sono molto minori. I risultati concordano con quanto osservato finora, anche se le mediane non sono più sufficienti per avere una visione completa delle differenze. Il comportamento a medio/lungo termine dei modelli, e le differenze, sono comunque fortemente influenzate dalle scelte arbitrarie compiute, legate all'importanza della componente stagionale e di trend (figura C.17).

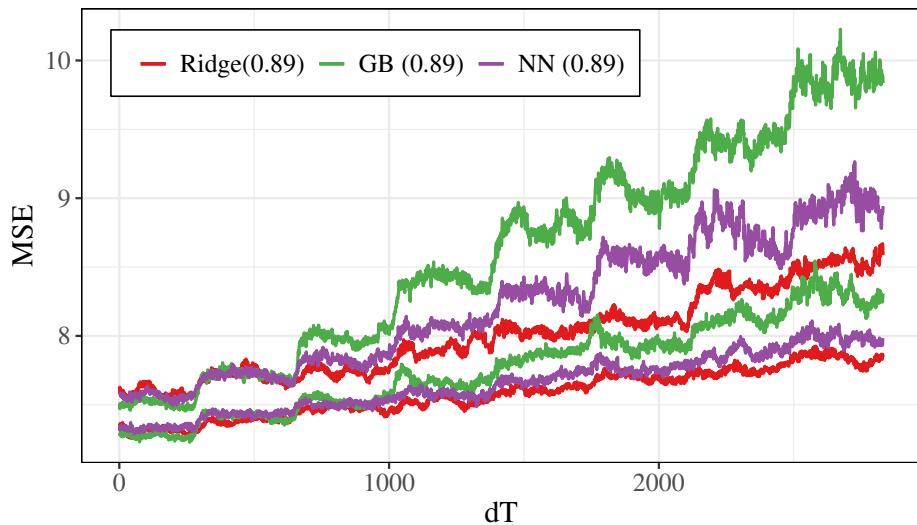


Figura 5.13: Confronto degli andamenti dei terzi quartili dell' MSE combinati, di *ridge*, *foresta casuale* e *gradient boosting*. I modelli presentano delle differenze importanti: le *reti neurali* si collocano in una posizione intermedia in termini di stabilità. Queste differenze non sono emerse nelle altre simulazioni condotte.

Riassunto dei risultati

Nella tabella 5.1 è riportato un riassunto delle simulazioni condotte in questo capitolo.

Simulazione	Risultati	Figure
Solo mese e giorno dell'anno	Il decadimento delle prestazioni è sostanzialmente identico per tutti i modelli	5.2, 5.5, 5.6, 5.8
Solo primo ritardo stagionale	Il decadimento delle prestazioni è molto ridotto per tutti i modelli. La variabilità dell'errore relativo è però maggiore per <i>RF</i> e <i>GB</i> , in quanto soffrono maggiormente del problema del <i>data drift</i> (sono quindi meno stabili). Per lo stesso motivo presentano un invecchiamento maggiore.	5.3, 5.7, C.2, C.4, C.5
Solo primi due ritardi giornalieri	Il decadimento delle prestazioni è molto ridotto per tutti i modelli. La variabilità dell'errore relativo è però maggiore per <i>RF</i> e <i>GB</i> , in quanto soffrono maggiormente del problema del <i>data drift</i> (sono quindi meno stabili). Per lo stesso motivo presentano un invecchiamento maggiore.	5.4, C.1, C.3, C.6, C.7
Tutte le variabili in contemporanea	Il comportamento a medio/lungo termine del <i>gradient boosting</i> si discosta notevolmente da quello degli altri modelli, che presentano invece un comportamento simile tra loro.	5.9-5.12
Mese, giorno dell'anno e primo ritardo stagionale	Il comportamento a medio/lungo termine del <i>gradient boosting</i> si discosta notevolmente da quello degli altri modelli, che presentano invece un comportamento simile tra loro.	C.8
Mese, giorno dell'anno e primi due ritardi giornalieri	Il comportamento a medio/lungo termine del <i>gradient boosting</i> si discosta notevolmente da quello degli altri modelli, che presentano invece un comportamento simile tra loro.	C.9
Tutte le variabili - Meno alberi per il <i>gradient boosting</i>	La riduzione della qualità iniziale del <i>GB</i> lo ha portato allo stesso livello della <i>RF</i> . Le differenze si sono ridotte notevolmente, specialmente nei casi mediani, ma non sono scomparse (la differenza nei tracciati del terzo quartile è notevole).	C.10, C.11

Tutte le variabili - Componente stagionale aumentata	Le differenze tra <i>GB</i> e <i>RF</i> sono scomparse, ma sono aumentate rispetto agli altri due modelli, che risultano molto più stabili nel medio/lungo periodo.	C.13
Tutte le variabili - Componente stagionale ridotta	La differenza tra <i>gradient boosting</i> e gli altri modelli è rimasta, ma in questo caso anche la <i>rete neurale</i> ha presentato delle differenze nei tracciati del caso peggiore, indicando un decadimento maggiore delle prestazioni rispetto allo stimatore <i>ridge</i> . Il comportamento di questi due modelli era sempre rimasto lo stesso nelle precedenti simulazioni.	C.14, C.16, 5.13
Tutte le variabili - Aggiunta componente di trend	I risultati concordano con quanto osservato nelle altre simulazioni, ma sono più complessi. A coppie di modelli, vengono osservate differenze negli andamenti degli <i>MSE</i> mediani, ma non in quelli dei terzi quartili combinati.	C.18- C.21

Tabella 5.1: Riassunto delle simulazioni condotte nel capitolo 5, relativamente al problema della stagionalità. Nella colonna “figure” sono indicate le figure che riportano i risultati. I numeri che iniziano con C indicano che le figure sono in appendice.

In questo capitolo è stata valutata la stabilità delle prestazioni a medio/lungo termine dei modelli in presenza di stagionalità, uno dei problemi che influiscono sull’evoluzione temporale delle prestazioni identificati in Vela et al. (2022).

Questa è valutata al variare delle variabili utilizzate per catturare il fenomeno (giorno dell’anno e mese, primo ritardo stagionale e primi ritardi giornalieri), permettendo di osservare come:

1. Le diverse variabili, considerate individualmente, permettono ai modelli diversi livelli di qualità iniziale (figure 5.2, 5.3 e 5.4) e diversi gradi di *degradazione temporale* (figure 5.5, 5.7 e C.1, in appendice). Sugli stessi dati però le differenze, anche sostanziali, sono state minime e prevedibili (figure 5.8, C.4 e C.6, in appendice).
2. Utilizzando le variabili mese e giorno dell’anno il decadimento delle prestazioni è stato molto maggiore, rispetto ad utilizzare i ritardi; questo

è atteso, in quanto in questi ultimi casi viene utilizzata informazione più recente nelle previsioni, e la differenza da un anno al successivo è molto ridotta.

3. Utilizzando il primo ritardo stagionale o i primi ritardi giornalieri, variabili che presentano *data drift*, la *foresta casuale* e il *gradient boosting*, e in misura minore la *rete neurale*, hanno presentato una variabilità maggiore delle prestazioni (figure C.2 e C.3, in appendice); i primi due anche una *degradazione temporale* maggiore.

Nessun singolo approccio per catturare le variazioni stagionali porta ad osservare differenze importanti nella stabilità a medio/lungo termine dei modelli. In tutti i casi, però, i risultati del “test” risultano poco interpretabili, in quanto indicano sempre differenze: tutte le conclusioni derivano dallo studio dell’errore non relativo.

Una volta identificato il comportamento dei modelli quando le variabili sono utilizzate singolarmente, queste sono state incluse assieme, come in un caso reale. I risultati sono riassunti di seguito:

1. Utilizzando tutte le variabili disponibili l’unico modello che ha presentato una differenza importante nell’evoluzione temporale delle prestazioni è stato il *gradient boosting*, una differenza non spiegabile sulla base della qualità iniziale raggiunta, allo stesso livello di *ridge* e *reti* (figura 5.11, per gli andamenti degli *MSE* mediani, e figura 5.12, per la singola replicazione).
2. Il livello iniziale, nonostante non possa spiegare la differenza, gioca un ruolo importante: limitare il numero di alberi (per il *GB*) ha portato ad una sua riduzione (figure C.10 e C.11, in appendice).
3. La differenza osservata rispetto agli altri modelli, in particolare la *foresta casuale*, il modello più simile in termini di logica matematica, dipende dalle specifiche arbitrarie delle simulazioni condotte: aumentare l’importanza della componente stagionale e il disturbo elimina le differenze tra i due (figura C.13, in appendice), mentre riducendone

l’importanza anche la *rete neurale* si è discostata dal modello più simile, lo stimatore *ridge* (figura 5.13).

In base a quanto osservato non ci sono le basi per poter affermare che il *gradient boosting* non vada utilizzato in caso di variazioni stagionali (anche se la sua stabilità, nei casi studiati, è sempre stata la peggiore): il comportamento a medio/lungo termine dipende dagli equilibri presenti nel dataset, dalla qualità iniziale raggiunta e dalle variabili utilizzate. Il manifestarsi della differenza è inoltre legato alla presenza di variabili alternative per spiegare il fenomeno: c’è ragione di pensare che anche gli altri modelli possano, in casi analoghi, compiere scelte che portano ad un maggiore decadimento della qualità; come è successo alla *rete neurale* nel caso riportato in figura 5.13.

Questo caso della *rete neurale* è interessante: il modello aveva già presentato variabilità maggiore del *ridge* (figura C.2, in appendice), ma le differenze in questi termini erano state risolte utilizzando i primi ritardi giornalieri (figura C.3, in appendice); che erano disponibili nella simulazione a cui la figura 5.13 fa riferimento. Come per il *gradient boosting* quindi, queste differenze non sono facilmente spiegabili.

Le simulazioni condotte permettono però di confermare che logiche matematiche differenti possono rispondere in modo differente ai *drift* nei dati. È quindi necessario valutare, caso per caso, il miglior modello da utilizzare, considerando anche l’adattamento iniziale raggiunto: ciò esclude l’utilizzo del solo “*test*” di *degradazione temporale*.

Un aspetto da considerare, importante per costruire un modello che resista adeguatamente alla prova del tempo, è la costruzione del dataset. Nei casi reali spesso viene realizzato un insieme di dati che contiene un gran numero di variabili, molta informazione, su cui i modelli vengono regolati e stimati fino ad ottenere prestazioni iniziali adeguate. Come affermato in Vela et al. (2022) la stabilità delle prestazioni nel tempo può quindi essere valutata per ciascun modello candidato all’uso, in modo da scegliere il migliore a medio/lungo termine; questo è anche l’approccio più facilmente realizzabile. In queste ultime simulazioni è però stato osservato come il *gradient boosting* abbia risentito notevolmente della presenza di troppe variabili: utilizzando

i soli primi ritardi giornalieri le prestazioni iniziali del modello, elevate (R^2 *predittivo* pari a 0.84), si sono mantenute molto bene nel tempo, ma il comportamento di medio/lungo periodo è cambiato notevolmente includendone di più, a parità di qualità iniziale (in figura 5.14 il confronto). I dataset generati, su cui sono state condotte le due simulazioni, sono assolutamente identici, cambiano solo le variabili disponibili ai modelli.

La costruzione del dataset può quindi essere tanto importante quanto la scelta del modello, per ottenere uno strumento con delle prestazioni a medio/lungo termine adeguate.

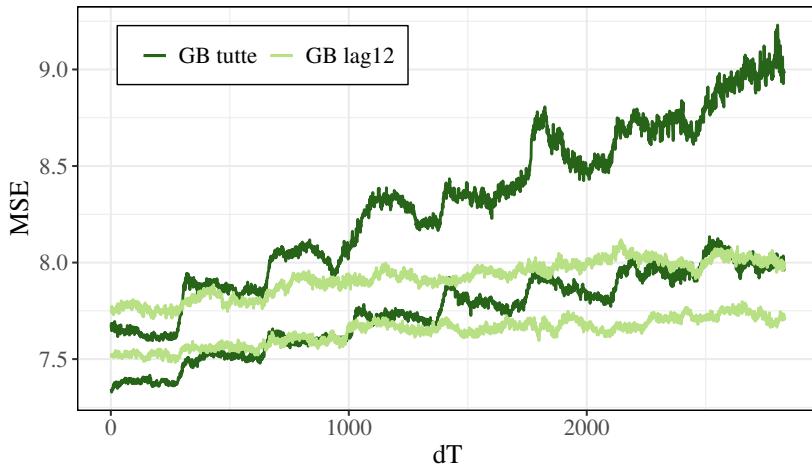


Figura 5.14: Confronto degli andamenti degli *MSE* mediani per il *gradient boosting*, utilizzato nella simulazione con stagionalità, con tutte le variabili possibili (*GB tutte*) e con i soli primi ritardi giornalieri (*GB lag12*). La qualità iniziale dei due modelli è la stessa, ma il peggioramento delle prestazioni è molto differente.

Conclusioni

L’obiettivo del lavoro svolto era quello di approfondire l’aspetto temporale dell’impiego di alcuni modelli di *machine learning* di comune utilizzo, quando questi vengono impiegati senza essere mai aggiornati. Tramite un approccio di simulazione si è cercato di evidenziare delle differenze nel modo in cui le prestazioni evolvono nel tempo, quando questi strumenti vengono applicati sugli stessi dati.

Utilizzando come punto di partenza i risultati presentati in Vela et al. (2022) e le analisi condotte nel capitolo 2, il “test” per lo studio della *degradazione temporale* è stato applicato in una serie di situazioni differenti: iniziando da alcuni casi in cui il processo che genera i dati non evolve nel tempo, passando a situazioni in cui invece ciò accade, e terminando con lo studio dei modelli in presenza di stagionalità.

Le simulazioni condotte non hanno permesso di evidenziare differenze o peculiarità nel modo in cui le prestazioni di un modello evolvono nel tempo, che possano essere associate alla logica matematica su cui si basano (ad eccezione che nei casi scontati, come quando c’è estrapolazione). Le differenze osservate sono infatti spesso attribuibili, o quantomeno associate, a differenze nella qualità iniziale dei modelli:

1. Quando il processo che genera i dati non evolve nel tempo (capitolo 3) i modelli inizialmente migliori sono spesso quelli che hanno presentato prestazioni meno volatili. Quando ciò non si è verificato le differenze osservate sono state, al più, minime.

2. Quando invece il processo evolve (capitolo 4) sono questi stessi modelli, inizialmente migliori, che risentono maggiormente del cambiamento.

Questo è quanto osservato nella maggior parte delle simulazioni condotte.

I livelli iniziali giocano quindi un ruolo molto importante nel determinare l'evoluzione futura delle prestazioni di un modello, a prescindere dalla tipologia, e differenze, anche ridotte, possono avere un effetto rilevante. Appurata l'esistenza di questa relazione, diventa molto difficile attribuire gli scostamenti osservati nel comportamento a medio/lungo termine alla logica matematica. Fare ciò è reso più complesso dal “test” di *degradazione temporale* stesso, che tende ad evidenziare sempre delle differenze per via del solo scostamento iniziale, che ha un impatto rilevante sui valori di errore relativo osservati. I modelli tendono infatti ad invecchiare in modo differente in ogni caso in cui lo scostamento è presente. Ciò rende impossibile ricavare informazioni utili dal confronto dei risultati, tra modelli sugli stessi dati, in quanto le differenze possono essere:

1. Dovute alla sola differenza in partenza, senza che questa esista veramente in termini di andamento dell'errore ($MSE(dT)$).
2. Dovute alla differenza in partenza, che porta a differenze effettive in termini di andamento dell'errore ($MSE(dT)$).
3. Dovute alla logica matematica, e al modo in cui risponde alla situazione.

Individuare le peculiarità specifiche dei modelli, con queste componenti di disturbo e questo strumento, diventa quindi meno fattibile di quanto inizialmente sperato.

Tutto ciò rende i concetti stessi di *degradazione temporale*, invecchiamento o stabilità di un modello privi di senso pratico, in quanto slegati dall'adattamento del modello ai dati ma profondamente influenzati da esso. Ciò mette in discussione l'utilità stessa del “test”.

A queste difficoltà è possibile aggiungere il fatto che l'approccio di simulazione stesso è impreciso: non ci sono metodi totalmente soddisfacenti per

combinare i risultati delle singole replicazioni, e ciò è tanto più vero nel capitolo 5, in cui i grafici di *AI Aging* (che non sono stati riportati) presentano forme più irregolari. Ciò rende i confronti meno precisi, rispetto a quanto può essere fatto sul singolo dataset.

In tutto questo, differenze nel modo in cui le logiche matematiche influenzano l’evoluzione delle prestazioni non possono essere escluse. Sono stati osservati casi specifici in cui la relazione tra adattamento iniziale e stabilità delle prestazioni non è stata rispettata:

1. Quando il processo non evolve (capitolo 3) sono state osservate situazioni (escludendo il caso di estrapolazione) in cui modelli migliori hanno presentato una maggiore volatilità dell’errore rispetto ad altri peggiori, talmente ridotte e sporadiche però da rendere difficile imputarle a qualcosa di sistematico.
2. Quando è presente (*real*) *concept drift* sono stati osservati casi in cui la perdita di prestazioni è maggiore per modelli inizialmente peggiori, ma non in modo abbastanza consistente per alcun modello. L’unica costante è per la *rete neurale*, modello per il quale è più difficile prevedere l’effetto del *drift* in base all’adattamento iniziale quando l’algoritmo di stima fatica a convergere. Questa è l’unica differenza consistentemente legata allo specifico modello che è stata osservata.
3. Nel caso di stagionalità, la scelta multipla tra le variabili disponibili per catturarla ha messo maggiormente in difficoltà il *gradient boosting*.

Complessivamente tutti i modelli hanno presentato delle eccezioni alla regola, in cui il comportamento è stato meno regolare e inatteso; per questo motivo la stabilità delle prestazioni dovrebbe essere studiata caso per caso.

Mentre questo può aiutare ad individuare un modello che resiste meglio alla prova del tempo, un aspetto da non sottovalutare è la costruzione del dataset: nel capitolo 5 è stato osservato come concedere al *gradient boosting* troppe variabili abbia minato la sua capacità di mantenere la qualità iniziale nel tempo. Scegliere un modello migliore per lo specifico dataset è sicuramente

l'approccio più comodo, ma correggere i dati per migliorare la tenuta del modello può rivelarsi altrettanto rilevante.

Tenere conto dell'evoluzione temporale delle prestazioni non è però semplice, viste le criticità del “test” di *degradazione temporale*. Una possibile alternativa consiste nel costruire il grafico di *AI Aging* utilizzando direttamente gli $MSE(dT)$, senza ricorrere all'errore relativo, in modo da visualizzare il vero cambiamento delle prestazioni e rendere il confronto dei risultati immediato. Un esempio di grafici prodotti in questo modo, costruiti sui dataset utilizzati nel capitolo 2, sono riportati in appendice nelle figure D.1 - D.4. La forma complessiva è la stessa e il confronto è più immediato (nella stessa sezione i risultati sui dati reali sono confrontati con quanto ricavato dalle simulazioni).

In conclusione, non è stato possibile capire come le logiche matematiche entrino in gioco nel determinare il comportamento a medio/lungo termine dei modelli, e non è stato possibile ottenere informazione utile che possa indicare una preferenza per una determinata logica rispetto ad un'altra. Invece, nelle simulazioni condotte, è stato osservato come i modelli rispondano in modo simile alle stesse problematiche, con le differenze principali in stabilità associate alla qualità iniziale.

Sulla base di quanto osservato la scelta basata sul modello inizialmente migliore tende anche ad essere la più valida, con una preferenza per quelli con una migliore estrapolazione; criterio di scelta difficile da applicare, in quanto si tratta di scegliere il modello migliore su valori non osservati. La stabilità delle prestazioni dipende quindi da una buona conoscenza del dominio di applicazione.

Nel caso il processo evolva le prestazioni di tali modelli tendono però a ridursi maggiormente, e ciò comporta un'esigenza ancora maggiore di implementare metodi di aggiornamento automatico.

Bibliografia

- Azzalini, A. & Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. OUP USA.
- Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5-32.
- Budiman, A., Fanany, M. I. & Basaruddin, C. (2016). Adaptive online sequential ELM for concept drift tackling. *Computational Intelligence and Neuroscience*.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29, 1189-1232.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys* 46.
- Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.
- Higham, N. J. (2002). Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis* 22.
- Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42.
- Jones, D. A. (1978). Nonlinear Autoregressive Processes. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 360.
- LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.
- Lima, M., Neto, M., Filho, T. S. & Fagundes, R. A. de A. (2022). Learning Under Concept Drift for Regression. *IEEE Access* 10.

- Liu, Z., Loo, C. K. & Seera, M. (2019). Meta-cognitive Recurrent Recursive Kernel OS-ELM for concept drift handling. *Applied Soft Computing* 75.
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J. & Zhang, G. (2018). Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31.
- Martínez-Rego, D., Pérez-Sánchez, B., Fontenla-Romero, O. & Alonso-Betanzos, A. (2011). A robust incremental learning method for non-stationary environments. *Neurocomputing* 74.
- Pianykh, O. S., Guitron, S., Parke, D., Zhang, C. et al. (2020). Improving healthcare operations management with machine learning. *Nature Machine Intelligence* 2.
- Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. Springer, New York.
- Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A. & Pianykh, O. S. (2022). Temporal quality degradation in AI models. *Nature Scientific Reports* 12.
- Xu, K., Zhang, M., Li, J., Du, S. S., Kawarabayashi, K. & Jegelka, S. (2020). How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks. *ArXiv*.
- Žliobaitė, I., Pechenizkiy, M. & Gama, J. (2016). *An overview of concept drift applications*. In *Big Data Analysis: New Algorithms for a New Society*, 91-114. Springer.

Appendice A

Materiale capitolo 3

In questo capitolo sono riportate delle figure aggiuntive relative alle simulazioni condotte nel capitolo 3. Le figure sono separate in due sezioni:

1. La prima contiene il materiale aggiuntivo relativo alle simulazioni in cui i flussi di dati sono indipendenti dalla componente temporale.
2. La seconda contiene il materiale aggiuntivo relativo alle simulazioni in cui la componente temporale viene inserita nei flussi di dati.

Assenza di concept drift: indipendenza dalla componente temporale

Effetto della matrice di correlazione e della componente non osservata

Elenco delle figure:

1. In figura A.1 è riportato un esempio di matrice utilizzata nelle simulazioni condotte (matrice di sinistra), mentre nel secondo caso una variazione per valutare l'impatto della scelta sui risultati delle simulazioni.
2. In figura A.2 sono riportati i risultati della simulazione in cui le variabili esplicative sono indipendenti e i loro effetti lineari.

3. In figura A.3 sono riportati i risultati della simulazione in cui le variabili esplicative sono maggiormente correlate (figura A.1, seconda matrice) e i loro effetti lineari.
4. In figura A.4 sono riportati i risultati della simulazione in cui le variabili esplicative hanno effetti lineari, ma la componente non osservata ha un'importanza maggiore.

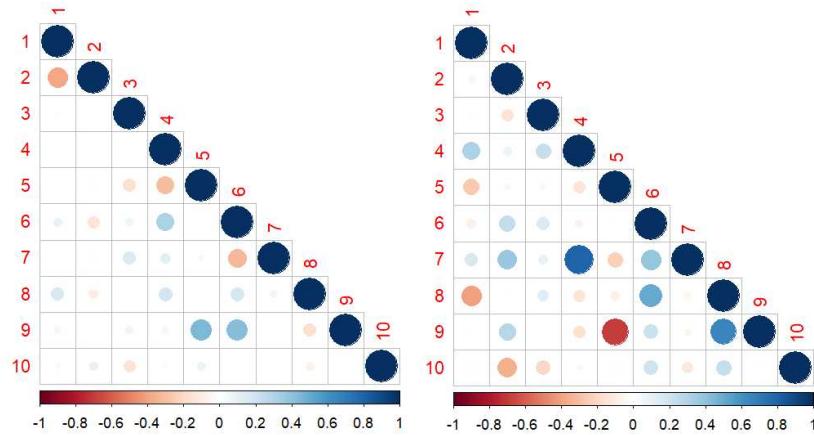


Figura A.1: Le matrici di correlazione utilizzate per la struttura lineare: a sinistra la matrice utilizzata nella prima simulazione (e utilizzata in quelle successive), che presenta una dipendenza lineare leggera tra le variabili esplicative, mentre a destra una matrice con correlazioni più forti. Con la matrice di destra la singola variabile non apporta nuova informazione: viene spiegata totalmente dalle altre in un modello lineare.

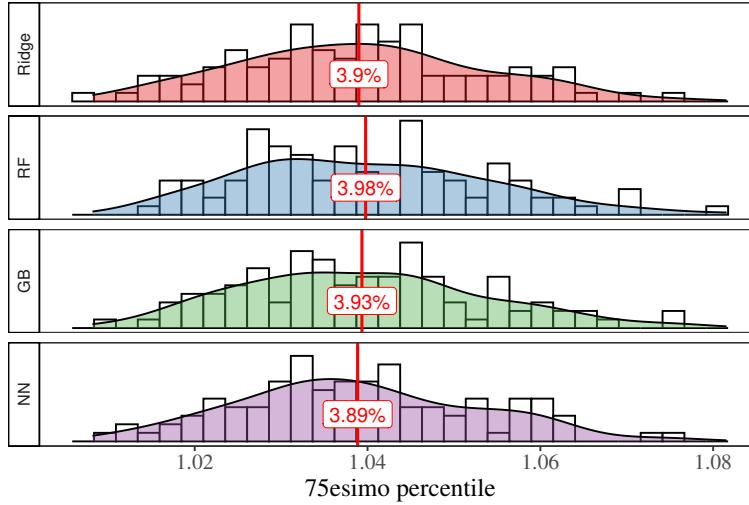


Figura A.2: Distribuzione del livello medio del 75esimo percentile, nella simulazione con relazione lineare e variabili esplicative indipendenti. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.87, RF = 0.81, GB = 0.87 e NN = 0.87.

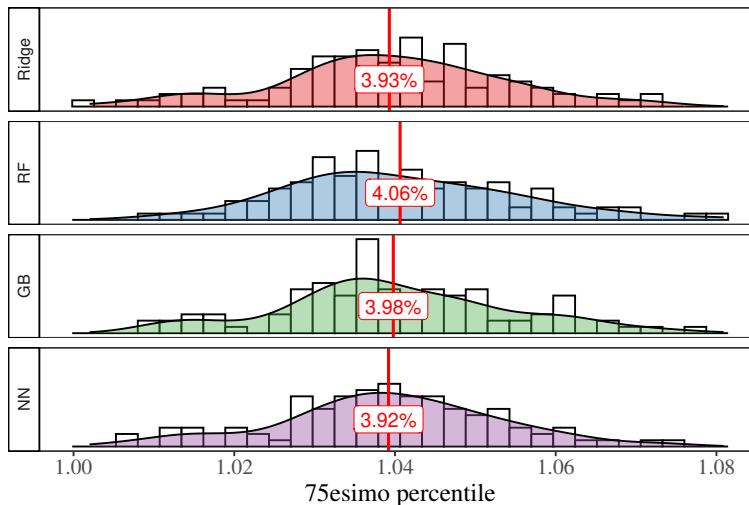


Figura A.3: Distribuzione del livello medio del 75esimo percentile, nella simulazione con relazione lineare e variabili esplicative maggiormente correlate. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.87, RF = 0.83, GB = 0.87 e NN = 0.87.

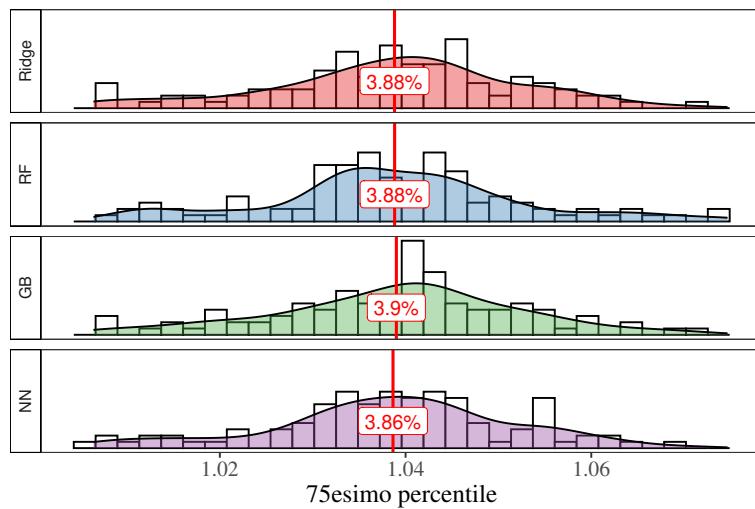


Figura A.4: Distribuzione del livello medio del 75esimo percentile, nella simulazione con relazione lineare e una maggiore componente non osservata. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.47, RF = 0.41, GB = 0.46 e NN = 0.47.

Situazioni aggiuntive: variabili di disturbo e maggiore errore

In questa breve sezione sono riportati i risultati di alcune simulazioni condotte con relazione lineare tra X ed Y :

1. In figura A.5 sono riportati i risultati di una simulazione in cui la componente di errore ha un'importanza notevole, e quindi la qualità iniziale dei modelli è molto ridotta. Le replicazioni dell'esperimento simulativo sono pari a 50.
2. In figura A.6 sono riportati i risultati di una simulazione in cui le variabili X sono altamente correlate, con valori attorno a ± 0.91 . Le replicazioni condotte in questo caso sono 50.
3. In figura A.7 sono riportati i risultati di una simulazione in cui le variabili X sono altamente correlate, e 8 di queste (su 10) sono associate ad un coefficiente pari a 0. Le replicazioni condotte in questo caso sono 100.

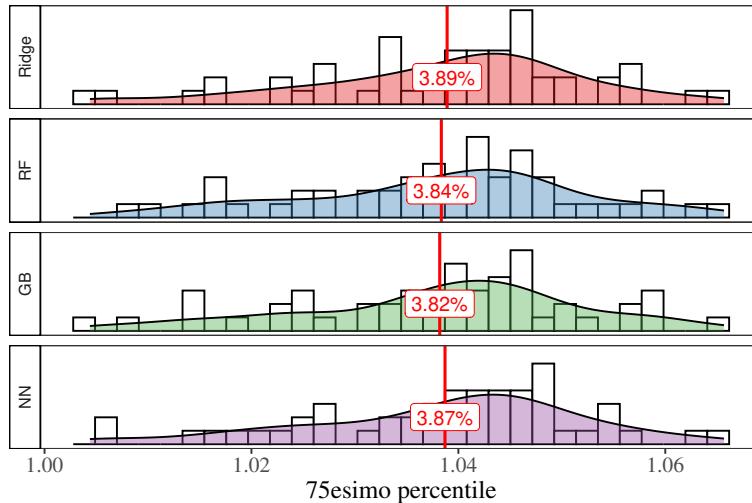


Figura A.5: Distribuzione del livello medio del 75esimo percentile, con una componente non osservata maggiore. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.18, RF = 0.15, GB = 0.17 e NN = 0.18. 50 replicazioni.

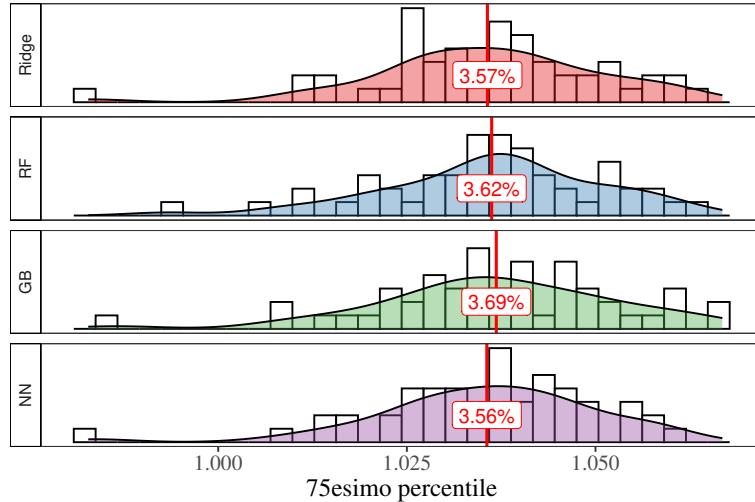


Figura A.6: Distribuzione del livello medio del 75esimo percentile, elevata correlazione tra le variabili esplicative. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.87, RF = 0.85, GB = 0.86 e NN = 0.87. 50 repliche.

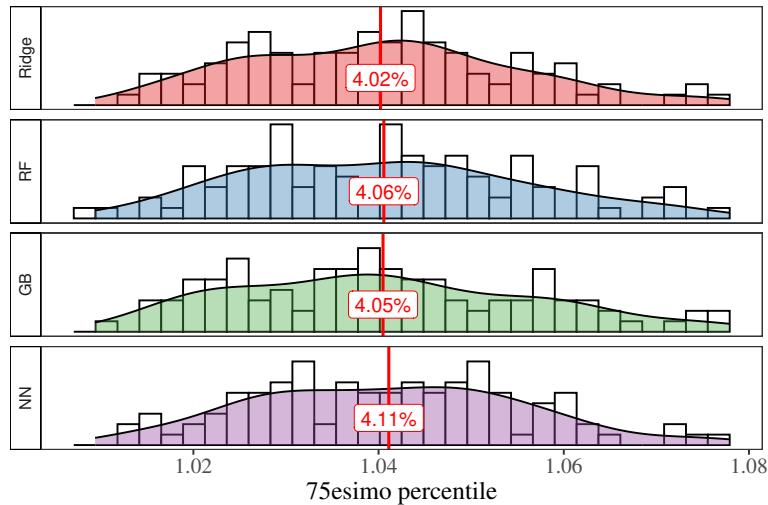


Figura A.7: Distribuzione del livello medio del 75esimo percentile, nel caso di variabili altamente correlate e di disturbo. I livelli di R^2 predittivo iniziale raggiunti sono: ridge = 0.83, RF = 0.82, GB = 0.83 e NN = 0.83. 100 repliche.

Relazioni non lineari

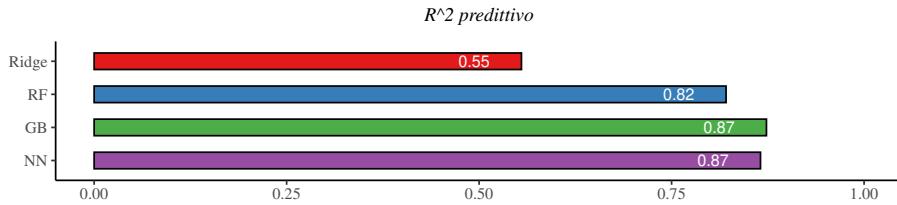


Figura A.8: Qualità iniziale dei modelli, nella simulazione caratterizzata da relazioni cubiche tra le variabili esplicative e la variabile risposta. La *rete neurale*, che stavolta utilizza una struttura a tre strati latenti e 200 nodi per strato, presenta una qualità iniziale maggiore.

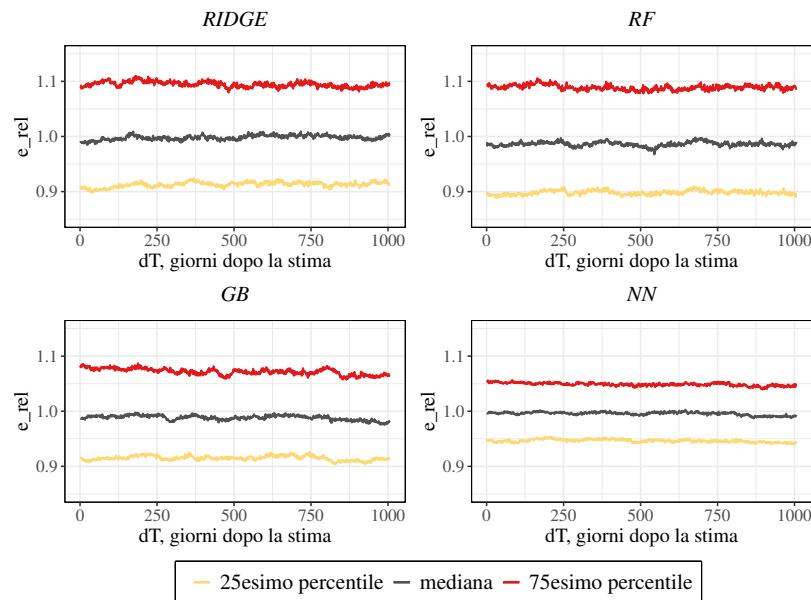


Figura A.9: Combinazione dei grafici di *AI Aging* relativi alla simulazione con relazioni cubiche, ottenuto utilizzando delle reti neurali più complesse (200 nodi latenti per strato), che mostrano come la *rete neurale* mantenga una variabilità dell'errore minore degli altri modelli.

Assenza di concept drift: dipendenza dalla componente temporale

In questa sezione sono riportate le simulazioni in cui sono incluse delle componenti associate all'aspetto temporale del flusso di dati. Questi sono processi autoregressivi, inclusi nella componente non osservata dell'equazione che definisce la struttura dei dataset. I processi utilizzati sono i seguenti, uno alla volta:

- Il primo è un processo AR(2), appartenente alla famiglia ARMA, una scelta ovvia per i processi stazionari. Questo processo è definito dalla seguente equazione:

$$(1 - \phi_1 B - \phi_2 B^2)u_i = \eta_i$$

dove $(\phi_1, \phi_2) = (0.6, 0.3)$, B rappresenta l'operatore di *backshift* e η_i costituisce il termine di innovazione con distribuzione normale. Un esempio di dataset che utilizza questa struttura è riportato in figura A.19, in cui è chiara l'importanza del termine all'interno del flusso di dati. Per fare in modo che i modelli catturino correttamente questa dipendenza, come in un caso reale, sono stati inseriti nei dataset utilizzati dai modelli i primi due ritardi della media giornaliera della variabile risposta (se non viene fatto tutti i modelli risentiranno in modo analogo della problematica, come osservato nel capitolo 4, relativo al *concept drift*). I risultati sono riportati nelle figure A.10 - A.13. Questo è un processo semplice, che tutti i modelli riescono a catturare correttamente.

- Il secondo è un processo ARMA(2,2), più difficile da catturare sulla base dei primi ritardi, e permette quindi di complicare la dipendenza temporale tra passato e futuro. Il processo scelto è definito dalla seguente equazione:

$$(1 - \phi_1 B - \phi_2 B^2)u_i = (1 + \theta_1 B + \theta_2 B^2)\eta_i$$

dove $(\phi_1, \phi_2) = (0.5, 0.3)$ e $(\theta_1, \theta_2) = (0.8, 0.5)$, stazionario ed invertibile. La scelta dei valori permette di avere una dipendenza temporale più

complessa da catturare sulla base dei soli ritardi (il grafico dell'autocorrelazione parziale è riportato in figura A.22); per enfatizzare questo aspetto, nella modellazione viene concesso ai modelli solamente il primo ritardo della media giornaliera (figura A.20, un esempio di dataset).

3. Una direzione differente per complicare la dipendenza tra passato e futuro consiste nel simulare un processo stazionario autoregressivo non lineare (processi NAR; Jones, 1978). Per semplicità è stato scelto un processo di tipo SETAR (Self-Exciting Threshold Autoregressive; Tong, 1983) a due regimi, per il quale il controllo della stazionarietà risulta molto semplice. Il processo è definito tramite la seguente formula:

$$\begin{cases} u_i = 0.3u_{i-1} + 0.1u_{i-2} + \eta_i & u_{i-1} < 1 \\ u_i = 0.8u_{i-1} + 0.15u_{i-2} + \eta_i & u_{i-1} \geq 1 \end{cases}$$

e ciò permette di definire il valore al giorno i come funzione non lineare dei due giorni precedenti, costruita affiancando due funzioni lineari (la figura A.25 mostra la relazione tra le medie giornaliere). Questo è un SETAR a due regimi: sulla base del valore al giorno precedente viene utilizzata una specifica regola tra due, mentre il valore di soglia scelto (1), che determina il regime, permette ad entrambi di comparire nel processo in egual misura. Il processo così ottenuto è stazionario, condizione per la quale è sufficiente che la somma dei coefficienti, in valore assoluto, di ciascun regime sia minore di 1. Un esempio di dataset è riportato in figura A.24. I dataset su cui sono stati stimati i modelli contengono i primi due ritardi della media giornaliera.

Iniziando dal primo caso, come osservato nel caso delle relazioni cubiche, le logiche basate sugli alberi risentono del problema delle osservazioni esterne all'intervallo osservato durante la stima, e ciò si può osservare fin da subito dalla qualità iniziale (figura A.10); questi modelli hanno prestazioni inferiori, soprattutto il *gradient boosting*, che nelle simulazioni precedenti raggiungeva gli stessi livelli degli altri modelli.

Dal punto di vista della stabilità a medio/lungo termine nessun modello tende a *degradare* (figura A.11), e ciò è atteso, considerando che il processo non

evolve; mentre in termini di variabilità dell'errore relativo (la misura più importante in assenza di decadimento) *ridge* e *rete neurale* sono i migliori, ma è presente una grossa differenza tra *foresta casuale* e *gradient boosting*: il primo mostra risultati migliori, anche se il livello iniziale è minore, in diretto contrasto a quanto osservato fino a questo punto (le simulazioni fanno parte del capitolo 3). Non solo, la stabilità è la stessa di *ridge* e *reti*, anche se le differenze iniziali sono elevate.

Osservando i tracciati dell'errore dei modelli, stimati sui singoli sottoinsiemi di stima, è però subito chiaro che questi possono essere raggruppati: *rete neurale* e *ridge*, *foresta casuale* e *gradient boosting* (figura A.14). I due modelli, basati su logiche simili, soffrono perdite di prestazioni negli stessi periodi. In alcuni casi però, come quello riportato nel grafico, calcolando l'errore relativo, dividendo per l'errore iniziale, il livello del *gradient boosting* rimane più alto (figura A.15). Sembra quindi che la similitudine tra i due modelli, che emerge osservando gli errore non-relativi, sia nascosta, stavolta, dalla diversa qualità iniziale. Concentrandoci sull'*errore quadratrico medio*, la situazione è comunque molto chiara: i modelli basati sugli alberi hanno prestazioni inferiori, e quindi non andrebbero scelti (figura A.23).

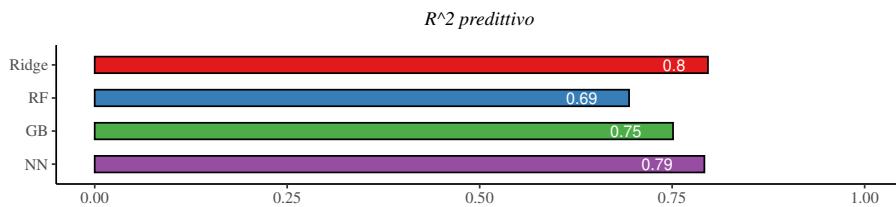


Figura A.10: Prestazioni iniziali medie dei modelli in presenza di un processo AR(2). Le differenze in qualità iniziali sono maggiori, soprattutto per i modelli basati sugli alberi. Questi modelli soffrono maggiormente quando operano all'esterno dell'intervallo osservato.

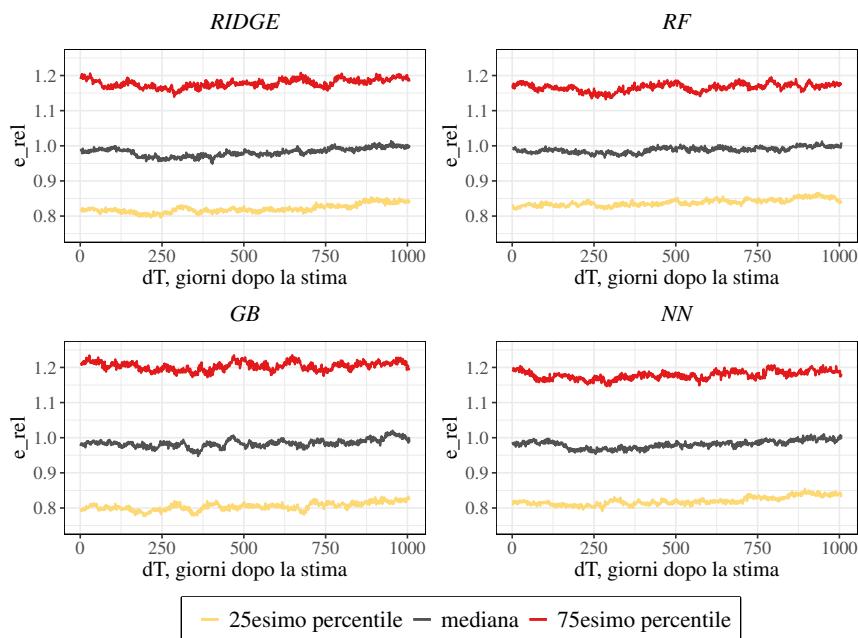


Figura A.11: Combinazione dei grafici di *AI Aging*, relativamente al caso dell'AR(2). I livelli mediani di errore mostrano un'assenza di trend. La tendenza complessiva, intesa come la forma dei grafici, coincide.

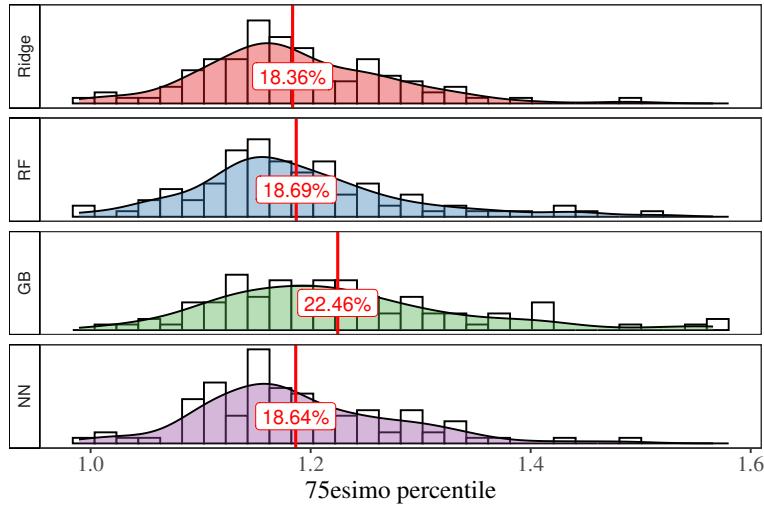


Figura A.12: Distribuzione del livello medio del 75esimo percentile nelle replicazioni della simulazione (AR(2)). Il *gradient boosting* si presenta come il meno stabile, nonostante la qualità iniziale sia maggiore di quella della *foresta casuale*.

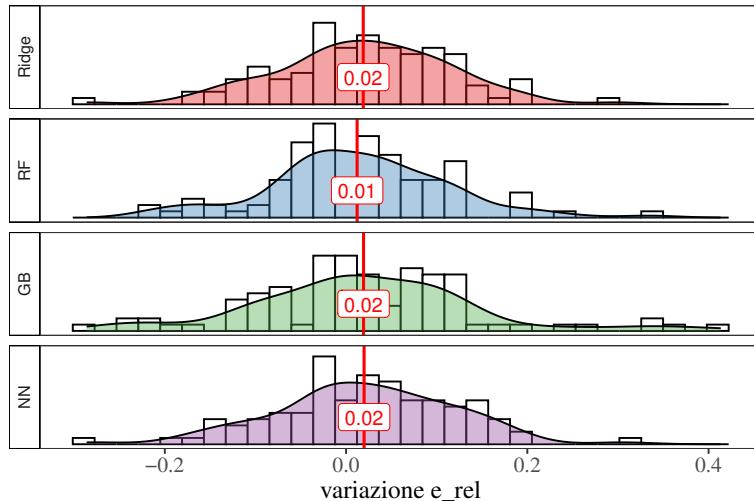


Figura A.13: Distribuzione della variazione dell'errore relativo nelle replicazioni della simulazione (AR(2)). I livelli medi di inclinazione sono gli stessi, con media attorno a 0, indicando assenza di *degradazione temporale*.

Per verificare se le differenze iniziali possono spiegare la diversa stabilità di

foresta e *GB*, nelle simulazioni successive il numero di variabili esplicative è stato ridotto a sette, in modo da permettere ai modelli con minore qualità un adattamento migliore.

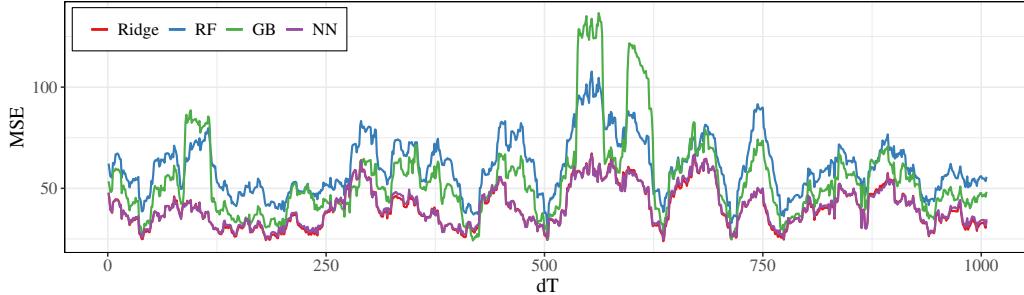


Figura A.14: Tracciati dell'errore ($MSE(dT)$) dei modelli (replicazione 31, insieme di stima 15). L'andamento dell'*errore quadratico medio*, non-relativo, è totalmente sovrapposto per *NN* e *ridge*, e molto simile per *RF* e *GB*.

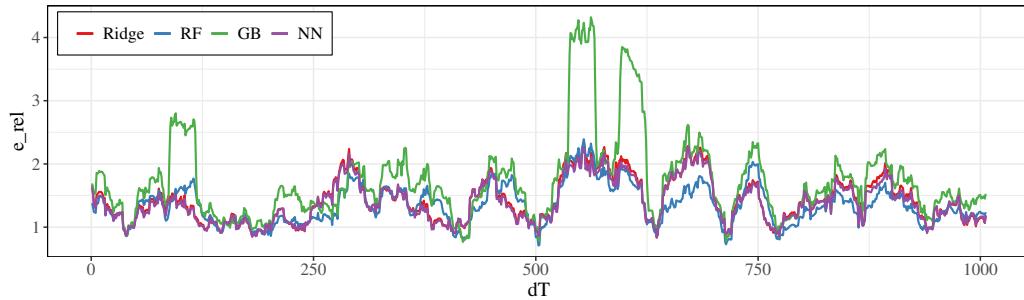


Figura A.15: Tracciati dell'errore relativo dei modelli (replicazione 31, insieme di stima 15). La sovrapposizione tra *RF* e *GB* scompare, a causa della diversa qualità iniziale.

I risultati, relativi alla variabilità dell'errore relativo negli altri casi considerati, sono riportati nelle figure A.16 e A.17. I grafici mostrano quanto atteso:

1. In presenza del processo ARMA(2,2) le considerazioni sono le stesse fatte per il caso del processo AR(2), con la differenza che la stabilità della *foresta* è molto più vicina a quella del *gradient boosting*, probabilmente grazie al miglioramento nei livelli iniziali.

2. Nel caso del *SETAR* il modello più stabile è la *rete neurale*, mentre gli altri presentano livelli di stabilità simili; a differenza dello stimatore *ridge*, questo riesce ad estrapolare in modo adeguato in presenza della relazione non lineare tra passato e futuro. È stato osservato infatti, in Xu et al. (2020), che in caso di estrapolazione le previsioni delle *reti neurali*, con la stessa architettura usata in questo lavoro, convergono ad una funzione lineare, e ciò funziona bene nel caso di un processo *SETAR*.

È interessante osservare come la stabilità dello stimatore *ridge* e quella della *rete neurale* sia la stessa, nei primi due casi, nonostante l'estrapolazione del secondo sia probabilmente meno affidabile.

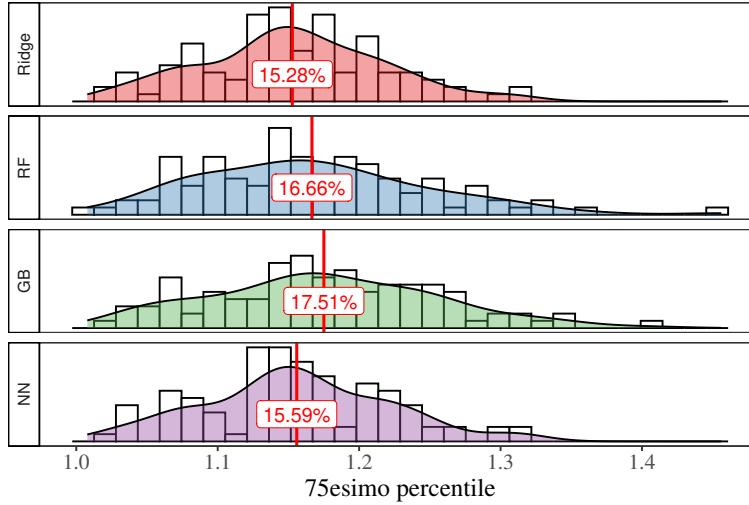


Figura A.16: Distribuzione del livello medio del 75esimo percentile, nella simulazione relativa al processo ARMA(2,2). *Forest casuale* e *gradient boosting* mostrano maggiore instabilità rispetto agli altri due modelli, e la similitudine tra i due modelli è ora più chiara, visti i livelli iniziali più simili. Questi sono riportati di seguito: *ridge* = 0.85, *RF* = 0.8, *GB* = 0.84, *NN* = 0.85.

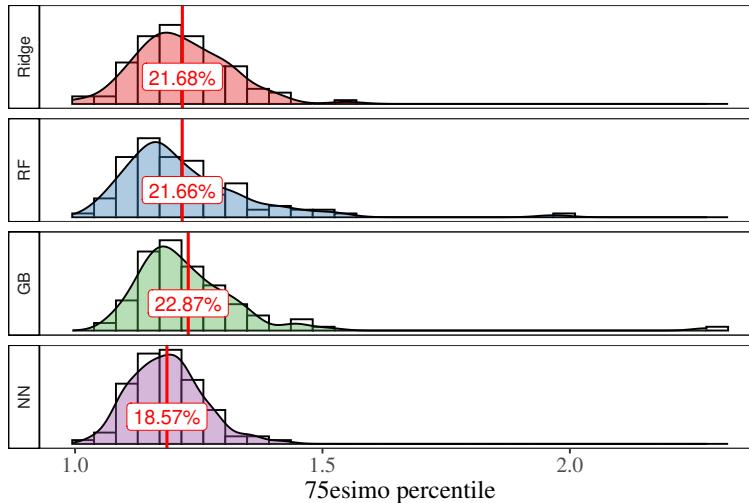


Figura A.17: Distribuzione del livello medio del 75esimo percentile, nella simulazione relativa al processo SETAR. La qualità iniziale dei modelli è riportata di seguito: *ridge* = 0.76, *RF* = 0.68, *GB* = 0.72, *NN* = 0.77.

Processi non stazionari

In questa sezione viene presentato il caso in cui viene inserito un processo non stazionario (in media) all'interno del flusso di dati, comportando la presenza di un trend lineare. Il processo è di tipo ARIMA(0,1,0) e i risultati sono riportati in figura A.18.

Il modello che invecchia meno è sicuramente lo stimatore *ridge*, molto più stabile della *rete neurale*. A differenza del primo, la *rete* presenta un andamento mediano con trend minimo ma un forte aumento della variabilità, indicando come possa estrapolare correttamente e seguire il trend, ma che sia meno affidabile nel farlo.

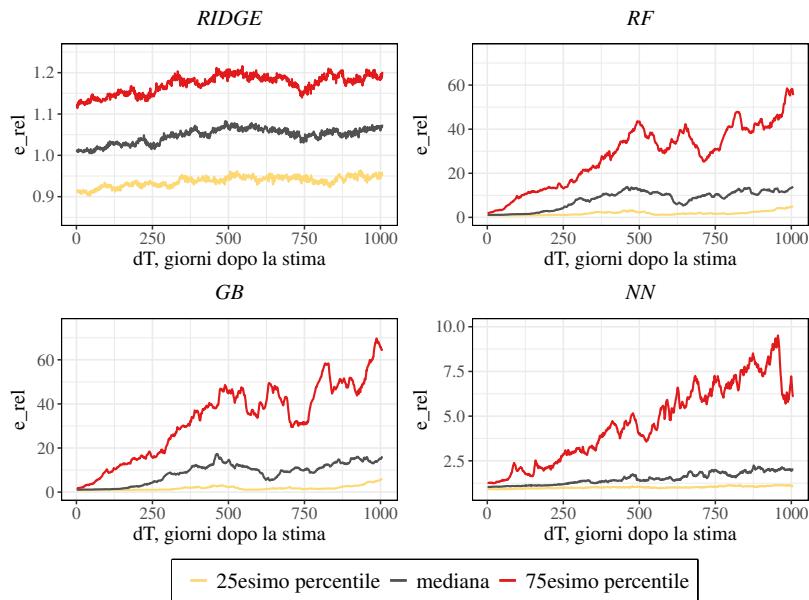


Figura A.18: *Degradazione temporale* dei modelli in presenza di trend lineare. Le prestazioni di tutti i modelli presentano segni di *degradazione*, con lo stimatore *ridge* che rimane il più stabile. Il comportamento di *RF* e *GB* è sostanzialmente identico, soffrendo dello stesso problema nell'estrapolazione. I livelli del *GB* sono più alti: come osservato in precedenza la diversa qualità iniziale nasconde in parte questo tipo di errore. Questa simulazione è stata condotta con 50 replicazioni, in quanto i comportamenti sono sufficientemente definiti.

Figure aggiuntive

Alcune figure aggiuntive relative alle prove condotte con processi autocorrelati:

1. In figura A.19 è riportato un esempio di dataset contenente il processo AR(2). L'importanza di questa componente è immediatamente chiara.
2. In figura A.20 è riportato un esempio di dataset contenente il processo ARMA(2,2). L'importanza di questa componente è immediatamente chiara.
3. In figura A.21 è riportato un esempio di dataset contenente il processo ARIMA(0,1,0).
4. In figura A.22 è riportato il grafico di autocorrelazione parziale del processo ARMA(2,2), che mostra come sarebbero necessari un gran numero di ritardi per catturare completamente la dipendenza temporale.
5. In figura A.23 è riportata la distribuzione del livello medio del terzo quartile dell'*MSE* (errore non-relativo), per mostrare la similitudine tra *RF* e *GB*, che dai risultati “relativi” veniva nascosta.
6. In figura A.24 è riportato un esempio di dataset contenente il processo SETAR, in cui è chiara la presenza dei due regimi.
7. In figura A.25 è riportato il lag-plot del processo SETAR, che mostra la relazione non lineare tra un giorno e il successivo.

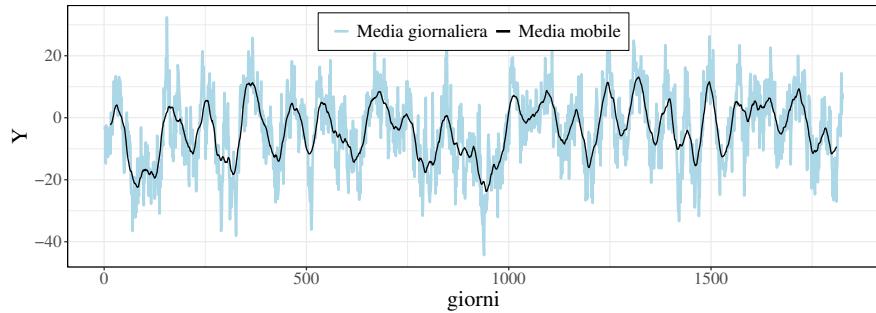


Figura A.19: Esempio di dataset contenente un processo AR(2) nella componente non osservata.

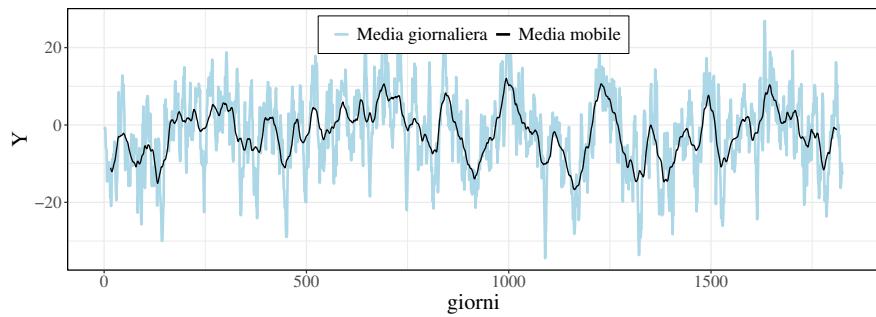


Figura A.20: Esempio di dataset contenente il processo ARMA(2,2) descritto, che provoca importanti oscillazioni nei valori della variabile risposta.

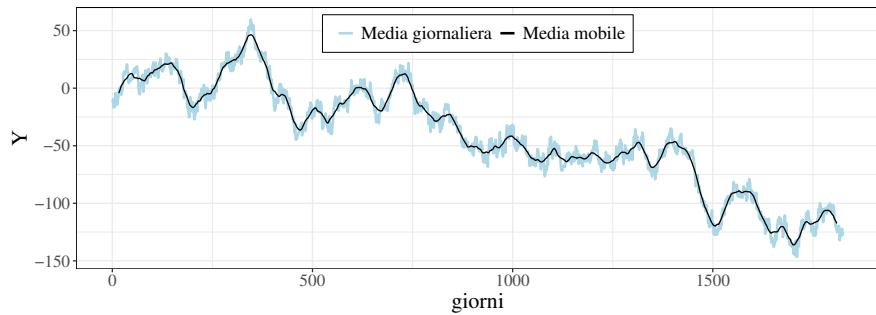


Figura A.21: Esempio di dataset contenente un processo ARIMA(0,1,0) nella componente non osservata; il trend è facilmente visibile.

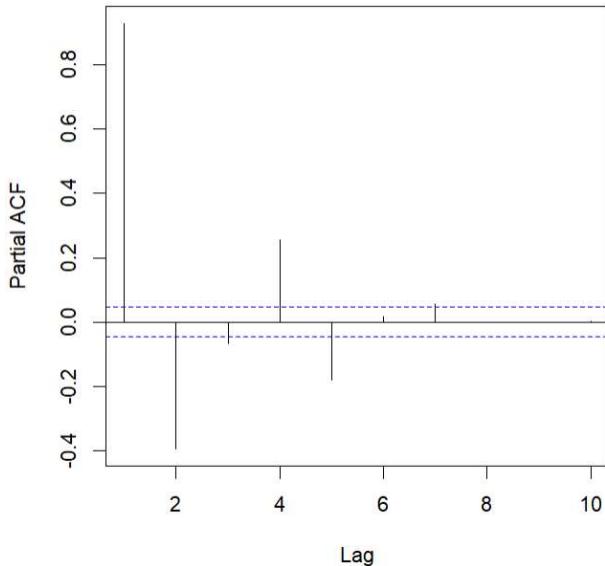


Figura A.22: Autocorrelazione parziale del processo ARMA(2,2). I ritardi sono significativi fino al quinto, rendendo meno efficace l'utilizzo dei soli ritardi della media giornaliera. Sulla base di questi valori dovrebbero esserne utilizzati 5 o 7.

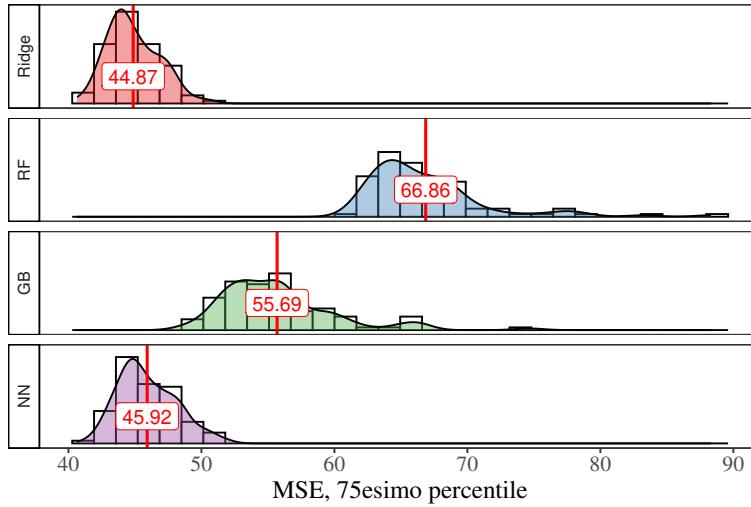


Figura A.23: Distribuzione del livello medio del 75esimo percentile, riferito all'errore quadratrico medio, non relativo, nel caso del processo AR(2). I modelli basati sugli alberi mostrano le stesse lunghe code a destra, indicando un comportamento complessivo simile.

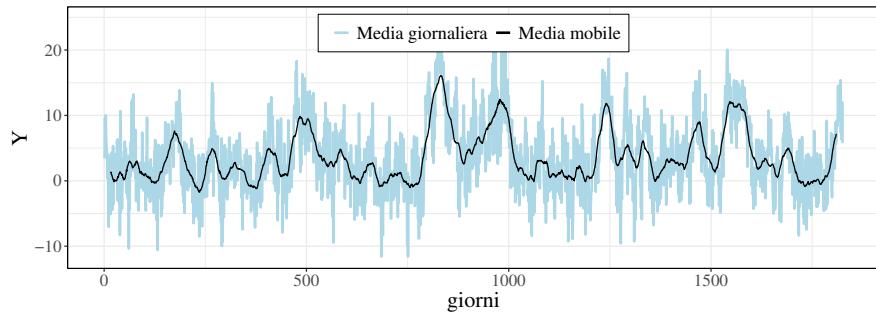


Figura A.24: Esempio di dataset contenente il processo autoregressivo non lineare (SETAR) descritto, in cui è chiara la presenza di due regimi.

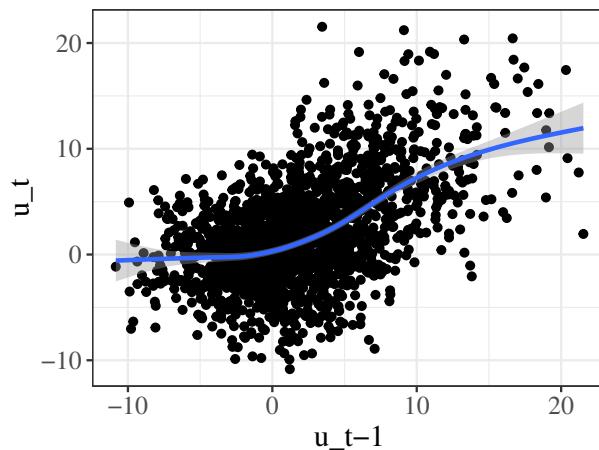


Figura A.25: Lag-plot del processo SETAR, che mostra una relazione non lineare con i valori precedenti. Sono chiari i due regimi lineari con cui la relazione è stata costruita.

Appendice B

Materiale capitolo 4

Data drift

Il contributo delle simulazioni, nel caso di *data drift*, è piuttosto limitato: ne sono state condotte alcune per confermare quanto discusso, relativamente alla relazione tra l'adattamento di un modello in una sezione specifica dell'insieme di stima e la *degradazione temporale*. Per limitare le conseguenze dell'estrapolazione e simulare *data drift* le osservazioni sono state generate da due possibili distribuzioni normali: una principale, centrata in zero, con una probabilità maggiore di essere utilizzata, e una seconda, con una diversa media, la cui probabilità di utilizzo è inizialmente più bassa, ma aumenta nel tempo. Le varianze delle singole variabili sono pari ad 1, e la correlazione è generata tramite la solita metodologia. I due casi presi in considerazione sono i seguenti:

1. Il primo, caratterizzato da relazioni cubiche. La probabilità iniziale di osservare la seconda distribuzione, in cui metà delle variabili sono centrate in 1.5, è pari a 0, ed aumenta linearmente fino a 0.8. L'insieme di stima meno recente contiene comunque il 7% di osservazioni generate dalla seconda distribuzione.
2. Il secondo, caratterizzato da effetti di interazione tra le X . La probabilità iniziale di generare osservazioni dalla seconda distribuzione, in cui metà delle variabili sono centrate in 2, è pari a 0.05 (più di 500 osser-

vazioni per insieme di stima), e rimane costante nei primi due anni, in cui i modelli vengono stimati. Successivamente, aumenta linearmente fino a 0.8. Questa correzione, rispetto al primo caso, permette di osservare meglio e con più precisione l'effetto del *data drift* dal grafico di *AI Aging*.

La lontananza della seconda distribuzione dalla prima, e la probabilità di utilizzo, determinano un *data drift* più o meno intenso. Queste scelte influiscono sulla forma complessiva osservata per i grafici di *AI Aging*, ma non sulla sostanza dei risultati, riportati nelle figure B.1 e B.2.

In entrambe le situazioni tutti i modelli risentono della sparsità iniziale delle regioni, mostrando segni di *degradazione temporale*, di diversa entità. Come atteso i modelli che risentono maggiormente del *data drift* sono lo stimatore *ridge*, che propone una struttura incompatibile con la relazione, e la *foresta casuale*. *Gradient boosting* e *rete neurale* presentano invece *degradazione temporale* della stessa entità (la maggiore ampiezza della banda di variabilità, con relazioni cubiche, è dovuta alla peggiore estrapolazione, come già introdotto nella sezione 3). *Gradient boosting* e *foresta casuale*, nonostante soffrano in modo simile per l'estrapolazione, mostrano invecchiamenti molto diversi: il secondo soffre maggiormente la sparsità, come osservato nel capitolo 3, e quindi maggiormente il *data drift*.

Per approfondire quale dei due modelli (tra *GB* e *NN*) soffra maggiormente del problema sono state condotte ulteriori prove, utilizzando la stessa procedura impiegata nel caso di interazioni, variando la sparsità iniziale e allontanando la seconda distribuzione, spostandone ulteriormente la media (metà in 3). I risultati sono riportati nella figura B.3 per il caso di interazioni e in figura B.4 per le relazioni cubiche.

Il *gradient boosting* risente del *data drift* in misura maggiore della *rete neurale*, nei casi valutati; sia con interazioni, che con relazioni cubiche, l'inclinazione della mediana è maggiore. Nei casi migliori (B.1 e B.2) la *degradazione* è al pari, nei peggiori è molto maggiore (B.3 e B.4). Nonostante sia stato limitato il problema dell'estrapolazione, questa è probabilmente la ragione

della differenza, vista la stessa qualità iniziale dei due modelli.

Una considerazione importante riguarda il livello iniziale dei modelli, che non va a costituire, in questo caso, un buon indicatore per prevedere la *degradazione*. Nelle ultime simulazioni condotte (figure B.3 e B.4) la qualità iniziale di *gradient boosting* e *rete neurale* è identica, o addirittura la prima è maggiore, ma la *degradazione* è molto differente. In parte ciò può essere spiegato dal fatto che le poche osservazioni generate dalla seconda distribuzione contribuiscono poco alla determinazione dell'indice iniziale, che non andrà a fornire una buona indicazione dell'adattamento nell'area; e in parte dal fatto che l'estrapolazione ha un effetto maggiore sul *gradient boosting*, ed è più facile avere osservazioni estreme nella parte finale del dataset.

Scegliere il miglior modello sulla base della qualità iniziale, in questo caso, non corrisponde necessariamente alla scelta migliore a medio/lungo termine. La *rete neurale*, nelle prove effettuate, è risultata essere il modello migliore, mentre il *gradient boosting* ha presentato, nei casi migliori (figure B.1 e B.2), la stessa *degradazione*; la differenza è però dovuta all'estrapolazione.
In situazioni meno complesse, con relazioni lineari, lo stimatore *ridge* presenterà un comportamento migliore.

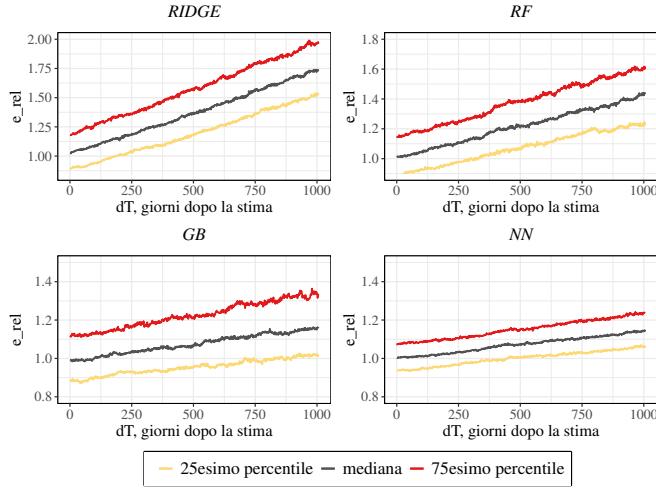


Figura B.1: Combinazione dei grafici di *AI Aging*, nella simulazioni con relazioni cubiche. Tutti i modelli presentano *degradazione temporale*. Qualità iniziale: *ridge* = 0.58, *RF* = 0.88, *GB* = 0.92 e *NN* = 0.92. Le pendenze medie, solitamente riportate nei grafici dedicati, di *GB* e *NN* sono pari a 0.18 e 0.15, indicando una tendenza del primo a *degradare* leggermente di più. 50 replicazioni.

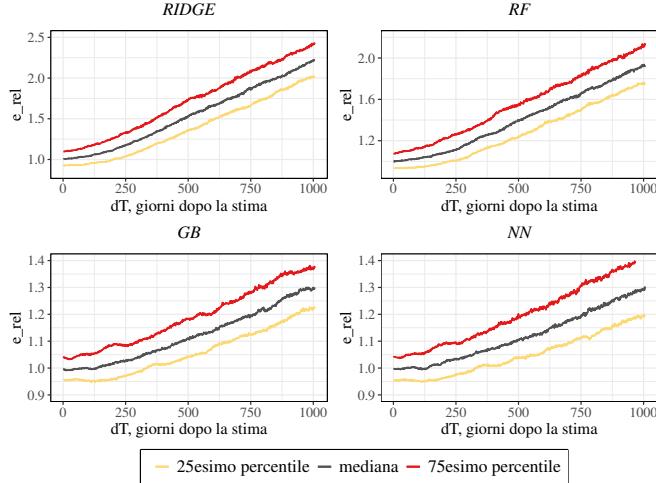


Figura B.2: Combinazione dei grafici di *AI Aging*, nelle simulazioni con effetti di interazione. Tutti i modelli presentano *degradazione temporale*. Qualità iniziale: *ridge* = 0.58, *RF* = 0.8, *GB* = 0.91 e *NN* = 0.91. L'inclinazione media della mediana: *GB* = 0.33, *NN* = 0.32, sostanzialmente identica. 50 replicazioni.

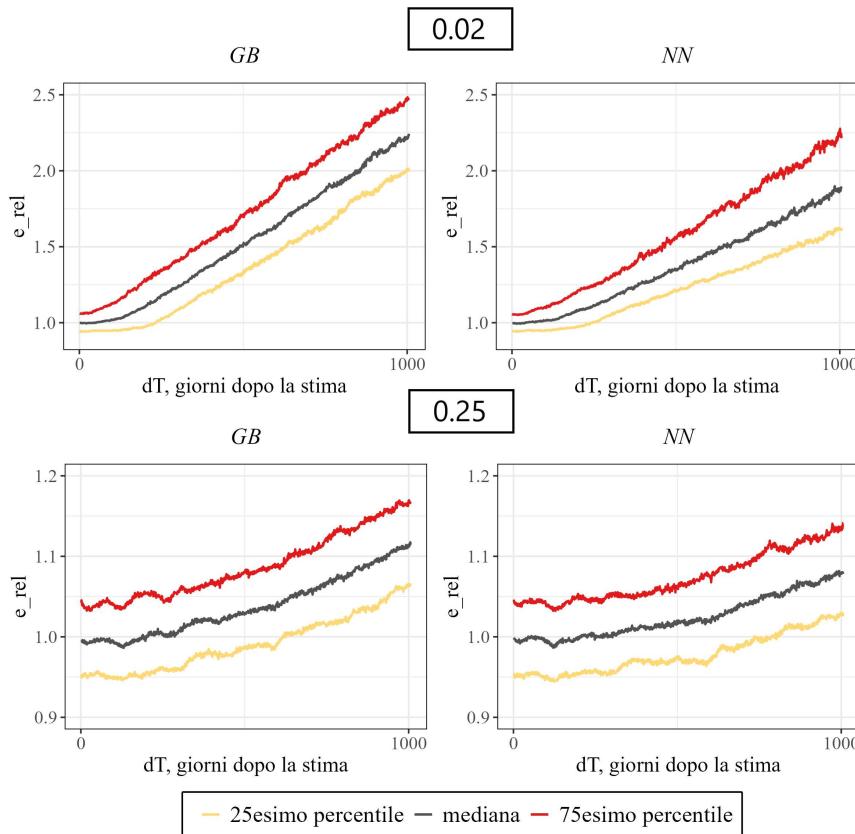


Figura B.3: Combinazione dei grafici di *AI Aging*, effetti di interazione, 50 repliche. La probabilità di osservare la seconda distribuzione nel primo caso è pari al 2% (circa 200 osservazioni per dataset di stima), mentre nel secondo caso è pari al 25%. La qualità iniziale è pari a 0.91 per entrambi i modelli nel primo caso, mentre 0.96 nel secondo, più alta per via della seconda distribuzione, che genera valori più grandi della variabile risposta, rendendo la componente non osservata meno rilevante.

Entrambi i modelli presentano *degradazione temporale*, il *gradient boosting* in misura maggiore.

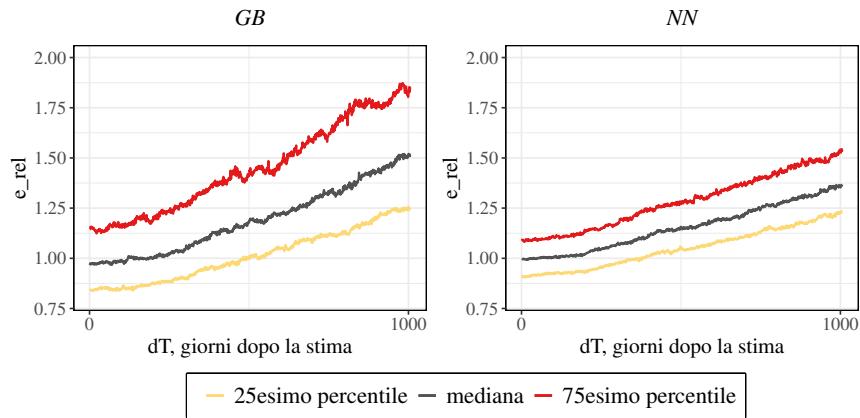


Figura B.4: Combinazione dei grafici di *AI Aging*, relazioni cubiche, 50 replicazioni. La probabilità di osservare la seconda distribuzione è pari al 20% (più di 2000 osservazioni per dataset di stima). La qualità iniziale dei due modelli è pari a 0.97 per il *GB* e 0.96 per la *NN*. Entrambi i modelli presentano *degradazione temporale*, che risulta maggiore per il *gradient boosting*. Non solo, la variabilità dell'errore è molto maggiore per quest'ultimo.

Real concept drift

Concept drift graduale - relazioni lineari

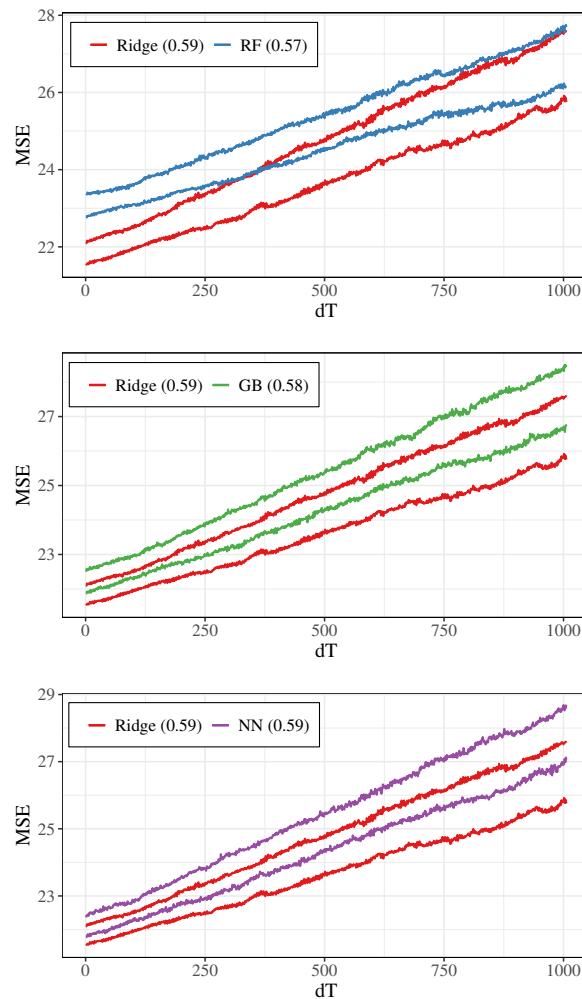


Figura B.5: Andamenti degli *MSE* mediani dei modelli in presenza di *drift graduale*, con 4 variabili esplicative. Le conclusioni sono simili a quelle già riportate: il modello peggiore, la *foresta casuale*, presenta un aumento minore dell'errore. Gli altri modelli, inizialmente simili, presentano un aumento simile dell'errore.

Concept drift graduale - relazioni non lineari

In modo analogo a quanto fatto per il caso lineare sono state condotte delle simulazioni con relazioni più complesse tra le X e la Y :

1. In un primo caso è stato simulato *concept drift graduale*, che coinvolge metà dei coefficienti, in una situazione caratterizzata dalla presenza di effetti di interazione (8 variabili e 8 interazioni di primo ordine) e 30 osservazioni al giorno. Le specifiche dell'approccio sono le stesse di quelle utilizzate nel caso lineare. Gli andamenti degli *MSE* mediani sono riportati nella figura B.6. In aggiunta, la simulazione è stata ripetuta riducendo il numero di osservazioni, in modo da ottenere un adattamento minore dei modelli (risultati in figura B.7).
2. In un secondo caso è stato simulato *concept drift graduale*, con le stesse specifiche, in una situazione caratterizzata da relazioni cubiche (risultati in figura B.8), e come già fatto il numero di osservazioni è stato successivamente ridotto (risultati in figura B.9).

Tutte le simulazioni indicate sono state condotte con 50 replicazioni.

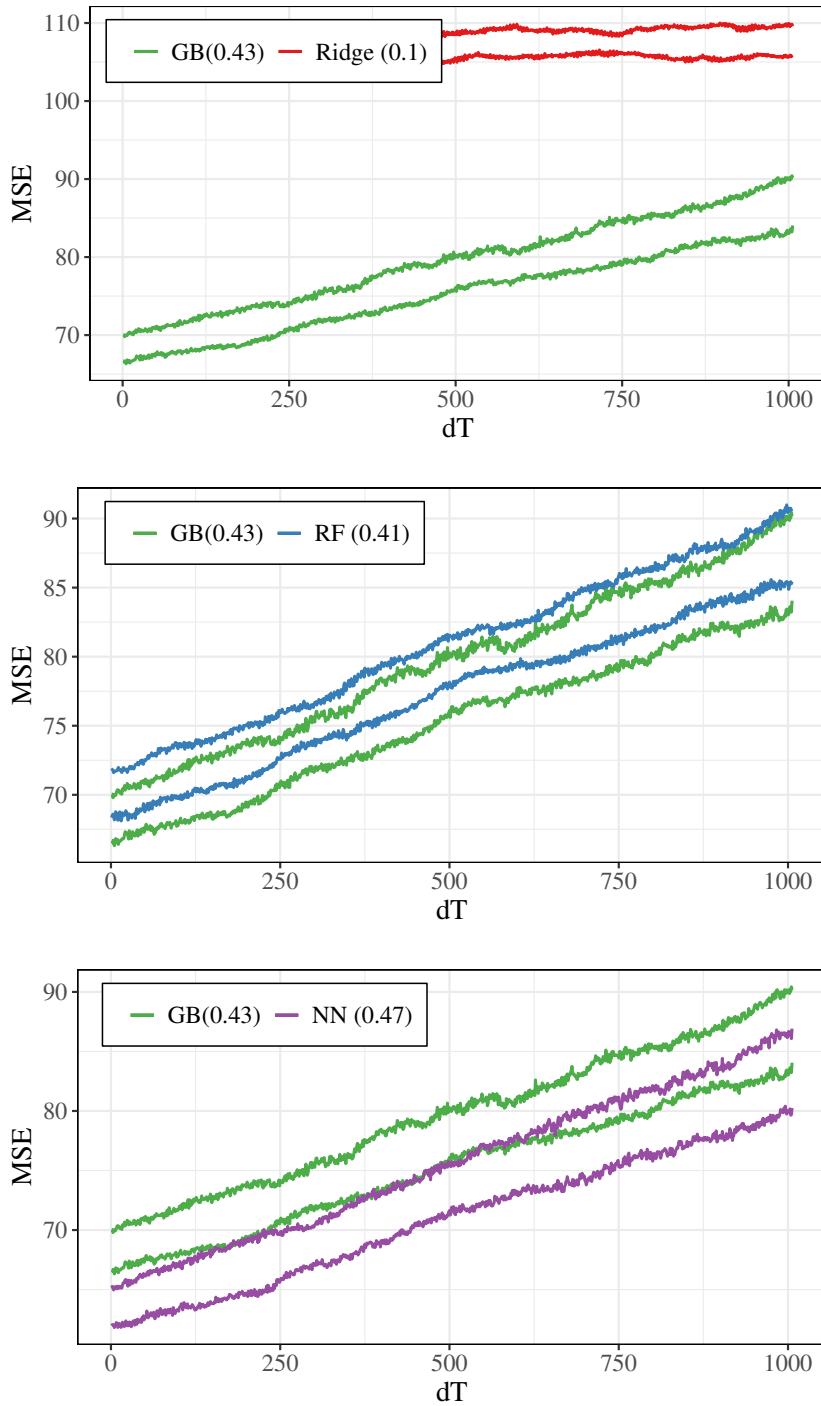


Figura B.6: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, con 8 variabili, 30 osservazioni al giorno e 8 interazioni di primo ordine. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

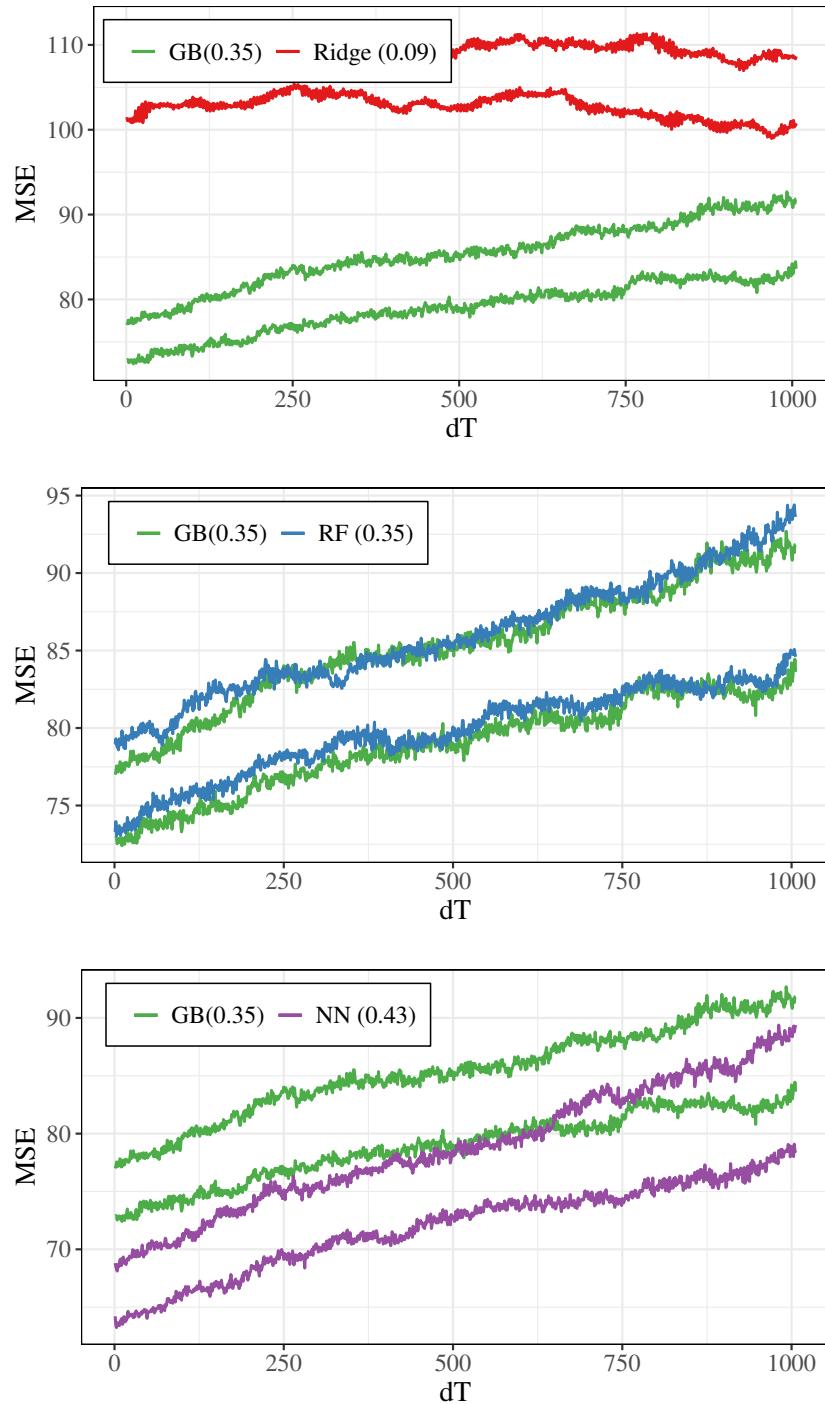


Figura B.7: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, con 8 variabili, 5 osservazioni al giorno e 8 interazioni di primo ordine. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

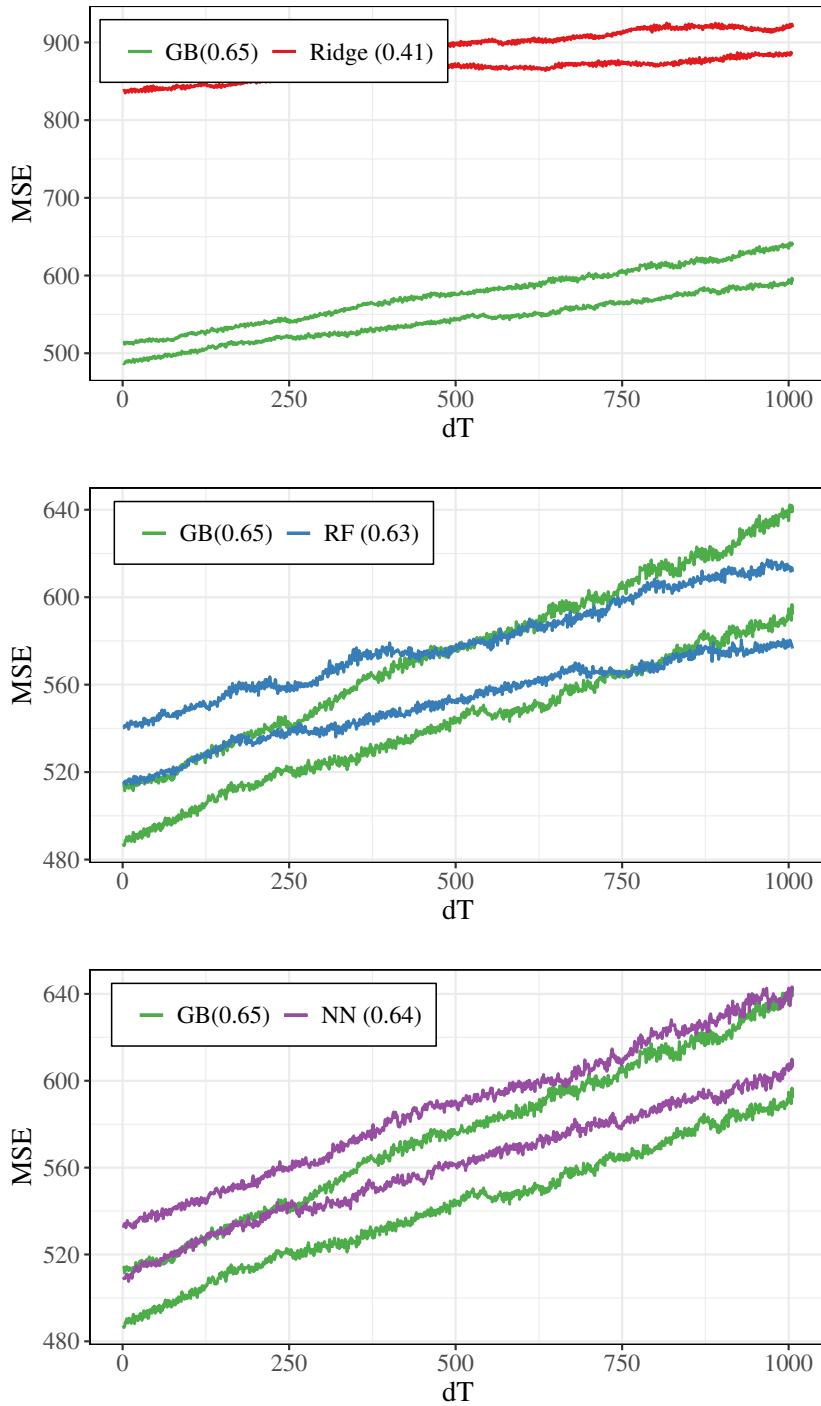


Figura B.8: Andamenti degli MSE mediani dei modelli in presenza di *drift graduale*, con 10 variabili, 30 osservazioni al giorno e relazioni cubiche. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

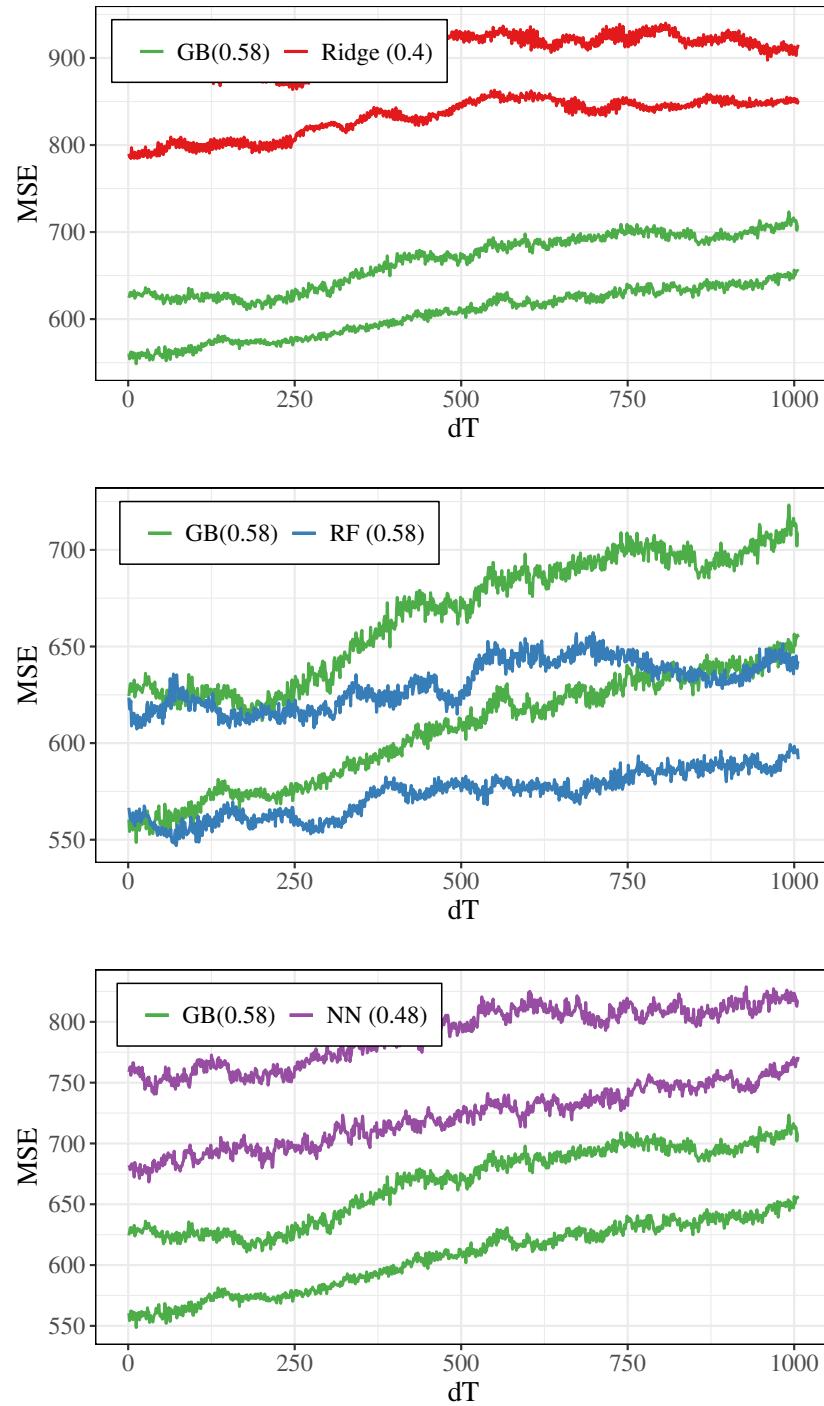


Figura B.9: Andamenti degli *MSE* mediani dei modelli in presenza di *drift graduale*, con 10 variabili, 5 osservazioni al giorno e relazioni cubiche. La *foresta casuale* risente del *drift* in misura minore, pur presentando un livello iniziale più alto.

Concept drift incrementale - relazioni lineari

La descrizione delle figure contenute in questa sezione:

1. In figura B.10 è riportata la seconda simulazione condotta nel caso lineare, con quattro variabili esplicative, e un differente insieme di coefficienti, iniziali e finali. Come nell'altro caso il comportamento dello stimatore *ridge* si distingue.
2. In figura B.11 è riportata una simulazione condotta nel caso lineare, con 60 variabili esplicative e 10 osservazioni al giorno, per cercare di aumentare il divario iniziale tra i modelli. La relazione tra qualità iniziale e decadimento si mantiene. La simulazione è stata condotta con 50 replicazioni.
3. In figura B.12 la simulazione con effetti lineari, *drift incrementale* e 30 osservazioni al giorno è stata ripetuta, riducendo il numero di alberi utilizzati dal *gradient boosting*. I modelli così ottenuti, con una qualità iniziale minore, soffrono in misura minore del cambio di regole.
4. In figura B.13 il confronto tra la *rete neurale*, in cui il numero di epoch è stato ridotto, con il *GB* con meno alberi (stessi dati della simulazione riportata in figura B.12). Nonostante le *reti* abbiano una qualità complessiva minore non risentono meno del *drift*.

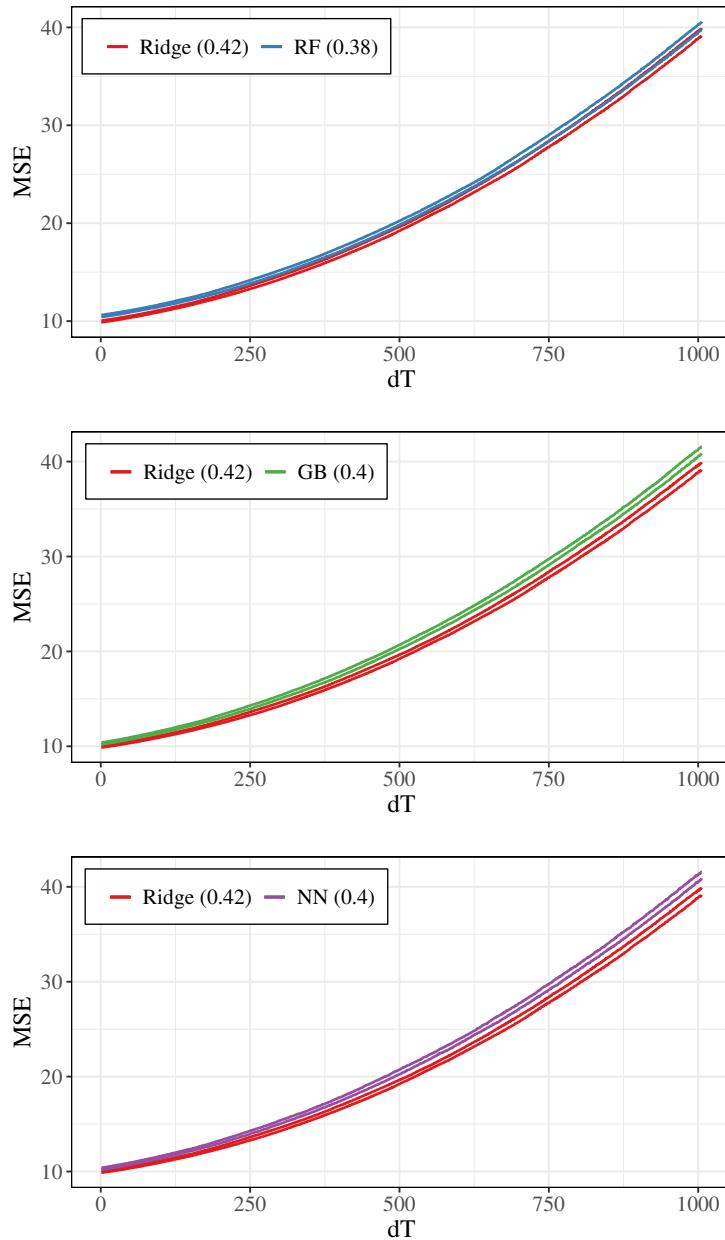


Figura B.10: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 4 variabili esplicative e un insieme di coefficienti, sia iniziali che finali, differente e scelto casualmente. Come nel caso presentato in figura 4.12 (pagina 108), *GB* e *NN* risentono maggiormente del *drift*, anche se presentano una qualità iniziale inferiore.

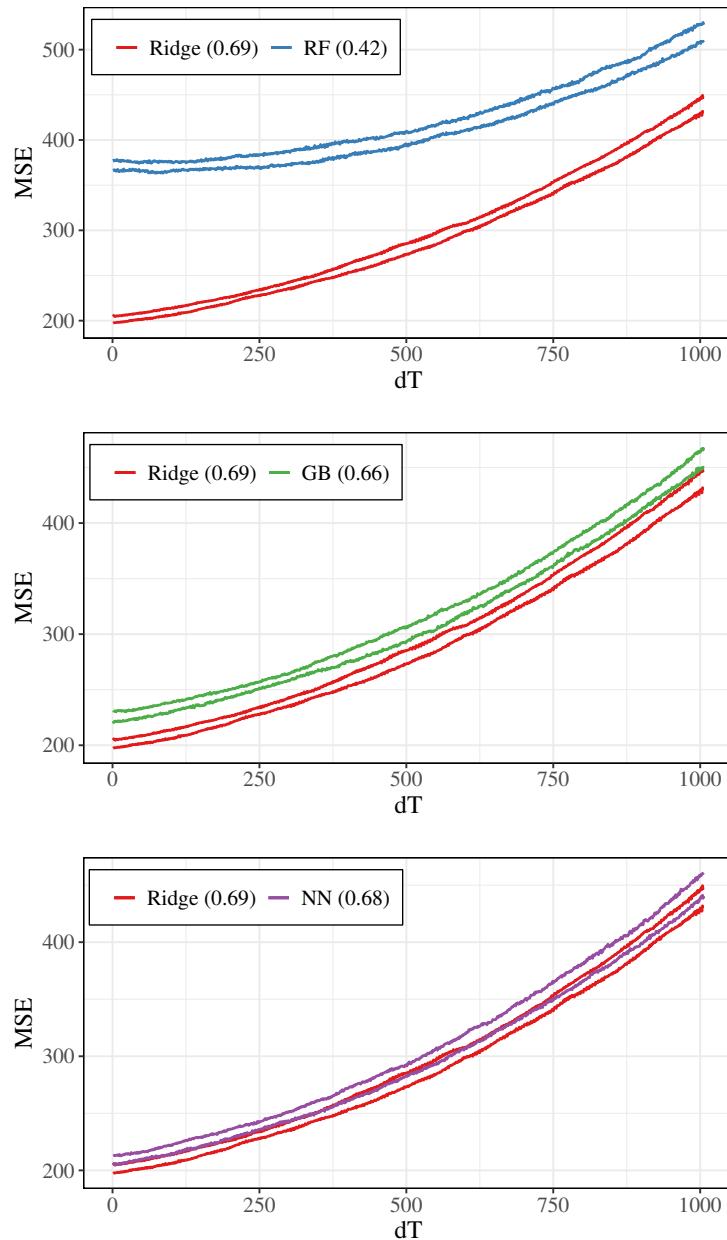


Figura B.11: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 60 variabili esplicative e 10 osservazioni al giorno.

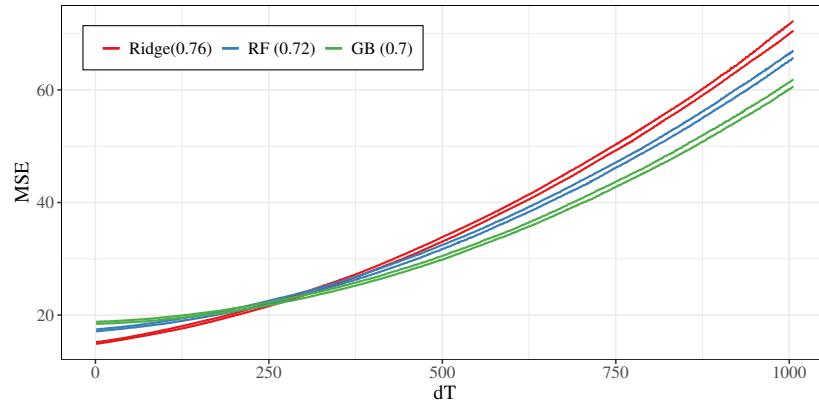


Figura B.12: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 10 variabili esplicative. Il numero di alberi utilizzati dal *GB* è stato ridotto, in modo da ridurre la qualità iniziale. Ciò ha portato il modello a soffrire in misura minore del *concept drift*.

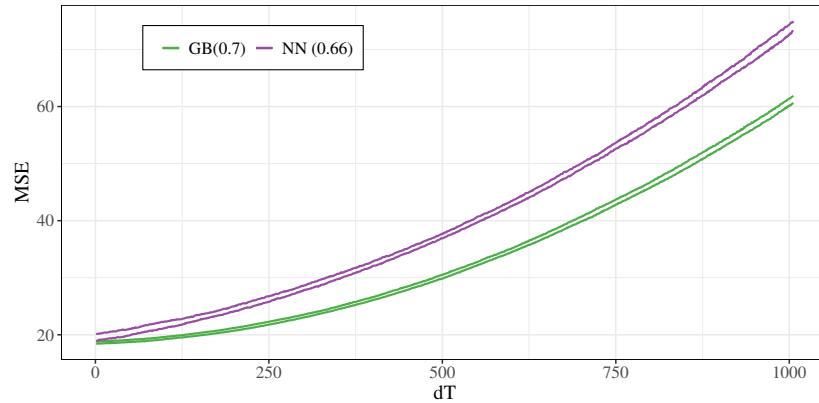


Figura B.13: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 10 variabili esplicative. Il numero di epoche utilizzate per la stima delle *reti neurali* è stato ridotto, e il confronto è fatto con il *gradient boosting*, con un numero di alberi ridotto.

Concept drift incrementale - relazioni non lineari

In modo analogo a quanto fatto per il caso lineare sono state condotte delle simulazioni con relazioni più complesse tra le X e la Y :

1. In un primo caso è stato simulato *concept drift incrementale*, che coinvolge metà dei coefficienti, in una situazione caratterizzata dalla presenza di effetti di interazione (8 variabili e 8 interazioni di primo ordine) e 30 osservazioni al giorno. Le specifiche dell'approccio sono le stesse di quelle utilizzate nel caso lineare. Gli andamenti degli *MSE* mediani sono riportati nella figura B.14. In aggiunta, la simulazione è stata ripetuta riducendo il numero di osservazioni, in modo da ottenere un adattamento minore dei modelli (risultati in figura B.15).
2. In un secondo caso è stato simulato *concept drift incrementale*, con le stesse specifiche, in una situazione caratterizzata da relazioni cubiche (risultati in figura B.16), e come già fatto il numero di osservazioni è stato successivamente ridotto (risultati in figura B.17).

Per verificare se la peculiarità osservata nel comportamento della *foresta casuale* sia strutturale (figura B.17) è stata condotta una seconda simulazione, prima con 30 osservazioni al giorno e poi 5 (sempre con relazioni cubiche), con un differente insieme di coefficienti iniziali e finali; l'irregolarità scompare (figure B.18 e B.19).

Nelle figure B.20 e B.21 sono riportati i risultati dell'applicazione del *gradient boosting*, con un numero di alberi ridotto, ai flussi di dati caratterizzati da relazioni non lineari.

Tutte le simulazioni indicate sono state condotte con 50 replicazioni.

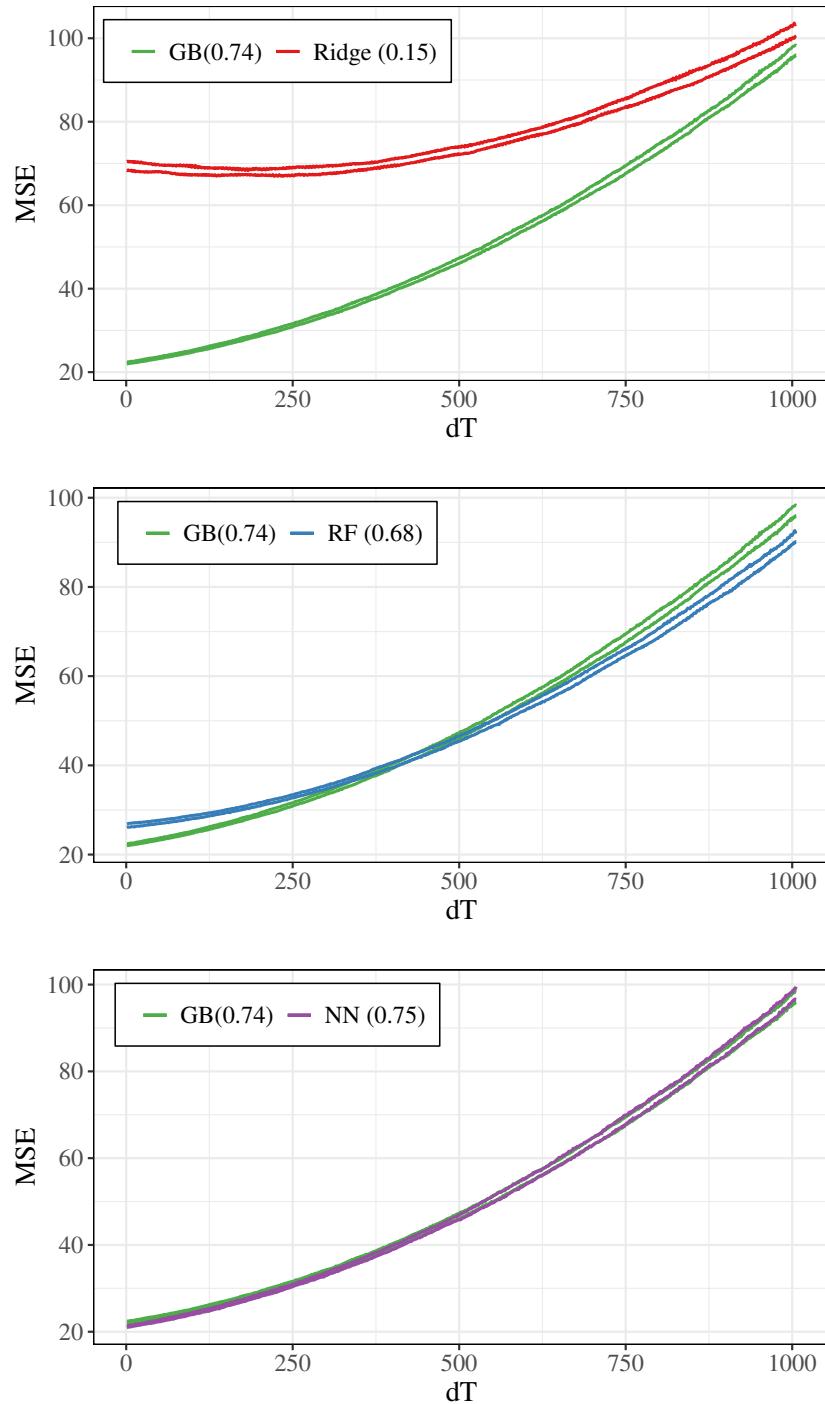


Figura B.14: Andamenti degli *MSE* mediani dei modelli in presenza di *drift incrementale*, con 8 variabili, 30 osservazioni al giorno e 8 interazioni di primo ordine. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

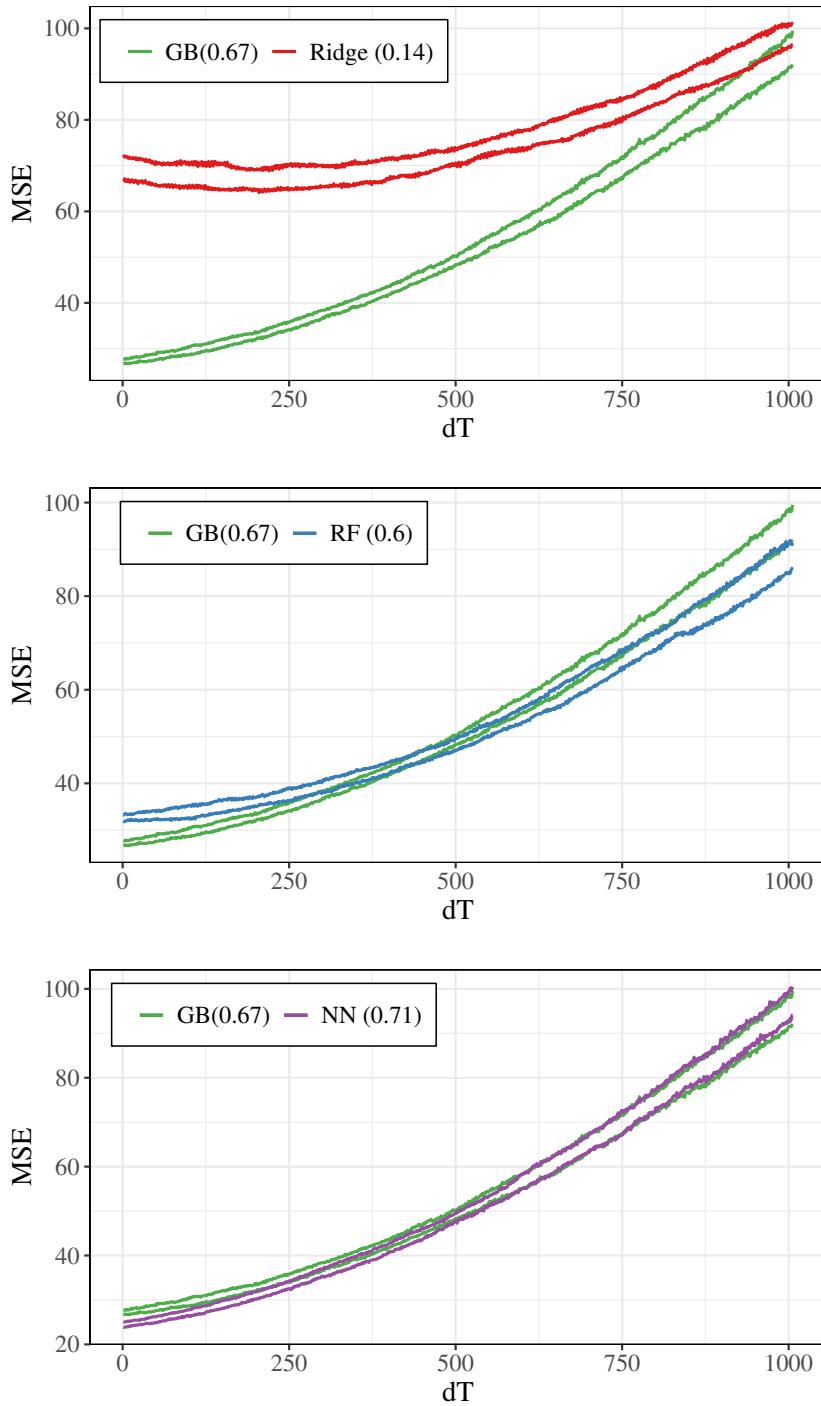


Figura B.15: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 8 variabili, 5 osservazioni al giorno e 8 interazioni di primo ordine. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

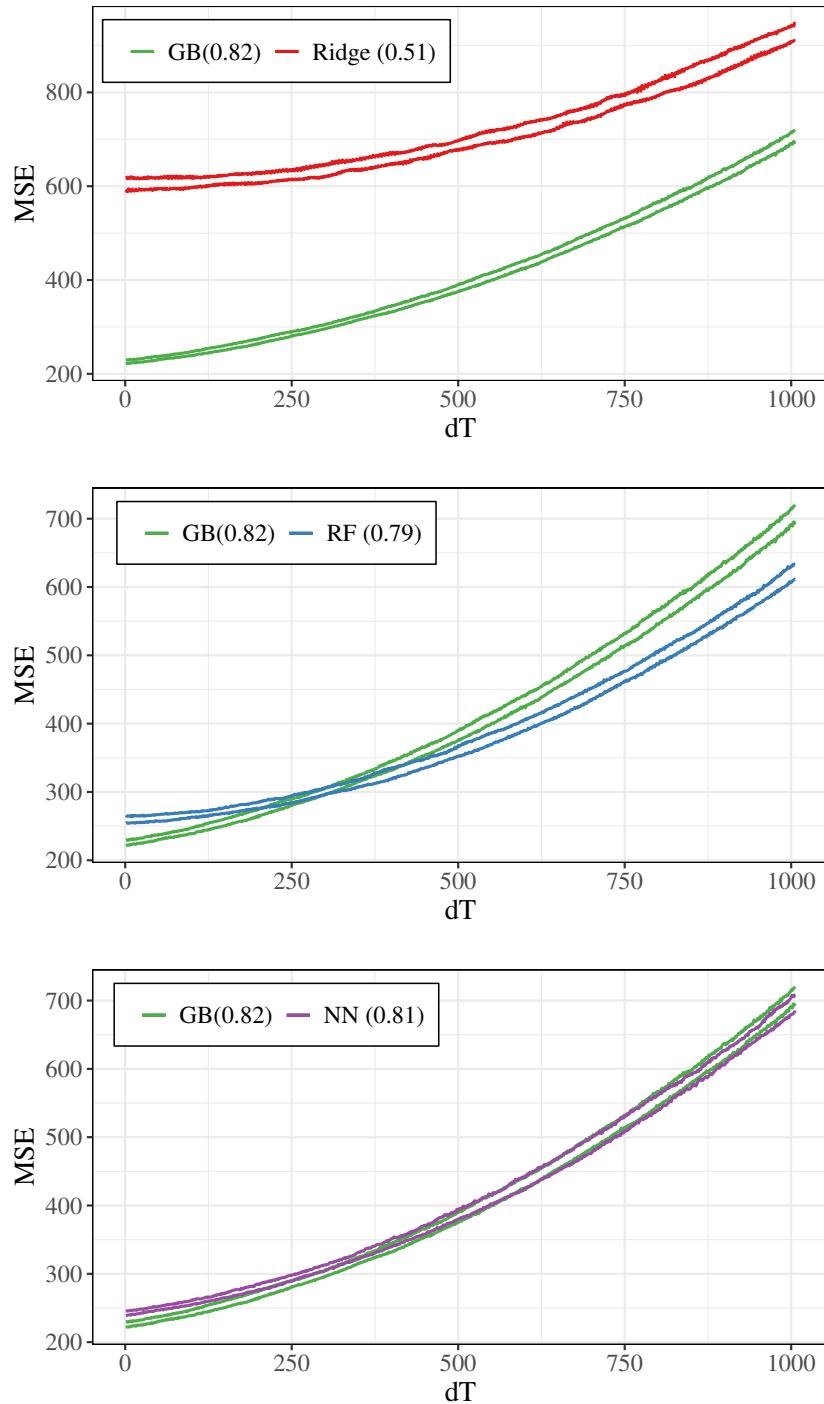


Figura B.16: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 10 variabili, 30 osservazioni al giorno e relazioni cubiche. I modelli con qualità iniziale maggiore risentono maggiormente del *drift*, a prescindere dalla logica matematica.

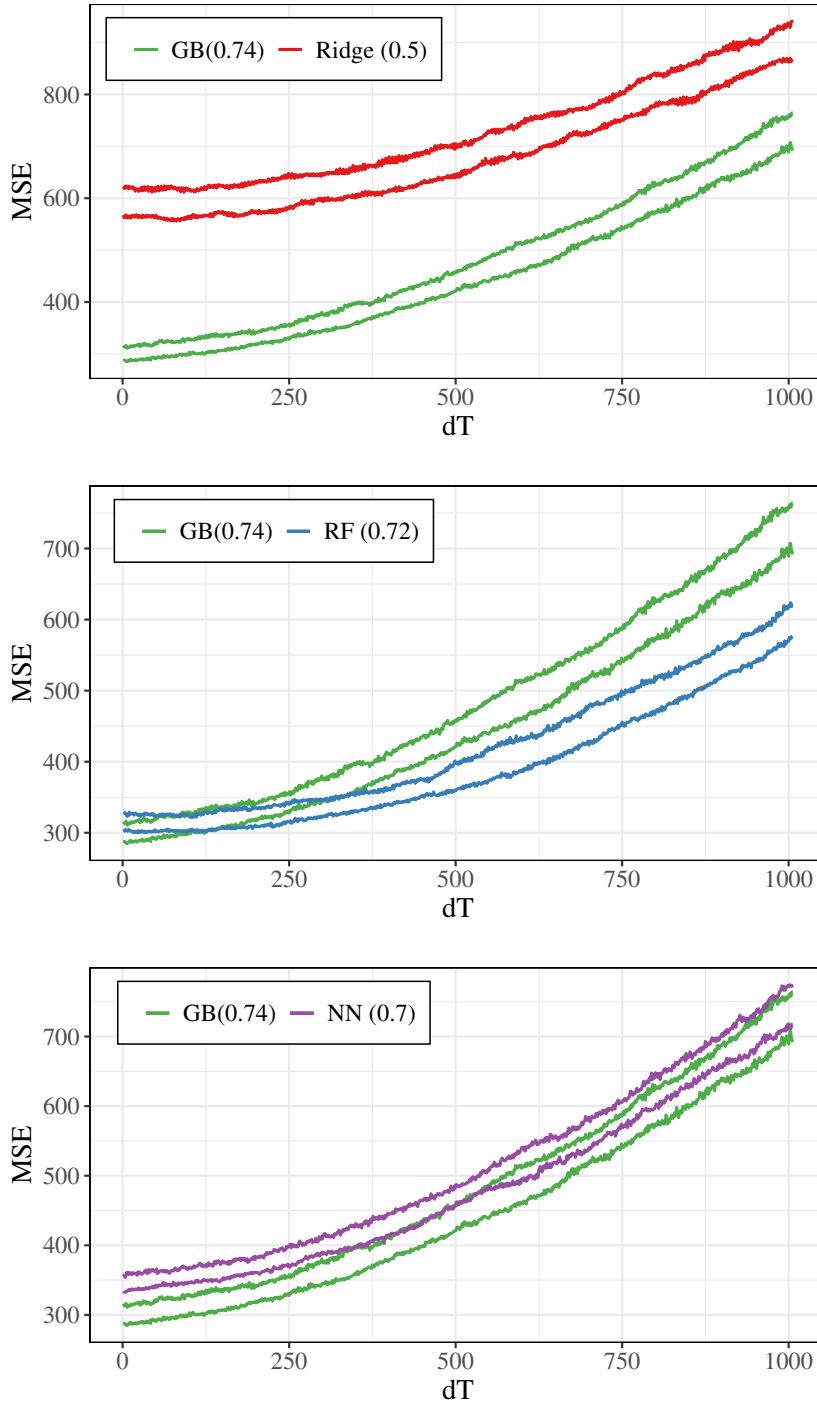


Figura B.17: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 10 variabili, 5 osservazioni al giorno e relazioni cubiche. La *foresta casuale* risente meno del *drift* rispetto alla *rete neurale* pur presentando un maggiore adattamento iniziale.

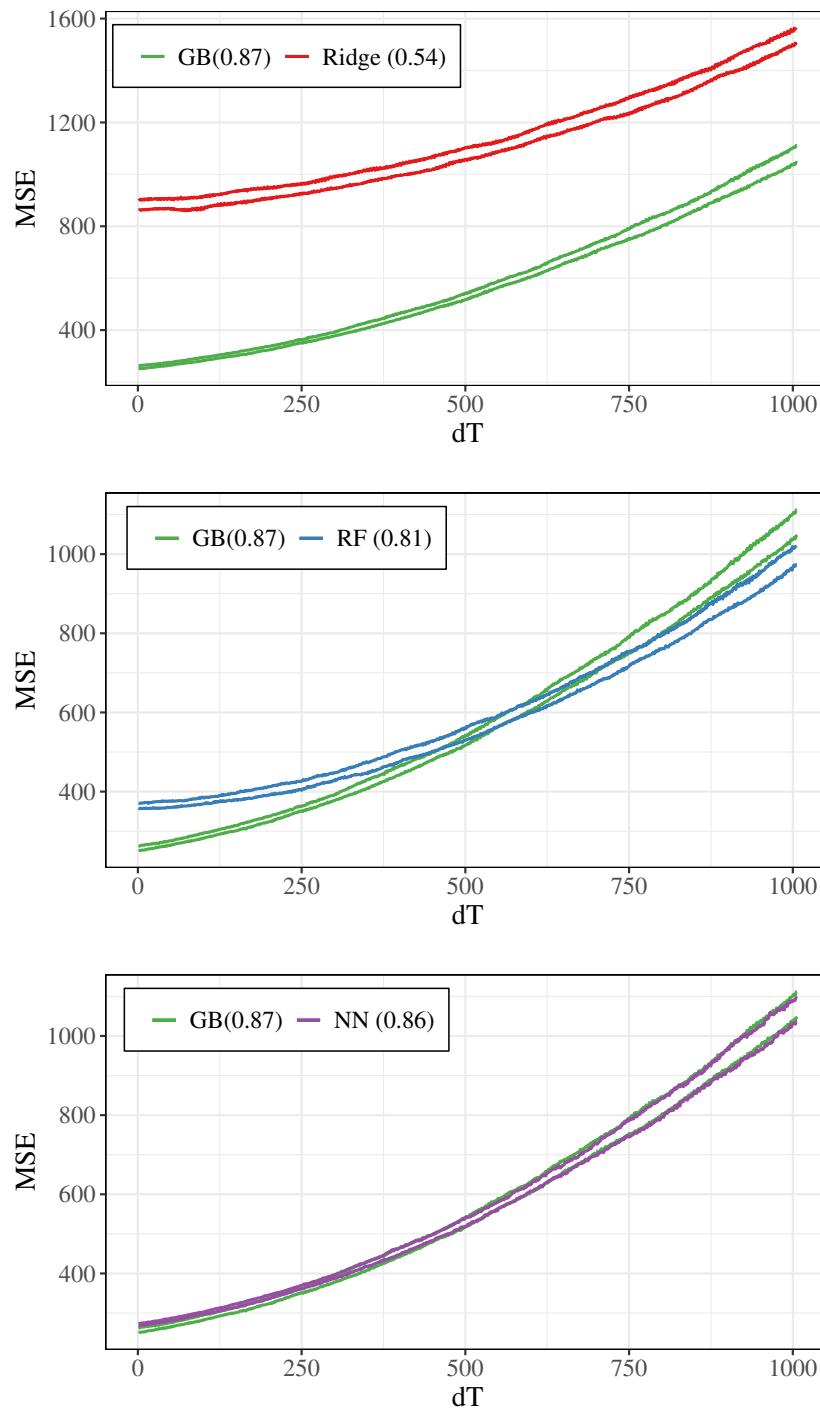


Figura B.18: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, con 10 variabili, 30 osservazioni al giorno e relazioni cubiche; diversi insiemi di coefficienti.

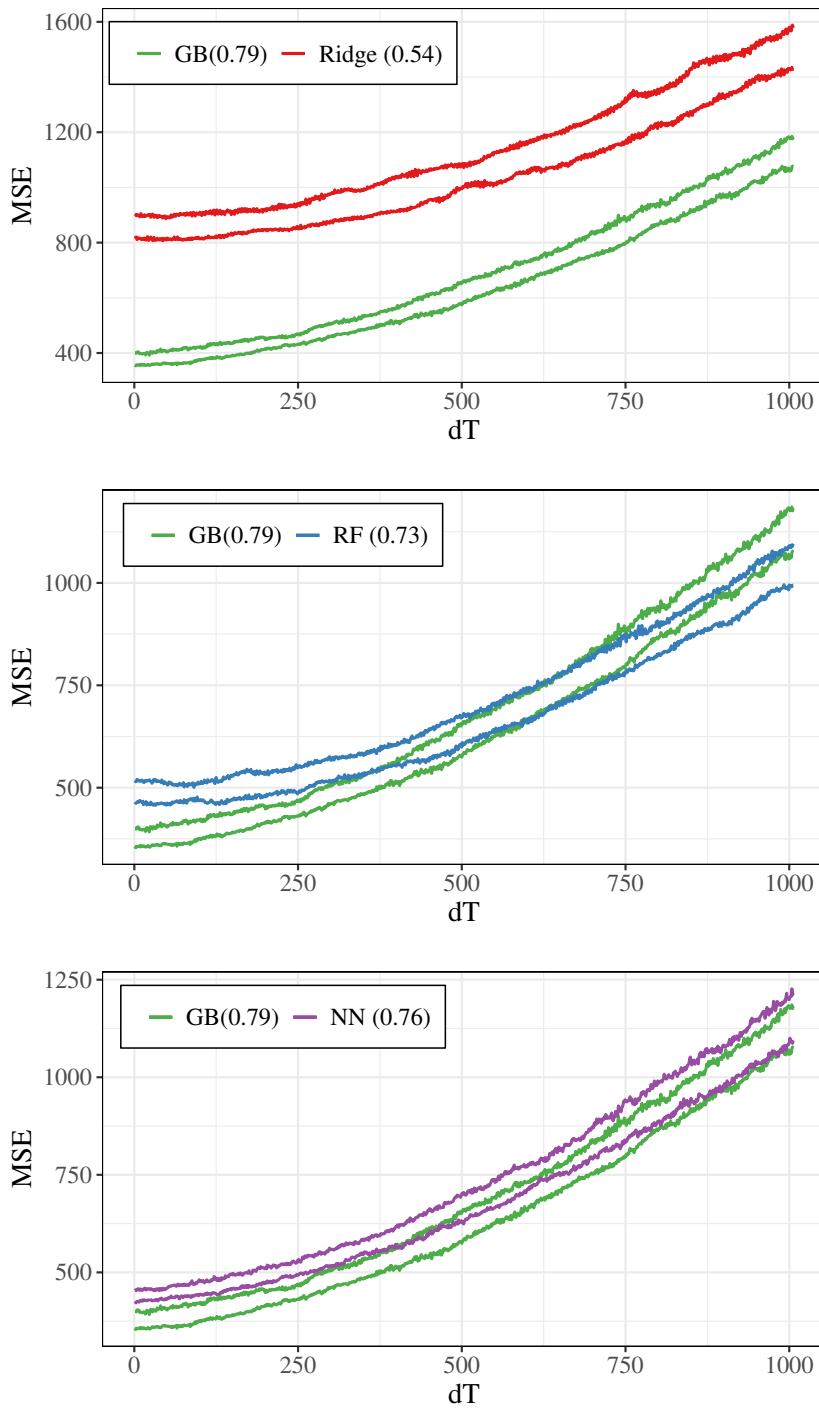


Figura B.19: Andamenti degli *MSE* mediani dei modelli in presenza di *drift incrementale*, con 10 variabili, 5 osservazioni al giorno e relazioni cubiche; diversi insiemi di coefficienti.

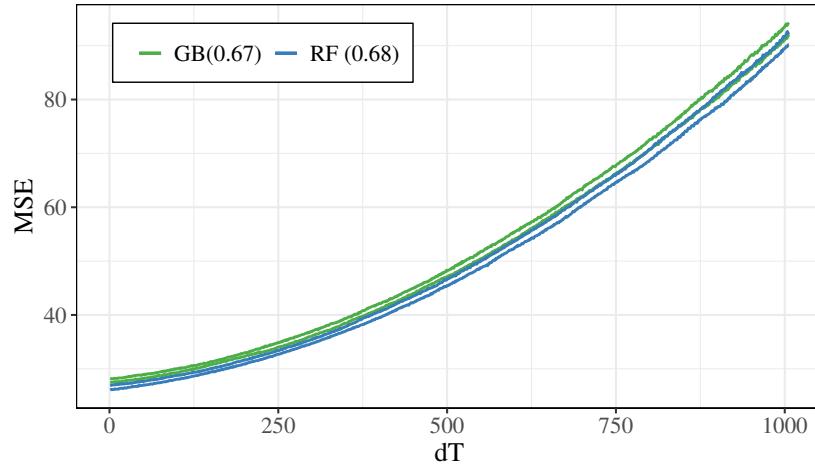


Figura B.20: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, 8 variabili esplicative e 8 interazioni di primo ordine. Il numero di alberi utilizzati dal *GB* è stato ridotto, in modo da ridurre la qualità iniziale. Ciò ha portato il modello a soffrire in misura minore del *concept drift*.

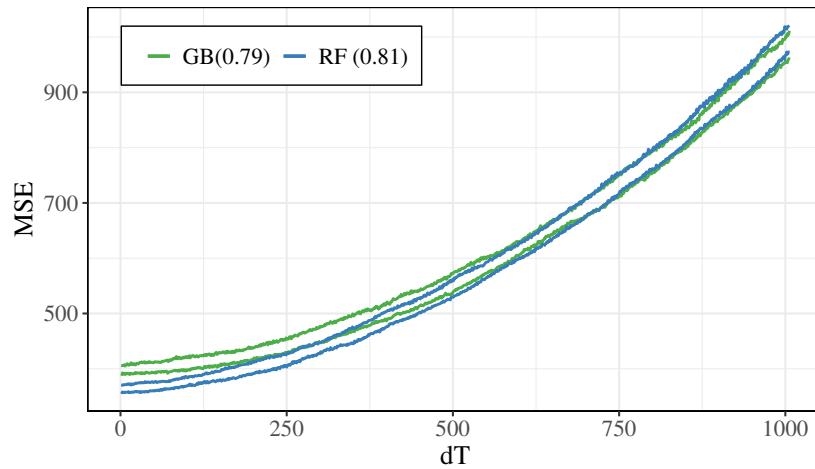


Figura B.21: Andamenti degli MSE mediani dei modelli in presenza di *drift incrementale*, relazioni cubiche e 10 variabili esplicative. Il numero di alberi utilizzati dal *GB* è stato ridotto, in modo da ridurre la qualità iniziale. Ciò ha portato il modello a soffrire in misura minore del *concept drift*.

Concept drift ricorrente

La scelta compiuta sulla metodologia utilizzata per simulare i movimenti nei valori dei coefficienti è legata a quanto presentato dagli autori dell'articolo tramite il grafico (figura 4.13, pagina 111); qui vediamo come la variazione sia stata quantificata tramite un modello *ridge*, adattato su finestre mobili di 3 mesi. Un'ipotesi ragionevole è che questo stimatore, in presenza di valori che variano, stimi una “media” del valore dei coefficienti nel periodo. Nel grafico è riportata quindi la media mobile del valore dei coefficienti (in rosso), calcolata su una finestra di 3 mesi, la cui variabilità è molto simile a quella presentata dagli autori.

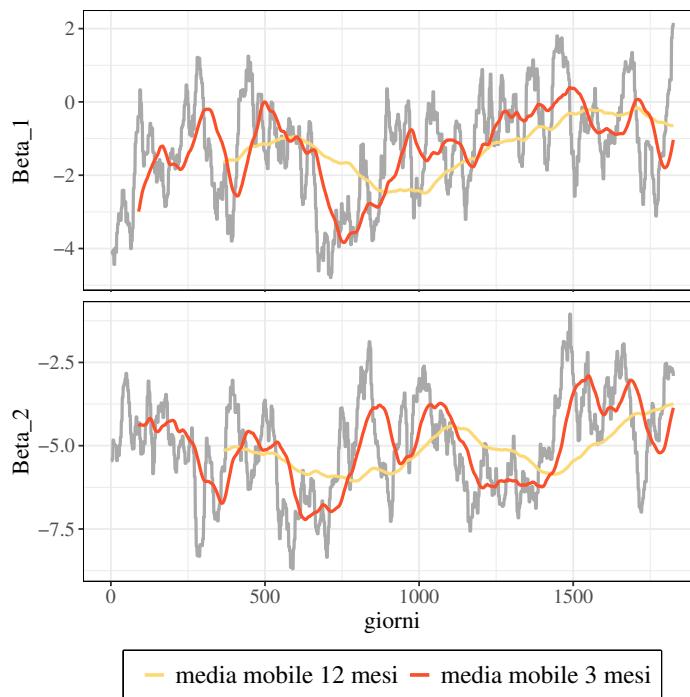


Figura B.22: Due coefficienti, tratti dalla prima replicazione, il cui valore varia seguendo un processo ARMA(2,2): ogni valore del processo è utilizzato per 3 giorni (linea grigia). La linea gialla rappresenta la media dei valori del coefficiente negli ultimi 12 mesi, quella rossa negli ultimi 3.
I coefficienti associati al processo sono $(\phi_1, \phi_2) = (0.6, 0.3)$ e $(\theta_1, \theta_2) = (0.6, 0.5)$
L'escursione dei valori è importante, considerando l'appartenenza dei coefficienti iniziali all'intervallo (± 5) .

Appendice C

Materiale capitolo 5

In questa appendice sono riportate delle figure aggiuntive relative alle simulazioni condotte nel capitolo 5.

Stagionalità - Singoli approcci

In questa sezione sono riportate alcune figure aggiuntive relative ai casi in cui i diversi approcci per catturare la componente stagionale sono utilizzati singolarmente.

L'elenco delle figure:

1. In figura C.1 è riportata la combinazione dei grafici di *AI Aging* nella simulazione in cui, oltre alle X , sono utilizzati i primi due ritardi giornalieri della variabile risposta. Solo 50 replicazioni.
2. In figura C.2 è riportata la distribuzione del terzo quartile quando, oltre alle X , è utilizzato il primo ritardo stagionale della variabile risposta. Solo 50 replicazioni.
3. In figura C.3 è riportata la distribuzione del terzo quartile quando, oltre alle X , sono utilizzati i primi due ritardi giornalieri. Solo 50 replicazioni.
4. In figura C.4 sono riportati gli andamenti degli *MSE* mediani, nel caso venga utilizzato il primo ritardo stagionale (oltre alle X , che saranno

omesse dall'essere menzionate da qui in avanti, ma che compaiono nei dataset). Solo 50 replicazioni.

5. In figura C.5 sono riportati gli andamenti degli *MSE* mediani, nel caso venga utilizzato il primo ritardo stagionale. Il confronto è, nello specifico, tra *RF* e *GB*. Solo 50 replicazioni.
6. In figura C.6 sono riportati gli andamenti degli *MSE* mediani, nel caso vengano utilizzati i primi ritardi giornalieri. Solo 50 replicazioni.
7. In figura C.7 sono riportati gli andamenti degli *MSE* mediani, nel caso vengano utilizzati i primi ritardi giornalieri. Il confronto è, nello specifico, tra *RF* e *GB*. Solo 50 replicazioni.

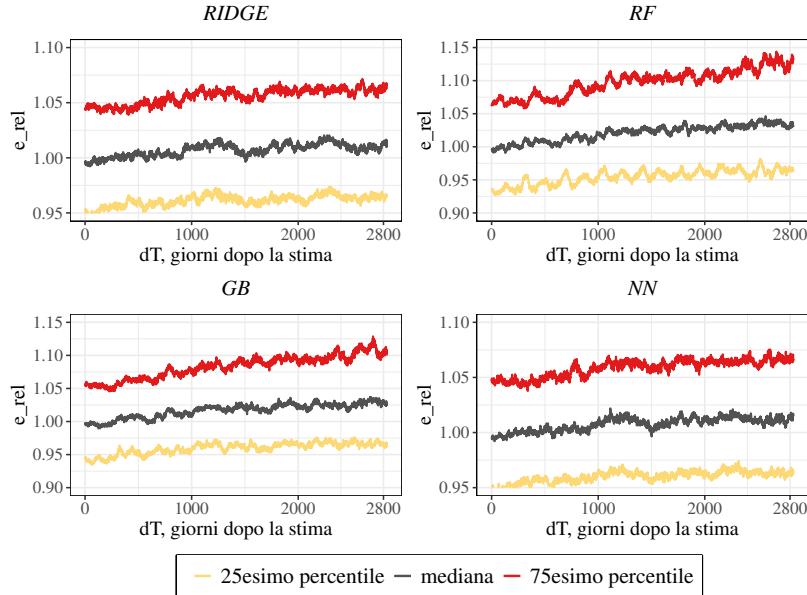


Figura C.1: Combinazione dei grafici di *AI Aging*, stagionalità, utilizzando solo i primi due ritardi giornalieri per catturare le variazioni stagionali. *Forest casuale* e *gradient boosting* presentano un comportamento simile.

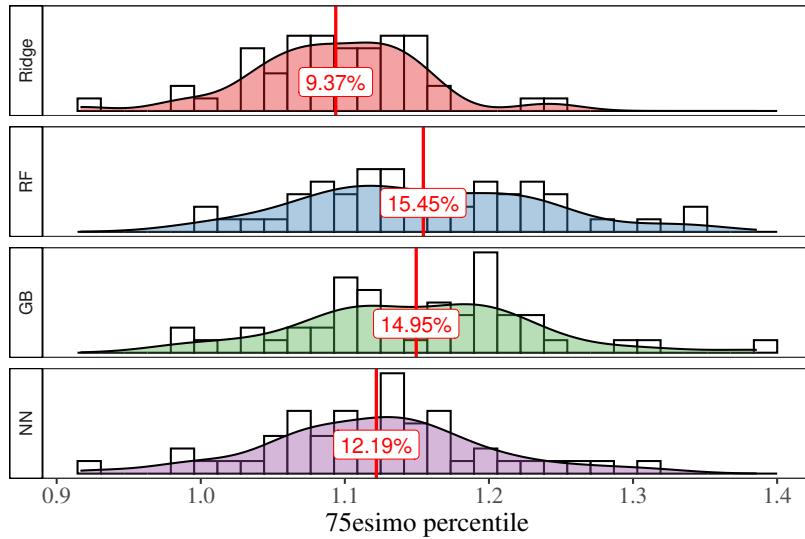


Figura C.2: Distribuzione dei livelli medi del terzo quartile, nel caso venga utilizzato solo il primo ritardo stagionale. *Forest casuale* e *gradient boosting* presentano livelli maggiori, seguiti dalla *rete neurale*.

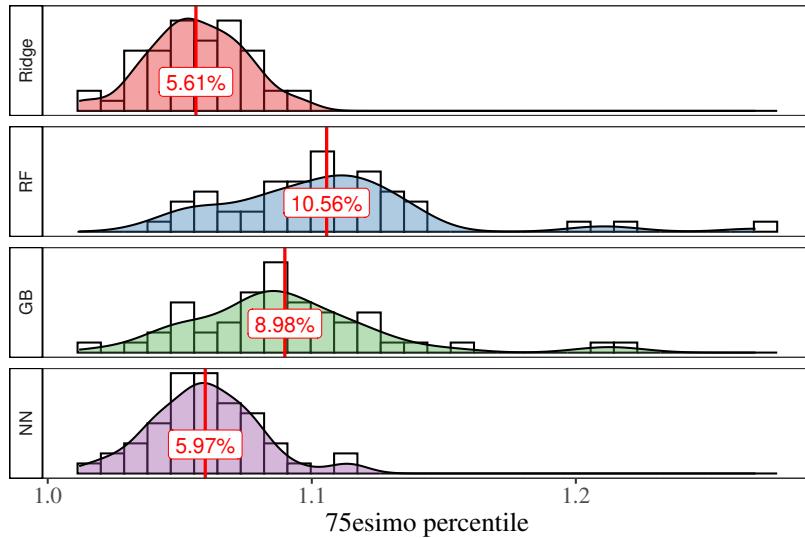


Figura C.3: Distribuzione dei livelli medi del terzo quartile, nel caso vengano utilizzati i soli primi ritardi giornalieri. *Forest casuale* e *gradient boosting* presentano livelli maggiori.

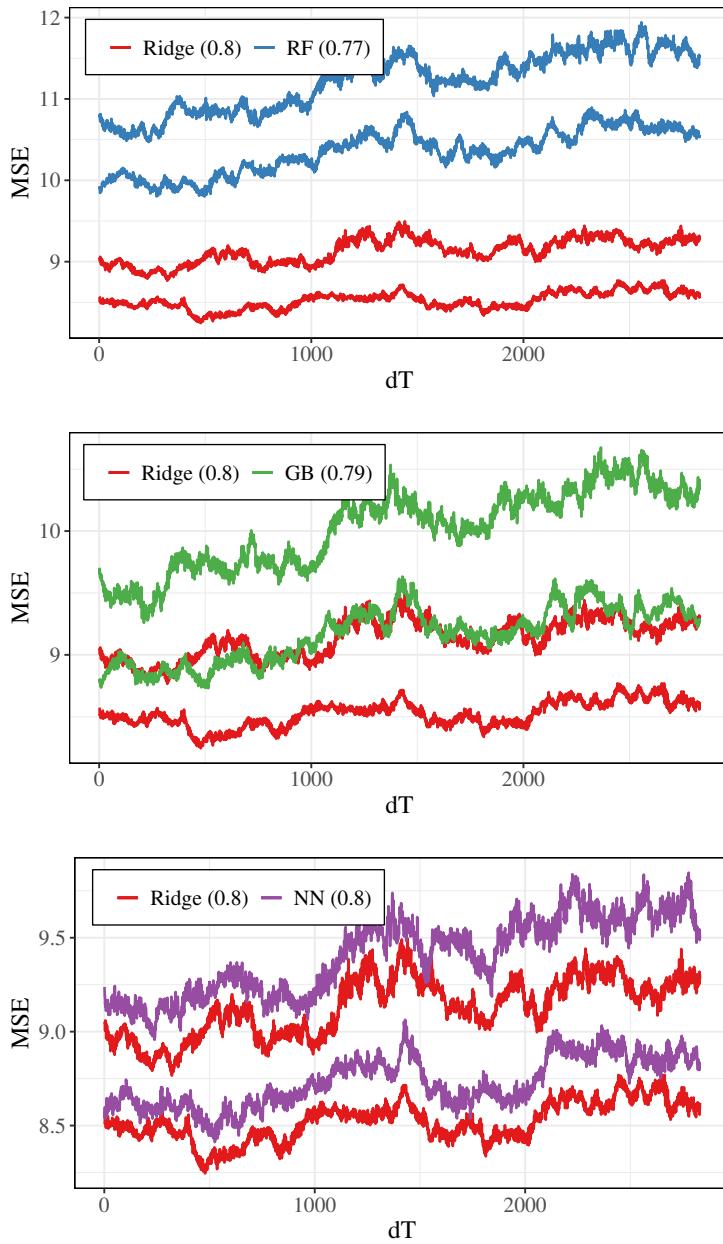


Figura C.4: Andamenti degli MSE mediani, variabili utilizzate: ritardo stagionale. I modelli presentano dei comportamenti di medio/lungo termini molto simili, e il peggioramento delle prestazioni è minimo

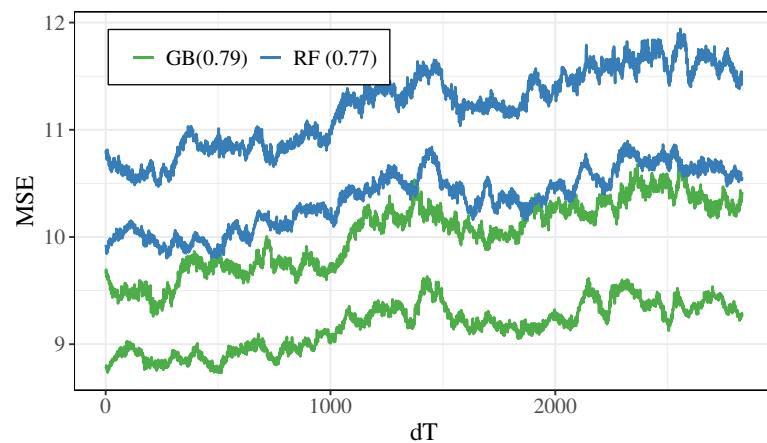


Figura C.5: Andamenti degli MSE mediani, variabili utilizzate: ritardo stagionale. Come nel grafico C.4, ma il confronto è, nello specifico, tra RF e GB , che presentano un comportamento di medio/lungo periodo simile.

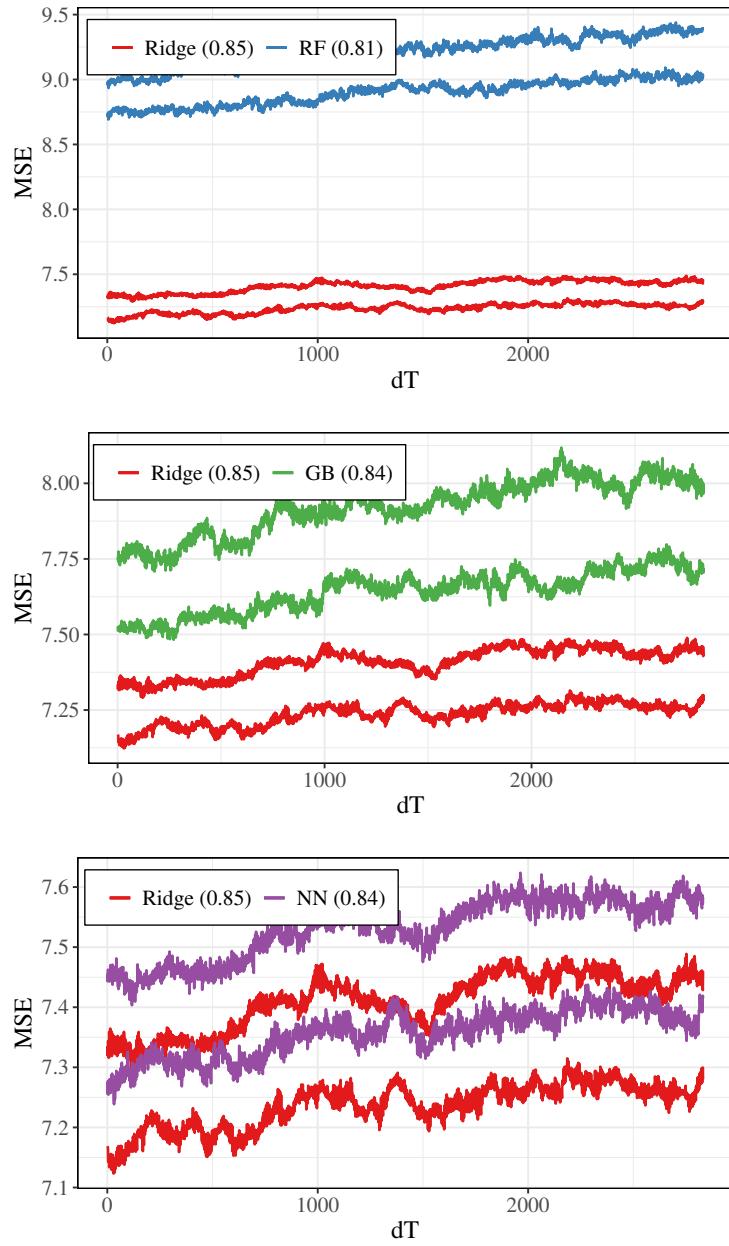


Figura C.6: Andamenti degli MSE mediani, variabili utilizzate: primi due ritardi giornalieri. I modelli presentano dei comportamenti di medio/lungo termini molto simili, e il peggioramento delle prestazioni è minimo.

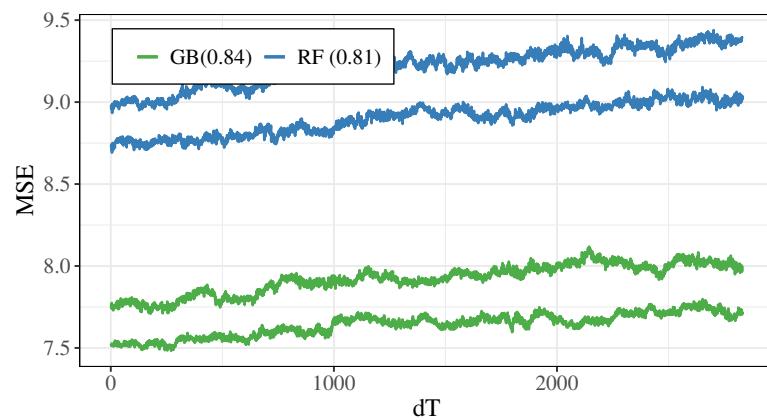


Figura C.7: Andamenti degli MSE mediani, variabili utilizzate: primi due ritardi giornalieri. Come nel grafico C.6, ma il confronto è, nello specifico, tra RF e GB , che presentano un comportamento di medio/lungo periodo simile.

Stagionalità - Più approcci in contemporanea

In questa sezione sono riportati i grafici aggiuntivi legati alle simulazioni in cui vengono utilizzati i diversi approcci assieme. L'elenco delle figure è il seguente:

1. In figura C.8 sono riportati gli andamenti degli MSE mediani nel caso vengano utilizzate le variabili mese, giorno dell'anno e ritardo stagionale.
2. In figura C.9 sono riportati gli andamenti degli MSE mediani nel caso vengano utilizzate le variabili mese, giorno dell'anno e primi ritardi giornalieri.
3. Nelle figure C.10 e C.11 sono riportati gli andamenti degli MSE mediani e dei terzi quartili (dell' MSE) nel caso vengano utilizzate tutte le variabili, ma il numero di alberi utilizzati dal *gradient boosting* venga ridotto. A parità di qualità iniziale la differenza nella mediana è molto ridotta (figura C.10), ma nel terzo quartile è più ampia (figura C.11).
4. Nella figura C.13 è presentata una simulazione differente: la componente stagionale e il disturbo sono maggiori. In questa è stato utilizzato un approccio differente, precedente rispetto a quello utilizzato negli altri casi (e meno adatto): il valore iniziale degli effetti stagionali cambia da una replicazione ad un'altra. Tutte le variabili sono state utilizzate per catturare l'effetto stagionale. Un esempio di dataset è riportato nella figura C.12.
5. Nelle figure C.14 e C.16 sono riportati i risultati della simulazione in cui gli effetti della componente stagionale sono stati ridotti. Nella prima sono riportati gli andamenti degli MSE mediani, mentre nella seconda vengono confrontati *ridge*, *foresta casuale* e *gradient boosting* dal punto di vista del caso peggiore, in cui le differenze sono più evidenti. In figura C.15 è riportato un esempio di dataset. Solo 50 replicazioni.
6. I risultati della simulazione che include una componente di trend sono riportati nelle figure C.18 (andamenti degli MSE mediani), C.19 (andamenti degli MSE mediani, *RF* e *GB*), C.20 (andamenti del terzo

quartile dell' MSE , RF e GB) e C.21 (andamenti del terzo quartile dell' MSE , $ridge$ e NN), con un esempio di dataset in figura C.17. Solo 50 repliche.

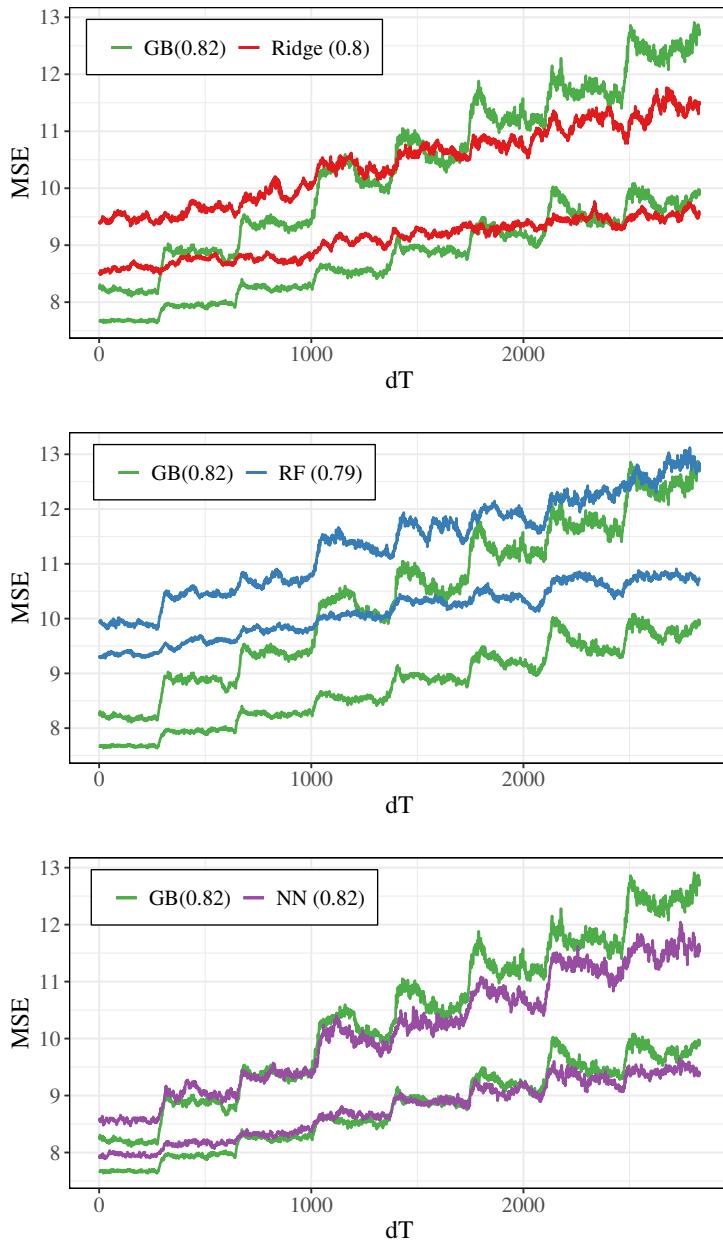


Figura C.8: Andamenti degli MSE mediani, variabili utilizzate: mese, giorno e ritardo stagionale. Il *gradient boosting* presenta un comportamento differente rispetto agli altri modelli.

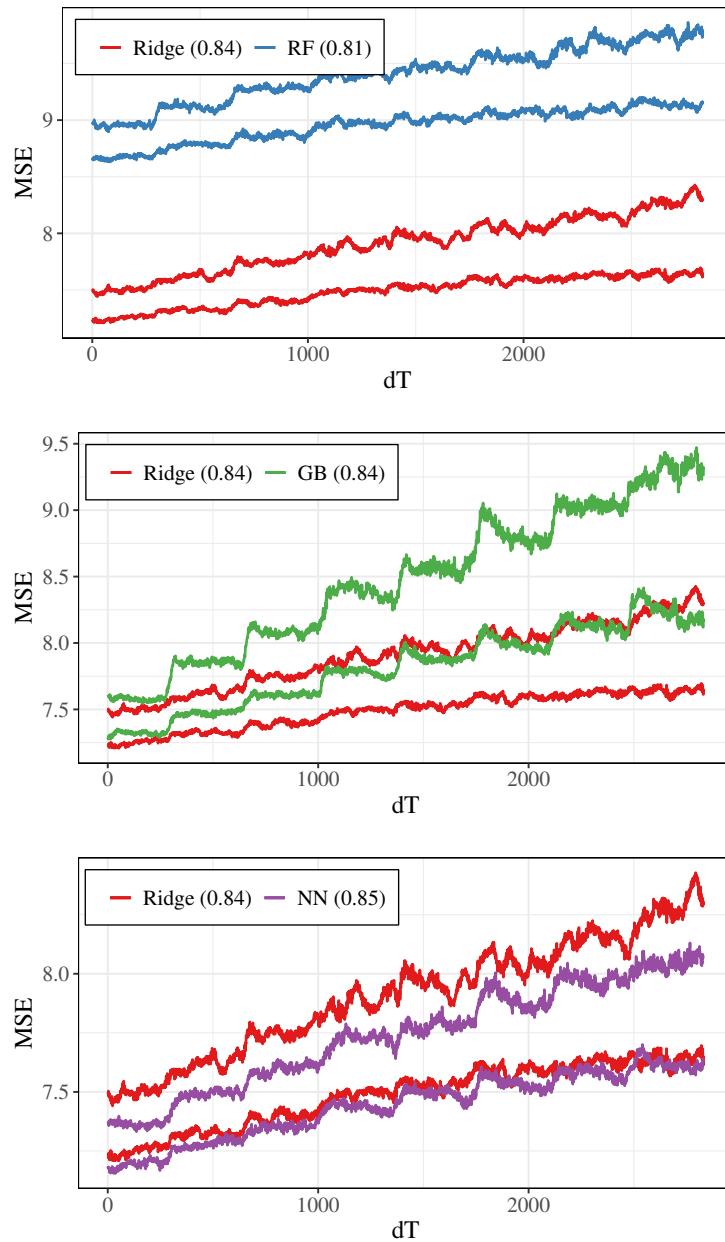


Figura C.9: Andamenti degli MSE mediani, variabili utilizzate: mese, giorno e primi ritardi giornalieri. Il *gradient boosting* presenta un comportamento differente rispetto agli altri modelli.

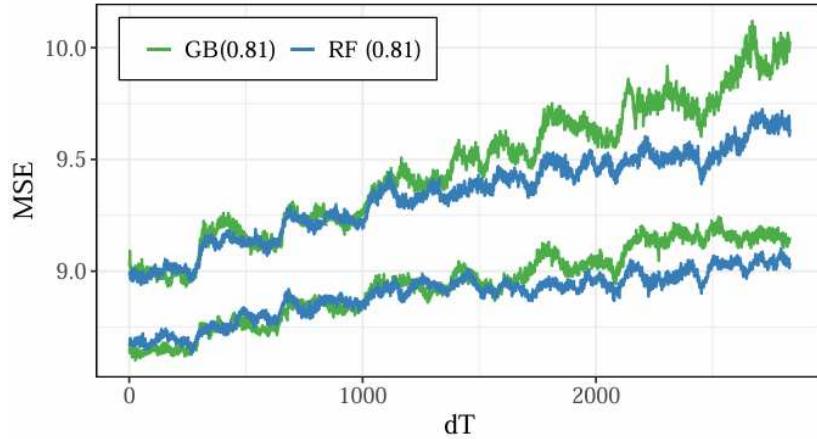


Figura C.10: Andamenti degli MSE mediani, nel caso vengano utilizzate tutte le variabili disponibili e il numero di alberi utilizzati dal *gradient boosting* venga ridotto. La differenza con la *foresta* si riduce moltissimo, ma non scompare. È più evidente dal grafico dei terzi quartili (figura C.11).

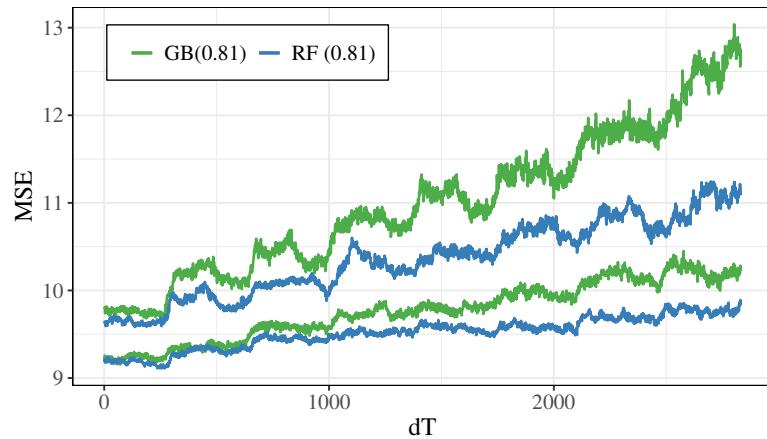


Figura C.11: Andamenti del terzo quartile dell' MSE , nel caso vengano utilizzate tutte le variabili disponibili e il numero di alberi utilizzati dal *gradient boosting* venga ridotto. La differenza tra *foresta casuale* e *gradient boosting* è molto più evidente.

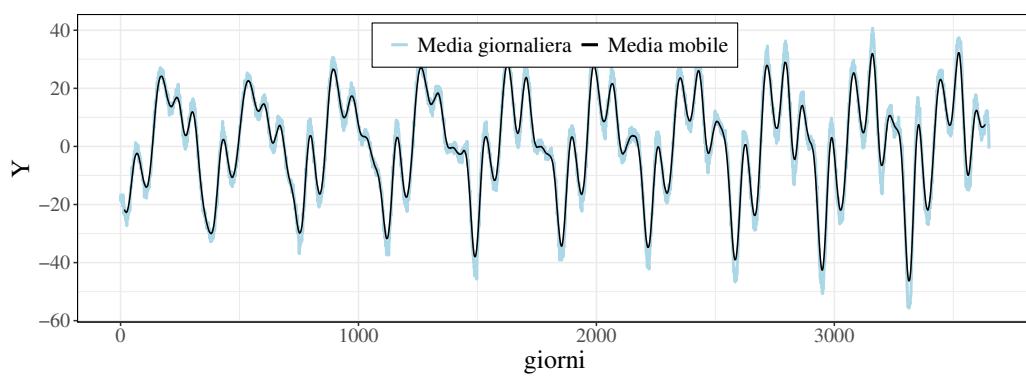


Figura C.12: Esempio di dataset in cui la componente stagionale ha un'importanza e un disturbo maggiori.

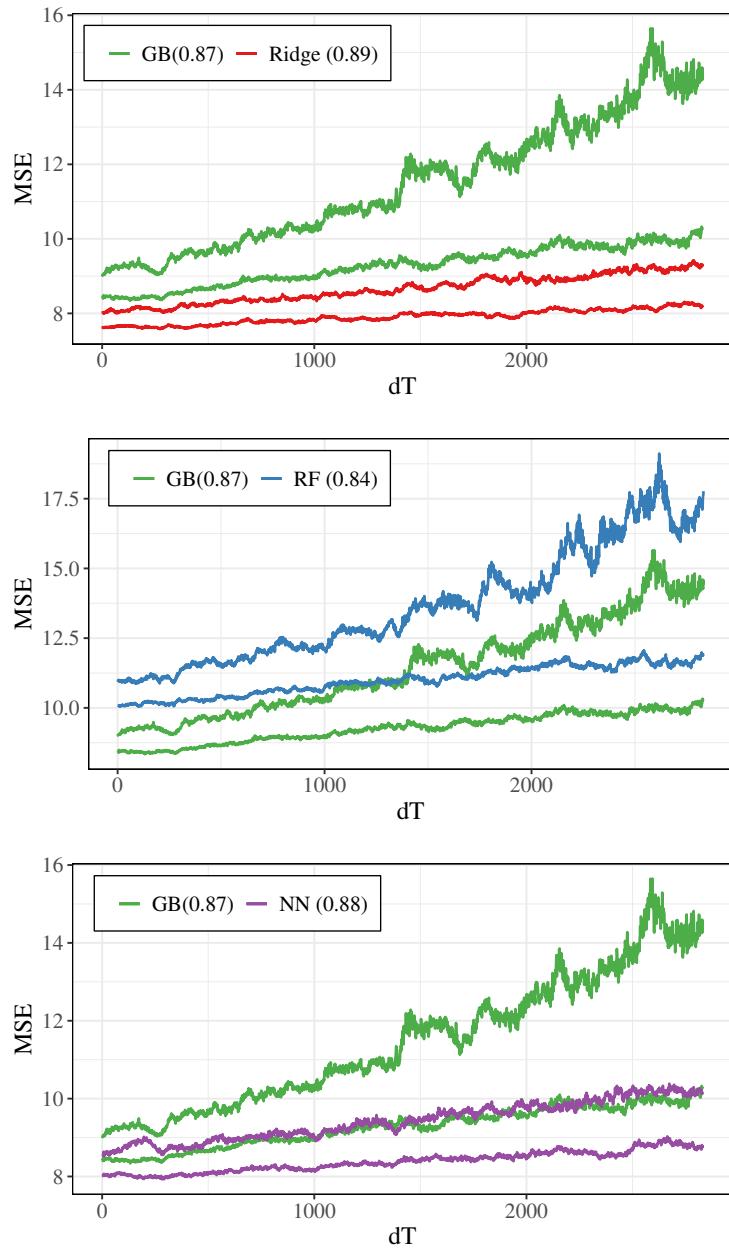


Figura C.13: Andamenti degli MSE mediani, stagionalità e disturbo maggiori, variabili utilizzate: tutte. La differenza tra *gradient boosting* e *foresta casuale* scompare.

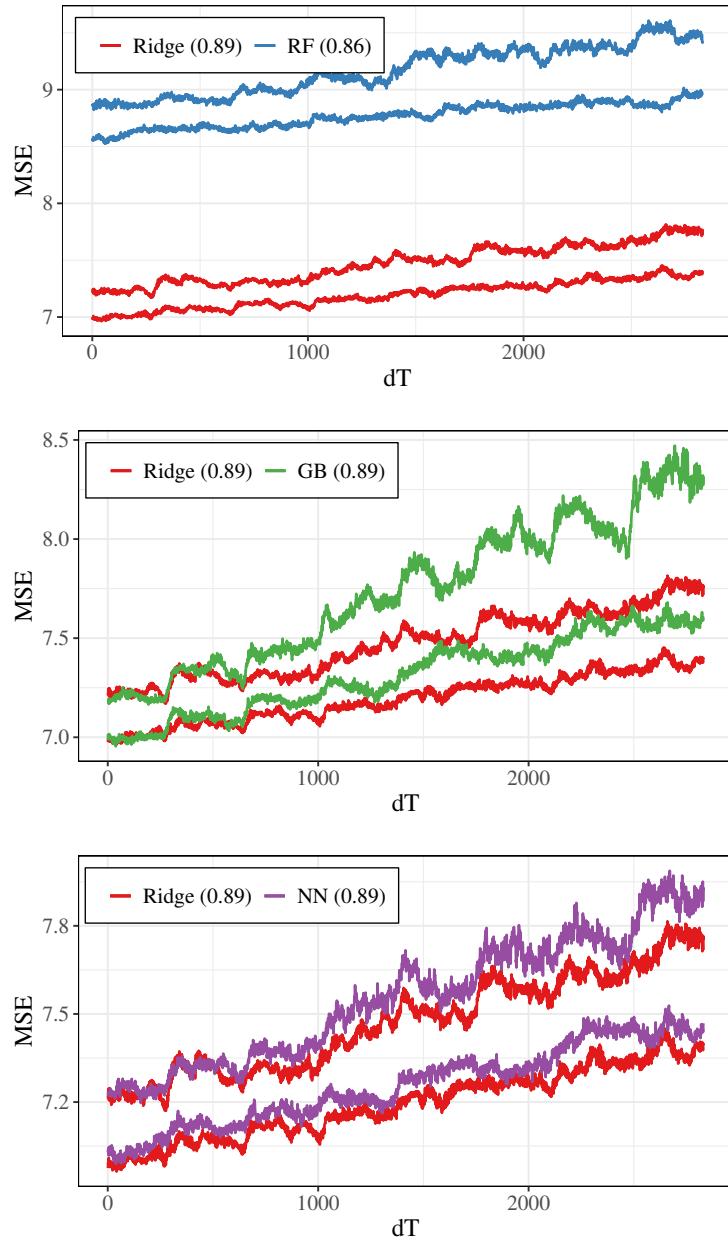


Figura C.14: Andamenti degli MSE mediani, nel caso di stagionalità e componente stagionale di importanza ridotta. Evitando di commentare i soliti risultati, in questo caso la differenza tra NN e *ridge* è maggiore, ma comunque minima.

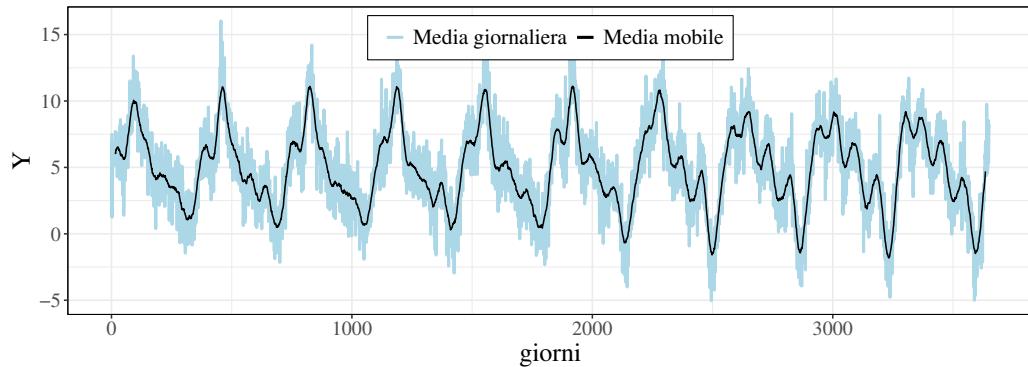


Figura C.15: Esempio di dataset in cui la componente stagionale ha un'importanza minore.

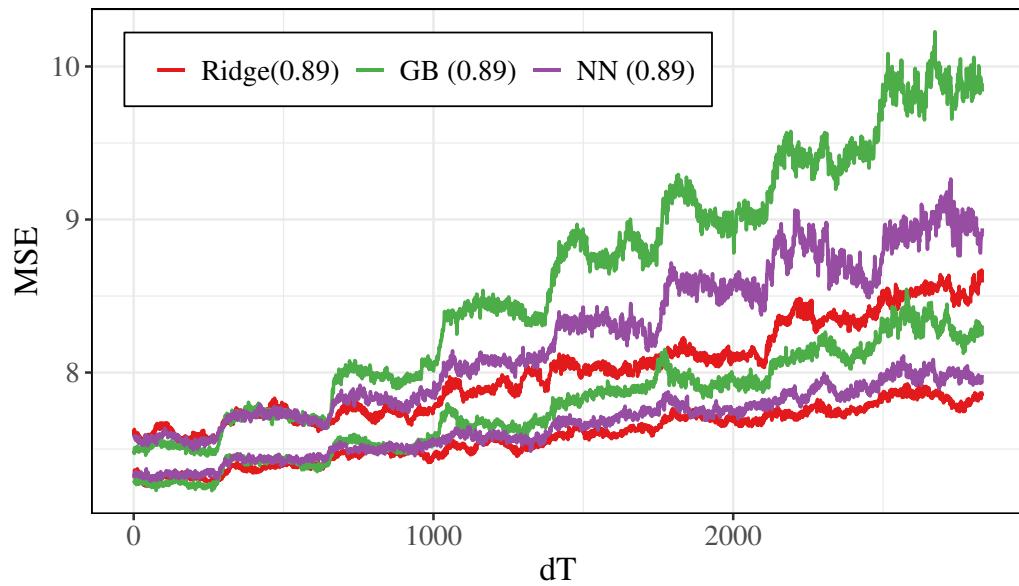


Figura C.16: Andamenti del terzo quartile dell'MSE di ridge, foresta casuale e gradient boosting. I modelli presentano delle differenze.

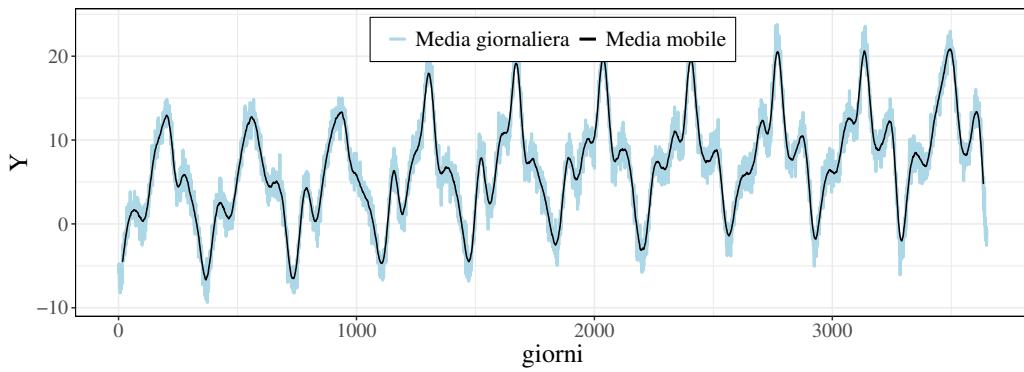


Figura C.17: Esempio di dataset in cui alla componente stagionale è aggiunta una di trend. Questa è simulata tramite un processo random walk con drift in modo analogo a quanto fatto nel capitolo 3 con i processi autoregressivi. Il trend è presente, ma non è eccessivo: in un caso di quel tipo le differenze sono scontate.

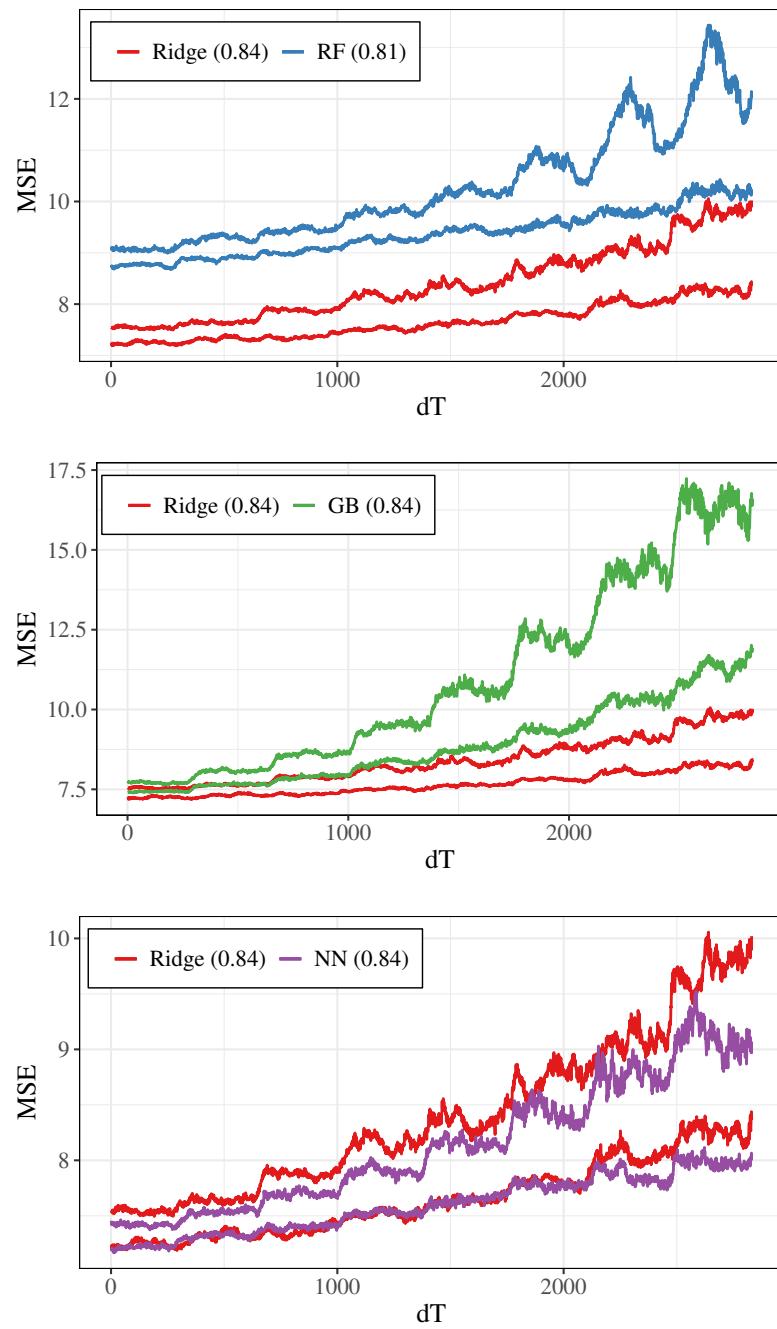


Figura C.18: Andamenti degli MSE mediani, nella simulazione in cui è presente la componente di trend.

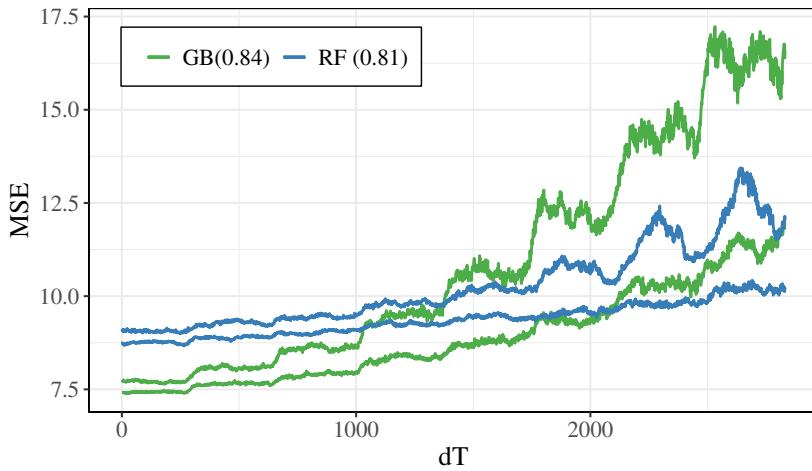


Figura C.19: Andamenti degli MSE mediani, nella simulazioni in cui è presente la componente di trend, nello specifico di RF e GB ; le differenze sono importanti.

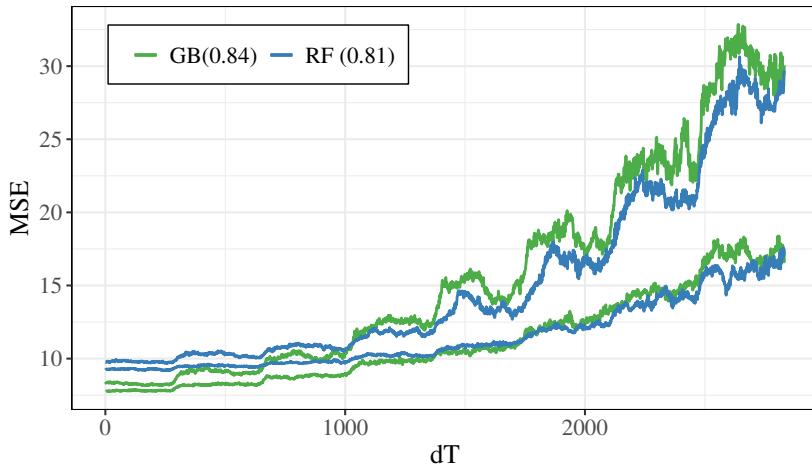


Figura C.20: Andamenti del terzo quartile dell' MSE , nella simulazioni in cui è presente la componente di trend, nello specifico di RF e GB ; le differenze sono minime.

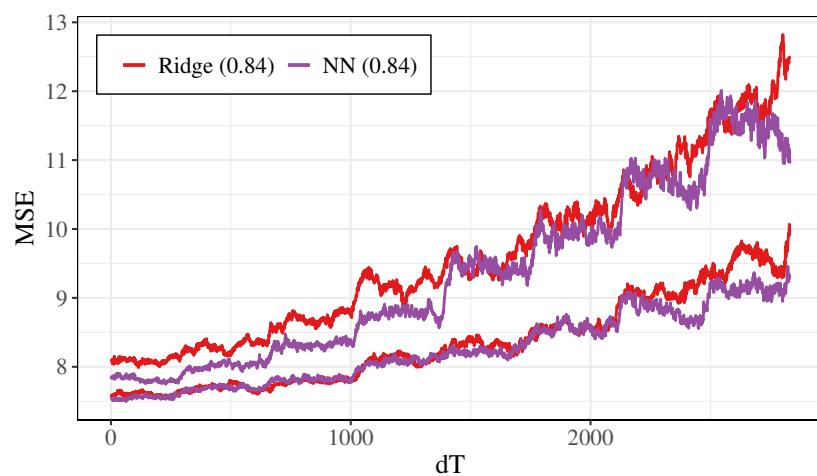


Figura C.21: Andamenti del terzo quartile dell'MSE, nella simulazione in cui è presente la componente di trend, nello specifico di *ridge* e *NN*.

Appendice D

Materiale aggiuntivo

In questa piccola sezione sono riportati i grafici di *AI Aging* costruiti sulla base dell'errore non relativo ($MSE(dT)$), utilizzando i risultati ottenuti dall'applicazione del “test” di *degradazione temporale* sui dataset reali del capitolo 2 (figure D.1 - D.4). In questi grafici non sono riportati i singoli punti, ma solo i tracciati dei quartili, per maggiore chiarezza.

Le figure sono molto simili a quelle ottenute utilizzando gli errori relativi, e riportate nel capitolo 2 (da pagina 50).

I risultati ottenuti dall'applicazione del “test”, nei casi reali, coincidono con quanto osservato nelle simulazioni? Nei casi F1 - F3 - F4 nessun modello si è veramente distinto, presentando comportamenti di medio/lungo periodo molto simili (basandosi sui risultati riportati nel capitolo 2):

1. Nel caso del dataset F1 la qualità iniziale dei modelli è molto simile, ma ci sono delle differenze nei livelli, soprattutto del terzo quartile. Certamente il modello inizialmente migliore è anche il più stabile (la *rete*). Le differenze sono comunque maggiori di quelle osservate nelle simulazioni, a parità di qualità iniziale.
2. In F3 l'errore iniziale prevede abbastanza bene la stabilità dei modelli, con i due modelli migliori molto simili in questi termini. La qualità iniziale è però praticamente la stessa per i quattro modelli.

3. In F4 il processo evolve, in quanto è presente *degradazione temporale*.

Il modello peggiore (la *rete*) presenta minore invecchiamento, mentre per gli altri modelli la stabilità è più simile. Le differenze non sono totalmente spiegabili sulla base del livello iniziale.

Rispetto alle simulazioni i risultati sono più difficili da spiegare in base ai soli livelli iniziali. In modo analogo, però, non è presente una logica più o meno stabile in ogni caso, e le differenze tra i modelli non si ripresentano allo stesso modo da un dataset all'altro.

Il caso F2 è più complicato, in quanto i punti di partenza dei modelli sono molto differenti. La scelta ricade sul modello inizialmente migliore (il *ridge*), e i livelli dei quartili indicano comunque, per i tre modelli *ridge*, *RF* e *GB*, come i meno stabili siano anche quelli inizialmente migliori. Questo è in linea con quanto affermato: il “test” evidenzia delle differenze quando la qualità iniziale è diversa. In termini assoluti (figure D.2) gli stessi modelli presentano differenze minime (e il *ridge*, il meno stabile, è da preferire; è quanto osservato nel capitolo 2).

L'eccezione è la *rete neurale*, con una qualità iniziale molto bassa e un grosso aumento dell'errore. Questo risultato non è spiegabile sulla base degli esperimenti condotti.

Certamente in un caso reale le differenze sono maggiori, e spiegarle diventa molto difficile, rendendo necessario uno studio caso per caso.

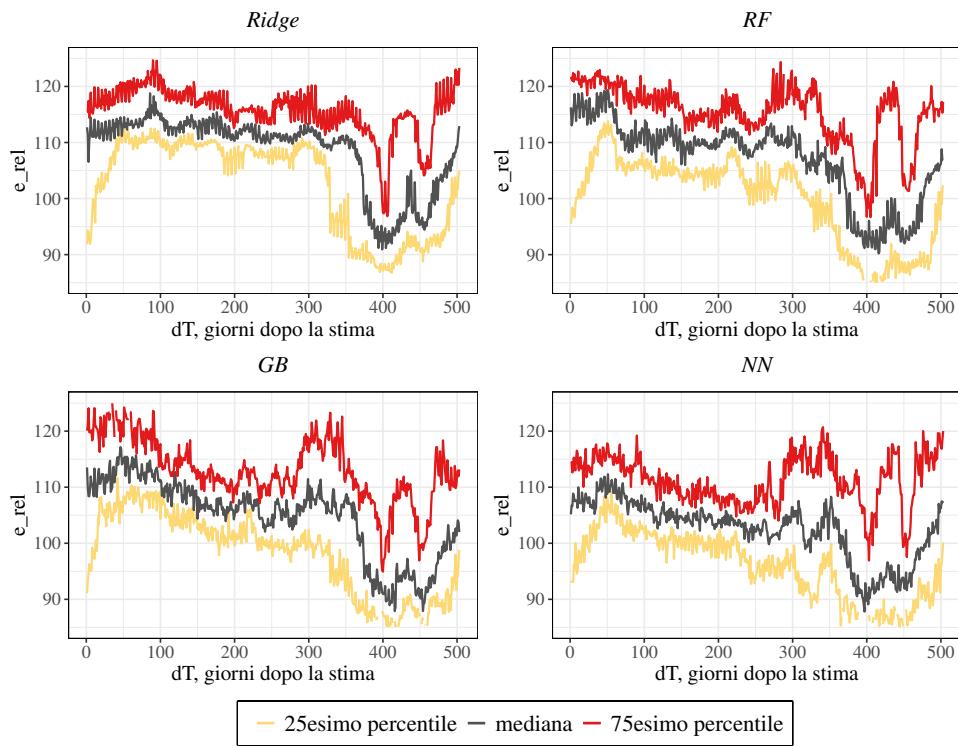


Figura D.1: Grafici di *AI Aging* costruiti a partire dagli $MSE(dT)$, gli errori non relativi, per il dataset F1, utilizzato nel capitolo 2.

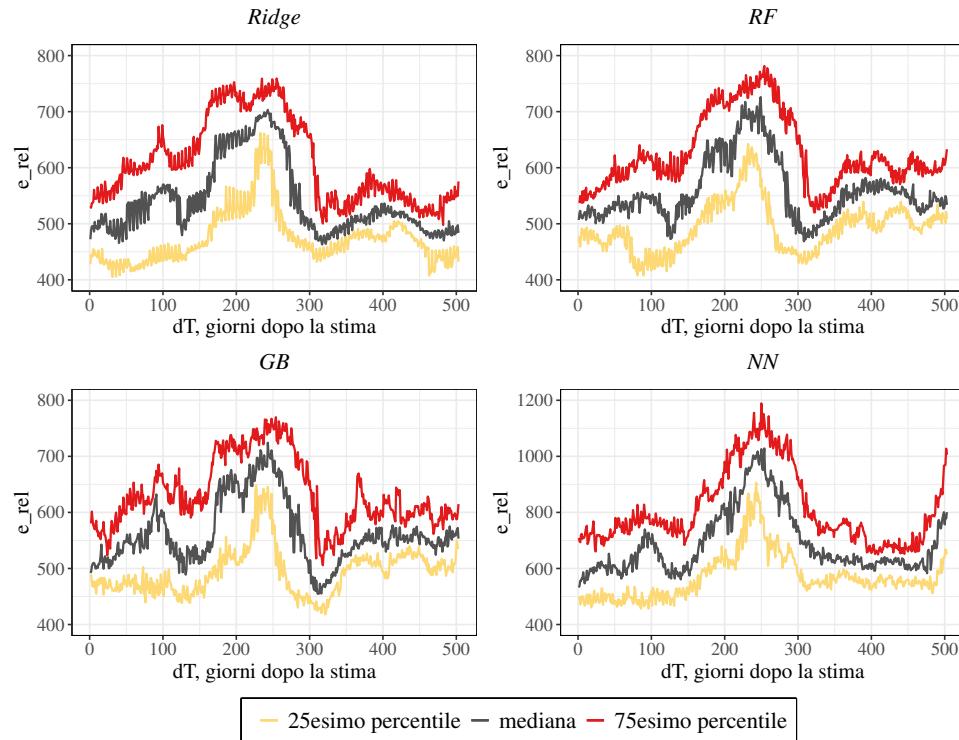


Figura D.2: Grafici di *AI Aging* costruiti a partire dagli $MSE(dT)$, gli errori non relativi, per il dataset F2, utilizzato nel capitolo 2. La *rete neurale* presenta dei livelli di errore molto più alti.

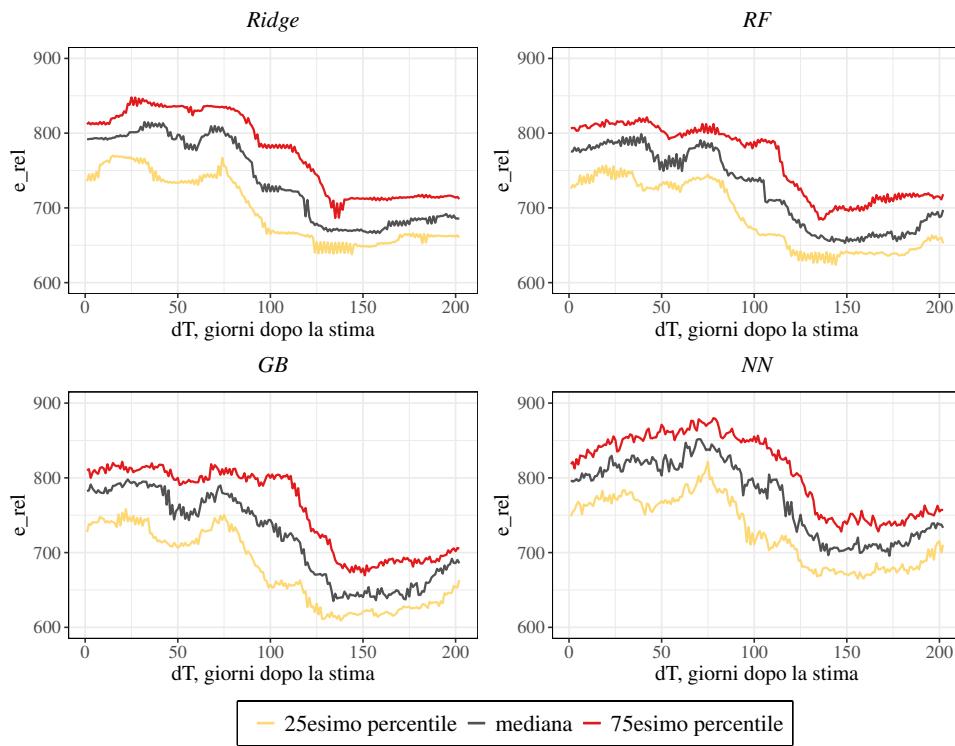


Figura D.3: Grafici di *AI Aging* costruiti a partire dagli $MSE(dT)$, gli errori non relativi, per il dataset F3, utilizzato nel capitolo 2.

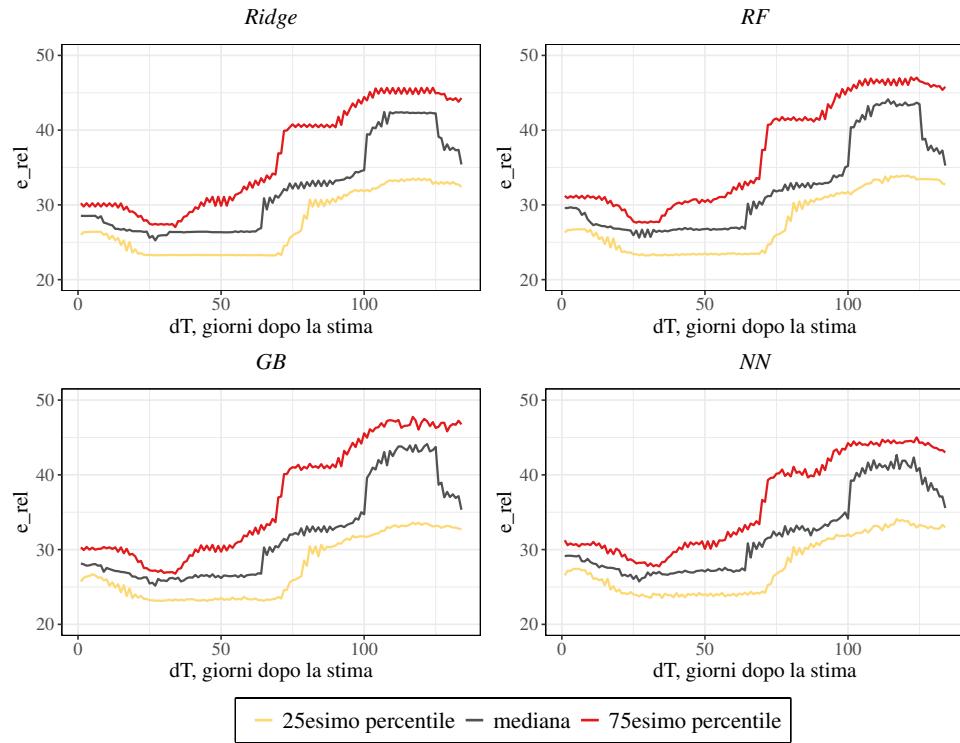


Figura D.4: Grafici di *AI Aging* costruiti a partire dagli $MSE(dT)$, gli errori non relativi, per il dataset F4, utilizzato nel capitolo 2.