Statistica Superiore - Relazione III

Vendita immobiliare negli Stati Uniti: previsioni per gli anni 2021 - 2022

Eleonora Basilico

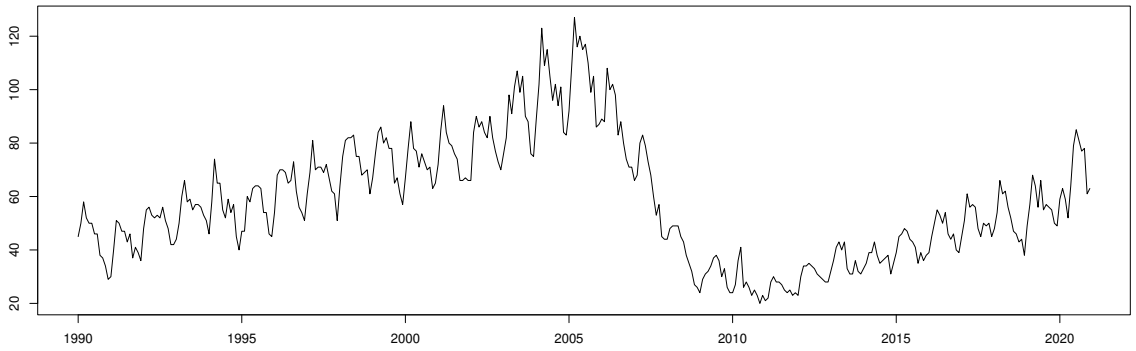
1 Descrizione del problema

La seguente analisi è rivolta alle società immobiliari americane e non, e più in generale a tutti coloro che desiderano investire sulla costruzione di nuove abitazioni unifamiliari negli Stati Uniti. L'obiettivo è quello di capire se e con quanta precisione l'andamento delle vendite di questo tipo di immobile negli ultimi 30 anni consenta di predire il numero di vendite future, ed aiutare così le società a cui ci rivolgiamo a compiere scelte più ponderate.

Abbiamo considerato la serie storica relativa al numero mensile di unità di migliaia di case unifamiliari nuove vendute negli Stati Uniti dal 1990 al 2020. La serie originale, reperibile al seguente link `https://fred.stlouisfed.org/series/HSN1FNSA`, contiene anche i dati relativi agli anni 1963-1989, che non abbiamo considerato in quanto troppo lontani nel tempo. Sono presenti anche i dati del 2021 fino al mese di novembre. Poichè manca il valore del mese di dicembre ancora in corso, abbiamo deciso di utilizzare i dati fino al dicembre 2020 e di prevedere a partire da questi l'andamento delle vendite relative al 2021. Confronteremo poi i valori previsti con quelli effettivi.

Cercheremo inoltre di prevedere, laddove abbia senso farlo, anche un possibile andamento delle vendite per il 2022. Bisogna, però, tenere conto del fatto che a causa della pandemia ancora in corso fare previsioni di questo tipo che siano affidabili è tutt'altro che facile.

Visualizziamo graficamente la nostra serie storica:

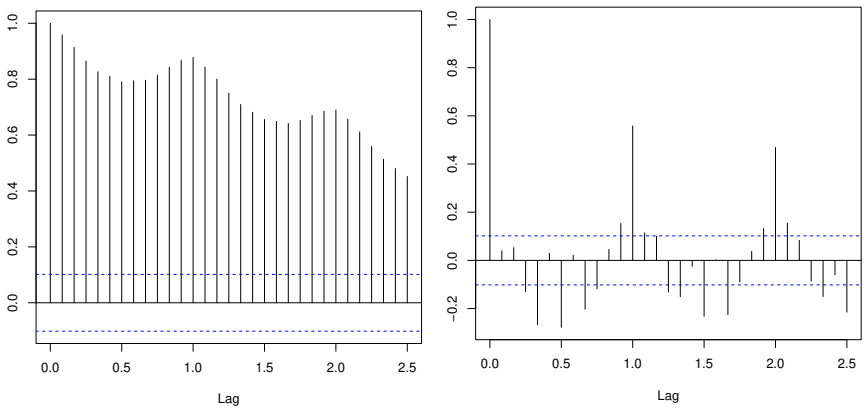


Notiamo la presenza di un pronunciato trend ascendente negli anni che vanno dal 1990 al 2005. A partire dal 2006 il numero di vendite cala rapidamente fino al 2008, per poi rimanere stabile negli anni immediatamente successivi. Dal 2013 c'è una leggera ripresa: le vendite tornano a crescere con un andamento simile a quello che si osserva nella prima metà del grafico. Il netto calo di vendite negli anni 2006 - 2008 è da attribuire alla pesante crisi economica che proprio in quel periodo ha colpito gli Stai Uniti.

Inoltre, tenendo conto della possibile presenza di rumore nei dati, sembra esserci anche una lieve stagionalità: il picco annuale di vendite è raggiunto il 71% delle volte nei mesi di marzo e aprile e il 19,3% nei mesi di maggio e giugno. Si discostano gli anni 2009 e 2020, che raggiungono il picco di vendite nella stagione estiva: è possibile che ciò sia dovuto nel primo caso agli effetti della crisi economica, nel secondo agli effetti della pandemia. Notiamo, infine, che, nonostante la pandemia, si registra nel 2020 un numero di vendite ancora abbastanza elevato.

2 Analisi preliminare

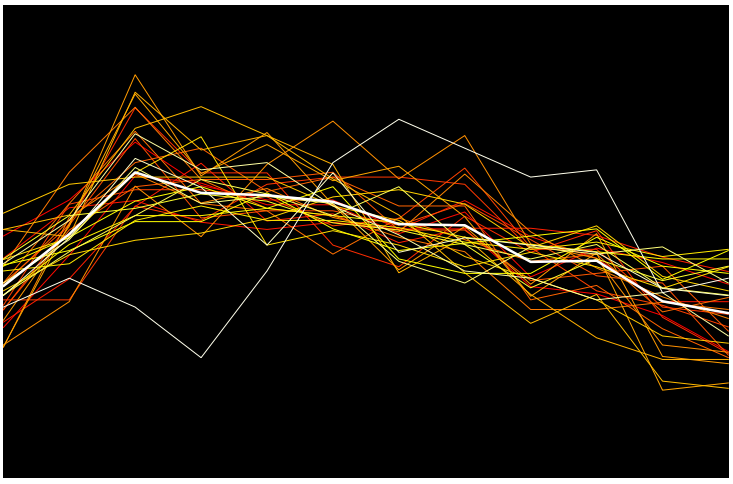
Analizziamo per prima cosa i grafici della funzione di autocorrelazione e della funzione di autocorrelazione al netto del trend al fine di capire quali sono le componenti della serie:



I grafici confermano quanto osservato in precedenza: la serie ha sia una componente di trend che una componente di stagionalità. Notiamo in particolare i seguenti fatti:

- La funzione di autocorrelazione ha un chiaro andamento discendente e oscillatorio. Presenta, inoltre, picchi piuttosto pronunciati in corrispondenza di ciascun periodo (della durata di 12 mesi);
- Il secondo grafico conferma la presenza di stagionalità: la funzione di autocorrelazione della serie detrendizzata ha l'andamento tipico delle serie stagionali. Inoltre, anche in questo caso i valori in corrispondenza di ciascun periodo risultano molto più alti rispetto agli altri valori. Notiamo, infine, la presenza di valori negativi più elevati in corrispondenza della metà di ciascun periodo.

Prima di separare le componenti della serie, confrontiamo graficamente l'andamento negli anni delle vendite, opportunamente riscalate, con l'obiettivo di capire se possiamo considerare uniforme la stagionalità. La curva in bianco rappresenta l'andamento medio:

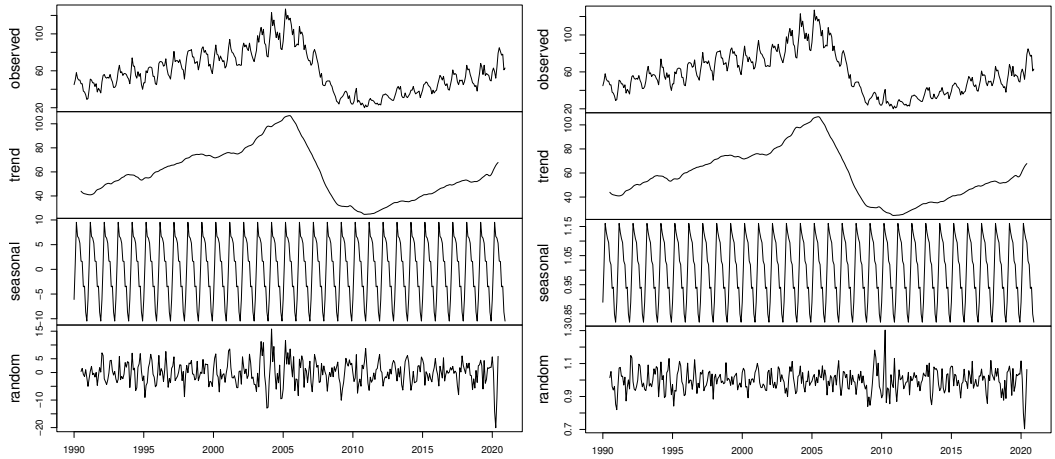


Quanto più il colore delle curve nel grafico si avvicina al rosso, tanto più si va indietro negli anni. Notiamo in particolare i seguenti fatti:

- La curva di colore più chiaro, che rappresenta il 2020, ha un andamento del tutto differente rispetto a quello degli altri anni, cosa che possiamo supporre essere dovuta agli effetti della pandemia ancora in corso. Non ci aspettiamo, dunque, una buona capacità di predizione dei nostri modelli su questo anno;
- Gli anni precedenti al 2020, rumore a parte, hanno un andamento piuttosto simile, ascendente nei primi mesi dell'anno e discendente nei mesi successivi.

In base a quanto osservato possiamo dunque supporre che la stagionalità sia stazionaria.[1]

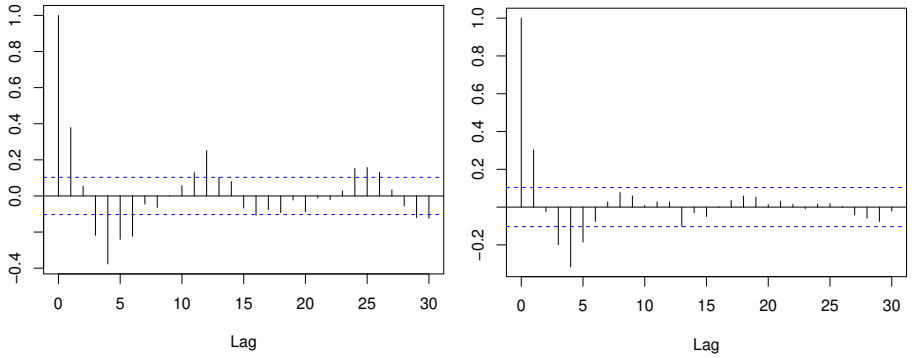
Utilizziamo dunque il comando *decompose* per separare le componenti della serie, e lo facciamo sia nel caso additivo (a sinistra) sia nel caso moltiplicativo (a destra):



Notiamo che in entrambi i casi l'ordine di grandezza dei residui è all'incirca paragonabile a quello della stagionalità. Tuttavia, otteniamo risultati abbastanza soddisfacenti:

- Nel caso additivo la stagionalità è compresa tra -10 e 10 e solo il 2,5% dei residui è in valore assoluto superiore a 10, mentre quasi il 79% dei residui è in valore assoluto minore di 5;
- Per facilitare il confronto con il caso additivo, abbaimo calcolato i logaritmi di stagionalità e residui del modello moltiplicativo. Otteniamo che la stagionalità è compresa tra -0.2 e 0.15 e che solo il 3,3% dei residui è in valore assoluto maggiore di 0.15, mentre il 77,5% è in valore assoluto minore di 0.075.

Per decidere qual è il modello che più si adatta alla serie confrontiamo le funzioni di autocorrelazione dei residui: scegliamo quello i cui residui mostrano meno struttura temporale. Il grafico a sinistra è relativo al modello additivo, quello a destra al modello moltiplicativo:



A differenza dei residui del modello additivo, che sembrano preservare una struttura temporale, i valori della acf dei residui del modello moltiplicativo, ad eccezione dei primi 6, sono interamente contenuti nella banda delimitata dalle due linee tratteggiate, e possono, quindi, essere approssimati a 0. Inoltre, nel primo caso abbiamo una deviazione standard della acf dei residui di 0.254 e una percentuale di varianza non spiegata di 0.035, nel secondo caso la deviazione standard è 0.227 e la percentuale di varianza non spiegata è 0.028.[2] Possiamo dire, dunque, che il modello moltiplicativo è quello che meglio si adatta alla nostra serie.

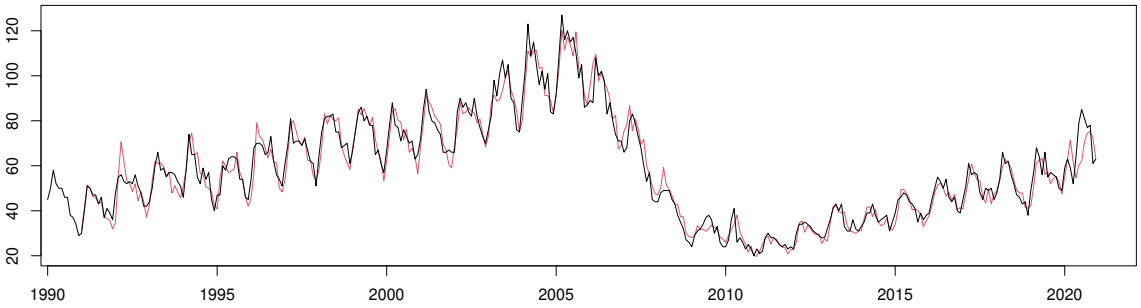
[1] Se anche utilizzassimo per decomporre la serie il comando *stl* (con finestre di dimensione 7 e 15), e lo facessimo sia nel caso additivo che nel caso moltiplicativo (filtrando la serie attraverso la funzione logaritmo), non otterremmo risultati migliori di quelli che si ottengono con il comando *decompose*.

[2] L'analisi "standard" dei residui (plot e istogramma dei residui, Q-Q plot e test di Shapiro) ci fa concludere che in entrambi i casi i residui non sono gaussiani, ma risultano comunque soddisfacenti.

3 Metodo di Holt-Winters

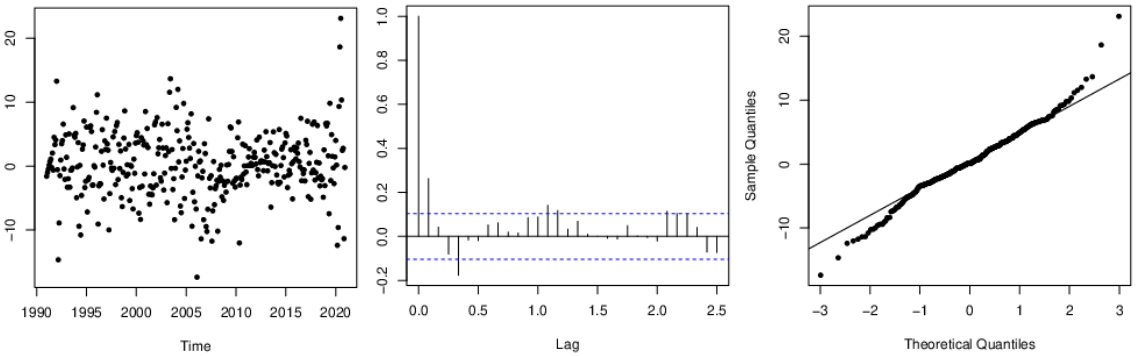
Per provare a predire l'andamento delle vendite del 2021 abbiamo implementato il metodo di smorzamento esponenziale con trend e stagionalità. Abbiamo sostituito i parametri ottimali trovati dal software R con una terna di parametri tale per cui si abbiano meno struttura nei residui e maggior capacità di previsione.[3]

Visualizziamo graficamente come il modello riesce a catturare l'andamento della serie:



Il modello riesce in generale a catturare abbastanza bene l'andamento della serie: quasi ogni anno non coglie alla perfezione l'andamento della curva nei punti in cui si hanno più vendite e meno vendite. Possiamo però ritenerci soddisfatti: data la presenza di rumore nei dati, non vogliamo che il modello catturi troppo bene la serie e che, quindi, filtri troppo poco il rumore.
Per quanto riguarda la parte finale del grafico, che è quella che maggiormente ci interessa, l'analisi risulta soddisfacente fino al 2018. Nel 2019 il metodo tende a semplificare l'andamento della curva, mentre l'andamento delle vendite nel 2020 non viene catturato appieno, come ci aspettavamo: a parte un breve tratto iniziale in cui la stima sembra perfetta, possiamo dire che il numero di vendite è sovrastimato nei primi mesi dell'anno e sottostimato nei mesi restanti.

Procediamo ora con l'analisi dei residui. Vediamo di seguito da sinistra verso destra plot dei residui rispetto al tempo, funzione di autocorrelazione dei residui e Q-Q plot:

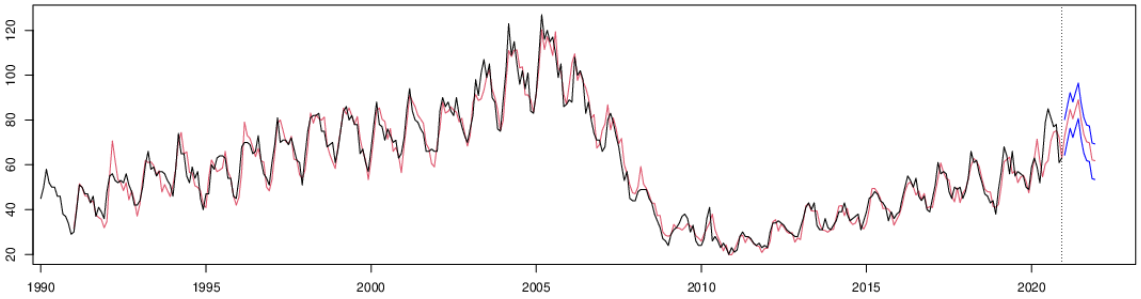


I residui non sono ottimali ma neppure pessimi. Osserviamo in particolare i seguenti fatti:

- Il grafico della acf non mostra la presenza nei residui di struttura temporale, sebbene ci siano alcuni valori non nulli. Inoltre, la deviazione standard della acf è pari a 0.2;
- La proporzione di varianza non spiegata è piuttosto bassa: è pari a 0.05;
- I residui non hanno distribuzione normale: il test di Shapiro ci porta a rigettare l'ipotesi nulla con un p-value di $7,54 \times 10^{-5}$. Inoltre, i punti nel Q-Q plot sono abbastanza addensati sulla diagonale, ma il plot dei residui rivela la presenza di valori più elevati in corrispondenza della metà del grafico e delle parti iniziale e finale, cosa che dà al grafico una forma riconoscibile.

[3]Per ulteriori dettagli sulla scelta dei parametri si rimanda allo script.r allegato alla relazione. Inoltre, non abbiamo modificato i valori dell'intercetta e del coefficiente angolare iniziali: abbiamo provato con una regressione lineare sui primi 20 valori ma non si ottiene alcun miglioramento significativo.

Facciamo ora una previsione sulle vendite del 2021. Nel grafico che segue l'incertezza della previsione (al 95%) è calcolata per via non parametrica, data la non normalità dei residui:



Le previsioni di vendita per il 2021 sono ottimistiche: ogni mese si prevede una vendita di più di 70.000 case unifamiliari. Fanno eccezione i mesi di ottobre, novembre e dicembre in cui il numero di vendite previste scende ma rimane comunque superiore a 60.000.

Per quanto riguarda la capacità di predizione del modello otteniamo i seguenti risultati:

- Se facciamo una predizione a 12 mesi utilizzando come train-set la serie fino al dicembre 2019 e come test-set i dati relativi al 2020, non otteniamo un buon risultato: la radice dell'errore quadratico medio è pari a 17.3.
  La scarsa capacità di predizione in questo caso è dovuta al fatto che a causa della pandemia l'andamento delle vendite nel 2020 è completamente diverso rispetto a quello degli anni precedenti, e risulta quindi praticamente impossibile predirlo basandosi sui soli dati fino al 2019[4];

- Con un'autovalidazione più robusta, e cioè facendo una previsione mese per mese per gli ultimi due anni, otteniamo un risultato migliore: la radice dell'errore quadratico medio è pari a 8.67, non troppo basso ma comunque abbastanza soddisfacente.

4 Metodi autoregressivi

Per provare a predire l'andamento delle vendite negli anni 2021 - 2022 abbiamo implementato quattro diversi metodi autoregressivi: modello diretto ridotto e non, metodo di Yule-Walker e metodo dei minimi quadrati.

Riportiamo nella relazione la sola analisi dettagliata del metodo migliore. La scelta è stata fatta confrontando la proporzione di varianza non spiegata, la deviazione standard della acf dei residui, il p-value restituito dal test di Shapiro e la capacità di predizione a 12 mesi[5]. I risultati ottenuti sono riportati nella seguente tabella:

	Var	**Sd**	**Shap**	**Prev**
Mod Dir r.	0.0516	0.207	0.0567	14.574
Mod Dir	0.0429	0.199	0.0299	13.285
Y-W	0.0428	0.197	0.028	13.327
Min Quad	0.0422	0.198	0.01	12.99

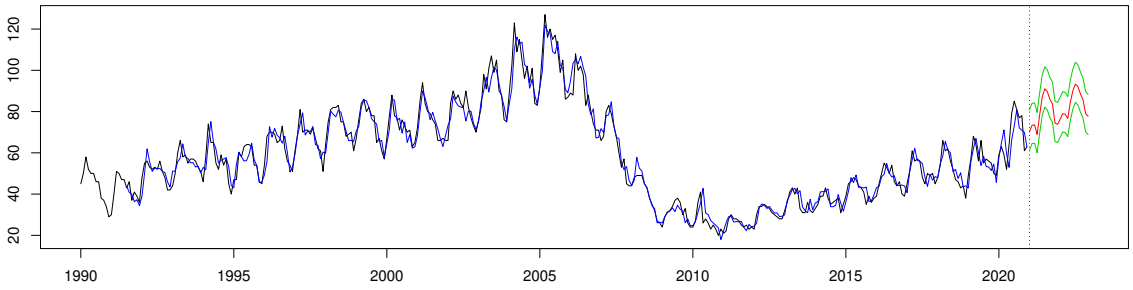
Sebbene il modello diretto ridotto sia l'unico ad avere residui gaussiani, esso è il peggiore in termini di proporzione di varianza non spiegata, capacità di previsione e struttura temporale dei residui.
I restanti metodi sono del tutto paragonabili: scegliamo il metodo dei minimi quadrati, che risulta essere migliore, seppur di poco, in termini di proporzione di varianza non spiegata e capacità di previsione.

[4] A riprova del fatto che il pessimo risultato non sia da attribuire al modello, si verifica che se si fa lo stesso tipo di predizione usando come train-set i dati fino al dicembre 2018 e come test-set i dati relativi al 2019 si ottiene un risultato migliore.

[5] Abbiamo confrontato anche i grafici delle acf dei residui: la acf dei residui del modello diretto ridotto presenta alcuni valori non nulli, invece le acf dei residui degli altri metodi sono approssimabili a 0.

4.1 Metodo dei minimi quadrati

Il seguente grafico mostra i risultati ottenuti dal metodo dei minimi quadrati in termini di analisi e previsione. L'incertezza della previsione è calcolata per via non parametrica, data la non normalità dei residui:

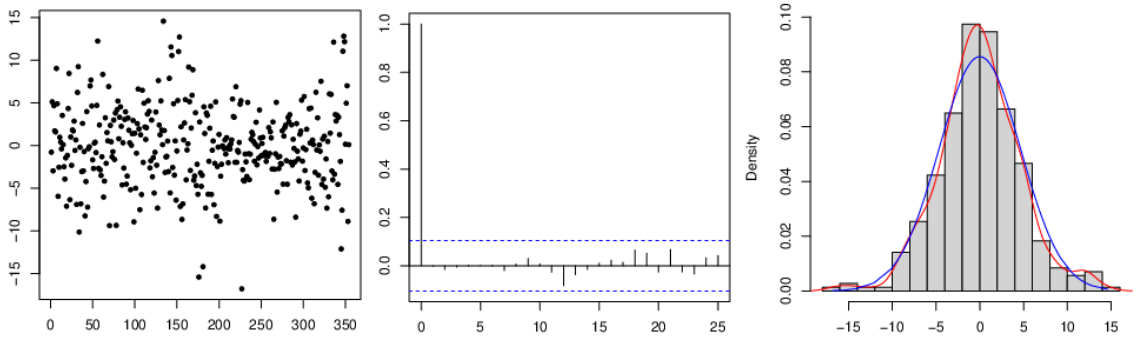


L'analisi segue l'andamento della serie abbastanza fedelmente: anche in questo caso l'andamento della curva nei punti in cui si hanno più vendite e meno vendite non viene colto alla perfezione, ma è comunque soddisfacente.
Rispetto al metodo di Holt-Winters si osserva una maggiore precisione sia nella parte iniziale del grafico che in quella finale: le stime sugli anni che vanno dal 2015 al 2018 sono pressochè simili, quelle relative agli anni 2019-2020 sono più realistiche per il metodo dei minimi quadrati, sebbene siano comunque lontane dall'essere perfette.

Le previsioni per il 2021 sono ottimistiche: si prevede un innalzamento delle vendite a partire dal mese di aprile fino a raggiungere il picco nel mese di maggio. Nei mesi successivi le vendite scendono ma sono comunque quasi sempre superiori a 75.000.
Per il 2022 si prevede un andamento simile, in lieve crescita rispetto al 2021: anche qui si dovrebbe raggiunguere il picco di vendite nel mese di maggio, per poi scendere nei mesi successivi. Ci si mantiene comunque sempre al di sopra delle 75.000 vendite per mese.

Abbiamo già visto che la capacità di predizione a 12 mesi del modello è abbastanza buona, ben migliore di quella del metodo di Holt-Winters: si osserva una radice dell'errore quadratico medio di 13 contro il 17.3 del metodo di Holt-Winetrs.
Anche in questo caso, con un'autovalidazione più robusta, e cioè facendo una previsione mese per mese sugli 11 mesi successivi, il risultato è migliore: la radice dell'errore quadratico medio è di 9.2.

Concludiamo con l'analisi dei residui. Vediamo di seguito da sinistra verso destra plot, funzione di autocorrelazione e istogramma dei residui:



I residui sono soddisfacenti, sebbene non abbiano distribuzione gaussiana. Notiamo in particolare i seguenti fatti:

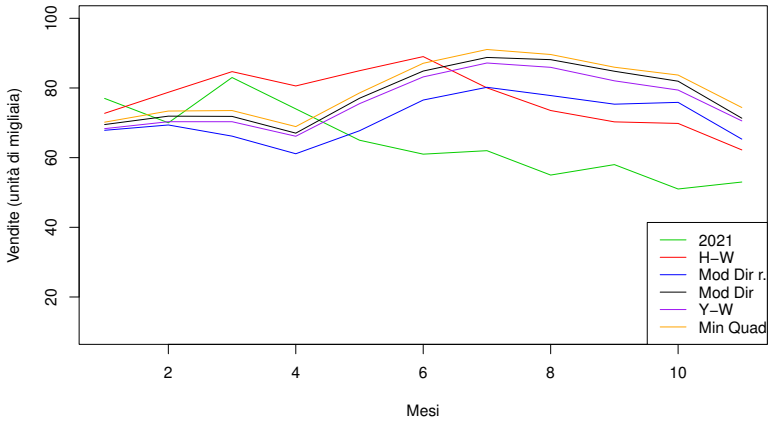
- Il grafico della acf mostra la totale mancanza nei residui di struttura temporale;
- Nell'istogramma dei residui vediamo densità empirica e densità gaussiana in rosso e in blu rispettivamente: le due curve sono piuttosto distinte attorno allo 0, meno

sugli altri valori. La differenza, però, non è sostanziale: il test di Shapiro ci porta a rigettare l'ipotesi nulla con un p-value di solo 0.01.

Rispetto al metodo di Holt-Winters, con il metodo dei minimi quadrati si ottengono risultati migliori in termini di residui. Entrambi i residui non hanno distribuzione gaussiana, ma nel primo caso il test di Shapiro ci porta a rigettare l'ipotesi nulla con un p-value ben più basso: 7.54×10^{-5} contro lo 0.01 del metodo dei minimi quadrati.
Inoltre, mentre nel primo caso la acf dei residui presenta alcuni valori non nulli, nel secondo caso è del tutto approssimabile a 0. Infine, nel primo caso si ottiene una proporzione di varianza non spiegata di 0.05, nel secondo di 0.042.

5 Le previsioni per il 2021 a confronto

Confrontiamo le previsioni per il 2021 dei cinque metodi con i valori noti (dal mese di gennaio al mese di novembre):



Nessuno dei metodi implementati è riuscito a prevedere correttamente l'andamento delle vendite del 2021. Osserviamo in particolare i seguenti fatti:

- I quattro metodi autoregressivi fanno previsioni molto simili: riescono ad essere precisi solo sul mese di febbraio; tendono invece a sottostimare il numero di vendite fino al mese di maggio e a sovrastimarlo nei restanti mesi;
- Il metodo di Holt-Winters, ad eccezione del mese di gennaio, tende in generale a sovrastimare il numero di vendite: si riscontra una precisione maggiore da gennaio ad aprile, minore nei restanti mesi.

6 Conclusioni

Al fine di capire se e con quanta precisione l'andamento passato delle vendite di case unifamiliari negli Stati Uniti consenta di predire il numero di vendite future, abbiamo implementato cinque diversi metodi di analisi e previsione (Holt-Winters, autoregressivo diretto ridotto e non, Yule-Walker e minimi quadrati). Inoltre, con il metodo dei minimi quadrati, che abbiamo decretato essere il migliore tra gli autoregressivi, abbiamo anche provato a predire un possibile andamento delle vendite per il 2022.

Il confronto delle previsioni per il 2021 dei nostri metodi con i valori noti non dà buoni risultati. Possiamo dire che in realtà ci aspettavamo questi errori di previsione, ma abbiamo visto che non sono dovuti allo scarso funzionamento dei nostri metodi, quanto al diverso andamento negli ultimi due anni del settore immobiliare considerato, a causa della pandemia ancora in corso.

Al momento risulta tutt'altro che facile fare previsioni affidabili sugli anni successivi al 2021. Nell'ultimo anno il numero di vendite è sceso, ma non in modo drastico. Il mercato immobiliare è sempre attivo, ma sarebbe opportuno aspettare che la situazione si ristabilizzi prima di fare investimenti di grosso calibro.