Analisi del numero di nascite in Francia a partire dal 1994

Alessandro La Farciola

28 dicembre 2021

1 Introduzione

1.1 Contesto e prospettiva

L'obiettivo della nostra analisi è quello di indagare l'andamento delle nascite in Francia dal 1994 al 2020 e provare a fare una previsione per l'anno 2021. Analisi di questo tipo sono svolte, ad esempio, dagli istituti di statistica nazionale per studiare i fenomeni demografici che caratterizzano un Paese.

1.2 Presentazione del *dataset* e link alle tabelle di dati

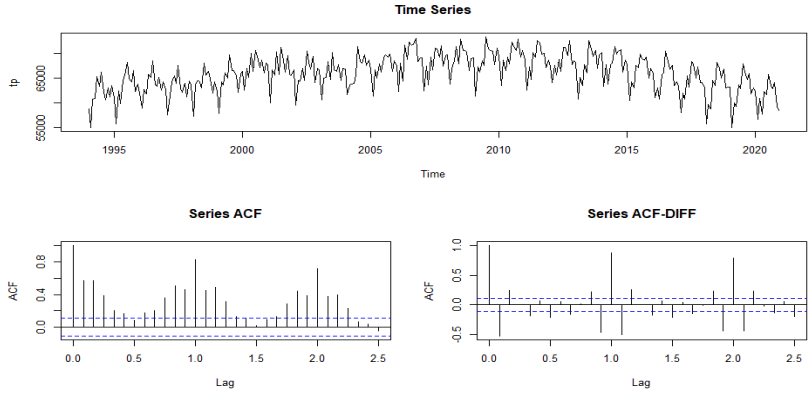
La tabella di dati da analizzare ha come osservazioni il numero di nascite avvenute in Francia a partire da Gennaio 1994 fino a Dicembre 2020, con occorrenza mensile.

I dati selezionati per la nostra analisi sono stati interamente reperiti sul sito *INSEE*, *Institut national de la statistique et des études économiques*. In particolare, è possibile accedere direttamente ad essi attraverso il seguente link: `https://www.insee.fr/en/statistiques/serie/001641601#Telechargement`.

2 Analisi

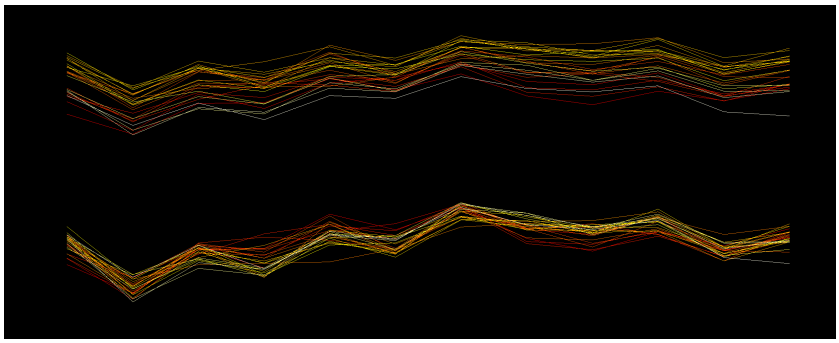
2.1 Indagine preliminare

La tabella selezionata per la nostra analisi è una serie storica che possiamo visualizzare graficamente nella seguente figura, insieme alle informazioni sull'autocorrelazione.



Osservando i grafici, emerge fin da subito la presenza di una forte stagionalità ricorrente. Ciò risulta evidente sia dalla visualizzazione della serie stessa, sia dall'andamento della funzione di autocorrelazione che presenta picchi significativi in corrispondenza del periodo. Inoltre, tenuto conto dell'occorrenza mensile dei dati, si è ritenuto opportuno assegnare un periodo pari a *12*.

D'altro canto, dal grafico che mostra l'andamento della serie si può notare un leggero trend, inizialmente crescente e poi decrescente, che può suggerirci di analizzare sia il trend che la stagionalità della serie. Tale osservazione può essere confermata, innanzitutto, dai seguenti grafici, in cui viene mostrato l'andamento annuale delle nascite. In particolare, in alto si mostra l'andamento annuale in cui ogni periodo ha la propria media, mentre in basso questi sono centrati. In aggiunta, i colori più scuri fanno riferimento ad anni più remoti, mentre quelli più chiari si riferiscono ad anni recenti.



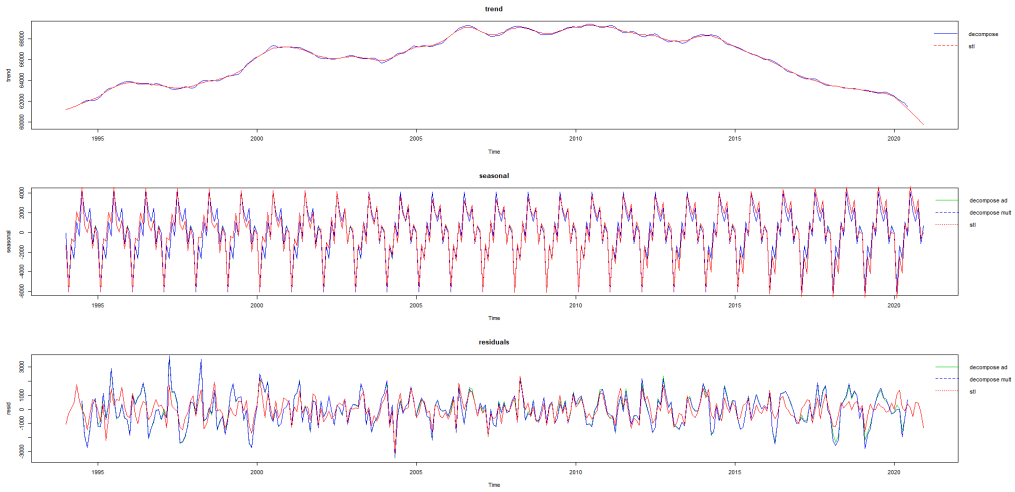
É possibile, infatti, notare nel primo grafico una lieve differenza di altezza tra i vari andamenti annuali. Questa viene leggermente annullata nel secondo, in cui i periodi sono centrati. Ciò evidenzia la presenza, lieve ma non trascurabile, di un trend della serie. La seconda figura, invece, dimostra evidentemente la presenza di stagionalità, poiché l'andamento annuale del numero di nascite si ripete in maniera molto fedele.

Infine, osservando i colori nel secondo grafico, si può evidenziare un andamento leggermente differente degli anni più remoti in corrispondenza dei mesi di *Aprile*, *Maggio*, *Agosto*, *Settembre*. In particolare, si evidenziano, rispetto ai dati più recenti, un calo di nascite minore nei mesi di *Aprile* e *Maggio* e uno più netto nei mesi *Agosto* e *Settembre*. Questo può suggerire una piccola differenza di stagionalità al passare degli anni, che necessita un maggiore approfondimento.

2.2 Decomposizione

Alla luce delle precedenti osservazioni, si procede con una decomposizione della serie. In particolare, si vuole mettere a confronto la decomposizione additiva, quella moltiplicativa (entrambe ottenute con il comando *decompose*) e quella ottenuta dal comando *stl* in cui si suppone una stagionalità non stazionaria[1]. In particolare, nella seguente figura, si mostrano sovrapposti il *trend*, la *stagionalità* e i *residui* ottenuti con i tre metodi.

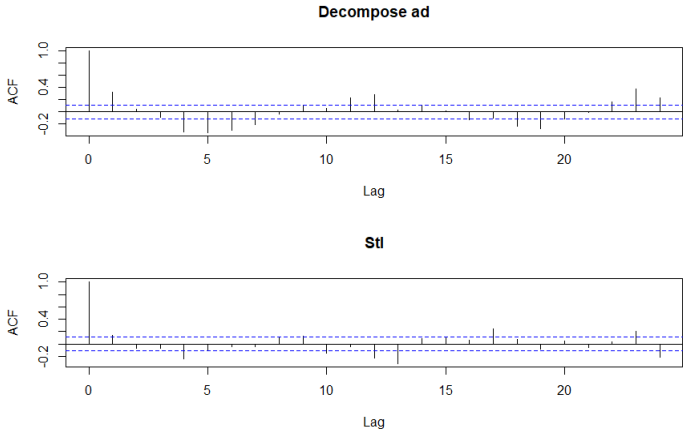
[1]É stata fatta un'analisi preliminare in cui si è scelto il parametro *s.window* pari a 9. Inoltre, per ora stiamo supponendo un modello additivo per la decomposizione mediante *stl*.



É evidente come i due metodi ottenuti con il comando *decompose* si equivalgono: infatti, a parte una differenza trascurabile nei residui, non si notano ulteriori cambiamenti. Per quanto riguarda la decomposizione ottenuta con il comando *stl*, essa sembra paragonabile agli altri sia sul trend che sulla stagionalità, mentre i residui mostrano maggiori differenze. In tutti i casi, però, la decomposizione ottenuta non è banale (basta guardare la scala di grandezza di stagionalità e residui) e pertanto si può procedere con un'analisi dei residui per scegliere il metodo più opportuno.

2.2.1 Analisi dei residui[2]

Innanzitutto, è necessario analizzare la funzione di autocorrelazione sulla serie dei residui per indagare se è presente una struttura temporale su di essi.



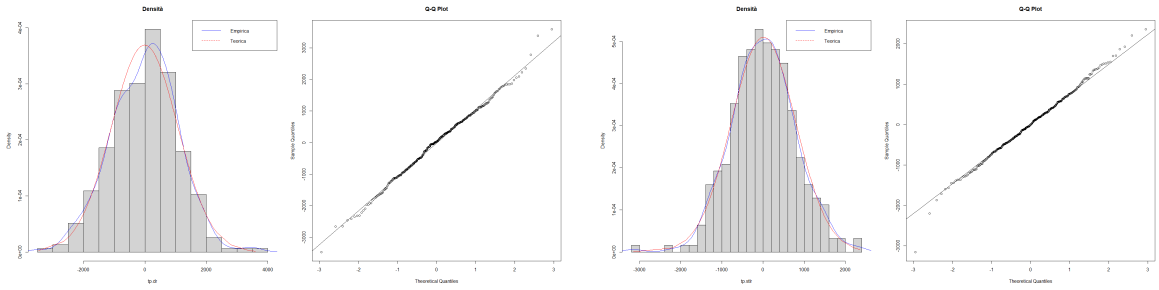
	sd acf decompose resid	sd acf stl resid
1	0.29	0.248

Dai grafici precedenti, si può osservare che in entrambi i residui vi è ancora una struttura temporale, che tuttavia non è pronunciata e quindi possiamo considerarla trascurabile. Inoltre, sembrerebbe

[2]Per motivi di spazio e chiarezza, si è preferito mostrare il confronto tra i due modelli additivi (decompose/stl) data la assoluta paragonabilità tra il modello additivo e quello moltiplicativo.

leggermente preferibile il metodo ottenuto con *stl*, come dimostra il valore della deviazione standard presentato.

Per realizzare una scelta più robusta si procede con l'analisi dei residui vera e propria.



In entrambi i casi, sia la densità che il grafico *quantile-quantile* mostrano una significativa aderenza dei residui ad una distribuzione gaussiana. Ciò, confermato anche dal risultato del test statistico Shapiro-Wilk, ci permette di accettare tali residui come componente di rumore casuale dei dati. Inoltre, alla luce delle figure precedenti, si ritiene il modello *stl* più efficace nel catturare la decomposizione della nostra serie storica, confermando l'osservazione di una stagionalità non stazionaria fatta in precedenza.

2.3 Calibrazione dei modelli

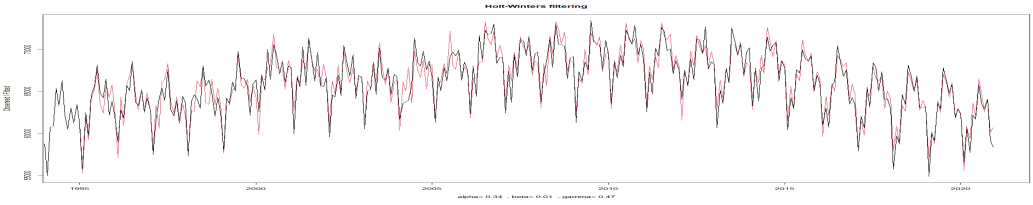
2.3.1 Metodi di Holt-Winters

Il nostro obiettivo, ora, è quello di calibrare un modello che riesca a rappresentare bene la serie, al netto del rumore, per realizzare un'accurata previsione per l'anno successivo. Si procede con i metodi di Holt-Winters e, in particolare, data la natura della serie già discussa, si cerca un modello con trend e stagionalità di tipo additivo[3].

Calibrando il modello attraverso i metodi di ottimizzazione utilizzato dal software si ottengono i seguenti parametri:

- $\alpha = 0.341$
- $\beta = 0.014$
- $\gamma = 0.466$

Per tutti i modelli, inoltre, si procede con una *mini* regressione lineare preliminare per determinare una pendenza e un'intercetta iniziale. Come si vede dal grafico successivo, il modello realizzato dal software tende a seguire la stagionalità della serie.



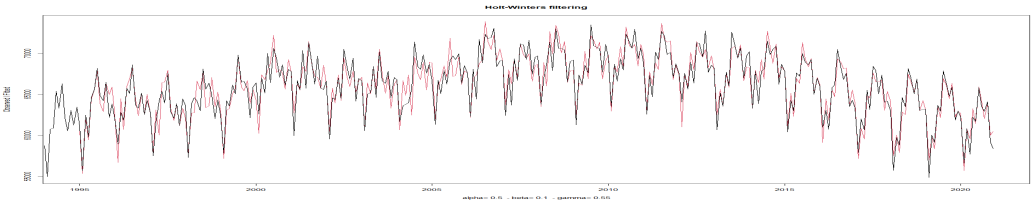
[3]In realtà, di fronte ad una scelta della decomposizione non netta, sono stati testati anche i modelli moltiplicativi ma i risultati, coerentemente con quanto detto prima, non si discostavano da quelli ottenuti con il metodo additivo. Pertanto, in questa fase, si è preferito considerare solo il modello additivo, così da poter scegliere successivamente, per il confronto, altri modelli che mostravano differenze più significative (evitando un confronto troppo numeroso e di conseguenza poco chiaro).

Questo risultato è coerente con la natura della nostra serie, in quanto la componente stagionale è molto più determinante rispetto al trend. Ciò si rispecchia anche nella scelta dei parametri: β, che determina la componente di trend nel modello, è vicino a 0.

Con un'analisi più approfondita svolta manualmente facendo variare i parametri, però, è possibile ottenere un modello che riesca a cogliere maggiormente, anche se di poco, i picchi negli anni più recenti[4]. In questo caso i parametri che realizzano tale modello sono:

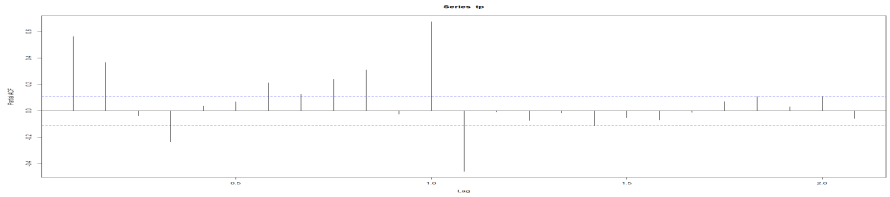
- $\alpha = 0.5$
- $\beta = 0.1$
- $\gamma = 0.55$.

Il grafico successivo mostra tale modello.



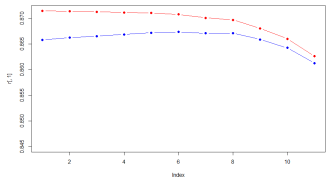
2.3.2 Metodi autoregressivi

Implementazione diretta Per il modello autoregressivo con implementazione diretta, è necessario, innanzitutto, analizzare la funzione di autocorrelazione parziale per decidere il numero di lag precedenti da cui vi è dipendenza.



Il grafico mostra una chiara dipendenza fino a 13 lag. Siamo, ora, pronti per implementare un modello regressivo. Le figure successive mostrano i risultati della regressione completa e del modello ristretto, con l'andamento della varianza spiegata (e varianza spiegata corretta) in rosso (in blu, rispettivamente) nel processo di riduzione del modello.

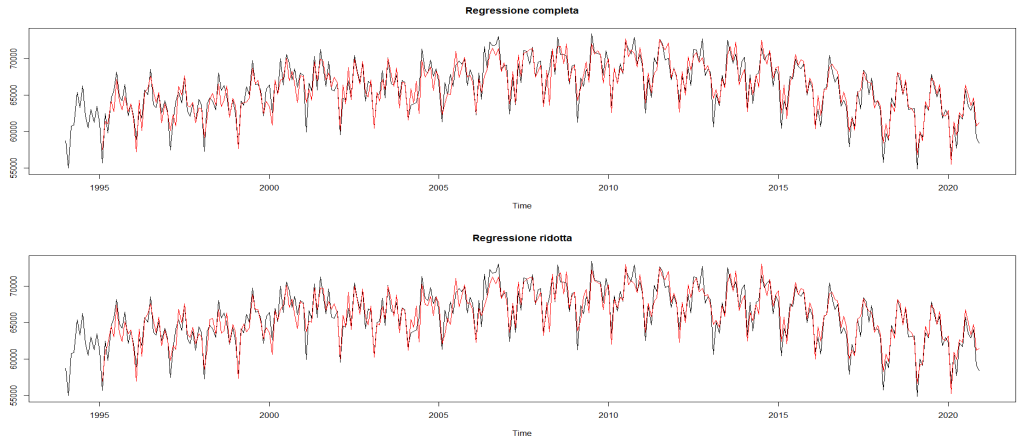
```
Call:
lm(formula = X14 ~ ., data = mtp)

Residuals:
     Min       1Q   Median       3Q      Max
-4016.0  -777.2   -26.1   765.7  4995.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  885.46429 2215.92752   0.400 0.689745
X1            -0.47322    0.05103  -9.273  < 2e-16 ***
X2             0.82987    0.03392  24.465  < 2e-16 ***
X3            -0.05101    0.03487  -1.463 0.144516
X4             0.02859    0.03474   0.823 0.411064
X5             0.04543    0.03481   1.305 0.192910
X6             0.04062    0.03475   1.169 0.243304
X7            -0.01236    0.03475  -0.356 0.722386
X8            -0.01840    0.03481  -0.528 0.597571
X9            -0.01030    0.03483  -0.296 0.767566
X10           -0.06578    0.03481  -1.890 0.059791 .
X11            0.02217    0.03480   0.637 0.524576
X12            0.13216    0.03402   3.885 0.000126 ***
X13            0.31849    0.05161  10.047  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1317 on 297 degrees of freedom
Multiple R-squared:  0.8715,    Adjusted R-squared:  0.8658
F-statistic: 154.9 on 13 and 297 DF,  p-value: < 2.2e-16
```



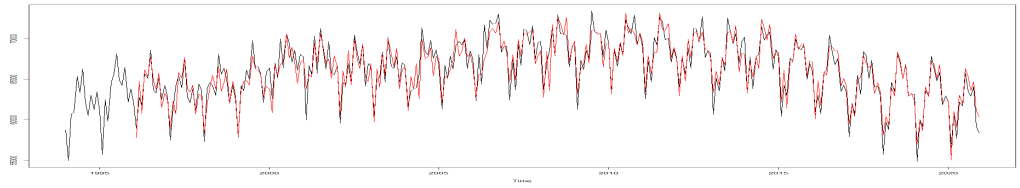

```
Call:
lm(formula = X14 ~ . - X9 - X7 - X11 - X8 - X4 - X5 - X3 - X6 -
    X10 - X12, data = mtp)

Residuals:
     Min       1Q   Median       3Q      Max
-4063.3  -881.0   -70.8   751.2  4851.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2769.29714 1633.73209   1.695   0.0911 .
X1            -0.51990    0.04943 -10.517   <2e-16 ***
X2             0.90131    0.02557  35.248   <2e-16 ***
X13            0.57640    0.04758  12.115   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1339 on 307 degrees of freedom
Multiple R-squared:  0.8626,    Adjusted R-squared:  0.8613
F-statistic: 642.5 on 3 and 307 DF,  p-value: < 2.2e-16
```

[4] A parte i modelli che palesemente sbagliavano, gli altri erano piuttosto simili e pertanto, per motivi di chiarezza, si è preferito prendere in analisi i parametri scelti dal software e quelli che si sono ritenuti i migliori, confermati anche da parte di una rapida analisi dei residui e capacità di predizione (che si omette per motivi di spazio: si è preferito dare precedenza al confronto finale).

Mentre, i modelli ottenuti sono i seguenti.



Metodo dei *minimi quadrati* Visti i risultati ottenuti nella decomposizione, può essere utile implementare un modello con il metodo dei *minimi quadrati*, in cui si tiene conto di una stagionalità non stazionaria.



Osservando il grafico, sembrerebbe emergere una minore efficacia del modello. Tuttavia, per giudizi più precisi si procede con il confronto tra modelli.

3 Confronto tra modelli

Dopo aver implementato diversi modelli con metodi differenti siamo pronti per scegliere quello più adatto attraverso un'analisi dei residui e testando la loro capacità di predizione.

Innanzitutto, si ritiene più efficace un modello che riesca a catturare bene i picchi stagionali, sia perché questi sono ritenuti parte strutturale della serie (e non rumore), sia perché risultano più appropriati ai fini della nostra analisi. Ciò è emerso già dall'indagine preliminare in cui, sovrapponendo l'andamento di nascite annuale, si osservava sempre la stessa struttura.

Tutti i modelli considerati finora, inoltre, tendono a sbagliare di più all'inizio della serie. Ciò, però, non incide sulla bontà di tali modelli, perché a noi interessa che riesca a cogliere bene l'andamento (e principalmente la stagionalità) negli anni più recenti, così da poter ottenere una previsione sul futuro in maniera più efficace.

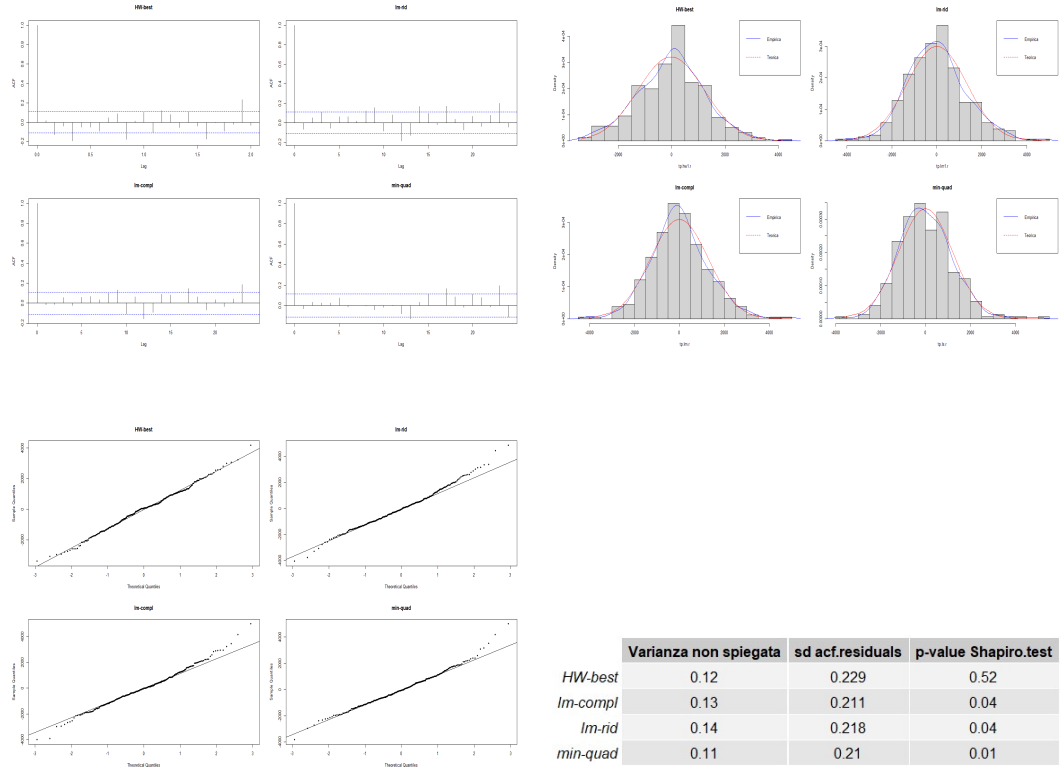
I metodi scelti per il confronto sono i seguenti:

- *HW-best*: Holt-Winters con i parametri $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.55$;
- *lm-compl*: modello autoregressivo completo con implementazione diretta;

- *lm-rid*: modello autoregressivo ridotto con implementazione diretta;
- *min-quad* modello autoregressivo implementato con la tecnica dei *minimi quadrati*.

3.0.1 Analisi dei residui

Innanzitutto, si procede con un'analisi dei residui. In questo caso, testiamo la *gaussianità* del rumore catturato dal modello al fine di determinarne la loro casualità.



	Varianza non spiegata	sd acf.residuals	p-value Shapiro.test
HW-best	0.12	0.229	0.52
lm-compl	0.13	0.211	0.04
lm-rid	0.14	0.218	0.04
min-quad	0.11	0.21	0.01

Questo tipo di analisi non sembra fornirci una risposta definitiva, in quanto i grafici non evidenziano drastiche differenze. La varianza non spiegata è paragonabile in tutti i casi, mentre il *p-value* restituito dal test statistico *Shapiro-Wilk* sembra suggerirci una maggiore aderenza nel caso *HW-best*. Osservando più attentamente, infatti, sia il grafico sulla densità empirica che il *qq-plot* mostrano una maggiore aderenza proprio in quel caso. Tuttavia, la funzione di autocorrelazione applicata ai residui evidenzia una maggiore struttura temporale proprio per il caso *HW-best*, anche se non si ritiene così significativa da invalidare il modello.

Alla luce di tali risultati, si reputa, quindi, determinante l'autovalidazione ai fini della scelta del modello più efficace.

3.0.2 Autovalidazione

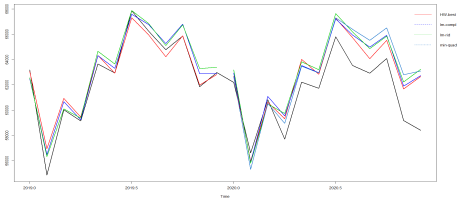
Per testare la validità di predizione del modello si è proceduto nel seguente modo:

1. Utilizzando come dati di addestramento del modello la serie storica fino all'intero 2018, si è proceduto con una predizione dell'anno 2019;

2. Analogamente, addestrando il modello fino all'intero 2019, si è proceduto con una predizione dell'anno 2020;

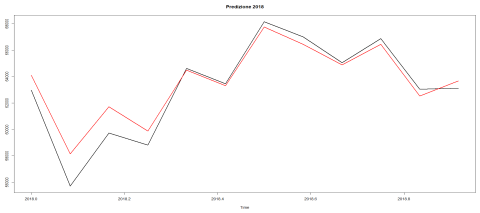
3. Infine, si è realizzata un'autovalidazione più robusta in cui la predizione, mese per mese, sfrutta il modello addestrato fino al mese precedente[5].

Per quanto riguarda i primi due punti precedenti, la seguente figura mostra le predizioni ottenute con i quattro metodi e lo scarto quadratico medio per entrambi gli anni testati.



	2019	2020
HW-best	730.15	1883.15
lm-compl	738.93	2028.86
lm-rid	790.25	2246.91
min-quad	1284.2	2321.81

Innanzitutto, si può osservare che le predizioni ottenute non sono radicalmente diverse le une dalle altre, tuttavia alla luce dei risultati ottenuti, il metodo più efficace sembrerebbe *HW-best*, coerentemente con quanto visto anche nell'analisi dei residui. Inoltre, si può sottolineare come la predizione per l'anno 2019 sia molto più precisa rispetto al 2020, dove tutti i modelli tendono a sovrastimare l'andamento. Questo risultato potrebbe essere approfondito, per provare a capire da cosa derivi tale errore, ipoteticamente riconducibile a diverse cause. La prima potrebbe essere una scarsa efficacia del modello scelto (anche se nel 2019 si era comportato nettamente meglio), magari dovuta al fatto che il trend discendente degli ultimi anni possa essere più pronunciato[6]. D'altro canto, potrebbe essere che, per il fenomeno che stiamo analizzando, il 2020 rappresenti un'annata particolarmente difficile da prevedere (dovuta, probabilmente, agli effetti della pandemia). Per provare a dare una risposta aggiuntiva si potrebbe vedere la previsione anche per l'anno 2018.



La previsione per l'anno 2018 con il metodo che abbiamo ritenuto più efficace (opportunamente addestrato fino a Dicembre 2017) sembra perfino migliore, dato anche uno scarto quadratico medio pari a 403.4, minore a quelli ottenuti per l'anno successivo. Questo potrebbe avvalorare la seconda ipotesi di peculiarità dell'anno 2020.

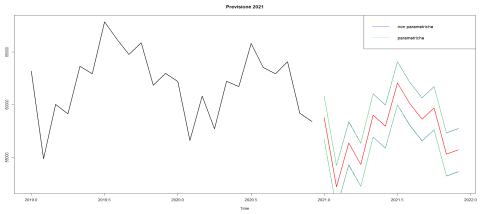
[5]Per motivi di spazio non si mostrano i risultati di tale confronto perché del tutto analoghi alle precedenti considerazioni.

[6]Ricordiamo che abbiamo calibrato il modello più sulla cattura della stagionalità, vista la netta preponderanza di tale componente nella natura della serie.

4 Conclusioni

L'andamento delle nascite, come già ampiamente osservato, assume una chiara struttura stagionale, con picchi più alti in corrispondenza dei mesi più lunghi. Si nota, però, un certo trend nel corso dell'anno con un massimo raggiunto nel mese di luglio. É proprio tale valore che si è cercato di prevedere meglio attraverso la calibrazione del modello, soprattutto per i metodi di *Holt-Winters*.

Il confronto tra modelli sviluppato al paragrafo precedente ha portato a scegliere come quello più efficace il metodo *HW-best*. Siamo, quindi, pronti per una predizione definitiva per l'anno 2021, in cui si mostrano sia le incertezze non parametriche che quelle parametriche, data l'ipotesi di gaussianità dei residui.



In conclusione, analizzando i dati reali che attualmente abbiamo a disposizione fino al mese di Ottobre 2021, è possibile mostrare che la predizione ottenuta non è molto soddisfacente, in quanto il modello realizzato sottostima l'andamento reale. Ciò è dovuto al fatto che l'anno precedente, come già visto, ha rappresentato un'anomalia evidente: i dati di fine 2020 e Gennaio 2021 sono nettamente più bassi di quanto il modello aveva previsto, probabilmente a causa del lockdown e della diffusione della pandemia. Il metodo di HW da noi scelto, infatti, presentava come parametro α un valore non trascurabile e di conseguenza il passato più recente assume una valenza importante nella predizione. Il grafico successivo, quindi, dimostra che lo stesso modello calibrato fino al 2019 assume una efficacia maggiore (seppur con evidenti limiti, soprattutto per il dato di Gennaio) per il 2021 rispetto a quello in cui si tiene conto anche dei dati relativi al 2020.