

# Analisi Mortalità Popolazione Norvegese

Andrea Mauro

2023-06-19

Nel seguente lavoro viene presentata l'applicazione di strumenti utili all'analisi e proiezione della mortalità: in particolare il progetto considera la popolazione Norvegese, le informazioni riguardanti il paese scandinavo sono state fornite dal'HMD, un database online gratuito che raccoglie dati sulla mortalità e sulla speranza di vita provenienti da fonti ufficiali di tutto il mondo. Il database contiene informazioni su oltre 40 paesi e fornisce dati storici e attuali sulla mortalità e sulla sopravvivenza, a partire dal 1950 o addirittura prima, fino ad oggi.

Dopo una fase iniziale di pre-processamento per preparare e definire gli aspetti principali da considerare, verranno definiti e stimati diversi modelli stocastici di mortalità. Questi modelli saranno valutati e selezionati utilizzando criteri standard ampiamente riconosciuti nella letteratura scientifica. Successivamente, basandosi sui modelli selezionati, verranno sviluppate le proiezioni, tenendo debitamente conto delle diverse fonti di incertezza che possono influenzare le previsioni

## Import Data

I dati vengono importati utilizzando la funzione 'read.demogdata' messa a disposizione dalla libreria 'Demography', pacchetto ampiamente utilizzato per l'analisi demografica, nel dettaglio fornisce molte funzioni utili per l'elaborazione, l'analisi e la visualizzazione dei dati demografici. Le informazioni importate riguardano la popolazione Norvegese tra gli anni 1846 e 2022, comprendendo soggetti di sesso maschile e femminile la cui età varia tra 0 e 110 anni.

```
# Importazione dati demografici HMD -----

data <- read.demogdata(file = "Mx_1x1.txt",
                      popfile = "Exposures_1x1.txt",
                      type = "mortality",
                      label = "Norway")

data

## Mortality data for Norway
##   Series: female male total
##   Years: 1846 - 2022
##   Ages:  0 - 110
```

## Pre-processing

Prima di procedere con la stima dei modelli di mortalità vengono definiti dei parametri per agevolare lo sviluppo del codice, in particolare definiamo età minima e massima utili alla definizione del vettore contenente le età da stimare, l'intervallo degli anni presi in considerazione per la stima(inizio:1965,fine:2019) e l'orizzonte temporale delle proiezioni.

```
# Data setting -----

a.min <- 0 # età minima dell'intervallo di fit
a.max <- 100 # età massima dell'intervallo di fit
A.fit <- c(a.min:a.max)
y.fit.min <- 1965 # anno minimo dell'intervallo di fit
y.fit.max <- 2019 # anno massimo dell'intervallo di fit
Y.fit <- c(y.fit.min:y.fit.max)
y.pred <- 30 # orizzonte di proiezione
Y.pred <- c((y.fit.max+1):(y.fit.max+y.pred))
```

Il problema di stima viene affrontato considerando la differenza di genere nella popolazione Novegese, per cui i modelli presi in esame vengono stimati in parallelo per la popolazione maschile e femminile, prima di procedere con la stima si rende necessaria l'estrapolazione delle esposizioni centrale dal dataframe in formato demogdata, e successivamente utilizzare le esposizioni centrali per ottenere le esposizioni iniziali, nello specifico questo viene effettuato per le due sotto-popolazioni.

Un ulteriore step di preparazione per la fase di stima è la costruzione della matrice dei pesi, costruita sull'età e gli anni inclusi nel problema di stima, nella matrice viene impostato a 0 il peso delle coorti con meno di 4 osservazioni. Questo passaggio è cruciale per la stima dei modelli che contengono il termine dell'effetto di coorte.

Infine vengono calcolati i tassi di mortalità utilizzando le esposizioni centrali, successivamente vengono selezionati solo quelli di interesse per il problema di stima.

```
# Handling data -----

# Estrazione esposizioni

# Centrali -----
datamStMoMoC <- StMoMoData(data, series = "male")
datafStMoMoC <- StMoMoData(data, series = "female")

# Iniziali-----
datamStMoMoI <- central2initial(datamStMoMoC)
datafStMoMoI <- central2initial(datafStMoMoC)
# Generazione matrice dei pesi
wxt <- genWeightMat(ages = A.fit, years = Y.fit, clip = 4)

# Tassi di mortalità

# -Male
mRates <- datamStMoMoC$Dxt/datamStMoMoC$Ext
mRates <- mRates[A.fit+1,tail(datamStMoMoC$years+1,
                             length(Y.fit))-datamStMoMoC$years[1]]

# -Female
fRates <- datafStMoMoC$Dxt/datafStMoMoC$Ext
fRates <- fRates[A.fit+1,tail(datafStMoMoC$years+1,
                             length(Y.fit))-datafStMoMoC$years[1]]
```

## Modelling

In questa sezione viene sviluppato il problema di stima, nel dettaglio vengono stimati diversi modelli di mortalità in parallelo sia per la popolazione femminile che per quella maschile. I modelli vengono quindi

valutati in base alla capacità di approssimare i tassi di mortalità osservati. I modelli presi in considerazione sono i seguenti:

Model	Predictor
LC	$\eta_{xt} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)}$
CBD	$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)}$
APC	$\eta_{xt} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$
RH	$\eta_{xt} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \gamma_{t-x}$
M7	$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + ((x - \bar{x})^2 - \hat{\sigma}_x^2) \kappa_t^{(3)} + \gamma_{t-x}$
PLAT	$\eta_{xt} = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x) \kappa_t^{(2)} + \gamma_{t-x}$

Figure 1: Model Structures

L'assunzione alla base dei modelli riguarda la componente randomica, in particolare si assume che il numero di morti segua una distribuzione Binomiale, quindi:

$$D_{xt} \sim \text{Binomial}(E_{0xt}, q_{xt}) \text{ con } E(D_{xt}/E_{0xt}) = q_{xt}$$

Data l'assunzione precedente la funzione  $g$  che collega la componente randomica e quella sistematica  $\eta_{xt}$  è la funzione "logit link":

$$g(E(D_{xt}/E_{0xt})) = \text{logit}(E(D_{xt}/E_{0xt})) = \eta_{xt}$$

Il problema di stima, per come specificato nella fase di pre-processing, include un indice  $x$  (età) che assume valori compresi nell'intervallo 0-100, mentre  $t$  (anno di calendario) assume valori a partire dal 1965 fino al 2019.

```
# Models Fitting -----

# Lee Carter

# - Male
LCfit_m <- fit(lc(link = "logit"),
               data = datamStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# - Female
LCfit_f <- fit(lc(link = "logit"),
               data = datafStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# Renshaw-Haberman

# - Male
RHfit_m <- fit(rh(link = "logit", cohortAgeFun="1"),
               data = datamStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt,
               start.ax = LCfit_m$ax,
```

```

        start.bx = LCfit_m$bx,
        start.kt = LCfit_m$kt)

# - Female
RHfit_f <- fit(rh(link = "logit", cohortAgeFun="1"),
              data = datafStMoMoI,
              ages.fit = a.min:a.max,
              years.fit=y.fit.min:y.fit.max,
              wxt = wxt,
              start.ax = LCfit_f$ax,
              start.bx = LCfit_f$bx,
              start.kt = LCfit_f$kt)

# APC

# - Male
APCfit_m <- fit(apc(link = "logit"),
               data = datamStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# - Female
APCfit_f <- fit(apc(link = "logit"),
               data = datafStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# CBD

# - Male
CBDfit_m <- fit(cbd(),
               data = datamStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# - Female
CBDfit_f <- fit(cbd(),
               data =datafStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# M7

# - Male
M7fit_m <- fit(m7(link = "logit"),
               data = datamStMoMoI,
               ages.fit = a.min:a.max,
               years.fit=y.fit.min:y.fit.max,
               wxt = wxt)

# - Female
M7fit_f <- fit(m7(link = "logit"),
               data = datafStMoMoI,

```

```

        ages.fit = a.min:a.max,
        years.fit=y.fit.min:y.fit.max,
        wxt = wxt)

# PLAT

# - Male
PLATfit_m <- fit(PLAT,
                data = datamStMoMoI,
                ages.fit = a.min:a.max,
                years.fit=y.fit.min:y.fit.max,
                wxt = wxt)

# - Female
PLATfit_f <- fit(PLAT,
                data = datafStMoMoI,
                ages.fit = a.min:a.max,
                years.fit=y.fit.min:y.fit.max,
                wxt = wxt)

```

## Goodness-of-fit analysis

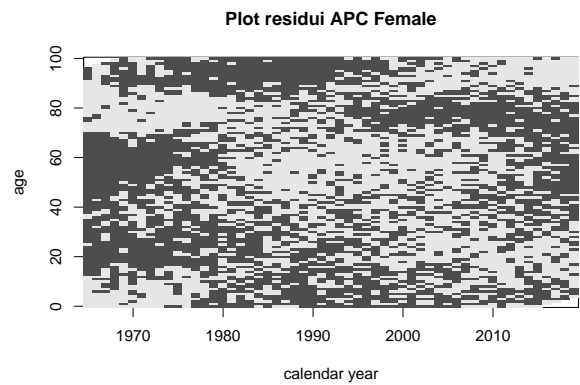
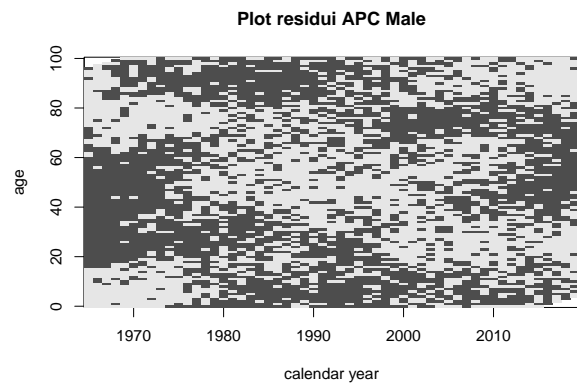
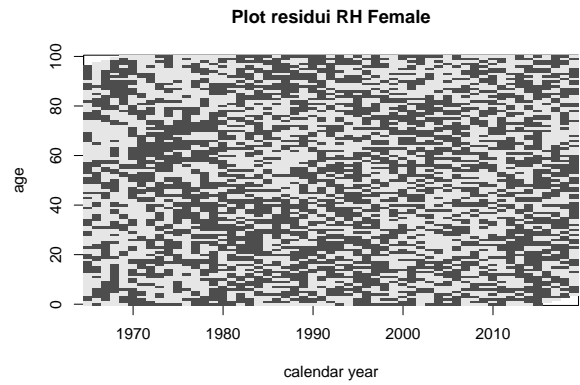
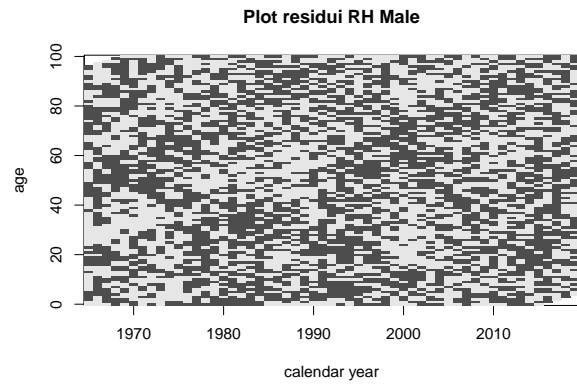
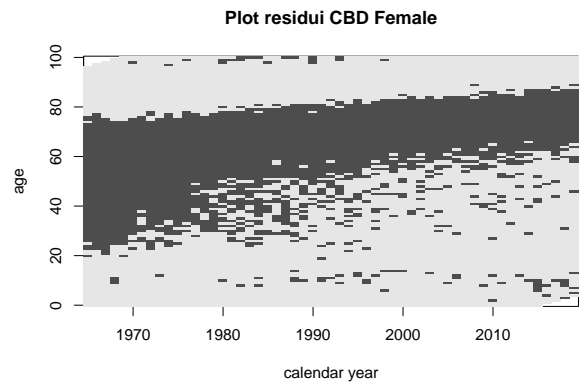
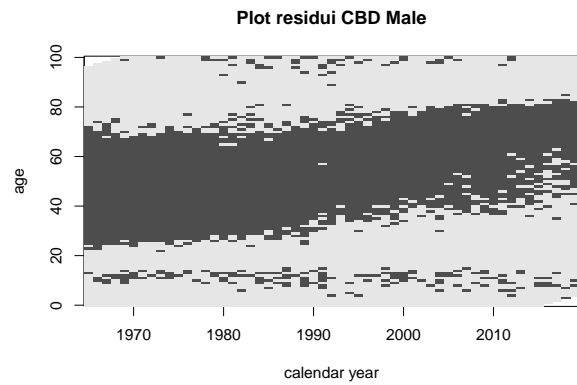
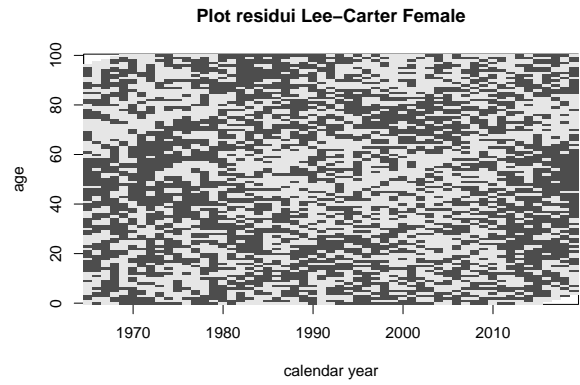
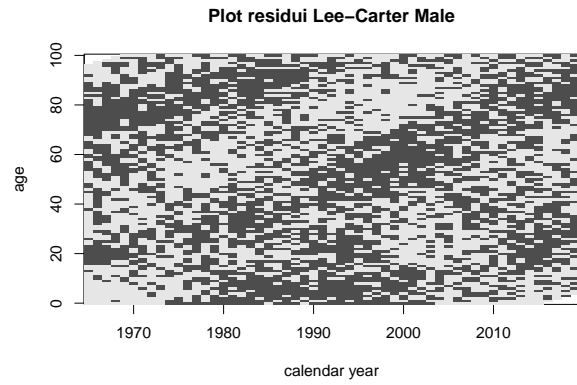
La bontà d'adattamento dei modelli di mortalità viene tipicamente analizzata ispezionando i residui del modello stimato, pattern regolari nei residui indicano l'incapacità del modello di descrivere adeguatamente tutte le caratteristiche dei dati. Nel caso di componente casuale di tipo Poisson o Binomiale, è opportuno esaminare le deviazioni residuali scalate.

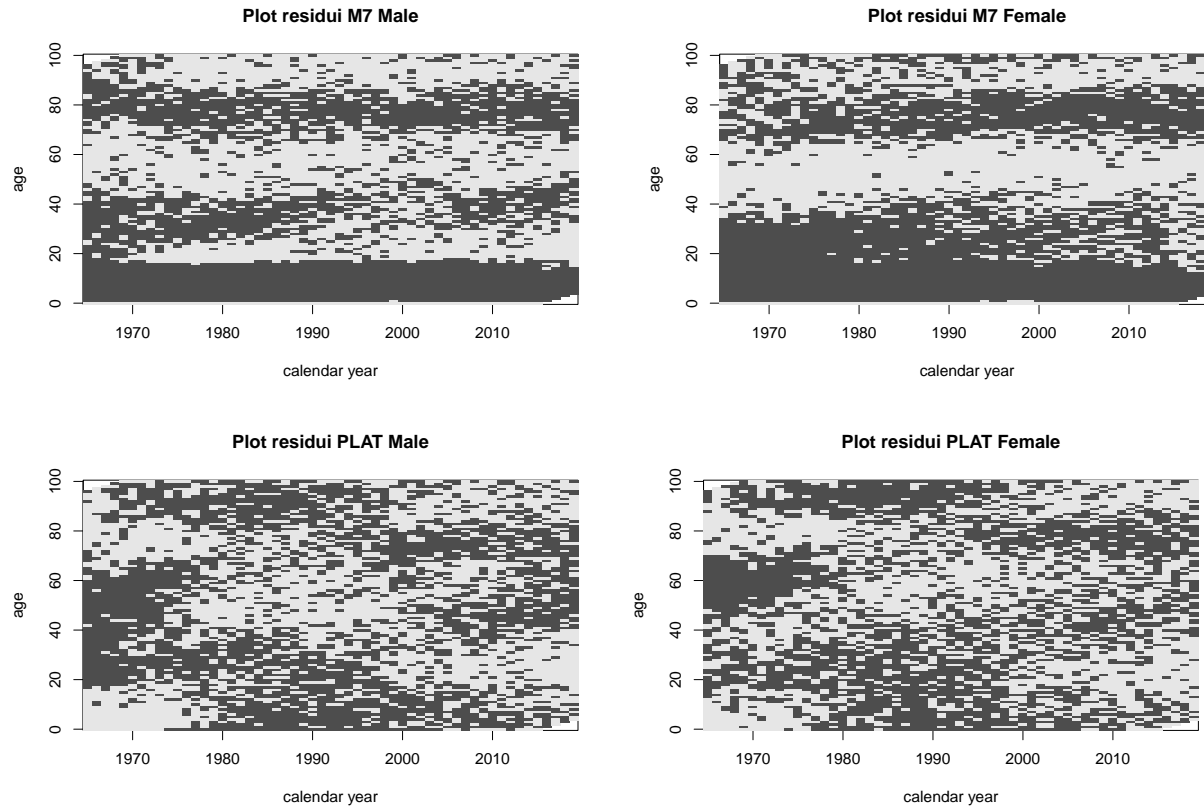
```

##### Residui
LCres_m <- residuals(LCfit_m)
LCres_f <- residuals(LCfit_f)
CBDres_m <- residuals(CBDfit_m)
CBDres_f <- residuals(CBDfit_f)
RHres_m <- residuals(RHfit_m)
RHres_f <- residuals(RHfit_f)
APCres_m <- residuals(APCfit_m)
APCres_f <- residuals(APCfit_f)
M7res_m <- residuals(M7fit_m)
M7res_f <- residuals(M7fit_f)
PLATres_m <- residuals(PLATfit_m)
PLATres_f <- residuals(PLATfit_f)

### rappresentazioni alternative dei residui
TYPE="signplot"
plot(LCres_m, type=TYPE, main = "Plot residui Lee-Carter Male")
plot(LCres_f, type=TYPE, main = "Plot residui Lee-Carter Female")
plot(CBDres_m, type=TYPE, main = "Plot residui CBD Male")
plot(CBDres_f, type=TYPE, main = "Plot residui CBD Female")
plot(RHres_m, type=TYPE, main = "Plot residui RH Male")
plot(RHres_f, type=TYPE, main = "Plot residui RH Female")
plot(APCres_m, type=TYPE, main = "Plot residui APC Male")
plot(APCres_f, type=TYPE, main = "Plot residui APC Female")
plot(M7res_m, type=TYPE, main = "Plot residui M7 Male")
plot(M7res_f, type=TYPE, main = "Plot residui M7 Female")
plot(PLATres_m, type=TYPE, main = "Plot residui PLAT Male")
plot(PLATres_f, type=TYPE, main = "Plot residui PLAT Female")

```





Dalle figure osserviamo i modelli CBD, APC, M7, PLAT presentare dei residui con pattern evidenti, mentre i modelli LC e RH appaiono ragionevolmente casuali, il fenomeno è condiviso per entrambe le popolazioni. In particolare osserviamo nei residui dei modelli CBD, APC ed M7 forti pattern orizzontali, dovuti probabilmente all'incapacità di tenere conto dei miglioramenti dei tassi di mortalità con il passare degli anni per le età.

Per completare l'analisi dei modelli vengono quindi prese in considerazione delle metriche utili per valutare e confrontare i risultati ottenuti dai modelli statistici in fase di stima, in particolare viene considerato il criterio d'informazione BIC e il valore della verosomiglianza.

```
library(knitr)
ris_model_m <- kable(model_m, caption = "Confronto Modelli Popolazione Maschile")
ris_model_f <- kable(model_f, caption = "Confronto Modelli Popolazione Femminile")

ris_model_m
```

Table 1: Confronto Modelli Popolazione Maschile

model	bic	loglik
LC	42754(2)	-20278(2)
RH	42523(1)	-19533(1)
APC	44038(3)	-20726(4)
CBD	170175(6)	-84613(6)
M7	79875(5)	-38606(5)
PLAT	44068(4)	-20513(3)

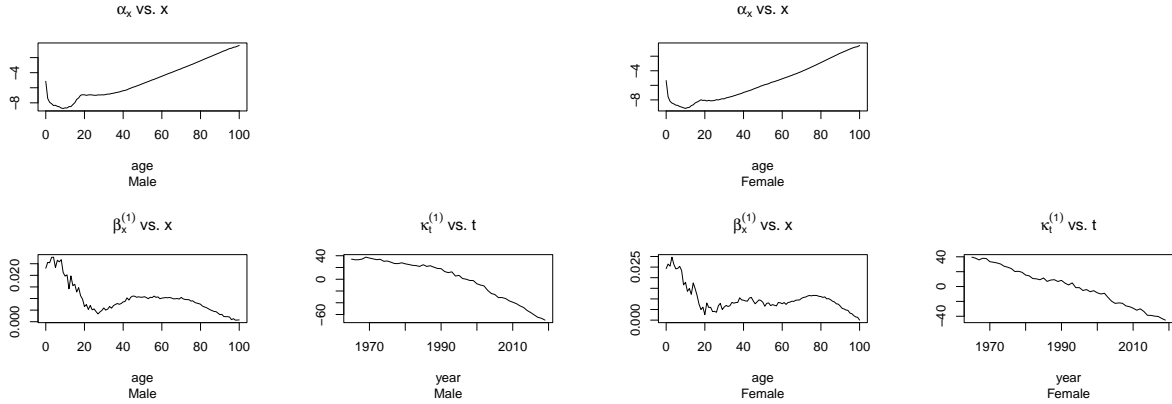
Table 2: Confronto Modelli Popolazione Femminile

model	bic	loglik
LC	40122(1)	-18962(2)
RH	40741(2)	-18643(1)
APC	42416(6)	-19915(4)
CBD	168115(3)	-83583(6)
M7	69997(5)	-33667(5)
PLAT	42137(4)	-19547(3)

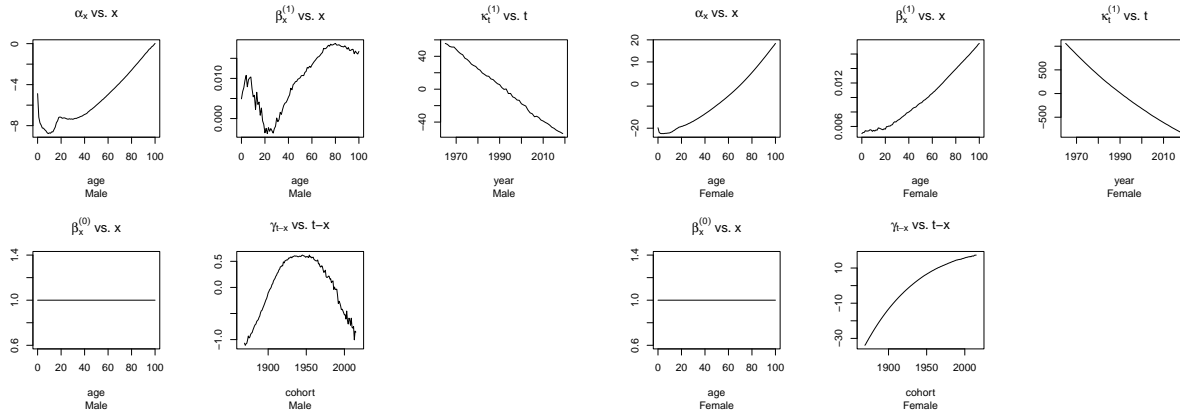
Dalle tabelle emerge che i modelli Lee-Carter e Renshaw&Haberman presentano valori di verosimiglianza e BIC migliori rispetto agli altri modelli considerati per entrambe le popolazioni (maschile e femminile).

Complessivamente, considerando i risultati ottenuti dall'analisi dei residui, dal BIC e il valore della verosimiglianza, la selezione per sviluppare proiezioni dei tassi di mortalità per la popolazione norvegese è ricaduta sui modelli Lee-Carter e Renshaw&Haberman.

Di seguito vengono mostrati i termini stimati del modello Lee-Carter:



Di seguito vengono mostrati i termini stimati del modello RH:





## Forecast

Nella famiglia dei modelli di mortalità stocastici, le dinamiche della mortalità vengono definite dagli indici di periodo  $k_t^{(i)}$  e dall'indice di coorte  $\gamma_{t-x}$ . Pertanto, le proiezioni e le simulazioni dei tassi di mortalità richiedono la modellizzazione di questi indici utilizzando tecniche di serie storiche.

Nella casistica delle proiezioni attraverso il modello Lee-Carter, viene adottato un approccio nel quale si assume un processo multivariato random walk per l'indice di periodo.

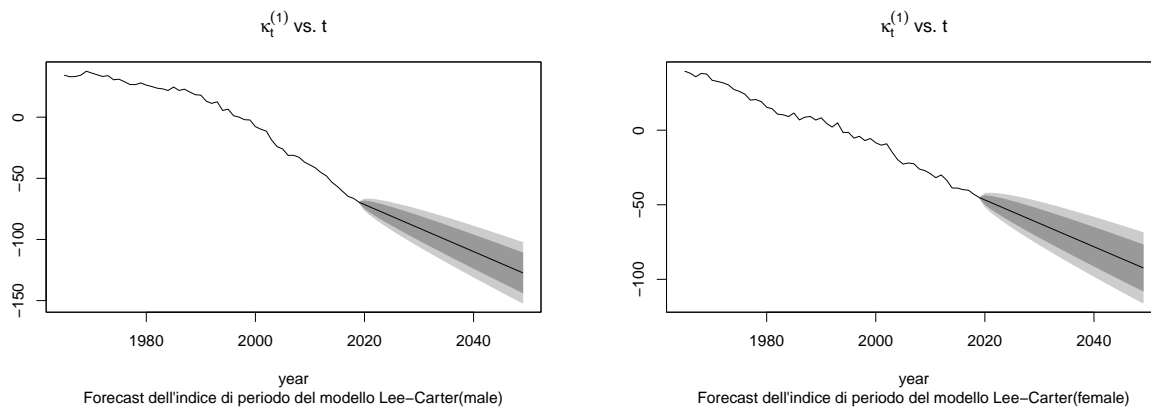
Nel caso di Renshaw & Haberman, per proiettare i tassi di mortalità della popolazione maschile viene adottato un approccio basato su un processo multivariato random walk per l'indice di periodo, mentre l'effetto di coorte viene considerato generato da un processo  $ARIMA(1, 1, 0)$ .

D'altra parte, per proiettare i tassi di mortalità della popolazione femminile vengono considerati processi  $ARIMA(2, 0, 1)$  indipendenti univariati per l'indice di periodo e un processo  $ARIMA(1, 1, 0)$  per l'effetto di coorte

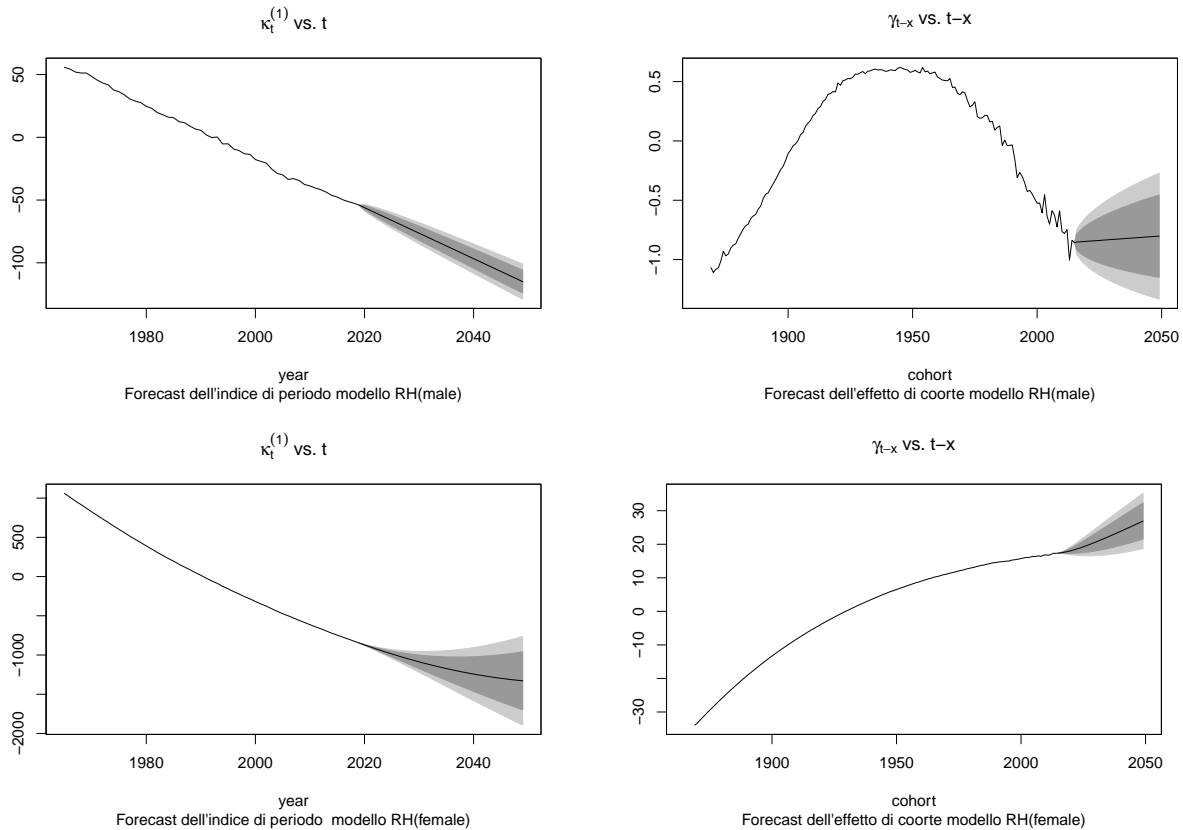
```
# Proiezioni -----
# -LC
LCfor_m <- forecast(LCfit_m, h=y.pred)
LCfor_f <- forecast(LCfit_f, h=y.pred)

# -RH
RHfor_m <- forecast(RHfit_m, h=y.pred, gc.order = c(1, 1, 0))
RHfor_f <- forecast(RHfit_f, h=y.pred, kt.method = "iarima", kt.order = c(2, 0, 1), gc.order = c(1, 1, 0))
```

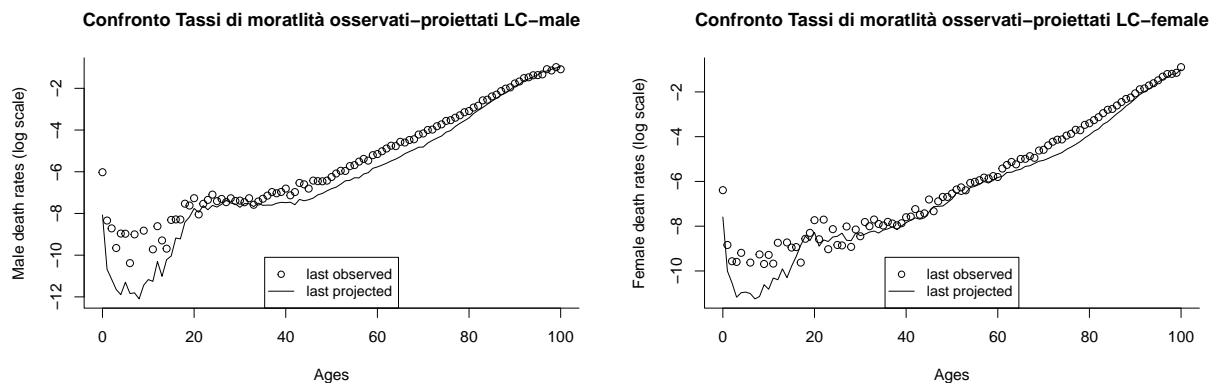
Di seguito vengono mostrate le proiezioni dell'indice di periodo del modello Lee-Carter:



Di seguito vengono mostrate le proiezioni dell'indice di periodo e dell'effetto di coorte del modello RH:

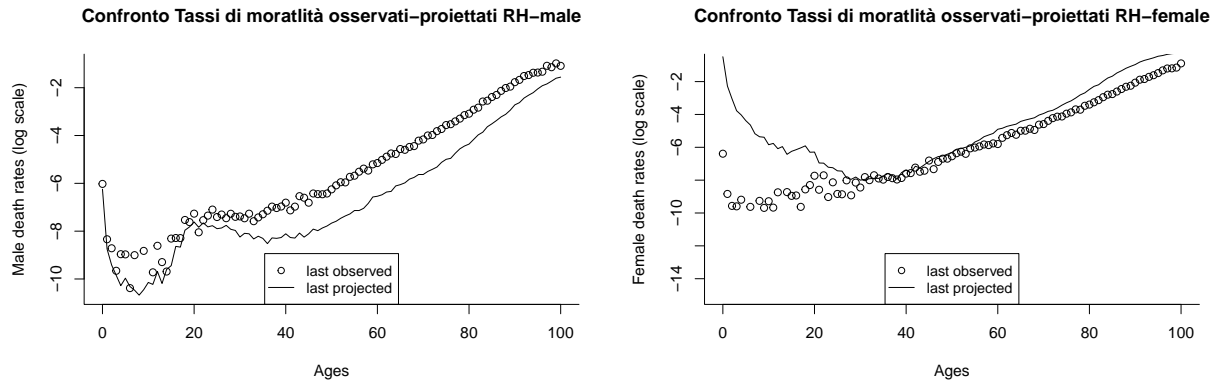


Successivamente, vengono confrontati i tassi di mortalità (espressi in scala logaritmica) osservati nell'ultimo anno disponibile (2019) con i tassi di mortalità proiettati nell'ultimo anno dell'orizzonte temporale considerato per il forecast, ovvero nel 2049.



Le proiezioni sviluppate a partire dal modello Lee-Carter sono generalmente in grado di catturare adeguatamente la dinamica dei tassi di mortalità nelle sottopopolazioni prese in considerazione. Si registrano, in entrambe le popolazioni, proiezioni dei tassi di mortalità generalmente ottimiste nei primi anni di vita fino all'adolescenza.

Nelle previsioni Lee-Carter per la popolazione maschile, tra il quarantesimo e l'ottantesimo anno di vita, le proiezioni risultano sistematicamente inferiori rispetto ai tassi di mortalità osservati nel 2019.



Le proiezioni dei tassi di mortalità della popolazione maschile, ottenute tramite il modello RH, mostrano un andamento sistematicamente inferiore rispetto agli ultimi tassi osservati, e questa differenza si amplifica ulteriormente dopo il 20° anno di vita. In altre parole, il modello RH considera un'evoluzione dei tassi di mortalità al ribasso nella popolazione maschile norvegese, soprattutto in età adulta.

Per quanto riguarda la popolazione femminile, le proiezioni dei tassi di mortalità del modello RH risultano irrealistiche. Il modello non è stato in grado di cogliere adeguatamente la dinamica della mortalità nei primi 30 anni di vita della popolazione femminile.

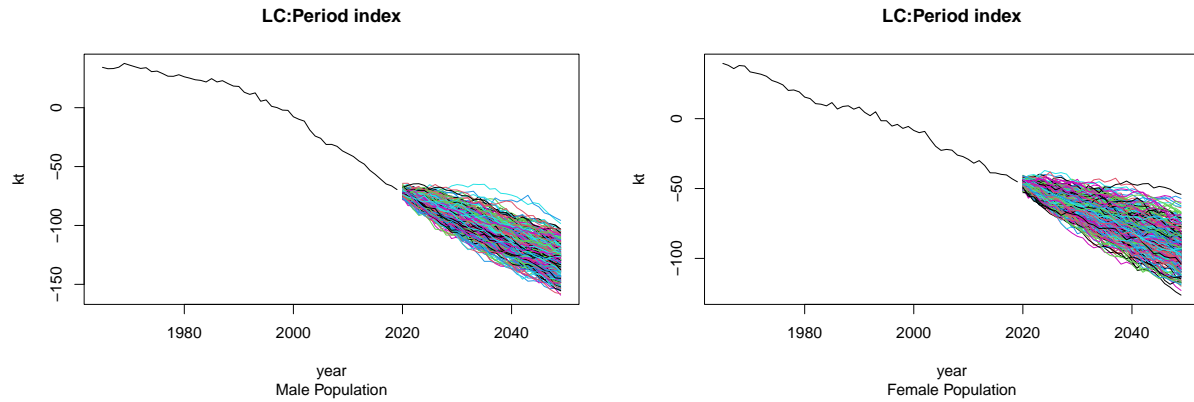
## Forecast uncertainty

Al fine di considerare l'incertezza derivante dagli errori di previsione, sono state simulate 300 traiettorie per i successivi 30 anni a partire dal 2019.

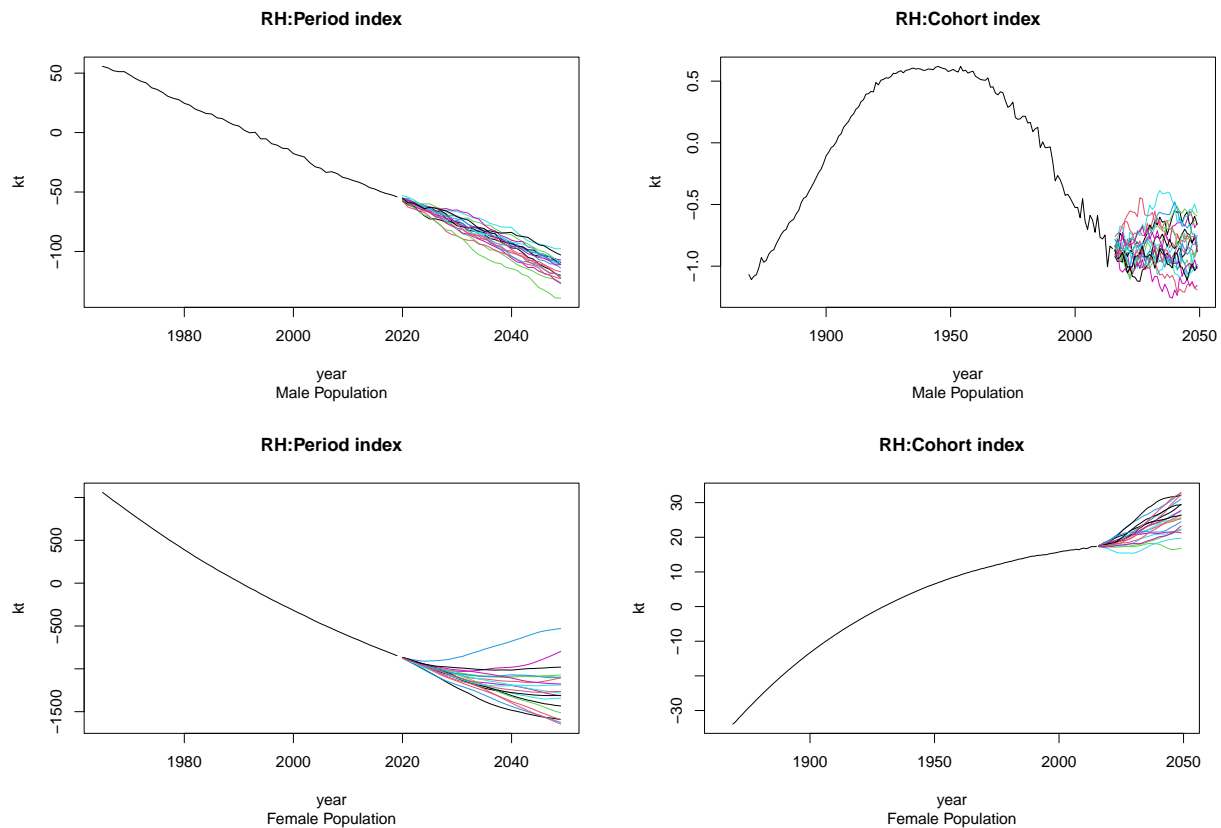
```
# Simulazioni -----
n.sim<-300
# Simulation LC
# -male
LCsim_m.mrwd <- simulate(LCfit_m, nsim = n.sim, h=y.pred)
# -female
LCsim_f.mrwd <- simulate(LCfit_f, nsim = n.sim, h=y.pred)

# Simulation RH
# -male
RHsim_m<- simulate(RHfit_m, nsim = n.sim, h=y.pred, gc.order = c(1, 1, 0))
# -female
RHsim_f <- simulate(RHfit_f, nsim = n.sim, h=y.pred, kt.method = "iarima", kt.order = c(2, 0, 1),
                    gc.order = c(1, 1, 0))
```

Di seguito vengono mostrate le simulazioni dell'indice di periodo del modello Lee-Carter:

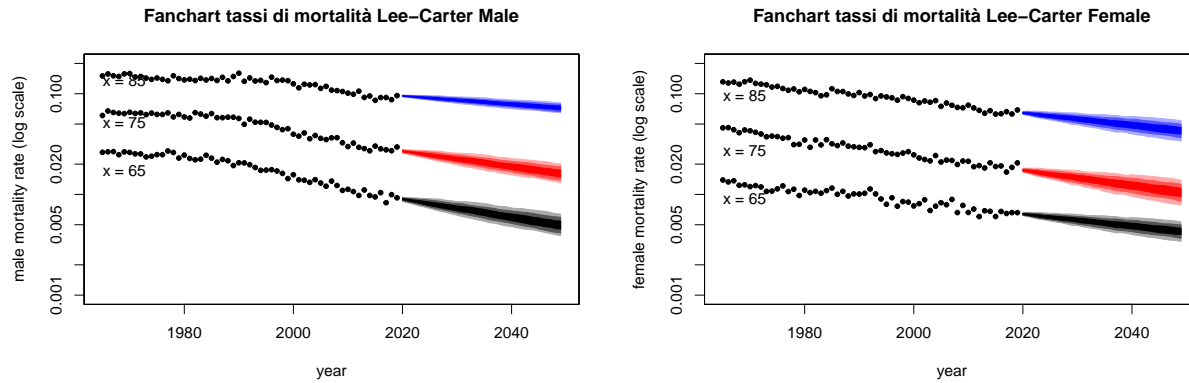


Di seguito vengono mostrate le simulazioni dell'indice di periodo e dell'effetto di coorte del modello RH:



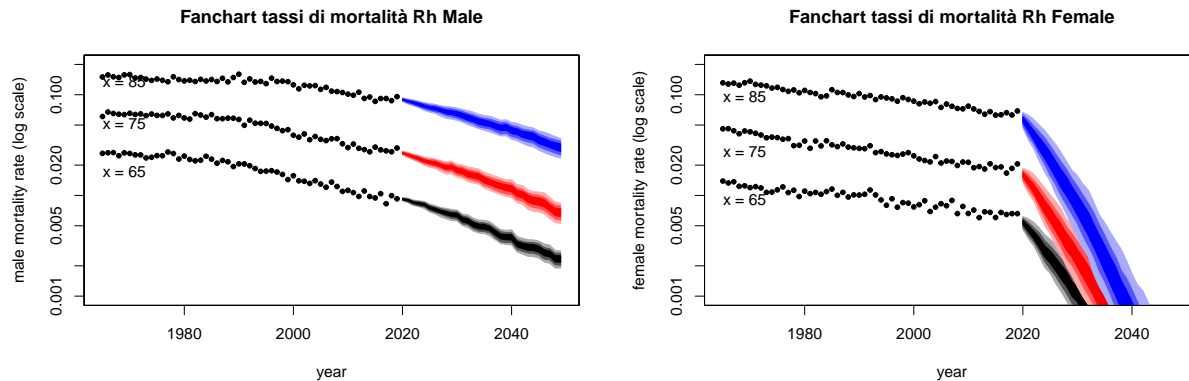
A partire dai tassi di mortalità simulati in precedenza, vengono calcolati gli intervalli di previsione per dare una misura dell'incertezza dovuta che caratterizza le proiezioni. Gli intervalli di proiezione vengono mostrati di seguito attraverso i fan chart. Nelle figure, i puntini rappresentano i tassi di mortalità osservati in Norvegia rispettivamente all'età di 65, 75 e 85 anni. L'area blu rappresenta gli intervalli di previsione dei tassi di mortalità proiettati all'età di 85 anni nel periodo di proiezione, e la gradazione del colore indica i diversi percentili: l'area più scura tra il 25% e il 75%, quella media tra il 10% e il 90%, quella più chiara tra il 2,5% e il 97,5%. L'area rossa si riferisce agli intervalli di previsione all'età di 75 anni e quella nera all'età di 65 anni.

Fan Chart Proiezioni Lee-Carter:



Si può notare che le proiezioni ottenute presentano un ragionevole livello di incertezza sia per la popolazione maschile che per quella femminile. Inoltre, l'andamento della mortalità sembra essere biologicamente plausibile.

Fan Chart Proiezioni Renshaw & Haberman:



Nella popolazione maschile, le proiezioni sviluppate tramite il modello Rh presentano un ragionevole livello di incertezza e l'andamento sembra biologicamente plausibile, con un declino dei tassi di mortalità maggiore rispetto ai risultati ottenuti dalle proiezioni Lee-Carter. Tuttavia, nel caso della popolazione femminile, si osserva un'evoluzione della mortalità irrealistica e biologicamente non plausibile.

## Parameter uncertainty

Quando si analizza l'incertezza nelle proiezioni di mortalità in un contesto attuariale, è importante considerare tutte le fonti di rischio. Ad esempio gli intervalli di previsione (fanchart) ottenuti nella sezione precedente tengono conto solo dell'incertezza derivante dall'errore nella previsione degli indici di periodo e di coorte e ignorano l'incertezza derivante dalla stima dei parametri del modello.

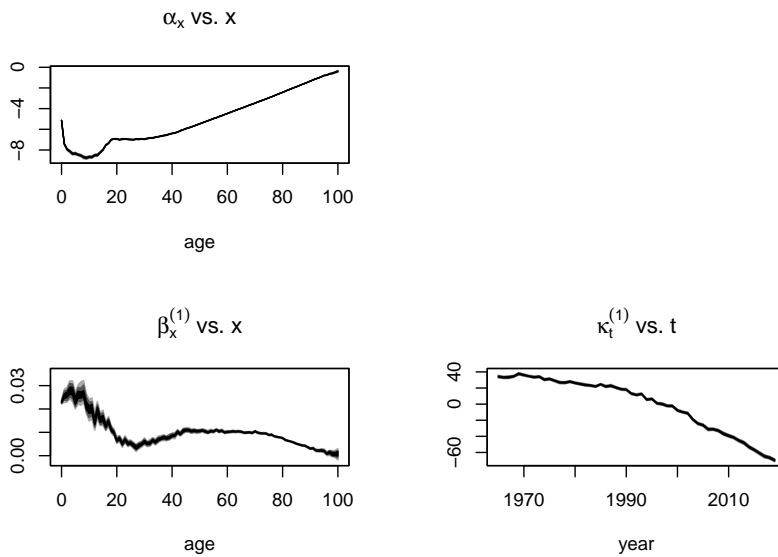
A causa dell'intrattabilità analitica di molti modelli stocastici di mortalità, l'incertezza dei parametri di solito viene presa in considerazione utilizzando procedure bootstrap. In questa applicazione viene considerato il bootstrap semiparametrico, dove vengono generati inizialmente  $B$  campioni del numero di decessi  $d_{xt}^b$ ,  $b = 1, \dots, B$ , mediante campionamento dalla distribuzione dei dati considerata (in questo caso Binomiale) con media  $d_{xt}$ . Dunque ogni campione bootstrap  $d_{xt}^b$ ,  $b = 1, \dots, B$ , viene quindi utilizzato per ristimare il modello al fine di ottenere  $B$  stime bootstrappate dei parametri, in seguito questi vengono utilizzati per produrre intervalli di confidenza e di previsione.

Considerate le proiezioni irrealistiche ottenute dal modello Renshaw & Haberman per la popolazione femminile norvegese, non è stata condotta l'analisi dell'incertezza dei parametri specifici di questo modello per

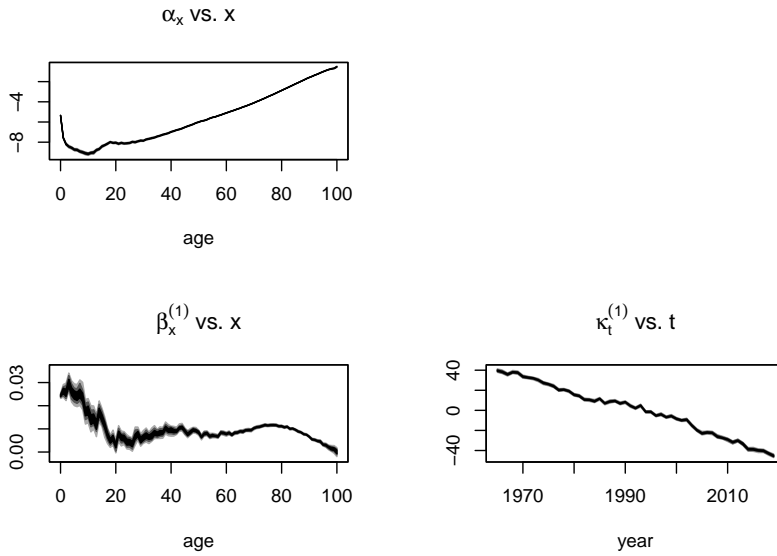
la popolazione in questione.

```
# Bootstrap -----
n.boot<-100
# LC
#   -male
LCboot_m <- bootstrap(LCfit_m, nBoot = n.boot, type = "semiparametric")
# LC
#   -female
LCboot_f <- bootstrap(LCfit_f, nBoot = n.boot, type = "semiparametric")
# RH
#   -male
RHboot_m <- bootstrap(RHfit_m, nBoot = n.boot, type = "semiparametric")
```

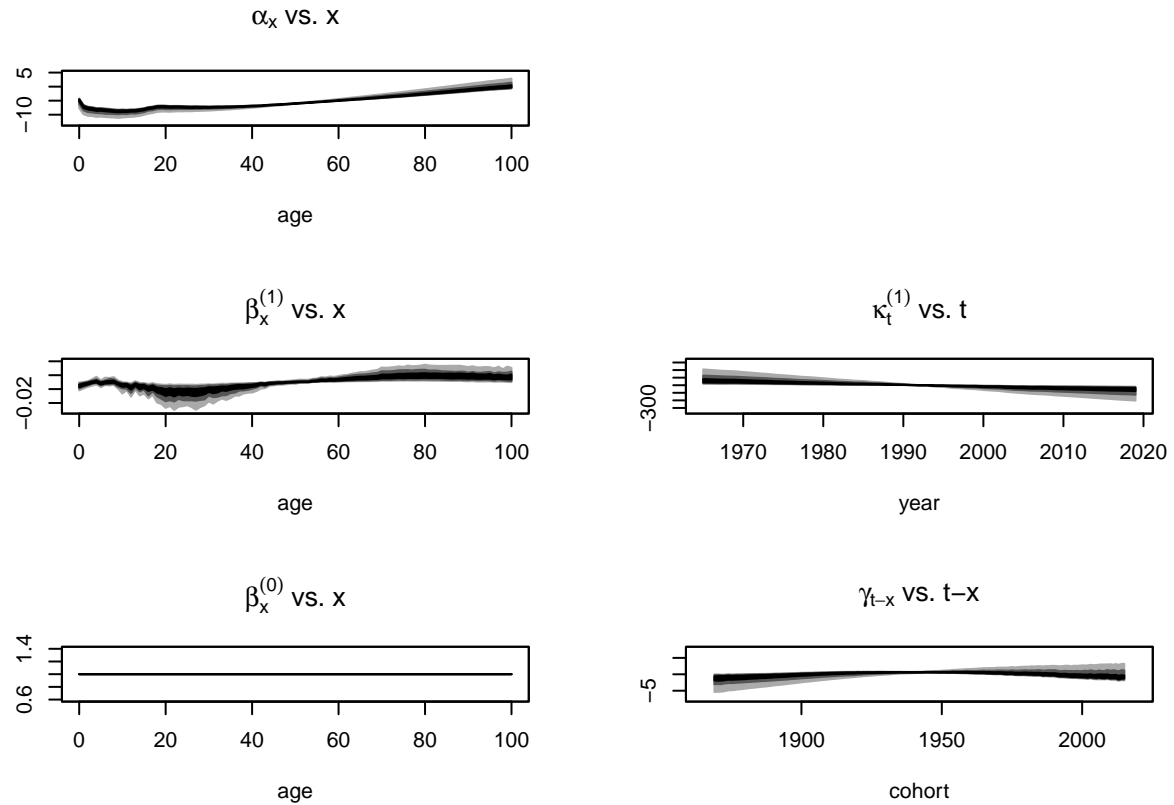
Termini “bootstrappati” del modello Lee Carter (popolazione maschile):



Termini “bootstrappati” del modello Lee Carter (popolazione femminile):



Termini “bootstrappati” del modello Renshaw&Haberman (popolazione maschile):

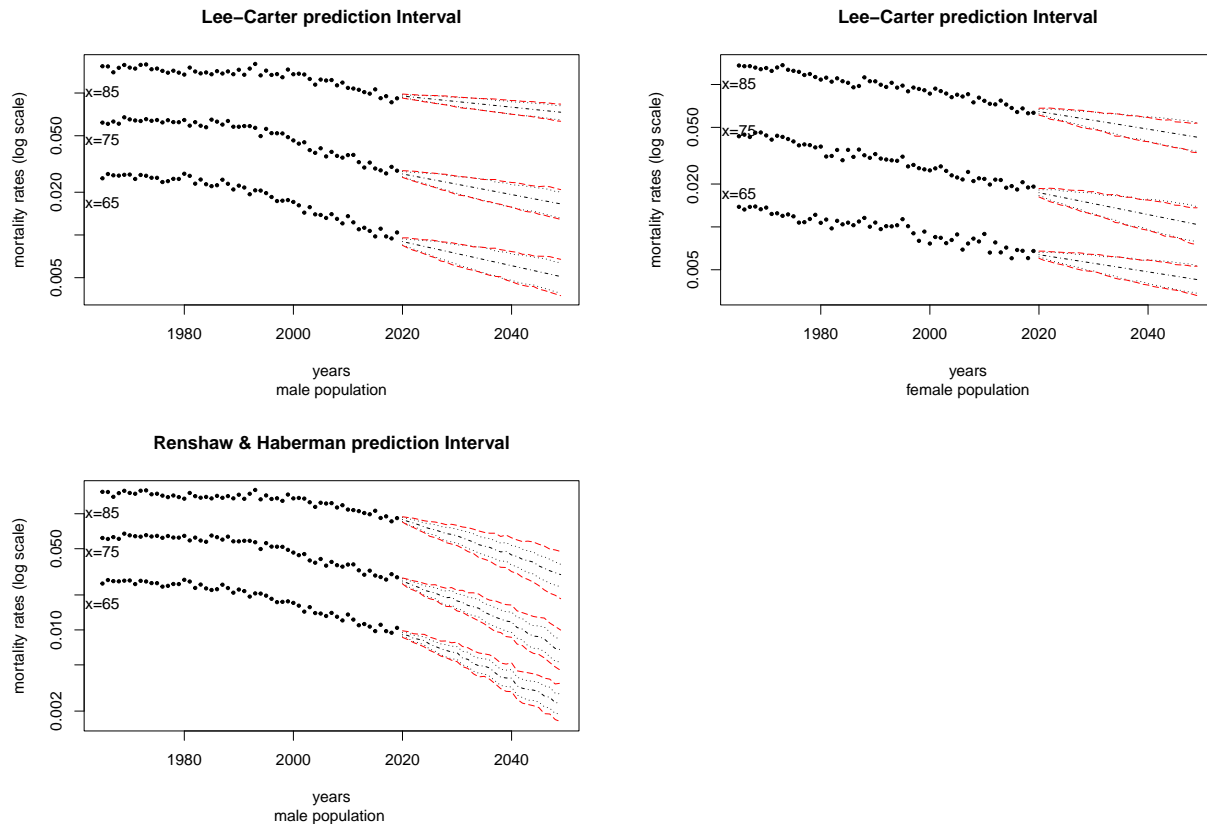


Dopo aver sottoposto il modello di mortalità stocastico alla procedura bootstrap, simuliamo le previsioni per ottenere traiettorie che tengono conto sia dell'errore nella previsione degli indici di periodo e coorte che dell'errore legato ai parametri del modello.

```

LCsim_m.boot <- simulate(LCboot_m, nsim = n.sim/n.boot, h = y.pred)
RHsim_m.boot <- simulate(RHboot_m, nsim = n.sim/n.boot, h = y.pred, gc.order = c(1, 1, 0))
LCsim_f.boot <- simulate(LCboot_f, nsim = n.sim/n.boot, h = y.pred)

```



Nelle figure precedenti:

- i punti rappresentano i tassi di mortalità storici per il periodo 1965-2019. Le linee tratteggiate rappresentano le previsioni centrali;
- le linee punteggiate nere rappresentano gli intervalli di previsione al 95% escludendo l'incertezza dei parametri;
- le linee tratteggiate rosse rappresentano gli intervalli di confidenza e previsione al 95% includendo l'incertezza dei parametri;

Nelle proiezioni effettuate utilizzando il modello Lee-Carter (per entrambe le popolazioni di riferimento), osserviamo che c'è una differenza minima tra gli intervalli di previsione al 95% che includono o meno l'incertezza di stima dei parametri.

Al contrario, nelle previsioni del modello Renshaw-Habermann, si evidenzia un effetto significativo associato all'incertezza dei parametri. Questo fenomeno è particolarmente evidente nelle previsioni dei tassi di mortalità per l'età  $x=85$ , dove l'ampiezza degli intervalli che includono l'incertezza di stima dei parametri è circa il doppio rispetto agli intervalli che non la includono.