Lee-Carter Mortality Forecasting and Smoothing Graphs: Application to ItalianPopulation

Contents

ABSTRACT:

This report will carry out an extensive analysis for mortality forecasting under mortality models like Lee-Carter, Cairns-Blake-Dowd, Hyndman-Ullah, Robust Hyndman-Ullah to the Italian Male population for the year 1950-2000 followed by mortality smoothing to check the feasibility of the estimates thereby obtained.

The report has been worked as an extension for the paper, "Lee-Carter mortality forecasting: application to the Italian population" by Steven Haberman and Maria Russolillo, (Nov 2005).

Their research investigates the feasibility of using the Lee-Carter methodology to construct the Italian population, which is forecasted under the ARIMA framework and is used to generate life tables. The research seemed to not provide a very clear comparison of the technique used with other widely recognized models and hence it gave me the scope to extend the analysis by using similar techniques on a part of the sample, that is, only on the Italian Male population.

INTRODUCTION:

1.1. Structure

The analysis carried out in the previous literature (Haberman and Russolillo, 2005) focuses explicitly on Lee carter model and it has been used to model the mortality of the Italian population between the years 1950-2000 followed by life table forecast.

I am going to extend this analysis by doing a similar analysis on Italian Male population with Lee carter and CBD model as well as smoothing the mortality data and applying the Lee Carter model along with its variants to forecast the life table thereby obtained. The time-varying index of mortality is forecasted in an ARIMA (0,1,0) framework, which is then used to obtain the projected life tables.

This paper has five sections. Section I gives a brief overview on what the previous research was about and how my analysis will give an extended perspective on it. Section II will focus on the methodology used for the analysis and the findings under Lee Carter model. Section III will focus on the CBD mortality model for Italian Male population. Section 4 will compare the models on Lee-Carter and CBD. Section V will expand the report on mortality smoothing. Section VI will conclude the report.

1.2. Previous Research Paper:

The research paper on which my report is based, deals with mortality forecasting of the Italian population under the Lee-Carter methodology and the forecasted Life Tables achieved thereby. The paper although has a detailed explanation of the Lee Carter model, it does not explicitly compare the findings with any other model in detail. There fore, there is a scope for research in that arena. Also, the results obtained from mortality forecasts have not been smoothened. This helps me to extend this analysis in two ways. First, I will compare the results with other models that is CBD. Based on the results of these models, I will smoothen the mortality rates for the model that fits the data best.

1.3. Methodology:

The model is fit to the Italian Male death rates from 1950 to 2000 from Human Mortality database maintained by University of California, Berkeley. The previous research was carried out to accommodate for the risks associated with a decrease in mortality rates especially the affect that it may have on the pricing for life annuities. Also, the risks associated with these mortality rates do not consider any particular trend into account. For instance, Longevity Risk in actuarial studies is a risk that is derived from future mortality rate, which does not accommodate for the forecasted rates. Thus, to cater to this risk a distinct projected table needs to be forecasted around trends pertaining to this form of risk.
The research paper that I have chosen uses the Lee-Carter approach extensively to forecast mortality by using the Italian mortality between the years 1950 to 2000.
The reasons for choosing this particular paper is that due to the relevance of Lee-Carter model which is by far the most distinguished developments in the field of mortality forecasting. It allows for uncertainty in forecasting mortality rates, which is also known

as the longevity risk. Finally, once we are able to calculate the precise time index value of *k*, the complete set of death probabilities can be calculated, and we can use the parameters for any particular year to calculate the life tables. The second most prominent model that has been an influential development in the field of mortality forecast is the CBD model or the Cairns-Blake-Dowd model. This model has been designed to provide for annuities and pensions with longevity risks and therefore, it fits the data better with higher aged population.

LEE-CARTER MORTALITY MODEL

2.1 Definition

$$In\ (m_{x,t})=\beta_x^{(1)} + \beta_x^{(2)}\, k_t^{(2)}$$

The Lee-Carter approach uses the sum of an age specific component ax which is independent of time and another factor which is the result of a product of a time-varying parameter kt to elucidate the log of a time series of age specific death rates mx,t. Further, Bx, helps us determine the rate at which each age varies with changes in general level of mortality. The last part explains the error term.

For the purpose of modeling, the data obtained from Human Mortality Database is segregated by gender and thus the death rate of a five-year-age sub groups from the year 1950 to 2000 and the same process is repeated for the exposure to risk.

2.2 Modeling

To model the data set of Italian Male population, I have used the Stochastic Mortality Model in R also commonly referred as Stmomo. The package has in built functions to help model the Italian Male mortality rates.

Figure 1: The following graph show the results of the historical male death rates in Italy.

2.3 Parameter Estimation:

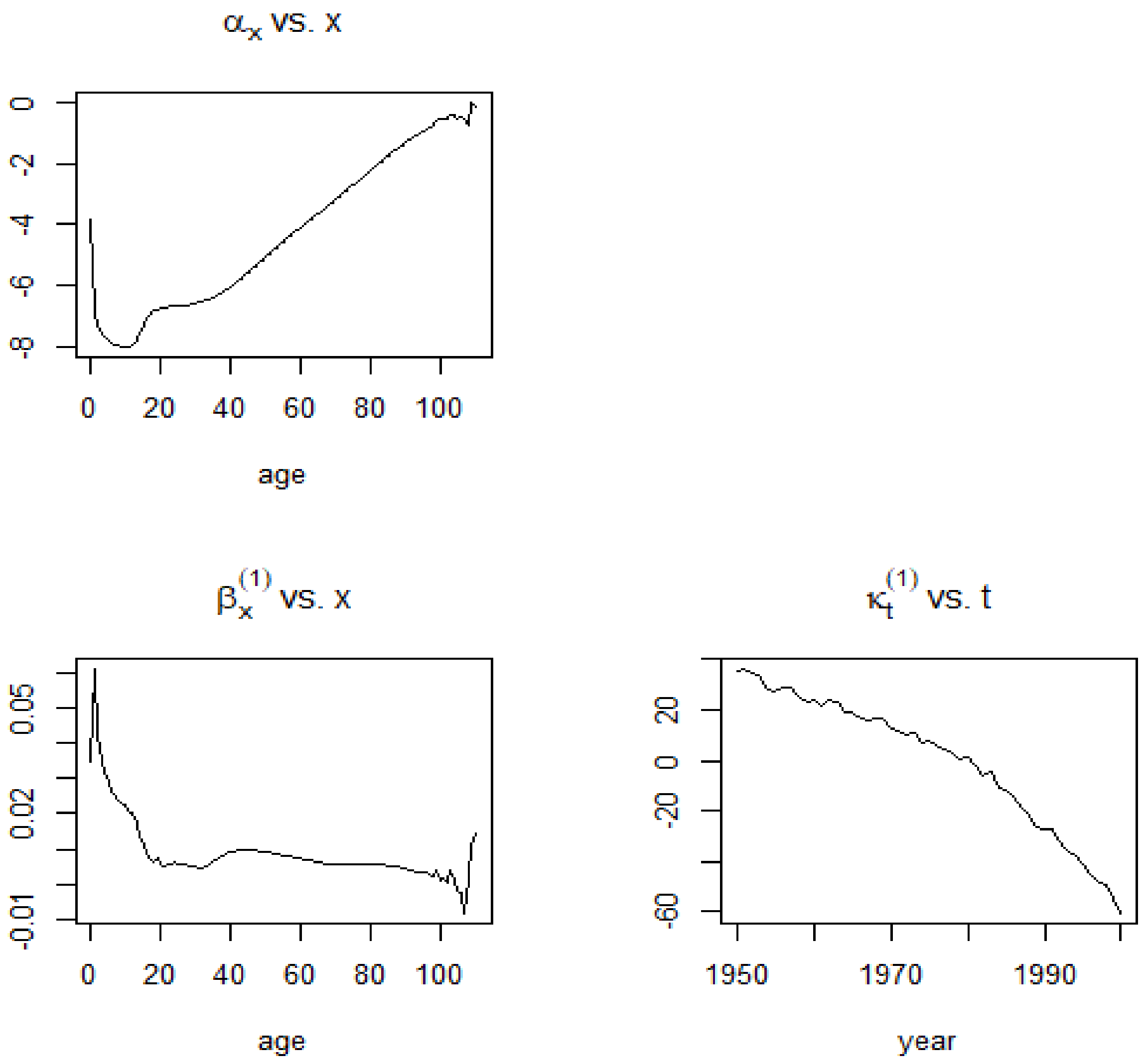


Figure 2: The parameters obtained after simulating the data are as follows

The figures above indicate that given that Bx is large for x at some parts, then the age x will vary by a large extent with the change in level of mortality and when Bx is small then the change in the level of mortality would lead to small variation in the death rate at that particular age.

Also, for the purpose of estimation of kt, it is quite indicative that the weightage of equalvalue to each data will result in bias estimation of kt because there is equal weightage given while fitting the log-transformed rates to the death rates of the youth which are very few in number as compared to the higher death rates of the older population.

The figure indicates that kt is declining almost linearly from 1950-2000 at a relatively constant rate. Thus, it can be said that the results obtained are in sync with the findings of Lee-Carter, which required k to decline in a linear way and have a relatively constant variance. Also, this infers that mortality in the male population is under a constant decline, which may be a very good thing for the Italian population.

CAIRNS-BLAKE-DOWD MODEL

3.1. Definition

The model expressed as:

$$In\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = k_t^{(1)} + k_t^{(2)}(x - \bar{x})$$

Where for a given year t, the value of qx,t is linearly related to age x after the logit transformation. $k_t^{(1)}$ is the mortality curve. $k_t^{(2)}$ represents the slope of the transformed mortality curve

The Cairns-Blake-Dowd model or the CBD model has been marked as an innovative tool to forecasting the mortality rates and also considered an extension of the lee carter method. The model is usually best described at higher age periods and has two correlated factors that is the level and the slope which allows for various improvements at different stages at various time periods. The model is very straightforward in incorporating any uncertainty associated with parameter estimation. CBD has been built to inculcate the deficiencies associated with Lee-Carter.

3.2. Modeling and parameter estimation:

The modeling of the Italian Male population was carried out in the StMoMo package that had an inbuilt function for carrying out the analysis under CBD.

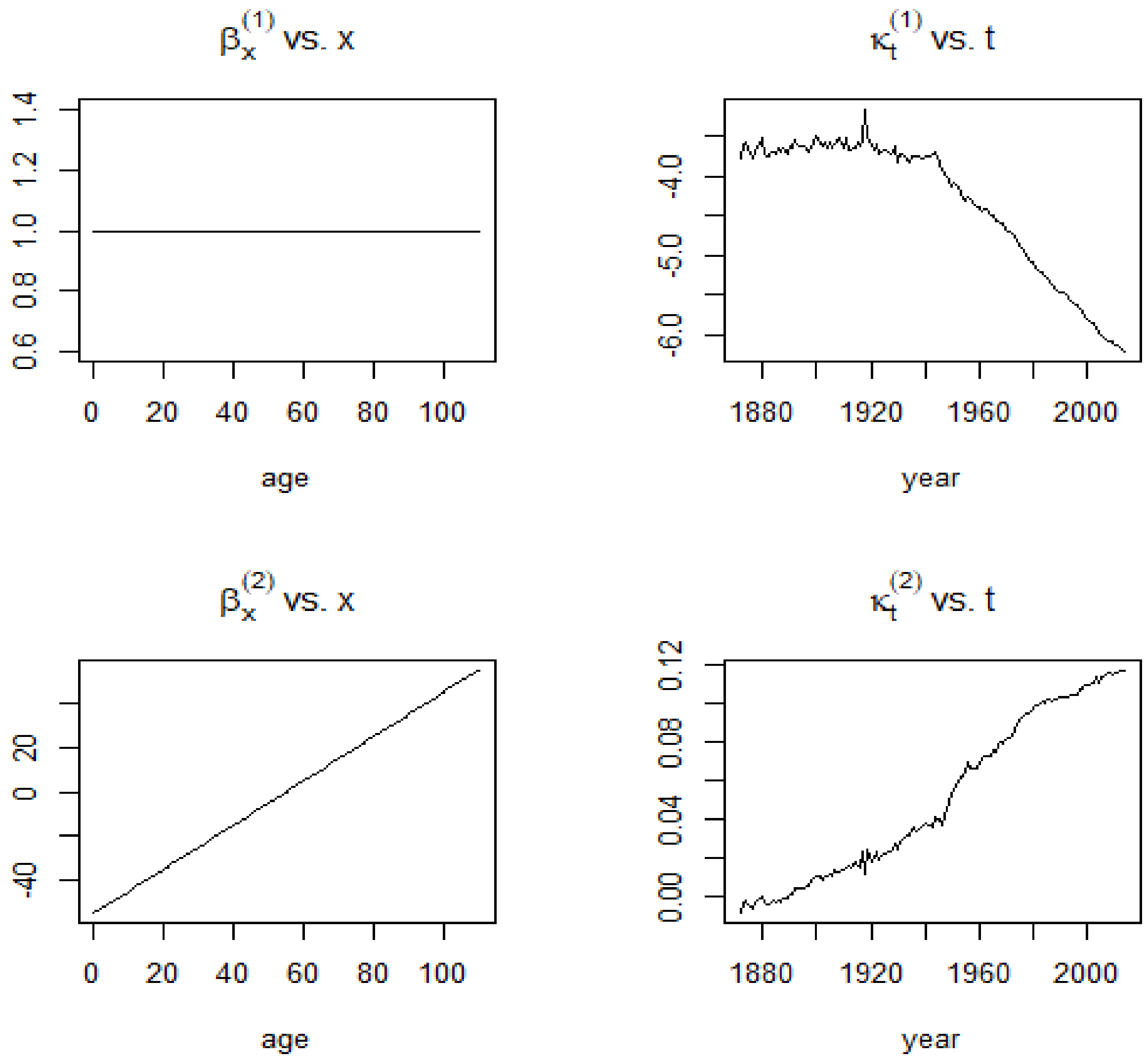
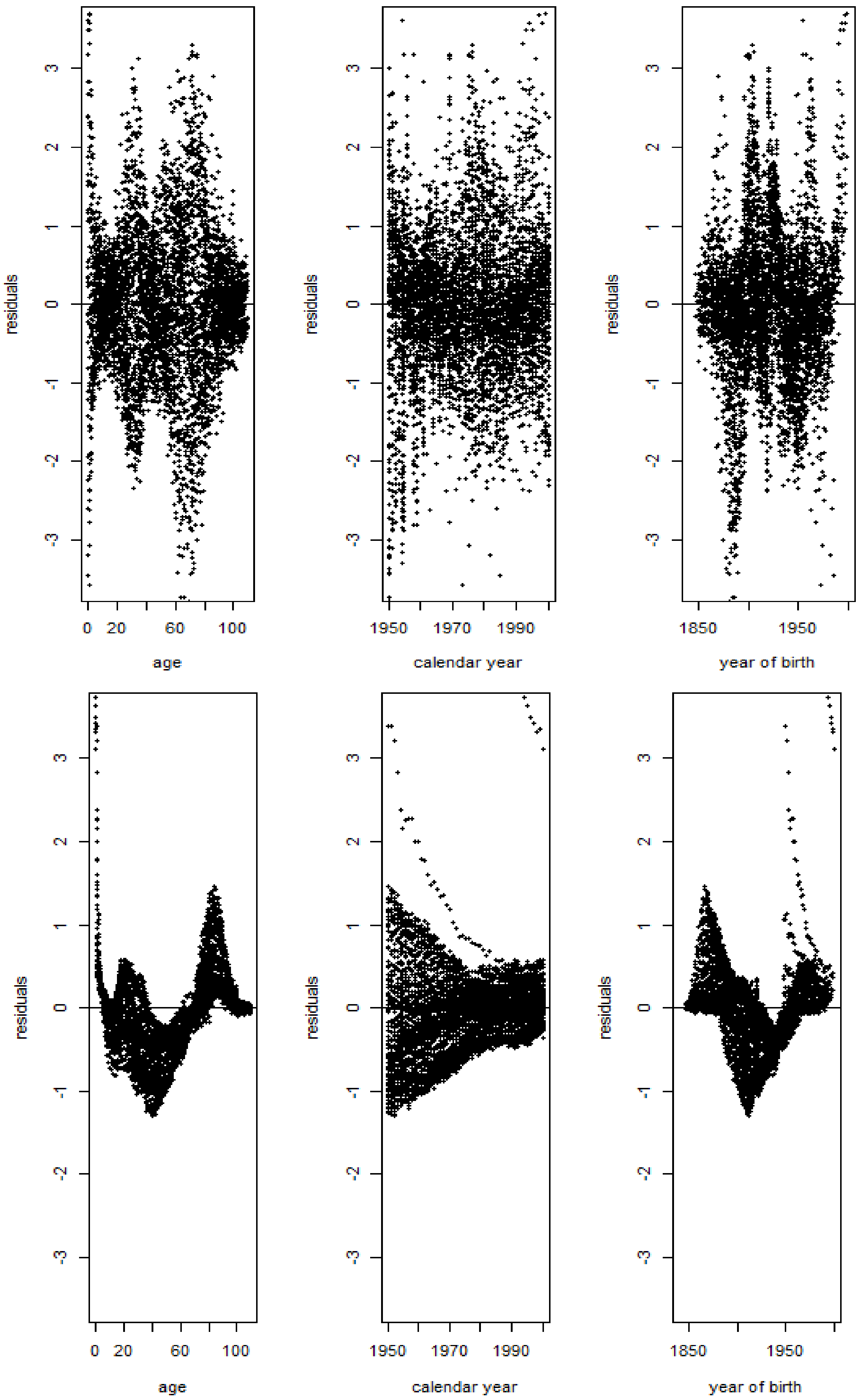


Figure 3: The parameters thus obtained are as follows

The figures above indicate a constantly declining kt(1) and constantly rising k(t2) when Bx(1) and Bx(2) are constant and linear respectively. This can be inferred as when the age specific parameters are constant then the time varying parameters are decreasing at a constant variability showing an overall mortality improvement but when the age specific parameters are increasing linearly then the time varying parameters are increasing at a steady rate show that the mortality at younger ages improves more rapidly than that at older ages. The CBD model is thus robust only with higher aged population and will give more robust estimates. This will also imply that the tractability issue related to providing a mortality index can only be catered by a model, which is robust with respect to the time period and has all round qualities.

The residual plots for both the models can be seen as below:

Figure 4&5 : Lee-Carter residuals



The residual plot indicates that the model which has data around zero scattered randomly will have a much better fit to the data than the one which is loosely scattered.

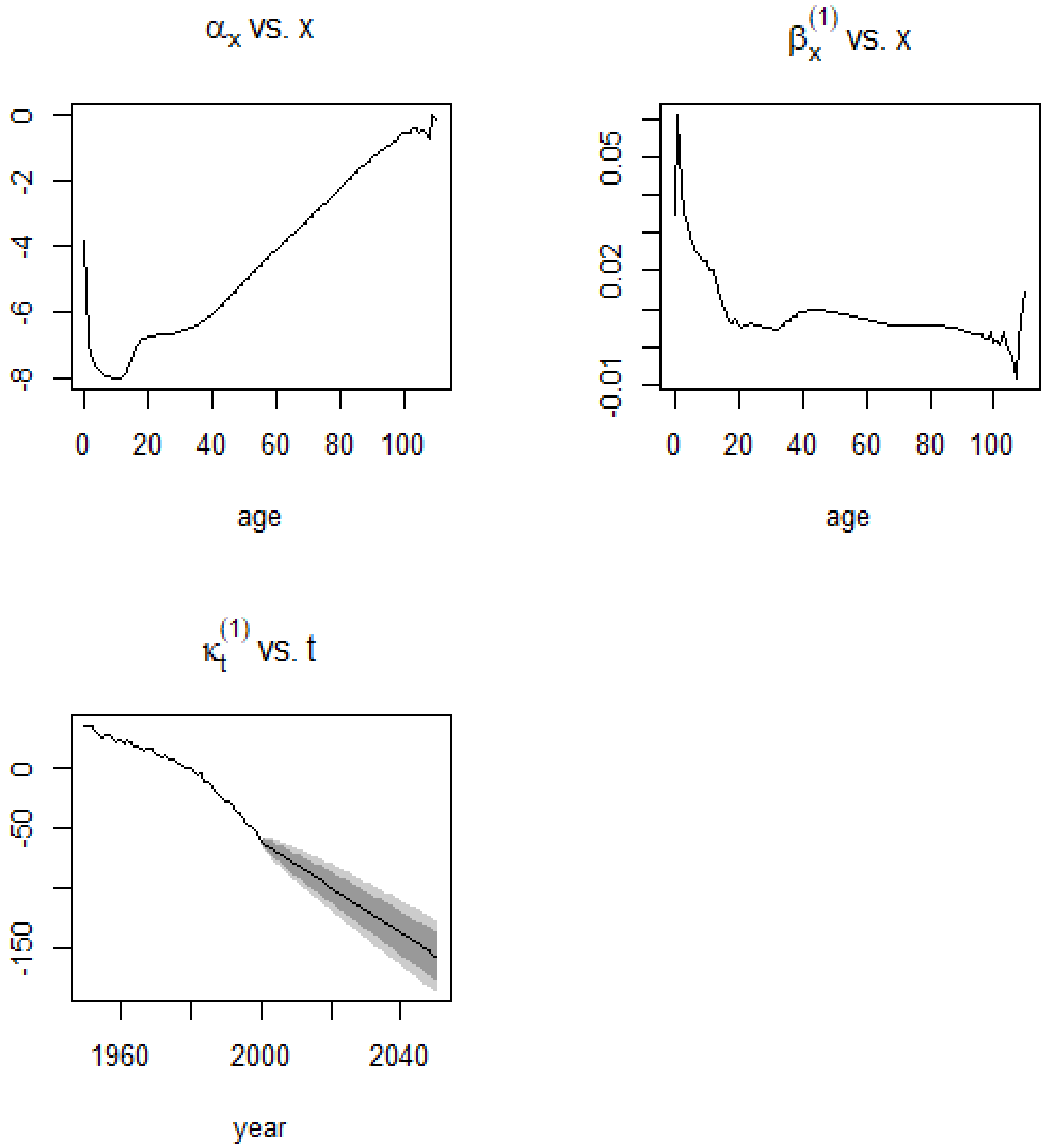
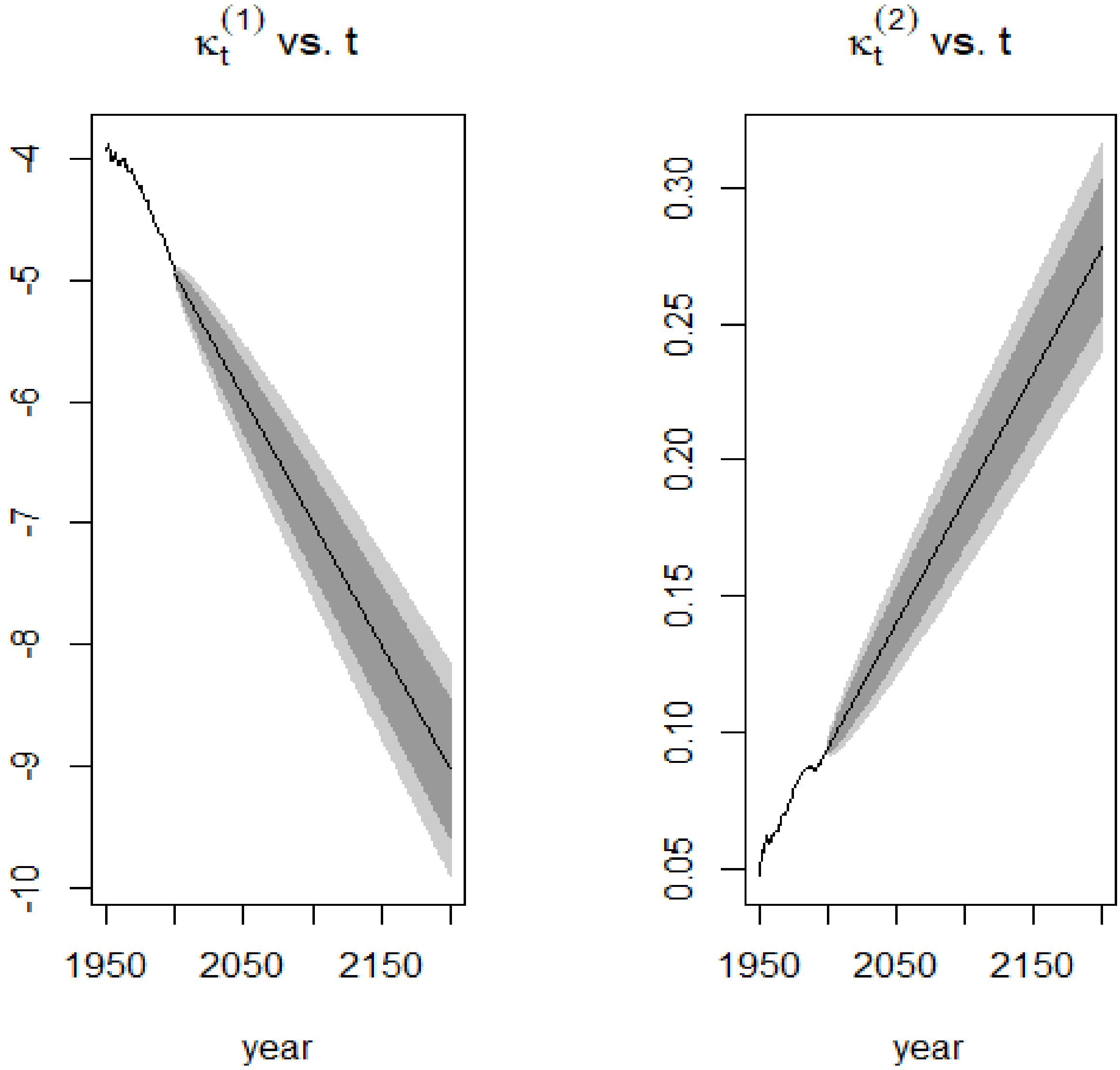


Figure 6: The forecasts of the parameters for the next 100 years can be seen as follows-Lee-Carter

CBD:



Finally, the forecasts of the parameters obtained from the Lee-Carter model and the CBD model seems to project the same trend for the parameter kt, kt1 and kt2 respectively. This estimation is quite accurate because the models are tentatively saying that in the next 50 years it is quite likely for the mortality to improve further and that the improvement would be more rapid in younger ages than older.

COMPARISON

4.1. AIC and BIC of LC and CBD

To fully interpret which model is a better fit for the Italian male data we need to calculate the Akaike information criterion or (AIC) and the Bayesian Information criterion (BIC).

	LEE CARTER	CAIRNS BLAKE DOWD
AIC	184309.8	5987654
BIC	190576.4	5988865

Thus, the above represents the AIC and BIC criterion of the model. AIC model estimates the quality of each model and hence the model with the lowest AIC is preferred. In the same way, the model with the lowest BIC is preferred.

Thus it can be seen that Lee Carter gives a better good ness of fit for the data of Italian male population. It also implies that it is a good fit over wide age ranges where as CBD is a good robust model for higher ages or data that involves a much bigger picture.

MORTALITY SMOOTHING

The analysis carried out above clearly outlines that Lee-Carter is a better approach to fit the Italian Male Population. Thus, the subsequent extension for the literature chosen by me would focus to single out the most suitable model amongst LC model and an off shoot of this model, that is the Functional Demographic Model by Hyndman-Ullah. This would give us the accurate fit by using the traditional P-splines approach and compare it with Weighted Constrained Penalized Regression Splines.

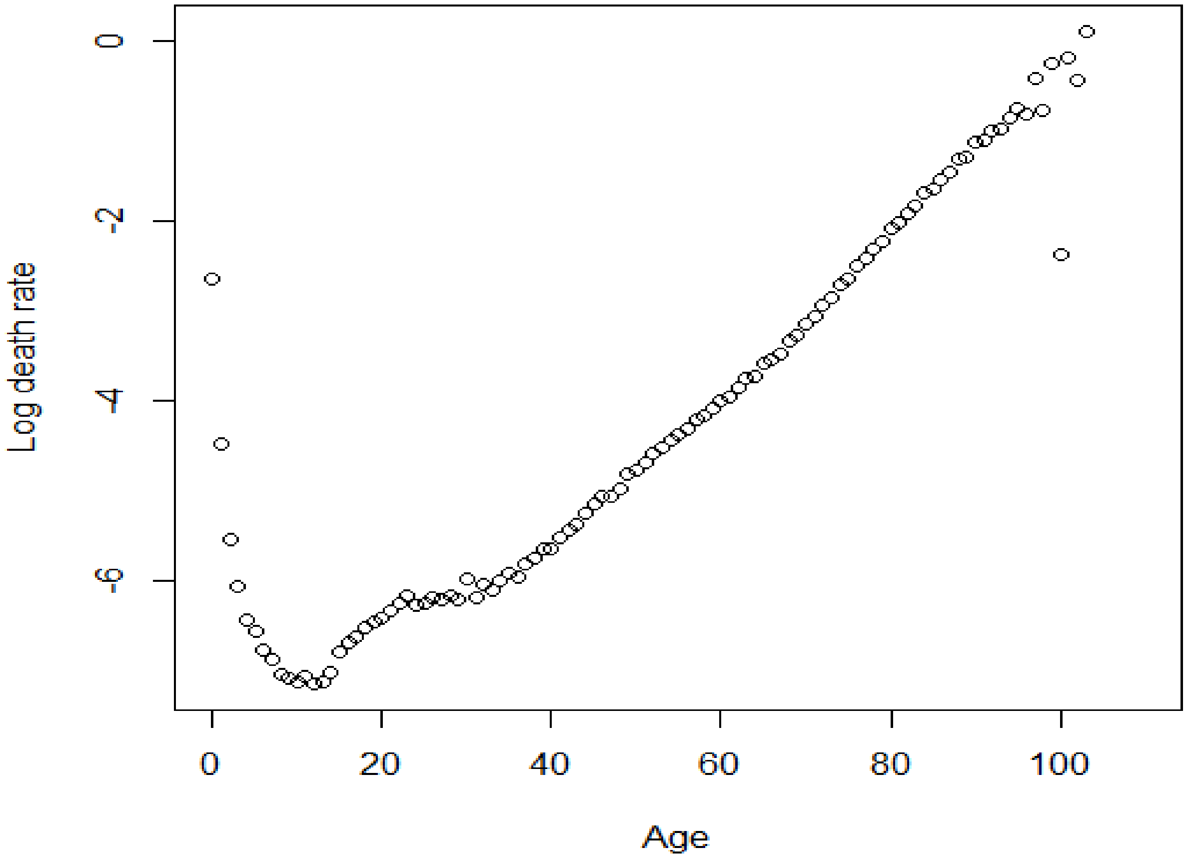


Figure 7: Weighted Constrained Penalized Regression splines

The figure above is an example of the fact that mortality rates follow a smooth function with some observational error, which has a high variance at very old ages dueto small population and also at young ages due to low rate of mortality.

The biggest advantage of smoothing the mortality data is that the smoothed data has alower variance than the tables constructed from the original observed mortality rates which implies that the former gives us a more accurate and unbiased estimate of life expectancy. Here *ft(x)*, is a smooth function of x and it can be estimated by kernel smoothing, loess or splines. The most common methods of estimating *ft(x)* are the regression splines.

5.1. Weighted Constrained Penalized Regression Splines:

Under the constrained and weighted penalized regression splines for estimating *ft(x)*, the weightage looks after the heterogeneity and monotonic constraints for higher age groups leading to better estimates.
The advantage of this technique is that it is quick to compute as well as it accommodates for monotonicity constraints to be imposed instantly.

Figure 7: It indicates the male death rates under the weighted constrained Regression splines.

5.2. P-Splines

P-Splines are effective in cases when there are outliers present in the data particularly when we observe older population, there is a high variability due to small group of survivors in that particular age range. This is when techniques related to smoothing comes into play and accommodates for this shortage of data. A higher variance at older ages has a huge impact on the fitting of mortality models and hence it is imperative for it to be taken care of.

Thus, Currie et al, (2004) suggested using two dimensional splines which adopts a Generalized linear modeling (GLM) framework. The fundamental principle involves using B-splines as the basis for regressions or improvise the log-likelihood function by applying a difference penalty on the coefficient of regression performed. The other approach of this is to follow Hyndman and Ullah (2007) approach where the smoothing is performed on the data first and then the values are fitted. This allows for placing a constraint on the smoothing more efficiently.

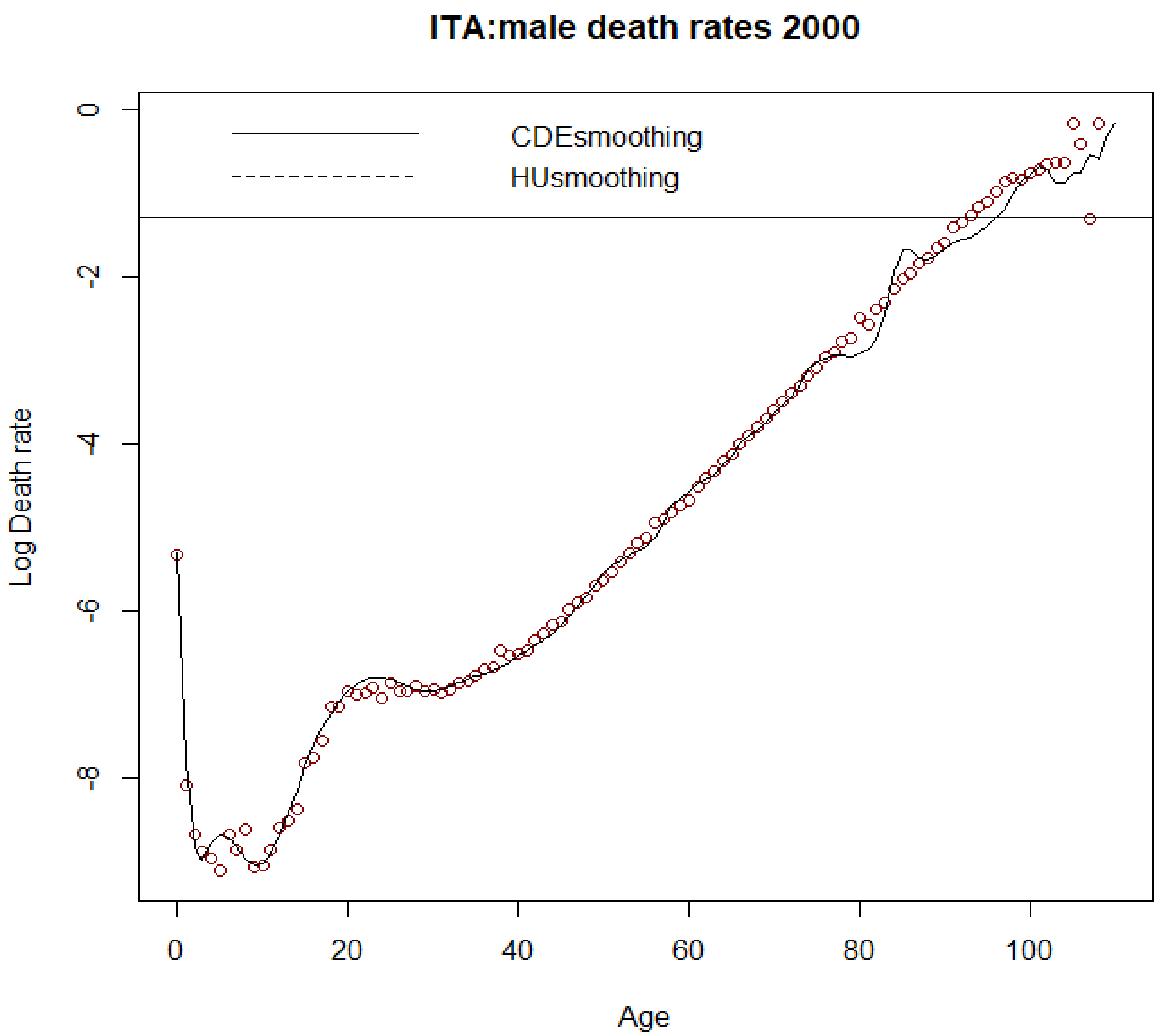


Figure 8: Currie Durban Eilers smoothing

5.3. Empirical Analysis

The analysis was carried out by fitting the LC model and the FDM version to the Italian Male Population. The percentage of variation as obtained by programing in R was 94.6%. The value of this could be high because the male mortality rates show a higher dispersion during older ages.

The coefficient that is fitted shows us that there has been significant improvement in mortality at a younger age. The graphs indicate a plummet in the mortality rate of children for the entire period and this is also understood clearly by the decreasing trend of the first coefficient. Thus for males the goodness of fit is better when moved from Lee Carter to Functional Demographic Model.

R codes showing the percentage of variation:

```
> legend("topright",lty=1:2,legend=c("ITAnone","LC"))
> lca(ita,series="male",adjust="dxt")
Lee-Carter analysis

Call: lca(data = ita, series = "male", adjust = "dxt")

Adjustment method: dxt
Region: ITA
Years in fit: 1872 - 2014
Ages in fit: 0 - 100

Percentage variation explained: 94.6%
```

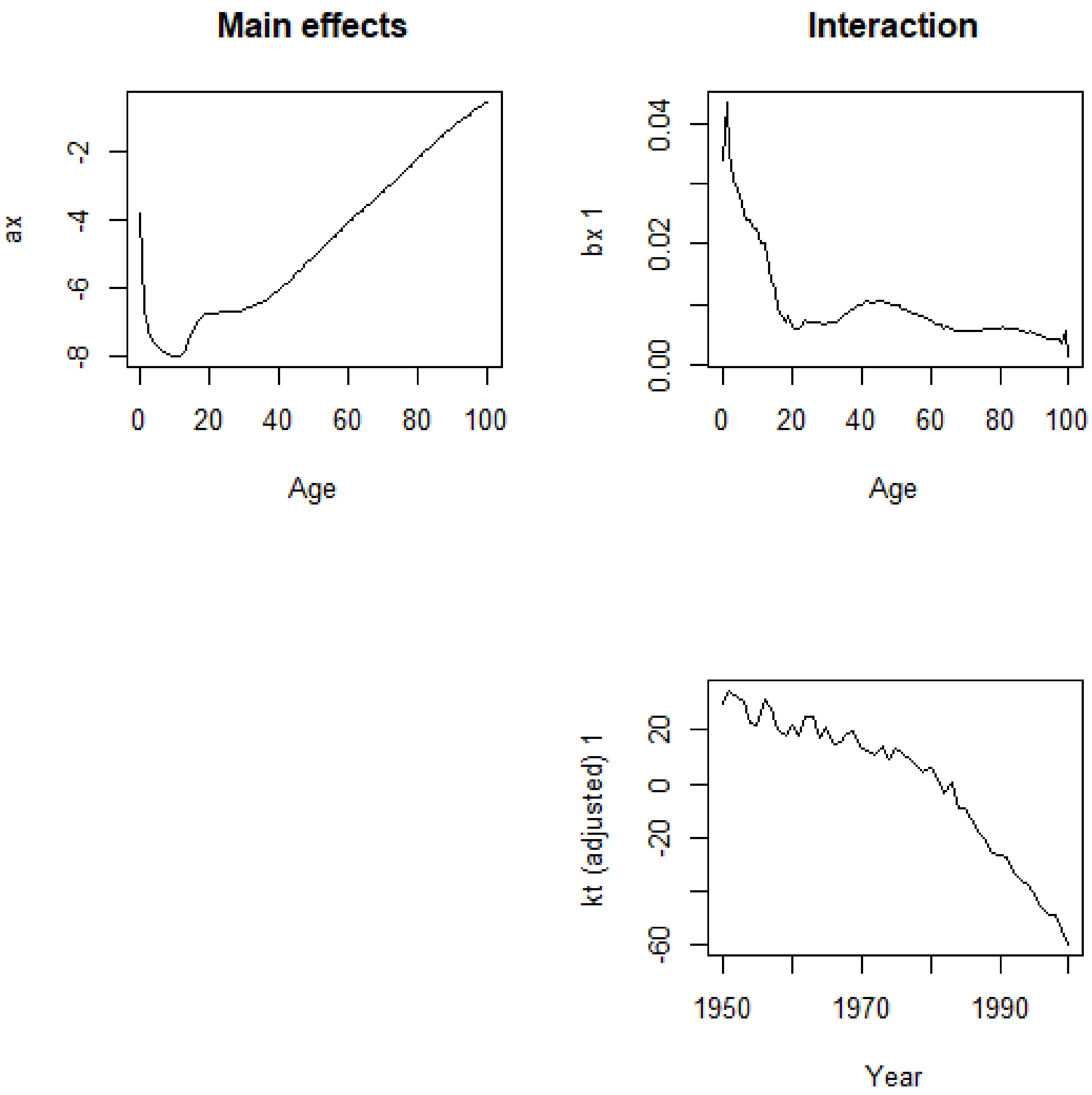


Figure 9:The parameters obtained

However, it is imperative to keep the analysis carrying forward to understand and also check if the fitting of FDM is really dependent on smoothing the data or is it due to other unknown factors.

Thus we next carry out smoothing the data using monotonic P-Splines after using the LC method to smooth the data. As mentioned earlier the percentage variation is 94.6% and it increases when we apply the P-splines to the LC model. This happens because there takes place a transformation of the data into a less variable data which leads to a higher Mean Squared Error for the LC model compared to the MSE of FDM model.

Lee-Carter model

	ME	MSE
Average across ages	0.000677	0.001397
Average across years	0.045943	0.872788

Functional Demographic Model

	ME	MSE
Average across ages	0.0000001	0.000543
Average across years	0.0000221	0.309439

Regardless of the values, it is quite plausible for a model to fit in the historical data well thereby having a goodness of fit but still produce poor forecasts which would be significantly different from the possible outcomes. This is when the technique of Backtesting is applied to check the outcomes that could have been produced if the model were used in the past.

Here, the focus of this report is to investigate the possibility of different mortality models by using various smoothing techniques instead of quantifying longevity risk and hence we will focus on approaching other methods of modeling the data.

Thus the next few sections would focus on the projections of the smoothed mortality rates using bivariate P-splines from 2017 to 2030 under the variants of Lee-Carter methods, that is the basic LC model, Hyndman-Ullah model and the Robust Hyndman-Ullah Method.

The basic feature of the Hyndman-Ullah approach is that the log mortality rates are smoothed before performing the modeling which involves the use of functional principle multiple components analysis.

The log mortality rates are smoothed using p-splines regression which also emphasizes that the age, x is assumed to be a constant variable. The homoscedastic error in the LC model is rectified in this case as the amount of noise is allowed to vary with various ages in various years.

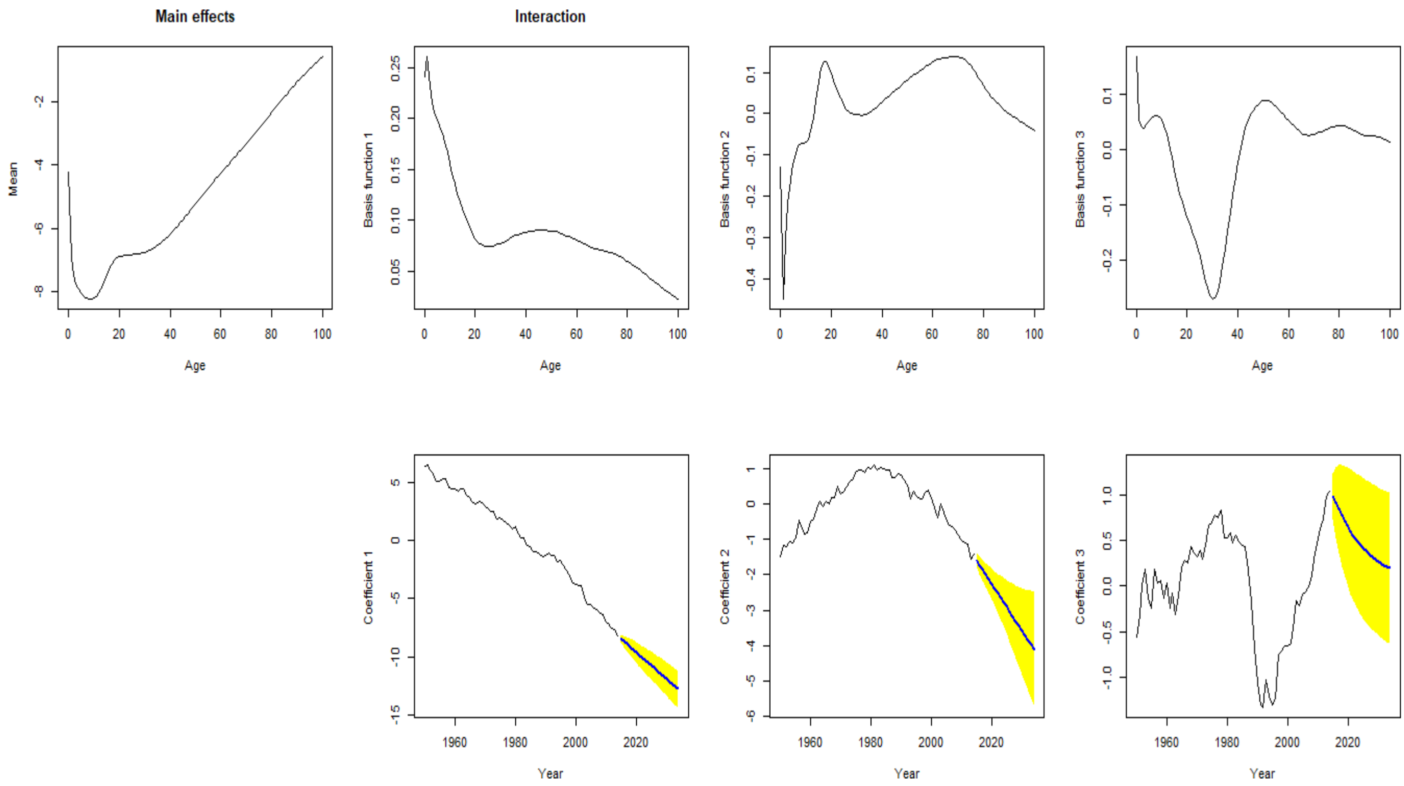


Figure 10: Ratio decomposition with forecast of Italian Male population

5.4. Robust Hyndman-Ullah (HUrob) Method

This method focuses solely on the presence of outliers that can seriously temper with the performance of the model and thereby the forecast results. this method obtains the projected pursuit estimates of the important components with the associated scores. The squared error which is integrated gives the accurate measure of the chief component approximation for every year. The years that lie in the outliers category would result in a larger integrated squared error when compared to the critical values obtained through normality of et(x). Thus this method assigns 0 weight to the values that are outliers to model and forecast mortality rates, reducing any possible chances of the presence of outliers.

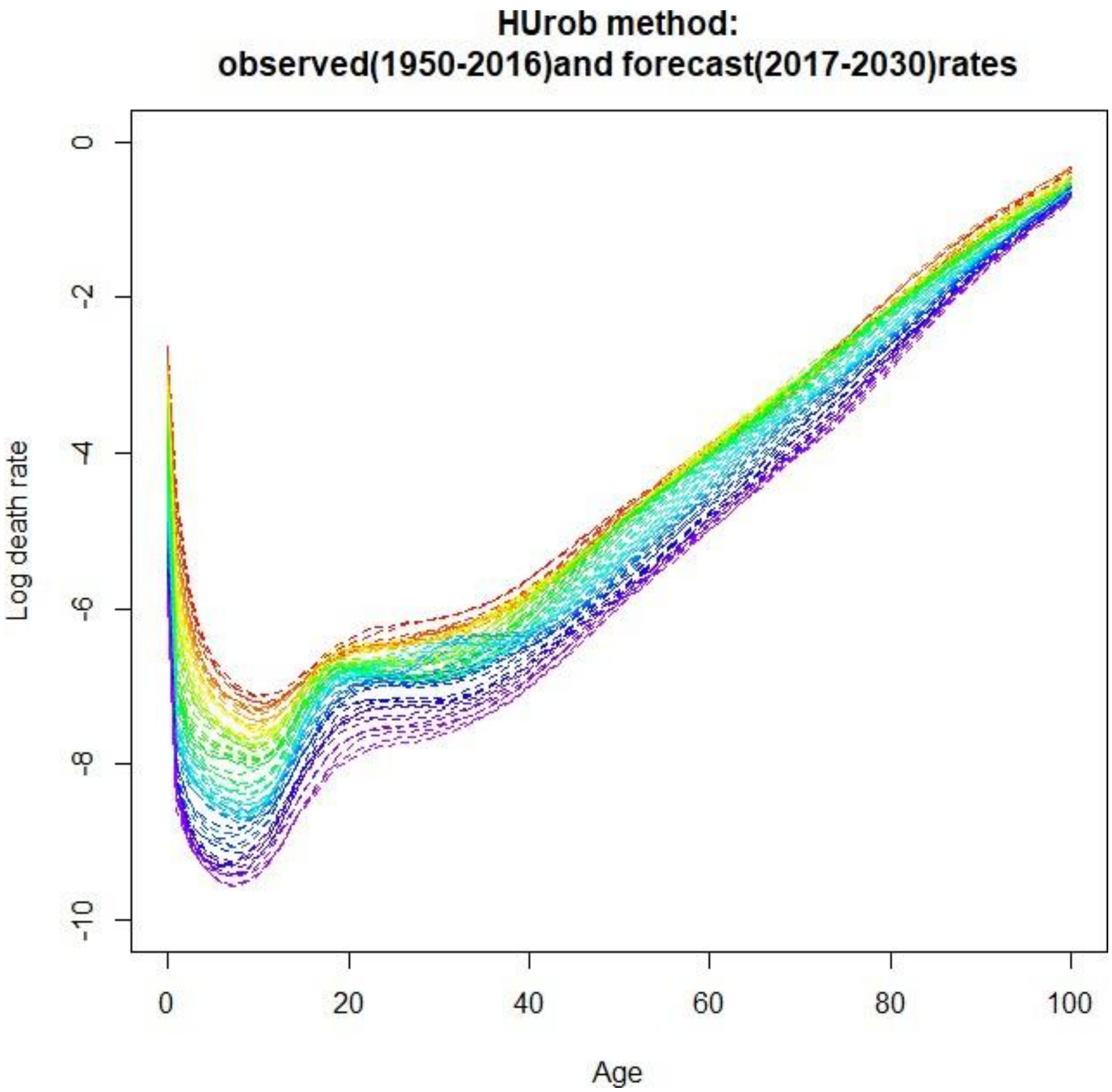


Figure 11: The results obtained under this model show a forecasted value along withobserved rates

5.5. Life table forecasting:

This section would look into the forecasting methodology involved in obtaining forecasted mx rates, which would be further used to forecast the complete life table functions, by involving standard methods.

To perform the analysis the lifetable function in the demography package is used and to forecast the expectancy, flife.expectancy function is used where the strting age is 0. The mx rates that are involved for the purpose of life table may be the rates being applied in a future forecast of t to generate a cross-sectional or a period life table (see appendix A) that focuses on a point forecast.

There are other ways of forecasting the prediction intervals for future life expectancies. For this purpose we have to simulate the forecasted log mortality rate as used un Hyndman& Booth 2008. There are various disturbances added to forecast basis function. Once the forecast of the parameter has been carried out, then the life expectancy of each set of simulated log mortality is calculated. This also helps us prepare the predicted intervals using the percentile that are obtained from the simulated life expectancies. All these calculations are performed in the demography package which has inbuilt functions to model the rates under various models and simulate using the life table functions.

CONCLUSION:

Therefore, it is very evident from the analysis that Lee-Carter is a robust method to fit a large population size. Lee Carter plays a vital role in undermining the forecasted mortality rates without any bias in the estimation of its parameters. This was evident because the residuals were mostly around 0 and scattered randomly. Although the basic Lee carter model did provide us with a goodness of fit, there are other variants of Lee carter like the Lee Miller Method or the Booth-Maindonald Smith Method which follow different assumptions like the jump off rates under Lee-Miller are the ones that are the observed rates unlike BMS where it is the fitted rates. Thus, it makes it compulsory to check for the data under various models to see if there is any closer variant giving the closest fit to the data to the model thereby improving the chances of mre accuracy in predicting the parameters and estimating the mortality rates of the future. Also, some approaches follow smoothing the mortality data before performing the simulation like in the Hyndman-Ullah and its variant Robust Hyndman-ullah approach. This approach can also be very meaningful in case any outliers existed as it would provide a higher level of accuracy of the principal component approximation for each year.

APPENDIX A:

Point Forecast

2000	10.523381
2001	10.644757
2002	10.766287
2003	10.895763
2004	11.041932
2005	11.171262
2006	11.308915
2007	11.434584
2008	11.564812
2009	11.693568
2010	11.821808
2011	11.938510
2012	12.049997
2013	12.164532
2014	12.268565
2015	12.373612
2016	12.474424
2017	12.574426
2018	12.673777
2019	12.772880
2020	12.871722
2021	12.970289
2022	13.068624
2023	13.166732
2024	13.264631
2025	13.362327
2026	13.459828
2027	13.557138
2028	13.654261

REFERENCES:

Alai DH, Sherris M (2014). "Rethinking Age-Period-Cohort Mortality Trend Models." Scandinavian Actuarial Journal, (3), 208–227.

BOOTH, H., MAINDONALD, J. and SMITH, L. (2002). Applying Lee–Carter under conditions of variable mortality decline. *Population Studies* **56** 325–336.

BOX, G. E. P and JENKINS, G. M. (1976). *Time Series Analysis for Forecasting and Control*. Holden-Day, San Francisco.

CAIRNS, A., BLAKE, D., DOWD, K., COUGHLAN, G. D., EPSTEIN, D., ONG, A., BALEVICH I., BROUHNS, N., DENUIT, M. and VERMUNT, J. K. (2009). A quantitative comparison of sto- chasting mortality models using data from England and Wales and the United States. *N. Amer. Actuar. J.* **13** 1–35.

Cairns, A.J.G. (2011) Modelling and management of longevity risk: approximations to survival functions and dynamic hedging. Insurance: Mathematics and Economics, 49, 438-453.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. Journal of Risk and Insurance, 73, 687-718.

Camarda CG (2012). "MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines." Journal of Statistical Software, 50(1), 1–24.

Clayton D, Schifflers E (1987) Models for temporal variation in cancer rates. II: Age-period- cohort models. Statistics in Medicine, 6, 469-81.

CMI (2006). Stochastic projections methodologies: Further progress and P-spline model feature, example results and implications. Working Paper 20, Continuous Mortality Investigation.

CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2004a). Smoothing and forecasting mortality rates. *Statist. Model.* **4** 279–298. MR2086492

Currie ID (2006). "Smoothing and Forecasting Mortality Rates with P-Splines." URL http://www.macs.hw.ac.uk/{~}iain/research/talks/Mortality.pdf.

DELWARDE, A., DENUIT, M. and EILERS, P. (2007). Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting. *Statist. Model.* **7** 29–48.

Delwarde A, Denuit M, Eilers P (2007). "Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized log-likelihood approach." Statistical Modelling, 7(1), 29–48.

Durban M, Currie ID (2003) A note on P-spline additive models with correlated errors. Com- putational Statistics, 18, 251-62.

Hunt A, Blake D (2015). "On the Structure and Classification of Mortality Models Mortality Models." Pension Institute Working Paper. URL http://www.pensions-institute.org/ workingpapers/wp1506.pdf.

Hurvich CM, Simonoff JS, Tsai C-L (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society, Series B, 60, 271-294.

Hobcraft J, Menken J, Preston S (1982). “Age, Period, and Cohort Effects in Demography: A Review.” Population Index, 48(1), 4–43.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute of Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 1 August 2012).

Hyndman RJ (2017). forecast: Forecasting Functions for Time Series and Linear Models. R package version 8.0, URL http://github.com/robjhyndman/forecast.

Hyndman RJ, Booth H, Tickle L, Maindonald J (2014). demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.18, URL http://CRAN. R-project.org/package=demography.

Hyndman RJ, Khandakar Y (2008). “Automatic Time Series Forecasting: The forecast Pack- age for R.” Journal of Statistical Software, 27(3), 1–22. URL http://www.jstatsoft. org/v27/i03/.

LEE R. D. and MILLER, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography* **38** 537–549.

Lee TCM (2003) Smoothing parameter selection for smoothing splines: a simulation study. Computational Statistics & Data Analysis, 42, 139-148.

Marx BD, Eilers PHC (1999) Generalized linear regression on sampled signals and curves: a P-spline approach. Technometrics, 41, 1-13.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Founda- tion for Statistical Computing

RENSHAW, A. E. and HABERMAN, S. (2003a). Lee–Carter mortality forecasting: A parallel generalised linear modelling approach for England and Wales mortality projections. *Appl. Statist.* **52** 119–137. MR1959085

RENSHAW, A. E. and HABERMAN, S. (2003b). Lee–Carter mortality forecasting with age specific enhancement. *Insurance Math. Econom.* **33** 255–272. MR2039286

Steven Haberman and Maria Russolillo (2005) Lee Carter Mortality Forecasting: An application to the Itaian population.

Turner H, Firth D (2015). Generalized Nonlinear Models in R: An Overview of the gnm Package. R package version 1.0-8, URL http://CRAN.R-project.org/package=gnm.

Villegas AM, Millossovich P, Kaishev V (2017). StMoMo: An R Package for Stochastic Mor- tality Modelling. R package version 0.4.0, URL http://CRAN.R-project.org/package= StMoMo.

WANG, C. and LIU, Y. (2010). Comparison of mortality modelling and forecasting—empirical evidence from Taiwan. Lee–Carter mortality forecasting with age specific enhancement. *Int. Res. J. Fin. Econ.* **37** 46–55.

Wand MP (2003) Smoothing and mixed models. Computational Statistics, 18, 223-50.

Wood SN (2003) Thin plate regression splines. Journal of the Royal Statistical Society, Series B, 65, 95-114.

Wikipedia (2014). "King Momo, in Wikipedia, The Free Encyclopedia." URL http://en.wikipedia.org/w/index.php?title=King{_}Momo{&}oldid=635124783.

Willets R (2004). "The Cohort Effect: Insights and Explanations." British Actuarial Journal, 10(4), 833–877.