Distr.: General
30 March 2016

English only

Economic Commission for Europe

Conference of European Statisticians

Joint Eurostat/UNECE Work Session on Demographic Projections
Geneva, 18-20 April 2016
Item 2 of the provisional agenda
Methodology

Population projections when using time series with extreme values

Note by STATISTICS ICELAND[1]

Summary

The central idea of this paper is that by rigorous analysis of the time series data on migration, births and deaths, one can build valid statistical models, which can then be used for calculating point estimates and prediction intervals of future values of population components.

The problem of outliers and/or extreme values arises in two types of data during this process: (i) values of migration (age-) time series in presence of strongly fluctuating exogenous, economic or social, variables; (ii) values of age-time series of mortality and fertility rates, where age is the age of death for mortality data and age of mothers for fertility data respectively. In addition, one has to analyse and give consistent forecasts of the net migration but also of its components like immigration and emigration by sex, citizenship and possibly other attributes. We describe our solutions to such challenges which consist of:

(i) modeling the non-stationary and auto-correlated time series and the impact of extreme values using:
i1) vector autoregressive distributed lags and general ARIMA models for short and long term migration components respectively, as well as an optimal combination approach for components' hierarchy
i2) functional data models (introduced by Hyndman in 2007) for fertility and mortality rates which use time series coefficient functions in orthonormal function expansions of the rates and which is robust with respect to extreme values

(ii) detecting and identifying the types of outliers/extreme values in migration series and associated exogenous variables, by hypotheses testing using the distributions of problem dependent models' residuals

(iii) calculating prediction intervals for population components and the probability of future shocks for migration components

The source of our demographic data is the Icelandic National Register, while the social and economic data and short term predictions are provided by Statistics Iceland. All time series are 45 years long.

[1] [Prepared by Violeta Calian, Senior Statistician].

I. Introduction

1. Population projections and forecasts are important for a wide spectrum of users, due to the social and economic impact of demographical changes. The methods used by Statistics Iceland have been improved continuously in recent years, in order to find solutions to recent data challenges. We focus in this paper on the way we use statistical models and hypothesis testing in order to provide predictions for population components, given strongly fluctuating, non-stationary and auto/cross-correlated time series data with extreme values and/or outliers. We analyse two types of problems encountered when forecasting: (i) the fertility and mortality rates and (ii) the (vector of-) migration components respectively.

2. Data on number of births by age of mothers and number of deaths by age and sex poses some challenges when calculating fertility, mortality rates and life tables. The main reason consists of zero counts for very small or very high ages. This happens due to the small size of the population and to the fact that most of these counts refer to rare events.

3. Standard solutions to such problem are to aggregate data over several years or to borrow data from similar and bigger populations, such as from other Nordic countries. Borrowing data is limited by the type of analysis though, and needs to be preceded by a proof of validity, i.e. testing hypotheses about the distributions of the needed variables in the populations.

4. Theoretically, a most reliable solution consists of calculating the probability of these rare events, based on their distributions most likely obtained by using bootstrap. This analysis is ongoing. The practical method we use at the present time is described in the following section and is based on smoothing and orthonormal expansions, in the context of functional data modelling of the rates.

5. The migration data presents the problem of hierarchical/grouped time series. One has to give consistent forecasts of the net migration but also of its components like immigration and emigration by sex, citizenship and possibly other attributes. In addition, strongly fluctuating data contains outliers, i.e. values generated by different underlying processes, and real life distributions of migration data often have significant tails, i.e. extreme values.

6. Therefore, before modelling migration we test a set of hypotheses, appropriately adjusting for multiple comparisons. They are related to the presence of outliers in: a) the time series of the net or various migration components as well as in b) the age patterns. The tests are based on estimates and confidence intervals of model parameters for: a) the stationary time series of migration rate differences and b) the stationary and ergodic processes underlying the age group processes. The empirical distributions of extreme values allow us to estimate probabilities of future extreme events. Our migration models for short term forecasting are joint (vector) auto-regressive distributed models, described in the next section. They take into account the hierarchical structure, complex correlations and non-stationarity of the migration components and influencing external factors.

II. Modelling solutions

a) Fertility and mortality rates

7. Let $y(t,x)$ denote the log of the observed mortality or fertility rate for age x and year t. In general, we could assume there exists an underlying smooth function $f(t,x)$ which is observed with error and at discrete points (t_i, x_a) of a (time-age) two-dimensional domain, giving the values $\{t_i, x_a, ; y_{ia}\}$, with $i = 1, \dots, N$ and $a = M_0, \dots, M$. We need to predict $y(t,x)$ for the same set of age values x_a ($a = M_0, \dots, M$) and for years t_i ($i = N+1, \dots, N+h$), where h is the length of the forecasting horizon. Due to the asymmetry of the age-temporal domain in a forecasting context, to consistency issues and correlations of the rates across time and age values, it is difficult to find a parametric model for the (vector-) function $f(t,x)$ without using simplifying assumptions. Ideally, a general decomposition of the (smooth) function should be $f(t,x)$ $=\sum_{j,b} \omega_{j,b} \varphi_{j,b}$ (t,x), with $\varphi_{j,b}$ being functions of an orthonormal basis over the bi-dimensional age-time domain.

8. A simpler and more efficient form is given by the factorization: $f_0(t,x)$=$\sum_{k=1,\dots K} \beta_k(t) \varphi_k$ (x), where the number of orthogonal functions K is reasonably small, and this is the solution proposed by Hyndman (2007, 2008), it is robust to outliers, and it has been tested in a sufficiently extensive way. In this case the observations are modelled as: $y_{ia} = \mu(x_a) + f_0(t_i, x_a) + \mathrm{e}_{t_i}(\mathrm{x}_a) + \alpha_{t_i}(\mathrm{x}_a)\varepsilon_{t_i, x_a}$, where $\mu(x)$ is the mean of $f_0(t,x)$ across time (years), $\mathrm{e}_t(\mathrm{x})$ is the residual modelling error (assumed serially uncorrelated), the coefficient functions $\beta_k(t)$ are independent (by construction), $\varepsilon_{t,x}$ represents the random variation in birth or death rates and $\alpha_t(\mathrm{x})$ allow the variance to change with age and time.

9. The method has several steps:

1) Smoothing the raw data, i.e. log of crude mortality or fertility rates, by using spline functions with constraints on concavity and monotonicity, as functions of time and age. This reduces the observational noise.

2) Expressing the smooth functions as series expansions over a basis of orthonormal functions ($\mathrm{f}_0(\mathrm{t},\mathrm{x})$) above). Fitting time series models for the coefficient functions $\beta_\mathrm{k}(\mathrm{t})$ of these series expansions and using these models for forecasting.

3) Using the forecast values of the coefficients to predict the values of the smooth functions and thus to predict mortality or fertility rates. Calculate prediction intervals based on the estimated variances of the error terms of step 1 and step 2.

10. The results are as follows:

(i) Figure 1 shows the past and forecast values of fertility rates by age. The variation due to orthonormal basis functions used in modelling is: 76.4%, 15.3%, 3.7%, 1.4%, 0.8% and 0.7%. We see that the increase in mothers' modal age with time will continue in the next 50 years. It is also clear that the fertility is predicted to decline for almost 30 years and then slightly increase again. This increase is due to the local peak in birth rate which occurred in 2008-2010 and to the average age (around 30 years) of mothers.

(ii) Figures 2 and figure 4 show the mean age pattern, the orthogonal basis functions and model coefficients for female and male mortality rates, respectively. The variation due to orthonormal basis functions is: 79.1%, 9.4%, 3.9%, 2.7%, 1.9%, 1.0% for the female model and 89.0%, 3.3%, 2.6%, 1.5%, 1.2%, 0.9% for the male model. The residuals (visualised in Figures 3 and 5) prove that we can use the fitted models to make predictions for future values of the mortality rates. The first coefficient function and basis function show in both cases that the mortality has consistently decreased over time but the speed of this improvement depends on age. Thus very small ages and people 40 to 80 years old are the main beneficiaries of the trend, a similar conclusion to another population analysed with the same method (Hyndman (2008)). In Figures 6 and 7, the past and forecast of female and male logarithm death rates are represented. We notice again the way the decrease in mortality over time depends on age but also that mortality changes are smaller around young adult ages than older.

(iii) For both fertility and mortality rates we obtain short and long term forecasts, i.e. point estimates and prediction intervals, based on functional data models with time series coefficients, which do not depend on exogenous variables or any subjective inputs.

11. In the case of fertility rates we also have three variants for the long term values, given by expert assumptions. The difference between the forecast point estimate and the medium value assumption is of the order of 10^{-2} and the differences between the lower/upper bounds of the prediction intervals and the low/high values given be experts are of the order of10^{-1}. Therefore, for the long term, our prediction intervals are smoothly connected to the expert assumption values. In the long term, total fertility rates are expected to converge to 1.8, 1.95 and 2.1 for the low, medium and high variants, respectively.

b) Migration rates

(i) Short term migration

12. We use the vector generalization of auto-regressive distributed lag models for the auto-correlated and non-stationary time series involved in migration processes, in order to give valid point estimates and prediction intervals of migration rates. We obtain short time predictions for the net migration and for the number of immigrants/emigrants of Icelandic and foreign citizenships as functions of several time series predictors: unemployment, change in GDP values, number of graduating high school students and dummy variables mirroring the EEA resizing in time and the Icelandic economic boom which ended in 2008.

13. We have built, in a parsimonious way, the vector dynamical model $y(t) \sim y(t-j) + x(t-l)$ where y and x are vectors of migration components and of exogenous variables. The vector components have similar forms, i.e. $y_i(t) \sim \sum y_k(t-j) + \sum x_p\,(t-l)$, where $j = 0,1,2, \dots$ up to maximum lag order in dependent variables, and $l = 0,1, \dots$ up to maximum lag order in exogenous variables.

14. We use the following notation - and an alternative one, for the ease of interpretation- for the components: y_1, y_2, - (*ImIceM*, *EmIceM*) - the number of

Icelandic immigrants/emigrants, men; y_3, y_4 – (*ImIceW*, *EmIceW*) - the number of Icelandic immigrants/emigrants, women; y_5, y_6 – (*ImForM*, *EmForM*) - the number of foreign immigrants/emigrants, men, y_7, y_8 – (*ImForW*, *EmForW*) - the number of immigrants/emigrants, women of foreign citizenship; x_4 - *UnEmpl* -the unemployment rate; x_8 - *GDP*, a measure of GDP, x_5, x_6 – (*GradM*, *GradF*), the number of graduating students, men and women respectively; $boom$ – an indicator variable coupled to the Icelandic economic boom, reflecting also temporary changes in the registration process; eea – an indicator variable which reflects the entrance of Iceland into the *EEA*, and thus free movement of persons within that area.

15. The models take particular simple forms, here written by using again the standard R notation (for dynamical models in our case) and the more meaningful variable names:

$$\begin{aligned}
y_1 &= ImIceM(t) \sim ImIceM(t-1) + UnEmpl(t) + UnEmpl(t-1) + EmIceM(t-1) \\
y_2 &= EmIceM(t) \sim EmIceM(t-1) + EmIceM(t-2) + GradM(t-2) + EmIceW(t) \\
y_3 &= ImIceW(t) \sim ImIceW(t-1) + UnEmpl(t) + GDP(t) + UnEmpl(t-1) + GDP(t-1) \\
y_4 &= EmIceW(t) \sim EmIceW(t-1) + EmIceW(t-2) + GradW(t-2) \\
y_7 &= ImForW(t) \sim ImForW(t-1) + UnEmpl(t) + GDP(t) + boom(t) + eea(t) + UnEmpl(t-1) \\
&\qquad + GDP(t-1) \\
y_8 &= EmForW(t) \sim EmForW(t-1) + ImForW(t) + ImForW(t-1) + UnEmpl(t) + GDP(t) \\
&\qquad + UnEmpl(t-1) + GDP(t-1) \\
y_9 &= y_{net}(t) \sim y_{net}(t-1) + y_{net}(t-2) + x_4 + x_8 + x_8(t-1) + x_4(t-1) + x_8(t-2) + x_4(t-2) \\
&\qquad + x_5(t-2) + x_6(t-2) + boom + bam + boom(t-1) + bam(t-1)
\end{aligned}$$

16. The variables y_5 and y_6 were obtained directly by using the empirically verified correlation between men and women migration numbers and the results of the models y_7, y_8. All univariate time series of 45 years length were tested for: (i) stationarity, by using augmented Dickey - Fuller and Kwiatkowski – Philips – Schmidt - Shin (KPSS) and (ii) auto-correlation of first and higher order, by using Durbin-Watson and Breusch - Godfrey tests. None of these series is I(2). These are necessary but not sufficient conditions (see Johansen 2010), for un-biased and consistent point estimates and independent and identically distributed residuals. The auto-regressive distributed lag models can be used to test for co-integration and to estimate long-run and short-run dynamics, even when the variables are stationary and non-stationary time series models (see Calian, V., Hardarson, O. (2015)). Choosing the structure and the order of the ARDL model by a consistent model selection criterion is a crucial step, too. We have applied standard tests to the residuals in order to establish the stationarity, normality, autocorrelation and goodness of fit of the models.

(ii) Long term migration

17. Structural models perform very well on short term but their main limitation is that they do require data on the future values of exogenous variables. These are not always easy to obtain. One could use instead purely probabilistic models, but one can argue that it is more efficient to take advantage, at least on the short term, of the information regarding various factors which influence the migration process. A new method based on alternative modelling is not yet sufficiently tested. Such a method could be based on a generalization of space-time series methods to the case

of age-time series such as the migration components, could use functional data models as for the fertility and mortality rates, or could use Bayesian priors as well as hierarchical modelling with carefully selected distributions for the factors which are significant when modelling migration. We are currently investigating the viability of these options.

18. We rely on both ARIMA modelling and on the expert opinion of Statistics Iceland’s advisory committee (on population projections) for predictions on long term migration. In the expert opinion, the three scenarios for the net migration are set by 400, 800 and 1,200 for the low, medium and high variants, respectively. In modelling the different migration patterns of Icelandic and foreign citizens, the long term net migration of Icelandic citizens is set as -800 for all variants. These are in very good agreement with the ARIMA point estimates and prediction intervals.

III. Conclusion

19. We have described in this paper the methodology used by Statistics Iceland for population projections, when using cross correlated, non-stationary time series data with extreme values and outliers. A detailed study of the performance of the employed models is the object of a future paper and it is based on using shorter time series in order to predict the (known) values of population components for recent years.

20. The dynamical models of short term migration can still be improved, especially if one aims to include more informative factors connected to internal and external social and economic processes. We are also investigating the possibility of building better models for long term migration. This can be done by using generalised age-time (analogue of spatial-temporal) autoregressive models and/or by using hierarchical or Bayesian models which take into account the distributions of exogenous factors. The functional data approach could be also improved by including the resampling based estimates of the rare events together with the crude rates before the smoothing stage and by choosing a more general type of orthonormal expansions.

21. Both classes of forecasts, functional data and dynamical model based, were shown to benefit from the information provided by the preliminary outlier and extreme value analysis.

References

Alho, J.M. (1997) Scenarios, uncertainty and conditional forecasts of the world population, *Journal of Royal Statistical Society* A 160, Part 1, 71–85.

Brunborg, H. and Cappelen, A. (2010) Forecasting migration flows to and from Norway using an economic model, *Joint Eurostat/UNECE work session on demographic projections*, Lisbon, Portugal, 28–20 April 2010.

Booth, H. (2006) Demographic forecasting: 1980 to 2005 in review, *International Journal of Forecasting*, 22(3), 547–581.

Calian, V. (2013) Dynamical models for migration projections, *proceedings of Joint Eurostat/UNECE Work Session on Demographic Projections*, Rome, 29.-31. október 2013, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.11/2013/WP_4.4.pdf

Calian, V., Hardarson, O, (2015) Methodology of population projections, Working Papers Statistical Series, http://www.hagstofa.is/media/49266/hag_151118.pdf

Hardarson, O. (2010) Long-term and short term migration in Iceland - Analysis of estimation methods of Statistics Iceland, *Working paper 13 at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses*, The Hague, The Netherlands, 10–11 May 2010.

Hyndman, R.J., & Ullah, M. S. (2007) Robust forecasting of mortality and fertility rates: a functional data approach, *Computational statistics & Data Analysis*, 61, 4942–4956.

Hyndman, R. J., Booth, H., (2008) Stochastic population forecasts using functional data models for mortality, fertility and migration, *International Journal of Forecasting* 24 (2008) 323–342.

Johansen, S. (1996) Likelihood-based inference in cointegrated vector autoregressive models, *Oxford University Press*, Oxford.

Johansen, S. (2010) The analysis of nonstationary time series using regression, correlation and cointegration with an application to annual mean temperature and sea level, *Discussion Papers*, no.10–27, Department of economics, University of Copenhagen.

Keilman, N., Pham, D. Q., Hetland, A. (2002) Why population forecasts should be probabilistic — illustrated by the case of Norway, *Demographic Research*, Vol. 6, Article 15.

Pesaran, M. H. (1999) An autoregressive distributed lag modelling approach to cointegration analysis, *Symposyum of the Norwegian Academy of Science and Letters*, Oslo, 3–5 March 1995, and DAE Working Paper Series No. 9514. Department of Econometrics, University of Cambridge, 1999.

Pesaran, M.H., Shin, Y. and Smith, R.J. (1996) Testing the existence of long-run relationships, *DAE Working Paper Series No. 9622*, Department of Econometrics, University of Cambridge.

Pesaran, M.H., Shin, Y. and Smith, R.J. (2001) Bounds testing approaches to the analysis of level relationships, *J. Applied Econometrics*, 16: 289–326.

Figures

Figure 1. Fertility rates 1971–2065

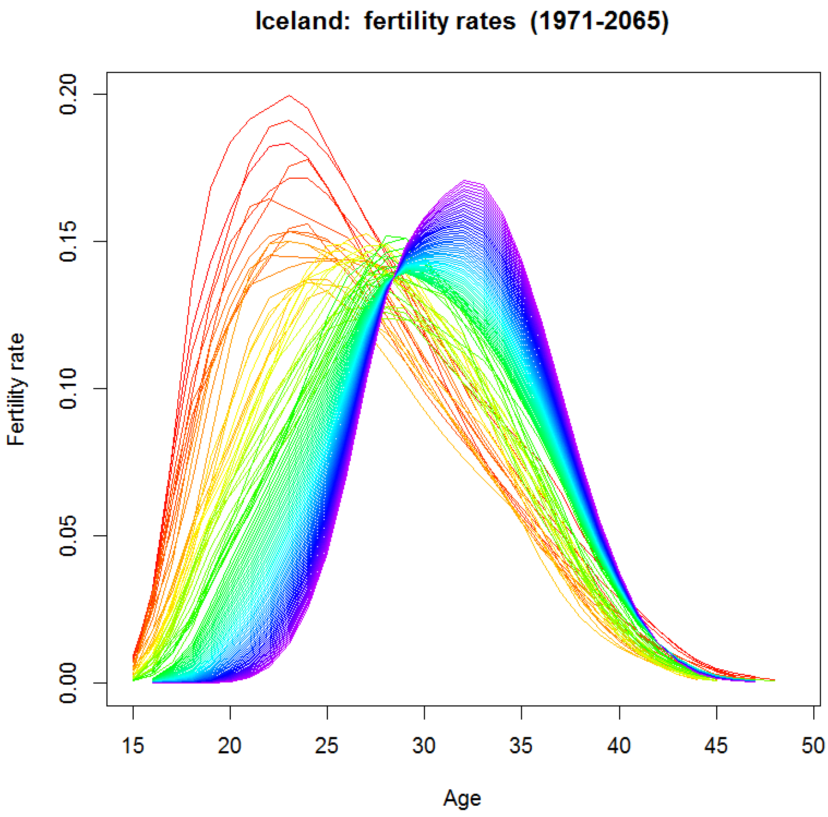


Figure 2. The basis functions and model coefficients for the female mortality rates

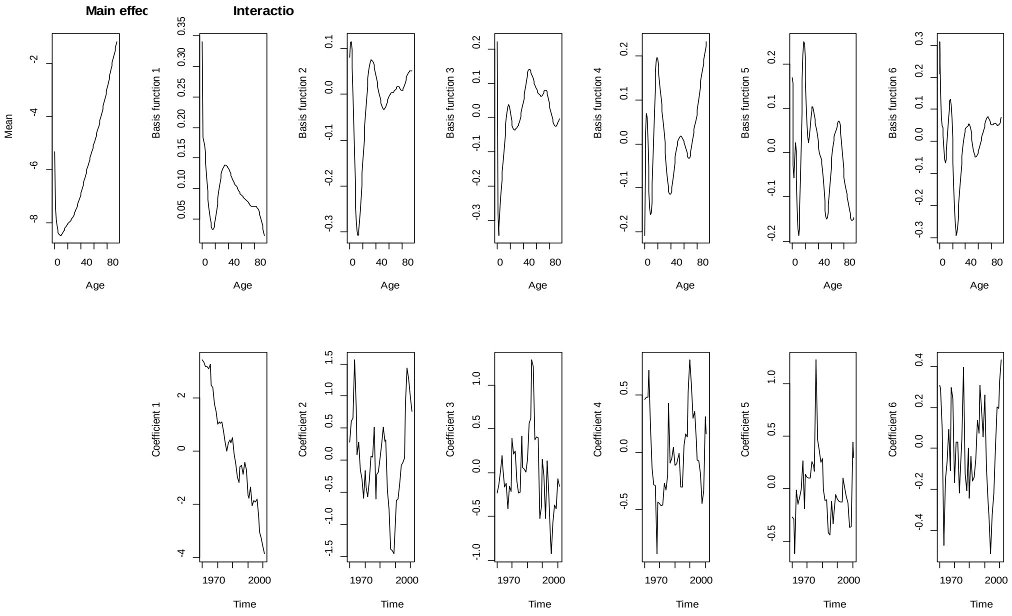


Figure 3. The residuals of the model for female mortality rates.
Darker colours mean higher values, in positive and negative directions (red or blue colours).

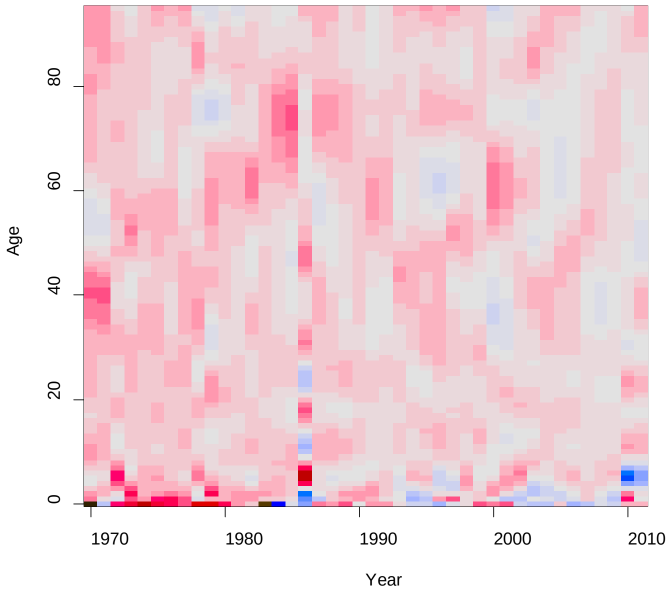


Figure 4. The basis functions and coefficients of the model for the mortality rates of men

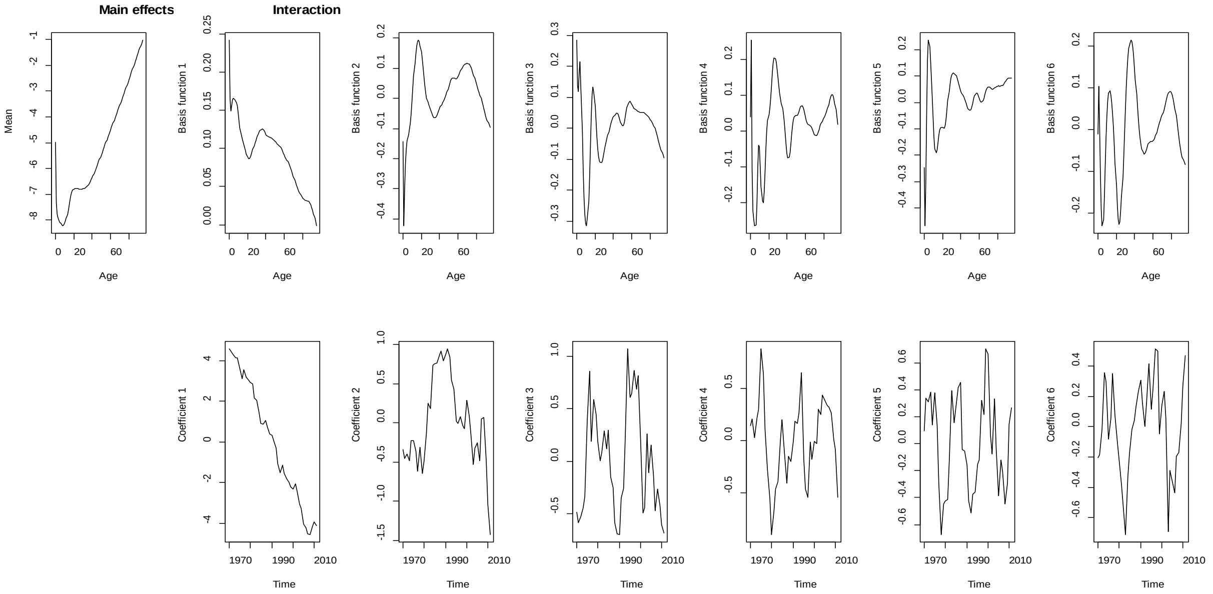


Figure 5. The residuals of the model for mortality rates of men
Darker colours mean higher values, in positive and negative directions (red or blue colours).

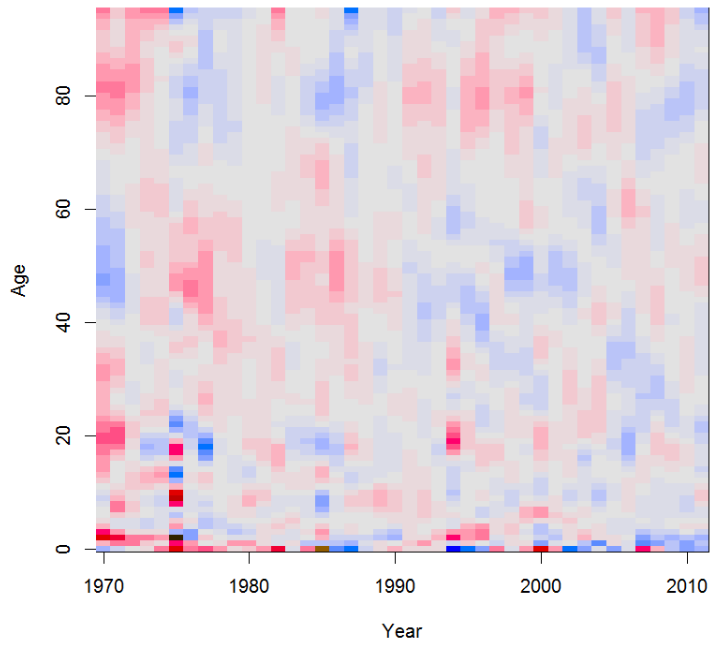


Figure 6. Logarithm of female death rates 1971–2065

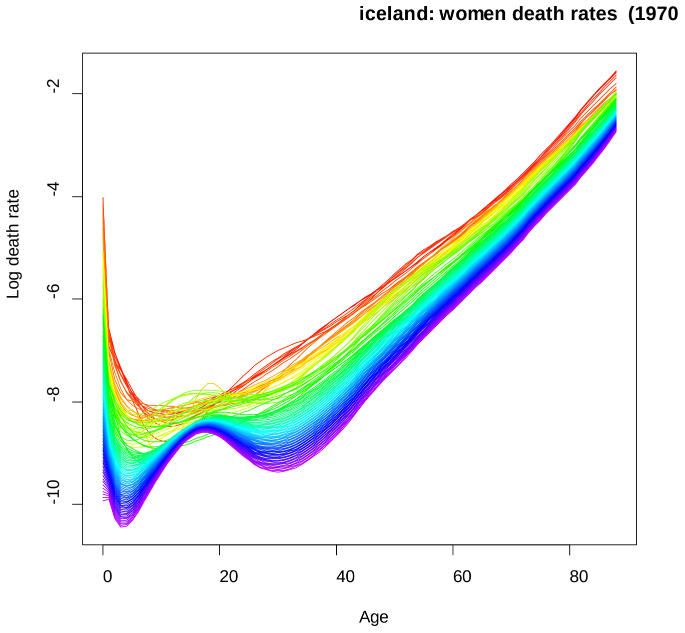
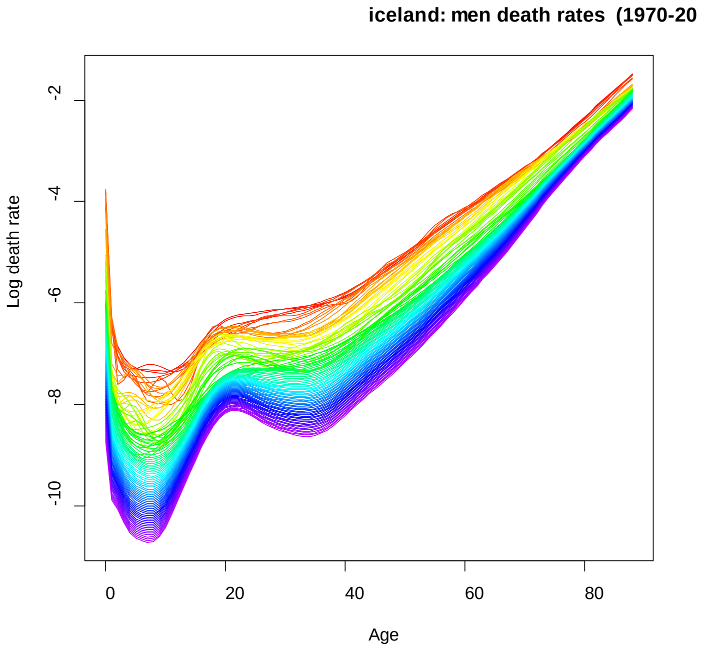


Figure 7. Logarithm of male death rates 1971–2065


iceland: men death rates (1970-20
Log death rate
Age