Appendix

A Data

Table A.1: Overview of countries included in the analysis

Country	HMD code (abbreviation)	Region	HMD data availability	# inclusions
Australia	AUS	America, Australia, Japan	1921 - 2018	4
Austria	AUT	Western Europe	1947 - 2017	3
Belarus	BLR	Eastern Europe	1959 - 2018	2
Belgium	BEL	Western Europe	1841 - 2018	4
Bulgaria	BGR	Eastern Europe	1947 - 2017	3
Canada	CAN	America, Australia, Japan	1921 - 2016	4
Czechia	CZE	Eastern Europe	1950 - 2018	3
Denmark	DNK	Scandinavia	1835 - 2019	4
Estonia	EST	Eastern Europe	1959 - 2017	2
Finland	FIN	Scandinavia	1878 - 2018	4
France	FRATNP	Western Europe	1816 - 2017	4
East Germany	DEUTE	Eastern Europe	1956 - 2017	2
West Germany	DEUTW	Western Europe	1956 - 2017	2
Hungary	HUN	Eastern Europe	1950 - 2017	3
Iceland	ISL	Scandinavia	1838 - 2018	4
Ireland	IRL	Great Britain	1950 - 2017	3
Italy	ITA	Southern Europe	1872 - 2017	4
Japan	JPN	America, Australia, Japan	1947 - 2018	3
Latvia	LVA	Eastern Europe	1959 - 2017	2
Lithuania	LTU	Eastern Europe	1959 - 2017	2
Luxembourg	LUX	Western Europe	1960 - 2017	2
Netherlands	NLD	Western Europe	1850 - 2016	4
Norway	NOR	Scandinavia	1846 - 2018	4
Poland	POL	Eastern Europe	1958 - 2016	2
Portugal	PRT	Southern Europe	1940 - 2018	3
Slovakia	SVK	Eastern Europe	1950 - 2017	3
Spain	ESP	Southern Europe	1908 - 2016	4
Sweden	SWE	Scandinavia	1751 - 2018	4
Switzerland	CHE	Western Europe	1876 - 2016	4
England & Wales	GBRTENW	Great Britain	1841 - 2016	4
Scotland	GBR_SCO	Great Britain	1855 - 2016	4
Northern Ireland	GBR_NIR	Great Britain	1922 - 2016	4
USA	USA	America, Australia, Japan	1933 - 2017	4

B Tree-Based Machine Learning Methods

B.1 Decision Trees

The decision tree method was first proposed by Breiman et al. 1984, and it can be applied to both regression and classification problems. According to Hastie et al. (2009), a decision tree is a nonparametric model that repeatedly splits the data into groups according to a set of feature/input variables, $\mathcal{F}$, in order to identify regions (subsets of the data) that are homogeneous in terms of the response variable, m. At every node, the algorithm identifies a set of possible splits for the feature variables.

It then chooses the feature variable and split point that maximizes the homogeneity of the response along each branch. Each branch creates new nodes that again can be split. The splitting continues until a stopping criterion is reached. For regression trees, the response variable is predicted using the mean response observed in each of the terminal nodes (i.e., in each region of the data).

Following, e.g., Hastie et al. (2009), suppose that data at a given node n can be represented by Q^n, and that there are $\theta = (f, s_n)$ candidate splits, each consisting of a feature variable f and a threshold/split point s_n. For each θ, Q^n is partitioned into two subsets:

$$Q^n_{left}(\theta) = (\mathcal{F}, m) \,|\, f \leq s_n, \tag{1}$$

$$Q^n_{right}(\theta) = Q^n \backslash Q^n_{left}(\theta). \tag{2}$$

The optimal split is chosen such that it minimizes the impurity at that node,

$$\theta^* = \arg\min_\theta Imp(Q^n, \theta) = \frac{N^n_{left}}{N^n} MSE\left(Q^n_{left}(\theta)\right) + \frac{N^n_{right}}{N^n} MSE\left(Q^n_{right}(\theta)\right), \tag{3}$$

where N^n_{left} and N^n_{right} are the total number of observations in each of the two subsets, $N^n = N^n_{left} + N^n_{right}$, and

$$MSE(Q^n(\theta)) = \frac{1}{N^n} \sum_{i \in N^n} (m_i - \bar{m}^n)^2, \tag{4}$$

with $\bar{m}^n = \frac{1}{N^n} \sum_{i \in N^n} m_i$ being the mean response in node n. These steps are repeated until a stopping criterion is met (e.g, maximum number of terminal nodes). The response is then predicted by $\bar{m}^n$ within each terminal node.

In contrast to the traditional stochastic mortality models that only include *year*, *age*, and *cohort* in the set of features, $\mathcal{F}$, this methodology allows the researcher to include *any* variable in $\mathcal{F}$ that he/she believes affects mortality.

B.2 Random Forests

Algorithm 1: Random forests for regression
(Algorithm 15.1 in Hastie et al. 2009)

1. For $k = 1, ..., K$:
   - 1.1. Draw a bootstrap sample from the training data.
   - 1.2. Construct a regression tree (denoted by T_k) from the bootstrap sample by recursively repeating the following steps for each node, n, of the tree, until a stopping criterion (e.g., maximum number of terminal nodes) is reached.
     - 1.2.1. Randomly select F_n feature variables from the F total number of features in $\mathcal{F}$.
     - 1.2.2. Choose the optimal feature and split point among the set of possible splits for the F_n features according to (3).
     - 1.2.3. Split the node into two daughter nodes according to the choice in 1.2.2.
2. Output the ensemble of regression trees $\{T_k\}_{k=1,...,K}$.

In step 1.2.2. of Algorithm 1, splitting a categorical variable (like *country*) is different from splitting a numerical or logical variable. For categorical variables, the split point is represented by an integer, whose base-2 (binary) representation defines the identities of the categories that go to the left (ones) and right (zeros).

B.3 Stochastic Gradient Boosting

Algorithm 2: Stochastic gradient boosting for regression
(See Friedman 2001 and Friedman 2002)

1. Initialize $\hat{m}_0(\mathcal{F}) = \arg\min_\gamma \sum_{i=1}^N L(m_i, \gamma)$, given a loss function $L(m, \gamma)$ as a function of the response variable, m, and the predicted values, γ.

2. For $k = 1, ..., K$:

   2.1. For $i = 1, ..., N$ compute the pseudo residuals

   $$r_{i,k} = -\left[\frac{\partial L(m_i, \hat{m}(\mathcal{F}_i))}{\partial \hat{m}(\mathcal{F}_i)}\right]_{\hat{m}=\hat{m}_{k-1}}. \tag{5}$$

   2.2. Randomly select a subsample of the training data of size $p \cdot N$, with p being a constant, pre-specified subsampling rate.

   2.3. Fit a regression tree to the $r_{i,k}$ values creating terminal regions $R_{j,k}$, $j = 1, 2, ..., J_k$ using only the subsample from the previous step.

   2.4. For $j = 1, 2, ..., J_k$ compute the output value

   $$\gamma_{j,k} = \arg\min_\gamma \sum_{\mathcal{F}_i \in R_{j,k}} L(m_i, \hat{m}_{k-1}(\mathcal{F}_i) + \gamma). \tag{6}$$

   2.5. Update $\hat{m}_k(\mathcal{F}) = \hat{m}_{k-1}(\mathcal{F}) + \nu \sum_{j=1}^{J_k} \gamma_{j,k} I(\mathcal{F} \in R_{j,k})$, where ν is a constant, pre-specified learning rate.

3. Output $\hat{m}(\mathcal{F}) = \hat{m}_K(\mathcal{F})$.

The optimal number of trees/iterations and the learning rate depend on each other. Smaller values of the learning rate almost always improves the predictive performance, but is associated with higher computational costs (more iterations are required). For small values of the learning rate and large number of iterations, the error rate is very flat. Thus, Friedman (2001) suggests that one should choose a small value for the learning rate while setting the number of iterations as large as is computationally feasible. For more details about the trade-off between number of trees and the learning rate, see Friedman (2001).

C Stochastic Mortality Models

The Augmented Common Factor (ACF) model developed by Li and Lee (2005) is an extended version of the Lee-Carter (LC) model built to handle multiple populations (e.g., men and women, different countries, etc.). In the ACF model, common mortality tendencies across populations are identified using a common factor approach, while at the same time, mortality schedules are allowed to vary between populations. The superscript i refers to a particular population. Thus, α_x^i and $\beta_x^i \kappa_t^i$ are population-specific, while $B_x K_t$ is specific to the 'pooled' population. The model is fit by first estimating the common factors, B_x and K_t, from a LC model of the pooled population,

Table A.2: Stochastic mortality models considered in this paper

Model and reference	**Formula**	**Identifiability constraints**
Lee-Carter (LC) Lee and Carter 1992	$\eta_{x,t} = \alpha_x + \beta_x \kappa_t$	$\sum_{t \in \mathcal{T}} \kappa_t = 0, \quad \sum_{x \in \mathcal{X}} \beta_x = 1$
Augmented Common Factor (ACF) Li and Lee 2005	$\eta_{x,t}^i = \alpha_x^i + B_x K_t + \beta_x^i \kappa_t^i, \quad \forall i \in \mathcal{R}$	$\sum_{t \in \mathcal{T}} K_t = 0, \quad \sum_{x \in \mathcal{X}} B_x = 1,$ $\sum_{t \in \mathcal{T}} \kappa_t^i = 0, \quad \sum_{x \in \mathcal{X}} \beta_x^i = 1, \; \forall i \in \mathcal{R}$
Cairns-Blake-Dowd (CBD) Cairns et al. 2006	$\eta_{x,t} = \kappa_t^{[1]} + \kappa_t^{[2]} (x - \bar{x})$	
Renshaw-Haberman (RH) Renshaw and Haberman 2006, Haberman and Renshaw 2011	$\eta_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x}$	$\sum_{t \in \mathcal{T}} \kappa_t = 0, \quad \sum_{x \in \mathcal{X}} \beta_x = 1,$ $\sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0$
Age-Period-Cohort (APC) Currie 2006	$\eta_{x,t} = \alpha_x + \kappa_t + \gamma_{t-x}$	$\sum_{t \in \mathcal{T}} \kappa_t = 0, \quad \sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x) \gamma_{t-x} = 0$
M6 Cairns et al. 2009	$\eta_{x,t} = \kappa_t^{[1]} + \kappa_t^{[2]} (x - \bar{x}) + \gamma_{t-x}$	$\sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x) \gamma_{t-x} = 0$
M7 Cairns et al. 2009	$\eta_{x,t} = \kappa_t^{[1]} + \kappa_t^{[2]} (x - \bar{x})$ $+ \kappa_t^{[3]} \left((x - \bar{x}) - \hat{\sigma}_x^2 \right) + \gamma_{t-x}$	$\sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x) \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x)^2 \gamma_{t-x} = 0$
(Full) Plat Plat 2009	$\eta_{x,t} = \alpha_x + \kappa_t^{[1]} + \kappa_t^{[2]} (\bar{x} - x)$ $+ \kappa_t^{[3]} (\bar{x} - x)^+ + \gamma_{t-x}$	$\sum_{t \in \mathcal{T}} \kappa_t^{[1]} = 0, \quad \sum_{t \in \mathcal{T}} \kappa_t^{[2]} = 0,$ $\sum_{t \in \mathcal{T}} \kappa_t^{[3]} = 0, \quad \sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x) \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x)^2 \gamma_{t-x} = 0$
(Reduced) Plat Plat 2009	$\eta_{x,t} = \alpha_x + \kappa_t^{[1]} + \kappa_t^{[2]} (\bar{x} - x) + \gamma_{t-x}$	$\sum_{t \in \mathcal{T}} \kappa_t^{[1]} = 0, \quad \sum_{t \in \mathcal{T}} \kappa_t^{[2]} = 0,$ $\sum_{t-x \in \mathcal{C}} \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x) \gamma_{t-x} = 0,$ $\sum_{t-x \in \mathcal{C}} (t-x)^2 \gamma_{t-x} = 0$

Notes: $\mathcal{T}$ is the set of calendar years, $\mathcal{X}$ is the set of ages, $\mathcal{C}$ is the set of cohorts, and $\mathcal{R}$ is the set of regions. Each model is fit and forecast in its pure form, as well as in combination with random forest, using Procedure 1.

which in this paper is regions (see Table A.1 in Appendix A). Next, α_x^i is estimated for each population as the average log-mortality rate at each age. Finally, β_x^i, and κ_t^i are estimated for each population by applying the singular value decomposition (SVD) to the residual matrix $\ln m_{x,t}^i - \alpha_x^i - B_x K_t$.

D The Model Confidence Set Procedure

Formally, the MCS procedure starts from an initial set of models, $\mathcal{M}_0$, consisting of all models described in Section 3. Assuming the total number of models is M, the MCS procedure then delivers a SSM, $\hat{\mathcal{M}}_{1-\alpha}^*$, consisting of $M^* \leq M$ models, given a user specified confidence level $1-\alpha$. Let $L_{i,t}$ be the loss function associated with model i at time t. The loss differential between model i and model j, $d_{ij,t}$, can then be defined as

$$d_{ij,t} = L_{i,t} - L_{j,t}, \qquad i,j = 1,...,M, \qquad t = 1,...,T. \tag{7}$$

The loss function applied in this paper is the squared error loss,

$$L_{i,t} = \left(\ln m_t - \widehat{\ln m}_t^i\right)^2. \tag{8}$$

The null hypothesis of equal predictive ability can be formulated based on the expected value of (7),

$$\begin{aligned} H_{0,\mathcal{M}} &: \mathbb{E}\left(d_{ij}\right) = 0, \qquad \text{for all } i,j = 1,...,M \\ H_{A,\mathcal{M}} &: \mathbb{E}\left(d_{ij}\right) \neq 0, \qquad \text{for some } i,j = 1,...,M. \end{aligned} \tag{9}$$

At each iteration of the MCS procedure, the hypothesis in (9) is tested for the remaining models.

Hansen et al. (2011) construct the t-statistic,

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\operatorname{var}}\left(\bar{d}_{ij}\right)}} \quad \text{for } i,j \in \mathcal{M}, \tag{10}$$

where $\bar{d}_{ij} = T^{-1}\sum_{t=1}^{T} d_{ij,t}$ is the relative sample loss between model i and model j, and $\widehat{\operatorname{var}}\left(\bar{d}_{ij}\right)$ is the estimated variance of $\bar{d}_{ij}$. Using (10), Hansen et al. (2011) argue that the null hypothesis in (9) maps naturally into the test statistic

$$T_{R,\mathcal{M}} = \max_{i,j\in\mathcal{M}} \left|t_{ij}\right|, \tag{11}$$

which has a non-standard, asymptotic distribution that can be estimated with bootstrap methods (see Kilian 1999; White 2000; Hansen 2003, 2005; Clark and McCracken

2005). With the test statistic, $T_{R,\mathcal{M}}$, the model to be eliminated can be identified via the elimination rule

$$e_{R,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \sup_{j\in\mathcal{M}} t_{ij}, \tag{12}$$

because the model, $e_{R,\mathcal{M}}$, is such that for some $j \in \mathcal{M}$, $t_{e_{R,\mathcal{M}},j} = T_{R,\mathcal{M}}$.

The algorithm behind the MCS procedure is summarized in the following:

Algorithm 3: The Model Confidence Set procedure
(Step 1 through 3, page 5 in Bernardi and Catania 2018)

1. Set $\mathcal{M} = \mathcal{M}_0$ (where $\mathcal{M}_0$ is the initial set of models).
2. Test $H_{0,\mathcal{M}} : \mathbb{E}\left(d_{ij}\right) = 0$ for all $i, j \in \mathcal{M}$, given the confidence level α. If $H_{0,\mathcal{M}}$ cannot be rejected, terminate the algorithm and set $\hat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$. If instead $H_{0,\mathcal{M}}$ is rejected, eliminate the worst-performing model from $\mathcal{M}$ according to the elimination rule in (12).
3. Go to step 2 using the reduced set of models.

E Robustness Checks

E.1 Forecasting Comparison Based on RMSE and MAPE

Table A.3 shows in percentage the frequency at which each model achieves the smallest root mean square error (RMSE) on the test set compared to all competing models. Similarly, Table A.4 shows in percentage the frequency at which each model achieves the smallest mean absolute percentage error (MAPE) on the test set compared to all competing models. The results are displayed for the two different age ranges: 59-89 and 20-89. In both tables, darker shadings are used to mark larger percentages.

Table A.3: Lowest RMSE for each training and test set combination

Forecast horizon:	30 years				16 years			
Fitting period:	1936-1986		1961-1986		1950-2000		1975-2000	
Age range:	59-89	20-89	59-89	20-89	59-89	20-89	59-89	20-89
LC	0%	0%	2%	2%	0%	0%	2%	0%
ACF	0%	0%	2%	0%	0%	0%	0%	3%
CBD	3%	3%	0%	3%	0%	20%	2%	12%
APC	0%	18%	3%	5%	0%	2%	2%	2%
RH	0%	0%	0%	2%	0%	0%	2%	0%
M6	0%	0%	0%	0%	0%	0%	0%	0%
M7	0%	0%	0%	0%	0%	0%	0%	3%
Plat (full)	3%	3%	5%	3%	2%	0%	0%	5%
Plat (reduced)	0%	9%	0%	6%	2%	2%	2%	0%
Pure RF	6%	21%	26%	29%	30%	6%	23%	18%
Pure GB	47%	12%	42%	20%	16%	12%	14%	5%
RF/ARIMA	6%	0%	0%	3%	0%	8%	0%	3%
RF/LC	0%	6%	2%	0%	2%	0%	0%	0%
RF/ACF	12%	3%	3%	2%	2%	2%	5%	8%
RF/CBD	0%	0%	0%	2%	0%	0%	0%	0%
RF/APC	0%	6%	3%	2%	0%	6%	5%	0%
RF/RH	0%	0%	0%	3%	0%	0%	0%	0%
RF/M6	0%	0%	0%	2%	0%	0%	0%	0%
RF/M7	0%	0%	0%	0%	0%	0%	0%	0%
RF/Plat (full)	3%	0%	0%	0%	2%	2%	0%	6%
RF/Plat (reduced)	3%	0%	0%	2%	0%	4%	0%	6%
GB/ARIMA	3%	0%	0%	3%	4%	8%	2%	3%
GB/LC	6%	6%	2%	2%	0%	0%	3%	0%
GB/ACF	6%	6%	11%	6%	6%	8%	11%	2%
GB/CBD	0%	0%	2%	2%	0%	0%	8%	0%
GB/APC	3%	3%	0%	3%	26%	2%	18%	14%
GB/RH	0%	3%	0%	0%	4%	6%	0%	0%
GB/M6	0%	0%	0%	0%	0%	4%	5%	3%
GB/M7	0%	3%	0%	2%	0%	2%	0%	2%
GB/Plat (full)	0%	0%	0%	0%	4%	0%	2%	5%
GB/Plat (reduced)	0%	0%	0%	2%	0%	6%	0%	3%
# country-gender combinations	34		66		50		66	

Notes: Within each column, the percentages are calculated as the frequency at which each model achieves the lowest RMSE across all country-gender combinations. Each column adds up to 100%, since only one model can have the lowest RMSE. The larger the percentage, the darker is the shade marking the cell.

Table A.4: Lowest MAPE for each training and test set combination

Forecast horizon:	30 years				16 years			
Fitting period:	1936-1986		1961-1986		1950-2000		1975-2000	
Age range:	59-89	20-89	59-89	20-89	59-89	20-89	59-89	20-89
LC	0%	0%	0%	5%	0%	0%	6%	0%
ACF	0%	3%	0%	0%	0%	0%	0%	2%
CBD	0%	3%	0%	8%	2%	6%	2%	3%
APC	0%	18%	3%	2%	0%	4%	2%	2%
RH	0%	0%	0%	2%	0%	0%	0%	0%
M6	0%	0%	0%	0%	0%	0%	2%	0%
M7	0%	0%	0%	0%	0%	2%	0%	2%
Plat (full)	0%	0%	5%	3%	2%	0%	0%	3%
Plat (reduced)	0%	3%	3%	2%	0%	4%	2%	5%
Pure RF	12%	24%	24%	35%	38%	8%	17%	24%
Pure GB	44%	12%	39%	14%	10%	24%	14%	5%
RF/ARIMA	3%	3%	0%	3%	0%	10%	0%	2%
RF/LC	0%	3%	5%	0%	4%	0%	0%	2%
RF/ACF	12%	3%	5%	0%	2%	0%	3%	2%
RF/CBD	3%	0%	0%	3%	0%	0%	0%	0%
RF/APC	0%	3%	2%	3%	2%	6%	5%	3%
RF/RH	0%	0%	2%	3%	0%	2%	0%	0%
RF/M6	0%	0%	0%	0%	0%	2%	2%	3%
RF/M7	0%	0%	2%	0%	0%	0%	0%	0%
RF/Plat (full)	9%	0%	2%	0%	2%	4%	2%	5%
RF/Plat (reduced)	3%	0%	0%	0%	0%	0%	0%	6%
GB/ARIMA	0%	0%	0%	3%	4%	6%	3%	3%
GB/LC	6%	6%	2%	2%	2%	0%	3%	0%
GB/ACF	9%	3%	8%	8%	4%	2%	12%	2%
GB/CBD	0%	0%	0%	0%	0%	0%	5%	0%
GB/APC	0%	12%	0%	3%	18%	2%	18%	15%
GB/RH	0%	0%	2%	0%	6%	4%	3%	3%
GB/M6	0%	0%	0%	2%	0%	2%	2%	0%
GB/M7	0%	3%	0%	2%	0%	2%	0%	2%
GB/Plat (full)	0%	0%	0%	0%	4%	2%	2%	8%
GB/Plat (reduced)	0%	3%	0%	2%	0%	8%	0%	3%
# country-gender combinations	34		66		50		66	

Notes: Within each column, the percentages are calculated as the frequency at which each model achieves the lowest MAPE across all country-gender combinations. Each column adds up to 100%, since only one model can have the lowest MAPE. The larger the percentage, the darker is the shade marking the cell.

E.2 Head-to-Head Forecasting Comparisons

This appendix provides additional details about the head-to-head comparison in Section 5.2. The head-to-head comparisons are based on the fitting period 1961-1986 with a forecast horizon of 30 years (both age ranges). For the Lee-Carter comparison, we also include the Levantesi and Pizzorusso (2019) random forests and gradient boosting improved Lee-Carter models. The procedure for estimating these models is similar to our approach with respect to the fitting part (although using a different transformation) but differs with respect to the forecasting part. In particular, the transformation step (step 2 of Procedure 1 in Section 3.2) is replaced with

2. Construct $\psi_{x,t}^{LC} = \frac{D_{x,t}}{\hat{D}_{x,t}^{LC}}$, where $D_{x,t}$ is the actual number of deaths, and $\hat{D}_{x,t}^{LC}$ is the estimated number of deaths according to the LC model.

Next, they fit $\psi_{x,t}^{LC}$ using random forests or gradient boosting (similarly to step 3 of Procedure 1) and obtain the fitted values of the random forests/gradient boosting estimator, $\psi_{x,t}^{LC,ML}$ where ML refers to either RF or GB. Forecasting of the random forests/gradient boosting estimator is based on the original LC framework, as opposed to the random forests/gradient boosting framework. Thus, step 4 of Procedure 1 is replaced with

4. Fit and forecast $\ln \psi_{x,t}^{LC,ML}$ using the LC framework, resulting in the following LC model improved by random forests/gradient boosting:

$$\widehat{\ln m}_{x,t}^{LC,ML} = \widehat{\ln m}_{x,t}^{LC} + \ln \psi_{x,t}^{LC,ML} = \left(\alpha_x^{LC} + \alpha_x^{\psi}\right) + \beta_x^{LC}\kappa_t^{LC} + \beta_x^{\psi}\kappa_t^{\psi}, \tag{13}$$

where both κ_t^{LC} and κ_t^{ψ} are forecast using random walks with drift.

These models are denoted "RF/LC (ψ)" or "GB/LC (ψ)" in Table 3 in Section 5.2 and in Table A.5.

Table A.5: RMSE for 30-year Lee-Carter forecasts with fitting period 1961-1986 for a selection of countries

Age range: 59-89	Italy		France		Denmark		USA	
	Female	Male	Female	Male	Female	Male	Female	Male
LC	**0.1554**	**0.3463**	**0.1269**	0.2172	0.1999	**0.3229**	0.0899	**0.1623**
RF/LC (r)	0.1593	0.3502	0.1289	0.2212	**0.1993**	0.3303	**0.0882**	0.1660
GB/LC (r)	0.1577	0.3489	0.1278	**0.2203**	0.2000	0.3347	0.0899	0.1626
RF/LC (ψ)	0.1566	0.3474	0.1274	0.2179	0.2034	0.3249	0.0898	0.1624
GB/LC (ψ)	0.1565	0.3474	0.1276	0.2182	0.2053	0.3259	0.0897	0.1625
Age range: 20-89	Italy		France		Denmark		USA	
	Female	Male	Female	Male	Female	Male	Female	Male
LC	0.1563	0.3134	0.1745	0.2942	0.3966	0.4154	**0.1682**	0.1767
RF/LC (r)	0.1644	**0.3105**	0.1659	0.2774	0.3768	0.3937	0.1848	0.1671
GB/LC (r)	**0.1536**	0.3117	**0.1636**	**0.2767**	**0.3594**	**0.3790**	0.1880	**0.1668**
RF/LC (ψ)	0.1559	0.3129	0.1688	0.2889	0.3938	0.4099	0.1698	0.1748
GB/LC (ψ)	0.1539	0.3128	0.1729	0.2916	0.3974	0.4144	0.1701	0.1748

Notes: Boldface indicates lowest RMSE within a column.

E.3 Forecasting Comparison using the `distRforest` Package for Random Forests

In this appendix, we produce forecasting results when the random forests algorithm used for estimating the RF variants of the stochastic mortality models is based on the `R` package `distRforest` (see Henckaerts 2019). The `distRforest` package is an extension of the `rpart` package (see Therneau and Atkinson 2019) that implements random forests with distribution-based loss functions. In particular, the `distRforest` package allows for a random forests implementation on count data using the Poisson distribution. The results are produced for the 30-year forecast with fitting period 1961-1986. We consider both age ranges (59-89 and 20-89).

The procedure differs slightly from Procedure 1 in Section 3.2. In particular, the number of deaths are modeled as in Deprez et al. (2017), i.e.

$$D(x,t) \sim Poisson\left(E(x,t) \cdot q(x,t) \cdot q_{RF}(x,t)\right) \tag{14}$$

The procedure for estimating and forecasting this model using the stochastic mortality models presented in Table A.2 is similar to Procedure 1, but replacing step 2 with

2. Construct $\psi_{x,t}^{model} = \frac{D_{x,t}}{\hat{D}_{x,t}^{model}}$, where $D_{x,t}$ is the actual number of deaths, and $\hat{D}_{x,t}^{model}$ is the estimated number of deaths according to some stochastic mortality model.

while step 4 is replaced with

4. Obtain the random forests forecast values $\hat{\psi}_{x,t}^{model,RF}$ and construct the random forests improved forecasts $\widehat{\ln m}_{x,t+h}^{model,RF} = \ln\left[\widehat{m}_{x,t+h}^{model} \cdot \hat{\psi}_{x,t+h}^{model,RF}\right]$

Table A.6: Lowest RMSE and superior set of models for 30-year forecast with fitting period 1961-1986 using `distRforest` for random forest models

	Lowest RMSE		Superior set of models	
Age range:	59-89	20-89	59-89	20-89
LC	2%	2%	3%	5%
ACF	2%	0%	3%	6%
CBD	0%	3%	0%	14%
APC	0%	2%	0%	2%
RH	0%	2%	0%	3%
M6	0%	0%	0%	0%
M7	0%	0%	0%	2%
Plat (full)	5%	0%	6%	2%
Plat (reduced)	0%	8%	2%	8%
Pure RF	26%	29%	27%	35%
Pure GB	35%	20%	39%	24%
RF/ARIMA	0%	3%	2%	6%
RF/LC	2%	0%	3%	0%
RF/ACF	9%	3%	15%	5%
RF/CBD	0%	2%	2%	3%
RF/APC	3%	3%	6%	3%
RF/RH	0%	2%	0%	2%
RF/M6	0%	2%	2%	2%
RF/M7	0%	0%	0%	0%
RF/Plat (full)	0%	3%	3%	6%
RF/Plat (reduced)	0%	0%	2%	2%
GB/ARIMA	0%	3%	2%	5%
GB/LC	2%	2%	2%	3%
GB/ACF	12%	5%	12%	9%
GB/CBD	2%	2%	2%	5%
GB/APC	3%	5%	5%	9%
GB/RH	0%	2%	0%	5%
GB/M6	0%	0%	2%	0%
GB/M7	0%	2%	0%	2%
GB/Plat (full)	0%	0%	3%	3%
GB/Plat (reduced)	0%	2%	2%	3%

Notes: Within each column, the percentages are calculated as the frequency at which each model achieves the lowest RMSE (columns 2-3) or is part of the SSM (columns 4-5) across all country-gender combinations. The larger the percentage, the darker is the shade marking the cell.

Additionally, the random forests algorithm uses the Poisson deviance (rather than MSE) when making variable split decisions. Table A.6 shows results based on the two performance measures: RMSE (columns 2-3) and MCS (columns 4-5). Comparing these results to the original results in Table A.3, columns 4-5 in Appendix E.1 and Table 2, columns 4-5 in Section 5 reveals that using the Poisson deviance and the procedure described above does not change the results significantly.

E.4 Forecasting Comparison when Including/Accounting for Mortality Shocks

In this appendix, we compare mortality forecasts that accounts for mortality shocks in the estimation phase. The results were produced for the 30-year forecast horizon with fitting period 1936-1986 for both age ranges. For the random forests and gradient

Table A.7: Lowest RMSE and superior set of models for 30-year forecast with fitting period 1936-1986 when accounting for mortality shocks

	Lowest RMSE		Superior set of models	
Age range:	59-89	20-89	59-89	20-89
LC	0%	3%	6%	6%
ACF	0%	0%	6%	3%
CBD	6%	12%	6%	18%
APC	3%	6%	3%	9%
RH	0%	0%	0%	6%
M6	0%	0%	0%	0%
M7	0%	0%	0%	0%
Plat (full)	0%	3%	0%	3%
Plat (reduced)	0%	12%	0%	15%
Pure RF	12%	26%	15%	35%
Pure GB	35%	12%	35%	18%
RF/ARIMA	6%	3%	6%	6%
RF/LC	3%	6%	6%	6%
RF/ACF	18%	0%	18%	0%
RF/CBD	0%	3%	0%	6%
RF/APC	0%	0%	0%	3%
RF/RH	0%	0%	0%	3%
RF/M6	0%	0%	0%	3%
RF/M7	0%	0%	0%	0%
RF/Plat (full)	0%	0%	3%	3%
RF/Plat (reduced)	0%	0%	0%	0%
GB/ARIMA	0%	0%	0%	0%
GB/LC	0%	0%	3%	3%
GB/ACF	12%	6%	15%	9%
GB/CBD	3%	0%	3%	0%
GB/APC	0%	9%	0%	15%
GB/RH	3%	0%	3%	6%
GB/M6	0%	0%	0%	3%
GB/M7	0%	0%	0%	0%
GB/Plat (full)	0%	0%	0%	3%
GB/Plat (reduced)	0%	0%	0%	6%

Notes: Within each column, the percentages are calculated as the frequency at which each model achieves the lowest RMSE (columns 2-3) or is part of the SSM (columns 4-5) across all country-gender combinations. The larger the percentage, the darker is the shade marking the cell.

boosting models, we include three mortality shock dummies in the set of features corresponding to World War II (1939-1945), the Asian Flu (1957-1958), and the Hong Kong Flu (1968-1969). The stochastic mortality models were fit for the entire period, but the refitting of the time components (κ_t-s) was based on 1972-1986 (15 years), thereby avoiding any of the mortality shocks mentioned above when refitting κ_t. Table A.7 presents the results based on the two performance measures: RMSE (columns 2-3) and MCS (columns 4-5). Comparing these results to the original results in Table A.3, columns 4-5 in Appendix E.1 and Table 2, columns 4-5 in Section 5 reveals that including/accounting for mortality shocks does not change the results significantly.

E.5 Forecasting Comparison when Excluding the *Cohort* Variable for RF and GB Estimation

In this appendix, we compare mortality forecasts when excluding *cohort* from the set of features used for estimating and forecasting by random forests and gradient boosting. The results are produced for the 30-year forecast with fitting period 1961-1986. We consider both age ranges (59-89 and 20-89). Table A.8 presents the results based on the two performance measures: RMSE (columns 2-3) and MCS (columns 4-5). Comparing these results to the original results in Table A.3, columns 4-5 in Appendix E.1 and Table 2, columns 4-5 in Section 5 reveals that excluding the *cohort* variable does not change the results significantly.

Table A.8: Lowest RMSE and superior set of models for 30-year forecast with fitting period 1961-1986 without the cohort variable for RF and GB

	Lowest RMSE		Superior set of models	
Age range:	59-89	20-89	59-89	20-89
LC	0%	3%	5%	3%
ACF	2%	0%	2%	3%
CBD	0%	3%	2%	11%
APC	0%	5%	2%	5%
RH	0%	2%	0%	5%
M6	0%	2%	2%	2%
M7	0%	0%	2%	0%
Plat (full)	5%	3%	6%	6%
Plat (reduced)	0%	5%	2%	8%
Pure RF	23%	26%	27%	35%
Pure GB	44%	24%	45%	27%
RF/ARIMA	0%	3%	2%	6%
RF/LC	0%	0%	2%	2%
RF/ACF	5%	5%	11%	11%
RF/CBD	0%	2%	2%	3%
RF/APC	0%	2%	0%	2%
RF/RH	0%	2%	0%	5%
RF/M6	0%	0%	2%	2%
RF/M7	2%	0%	3%	0%
RF/Plat (full)	2%	0%	2%	2%
RF/Plat (reduced)	0%	3%	2%	6%
GB/ARIMA	0%	3%	2%	6%
GB/LC	5%	0%	6%	2%
GB/ACF	8%	3%	11%	6%
GB/CBD	0%	0%	2%	5%
GB/APC	6%	3%	8%	5%
GB/RH	0%	2%	0%	2%
GB/M6	0%	0%	0%	0%
GB/M7	2%	2%	3%	2%
GB/Plat (full)	0%	0%	3%	2%
GB/Plat (reduced)	0%	2%	2%	5%

Notes: Within each column, the percentages are calculated as the frequency at which each model achieves the lowest RMSE (columns 2-3) or is part of the SSM (columns 4-5) across all country-gender combinations. The larger the percentage, the darker is the shade marking the cell.

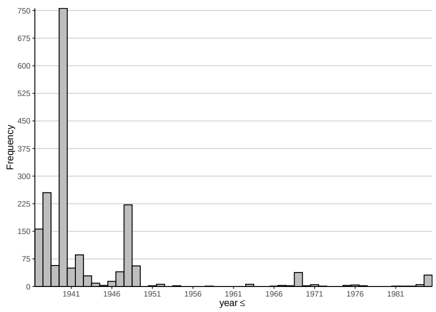
F Random Forests Additional Results

F.1 50 Most Frequent Country Combinations

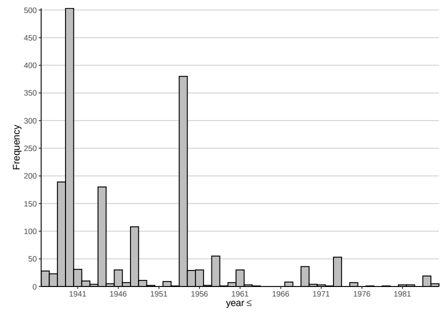
Table A.9: 50 most frequent 4-way groupings of countries for random forests. Fitting period: 1961-1986, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	CZE	HUN	159	742	23%
BGR	BLR	CZE	LTU	135	699	22%
BGR	BLR	CZE	LVA	137	715	22%
BGR	BLR	CZE	SVK	153	742	23%
BGR	BLR	HUN	LTU	195	763	24%
BGR	BLR	HUN	LVA	185	766	24%
BGR	BLR	HUN	POL	112	654	20%
BGR	BLR	HUN	SVK	236	827	26%
BGR	BLR	LTU	LVA	155	723	23%
BGR	BLR	LTU	SVK	192	765	24%
BGR	BLR	LVA	SVK	187	775	24%
BGR	CZE	HUN	LTU	142	712	22%
BGR	CZE	HUN	LVA	146	736	23%
BGR	CZE	HUN	POL	120	693	22%
BGR	CZE	HUN	SVK	166	780	24%
BGR	CZE	LTU	LVA	123	687	21%
BGR	CZE	LTU	SVK	132	709	22%
BGR	CZE	LVA	POL	116	670	21%
BGR	CZE	LVA	SVK	142	740	23%
BGR	CZE	POL	SVK	114	692	22%
BGR	HUN	LTU	LVA	155	722	23%
BGR	HUN	LTU	SVK	194	772	24%
BGR	HUN	LVA	POL	118	670	21%
BGR	HUN	LVA	SVK	192	791	25%
BGR	HUN	POL	SVK	123	697	22%
BGR	LTU	LVA	SVK	151	725	23%
BGR	LVA	POL	SVK	115	672	21%
BLR	CZE	HUN	LTU	135	698	22%
BLR	CZE	HUN	LVA	140	716	22%
BLR	CZE	HUN	SVK	153	739	23%
BLR	CZE	LTU	LVA	122	681	21%
BLR	CZE	LTU	SVK	127	693	22%
BLR	CZE	LVA	SVK	135	719	22%
BLR	HUN	LTU	LVA	150	713	22%
BLR	HUN	LTU	SVK	185	757	24%
BLR	HUN	LVA	POL	112	648	20%
BLR	HUN	LVA	SVK	183	770	24%
BLR	LTU	LVA	SVK	148	716	22%
CZE	DEUTE	HUN	LVA	112	684	21%
CZE	DEUTE	LTU	POL	112	650	20%
CZE	DEUTE	LVA	POL	118	696	22%
CZE	HUN	LTU	LVA	133	694	22%
CZE	HUN	LTU	POL	112	644	20%
CZE	HUN	LTU	SVK	134	705	22%
CZE	HUN	LVA	POL	122	678	21%
CZE	HUN	LVA	SVK	144	739	23%
CZE	HUN	POL	SVK	116	699	22%
CZE	LTU	LVA	SVK	122	692	22%
CZE	LVA	POL	SVK	113	677	21%
HUN	LTU	LVA	SVK	153	723	23%
HUN	LVA	POL	SVK	118	674	21%

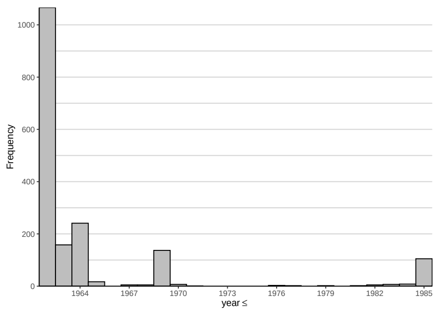
F.2 All Fitting Periods and Both Age Ranges



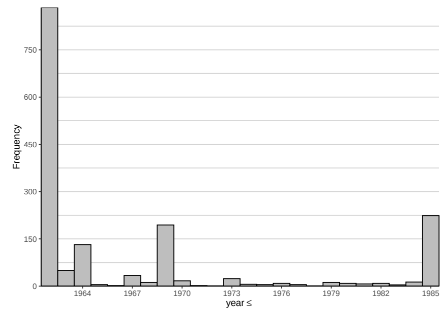
(a) Age range: 59-89, Fitting period: 1936-1986



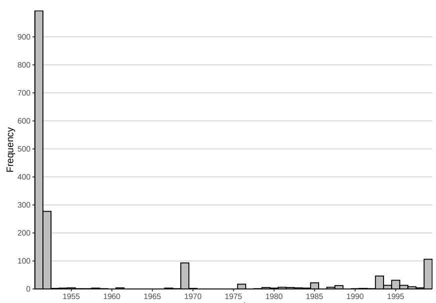
(b) Age range: 20-89, Fitting period: 1936-1986



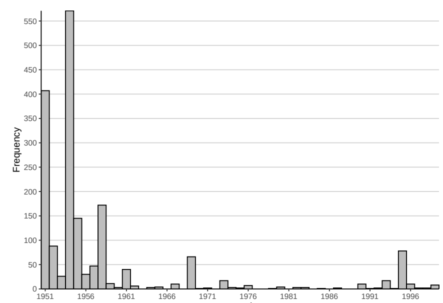
(c) Age range: 59-89, Fitting period: 1961-1986



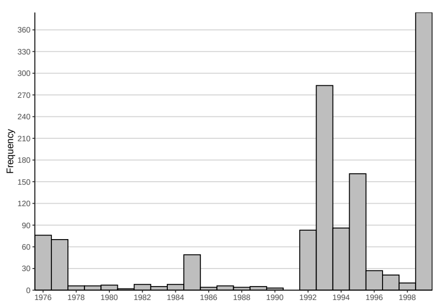
(d) Age range: 20-89, Fitting period: 1961-1986



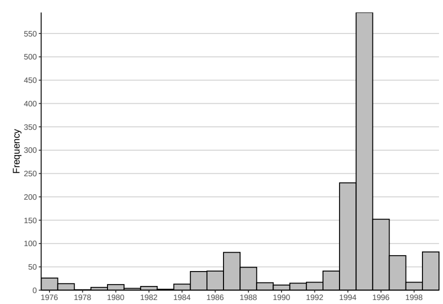
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000

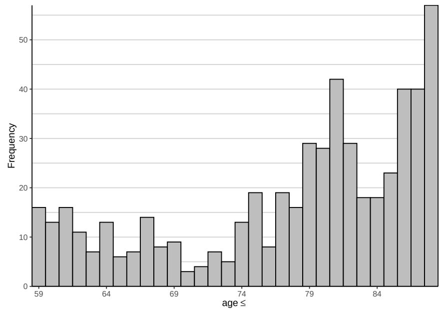


(g) Age range: 59-89, Fitting period: 1975-2000

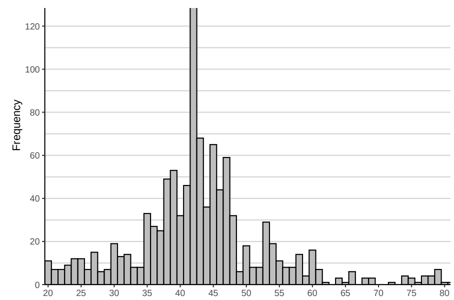


(h) Age range: 20-89, Fitting period: 1975-2000

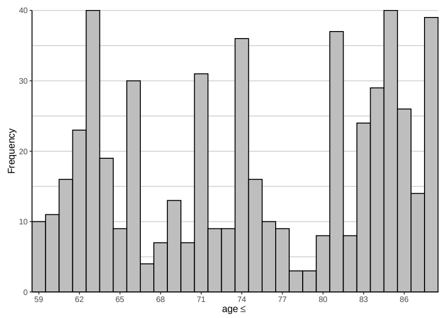
Figure A.1: Distribution of *year* split points for random forests



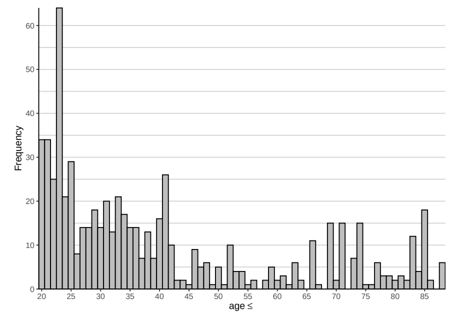
(a) Age range: 59-89, Fitting period: 1936-1986



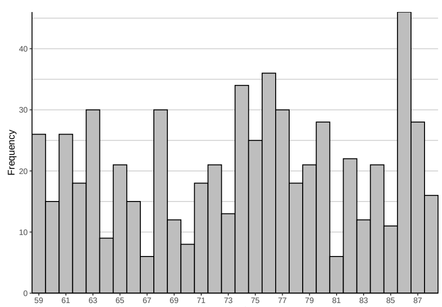
(b) Age range: 20-89, Fitting period: 1936-1986



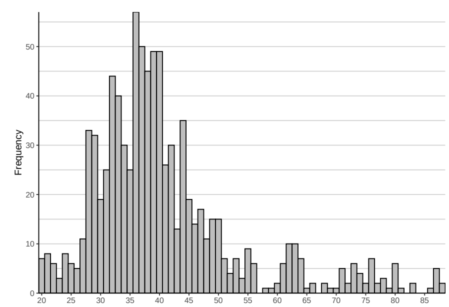
(c) Age range: 59-89, Fitting period: 1961-1986



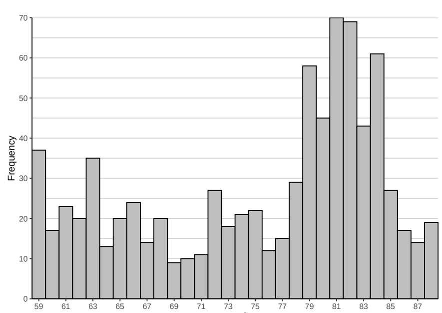
(d) Age range: 20-89, Fitting period: 1961-1986



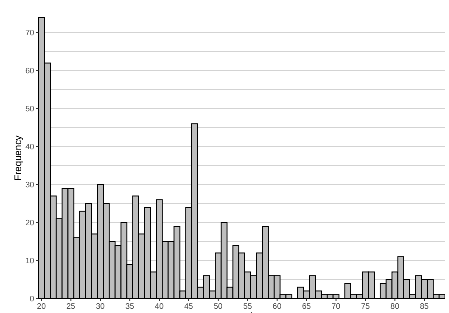
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000

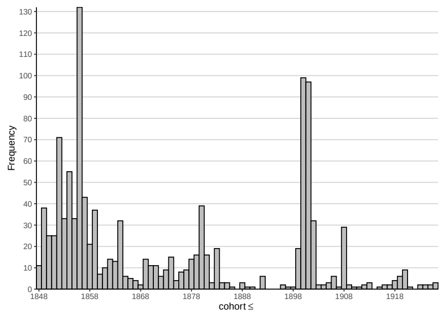


(g) Age range: 59-89, Fitting period: 1975-2000

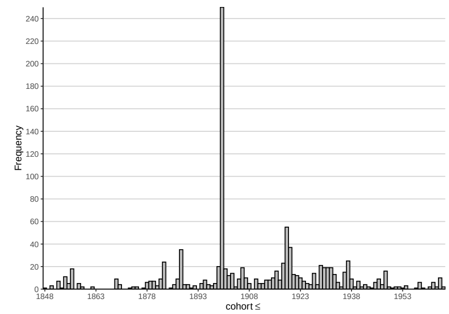


(h) Age range: 20-89, Fitting period: 1975-2000

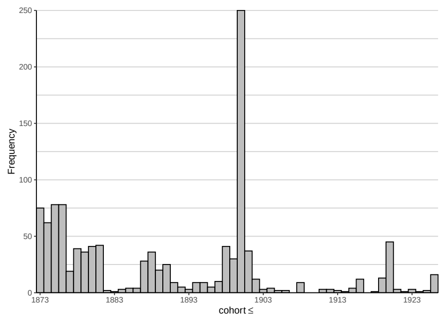
Figure A.2: Distribution of *age* split points for random forests



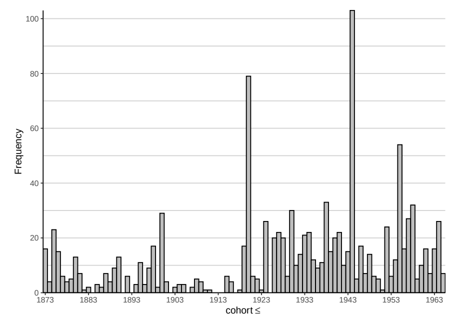
(a) Age range: 59-89, Fitting period: 1936-1986



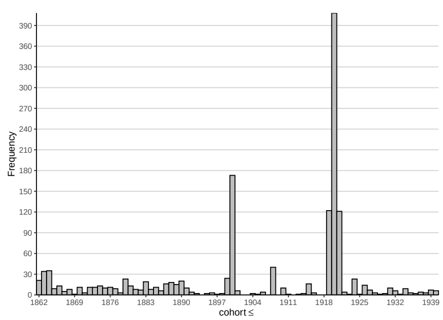
(b) Age range: 20-89, Fitting period: 1936-1986



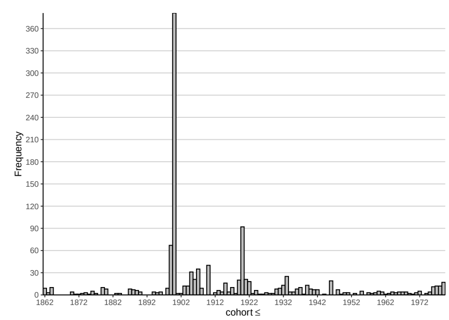
(c) Age range: 59-89, Fitting period: 1961-1986



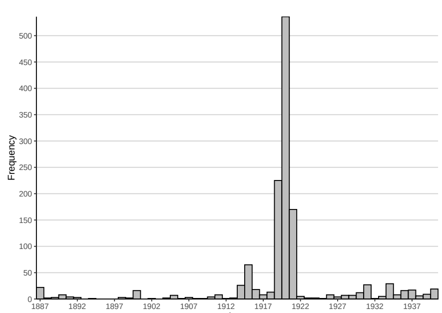
(d) Age range: 20-89, Fitting period: 1961-1986



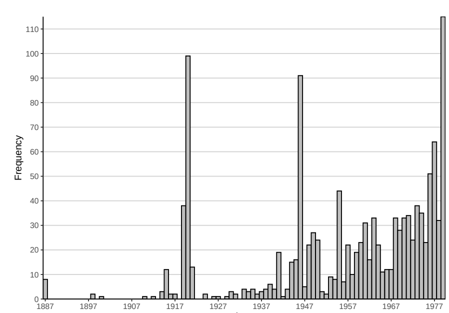
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000



(g) Age range: 59-89, Fitting period: 1975-2000



(h) Age range: 20-89, Fitting period: 1975-2000

Figure A.3: Distribution of *cohort* split points for random forests

Table A.10: Most frequent 4-way groupings of countries for random forests. Fitting period: 1936-1986, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
CHE	ESP	FIN	FRATNP	226	345	11%
CHE	ESP	FIN	ITA	159	343	11%
CHE	ESP	FRATNP	ITA	210	344	11%
CHE	FIN	FRATNP	ITA	160	362	11%
DNK	GBRNIR	GBRSCO	GBRTENW	159	474	15%
DNK	GBRNIR	GBRSCO	NOR	177	562	18%
DNK	GBRSCO	GBRTENW	NOR	159	485	15%

Table A.11: Most frequent 4-way groupings of countries for random forests. Fitting period: 1950-2000, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	DNK	HUN	NLD	221	458	14%
BGR	DNK	HUN	SVK	244	480	15%
BGR	DNK	NLD	SVK	224	467	15%
BGR	HUN	NLD	SVK	230	474	15%
DNK	HUN	NLD	SVK	232	497	16%

Table A.12: Most frequent 4-way groupings of countries for random forests. Fitting period: 1975-2000, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	HUN	POL	SVK	227	605	19%
DNK	EST	HUN	POL	210	611	19%
DNK	EST	LTU	LVA	213	668	21%
DNK	HUN	LVA	POL	201	613	19%
DNK	HUN	POL	SVK	224	630	20%

Table A.13: Most frequent 4-way groupings of countries for random forests. Fitting period: 1936-1986, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
AUS	CAN	DNK	USA	141	448	14%
BEL	FIN	FRATNP	GBRTENW	161	326	10%
BEL	FIN	FRATNP	NLD	143	266	8%
BEL	FRATNP	GBRTENW	NLD	168	322	10%
CAN	DNK	SWE	USA	139	450	14%

Table A.14: Most frequent 4-way groupings of countries for random forests. Fitting period: 1961-1986, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	HUN	LTU	112	835	26%
BGR	BLR	HUN	SVK	112	853	27%
BGR	CZE	HUN	SVK	96	862	27%
BGR	HUN	LTU	SVK	94	805	25%
BLR	HUN	LTU	SVK	104	818	26%

Table A.15: Most frequent 4-way groupings of countries for random forests. Fitting period: 1950-2000, age range: 20-89

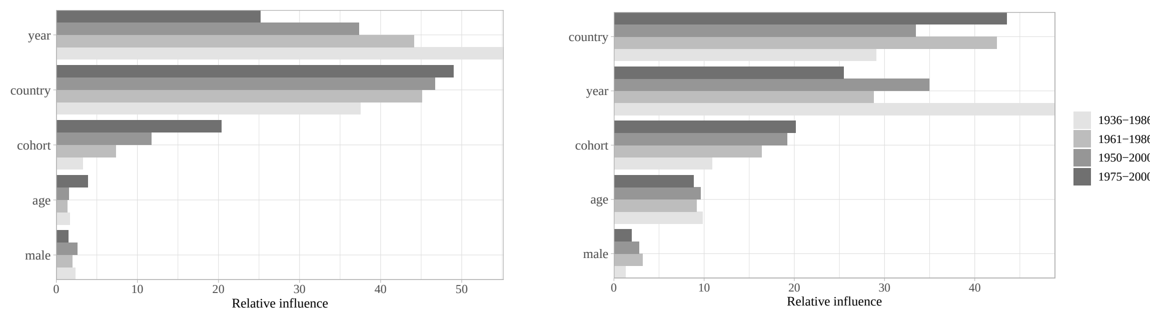
4-way grouping				10-group freq.	Total freq.	%total
BGR	DNK	HUN	NLD	206	526	16%
BGR	DNK	HUN	NOR	190	500	16%
BGR	DNK	HUN	SVK	209	518	16%
BGR	HUN	NLD	SVK	190	508	16%
DNK	HUN	NLD	SVK	193	540	17%

Table A.16: Most frequent 4-way groupings of countries for random forests. Fitting period: 1975-2000, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	EST	LVA	284	494	15%
BGR	BLR	LTU	LVA	287	521	16%
BGR	EST	HUN	LVA	283	491	15%
BLR	EST	LTU	LVA	321	537	17%
EST	HUN	LTU	LVA	310	609	19%

G Gradient Boosting Results: Opening the Box

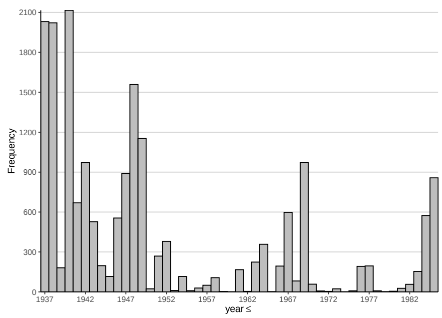
Figure A.4 plots the relative influence (see Friedman 2001) of each variable in the gradient boosting model. The relative influence is provided by the `gbm` package in `R`. The gradient boosting settings used to fit the gradient boosting model (6,000 trees and maximum tree depth of 4) resulted in a total of 30,000 terminal conditions to be analyzed.



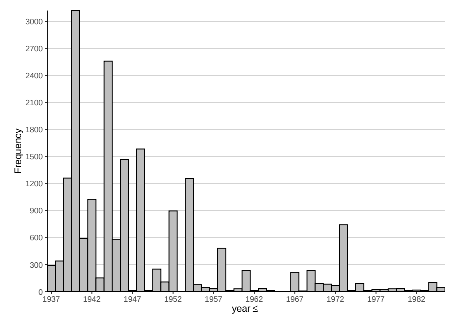
(a) Age range: 59-89

(b) Age range: 20-89

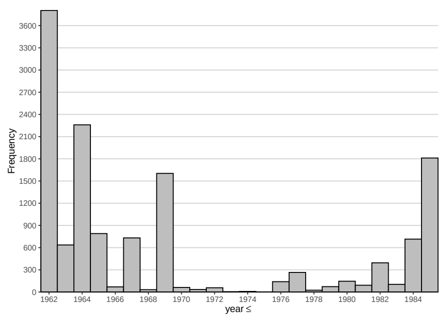
Figure A.4: Relative influence of each variable in the gradient boosting model for all fitting periods and both age ranges



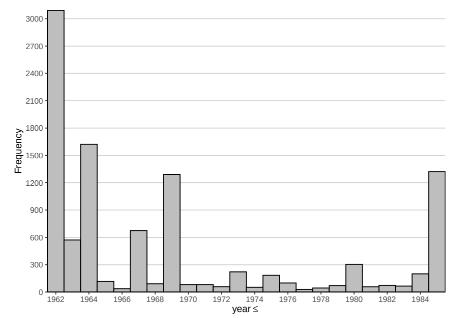
(a) Age range: 59-89, Fitting period: 1936-1986



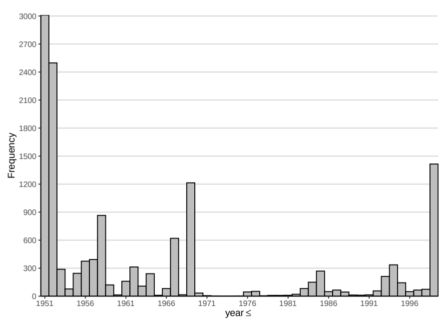
(b) Age range: 20-89, Fitting period: 1936-1986



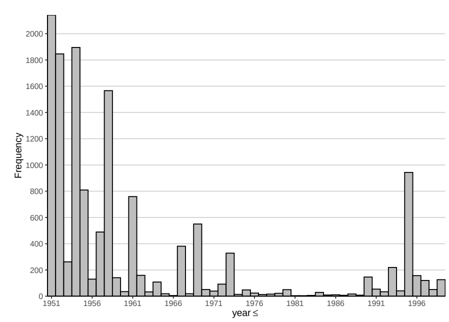
(c) Age range: 59-89, Fitting period: 1961-1986



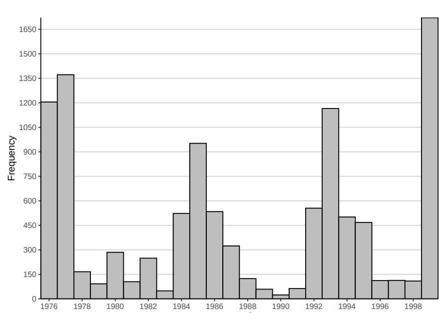
(d) Age range: 20-89, Fitting period: 1961-1986



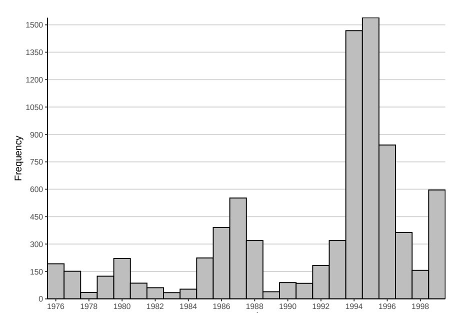
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000

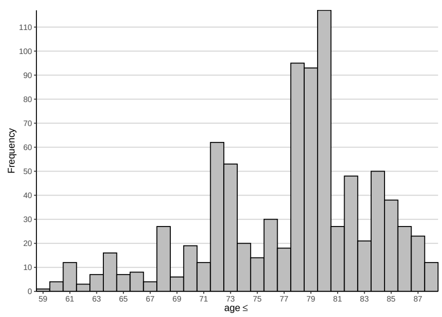


(g) Age range: 59-89, Fitting period: 1975-2000

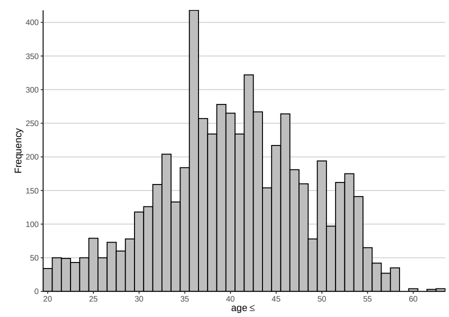


(h) Age range: 20-89, Fitting period: 1975-2000

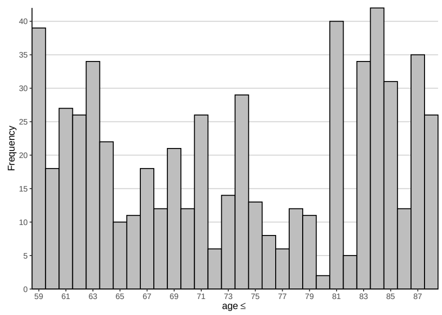
Figure A.5: Distribution of *year* split points for gradient boosting



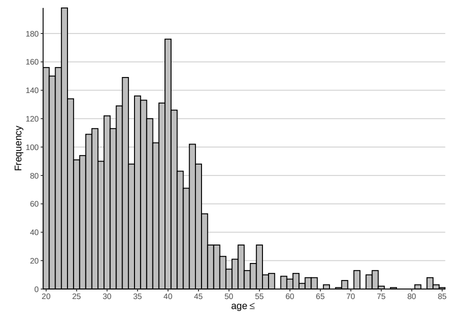
(a) Age range: 59-89, Fitting period: 1936-1986



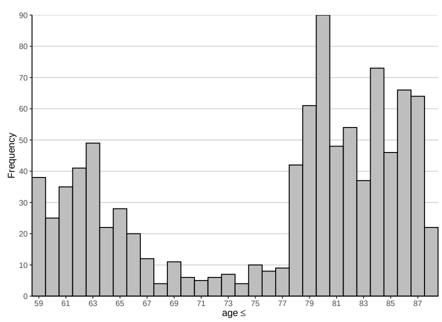
(b) Age range: 20-89, Fitting period: 1936-1986



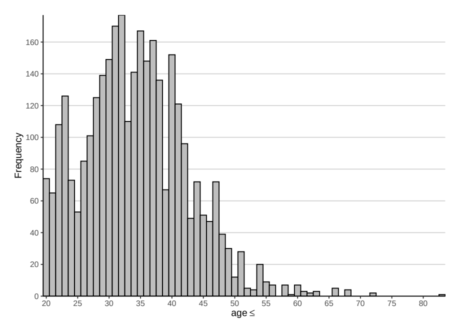
(c) Age range: 59-89, Fitting period: 1961-1986



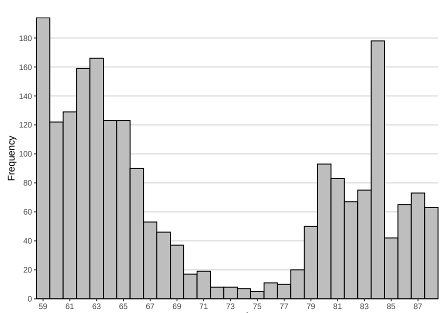
(d) Age range: 20-89, Fitting period: 1961-1986



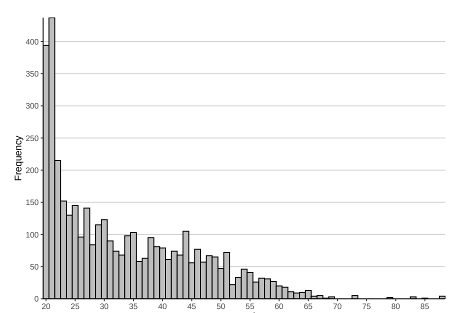
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000

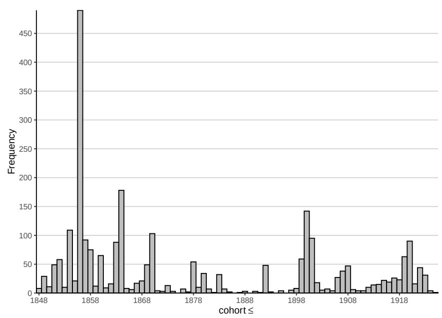


(g) Age range: 59-89, Fitting period: 1975-2000

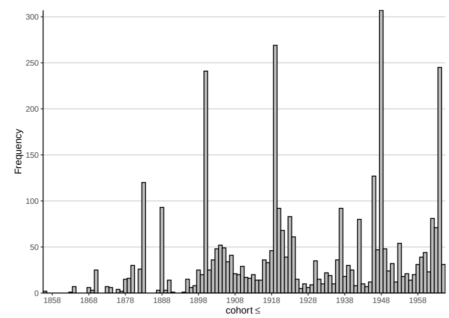


(h) Age range: 20-89, Fitting period: 1975-2000

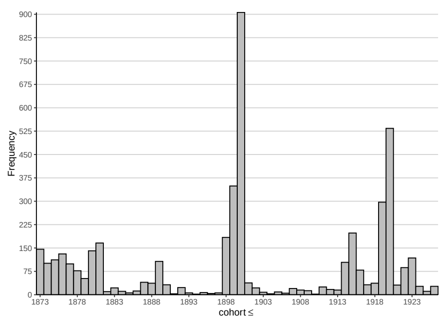
Figure A.6: Distribution of *age* split points for gradient boosting



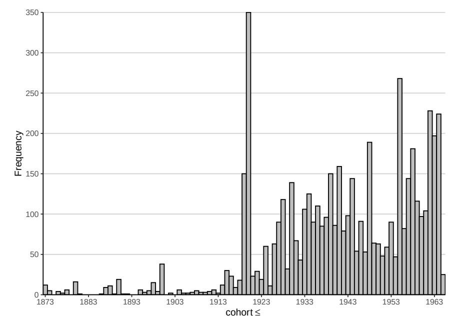
(a) Age range: 59-89, Fitting period: 1936-1986



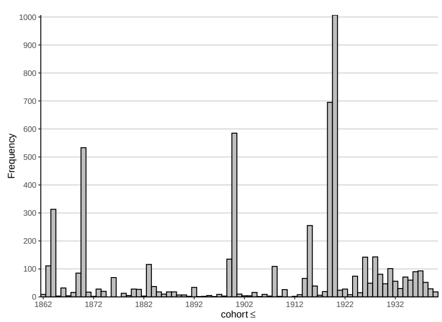
(b) Age range: 20-89, Fitting period: 1936-1986



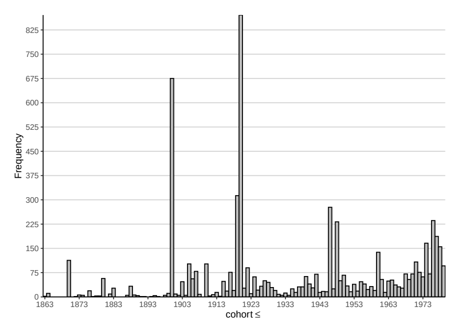
(c) Age range: 59-89, Fitting period: 1961-1986



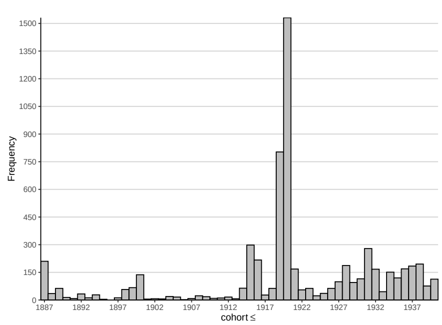
(d) Age range: 20-89, Fitting period: 1961-1986



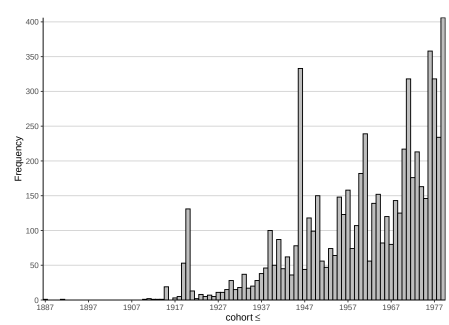
(e) Age range: 59-89, Fitting period: 1950-2000



(f) Age range: 20-89, Fitting period: 1950-2000



(g) Age range: 59-89, Fitting period: 1975-2000



(h) Age range: 20-89, Fitting period: 1975-2000

Figure A.7: Distribution of *cohort* split points for gradient boosting

Table A.17: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1936-1986, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
DNK	GBRNIR	GBRSCO	GBRTENW	1426.00	4492.00	15%
DNK	GBRNIR	GBRSCO	NOR	1114.00	4272.00	14%
DNK	GBRSCO	GBRTENW	NOR	1240.00	4573.00	15%
GBRNIR	GBRSCO	GBRTENW	NLD	1180.00	4128.00	14%
GBRNIR	GBRSCO	GBRTENW	NOR	1258.00	4452.00	15%

Table A.18: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1961-1986, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	EST	LVA	990.00	5597.00	19%
BGR	BLR	HUN	SVK	1067.00	6164.00	21%
BGR	BLR	LTU	LVA	1262.00	6296.00	21%
BGR	BLR	LTU	SVK	979.00	5713.00	19%
BLR	EST	LTU	LVA	1022.00	5545.00	18%

Table A.19: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1950-2000, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	CZE	HUN	SVK	538.00	3923.00	13%
BGR	DNK	HUN	SVK	546.00	3491.00	12%
BGR	HUN	IRL	SVK	511.00	3050.00	10%
BGR	HUN	ISL	SVK	689.00	1947.00	6%
ESP	GBRNIR	GBRTENW	IRL	626.00	3914.00	13%

Table A.20: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1975-2000, age range: 59-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	EST	LTU	732.00	4151.00	14%
BGR	BLR	EST	LVA	664.00	4061.00	14%
BGR	BLR	LTU	LVA	794.00	4658.00	16%
BLR	EST	LTU	LVA	891.00	4800.00	16%
BLR	ISL	LTU	LVA	663.00	2512.00	8%

Table A.21: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1936-1986, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BEL	FIN	FRATNP	GBRTENW	1673.00	3106.00	10%
BEL	FIN	FRATNP	NLD	1838.00	2928.00	10%
BEL	FIN	GBRTENW	NLD	1462.00	2617.00	9%
BEL	FRATNP	GBRTENW	NLD	1509.00	3118.00	10%
FIN	FRATNP	GBRTENW	NLD	1579.00	2691.00	9%

Table A.22: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1961-1986, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	HUN	LTU	644.00	5617.00	19%
BLR	EST	LTU	LUX	675.00	2383.00	8%
BLR	EST	LTU	LVA	1270.00	5048.00	17%
BLR	EST	LUX	LVA	676.00	2397.00	8%
BLR	LTU	LUX	LVA	752.00	2862.00	10%

Table A.23: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1950-2000, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	DNK	HUN	SVK	619.00	3614.00	12%
BGR	HUN	IRL	SVK	615.00	3464.00	12%
BGR	HUN	ISL	SVK	794.00	2532.00	8%
ESP	GBRNIR	IRL	ISL	606.00	1945.00	6%
ESP	GBRNIR	IRL	PRT	713.00	3618.00	12%

Table A.24: Most frequent 4-way groupings of countries for gradient boosting. Fitting period: 1975-2000, age range: 20-89

4-way grouping				10-group freq.	Total freq.	%total
BGR	BLR	EST	LVA	1229.00	4408.00	15%
BGR	BLR	HUN	LVA	1155.00	4877.00	16%
BGR	BLR	LTU	LVA	1214.00	4722.00	16%
BLR	EST	LTU	LVA	1476.00	4577.00	15%
EST	HUN	LTU	LVA	1337.00	4779.00	16%

References

Bernardi, M. and Catania, L. (2018). "The Model Confidence Set package for R". In: *International Journal of Computational Economics and Econometrics* 8.2, pp. 144–158. DOI: 10.1504/IJCEE.2018.091037.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). "A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration". In: *Journal of Risk & Insurance* 73.4, pp. 687–718. DOI: 10.1111/j.1539-6975.2006.00195.x.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). "A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States". In: *North American Actuarial Journal* 13.1, pp. 1–35. DOI: 10.1080/10920277.2009.10597538.

Clark, T. E. and McCracken, M. W. (2005). "Evaluating Direct Multi-Step Forecasts". In: *Econometric Reviews* 105.5, pp. 85–110. DOI: 10.1080/07474930500405683.

Currie, I. D. (June 2006). *Smoothing and forecasting mortality rates with P-splines*. Talk given at the Institute of Actuaries. URL: http:www.ma.hw.ac.uk/~iain/research/talks.html (visited on 11/03/2020).

Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). "Machine learning techniques for mortality modeling". In: *European Actuarial Journal* 7.2, pp. 337–352. DOI: 10.1007/s13385-017-0152-4.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451.

— (2002). "Stochastic gradient boosting". In: *Computational Statistics & Data Analysis* 38.4, pp. 367–378. DOI: 10.1016/S0167-9473(01)00065-2.

Haberman, S. and Renshaw, A. E. (2011). "A Comparative Study of Parametric Mortality Projection Models". In: *Insurance: Mathematics and Economics* 48.1, pp. 35–55. DOI: 10.1016/j.insmatheco.2010.09.003.

Hansen, P. R. (2003). *Regression Analysis With Many Specifications: A Bootstrap Method to Robust Inference*. Working Paper. Brown University.

— (2005). "A test for superior predictive ability". In: *Journal of Business & Economic Statistics* 23.4, pp. 365–380. DOI: 10.1198/073500105000000063.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497. DOI: 10.3982/ECTA5771.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition. New York: Springer. DOI: 10.1007/978-0-387-84858-7.

Henckaerts, R. (2019). *distRforest: Distribution-based Random Forest*. URL: https://www.github.com/henckr/distRforest (visited on 05/22/2021).

Kilian, L. (1999). "Exchange rates and monetary fundamentals: What do we learn fromlong-horizon regressions?" In: *Journal of Applied Econometrics* 14.5, pp. 415–510. DOI: 10.1002/(SICI)1099-1255(199909/10)14:5<491::AID-JAE527>3.0.CO;2-D.

Lee, R. and Carter, L. R. (1992). "Modeling and forecasting of U.S. mortality". In: *Journal of the American Statistical Association* 87.419, pp. 659–675. DOI: 10.1080/01621459.1992.10475265.

Levantesi, S. and Pizzorusso, V. (2019). "Application of Machine Learning to Mortality Modeling and Forecasting". In: *Risks* 7.1, p. 26. DOI: 10.3390/risks7010026.

Li, N. and Lee, R. (2005). "Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method". In: *Demography* 42.3, pp. 575–594. DOI: 10.1353/dem.2005.0021.

Plat, R. (2009). "On Stochastic Mortality Modeling". In: *Insurance Mathematics and Economics* 45.3, pp. 393–404. DOI: 10.1016/j.insmatheco.2009.08.006.

Renshaw, A. E. and Haberman, S. (2006). "A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors". In: *Insurance: Mathematics and Economics* 38.3, pp. 556–570. DOI: 10.1016/j.insmatheco.2005.12.001.

Therneau, T. M. and Atkinson, E. J. (2019). "An Introduction to Recursive Partitioning Using the RPART Routines". In:

White, H. (2000). "A reality check for data snooping". In: *Econometrica* 68.5, pp. 1097–1126. DOI: 10.1111/1468-0262.00152.