Ma thèse d'Actuariat en 10 min
Extensions multivariées pour la modélisation de la mortalité

Antoine Burg, CEREMADE Université Paris Dauphine, SCOR

Congrès des Actuaires, 17 juin 2025

The Art & Science of Risk

Mortality is driven by several major causes of death

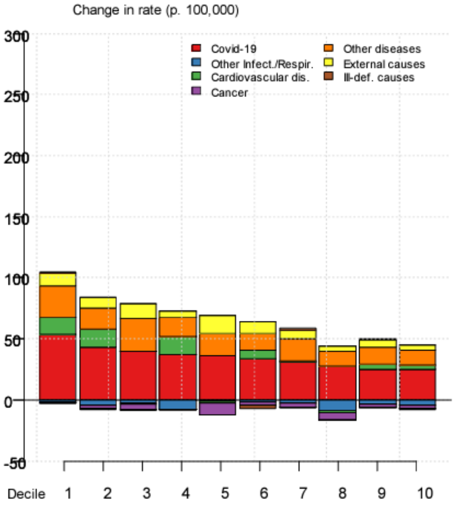


Figure 1: US mortality between 2019 and 2022 by socio-economic groups, Females [Barbieri, 2024].

Outline

1. Closed-form estimators for multivariate regressions models - a single categorical variable approach

2. Alternative estimation framework based on GLM for mortality models: from single population to Cause-of-Death modelling

3. Practical aspects of multivariate LSTM for mortality forecasts

Closed-form estimators for multivariate regressions models - a single categorical variable approach

Motivation: get closed-form formulas for repeated GLM fits for model selection or ensemble methods.

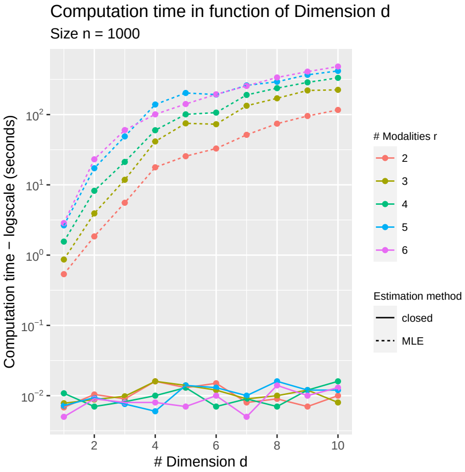
Assumptions:

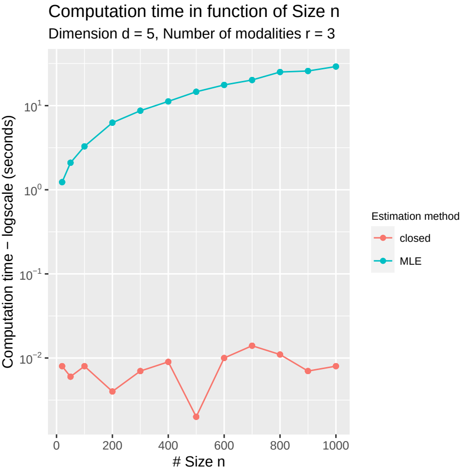
- Multivariate GLM framework
- Multinomial or Dirichlet distribution for the response variable
- Explanatory variables consist of a single categorical variable with a dummy structure

Main results:

- Closed-form formulas for MGLM regression parameters in special case of categorical variables for both Multinomial and Dirichlet distributions
- Faster computation time, especially in high dimensions
- New class of estimators for Dirichlet distribution with asymptotic properties

Computation time on Simulated Data for MGLM regression parameters with multinomial distribution





Alternative estimation framework based on GLM for mortality models: from single population to Cause-of-Death modelling

Motivation: provide a unified framework for single population, multi-population and causes-of-deaths modelling.

Assumptions:

- Death counts $\boldsymbol{D}_{x,t}$ follow a multinomial distribution
- Only mortality models with linear components $\eta_{x,t}^{(j)} = \alpha_x^{(j)} + \sum_i \beta_x^{i,(j)} \kappa_t^{i,(j)} + \beta_x^{(0)} \gamma_{t-x}^{(j)}$

Main results:

- Alternative estimators for mortality models involving single population, multi-population and causes-of-death.
- Explicit estimators
- For multi-populations models:1-step procedure instead of 2-step Li-Lee calibration, with no convergence issue.
- Better performance for causes-of-death modelling.

Multi-populations example

- Countries: France, Italy, Spain
- Age band: 50-84
- Period: 2001-2020
- Sex: Male
- Source: Human Mortality Database
- Model: Li-Lee structure [Li and Lee, 2005] with M3 models and one common factor K_t

$$\begin{cases} \eta_{x,t,g_1} = \alpha_{x,g_1} + K_t + \kappa_{t,g_1} + \gamma_{t-x,g_1} \\ \eta_{x,t,g_2} = \alpha_{x,g_2} + K_t + \kappa_{t,g_2} + \gamma_{t-x,g_2} \\ \eta_{x,t,g_3} = \alpha_{x,g_3} + K_t + \kappa_{t,g_3} + \gamma_{t-x,g_3} \end{cases}$$

Multi-populations example - trend factors

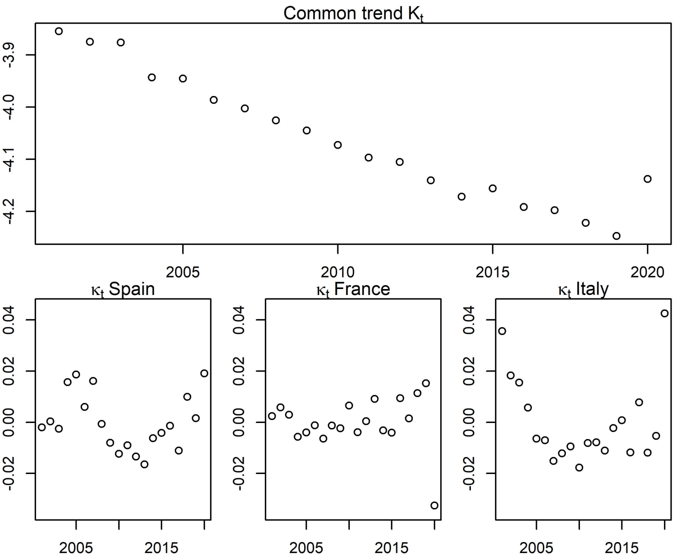


Figure 2: Common and specific trend parameters, Males.

Causes-of-death example

- Country: USA
- Age band: 50-84
- Period: 2001-2020
- Sex: Male
- Source: Human Mortality Database, based on ICD10 classification
- For each of the 16 selected causes, M3 model with cause-specific parameters is assumed

$$\eta_{x,t}^{(j)} = \alpha_x^{(j)} + \kappa_t^{(j)} + \gamma_{t-x}^{(j)}.$$

- Estimators: comparison between univariate independent single population VS multinomial assumption (exact same numbers of parameters).

Table 1: Goodness of fit

Parameters	Deviance	AIC	MSE	MAE	MAPE (%)
Independent univariate	11883	107342	2.55E-07	1.58E-04	7,03%
Multivariate	5149	97172	8.88E-09	2.56E-05	2.14%

State of the art of mortality forecasting

Usual forecasting framework consists in assuming a random walk or another time series model for time-dependent parameters κ_t^i of a GAPC

$$\kappa_t^i = \kappa_{t-1}^i + \nu + \epsilon_t.$$

Deep learning literature provide several architectures to deal with time-series

- Recurrent Neural Networks
- Convolutional Networks
- Generative Adversarial Networks
- Transformers
- Long Short Term Memory (LSTM) Networks

LSTM architecture

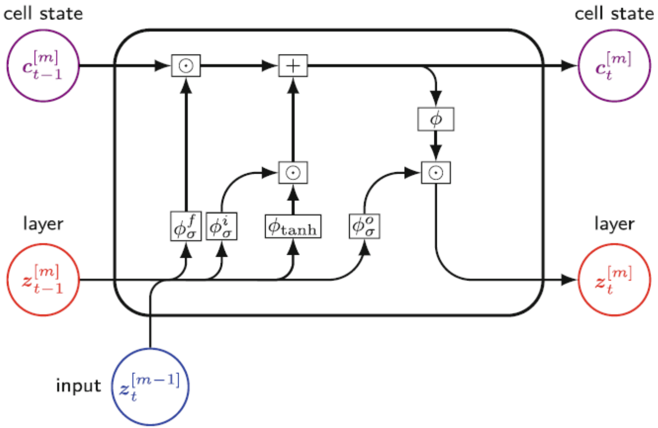


Figure 3: A LSTM cell. Source: [Wüthrich and Merz, 2023].

Historical data, USA, 1970-2022

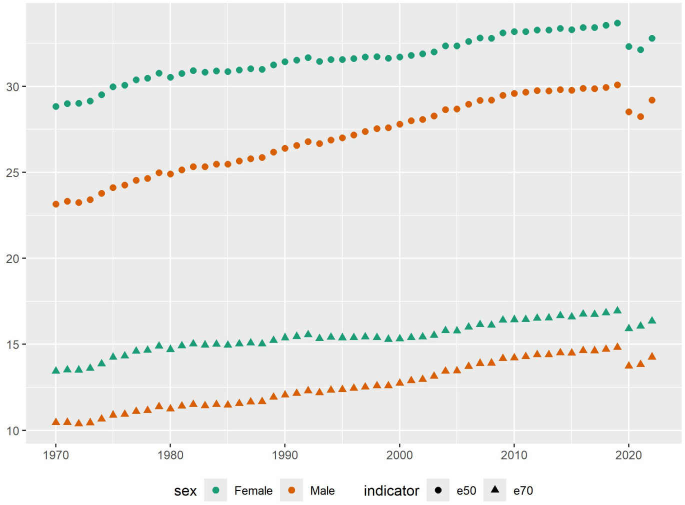


Figure 4: Evolution of residual life expectancy at ages 50 and 70.

Use case: stationary (no mortality shock), USA, Female

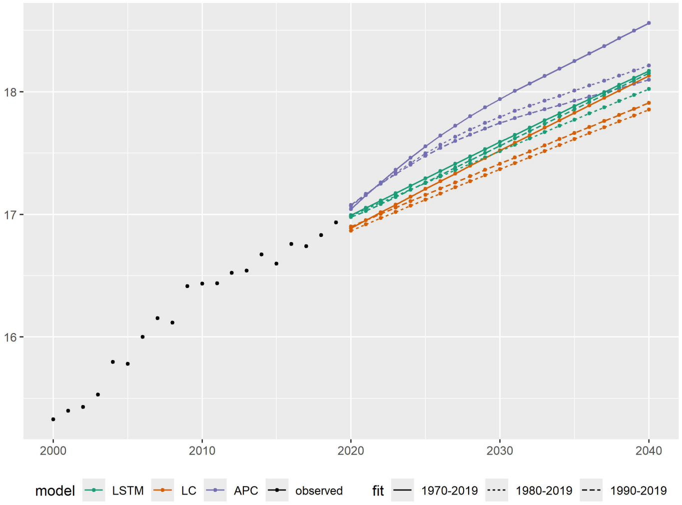


Figure 5: Forecasts of residual life expectancy at age 70 over several calibration periods.

Use case: presence of a mortality shock, USA, Female

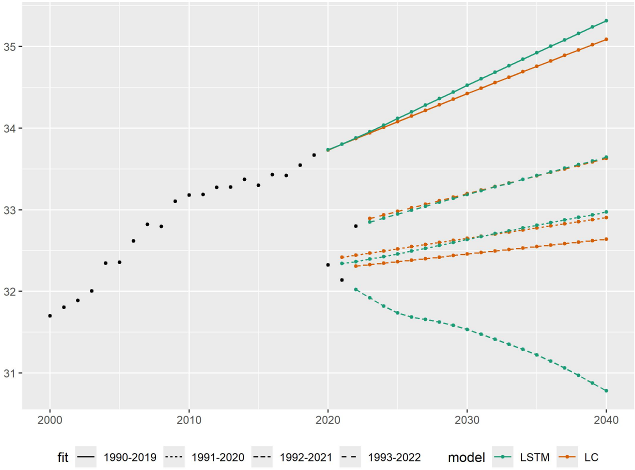


Figure 6: Forecasts of residual life expectancy at age 50 in presence of a pandemic shock.

Main results

- Analysis of Multivariate LSTM in multiple use cases with emphasis on forecasts
- Forecasts appear more influenced by recent data points than with other models
- In particular, very sensitive to short-term fluctuations

Merci pour votre attention !

Bibliography I

Barbieri, M. (2024).
The impact of the covid-19 pandemicon socio-economic disparities in mortality in the united states.
Longevity 19 conference.

Euthum, M., Scherer, M., and Ungolo, F. (2024).
A neural network approach for the mortality analysis of multiple populations: a case study on data of the italian population.
European Actuarial Journal, pages 1–30.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017).
A review of multivariate distributions for count data derived from the poisson distribution.
Wiley Interdisciplinary Reviews: Computational Statistics, 9(3):e1398.

Li, N. and Lee, R. (2005).
Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method.
Demography, 42:575–594.

Bibliography II

Lindholm, M. and Palmborg, L. (2022).
Efficient use of data for LSTM mortality forecasting.
European Actuarial Journal, pages 1–30.

Marino, M., Levantesi, S., and Nigri, A. (2021).
Deepening lee-carter for longevity projections with uncertainty estimation.
arXiv preprint arXiv:2103.10535.

Nigri, A., Levantesi, S., and Marino, M. (2021).
Life expectancy and lifespan disparity forecasting: a long short-term memory approach.
Scandinavian Actuarial Journal, 2021(2):110–133.

Richman, R. and Wuthrich, M. V. (2019).
Lee and carter go machine learning: recurrent neural networks.
Available at SSRN 3441030.

Rusch, T. and Zeileis, A. (2013).
Gaining insight with recursive partitioning of generalized linear models.
Journal of Statistical Computation and Simulation, 83(7):1301–1315.

Bibliography III

Seibold, H., Hothorn, T., and Zeileis, A. (2019).
Generalised linear model trees with global additive effects.
Advances in Data Analysis and Classification, 13(3):703–725.

Wüthrich, M. V. and Merz, M. (2023).
Statistical foundations of actuarial learning and its applications.
Springer Nature.

Perspectives

- Study more asymptotic properties of the regression estimators, e.g. comparison to MLE, speed of convergence
- Leverage closed-form formulas within ensemble methods, especially GLM-trees and GLM-forests
- Study the forecasts based on the alternative estimator
- Perform studies assuming a Dirichlet distribution instead of multinomial
- Apply multivariate LSTM for multi-populations or cause-of-death forecasts

Multivariate Poisson extensions

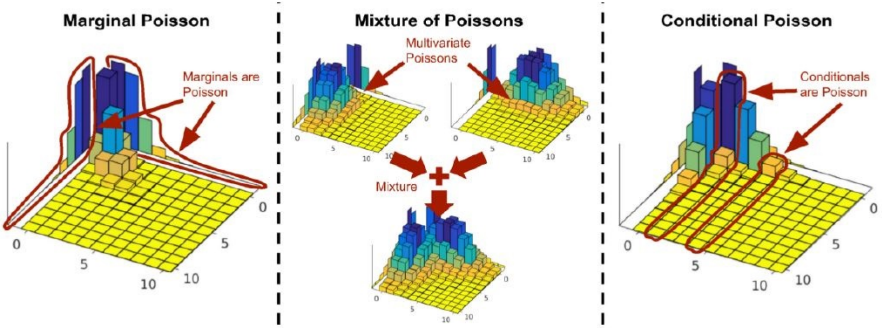


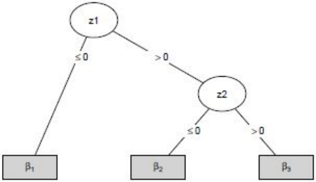
Figure 7: Source: [Inouye et al., 2017].

Motivation - Use of Generalized Linear Models (GLM)-trees

The GLM-based tree algorithm [Rusch and Zeileis, 2013] is a recursive split of the dataset based on partitioning variables, similar to other tree algorithm, e.g. CART.
A GLM is fitted at each node on a set of explanatory variables.

Main steps are:

1. Fit the GLM on the current sample
2. Asses parameter stability for each partition variable
3. Choose the best splitting point



Example from [Seibold et al., 2019]:GLM tree with 2 partition variables

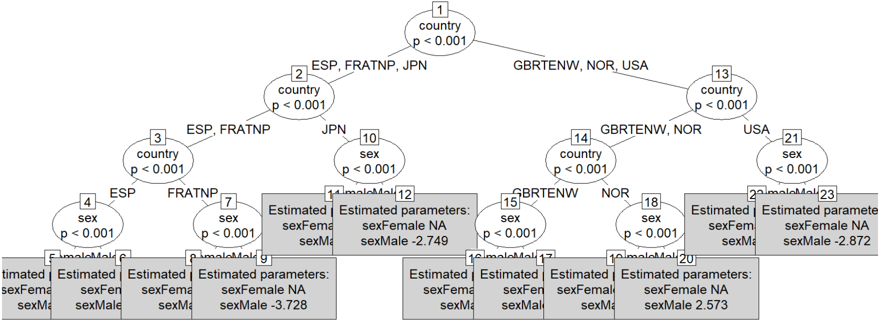
$$g(E(Y_i)) = \begin{cases} \beta_1 & \text{if } z_1 \leqslant 0 \\ \beta_2 & \text{if } z_1 > 0 \text{ and } z_2 \leqslant 0 \\ \beta_3 & \text{if } z_1 > 0 \text{ and } z_2 > 0 \end{cases}$$

Advantages: Automatic partitioning → enables classification + increased accuracy.

Drawbacks: Numerical optimization of the MLE → high computational time.

Use case - classification based on all-cause mortality

- Data source: Human Mortality Database
- Explicative variables: age (50-84) and year (1990-2019)
- Partitioning variables: country (France, Japan, Norway, Spain, UK, US) and sex (Male or Female)
- Computation time: 39s
- Convergence: numerical optimization may fail to converge to a solution
  → known issue with glm.fit in R when data are already well partitioned.



ICD10 classification

	Disease	ICD10
1	Certain infectious diseases	A00-B99
2	Neoplasms	C00-D48
3	Diseases of the blood and blood-forming organs	D50-D89
4	Endocrine, nutritional and metabolic diseases	E00-E88
5	Mental and behavioural disorders	F01-F99
6	Diseases of the nervous system and the sense organs	G00-G44, G47-H93
7	Heart diseases	I00-I51
8	Cerebrovascular diseases	G45, I60-I69
9	Other and unspecified disorders of the circulatory system	I70-I99, K64
10	Acute respiratory diseases	J00-J22, U04, U07
11	Other respiratory diseases	J30-J98
12	Diseases of the digestive system	K00-K63, K65-K92
13	Diseases of the skin	L00-M99
14	Diseases of the genitourinary system	N00-O99
15	Certains conditions originating in the perinatal period	P00-Q99, R95
16	External causes	V01-Y89

Back to MGAPC.

New approach with Multivariate LSTM

Several strategies are available with LSTM depending on input/output dimension

- "one-to-one", e.g. univariate forecasts of (κ_t) [Nigri et al., 2021, Marino et al., 2021, Lindholm and Palmborg, 2022]
- "many-to-one", e.g. forecasts of mortality rate by age [Richman and Wuthrich, 2019, Euthum et al., 2024]
- "many-to-many": proposed model

Table 2: Dataset characteristics

Model	Input size τ_0	Output size
[Nigri et al., 2021, Marino et al., 2021, Lindholm and Palmborg, 2022]	1	1
[Richman and Wuthrich, 2019, Euthum et al., 2024]	5	1
Proposed model	71	71

- Data transformation: log-ratio transform and normalization,
- Hyperparameters tuned using Bayesian Optimization: number of cells, number of layers, learning rate and lookback period.

Parameters and Hyper-parameters

Model	population	$\mathcal{X}_0$	$\mathcal{T}_0$	T	τ_0	$\sharp\mathcal{X}_0$	$\sharp\mathcal{T}_0$	$\lvert\mathcal{T}\rvert$
[Richman and Wuthrich, 2019]	Switzerland	0-99	1950-1999	10	5	100	40	4000
[Euthum et al., 2024]	Italy	50-95	1982-2014	10	5	45	23	9522
Proposed model	USA	30-100	1970-2019	$\{3,\ldots,7\}$	71	71	$\{30;40;50\}$	$\sharp\mathcal{T}_0 - 7$

Table 3: Summary of dataset characteristics

	Calibration period			
Hyperparameter	1990-2019	1991-2020	1992 - 2021	1993 - 2022
Learning rate	1.75E-03	6.27E-04	4.93E-03	1E-05
Number of neurons	20	15	6	40
Number of layers	1	1	2	2
Lookback period	3	3	7	5
Number of parameters	7440	5280	2232	31200

Table 4: Selected hyperparameters.

LSTM and confidence intervals

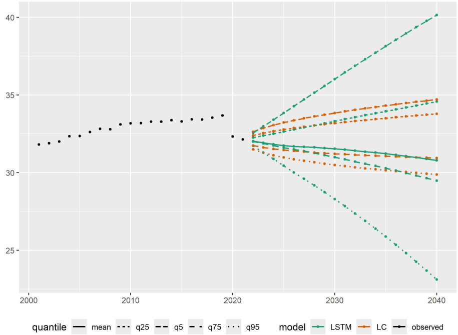


Figure 8: Forecasts of residual life expectancy at age 50 with confidence intervals, Female, 1992-2021.

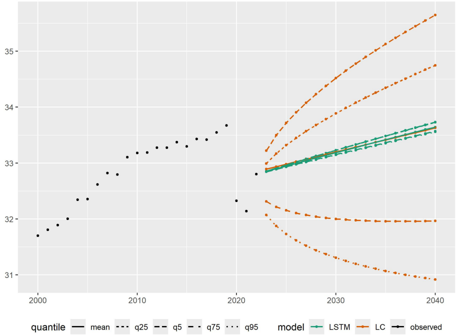


Figure 9: Forecasts of residual life expectancy at age 50 with confidence intervals, Female, 1993-2022.

LSTM for cause-of-death mortality forecasts

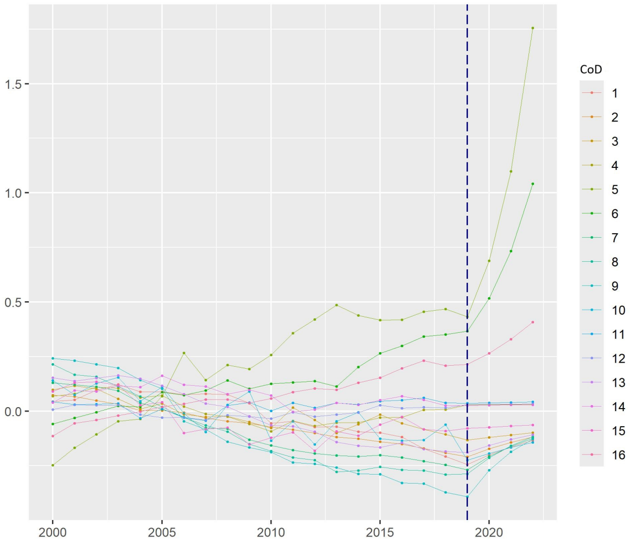


Figure 10: Forecasts of trend parameters, M3 assumed for each cause, Females, 1990-2019.