



HAL
open science

Contributions of Statistical Learning to Actuarial Sciences and Financial Risk Management

Pierrick Piette

► **To cite this version:**

Pierrick Piette. Contributions of Statistical Learning to Actuarial Sciences and Financial Risk Management. Risk Management [q-fin.RM]. Université Lyon 1 - Claude Bernard, 2019. English. NNT : . tel-02424550

HAL Id: tel-02424550

<https://hal.science/tel-02424550v1>

Submitted on 27 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N°d'ordre NNT : 2019LYSE1250

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale ED 486

Sciences Économiques et de Gestion

Spécialité de doctorat : Sciences de Gestion

Soutenue publiquement le 02 Décembre 2019 par :
Pierrick Piette

**Contributions de l'Apprentissage
Statistique à l'Actuariat et la Gestion des
Risques Financiers**

Devant le jury composé de :

Arthur Charpentier, Professeur à l'Université du Québec à Montréal
Caroline Hillairet, Professeure à l'ENSAE Paris
Christophe Gouel, Directeur de Recherche à l'INRA Paris
Aurélie Lemmens, Professeure à Erasmus University Rotterdam
Catherine Viot, Professeure à Université Lyon 1

Stéphane Loisel, Professeur à l'Université Lyon 1
Olivier Lopez, Professeur à Sorbonne Université
Hervé Morand, Président à Sinalys

Rapporteur
Rapporteuse
Examinateur
Examinatrice
Examinatrice

Directeur de thèse
Co-directeur de thèse
Invité

Remerciements

Tout d'abord, je tiens à remercier très chaleureusement Stéphane Loisel et Olivier Lopez, mes directeurs de thèse, qui m'ont accompagné durant ces années de doctorat. Votre soutien et votre disponibilité ont été très précieux pour ces travaux. Chacune de nos discussions furent instructives et stimulantes. Elles m'ont permis de mieux appréhender les problématiques et de pousser toujours plus loin mes réflexions. Mais vous avez aussi et surtout su faire de cette thèse une extraordinaire aventure humaine, et pour cela mille mercis !

Un grand merci à Arthur Charpentier et Caroline Hillairet d'avoir accepté de rapporter ma thèse et d'y avoir consacré de votre temps. J'en suis très honoré. Merci également à Aurélie Lemmens, Christophe Gouel et Catherine Viot de faire partie du jury aujourd'hui.

Cette thèse n'aurait pas été possible sans les excellentes conditions dont j'ai pu bénéficier à Sinalys. Merci beaucoup Hervé pour ta confiance et ta passion pour la recherche. Un gros clin d'oeil au pôle actuariat, et spécialement à Mourad pour nos nombreuses discussions.

Je remercie l'ensemble des membres du laboratoire SAF pour votre accueil. J'ai fortement apprécié chacun de mes passages à Lyon. Grand merci à Christian, Anne, Nicolas et Jean-Louis pour votre aide. Des remerciements spéciaux à Yahia pour tous nos échanges, de la mortalité au vin. Pour ces excellents moments de recherche et bien plus encore, merci à Quentin, Xavier, Pierre, Patrick, P-O, Julien sans oublier les co-doctorants : Claire, Junior, Gwladys, Sarah, Steve, Romain, Morgane, Maximilien, Arthur . . . et un énorme big up à Baptiste pour tout ce que l'on a partagé !

Merci aux membres du LPSM, notamment Nicole, Alexandre, Maud et Sarah, pour toutes nos discussions et vos précieuses remarques lors des groupes de travail à Jussieu. I also want to thank the NCCU team, and more specifically Jason, Morgan, Jin-Lung and Jennifer, for their wonderful welcome. Xie Xie !

Je ne serai sûrement pas arrivé là aujourd'hui sans la rencontre fortuite avec quelques personnes qui m'ont beaucoup appris et que je tiens à remercier ici : Christophe Boilley pour le goût de la recherche et des mathématiques, Jean Nicolini et Mohamed Benkhalfa pour l'initiation à l'actuariat, mais aussi Chloé Cadene, Fatima Roudani, Kévin Dedieu, Tanguy Touffut et Claudio Busarello.

Dans cette période post-thèse, je remercie grandement Jean, Philippe et Guillaume de me donner l'occasion de continuer la recherche en parallèle de la passionnante aventure qu'est Seyna. Et merci à Arnaud, Stéphanie, Guichou, Max et Yves pour votre accueil.

Cette thèse n'aurait certainement pas eu la même saveur si je n'avais pas pu la partager avec mes amis et compter sur leur soutien quotidien. Merci!! À Yolán, pour ta relecture attentive qui mérite bien un diabolo citron. À Virgile, et tes précieux conseils transatlantiques. À Thomas, pour nos échanges nocturnes sur les statistiques et le monde. À Damien, ami de plus de 20 ans déjà et autant de souvenirs mémorables. À Juliette, ma soeur d'armes toujours présente quelle que soit la distance.

Je ne pourrai tous vous citer (ceci n'est pas une revue exhaustive de la littérature) mais merci à mes amis de l'ENSAE Morgane, Caël, Raphaëlle, Ismaël, Loïc, Benjamin, Thomas ... de prépa Armelle, Lucas, Manu, Antoine, Joris, Zaz, ... la Shifumi team Sébastien et Arnaud ainsi que les bretteurs Damien, Élie, Astrid, Tiphaine ... les cartésiens Guillaume, Clément, Joss, Juan ... les nordistes Jean, Mickael, Rémy ... et les stambouliotes Ilos, Cans et Gustave.

Enfin, rien n'aurait été possible sans ma famille. Un immense merci à mes parents et à mon frère, à qui je dois tout. Et bien sûr à mes grands-parents, à qui je dédie cette thèse.

*À Frieda, Jules,
Étienne et Micheline.*

In God we trust,
all others bring data.

William Edwards Deming

Résumé

L'augmentation continue des performances informatiques des dernières décennies a permis une application répandue de la théorie de l'apprentissage statistique dans de multiples domaines. Les actuaires, experts historiques des statistiques, se tournent en particulier de plus en plus fréquemment vers ces algorithmes novateurs pour l'évaluation des risques auxquels ils sont confrontés. Ainsi, dans cette thèse, nous examinons comment l'intégration de méthodologies issues de l'apprentissage statistique peut contribuer au développement des sciences actuarielles et de la gestion des risques au travers de l'étude de trois problématiques indépendantes, présentées au préalable dans une introduction générale.

Les deux premiers chapitres proposent de nouveaux modèles de projection de mortalité dans le cadre de l'évaluation du risque de longévité porté par les compagnies d'assurances ou les fonds de pensions. Le Chapitre 1 s'attarde sur le cas où une seule population est étudiée, alors que le Chapitre 2 étend l'analyse à la multi-population. Dans les deux situations, la problématique de la grande dimension apparaît centrale et nous l'abordons à l'aide d'un vecteur autorégressif (VAR) pénalisé. Ce modèle est appliqué directement sur les taux d'amélioration de mortalité dans le premier chapitre, et sur les séries temporelles issues de l'estimation d'un modèle Lee-Carter pour le second. La pénalisation elastic-net permet de garder la grande liberté de structure de dépendance spatio-temporelle qu'offre le VAR tout en restant parcimonieux dans le nombre de paramètres, et donc éviter le surapprentissage.

Dans le Chapitre 3 nous analysons le risque de rachat des contrats d'assurance vie par l'utilisation d'algorithmes de classification supervisée. Nous y appliquons entre autres le séparateur à vaste marge (SVM) et l'extreme gradient boosting (XGBoost). Afin de comparer les performances des différents classificateurs nous adoptons une vision économique issue de la littérature du marketing et basée sur les profits potentiels d'une campagne de rétention. Nous insistons sur l'importance de la fonction de perte retenue dans les algorithmes d'apprentissage statistique suivant l'objectif recherché : l'utilisation d'une fonction de perte en lien avec la mesure de performance amène une amélioration significative dans l'application de l'XGBoost dans notre étude.

Enfin, dans le cadre de la gestion des risques financiers, nous étudions les dynamiques des prix agricoles lors de sessions de bourse particulières où des rapports gouvernementaux, conte-

nant des informations précieuses pour les agents, sont publiés. Nous examinons le potentiel des données en libre accès, en particulier les images satellites d'indice de végétation rendues disponibles par la NASA, pour la prédiction des réactions des marchés. Nous proposons alors des pistes d'amélioration à considérer pour une mise en œuvre pratique de cette méthodologie d'enrichissement des données dans la gestion des risques.

Mots-clés : Apprentissage statistique ; Projection de mortalité ; Rachat ; Marché des commodités ; Elastic-net ; Gradient Boosting ; NDVI.

Abstract

The continuous enhancement of computer performances during the last decades has favored a widespread application of statistical learning theory in multiple domains. Actuaries, long-standing statistical experts, notably turn more and more frequently to these new algorithms for the evaluation of risks they are facing. Thus, in this thesis, we examine how the integration of methods derived from statistical learning can contribute to the development of actuarial science and risk management through the study of three independent problematics, preliminarily presented in a general introduction.

The first two chapters propose new mortality forecasting models within the longevity risk evaluation framework that insurance companies and pension funds encounter. Chapter 1 focuses on the single population case, while Chapter 2 extends the study to the multi-population. In both situations, high-dimensionality appears to be a major concern. We tackle this issue thanks to a penalized vector autoregressive (VAR). This model is directly applied to the mortality improvement rates in the first chapter, and to the time series derived from Lee-Carter's model fits in the second one. The elastic-net penalization preserves the large freedom in the spatio-temporal dependence structure offered by the VAR, while remaining sparse in terms of estimated parameters, thereby avoiding overfitting.

In Chapter 3, we analyze lapse risk in life insurance policies through the use of supervised classification algorithms. We apply, among others, support vector machine (SVM) and extreme gradient boosting (XGBoost). In order to compare performances from one classifier to another, we adopt an economic point of view derived from the marketing literature and based on potential profits of a retention campaign. We insist on the importance of the loss function retained in the statistical learning algorithms according to the pursued objective: the use of a loss function linked to the performance measure leads to a significant enhancement in the application of the XGBoost in our study.

In the last Chapter, in the financial risks' management framework, we study agricultural commodity price dynamics throughout specific trading sessions where governmental reports, that contain valuable information for the agents, are released. We examine the potential of open data, in particular the satellite data on vegetation index made available thanks to the NASA, for market reactions forecast. We then suggest some improvements to be considered

before an operational implementation of this forecasting method.

Keywords: Statistical learning; Mortality forecasting; Lapse; Commodities market; Elastic-net; Gradient Boosting; NDVI.

Table des matières

| | |
|--|-------------|
| Remerciements | iii |
| Résumé | viii |
| Abstract | x |
| Table des matières | xiii |
| Liste des tableaux | xv |
| Liste des figures | xvii |
| Introduction Générale | 1 |
| Preliminaires | 1 |
| Actuariat Vie | 3 |
| Risque de Rachat | 20 |
| Informations sur les marchés agricoles | 29 |
| Introduction des outils statistiques | 36 |
| Contributions | 46 |
| Bibliographie | 51 |
| 1 Modèle VAR pénalisé pour la projection des taux d'amélioration de mortalité | 73 |
| Abstract | 73 |
| 1.1 Introduction | 74 |
| 1.2 A Vector Autoregression approach for mortality rate improvements | 77 |
| 1.3 High-dimensional estimation of the VAR model | 78 |
| 1.4 Empirical analysis | 81 |
| 1.5 A multi-population extension | 94 |
| 1.6 Conclusion | 95 |
| 1.7 Appendix | 98 |
| Bibliography | 104 |
| 2 Elastic-net et cohérence locale pour la modélisation des dynamiques de mortalité inter-population | 111 |
| Abstract | 111 |
| 2.1 Introduction | 112 |
| 2.2 The mortality dispersion issue | 114 |
| 2.3 A locally coherent approach | 118 |

| | | |
|----------|---|------------|
| 2.4 | Population clustering | 121 |
| 2.5 | Numerical applications | 124 |
| 2.6 | Toward a dynamic local coherence | 127 |
| 2.7 | Conclusions | 130 |
| 2.8 | Appendix | 132 |
| | Bibliography | 133 |
| 3 | Apprentissage statistique et mesures économiques pour la gestion du risque de rachat | 138 |
| | Abstract | 138 |
| 3.1 | Introduction | 139 |
| 3.2 | Binary classification algorithms | 140 |
| 3.3 | Validation metrics | 146 |
| 3.4 | Data | 148 |
| 3.5 | Result with respect to statistical and economic metrics | 151 |
| 3.6 | Optimization on profitability instead of classification | 156 |
| 3.7 | Conclusion | 159 |
| 3.8 | Appendix | 162 |
| | Bibliography | 163 |
| 4 | Potentiel des données satellite dans la gestion des risques financiers agricoles | 169 |
| | Abstract | 169 |
| 4.1 | Introduction | 170 |
| 4.2 | Data | 173 |
| 4.3 | Methodology | 174 |
| 4.4 | Results | 178 |
| 4.5 | Perspectives | 183 |
| 4.6 | Conclusion | 185 |
| | Bibliography | 187 |
| | Conclusion | 193 |

Liste des tableaux

| | | |
|------|--|-----|
| 0.1 | Matrice de confusion. | 26 |
| 1.1 | The estimated VAR-ENET hyper-parameters. | 84 |
| 1.2 | The RMSE of the VAR-ENET and benchmark models. | 87 |
| 1.3 | Summary statistics for the RMSE of the VAR-ENET and the benchmark models. | 87 |
| 1.4 | The RMSFE of the VAR and the benchmark models estimated on the period 1950 – 2000. | 90 |
| 1.5 | Summary statistics for the RMSFE of the VAR-ENET and the benchmark models. | 91 |
| 1.6 | The RMSFE of the VAR and the benchmark models estimated on the period 1970 – 2000. | 101 |
| 1.7 | Summary statistics for the RMSFE of the VAR-ENET and the benchmark models estimated on the period 1970 – 2000. | 101 |
| 1.8 | The RMSFE of the VAR and the benchmark models estimated on the period 1980 – 2000. | 102 |
| 1.9 | Summary statistics for the RMSFE of the VAR-ENET and the benchmark models estimated on the period 1980 – 2000. | 102 |
| 2.1 | 8 groups of populations determined by HCA on the LC's κ_t | 123 |
| 2.2 | Provisions and Solvency capital requirement estimates. | 126 |
| 3.1 | Descriptive statistics of categorical explanatory variables. | 150 |
| 3.2 | Descriptive statistics of continuous explanatory variables. | 151 |
| 3.3 | Cross-validated statistical accuracies. | 151 |
| 3.4 | Average confusion matrix of XGB. | 152 |
| 3.5 | Average confusion matrix of SVM. | 152 |
| 3.6 | Average confusion matrix of CART. | 153 |
| 3.7 | Average confusion matrix of LR. | 153 |
| 3.8 | Estimated retention probability r_{lapse} | 153 |
| 3.9 | Incentive strategies. | 153 |
| 3.10 | Cross-validated retention gains with the aggressive strategy. | 154 |
| 3.11 | Cross-validated retention gains with the moderate strategy. | 155 |
| 3.12 | Cross-validated statistical accuracies obtained with the economic loss functions. | 157 |
| 3.13 | Average confusion matrix of XGB_R1. | 158 |
| 3.14 | Average confusion matrix of XGB_R1. | 158 |
| 3.15 | Cross-validated retention gains with the aggressive strategy. | 158 |
| 3.16 | Cross-validated retention gains with the moderate strategy. | 159 |
| 4.1 | NDVI image periods and corresponding calendar dates. | 173 |
| 4.2 | Descriptive statistics of corn future returns over the period 2000-2016. | 175 |

| | | |
|-----|--|-----|
| 4.3 | Future return volatility test results for WASDE reports in Corn Markets, August-October, 2000-2016. | 179 |
| 4.4 | Regression results of future prices impacts to early corn yield estimate changes, and their Kendall τ , August to October, 2000-2016 | 180 |
| 4.5 | Regression results of final corn yield estimate to NDVI time series, 2000-2016 | 182 |
| 4.6 | Area planted for corn, soybeans and winter wheat in Iowa and Kansas in 2016 (thousands of acres). | 182 |
| 4.7 | Regression results of future prices impacts to early NDVI-based corn yield estimate changes, and their Kendall τ , August to October, 2000-2016 | 183 |

Liste des figures

| | | |
|------|--|----|
| 0.1 | Évolution de l'espérance de vie à la naissance. | 5 |
| 0.2 | Projection du rapport de dépendance en France entre 2008 et 2108. | 6 |
| 0.3 | Taux de mortalité en France en 2017. | 8 |
| 0.4 | Taux de mortalité en France en 2017 par sexe. | 9 |
| 0.5 | Logarithme des taux d'amélioration de la mortalité en France entre 1980 et 2016. | 10 |
| 0.6 | Logarithme des taux d'amélioration de la mortalité en Angleterre et Pays de Galles entre 1980 et 2016. | 10 |
| 0.7 | Fonctionnement d'un q -forward. | 15 |
| 0.8 | Taux de mortalité avant, pendant et après la période à risque. | 16 |
| 0.9 | Projection sur 50 ans des taux de mortalité des hommes français de la cohorte d'âge 50 ans en 2016 selon plusieurs modèles de mortalité. | 17 |
| 0.10 | Représentation schématique des gains d'une campagne de rétention. | 27 |
| 0.11 | Gestion de la rétention client. | 29 |
| 0.12 | Variance des rendements fermeture–ouverture des futures sur le maïs et le soja pour tous les mois de communication du WASDE. Janvier 1985 - Décembre 2006. | 35 |
| 0.13 | Gauche : deux hyperplans différents séparant le même ensemble de données. Droite : l'hyperplan à marge maximale. | 40 |
| 0.14 | Gauche : cas non séparable linéairement en dimension 1. Droite : séparation linéaire possible après projection de l'ensemble de données en dimension 2. | 40 |
| 0.15 | Spectres de réflectivité de la végétation en bonne et mauvaise santé, et du sol. | 44 |
| 1.1 | The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for overall populations. | 81 |
| 1.2 | The Granger causality matrix \mathbf{A}_1 for England and Wales (UK), the United States (US) and France (FR) on the age range 45-99 for females, males and the overall population. | 85 |
| 1.3 | The Granger causality matrix \mathbf{A}_1 for the French males on the age range 45-99 estimated with a VAR-ENET(1) over the period 1950-2012, on the HMD data downloaded on the 2 nd October 2017 (Before correction) and the 28 th January 2019 (After correction). | 86 |
| 1.4 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the overall populations. | 88 |
| 1.5 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the overall populations. | 89 |
| 1.6 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the overall populations. | 91 |

| | | |
|------|--|-----|
| 1.7 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the overall populations. | 92 |
| 1.8 | The observed and the projected log of death rates for British (UK), American (US) and French(FR) females and males with the 97.5% prediction intervals, obtained from the VAR-ENET model. | 93 |
| 1.9 | The observed and the projected log of death rates for British (UK), American (US) and French (FR) females and males. This figure compares trends obtained with the HU, the LC and the VAR-ENET models. | 94 |
| 1.10 | The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for females. | 98 |
| 1.11 | The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for males. | 98 |
| 1.12 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the female populations. | 99 |
| 1.13 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the female populations. | 99 |
| 1.14 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the male populations. | 99 |
| 1.15 | The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the male populations. | 100 |
| 1.16 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the female populations. | 100 |
| 1.17 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the female populations. | 103 |
| 1.18 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the male populations. | 103 |
| 1.19 | The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the male populations. | 103 |
| 2.1 | Dispersion at age 85 in the western European populations: historical data and median projections by Lee-Carter and Li-Lee models with the corresponding 95% prediction intervals. | 118 |
| 2.2 | Estimated autoregressive matrices of the VAR-ENET from the LC-LL(MF) model. | 122 |
| 2.3 | Dendrogram associated with the HCA applied on the LC's κ_t , colored according to 8 final clusters. | 124 |
| 2.4 | Dispersion at age 85 in the western European populations: historical data and median projections by LC, LL and LC-LL(MF) models with the corresponding confidence intervals at 5% and 95%. | 125 |
| 2.5 | Dispersion at age 85 in the western European populations: historical data and median projections by LC, LL and LC-LL(HCA8) models with the corresponding confidence intervals at 5% and 95%. | 125 |
| 2.6 | Median European LDIV(2024) according to the switching time of the French female population. | 130 |

| | | |
|-----|--|-----|
| 2.7 | Dispersion in the western European populations: historical data and median projections by LC, LL and LC-LL(MF) models with the corresponding confidence intervals at 5% and 95%. | 132 |
| 2.8 | Dispersion in the western European populations: historical data and median projections by LC, LL and LC-LL(HCA8) models with the corresponding confidence intervals at 5% and 95%. | 133 |
| 3.1 | Pseudocode of the Gradient Tree Boosting algorithm. | 142 |
| 3.2 | Box plot of statistical accuracies. | 152 |
| 3.3 | Box plot of retention gains with the aggressive strategy. | 154 |
| 3.4 | Box plot of retention gains with the moderate strategy. | 155 |
| 3.5 | Box plot of statistical accuracies obtained with the economic loss functions. | 157 |
| 3.6 | Box plot of retention gains with the aggressive strategy. | 159 |
| 3.7 | Box plot of retention gains with the moderate strategy. | 160 |
| 4.1 | Mean absolute future returns relative to the August reports session during the period 2000-2016 | 179 |
| 4.2 | Mean absolute future returns relative to the September reports session during the period 2000-2016 | 179 |
| 4.3 | Mean absolute future returns relative to the October reports session during the period 2000-2016 | 180 |
| 4.4 | NASS yield estimation changes to NDVI-based yield estimation changes, August to October, 2000-2016 | 183 |
| 4.5 | MODIS NDVI time series of Kansas over the year 2016. | 184 |

Introduction Générale

Préliminaires

L'une des plus anciennes applications connues de l'apprentissage par la statistique nous vient de la Grèce antique, lors de la guerre du Péloponnèse entre Athènes et Sparte, au V^{ème} siècle avant notre ère. [Thucydide \(411 av. J.-C.\)](#) décrit comment, au siège de Platée durant l'hiver 428 avant J.-C., les Athéniens mesurèrent les murs ennemis. Pour ce faire, plusieurs soldats comptèrent le nombre de rangées de briques. Même si plusieurs d'entre eux purent se tromper, la majorité obtinrent le juste compte. Ainsi, en multipliant le nombre le plus fréquemment observé par ses soldats, i.e. le mode en vocabulaire statistique, par l'épaisseur d'une brique, le commandant athénien put en déduire la hauteur du mur adverse afin d'évaluer si une percée des lignes ennemies était possible. Avec une terminologie spécifique à l'apprentissage statistique, anglophone et anachronique, nous pourrions dire que le commandant grec a utilisé une méthode d'*ensemble learning* dans laquelle chaque soldat est un *weak learner*.

Près de 2 500 ans plus tard, le domaine de la statistique a considérablement évolué de par ses champs d'application, les sources de données à disposition et les algorithmes utilisés. Hier le commandant athénien estimait la hauteur du mur adverse grâce à sa garnison, aujourd'hui l'actuaire calcule la prime d'assurance grâce à une forêt aléatoire. Ces dernières décennies, l'augmentation exponentielle des flux de données et de la puissance de calcul disponible sont à la fois une aubaine et un défi constant pour le statisticien, dont le travail consiste à "faire parler les données" en y extrayant et apprenant le maximum d'information. La partie quantitative jouant un rôle majeur dans cet apprentissage, de nombreux algorithmes ont été développés pour répondre aux différentes problématiques rencontrées ([Hastie et al., 2016](#)). Néanmoins, cette science des données ne se limite pas à la continuelle sophistication des méthodologies calculatoires appliquées : le choix et la qualité des sources d'information utilisées ne doivent en aucun cas être négligés.

Dans une économie engendrant toujours plus de données, il n'est pas étonnant d'observer un domaine toujours plus vaste d'application de l'apprentissage statistique, dont nous donnons ici quelques exemples :

- reconnaissance de code postal écrit à la main ([Scholkopf et al., 1997](#)) ;

- détection de fraudes (Bolton et Hand, 2002) ;
- prédiction de la résiliation d'un contrat de télécommunication à partir des données clients (Lemmens et Croux, 2006) ;
- prédiction de l'évolution des prix de marché en analysant les sentiments sur un média social (Nguyen et al., 2015) ;
- orchestration automatique d'un morceau de piano (Crestel et Esling, 2016).

Logiquement, les sciences actuarielles sont l'un des domaines impactés par ce récent développement de l'apprentissage statistique (e.g., Charpentier, 2014; Bellina, 2014; Lopez et al., 2016). En effet, la genèse de cette discipline au XVII^{ème} siècle est intimement liée à celle de la statistique. Elle remonte à la publication à Londres en 1662 par John Graunt, aidé par William Petty, de *Observations naturelles et politiques sur les bulletins de mortalité*, qui contient la première table de mortalité (Le Bras, 2013). Cet ouvrage est souvent reconnu comme marquant la naissance commune de l'actuariat, de la démographie mais aussi de la statistique moderne. Il est donc naturel que, de nos jours, les sciences actuarielles bénéficient de ce nouvel essor dans les domaines de l'informatique et de la statistique, que l'on englobe communément par les termes anglophones de *Data Science* ou *Machine Learning*.

Structure

Dans cette thèse nous présentons quelques contributions de l'application des méthodologies d'apprentissage statistique aux sciences actuarielles et à la gestion des risques. Cette thèse s'articule en quatre chapitres indépendants, précédés d'une introduction générale. Dans cette dernière nous y décrivons notamment les motivations sous-jacentes à nos recherches et résumons les principales contributions de nos résultats. Les contributions issues de la thèse sont, quant à elles, détaillées dans les quatre chapitres.

- Le Chapitre 1 reprend l'article "Forecasting mortality rate improvements with a high-dimensional VAR", co-écrit avec Quentin Guibert et Olivier Lopez, publié dans *Insurance : Mathematics & Economics*. Nous y présentons un nouveau modèle de projection de taux de mortalité à l'aide d'un vecteur autorégressif dont l'estimation des coefficients est régularisée pour prendre en compte les problématiques de grande dimension.
- Le Chapitre 2 est issue du document de recherche "Bridging the Li-Carter's gap : a locally coherent mortality forecasting approach", co-écrit avec Quentin Guibert, Stéphane Loisel et Olivier Lopez. Nous y introduisons une modélisation des dynamiques de mortalité entre de multiples populations à l'aide d'un vecteur autorégressif pénalisé.
- Le Chapitre 3 est tiré de l'article "Applying economic measures to lapse risk management with machine learning approaches", co-écrit avec Stéphane Loisel et Cheng-Hsien Jason Tsai, soumis. Nous y analysons la détection des rachats de contrats d'assurance vie à l'aide d'algorithmes de classification supervisée que nous comparons entre eux à

l'aide notamment de mesures de performance économiques.

- Le Chapitre 4 reprend le papier "Can satellite data forecast valuable information from USDA reports? Evidences on corn yield estimates", publié dans les actes de la *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*. Nous y étudions le potentiel des données satellite pour prédire les informations contenues dans les rapports du Département d'agriculture des États-Unis et ayant de la valeur pour les agents des marchés financiers agricoles.

Dans la suite de cette introduction générale, nous commençons par exposer notre premier sujet d'étude qu'est l'actuariat vie. Plus précisément nous nous intéressons à deux risques particuliers auxquels cette discipline est confrontée. Ainsi nous introduisons dans un premier temps la problématique du risque de longévité. Puis nous décrivons le risque inhérent aux comportements des assurés en assurance vie, en particulier le rachat, tout en faisant le parallèle avec le risque de comportement d'attrition étudié dans la littérature du marketing quantitatif. Dans une troisième partie, nous nous attardons à l'importance de l'information, i.e. des données, dans les marchés financiers et plus particulièrement dans les marchés agricoles. Ensuite, nous introduisons les outils d'apprentissage statistique (algorithmes et sources de données) que nous avons utilisés lors de nos travaux. Finalement, nous résumons les contributions aux sciences actuarielles et à la gestion du risque découlant de l'application de telles méthodologies.

Actuariat Vie

L'actuariat vie est la discipline qui porte sur la gestion des risques de l'ensemble des contrats faisant naître des engagements dont l'exécution dépend de la durée de la vie humaine. A ce titre, nous pouvons citer quelques exemples connus de produit d'assurance comme l'assurance vie, la retraite, l'assurance décès ou l'assurance emprunteur. Notons que d'autres formes de contrat, moins répandus, existent aussi tels que la tontine ou le *longevity swap*. Ces contrats peuvent être gérés par des compagnies d'assurances, des mutuelles, des fonds de pension, mais aussi par des organisations étatiques comme par exemple la Caisse nationale d'assurance vieillesse pour une partie des retraites françaises.

Le produit d'assurance vie est depuis plusieurs années le placement d'épargne préféré des français. Ainsi, en 2018, la collecte nette pour l'ensemble des 54 millions de contrats a atteint 22,4¹ milliards d'euros. L'encours sous-jacent, i.e. la somme des provisions mathématiques et des provisions pour participation aux bénéfices, s'élevait à 1 700 milliards d'euros à fin décembre 2018. En comparaison le montant des engagements des compagnies d'assurance au titre de l'assurance retraite ne s'élève qu'à 190 milliards à fin 2016. Cependant ce chiffre est à prendre avec du recul dans le cas de la France car les retraites sont essentiellement financées par un système dit par répartition. En 2015, les versements de prestations au titre de la

1. Source : Fédération Française de l'Assurance

retraite par les régimes obligatoires ont été de 302 milliards d’euros alors que les organismes d’assurances complémentaires n’ont versé que 5,7 milliard (soit environ 2% du montant global). Au contraire, le système de retraite du Royaume-Uni est majoritairement un système dit par capitalisation, reposant notamment sur les quelques 100 000 fonds de pension que compte le pays. Ces fonds de pension sont parmi les plus grands investisseurs institutionnels, détenant plus de 2 900² milliards de dollars en divers actifs dont un tiers de l’ensemble des actions du pays et un cinquième des obligations d’État.

Les montants colossaux sous-jacents à ces contrats font que le système assurantiel présente un risque systémique, i.e. qu’un événement particulier pourrait avoir des conséquences négatives considérables sur l’ensemble de l’économie (Harrington, 2009). Le rôle de la gestion des risques, et donc de l’actuaire en particulier, y est alors primordial. A l’instar du système bancaire, le milieu de l’assurance s’est doté d’institutions de régulation comme le *European Insurance and Occupational Pensions Authority* (EIOPA) au niveau européen ou de l’Autorité de Contrôle Prudentiel et de Résolution (ACPR) en France. L’un des principaux rôles de ces régulateurs est de s’assurer de la bonne évaluation des risques portés par la société et de l’adéquation des fonds propres disponibles à ces derniers. Plusieurs référentiels réglementaires ont été mis en place pour évaluer l’exigence en capital dont la directive Solvabilité II pour l’Union Européenne, le *Swiss Solvency Test* en Suisse ou le *Risk-Based Capital* aux États-Unis (voir Holzmüller, 2009, pour une comparaison de ces trois régulations). Même si certaines hypothèses effectuées par ces cadres réglementaires peuvent être critiquées (e.g., Vedani *et al.*, 2017), ces dispositifs n’en soulignent pas moins la grande importance de l’actuariat. La modélisation de ces risques, principalement par des outils statistiques et probabilistes, est fondamentale pour leur évaluation quantitative.

Les risques que l’actuariat vie se doit d’étudier sont de diverses natures. Ils peuvent, entre autres, être biométriques, financiers ou comportementaux. Dans le cadre de cette thèse nous nous intéressons à l’analyse de deux risques particuliers : le risque de longévité et le risque de rachat. Le premier impacte principalement les produits incluant le paiement de rentes viagères, en particulier l’assurance retraite, et est donc supporté à la fois par des acteurs privés et publics. Au contraire, le rachat n’affectant que les produits d’assurance vie souscrits auprès d’une compagnie ou mutuelle, c’est exclusivement le secteur privé qui est touché par ce risque de comportement des assurés. La cinquième étude quantitative d’impact (QIS 5) de la directive Solvabilité II estime que ces deux risques sont les deux plus substantiels portés par les compagnies d’assurances vie, significativement plus importants que le risque de mortalité, d’invalidité, de coût ou de catastrophe (EIOPA, 2011).

2. Source : Organisation de coopération et de développement économiques

Risque de Longévité

Depuis plus d'un demi-siècle, nous observons une augmentation de l'espérance de vie dans la majorité des pays développés (cf. Figure 0.1). Cet accroissement de la longévité humaine est un défi pour les systèmes de retraite, qu'ils soient par répartition ou capitalisation. L'incertitude sur la projection future de la longévité est d'autant plus risquée que la valeur des passifs est importante et sensible aux changements de tendance. En effet, l'actuariat vie doit, entre autres, se soucier de l'évaluation des retraites des individus qui viennent de commencer à travailler : en notant qu'en 2017 l'espérance de vie résiduelle d'une femme française âgée de 20 ans est de plus de 65 ans (en supposant un arrêt de l'accroissement de la longévité), nous comprenons l'ampleur de la tâche de l'actuaire. En 2017, l'ensemble des actifs des fonds de pension de l'Organisation de Coopération et de Développement Économique (OCDE) représentait une valeur supérieure à 28 000³ milliards de dollars. Une année supplémentaire, et non anticipée par les actuaires, d'espérance de vie à 65 ans représente une augmentation d'environ 5% des passifs de ces fonds (Blake *et al.*, 2018), soit 1 400 milliards de dollars. Michaelson et Mulholland (2014) estiment qu'un événement de queue de distribution de l'accroissement de la longévité (i.e., 2,5 fois l'écart type) provoquerait une augmentation pouvant atteindre 10%, soit 2 800 milliards, i.e. de l'ordre de grandeur du PIB de la France. Notons bien que nous nous focalisons seulement sur le système des fonds de pension dans ces chiffres : nous n'y incluons pas les autres systèmes de retraite.

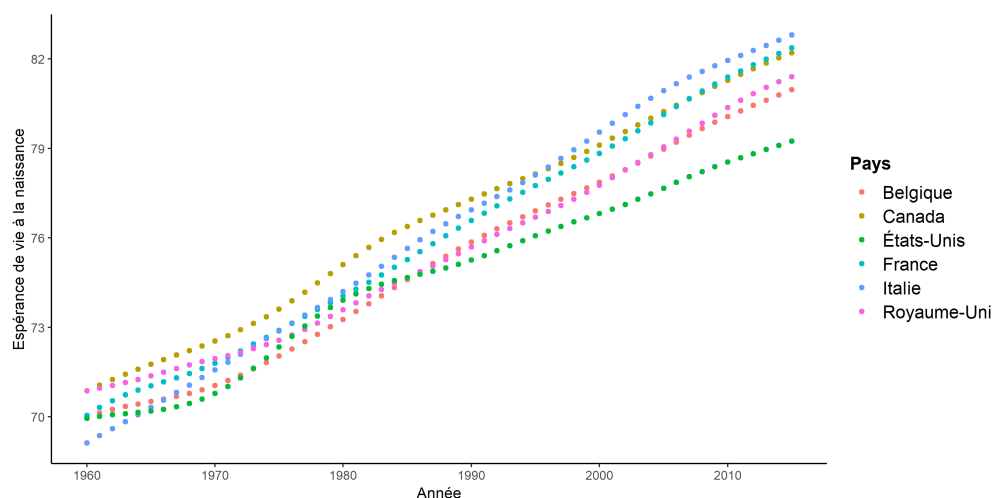


FIGURE 0.1 – Évolution de l'espérance de vie à la naissance.

Source : Nations Unies.

D'autres risques, nécessitant aussi la modélisation et la projection de la durée de vie humaine, peuvent venir mettre à mal les systèmes de retraites. Lorsque l'on s'attarde à l'analyse de la pyramide des âges et la projection de la population, des changements sur la structure en âge de

3. Source : OCDE *Global Pension Statistics*

la population peuvent compromettre les systèmes dit par répartition. C’est le cas par exemple de la France où la génération vieillissante dite du *baby-boom*, on parle d’ailleurs aujourd’hui de *papy-boom*, augmente significativement le rapport de dépendance (cf. Table 0.2), i.e. le rapport entre le nombre de personnes âgées entre 15 et 65 ans exclu et les individus âgés de plus de 65 ans. Ce dernier est une estimation de la taille de la population active par rapport à la population des retraités. Cet indicateur est donc crucial pour la stabilité des systèmes par répartition, son évolution doit de fait être modélisée et analysée attentivement. Ainsi en France, le risque de longévité pesant sur les retraites est d’autant plus fort pour cette cohorte particulière du *baby-boom*.

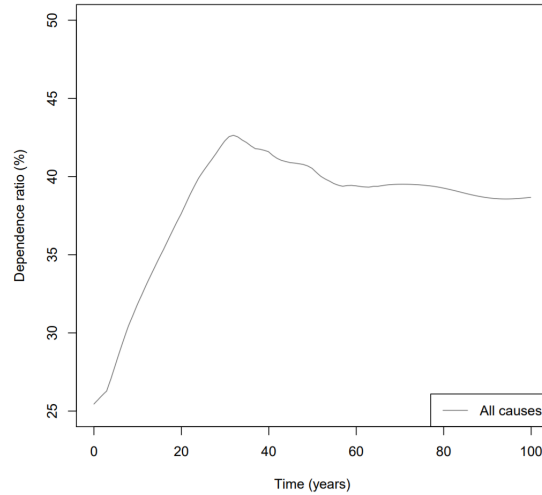


FIGURE 0.2 – Projection du rapport de dépendance en France entre 2008 et 2108.

Source : Boumezoued (2016a).

Quantification de la mortalité

L’évaluation du risque de longévité, i.e. l’étude de la durée de vie humaine, repose sur l’estimation et la projection des taux de mortalité, que l’on retrouve sous forme de tables de mortalité depuis les débuts de l’actuariat et de la démographie (Le Bras, 2013). Notons τ la durée de vie d’un individu. La loi de cette variable aléatoire, en supposant qu’elle admette une densité, est définie par sa probabilité de survie à l’âge a :

$$S(a) = \mathbb{P}(\tau > a) = \exp\left(-\int_0^a \mu(x) dx\right), \quad (0.1)$$

où $\mu(x)$ est la force de mortalité à l’âge exact x . La fonction $\mu(x)$ peut aussi être vue comme la probabilité instantanée de décès à l’âge x en remarquant que $\mu(x) dx = \mathbb{P}(x \leq \tau < x + dx \mid \tau \geq x)$. Néanmoins, à la vue de l’évolution de l’espérance de vie (cf. Figure 0.1), il est naturel de supposer que cette force de mortalité μ soit non seulement dépendante en âge x , mais aussi en fonction du temps t . Nous définissons alors $\mu(x, t)$ la force de mortalité des individus d’âge x au

temps t . En notant ν la date de naissance d'un individu, sa durée de vie est alors caractérisée par la fonction de survie :

$$S(a | \nu) = \mathbb{P}(\tau > a | \nu) = \exp\left(-\int_0^a \mu(x, \nu + x) dx\right). \quad (0.2)$$

Cependant cette fonction $\mu(x, t)$ est difficilement estimable en pratique par les statisticiens car elle est définie en temps continu. Ainsi, l'hypothèse simplificatrice régulièrement retenue lors de l'analyse des dynamiques de mortalité est que pour tout âge entier x et année entière t , et pour tout $0 \leq u, s < 1$, on a $\mu(x + u, t + s) = \mu(x, t)$, i.e. que la force de mortalité est constante sur chaque âge entier et sur chaque année calendaire. Nous définissons alors $m(x, t) = \mu(x, t)$ comme le taux central de mortalité. Une estimation de ce taux est donnée par :

$$m(x, t) = m_{x,t} = \frac{D(x, t)}{E(x, t)}, \quad (0.3)$$

où $D(x, t)$ représente le nombre d'individus décédés l'année t et d'âge x à leur dernier anniversaire avant leur mort, et $E(x, t)$ est l'exposition centrale au risque. Cette exposition est théoriquement définie comme la durée totale vécue dans l'année t par les individus d'âge x à leur dernier anniversaire. Mathématiquement, nous écrivons donc

$$E(x, t) = \int_0^1 P(x, t + s) ds, \quad (0.4)$$

où $P(x, s)$ est le nombre de personnes d'âge x à leur dernier anniversaire au temps exact s . Pratiquement, les données disponibles, notamment au niveau national, ne sont pas suffisantes pour estimer $E(x, t)$ de manière fiable. Des hypothèses simplificatrices supplémentaires sont alors posées. Ainsi, au Royaume-Uni l'*Office for National Statistics* (ONS) estime l'exposition centrale par la population à mi-année, i.e.

$$E(x, t) = P\left(x, t + \frac{1}{2}\right). \quad (0.5)$$

Même si elles semblent n'avoir que peu d'impact à première vue, certaines des hypothèses effectuées par les organismes nationaux peuvent avoir des conséquences significatives sur l'estimation des taux de mortalité, qui servent ensuite de base aux études de longévité (cf. [Cairns et al., 2016](#); [Boumezoued, 2016b](#)).

Une seconde mesure possible de la mortalité est la probabilité de décès $q(x, t)$, i.e. la probabilité qu'un individu d'âge exact x à l'année exacte t décède entre t et $t + 1$. Sous les hypothèses énoncées précédemment, nous avons alors la relation

$$q(x, t) = q_{x,t} = 1 - \exp(-m_{x,t}). \quad (0.6)$$

Structure des taux de mortalité

Analysons à présent la courbe des taux de mortalité ainsi définis. Les données que nous utilisons proviennent de la *Human Mortality Database* (HMD, 2019). La Figure 0.3 représente les taux de mortalité par âge pour l'ensemble de la population française en 2017, ainsi que leur logarithme. En effet, les taux de mortalité croissant exponentiellement aux âges élevés, il est usuel de les analyser en échelle logarithmique pour une meilleure comparaison aux différents âges. La première année de vie nous remarquons une mortalité infantile élevée comparée aux autres âges, suivie de faible taux aux alentours de 10 ans. Notons ensuite la bosse des accidents dans le voisinage des 20 ans, et enfin une augmentation environ linéaire du logarithme de la force de mortalité en fonction de l'âge au-delà de 40 ans. Cette dernière observation nous renvoie à l'un des premiers modèles de mortalité celui de Gompertz (1825) selon lequel la force de mortalité croit exponentiellement à partir d'un certain âge, i.e. $\mu(x) = \alpha \exp(\beta x)$.

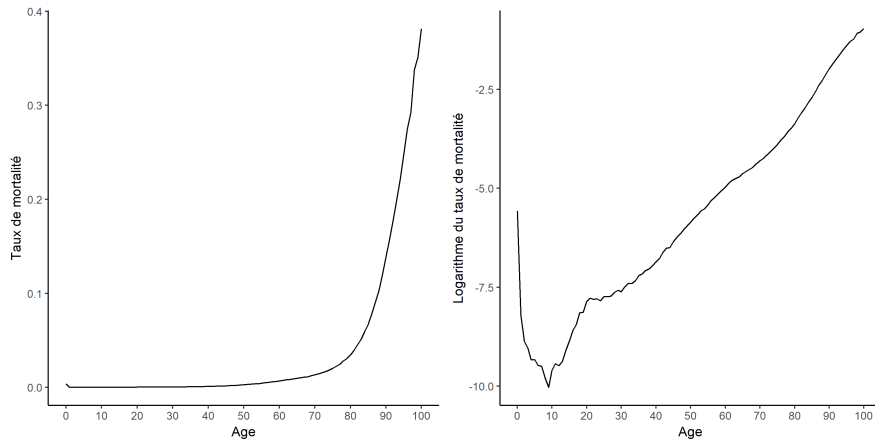


FIGURE 0.3 – Taux de mortalité en France en 2017.

En estimant les taux de mortalité non plus sur l'ensemble de la population française, mais en différentiant cette dernière par le sexe, nous observons des différences significatives entre ces deux populations. Sur la Figure 0.4 nous remarquons premièrement une mortalité plus élevée pour les hommes à partir de 15 ans environ. Au contraire, la mortalité infantile est relativement identique pour les deux sexes. Enfin, la bosse des accidents apparaît beaucoup plus marquée pour les hommes. Ces remarques nous amènent à un point crucial dans la gestion du risque de longévité : l'hétérogénéité des risques biométriques dans les populations. Ici nous en avons un exemple simple bien connu sur la différence d'espérance de vie entre les hommes et les femmes. Mais des facteurs socio-économiques peuvent aussi influencer sur cette hétérogénéité de la mortalité comme par exemple le niveau de revenu ou le statut marital (voir par exemple Boumezoued *et al.*, 2017; Cairns *et al.*, 2019). Nous reviendrons par la suite sur

les problématiques concrètes qu’une telle hétérogénéité peut induire dans la gestion du risque de longévité.

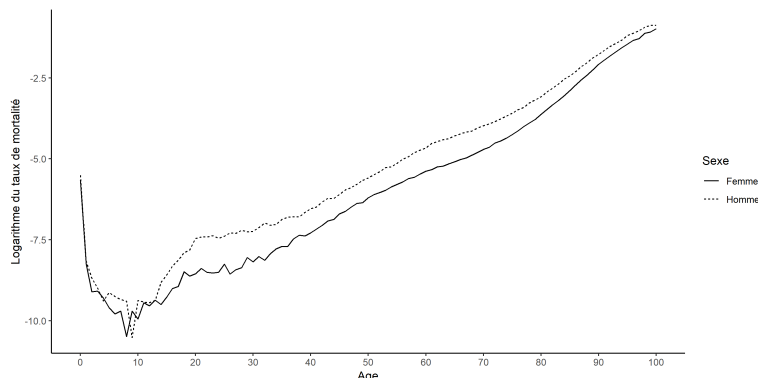


FIGURE 0.4 – Taux de mortalité en France en 2017 par sexe.

Comme nous l’avons remarqué précédemment, les taux de mortalité $m_{x,t}$ dépendent non seulement de l’âge x mais aussi de l’année calendaire t . Pour analyser l’évolution temporelle de la mortalité au sein d’une population, une variable d’intérêt est le logarithme du taux d’amélioration de la mortalité $\Delta \log(m_{x,t}) = \log(m_{x,t}) - \log(m_{x,t-1}) = \log\left(\frac{m_{x,t}}{m_{x,t-1}}\right)$. Les Figures 0.5 et 0.6 représentent cette variable respectivement pour la France et l’Angleterre et pays de Galles. Notons d’abord que la majorité des logarithmes des améliorations sont négatifs, i.e. les taux de mortalité baissent. Cette tendance est la cause directe de l’accroissement de l’espérance de vie. Ensuite, deux types de motifs ressortent de l’analyse de cette surface d’amélioration de la mortalité.

Premièrement, nous constatons des structures verticales correspondant à des effets dit de période. Pour certaines années, nous observons que la mortalité évolue de la même façon pour un ensemble d’âges voisins. Certains effets sont expliqués par des événements spécifiques comme une épidémie de grippe ou une vague de chaleur (Huynen *et al.*, 2001). Ces événements exogènes peuvent avoir des conséquences à court terme tel qu’une sur-mortalité dans l’année d’occurrence, mais aussi à moyen terme. C’est l’une des explications de la sous-mortalité en France en 2004, visible sur la Figure 0.5 par une verticale rouge, i.e. une forte diminution de la mortalité. En 2003, une canicule a engendré l’un des étés les plus chauds jamais enregistrés en France. La conséquence directe fut une sur-mortalité constatée sur cette année, notamment aux âges les plus élevés. Ensuite, en 2004, la mortalité fut nettement plus basse à cause de l’effet moisson (Toulemon et Barbieri, 2008; Izraelewicz, 2012). Pratiquement, une partie des personnes décédées à cause de la vague de chaleur à l’été 2003 avaient une faible espérance de vie résiduelle, et "auraient dû" mourir en 2004.

Le second effet observé se traduit graphiquement par des diagonales. Il s’agit de l’effet cohorte, i.e. des individus nés à la même date, ou par extension dans la même période de temps, présentent une dynamique de mortalité significativement différente des autres générations

(Willets, 2004). Nous remarquons que les taux d'améliorations de la mortalité en l'Angleterre et du pays de Galles, représentés dans la Figure 0.6, exhibent un effet cohorte significatif pour les individus né en 1920, représentée graphiquement par une diagonale bleue.

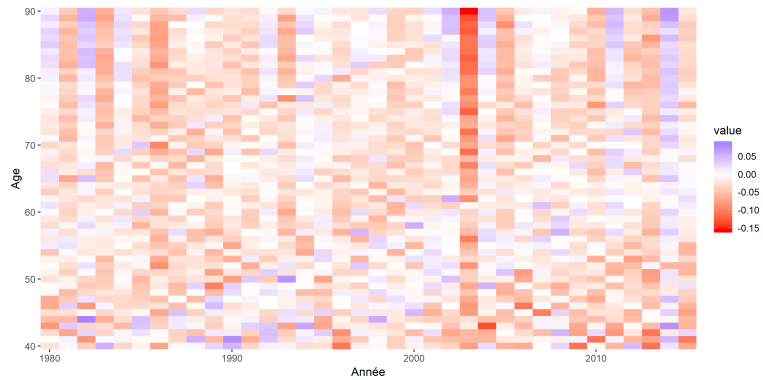


FIGURE 0.5 – Logarithme des taux d'amélioration de la mortalité en France entre 1980 et 2016.

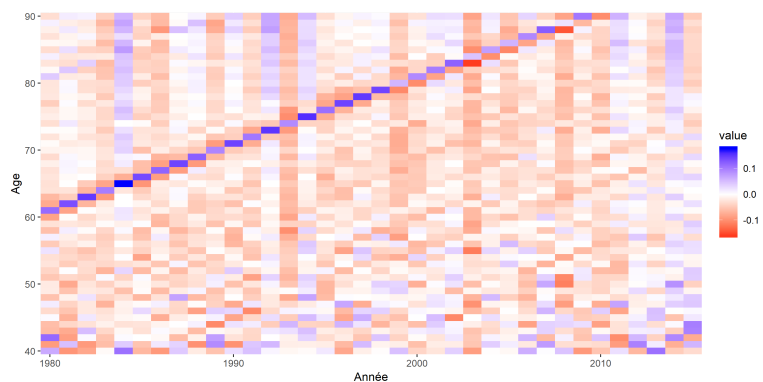


FIGURE 0.6 – Logarithme des taux d'amélioration de la mortalité en Angleterre et Pays de Galles entre 1980 et 2016.

La modélisation et la projection des taux de mortalité est donc une problématique faisant intervenir l'analyse de séries temporelles selon plusieurs dimensions. Nous venons de mettre en avant certaines d'entre elles :

- la dimension temporelle, par l'analyse de la tendance à long-terme mais aussi d'événements plus ponctuels ;
- la dimension âge ;
- la dimension cohorte ;
- la dimension population, due à la problématique d'hétérogénéité.

Modèles de projection

Cette complexité démographique, couplée à l'intérêt toujours grandissant des institutions gouvernementales, fonds de pensions et compagnies d'assurances, a engendré le développement

d'un grand nombre de modèles pour la compréhension, la modélisation et la gestion du risque de longévité (Barrieu *et al.*, 2012).

Parmi les modèles de mortalité les plus utilisés en actuariat vie de nos jours, nous présentons dans un premier temps la famille reposant sur la décomposition des taux de mortalité en facteurs âge-période-cohorte, permettant ainsi de réduire les dimensions d'analyse. A l'origine de cette famille de modèles, nous retrouvons le Lee-Carter, sûrement le plus connu et utilisé en actuariat, introduit par Lee et Carter (1992). Ce dernier décrit la dynamique des taux de mortalité par deux séries d'effet âge α_x et β_x , ainsi qu'une série d'effet période κ_t :

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t. \quad (0.7)$$

Cependant, nous pouvons remarquer que pour tout réel a et b et étant donné une solution formée des séries $\alpha_x, \beta_x, \kappa_t$; alors les nouveaux estimateurs $\tilde{\alpha}_x, \tilde{\beta}_x$ et $\tilde{\kappa}_t$, définis par

$$\begin{aligned} \tilde{\alpha}_x &= \alpha_x + b\beta_x, \\ \tilde{\beta}_x &= \beta_x a, \\ \tilde{\kappa}_t &= a(\kappa_t - b), \end{aligned} \quad (0.8)$$

sont aussi une solution du modèle. Ceci implique donc un problème d'identifiabilité dans l'estimation des différents facteurs. Pour résoudre ce dernier, nous avons besoin d'imposer des contraintes. Les plus régulièrement retenue sont les suivantes :

$$\begin{aligned} \sum_t \kappa_t &= 0 \\ \text{et } \sum_x \beta_x &= 1. \end{aligned} \quad (0.9)$$

Ainsi définis, α_x représente le niveau moyen de mortalité à chaque âge au cours du temps; la série κ_t décrit la dynamique temporelle d'amélioration de la mortalité et β_x précise la sensibilité de la mortalité à chaque âge par rapport à κ_t . A l'origine, Lee et Carter (1992) proposent une estimation des paramètres α_x, β_x et κ_t par une méthode de décomposition en valeurs singulières. Depuis, l'estimation par vraisemblance est plus commune, notamment dans les packages informatiques qui ont été développés (e.g., Villegas *et al.*, 2017b). Cette estimation est fondée sur l'hypothèse que le nombre de décès $D(x, t)$ suit une loi de Poisson ayant pour moyenne la force totale de mortalité, i.e. $E(x, t) \times m(x, t)$. Enfin, une fois ces séries de paramètres estimées, une seconde étape d'analyse des séries temporelles des différents facteurs permet une projection stochastique des taux de mortalité.

Depuis sa publication, le modèle Lee-Carter a engendré de nombreuses variantes et extensions. Ainsi [Renshaw et Haberman \(2006\)](#) proposent d'inclure un effet cohorte, non présent dans le modèle initial, par l'ajout d'une nouvelle série de paramètres d'effet cohorte γ_{t-x} :

$$\log m_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t + \beta_x^{(2)} \gamma_{t-x}. \quad (0.10)$$

A l'instar du Lee-Carter, l'estimation des facteurs du modèle de [Renshaw et Haberman \(2006\)](#) nécessite l'ajout de contraintes pour des questions d'identifiabilité :

$$\begin{aligned} \sum_t \kappa_t &= 0, \quad \sum_x \beta_x^{(1)} = 1, \\ \text{et } \sum_{x,t} \gamma_{t-x} &= 0, \quad \sum_x \beta_x^{(2)} = 1. \end{aligned} \quad (0.11)$$

Une autre variante de ces modèles factoriels est le modèle Age-Période-Cohorte, introduit par [Currie \(2006\)](#), dans lequel un seul jeu de facteurs d'effet âge est estimé :

$$\log m_{x,t} = \alpha_x + \kappa_t + \gamma_{t-x}. \quad (0.12)$$

La famille de modèles basés sur le Lee-Carter est bien entendu encore plus vaste de par ses extensions, dont nous donnons ici seulement deux des exemples les plus utilisés, mais aussi par ses procédures alternatives d'estimation. Le choix de la dynamique des séries temporelles alors obtenues peut aussi varier, allant d'une simple marche aléatoire avec dérive au modèle ARIMA plus général.

Un second ensemble de modélisations de la mortalité reprend l'idée fondatrice du modèle de [Gompertz \(1825\)](#) qu'à partir d'un certain âge la force de mortalité s'accroît exponentiellement avec l'âge de l'individu. Les modèles Cairns, Blake et Dowd (CBD) étendent cette hypothèse au cadre dynamique. Ainsi [Cairns *et al.* \(2006\)](#) proposent la formulation suivante de la probabilité de décès :

$$\text{logit } q_{x,t} = \kappa_t^{(1)} + \kappa_t^{(2)} (x - \bar{x}), \quad (0.13)$$

où \bar{x} est l'âge moyen dans l'ensemble des âges modélisés. En réécrivant la loi de Gompertz avec la dimension temporelle par $\ln \mu(x, t) = \ln \alpha_t + \beta_t x$, nous pouvons remarquer les similitudes avec la structure de [Cairns *et al.* \(2006\)](#). Dans la lignée de ce premier modèle, [Cairns *et al.* \(2009\)](#) introduisent plusieurs extensions, en ajoutant notamment un effet cohorte γ_{t-x} ainsi qu'un terme quadratique. Le modèle dit M7 est alors décrit par

$$\text{logit } q_{x,t} = \kappa_t^{(1)} + \kappa_t^{(2)} (x - \bar{x}) + \kappa_t^{(3)} ((x - \bar{x}) - \hat{\sigma}_x^2) + \gamma_{t-x}, \quad (0.14)$$

où $\hat{\sigma}_x$ est l'écart-type des âges utilisés dans l'estimation. L'ajout d'un terme quadratique est dû historiquement à l'observation empirique de la surface des logit $q_{x,t}$ pour les données des États-Unis.

En se fondant sur l'observation que les taux de mortalité sont en pratique des données comportant du bruit, une troisième famille de modèle est apparue. Si aucun événement exceptionnel ne se produit, nous pouvons supposer que la surface de mortalité est plutôt lisse dans les dimensions de la période et de l'âge. Ainsi, des méthodes d'analyse fonctionnelle et des techniques de lissage non-paramétrique ont été appliquées à la modélisation de la mortalité. [Currie *et al.* \(2004\)](#) introduisent l'utilisation de P-spline pour ajuster les taux de mortalité. Globalement, ces derniers sont modélisés par :

$$\log m_{x,t} = \sum_{i,j} \theta^{i,j} B_t^{i,j}(x), \quad (0.15)$$

où $B^{i,j}$ sont une base de fonctions cubiques et $\theta^{i,j}$ des paramètres à estimer. Les taux de mortalité sont ainsi lissés sur l'historique des données, permettant d'en extraire les différents chocs. La tendance générale est alors déduite et les taux futurs sont considérés comme des données manquantes.

[Hyndman et Ullah \(2007\)](#) proposent un lissage non-paramétrique $f_t(x)$ de la courbe des taux de mortalité $\log m_{x,t}$ pour chaque période t de l'historique. Ces courbes sont ensuite décomposées par une approche d'analyse fonctionnelle de la manière suivante :

$$f_t(x) = \mu(x) + \sum_{k=1}^K \beta_{t,k} \phi_k(x), \quad (0.16)$$

où $\mu(x)$ est une mesure de position de $f_t(x)$ et $(\phi_k)_{k=1,\dots,K}$ est une base orthonormée de fonctions. La projection des futurs taux de mortalité passe ici par la modélisation univariée de la dynamique temporelle des K séries $\beta_{t,k}$. D'autres modèles dit de lissage ont été développés par la suite (e.g., [Li *et al.*, 2016](#); [Dokumentov *et al.*, 2018](#)), même si cette troisième famille de modèles a rencontré moins de succès que les modèles à facteurs de type Lee-Carter ou CBD dans la pratique actuarielle.

D'autres méthodes plus récentes s'attardent sur la structure de dépendance spatio-temporelle de la mortalité. Ainsi [Christiansen *et al.* \(2015\)](#) utilisent des techniques de statistiques spatiales, le krigeage plus précisément, pour la projection des taux d'amélioration de la mortalité suivant les dimensions âge et période. [Doukhan *et al.* \(2017\)](#) s'appuient aussi sur cette surface

d'amélioration de mortalité et la considèrent comme un champ aléatoire avec une structure causale. La dépendance entre cohortes adjacentes y est modélisée par un AR-ARCH permettant alors de capturer l'effet cohorte et les corrélations entre générations. [Li et Lu \(2017\)](#) appliquent un processus de type Vecteur Autoregressif (VAR) sur la surface des taux de mortalité logarithmiques. Afin d'obtenir un modèle parcimonieux et stationnaire, ils contraignent *a priori* la forme de la matrice de causalité de Granger ([Granger, 1969](#)). Enfin, plus récemment encore, des techniques d'apprentissage statistique ont été appliquées à la modélisation et à la projection de la mortalité. Ainsi des modèles de type Lee-Carter ont été améliorés dans l'estimation de leurs paramètres et la projection de ceux-ci par des méthodes de bagging et de boosting⁴ ([Deprez et al., 2017](#); [Levantesi et Pizzorusso, 2019](#)), ou par des réseaux de neurones artificiels ([Nigri et al., 2019](#)). [Hainaut \(2018\)](#) proposent une approche de projection grâce à des réseaux de neurones directement entraînés sur la surface de mortalité.

Si les sciences actuarielles se sont dotées d'un nombre important de modèles de projection de mortalité, dont nous avons donné ici seulement quelques exemples (voir par exemple [Janssen, 2018](#), pour une revue récente et plus exhaustive de la littérature), nous remarquons qu'une grande partie d'entre eux repose sur l'hypothèse centrale, et commune en statistique, que l'étude des dynamiques passées permettent de déduire la tendance future. Cependant, dans un contexte sanitaire futur toujours plus incertain, les projections de la longévité à long terme sont un exercice extrêmement difficile. Une technique possible est d'ajouter un avis d'expert à la modélisation ([Kamega et Planchet, 2011](#)). Cependant les avis d'experts divergent fortement sur le sujet de l'évolution de la longévité à moyen-long terme ([Debonneuil et al., 2018](#)). Plusieurs scénarios macros s'opposent :

- la vision du transhumanisme présage une accélération exponentielle de la durée de vie humaine grâce aux futurs progrès de la médecine ;
- les modèles que nous avons introduits sous-tendent généralement une augmentation linéaire de la longévité, tout du moins à l'échelle de projection d'une vie humaine ;
- certaines analyses de la mortalité aux âges élevés montrent que la longévité humaine semble atteindre une limite biologique ([Gavrilova et Gavrilov, 2019](#)) ;
- le changement climatique et l'augmentation globale de la pollution de l'environnement peuvent se traduire par une diminution de la longévité dans le futur.

Transfert du risque de longévité

Ainsi, même si les compagnies d'assurances et les fonds de pensions ont à leur disposition de nombreuses méthodes d'évaluation quantitative, l'incertitude reste omniprésente dans la projection de la longévité. Comme nous l'avons déjà noté, le risque de longévité des retraites porte sur des montants colossaux. [Michaelson et Mulholland \(2014\)](#) remarquent d'ailleurs que

4. Nous introduirons dans un deuxième temps les différents algorithmes d'apprentissage statistique sur lesquels notre thèse se focalise. A ce point du manuscrit, le lecteur peut se référer à [Hastie et al. \(2016\)](#) ou [Charpentier et al. \(2018\)](#) pour les détails des méthodologies citées.

le risque de longévité est bien supérieur aux capitaux que le secteur global de l'assurance peut y consacrer de manière réaliste. Ils soulignent alors que les marchés financiers globaux sont la seule source pouvant offrir les montants substantiels nécessaires au risque de longévité. Il est donc nécessaire de transformer le risque de longévité en un actif suffisamment attractif pour drainer d'importants capitaux. C'est dans cette optique que plusieurs outils de transfert de risque de longévité ont été développés depuis le début des années 2000 (e.g., [Blake et Burrows, 2001](#); [Blake et al., 2006](#); [Barrieu et al., 2012](#); [Blake et al., 2018](#)).

Parmi ces instruments de transfert de risques, notons en premier le q -forward ([Coughlan et al., 2007](#)), dont le nom provient de la notation de la probabilité de décès $q_{x,t}$. Il s'agit d'un contrat entre deux parties qui permet d'échanger à maturité un montant proportionnel à la mortalité réalisée d'une certaine population contre un montant proportionnel à une mortalité fixe définie à la signature du contrat. Le fonctionnement du q -forward est schématisé dans la Figure 0.7. Avec cette mécanique un fond de pension peut transférer une partie de son risque de longévité en choisissant de payer le taux variable et en recevant le taux fixe.

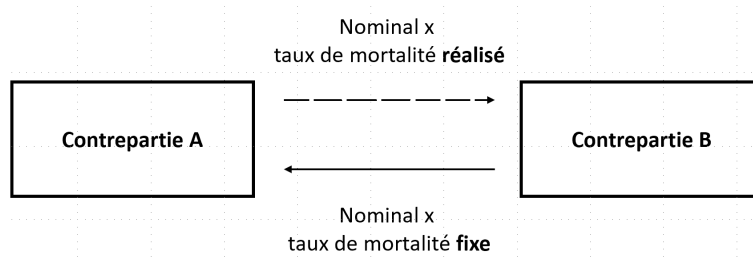


FIGURE 0.7 – Fonctionnement d'un q -forward.

Une extension naturelle du q -forward est le *longevity swap* qui est aussi un contrat entre deux parties. L'une d'entre elle accepte de payer une série de flux monétaires prédéterminés et de recevoir en échange une série de flux calculés sur la mortalité réalisée. La plupart de ces contrats ont une maturité relativement importante. Ainsi, l'un des premiers *longevity swap* s'est conclu en 2008 entre J.P. Morgan et Canada Life pour un nominal de 500 millions de livres sterling et une maturité de 40 ans.

La problématique des deux instruments de transfert de risque de longévité que nous venons de citer est leur maturité. Cette dernière est en effet choisie pour couvrir le risque de longévité pesant sur les retraites, i.e. sur des périodes d'environ 50 ans. Or la majorité des investisseurs présents sur les marchés financiers souhaitent plutôt un horizon de placement n'excédant pas les 15 ans. Par exemple, le point de repère pour les taux d'intérêt en France est l'Obligation Assimilable du Trésor (OAT) à maturité 10 ans, celles de 30 ou 50 ans ne sont pas des indicateurs économiques fortement analysés ni fortement attractives pour le marché. Dans ce sens, [Michaelson et Mulholland \(2014\)](#) décrivent la structuration d'une protection qui a été mise en œuvre entre Aegon et la Société Générale en 2013. Afin de combiner le risque à long terme du fond de pension et la relative courte maturité acceptée, la couverture proposée est

fondée sur deux éléments :

- la mortalité réalisée durant la période à risque ;
- la valeur de l'exposition résiduelle après une ré-estimation du modèle de longévité.

Cette procédure de ré-estimation, schématisée dans la Figure 0.8, implique l'utilisation de plusieurs jeux de données. Les différentes étapes sont alors :

1. le choix d'un modèle de longévité ainsi que l'estimation de ses paramètres à partir des données disponibles au début de la période à risque ;
2. la définition objective, avant le début de la période à risque, du processus de ré-estimation qui sera appliqué ;
3. la ré-estimation du modèle de longévité en utilisant les données supplémentaires obtenues durant la période à risque ;
4. le calcul la valeur de l'exposition résiduelle en utilisant le modèle ré-estimé.

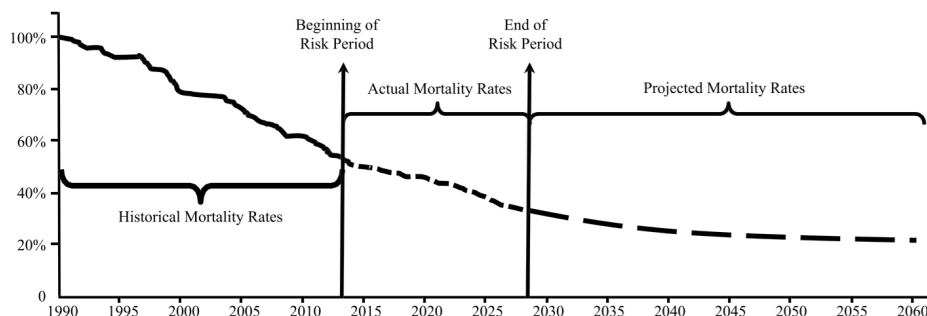


FIGURE 0.8 – Taux de mortalité avant, pendant et après la période à risque.

Source : [Michaelson et Mulholland \(2014\)](#)

Ce dernier instrument de transfert de risques, bien qu'il n'y en ait pas eu un grand nombre de conclus, nous interpelle sur le choix du modèle de longévité. Comme nous l'avons vu, il en existe une multitude mais aucune hiérarchie n'a été clairement établie entre eux vis-à-vis de leur précision, et de nouvelles méthodologies sont développées chaque année. Or, dans ce dernier outil de transfert de risques de longévité, il est nécessaire que toutes les parties en présence, i.e. la contrepartie actuarielle et la contrepartie financière, s'accordent sur le choix d'un unique modèle de longévité. [Blake *et al.* \(2018\)](#) remarquent d'ailleurs que cette problématique, loin d'être spécifique à cet instrument, est un défi pour l'ensemble du marché du transfert du risque de longévité. En effet, à long terme, i.e. à maturité d'un *longevity swap*, un changement de modèle peut conduire à des projections significativement différentes, et donc à des évaluations divergentes de l'instrument de transfert de risques. Pour illustrer ce point, nous projetons les taux de mortalité d'un homme français âgé de 50 ans en 2016 sur 50 ans, i.e. jusqu'à ses 100 ans. Pour ce faire nous utilisons les données de la HMD ([HMD, 2019](#)) ainsi que le package *StMoMo* ([Villegas *et al.*, 2017b](#)) sous R pour l'estimation et la projection des modèles. Ces derniers sont estimés sur les données historiques allant de 1960 à 2016 sur la

plage d'âge 50-100. Nous comparons le Lee-Carter (LC) (Lee et Carter, 1992), l'Age-Période-Cohorte (APC) (Currie, 2006), le Cairns-Black-Dowd (CBD) à deux facteurs (Cairns *et al.*, 2006) et le M7 (Cairns *et al.*, 2009). Si les projections des taux de mortalité, représentées sur la Figure 0.9, sont relativement proches pendant les 15 premières années, à long terme nous observons une nette différenciation. Cette dernière est d'autant plus visible si nous comparons la probabilité de survie entre 51 et 101 ans estimée suivant les différents modèles : elle est de 13,3% pour l'APC, 8,2% pour le CBD, 3% pour le LC et seulement de 0,5% pour le M7.

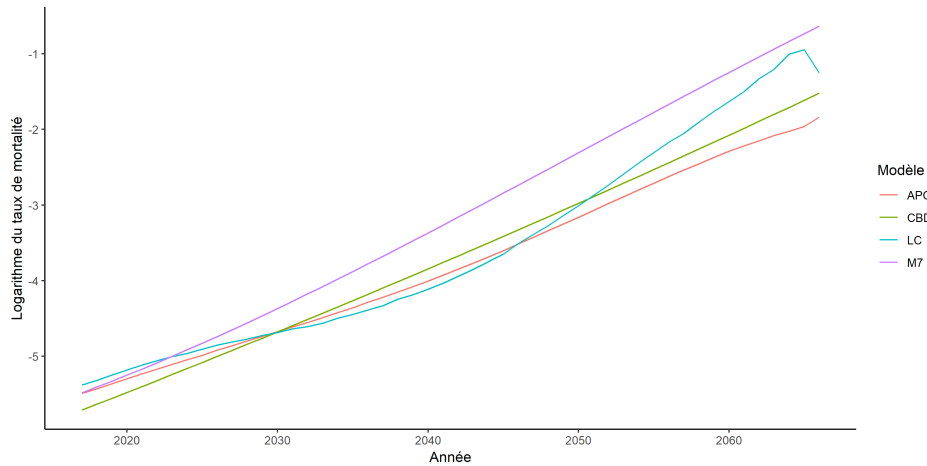


FIGURE 0.9 – Projection sur 50 ans des taux de mortalité des hommes français de la cohorte d'âge 50 ans en 2016 selon plusieurs modèles de mortalité.

Hétérogénéité de la mortalité

La nécessité de déterminer un modèle de référence pour tous les participants soulève la problématique des différences dans les dynamiques de la longévité entre les populations. Nous en avons vu un exemple en comparant les taux d'amélioration de la mortalité de la France et de l'Angleterre et pays de Galles. Dans la seconde population nous pouvons voir apparaître des effets cohortes fortement marqués, au contraire de la population française. Ainsi, l'actuaire choisirait sûrement un modèle incluant un facteur cohorte pour l'Angleterre, mais pas nécessairement pour la France. Si un tel modèle de référence est défini, il se doit donc d'être capable de s'adapter aux spécificités de chaque population. Les modèles les plus utilisés, i.e. de type Lee-Carter ou CDB, tendent parfois à se spécialiser pour une population de par leur construction historique. Par exemple le terme quadratique $\kappa_t^{(3)} ((x - \bar{x}) - \hat{\sigma}_x^2)$ du M7 "a été inspiré par la possibilité d'une certaine courbature identifiée dans les représentations graphiques du logit $q(t, x)$ pour les données US" (Cairns *et al.*, 2009).

Cette différence entre les populations nous renvoie alors à la question de l'hétérogénéité en terme de mortalité que nous avons déjà abordée rapidement en constatant les différences entre les sous-populations définies par le sexe. Pratiquement, pour une compagnie d'assurances ou un fonds de pension particulier, la population à risque est en général différente de la population

agrégée du pays en terme de mortalité. Ainsi lorsqu’une compagnie d’assurances utilise les tables de mortalité nationales pour l’évaluation de ses passifs, il y a un risque que les taux de mortalité soient significativement biaisés par rapport à ceux de la population à risque, i.e. celle du portefeuille de la compagnie. De même si un fonds de pension conclut un *longevity swap* dont le taux de mortalité de référence est celui de la population nationale, il y a un risque pour que la couverture ne soit pas parfaite car non alignée sur la mortalité réellement expérimentée par la population à risque. Ce risque de base est un sujet qui a récemment gagné en intérêt dans la littérature (e.g., [Li et Hardy, 2011](#); [Coughlan et al., 2011](#); [Villegas et al., 2017a](#); [Salhi et Loisel, 2017](#)). Dans le cas des transferts de risques aux marchés financiers, il peut cependant être économiquement intéressant pour les fonds de pensions d’accepter de garder une partie de ce risque de base. En effet, l’utilisation d’indice de longévité standardisé plutôt que la mortalité d’une population spécifique permet en général une couverture certes moins efficace mais moins chère. [Blake et al. \(2018\)](#) soulignent que, dans l’objectif du développement d’un marché de la longévité viable et efficace, les participants se doivent de trouver un arbitrage optimal entre le risque de base d’un indice standard et le risque de liquidité d’un indice fait sur mesure. Néanmoins, il convient dans tous les cas d’évaluer et le plus précisément possible le risque de base.

Modèles multi-population

Pour quantifier ce risque, de nombreux modèles stochastiques de mortalité à deux populations ont été développés. Ils permettent de projeter et de comparer les dynamiques de mortalité de différentes populations tout en prenant en compte les potentielles dépendances entre ces dernières. La problématique revient à rajouter la dimension population dans la modélisation, nous nous intéressons donc maintenant aux taux de mortalité $m(x, t, i) = m_{x,t}^i$ des individus d’âge x durant l’année calendaire t pour la population i . La majorité des modèles multi-population utilisés actuellement peuvent être vus comme des généralisations des modèles de mortalité à une population que nous avons introduits précédemment. Ainsi [Li et Lee \(2005\)](#) proposent une extension du Lee-Carter en y ajoutant une tendance temporelle commune K_t aux populations modélisées :

$$\log m_{x,t}^i = \alpha_x^i + B_x K_t + \beta_x^i \kappa_t^i, \quad (0.17)$$

où α_x^i , β_x^i et κ_t^i sont des effets âge et période spécifiques à chaque population, et B_x un effet âge global. Ces séries de paramètres sont estimées en suivant une procédure en deux étapes basée sur la décomposition en valeur singulière : d’abord les effets communs B_x et K_t sont obtenus sur les données agrégées, puis les effets spécifiques de chaque pays sont calculés. Comme dans le cas du Lee-Carter, des contraintes doivent être imposées pour des raisons d’identifiabilité des paramètres :

$$\begin{aligned} \sum_t K_t = 0 \text{ et } \sum_x B_x = 1 \text{ pour les termes communs,} \\ \sum_t \kappa_t^i = 0 \text{ et } \sum_x \beta_x^i = 1 \text{ pour chaque } i. \end{aligned} \tag{0.18}$$

Les dynamiques des séries temporelles K_t et κ_t^i sont ensuite modélisées indépendamment, et notamment de telle sorte que les κ_t^i tendent vers une constante à long terme. Une hypothèse fondamentale de ce modèle est la notion de cohérence dans les projections. Celle ci suppose qu'à long terme, les taux de mortalité des différentes populations ne peuvent diverger indéfiniment, ce qui n'est pas raisonnable d'un point de vue biologique. Bien que cette hypothèse soit parfois critiquée (par exemple [Li et al., 2017](#), proposent une hypothèse plus faible de semi-cohérence), la majorité des modèles existant repose sur cette hypothèse centrale.

Alors que le modèle introduit par [Li et Lee \(2005\)](#) s'assure de la cohérence par la présence d'un effet période commun K_t , d'autres extensions du Lee-Carter à la multi-population proposent un effet âge commun afin de vérifier la propriété de cohérence. Ainsi [Zhou et al. \(2014\)](#) présentent la modélisation suivante :

$$\log m_{x,t}^i = \alpha_x^i + \beta_x \kappa_t^i, \tag{0.19}$$

où l'effet âge β_x est commun à toutes les populations considérées. Notons que [Kleinow \(2015\)](#) propose parallèlement un modèle similaire où le terme $\beta_x \kappa_t^i$ est remplacé par $\sum_p \beta_{x,p} \kappa_{t,p}^i$ plus général. Pour atteindre la cohérence il ne suffit plus que de s'assurer que la différence $\kappa_t^1 - \kappa_t^2$ soit modélisée de telle sorte qu'il y ait un retour à la moyenne. [Zhou et al. \(2014\)](#) soumettent alors trois dynamiques temporelles, dont le vecteur autorégressif (VAR) selon lequel nous avons :

$$\begin{aligned} \Delta \kappa_t^1 &= \phi_0 + \phi_1 \Delta \kappa_{t-1}^1 + \phi_2 \Delta \kappa_{t-1}^2 + \epsilon_t^1, \\ \Delta \kappa_t^2 &= \theta_0 + \theta_1 \Delta \kappa_{t-1}^1 + \theta_2 \Delta \kappa_{t-1}^2 + \epsilon_t^2, \end{aligned} \tag{0.20}$$

où $\Delta \kappa_t^i = \kappa_t^i - \kappa_{t-1}^i$ est le changement de la mortalité au niveau global dans la population i entre $t-1$ et t ; $\phi_0, \phi_1, \phi_2, \theta_0, \theta_1$ et θ_2 sont des paramètres du modèle à estimer; ϵ_t^1 et ϵ_t^2 sont des chocs d'innovations suivant une loi normale bivariée centrée avec une matrice de covariance constante. L'obtention de la cohérence se fait alors en imposant la contrainte suivante dans le processus d'estimation :

$$\frac{\phi_0}{1 - \phi_1 - \phi_2} = \frac{\theta_0}{1 - \theta_1 - \theta_2}. \tag{0.21}$$

Le risque de base est souvent rencontré entre une population spécifique de taille relativement petite par rapport à une population de référence à l'échelle nationale. Ainsi, plusieurs modèles de mortalité en multi-population font l'hypothèse qu'il existe *a priori* une population dominante, la plus grande, et une population dominée dont la mortalité suit alors un processus stochastique de retour à la moyenne. C'est le cas par exemple du modèle GRAVITY proposé par Dowd *et al.* (2011) qui est une extension du modèle Age-Période-Cohorte, ou du similaire modèle SAINT décrit par Jarner et Kryger (2011). De nombreux autres modèles multi-population ont récemment été développés. Nous pouvons y retrouver d'autres variantes d'extensions du Lee-Carter (e.g., Danesi *et al.*, 2015; Enchev *et al.*, 2016), mais aussi des modèles qui ajoutent des effets cohortes (e.g., Villegas et Haberman, 2014; Yang *et al.*, 2016), ou encore des extensions de modèle de type CBD (e.g., Hahn, 2014; Li *et al.*, 2015). Villegas *et al.* (2017a) proposent une revue récente de la littérature sur ce sujet.

Au-delà de l'évaluation du risque de base, les modèles de multi-population sont aussi aujourd'hui utilisés pour améliorer la projection de la mortalité pour une population spécifique. En effet, ils permettent de prendre en compte des informations sur des populations avec des caractéristiques similaires. Ainsi, Antonio *et al.* (2017) utilisent un modèle de type Li et Lee (2005) pour projeter la mortalité des Pays-Bas et de la Belgique en s'appuyant sur les données agrégées d'un ensemble de 14 pays européens. L'idée sous-jacente n'est pas sans rappeler la théorie de la crédibilité (Bühlmann et Gisler, 2006) et ses applications dans l'évaluation de la mortalité (e.g., Hardy et Panjer, 1998; Salhi *et al.*, 2016; Salhi et Thérond, 2018). La décomposition d'une population hétérogène en plusieurs sous-populations plus homogènes, dont les dynamiques de mortalité sont ensuite étudiées au sein d'un modèle multi-population, peut aussi apporter de l'information et une meilleure précision dans les projections (Danesi *et al.*, 2015).

Cette prise en compte des relations de dépendance entre les populations reste cependant un défi pour la modélisation mathématique de la mortalité. En effet, elle rajoute une dimension supplémentaire à la structure, déjà complexe, de la dynamique de longévité en âge, période et cohorte. L'estimation d'un système complexe de dépendance implique un grand nombre de paramètres. Ainsi, nombre de modèles ne sont d'ailleurs seulement conçus que pour deux populations, ou imposent des contraintes *a priori* sur le système d'interdépendance comme par exemple la présence d'une population dominante. De plus, la profondeur des historiques des données est limitée : il est peu fréquent d'estimer les paramètres d'un modèle avec les données d'avant 1950, notamment à cause des changements importants qu'ont impliqués les deux Guerres mondiales. La problématique de la grande dimension nous apparaît alors d'autant plus centrale dans le cas de la multi-population.

Risque de Rachat

Dans le cas de la longévité, le risque repose entièrement sur la variable de la durée de vie humaine de l'assuré qui est aléatoire à la fois du point de vue de la compagnie d'assurances mais aussi de l'individu qui reçoit la rente viagère. Ce dernier peut essayer d'augmenter son espérance de vie en adoptant un mode de vie plus sain avec par exemple la pratique d'une activité sportive et la non consommation de tabac. Cependant, le comportement de l'assuré, si nous excluons le suicide, ne peut pas avoir d'influence certaine sur sa durée de vie qui reste une variable aléatoire. Au contraire, le risque de rachat repose sur l'asymétrie de choix possible entre l'assuré et la compagnie d'assurances. Le contrat d'assurance vie prévoit en effet que la compagnie garantit définitivement le versement de prestations en fonction d'événements viagers de l'assuré : son décès ou sa survie. En retour le souscripteur s'engage à s'acquitter des cotisations, mais cet engagement est révocable : il peut choisir à tout moment de racheter son contrat avant l'échéance, et ainsi récupérer son épargne en totalité (rachat total) ou seulement en partie (rachat partiel).

L'incertitude de la compagnie d'assurances sur le comportement de rachat de son portefeuille d'assurés a un fort impact sur l'anticipation de ses flux financiers futurs, et donc constitue un risque économique important pour l'entreprise au niveau de sa rentabilité ou même de sa solvabilité. Le rapport QIS 5 (EIOPA, 2011) souligne d'ailleurs que ce risque de rachat est le risque le plus important selon la mesure décrite par la directive Solvabilité II. Par exemple, des rachats précoces peuvent compromettre la capacité de l'assureur à recouvrer ses frais d'acquisition et de gestion du contrat (Tsai *et al.*, 2009; Pinquet *et al.*, 2011). La compagnie prévoit de rentabiliser ses frais par les profits obtenus tout au long de la vie du contrat. Une mauvaise projection des rachats, et donc des flux financiers sortants à dégager par l'assureur, peut entraîner une inadéquation avec la stratégie d'investissement de la compagnie. L'efficacité de la gestion actif-passif s'en retrouve alors impactée (Kim, 2005b; Eling et Kochanski, 2013). Dans le cas d'un scénario de rachat de masse, la société fait face à un grand nombre de demandes de rachat et peut manquer de liquidité, de la même façon où un grand nombre de clients d'une banque viendraient retirer leur argent dans une période relativement courte. Elle doit alors vendre ses actifs à un moment qui n'est pas forcément optimal d'un point de vue financier.

La dynamique des taux d'intérêt joue un rôle important dans le risque de rachat. Par exemple nous connaissons actuellement une période de taux historiquement faible : le taux OAT à 10 ans de la France est de $-0,19\%$ au 1^{er} août 2019 ! Ainsi les sociétés d'assurances vie ne proposent pas de souscription avec des taux garantis élevés. Un scénario de hausse des taux serait alors théoriquement bénéfique aux assureurs par rapport à ces dernières garanties, car les actifs auraient un rendement supérieur aux engagements pris. Une augmentation des rachats sur ces contrats diminuerait donc les profits futurs que la compagnie d'assurances pourrait obtenir. Au contraire, actuellement un grand nombre d'anciens contrats d'assurance vie ont des taux

garantis élevés en comparaison du rendement des actifs. A l'époque de la souscription les taux étant plus hauts, par exemple celui de l'OAT 10 ans était de 9,22% au 1^{er} août 1991, les assureurs n'hésitaient pas à garantir des taux de capitalisation plus importants pour l'épargne de leurs assurés. Dans le contexte économique actuel, le risque de rachat est donc un risque de baisse du nombre de rachats au sein de ces contrats. Ces derniers sont en effet déficitaires pour l'assureur qui doit continuer à tenir ses engagements.

Ainsi les rachats peuvent être un risque pour la compagnie d'assurances dès qu'ils dévient des projections effectuées par les actuaires, que cela soit à la hausse ou à la baisse, et sur lesquelles la stratégie d'actif-passif repose en grande partie. Les trois scénarios de chocs prévus par Solvabilité II pour l'estimation du SCR (*Solvency Capital Requirement*) sont une hausse durable du taux de rachat, une baisse durable du taux de rachat et un scénario de rachat de masse. La compagnie gardant alors le scénario le plus défavorable pour le calcul final du capital réglementaire. Ceci montre l'importance de la modélisation et la projection les plus précises possibles de la dynamique des taux de rachat, d'autant plus lorsque l'on connaît les sommes colossales que l'assurance vie représente. Il n'est donc pas étonnant qu'une large littérature se soit développée sur ce sujet (e.g., [Eling et Kochanski, 2013](#); [Bauer et al., 2017](#))

Modélisation des rachats

Dans la littérature actuarielle, la modélisation des causes de rachat se fonde principalement sur deux théories : l'hypothèse des taux d'intérêt ([Pesando, 1974](#); [Dar et Dodds, 1989](#)) et l'hypothèse du besoin urgent de ressources ([Outreville, 1990](#)). Dans cette seconde approche, les fonds de l'assurance vie sont rachetés dans le but de faire face un événement coûteux dans la vie personnelle de l'assuré. Par exemple ils peuvent permettre de surmonter des difficultés financières rencontrées lors d'une période de chômage, ou de financer l'achat d'un bien immobilier. Ainsi, les caractéristiques particulières de l'assuré, qui peuvent être influencées par le cycle économique, sont intuitivement des causes probables de rachat. L'approche des taux d'intérêt repose sur l'hypothèse que les assurés sont des agents de marché rationnels. Si les taux d'intérêt du marché augmentent, des opportunités d'arbitrage apparaissent. L'assuré rationnel, souhaitant maximiser ses gains, résilie alors son contrat d'assurance vie afin d'obtenir de meilleurs rendements financiers.

Dans cette continuité de l'assuré rationnel et d'un monde risque neutre, un grand nombre d'études s'attardent sur l'évaluation financière du rachat d'assurances vie. Dans cette littérature, le rachat est vu comme une option (par exemple américaine, bermudéenne ou encore quanto asiatique) dont l'assuré cherche le temps d'arrêt optimal pour maximiser son épargne. Plusieurs modèles financiers ont été appliqués à l'évaluation de cette option de rachat pour différents types de contrats d'assurance vie. A titre d'exemples, nous pouvons citer :

- le modèle de Cox-Ross-Rubinstein qui se fonde sur des arbres binomiaux ([Grosen et Løchte Jørgensen, 2000](#); [Bacinello, 2003, 2005](#)) ;

- le modèle de Longstaff-Schwartz qui repose sur un algorithme des moindres carrés Monte Carlo (Nordahl, 2008; Kling *et al.*, 2014);
- l'utilisation d'équation différentielle partielle (Jensen *et al.*, 2001; Bauer *et al.*, 2006; MacKay *et al.*, 2017);
- l'analyse de portefeuille répliquant (Consiglio et Giovanni, 2010).

Malgré leur grande utilité pour la compréhension de la dynamique de rachat, ces modèles d'évaluation d'option de rachat reposent sur des hypothèses qui ne reflètent pas toujours la réalité. Des améliorations à ce cadre théorique ont récemment été proposées pour mieux prendre en compte les comportements des assurés. Par exemple, la question de la fiscalité spécifique des contrats d'assurance vie, qui est une des raisons de leur popularité, impacte fortement l'évaluation *subjective* que l'assuré se fait de son épargne. Moenig et Bauer (2016) proposent alors une méthodologie d'évaluation risque neutre *subjective* en prenant en compte les différences de fiscalité entre les choix d'investissement pour plus de réalisme. De même la rationalité des assurés peut être restreinte (Li et Szimayer, 2014) pour mieux représenter le comportement des assurés.

En parallèle de ces cadres théoriques stochastiques, de nombreuses études empiriques ont été menées pour analyser le rachat en assurance vie d'un point de vue statistique. Cette littérature peut être séparée en deux parties suivant les variables explicatives étudiées. La première branche des recherches s'intéresse à l'impact de facteurs conjoncturels, tels que le contexte économique (e.g. l'évolution des taux d'intérêts, le taux de chômage, la croissance, le rendement des marchés) les caractéristiques de la compagnie d'assurances (e.g. son âge, son rating, sa taille, sa forme juridique) ou encore des variables de concurrence commerciale (e.g. le spread de taux de rendement, les frais de rachat). L'étude des rachats se fait alors d'un point de vue macro. À l'inverse, nous trouvons des études analysant le rachat à un niveau micro, i.e. en se focalisant sur des facteurs explicatifs structurels. Dans cette seconde partie de la littérature, les variables d'intérêt incluent généralement les caractéristiques de l'assuré (e.g. le sexe, la catégorie socio-professionnelle, l'âge), le type de produit (e.g. fonds euros, unités de compte, possibilité d'arbitrage) ou les caractéristiques du contrat (e.g. son ancienneté, la fréquence de la prime, le réseau de distribution).

Dans les études statistiques conjoncturelles des rachats, nous retrouvons logiquement l'utilisation de régressions temporelles classiques pour tester les hypothèses sur les conditions de rachat. Par exemple Dar et Dodds (1989) et Outreville (1990) présentent des résultats validant l'hypothèse du besoin urgent de ressources en utilisant respectivement des données du Royaume-Uni d'une part, et des Etats-Unis et Canada d'autre part. En utilisant les travaux sur la cointégration d'Engle et Granger (1987), d'autres études se proposent de séparer la modélisation de la dynamique court-terme et les potentielles relations de long-terme des rachats en fonction des variables économiques. Ainsi Tsai *et al.* (2002) et Kuo *et al.* (2003) montrent des relations de long-terme significatives entre les taux de rachat et les taux d'intérêt, vali-

dant alors l'hypothèse éponyme. De potentielles relations entre les rachats d'assurances vie et d'autres variables conjecturales ont aussi été examinées par l'utilisation de modèles linéaires généralisés, introduits par [Nelder et Wedderburn \(1972\)](#). A ce titre nous pouvons citer [Kim \(2005a\)](#) et [Kiesenbauer \(2012\)](#) qui utilisent la régression logistique, ou [Cox et Lin \(2006\)](#) qui proposent l'application d'un modèle Tobit.

L'utilisation des modèles linéaires généralisés (GLM) pour l'étude statistique des rachats fait écho aux travaux précurseurs de [Renshaw et Haberman \(1986\)](#) qui appliquent une régression logistique sur les caractéristiques particulières des assurés et des produits à partir des données de sept compagnies d'assurances vie en Écosse pour l'année 1976. Leur analyse met en avant quatre facteurs importants segmentant le risque de rachat : l'âge à la souscription, la compagnie ayant émis le contrat, l'ancienneté du contrat et le type de contrat. La problématique de segmentation du risque de rachat grâce à des facteurs structurels par l'utilisation de GLM au sein d'un portefeuille de contrats d'assurance vie a regagné en intérêt à partir de la fin des années 2000 (e.g., [Kagraoka, 2005](#); [Cerchiara et al., 2008](#); [Eling et Kiesenbauer, 2014](#)). Depuis, le spectre des techniques statistiques appliquées s'est agrandi. Ainsi, dans la suite des GLM, [Milhaud \(2013\)](#) propose l'application de mélange de régressions logistiques pour la prédiction des rachats. [Milhaud et al. \(2011\)](#) comparent la régression logistique avec les arbres de classification et de régression (CART), introduits par [Breiman et al. \(1984\)](#), pour la détection des variables explicatives des rachats. Très récemment, des méthodes d'apprentissage statistique plus sophistiquées, comme par exemple le boosting et le bagging, ont été utilisées pour à la modélisation du risque de rachat en fonction de facteurs structurels ([Jamal, 2017](#); [Aleandri, 2017](#)).

Risque d'attrition

Ces modèles statistiques et leur contexte d'utilisation, i.e. la segmentation du risque de rachat en fonction de facteurs sectoriels, nous renvoient à une discipline *a priori* éloignée des sciences actuarielles : le marketing quantitatif. Plus exactement, nous nous intéressons à la littérature concernant le risque d'attrition (*churn* en anglais). Prenons le temps d'une courte analogie. Le rachat est le fait qu'un assuré résilie son assurance vie afin de récupérer son épargne. En remplaçant la terminologie actuarielle, "assuré" et "assurance vie", par la terminologie plus généraliste du marketing, respectivement "client" et "contrat avec l'entreprise", nous obtenons une définition de l'attrition. La définition générale d'attrition est néanmoins encore plus large, et se focalise plutôt sur l'arrêt des relations client-entreprise, afin d'englober les secteurs où le terme de contrat n'est pas forcément pertinent. Cette littérature s'est attardée sur le risque d'attrition dans différents secteurs tels que la télécommunication, la banque de détail, la presse écrite, le jeu vidéo, etc. Le milieu assurantiel, non-vie et santé essentiellement, est aussi impacté par cette problématique que les actuaires commencent à s'approprier (e.g., [Günther et al., 2014](#); [Gerber et al., 2018](#))

En conséquence des similitudes dans les risques de rachat et de l'attrition, il est cohérent de constater des réponses méthodologiques similaires. Ainsi les modèles de prédictions de la littérature du marketing quantitatif analysent aussi des facteurs structurels comme les caractéristiques des contrats et des clients afin de prévoir le comportement de ces derniers, i.e. les futures résiliations. Les techniques d'apprentissage statistique employées dans les modèles de prédictions d'attrition sont aussi logiquement semblables à ceux de la littérature sur les rachats et de la détermination de leurs facteurs structurels. Les méthodes quantitatives appliquées pour l'attrition forment d'ailleurs un éventail historiquement plus large d'algorithmes qui inclue notamment :

- la régression logistique (Eiben *et al.*, 1998; Hadden *et al.*, 2006);
- les arbres de décisions (Mozer *et al.*, 2000; Hwang *et al.*, 2004);
- le bagging et les forêts aléatoires (Larivière et Van den Poel, 2005; Burez et Van den Poel, 2007);
- le boosting et le gradient boosting (Lemmens et Croux, 2006; Burez et Van den Poel, 2009)
- les machines à vecteurs de support (SVM⁵) (Coussement et Van den Poel, 2008; Xia et Jin, 2008)
- les réseaux de neurones artificiels (Hung *et al.*, 2006; Tsai et Lu, 2009);
- la classification naïve bayésienne (Nath et Behara, 2003; Huang *et al.*, 2012).

Malgré leurs ressemblances méthodologiques apparentes, il existe néanmoins une différence significative entre l'analyse des comportements de rachat et d'attrition : l'objectif final. Dans le cas des sciences actuarielles, les études sur les rachats servent essentiellement pour la compagnie d'assurances à évaluer ses passifs, ajuster sa stratégie de gestion actif-passif et estimer son besoin en capital réglementaire. L'actuaire se place dans une situation passive par rapport au rachat aléatoire de l'assuré. Au contraire, la vision du marketing quantitatif est plus active. La prédiction de l'attrition est un des piliers principaux de gestion de la relation client (CRM⁶), et consiste seulement en la première étape de l'objectif final qu'est la rétention du client.

Customer lifetime value

En effet, les clients sont modélisés comme des actifs dans la vision marketing (Gupta et Lehmann, 2003), il est donc logique de vouloir les conserver pour en obtenir le maximum de profit. Comme tout actif, il s'agit alors d'évaluer leur valeur pour une gestion des risques efficaces. En ce sens, la valeur vie client (CLV⁷) est définie comme la valeur actuelle de tous les futurs profits obtenus d'un client durant la vie de sa relation avec l'entreprise (Berger et Nasr, 1998; Gupta *et al.*, 2004, 2006). Mathématiquement, la CLV pour un client peut être

5. *Support Vector Machine*

6. *Customer Relationship Management*

7. *Customer lifetime value*

estimée par

$$CLV = \sum_{t=0}^{\infty} \frac{(p_t - c_t) r_t}{(1 + i)^t} - AC, \quad (0.22)$$

où p_t et c_t représentent respectivement les prix payés par le client et les coûts de service pour le client au temps t ; i est le taux d'actualisation; r_t est la probabilité de survie de la relation client-entreprise, i.e. sa probabilité cumulative de rétention au temps t ; AC exprime le coût d'acquisition. Dans ce cas général, nous nous intéressons à la valeur obtenue sur l'ensemble de la relation client-entreprise. Suivant le contexte, des variantes de la CLV peuvent être préférables comme le choix d'un horizon temporel fini T , ou la non prise en compte du coût d'acquisition AC lorsque l'étude s'attarde sur les clients déjà en portefeuille, par exemple dans le cas de l'attrition. Le concept de CLV a reçu un grand intérêt de la part des chercheurs et des professionnels, notamment puisque cette métrique fournit des informations pertinentes sur la valeur financière de l'entreprise (Gupta et Lehmann, 2006). D'ailleurs nous noterons les fortes ressemblances entre la CLV et la méthode des flux de trésorerie actualisés (DCF⁸) utilisée pour la valorisation d'une entreprise.

Avec la définition de la CLV nous remarquons encore plus l'importance de prédire correctement l'attrition. Il est donc crucial pour les responsables CRM d'avoir à disposition le meilleur modèle de détection d'attrition possible. Les techniques d'apprentissage statistique appliquées à la détection d'attrition que nous avons citées sont des modèles de classification binaire supervisée : la variable d'intérêt Y représentant si l'attrition a lieu ou non. L'application d'une telle méthodologie amène à présenter les résultats sous la forme d'une matrice de confusion (cf. Table 0.1).

TABLE 0.1 – Matrice de confusion.

| | | Prédite | |
|----------|---------------|---------------------|---------------------|
| | | Attrition | Non Attrition |
| Réalisée | Attrition | Vrais Positifs (TP) | Faux Négatifs (FN) |
| | Non Attrition | Faux Positifs (FP) | Vrais Négatifs (TN) |

Dans le cadre général des statistiques, de nombreuses mesures ont été développées, afin de comparer ces matrices d'un modèle à l'autre. Certaines des plus communes incluent par exemple la précision, i.e. la proportion de bonnes classifications parmi l'ensemble des éléments classifiés comme pertinents,

$$précision = \frac{TP}{TP + FP}, \quad (0.23)$$

8. *Discounted cash flows*

le rappel, i.e. la proportion de bonnes classifications parmi l'ensemble des éléments pertinents,

$$rappel = \frac{TP}{TP + FN}, \quad (0.24)$$

ou encore la précision totale (*accuracy* en anglais) qui mesure la proportion total de bonnes classifications

$$précision\ totale = \frac{TP + TN}{TP + FP + TN + FN}. \quad (0.25)$$

Devant certains biais que ces mesures simples peuvent présenter (Powers, 2011), d'autres méthodologies d'évaluation de classificateur binaire ont pris de l'importance ces dernières années. Notamment, les utilisations de la courbe ROC (*receiver operating characteristic*), qui analyse le taux de vrais positifs en fonction du taux de faux positifs, et de l'AUC (*area under curve*), i.e. l'aire en dessous de la courbe ROC, sont de plus en plus répandues. Ainsi, nous retrouvons naturellement des études de comparaison des modèles de prédiction d'attrition reposant sur ces différentes mesures (e.g., Kumar et Ravi, 2008; Xia et Jin, 2008; Vafeiadis et al., 2015).

Néanmoins l'utilisation des modèles de prédiction d'attrition se fait généralement dans un contexte économique pratique bien spécifique : celui des campagnes marketing de rétention. L'objectif final est d'augmenter la CLV du portefeuille client de l'entreprise. Il est donc raisonnable de souhaiter comparer les différents classificateurs non pas d'un point de vue purement statistique, mais à l'aide d'une mesure permettant d'évaluer la performance économique de l'utilisation effective des modèles. Dans ce contexte, Neslin et al. (2006) proposent une mesure de rentabilité de la gestion de l'attrition en estimant la variation de la CLV d'une campagne de rétention. Le profit Π de cette campagne marketing est défini par

$$\begin{aligned} \Pi &= N\alpha [\beta\gamma (CLV - c - \delta) - \beta(1 - \gamma)c - (1 - \beta)(c + \delta)] - A \\ &= N\alpha [\beta[\gamma CLV + \delta(1 - \gamma)] - \delta - c] - A, \end{aligned} \quad (0.26)$$

où N est le nombre total de clients ; α la fraction des clients ciblés par la campagne de rétention ; β la fraction des clients ciblés qui auraient été touchés par l'attrition (*would-be churners*), i.e. qui auraient arrêté leur relation avec l'entreprise ; δ le coût pour l'entreprise de l'incitation client ; γ la fraction des clients ciblés *would-be churners* qui décident de rester à cause de l'incitation, i.e. le taux de succès de l'incitation ; c le coût de contact du client pour lui offrir l'incitation ; CLV la valeur vie client, i.e. la valeur pour l'entreprise si le client est retenu ; A les coûts administratifs fixes de la campagne de rétention. Cette modélisation de la campagne de rétention est représentée schématiquement dans la Figure 0.10. Ce cadre a

permis le développement de mesures économiques spécifiques à la problématique de la détection d'attrition pour comparer les modèles mis en place : le meilleur est alors celui produisant le plus de profits pour l'entreprise (e.g., Burez et Van den Poel, 2007; Verbeke *et al.*, 2012; Tamaddoni *et al.*, 2016).

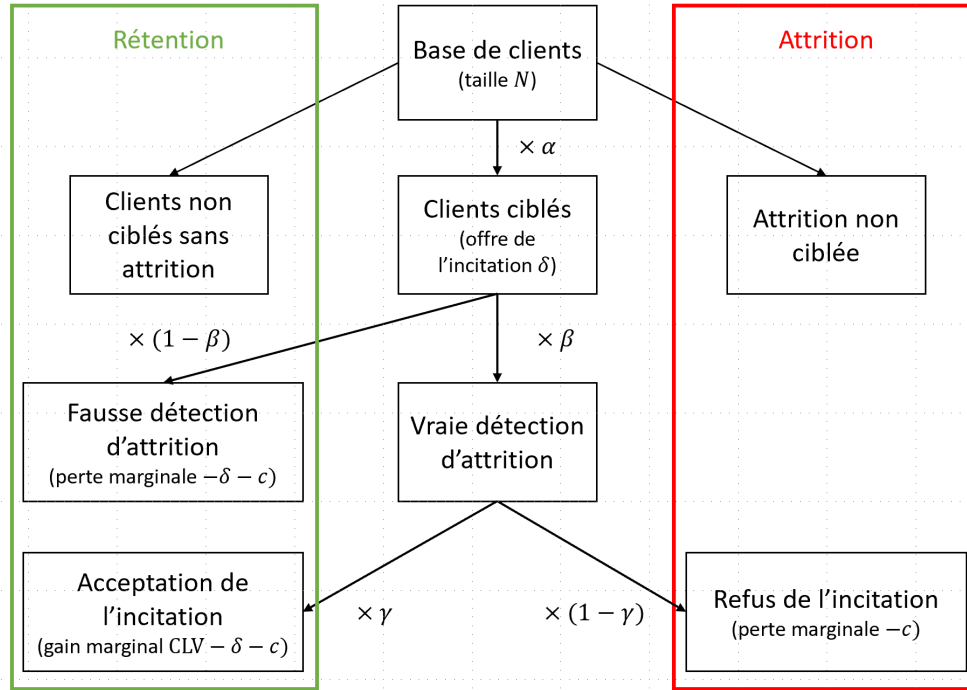


FIGURE 0.10 – Représentation schématique des gains d'une campagne de rétention.

Optimisation de l'approche

Prenons à présent un peu de recul sur les méthodes d'apprentissage statistique qui sont appliquées. La problématique est de trouver le meilleur modèle suivant une mesure d'évaluation spécifique déterminée par le praticien. Or les algorithmes employés sont mathématiquement conçus pour minimiser la mauvaise classification. Ceci est du à leur vraisemblance ou à leur fonction de perte. Or il est connu depuis longtemps que le choix de la fonction de perte pour l'estimation des paramètres est cruciale et définit implicitement le modèle (Engle, 1993; Christoffersen et Jacobs, 2004). Prenons l'exemple simple de la régression logistique, étant donné un échantillon d'apprentissage $\{y_i, \mathbf{x}_i\}_1^N$ où $\mathbf{x}_i \in \mathbb{R}^n$ et $y_i \in \{0, 1\}$, le modèle est spécifié de la façon suivante :

$$\ln \left(\frac{\mathbb{P}[Y = 1 | X = \mathbf{x}]}{\mathbb{P}[Y = 0 | X = \mathbf{x}]} \right) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}, \quad (0.27)$$

où les paramètres $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^n$ sont estimés par maximisation de la vraisemblance

$$\mathcal{L} = \prod_{i=1}^N \left(\frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)^{1-y_i}. \quad (0.28)$$

En d'autres termes, les modèles de prédiction d'attrition sont entraînés de façon à bien classer les clients qui vont arrêter leur relation avec l'entreprise et ceux qui vont la continuer ; mais ils ignorent complètement l'aspect économique de la campagne de rétention comme par exemple la valeur spécifique de chaque client pour l'entreprise, le coût économique d'une mauvaise classification, le profit potentiel d'une rétention réussie, etc. Le problème à maximiser n'est pas équivalent dans les deux cas, et par conséquent, la solution optimale de l'un n'est *a priori* pas celui de l'autre. Quelques réponses ont été proposées par certains chercheurs dans le cas de la détection de l'attrition. [Glady et al. \(2009\)](#), en redéfinissant l'attrition comme un changement de comportement du client diminuant le profit, proposent un algorithme de boosting dont la fonction de perte est définie à partir du gain en CLV obtenu par une action de rétention. Plus récemment, [Lemmens et Gupta \(2017\)](#) introduisent une fonction de perte fondée sur le profit ainsi qu'une méthodologie pour son application aux différents modèles de classification. Ils montrent alors empiriquement que le gain en profit peut atteindre 62% pour une campagne de rétention.

Dans cette volonté d'optimiser une campagne de rétention, d'autres paramètres restent à étudier plus précisément, et notamment tout ce qui concerne l'hétérogénéité des clients dans leur comportement. Le modèle proposé par [Neslin et al. \(2006\)](#) ne fait pas clairement apparaître cette notion dans son écriture mathématique, et peut être trompeur à première vue. Par exemple l'offre d'une même incitation par l'entreprise ne va pas être perçue de la même manière par l'ensemble des clients. [Ascarza \(2018\)](#) remarque à ce propos que les clients dont le risque d'attrition est le plus fort, ne sont pas forcément ceux sur qui l'incitation aura le plus de succès. Cibler ces derniers dans une campagne de rétention, et plus généralement faire abstraction de l'hétérogénéité de la sensibilité des client, peut amener à une campagne inefficace. La segmentation du risque d'attrition n'est pas suffisant pour une bonne gestion. [Ascarza et al. \(2018\)](#) soulignent l'importance de bien distinguer la population à risque de la population à cibler : les clients ont des comportements hétérogènes et tous n'ont pas la même valeur pour l'entreprise. Les auteurs proposent alors une série de points d'attention pour la gestion d'une campagne de rétention efficace, schématisée dans la Figure 0.11.

Revenons maintenant à la modélisation du risque de rachat à partir de facteurs structurels par méthode d'apprentissage statistique. Actuellement, les précisions des méthodes de prédiction sont comparées avec des mesures statistiques classiques (voir par exemple [Milhaud et al., 2011](#); [Aleandri, 2017](#); [Jamal, 2017](#)). Or l'impact pour la compagnie d'assurances vie du rachat d'un contrat fortement valorisé est différent du rachat d'un contrat de faible valeur. Pratiquement, un actuaire pourrait vouloir une meilleure précision sur la prédiction des rachats des assurés avec une épargne importante, quitte à perdre un peu en exactitude sur le risque de rachat

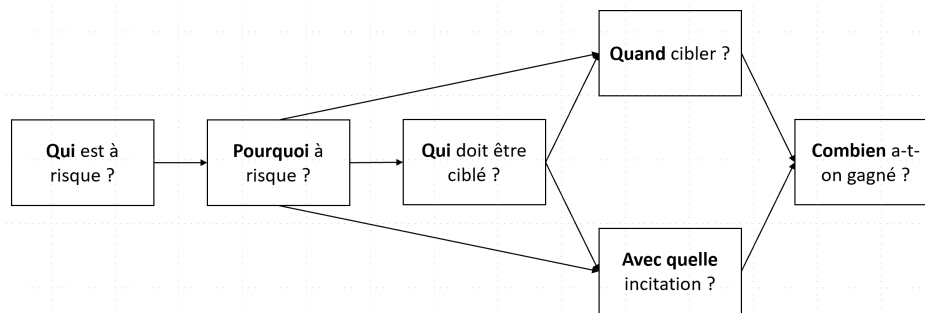


FIGURE 0.11 – Gestion de la rétention client.

Source : traduit d'Ascarza *et al.* (2018)

des contrats peu valorisés. Le choix du modèle retenu peut alors en être influencé. De même que dans l'analyse de l'attrition, le développement d'une fonction de perte personnalisée pour l'apprentissage statistique est la suite logique à ces questions.

A l'instar de la gestion d'une campagne de rétention (Figure 0.10), l'actuaire doit avoir à l'esprit l'ensemble des tenants et aboutissants des études de comportement de rachat qu'il effectue. Campbell *et al.* (2014) soulignent d'ailleurs que l'un des défis dans la compréhension du comportement des assurés est que très souvent, les actuaires sont encore trop concentrés sur l'objectif "d'ajuster une courbe" à partir des expériences passées, et moins portés sur des questions telles que "pourquoi" ou "et alors?"⁹.

Informations sur les marchés agricoles

L'apprentissage statistique ne se limite pas à l'application du meilleur modèle possible : choix de l'algorithme, de la méthodologie d'estimation des paramètres, de la fonction de perte, etc. Le rôle des données primaires est crucial. Le statisticien doit avant tout s'assurer que les données utilisées sont de bonne qualité, sans quoi il ne peut obtenir de bons résultats, et ce quel que soit le niveau de sophistication des techniques mathématiques appliquées. L'expression consacrée en informatique, généralement attribuée au programmeur d'IBM George Fuechsel, nous résume cette problématique : "*Garbage in, garbage out*"¹⁰. L'ingénierie des données est alors une étape initiale indispensable et se doit de collecter, homogénéiser et nettoyer les données avant leur analyse statistique. Un autre objectif du traitement des données, moins primordial mais qui peut s'avérer fortement bénéfique, est d'établir des liens pertinents entre les différentes sources de données afin d'augmenter l'information disponible pour la prédiction. Ces liens peuvent être établis entre des environnements plus ou moins disjoints au sein d'une même entreprise, mais peuvent aussi faire appel à des bases de données externes, acquises

9. Campbell *et al.* (2014), page 14, *Very often actuaries are still more focused on "fitting a curve" to past experience, with less emphasis on the "why" and "so what?" aspects*

10. Littéralement : des déchets en entrée, des déchets en sortie.

auprès d'un organisme privé ou disponibles publiquement. Le principe d'utiliser des données extérieures à un problème spécifique rencontré est bien connu des actuaires qui l'utilisent d'une certaine façon dans la théorie de la crédibilité (Bühlmann et Gisler, 2006). Un cadre d'application apparaît par exemple lorsque l'entreprise ne possède pas beaucoup d'informations sur ses clients, mais connaît tout de même son code postal. En utilisant les données socio-économiques disponibles à la maille géographique d'intérêt par le biais d'un organisme étatique (par exemple l'INSEE en France), la compagnie peut alors augmenter sa connaissance de ses clients.

Depuis plusieurs années, on constate que le volume de données générées augmente exponentiellement, et en particulier les données dont l'accès est totalement public et libre de droit. Nous sommes entrés non seulement dans l'ère du *Big data* mais aussi de l'*Open data*. Le secteur privé a bien compris l'intérêt grandissant pour l'*open data*, et a récemment développé des moteurs de recherche de bases de données comme par exemple Quandl¹¹, spécialisé dans les séries financières et économiques, ou Google Dataset Search¹² plus généraliste dans les bases de données référencées. Néanmoins, la multiplication des informations en libre accès est avant tout une volonté gouvernementale. Plusieurs pays proposent des plateformes d'accès à une multitude de bases de données gouvernementales : data.gouv.fr pour la France, data.gov pour les États-Unis, data.gov.uk pour le Royaume-Uni, ouvert.canada.ca pour le Canada. Les villes ont aussi développé des outils similaires : opendata.paris.fr pour Paris, opendata.cityofnewyork.us pour New-York, data.london.gov.uk pour Londres. A une plus grande échelle, les institutions internationales proposent les mêmes services : data.worldbank.org pour la Banque Mondiale, data.imf.org pour le Fonds Monétaire International. Des agences gouvernementales plus spécifiques se sont dotées de portails avec ce même objectif de simplifier l'accès à la donnée, nous notons en particulier les agences spatiales : search.earthdata.nasa.gov pour la NASA, scihub.copernicus.eu pour Copernicus (le programme européen de surveillance de la Terre).

Cet avènement du *Big Open data* est évidemment à la fois source d'opportunités et de défis pour l'apprentissage statistique. Mais l'ouverture des données au plus grand nombre a bien d'autres avantages : politiques, sociaux et économiques (Janssen *et al.*, 2012). En effet, les objectifs initiaux principaux sont souvent, pour les initiatives gouvernementales, l'amélioration de l'efficacité de l'action publique et du fonctionnement démocratique en offrant plus de transparence aux citoyens. L'*open data* a aussi des bénéfices économiques, par exemple en réutilisant ces données dans de nouveaux services à forte valeur ajoutée, elle stimule l'innovation économique et sociale. Enfin, toujours d'un point de vue économique, l'ouverture des données améliore la disponibilité de l'information et notamment son caractère d'accès égalitaire pour les entreprises et les investisseurs.

Asymétrie d'information

11. quandl.com

12. toolbox.google.com/datasetsearch

Cette question de l'homogénéité de l'information disponible est l'une des hypothèses fondamentales dans nombre de travaux fondateurs de la théorie actuelle de l'économie de marché. Ainsi, en économie, les travaux d'Arrow et Debreu, récompensés par le prix Nobel d'économie respectivement en 1972 et 1983, sur la théorie de l'équilibre général (Arrow et Debreu, 1954), supposent une concurrence pure et parfaite où les agents sont parfaitement informés. En finance, la grande majorité des modélisations stochastiques, dont la théorie de la spéculation de Louis Bachelier (Bachelier, 1900) ou celle de la couverture d'options introduite par Black, Scholes et Merton (Black et Scholes, 1972, 1973; Merton, 1973), admettent une symétrie d'information parmi les agents sur le marché.

Pratiquement, cette hypothèse n'est que rarement vérifiée, et ne reflète donc pas la réalité des marchés, économiques et financiers, où certains agents disposent d'informations pertinentes supplémentaires alors que d'autres n'ont accès qu'à des informations incomplètes. La recherche académique s'est bien entendu attardée sur le sujet afin de proposer des modèles convenables face à cette problématique. En économie, le prix Nobel remis à Stiglitz, Akerlof et Spence en 2001 vient d'ailleurs récompenser leurs travaux sur les marchés avec asymétrie d'information depuis le début des années 1970 (e.g., Akerlof, 1970; Spence et Zeckhauser, 1971; Rothschild et Stiglitz, 1976). Dans les approches probabilistes de la finance, les premières recherches sur la modélisation de marché en présence d'asymétrie d'information arrivent quelques années plus tard avec notamment les travaux de Kyle (1985) et Back (1992) qui étudient l'impact d'un agent initié sur le marché, i.e. qui possède initialement de l'information supplémentaire. L'une des modélisations possibles de l'agent initié, d'un point de vue probabiliste, repose sur l'utilisation dans les calculs stochastiques du grossissement de filtration, méthode développée au début des années 1980 par l'école française de probabilité (Yor, 1978; Jeulin, 1980; Jeulin et Yor, 1985). L'application de ce dernier cadre a permis l'étude détaillée des marchés financiers en présence d'asymétrie d'information, avec par exemple des travaux sur l'équilibre (Hillairet, 2004) ou sur la couverture (Eyraud-Loisel, 2005) des instruments financiers.

En parallèle du développement des modèles théoriques, nécessaires à la compréhension des conséquences de l'asymétrie d'information sur les marchés, les institutions gouvernementales tentent pratiquement de limiter cette problématique. Ainsi, d'un point de vue légal, l'Autorité des marchés financiers (AMF) définit d'"*initiée* une personne qui détient une information précise sur un instrument financier, qui n'a pas encore été rendue publique, et qui, si elle l'était, aurait un impact significatif sur le cours de cet instrument financier", et de continuer par "un *initié* commet un délit d'initié s'il utilise cette information ou la transmet à une autre personne"¹³. En effet, dans le système judiciaire français, l'utilisation ou la communication d'une telle information est considérée comme un délit, et est passible, au titre de l'article L465-1 du Code monétaire et financier, de cinq ans d'emprisonnement et de 100 millions d'euros

13. Pour la définition légale en vigueur d'information privilégiée et d'opérations d'initiés, le lecteur pourra se référer au Règlement No 596/2014 du Parlement européen, articles 7 et 8.

d'amende. Outre le fait de sanctionner les délits d'initiés, les pouvoirs publics obligent les entreprises faisant appel aux investisseurs financiers à faire preuve de transparence, et notamment de communiquer sur leurs performances, positions financières ou modifications importantes de l'actionnariat à l'ensemble des agents du marché. Ce devoir de déclaration d'informations est aussi familier des actuaires, particulièrement depuis la mise en place de la directive Solvabilité II, dont le troisième pilier décrit les obligations de communication au public. De même, l'ouverture des données gouvernementales s'inscrit ainsi dans cette volonté institutionnelle de diminuer l'asymétrie d'information sur les marchés.

Information et marchés financiers agricoles

L'un des marchés où le gouvernement s'investit historiquement le plus dans la diffusion de l'information est le marché des biens agricoles. L'agriculture est en effet un secteur stratégique, qui implique des volumes importants d'échanges et dont les prix sont fortement volatiles, ce qui est notamment dû à la forte météo-sensibilité des rendements et donc de l'offre agricole. Ceci est d'autant plus vrai si l'on se replace dans le contexte de la fin du XIX^{ème}, où la majorité de la population était encore dépendante de l'agriculture dans l'ensemble des pays. Ainsi, le Département de l'Agriculture des États-Unis (USDA), publie chaque mois le rapport *Crop Production* qui rassemble les données sur les récoltes à travers l'ensemble du pays, et ce depuis 1866 (USDA, 2018). L'ensemble de ces données est d'ailleurs toujours en libre accès via le portail quickstats.nass.usda.gov dédié à cet effet. De plus, l'USDA produit aussi, dès 1893, des rapports sur la production des cultures les plus importantes à travers le monde (USDA, 1893). Ces rapports jouent un rôle central dans la détermination des prix sur les marchés agricoles (Schnepf, 2005). Le début de la publication de ces rapports coïncide avec la création du *Chicago Board of Trade* (CBOT). En effet à la fin des années 1840, les agriculteurs commencèrent à vendre leurs récoltes aux marchands de Chicago en utilisant des contrats à terme afin de se protéger de prix défavorables sur les marchés des céréales. Cependant le risque de crédit posant encore des difficultés, le CBOT fut alors créé en 1848.

Parmi les rapports gouvernementaux publiés par l'USDA, nous notons donc le mensuel *Crop Production*, dont la préparation est assurée par le *National Agricultural Statistics Service* (NASS), et qui couvre les céréales, les plantes oléagineuses et le coton cultivés aux États-Unis. Les données sont disponibles à plusieurs mailles géographiques : au niveau du pays, des régions, des États ou encore des comtés. Suivant la saisonnalité des cycles de chaque culture, les informations sont publiées à différents moments de l'année et peuvent être des estimations, i.e. une observation d'un fait accompli, ou des prévisions, i.e. à propos d'un processus non fini. Dans les statistiques mises à disposition par l'USDA au marché nous retrouvons :

- les rendement agricoles ;
- les surfaces plantées et récoltées ;
- les productions.

A ces *Crop Production* mensuels s'ajoutent aussi des rapports hebdomadaires :

- le *Crop Progress*, communiqué pendant la saison de croissance (d'avril à novembre), qui résume les conditions générales (exprimées en pourcentage des cultures réparties en cinq classes allant de "très mauvaise" à "excellente") et la progression des semis et de la moisson ;
- le *Weekly and Crop Bulletin*, publié toute l'année, qui rassemble des informations sur la météo pertinentes pour les cultures, dont la température et les précipitations par exemple.

Enfin, le *World Agricultural Supply and Demand Estimates* (WASDE), décrit comme la "pierre angulaire des rapports" fournis par l'USDA, apporte une vision globale de l'offre et de la demande. Ce rapport mensuel contient des statistiques sur l'ensemble des marchés mondiaux dont notamment :

- la production ;
- la consommation ;
- les imports et exports ;
- les stocks ;
- les prix.

La préparation de ce dernier rapport, publié en même temps que le *Crop Production*, se fait dans la plus grande vigilance pour éviter tout risque de délit d'initié. Ainsi pour la préparation du WASDE nous pouvons lire sur le site de l'USDA que "*pour s'assurer que les informations fortement sensibles pour le marché soient communiquées simultanément à l'ensemble des utilisateurs finaux, et non prématurément à aucune personne en particulier, le rapport WASDE est préparé sous haute sécurité dans une zone conçue spécialement dans le Bâtiment Sud de l'USDA. Le matin de la publication, les portes dans la zone du huis clos sont sécurisées, les stores des fenêtres sont scellés, le téléphone et les communications Internet sont bloqués. Une fois que les analystes ont présenté leur accréditation à un garde, ils entrent dans la zone sécurisée pour finaliser le rapport WASDE. Les communications avec le monde extérieur sont suspendues jusqu'à ce que le rapport soit publié à midi pile, heure de la côte Est*¹⁴". Des mesures draconiennes s'il on peut dire...

Les précautions vont même jusqu'au choix de l'heure de la parution du rapport. Jusqu'au 1^{er} janvier 2013, elle avait lieu à 8 :30. Mais, à cause de changements d'horaires de marché, l'USDA a sollicité les observations du public sur l'heure de communication entre le 8 juin et le 9 juillet

14. To assure the highly market-sensitive information is released simultaneously to all end-users, and not prematurely to any one, the WASDE report is prepared under tight security in a specially designed area of USDA's South Building. The morning of release, doors in the "lockup" area are secured, window shades are sealed, and telephone and Internet communications are blocked. Once analysts present their credentials to a guard, they enter the secured area to finalize the WASDE report. Communications with the outside world are suspended until the report is released at 12 :00 noon Eastern time.

2012. Après avoir reçu 147 commentaires, consultables sur le site internet du NASS, l'USDA a choisi d'un "*changement pour une publication à midi qui permet la meilleure liquidité dans les marchés, fournit le meilleur accès au rapport durant les heures ouvrées aux États-Unis, et poursuit un accès égalitaire aux données pour toutes les parties*"¹⁵. L'impact de changement a d'ailleurs étudié *a posteriori*. [Adjemian et Irwin \(2018\)](#) montrent que même si les agents ont plus de difficultés à distinguer les informations de valeurs contenues dans les rapports, le marché met à peu près le même temps pour les assimiler ; et de conclure que remettre une publication hors des horaires de marché allongerait le processus de découverte des prix.

Impact des rapports USDA

Toutes ces considérations nous laissent entrevoir l'importance des informations contenues dans ces différents rapports de l'USDA. La valeur de ces dernières est un sujet de grand intérêt dans la littérature économique agricole depuis les années 1980 (e.g., [Choi, 1982](#); [Milonas, 1987](#)). La majorité de ces travaux s'inscrivent dans le cadre de l'étude événementielle, i.e. de "*déterminer si un effet prix "anormal" associé avec un événement inattendu*" ([McWilliams et Siegel, 1997](#)). La première étude de ce genre remonte aux travaux de [Dolley \(1933\)](#), qui analyse l'impact sur les rendements des fractionnements d'action. Pour notre sujet d'intérêt, [Sumner et Mueller \(1989\)](#) ont été les premiers à utiliser ce cadre méthodologique afin d'examiner l'effet des rapports *Crop Production* de juillet à novembre sur les prix futures du maïs et du soja pendant la période 1961–1982. Pour ce faire, ils utilisent divers tests statistiques (*t*-tests, *F*-tests et test du χ^2) pour déterminer si la variance des rendements financiers des futures est différente entre les séances liées à une annonce de l'USDA et les autres séances de bourse. Ils concluent alors que les informations incluses dans les rapports gouvernementaux sont considérées comme nouvelles et fiables par les agents du marchés, car les changements des prix durant les séances liées aux publications sont significativement plus importantes que les autres sessions.

[Fortenbery et Sumner \(1993\)](#) notent qu'à partir de 1982, de nombreux changements ont eu lieu dans le milieu de la bourse, des technologies de communication et dans les rapports de l'USDA. Ils analysent alors les effets de la publications du WASDE sur les prix futures et les options pendant la période 1982-1989 sur les mêmes denrées (maïs et soja). Pour ce faire, ils utilisent une méthodologie similaire de tests statistiques, mais s'appuient aussi sur un modèle de régression. Dans ce dernier, le changement des prix journaliers est régressé, entre autres, sur une variable binaire indiquant la communication ou non d'un rapport WASDE cette séance là. Les résultats des deux approches les amènent à conclure que le WASDE n'apporte plus d'information pour les marchés financiers agricoles.

L'étroitesse de la fenêtre temporelle considérée dans les travaux de [Fortenbery et Sumner](#)

15. The shift to a noon release allows for the greatest liquidity in the markets, provides the greatest access to the reports during working hours in the United States, and continues equal access to data among all parties.

(1993) a poussé les chercheurs à itérer l'étude en incluant une plus longue période d'observation quelques années plus tard. Ainsi Garcia *et al.* (1997) examinent la valeur informationnelle des prévisions de production de l'USDA sur la période 1971-1992, en prenant aussi en compte les prévisions privées dont celles de Conrad Leslie considéré comme "le plus grand prévisionniste de récolte de la nation" par le New York Times. Ils concluent que les rapports USDA ont toujours un impact significatif, et que les traders sur les futures de maïs et de soja seraient prêts à payer pour connaître les prévisions de l'USDA en avance. Dans la même lignée, Isengildina-Massa *et al.* (2008) étudient l'impact du WASDE au cours de la période 1985-2006 sur ces mêmes futures en analysant la variance de leurs rendements fermeture-ouverture. La Figure 0.12 donne une représentation graphique des résultats. Ils montrent alors que l'effet de l'annonce du WASDE est d'autant plus important lorsqu'il est publié avec le *Crop Production*. Et de conclure que non seulement les rapports USDA apportent toujours de l'information, mais que l'effet de l'annonce du WASDE a augmenté au cours du temps.

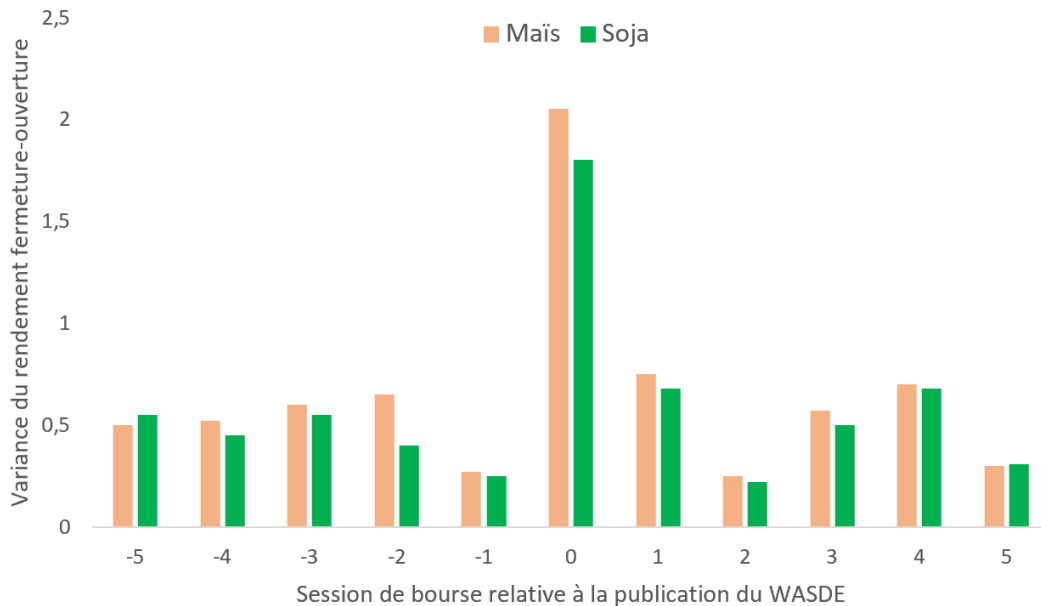


FIGURE 0.12 – Variance des rendements fermeture-ouverture des futures sur le maïs et le soja pour tous les mois de communication du WASDE. Janvier 1985 - Décembre 2006.

Source : Isengildina-Massa *et al.* (2008)

De nombreuses autres études ont été menées sur ce sujet pour lequel la littérature est vaste (e.g., McNew et Espinosa, 1994; Colling *et al.*, 1996; Irwin *et al.*, 2001; McKenzie, 2008; Dorfman et Karali, 2015; Gouel, 2018). Certaines analysent des biens agricoles différents du maïs et du soja comme par exemple le riz (McKenzie et Darby, 2017), ou des animaux comme le porc ou le bétail (Colling et Irwin, 1990; Grunewald *et al.*, 1993; Isengildina *et al.*, 2006). D'autres utilisent de nouvelles techniques statistiques pour évaluer l'impact des informations communiquées par l'USDA comme des systèmes de régressions (Adjemian, 2012) ou des mo-

dèles GARCH multivariés (Karali, 2012). Plus récemment, Abbott *et al.* (2016) quantifient à 301 millions de dollars la valeur de l'information contenue dans le WASDE pour l'ensemble des agents du marché dans le cas du maïs. En particulier, la valeur des statistiques sur les rendements agricoles est estimée à 188 millions de dollars.

La majorité de ces recherches s'intéressent aux effets de la publication des rapports mensuels de l'USDA (WASDE et *Crop Production*) sur les marchés. Lehecka (2014), quant à lui, s'attarde sur la valeur des informations contenues dans les rapports *Crop Progress* hebdomadaires au cours de la période 1986-2012 pour le maïs et le soja. L'auteur régresse le rendement fermeture-ouverture des futures en fonction du pourcentage des cultures en excellente ou bonne condition. Les résultats présentent une corrélation significative et négative entre l'amélioration des conditions des cultures et les changements de prix, soulignant ainsi une réaction rapide et rationnelle des marchés face cette information de l'USDA. Les réactions sont d'autant plus fortes lorsque les conditions météorologiques sont les plus critiques pour les cultures, i.e. en juillet et en août.

Plus récemment, avec l'essor du *Big data*, les chercheurs se demandent, avec les mêmes motivations que Fortenbery et Sumner (1993), si les rapports de l'USDA contiennent toujours des informations utiles pour le marché. Intuitivement, la multiplication significative des sources d'informations devrait en effet diminuer les effets de la publication des rapports comme le WASDE ou le *Crop Production*. Karali *et al.* (2019) et Ying *et al.* (2019) examinent la question et concluent que l'impact de la communication de ces rapports n'a pas diminué, certains effets semblant même gagner en importance. Si le *Big data* ne semble pas altérer l'influence des rapports gouvernementaux sur les marchés, il est cependant peut-être possible capable de prédire les informations contenues dans ces publications grâce à l'apprentissage statistique. En partant de l'hypothèse qu'une entreprise ait développé un modèle prédictif pour le WASDE, Milacek et Brorsen (2017) évaluent le gain potentiel qu'elle en tirerait sur le maïs et le soja. Un tel modèle serait bénéfique pour les traders présents sur les marchés. En particulier, nous noterons que les assureurs agricoles sont concernés dans le cas des États-Unis car la majorité des produits d'assurance agricole y sont de type revenus, i.e. que l'agriculteur transfère à la fois son risque de rendement agricole et son risque de prix de vente.

Aujourd'hui les données en libre accès sont légion, et même si elles sont publiques, le *big open data* peut amener certains agents à détenir de l'information supplémentaire sans qu'elle ne soit considérée comme privilégiée, dans le sens où le reste des agents y ont aussi théoriquement accès. À l'instar du risque d'obfuscation où un demandeur d'assurance se retrouve noyé sous la multitude d'offres et choisit de façon sous-optimale sa couverture (Mouminoux *et al.*, 2018), l'agent des marchés financiers est submergé dans les téraoctets en accès libre et choisit peut-être de façon sous-optimale les sources d'informations. Ainsi, les données nécessaires à la création d'un modèle de prédiction du WASDE sont potentiellement déjà disponibles dans l'océan de l'*open data*.

Introduction des outils statistiques

Nous introduisons maintenant les algorithmes ainsi que les données issues de l'*open data* utilisés dans les méthodologies d'apprentissage statistique sur lesquels nous nous focalisons dans cette thèse. L'objectif de cette partie n'est pas leur application spécifique à nos cas d'étude, mais plutôt de fournir une description des techniques employées dans un cas général. Sans pleinement détailler les théories mathématiques, nous donnons les principales idées sous-jacentes afin de comprendre leurs fonctionnements. L'ensemble des travaux statistiques de cette thèse ont été effectués à l'aide du logiciel R (R Core Team, 2019).

Régression pénalisée

La régression linéaire est le modèle statistique le plus simple pour envisager l'impact de variables explicatives $\mathbf{x} \in \mathbb{R}^p$ sur une variable à expliquer y . Soient les données d'apprentissage $D = \{(y_i, \mathbf{x}_i)\}$, pour les observations $i = 1, 2, \dots, N$; le statisticien estime alors les paramètres $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^n$ du modèle

$$y = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}, \quad (0.29)$$

en minimisant le critère de perte des moindres carrés (OLS¹⁶)

$$L_{OLS}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2. \quad (0.30)$$

Les estimateurs OLS sont connus pour avoir généralement un biais nul. Néanmoins, dans le cas où le nombre p de variables explicatives est important par rapport au nombre d'observations N , ils ont une forte variance, ce qui n'est pas souhaitable dans le cadre de la prédiction. Par ailleurs, plus la dimension p de l'espace des variables explicatives est grand, moins le modèle est interprétable. En fixant certains coefficients à 0, i.e. en se restreignant à un sous-ensemble de variables explicatives, la précision de prédiction ainsi que l'interprétation du modèle peuvent s'en retrouver améliorées, quitte à augmenter le biais des coefficients.

Dans ce cadre, Tibshirani (1996) propose une nouvelle méthode d'estimation : le *Least Absolute Shrinkage and Selection Operator* ou "*lasso*". Le principe est de forcer l'estimation de certains coefficients à zéro lors de leur estimation. Pour ce faire Tibshirani (1996) ajoute une pénalisation portant sur la somme des valeurs absolues des coefficients au critère de perte OLS :

16. Ordinary Least Square

$$L_{lasso}(\beta_0, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|, \quad (0.31)$$

où $\lambda > 0$ est le paramètre de régularisation. Intuitivement, les coefficients apportant le moins d'information ne font pas assez diminuer le terme OLS par rapport à leur pénalisation en norme \mathcal{L}_1 . Ainsi, plus la valeur de λ est grande, plus nombreux sont les coefficients β_j estimés à exactement 0. Les variables explicatives sont alors sélectionnées automatiquement selon un critère de minimisation.

Cependant, il nous reste encore une question à ce problème : quel niveau de régularisation λ faut-il choisir ? Une méthode commune est la validation croisée, plus particulièrement la *K-fold cross validation*. Elle consiste à séparer aléatoirement ses données initiales D en K sous-échantillons $\{D_1, \dots, D_k\}$ de taille identique. Soit un λ fixé, pour chaque $k \in \{1, \dots, K\}$ nous entraînons le modèle, i.e. nous estimons $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})_\lambda$ avec la pénalité λ , sur l'ensemble d'apprentissage $D - D_k$. Puis nous appliquons le modèle sur l'ensemble test D_k pour obtenir les prédictions \hat{y}_λ de la variable dépendante. En répétant ces opérations K fois, nous obtenons des prédictions sur l'ensemble total D des données. Nous appliquons alors une mesure de validation $\rho(y, \hat{y}_\lambda)$ afin de comparer la précision du modèle selon le choix du paramètre λ . Dans le cas spécifique de la régression linéaire, nous utilisons l'erreur quadratique en tant que mesure de validation, i.e. $\rho(y, \hat{y}) = \sum_{i=1}^N (y_i - \hat{y}_i)^2$.

Le lasso a deux limites principales :

- en très grande dimension où $p \gg N$, le lasso ne peut sélectionner que N variables au maximum, i.e. seulement N coefficients seront non nuls ;
- si des variables explicatives sont fortement corrélées, le lasso a tendance à ne sélectionner qu'une seule variable parmi ce groupe.

Une autre forme de pénalisation, similaire au lasso, répond à ces deux problématiques : la régression *ridge* introduite par [Hoerl et Kennard \(1970\)](#). Dans ce cas les coefficients ne sont plus régularisés selon la norme \mathcal{L}_1 mais \mathcal{L}_2 . Le critère de perte est alors

$$L_{ridge}(\beta_0, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|^2. \quad (0.32)$$

Comme le lasso, la régression ridge permet de diminuer la variance des coefficients en augmentant leur biais. Elle se comporte cependant différemment. Ainsi aucun coefficient n'est exactement estimé à 0, donc aucune sélection de variable n'est effectuée, et ce même dans le cas où $p \gg N$. Par ailleurs, si des variables explicatives sont fortement corrélées, le ridge aura tendance à estimer les coefficients concernés à la même valeur.

Partant des différences de chacune de ces approches dans leurs limites et leurs forces , [Zou et Hastie \(2005\)](#) proposent un mélange entre les deux démarches : l'*elastic-net*. Il est ajouté au critère de perte OLS des termes de pénalisation des coefficients à la fois en norme \mathcal{L}_1 et en norme \mathcal{L}_2 :

$$L_{enet}(\beta_0, \boldsymbol{\beta}, \lambda, \alpha) = \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1 - \alpha}{2} \sum_{j=1}^p |\beta_j|^2 \right), \quad (0.33)$$

où λ est le paramètre de régularisation général et α le paramètre de mélange entre le cas du lasso ($\alpha = 1$) et du ridge ($\alpha = 0$).

Avec le bon choix de mélange α , la pénalité elastic-net permet alors de surmonter les limites du lasso simple. Si des variables explicatives sont fortement corrélées, l'elastic-net aura tendance à soit sélectionner l'ensemble du groupe soit toutes les rejeter. De plus, il est possible à l'aide de l'elastic-net de sélectionner plus de N variables tout en estimant certains coefficients à 0.

Support Vector Machine

Le *Support Vector Machine* (SVM) est un algorithme de classification en apprentissage statistique introduit au début des années 1990 par [Boser et al. \(1992\)](#) et [Cortes et Vapnik \(1995\)](#). En alternative des réseaux de neurones, fortement appréciés à l'époque pour la classification, le SVM est devenu un algorithme populaire avec des domaines d'applications très variés : la reconnaissance de chiffres écrits à la main ([Scholkopf et al., 1997](#)) ou de visage ([Guo et al., 2000](#)), la détection de fraude ([Chen et al., 2004](#)) ou d'attrition ([Zhao et al., 2005](#)), l'analyse sentimentale de texte ([Mullen et Collier, 2004](#)) ou encore l'analyse de gènes en bio-informatique ([Noble, 2004](#)).

Même si des variantes existent, nous décrivons son application simple dans le cas de la classification binaire. L'algorithme peut être décrit en terme géométrique. L'objectif est de classer N observations binaires étant donné un espace de variables explicatives de dimension p . Pour ce faire le SVM se propose de trouver un hyperplan séparateur. Dans le cas où $p = 2$, il s'agit alors d'une droite. Néanmoins, il se peut que plusieurs hyperplans puissent parfaitement séparer les données, or nous voudrions choisir le "meilleur" d'entre eux. Ainsi, la marge est définie comme la distance entre l'hyperplan et le point d'observation le plus proche. La séparation retenue est alors celle qui maximise cette marge : l'hyperplan à marge maximale. La Figure 0.13 représente graphiquement ce processus.

Cependant, dans les cas réels de classification, il est très rare de pouvoir séparer parfaitement son ensemble de données avec un hyperplan. L'algorithme SVM est modifié en conséquence par l'ajout d'une marge "souple" (*soft-margin* en anglais). En d'autres termes, cette dernière tolère les mauvaises catégorisations, à condition de ne pas être trop loin de la frontière définie

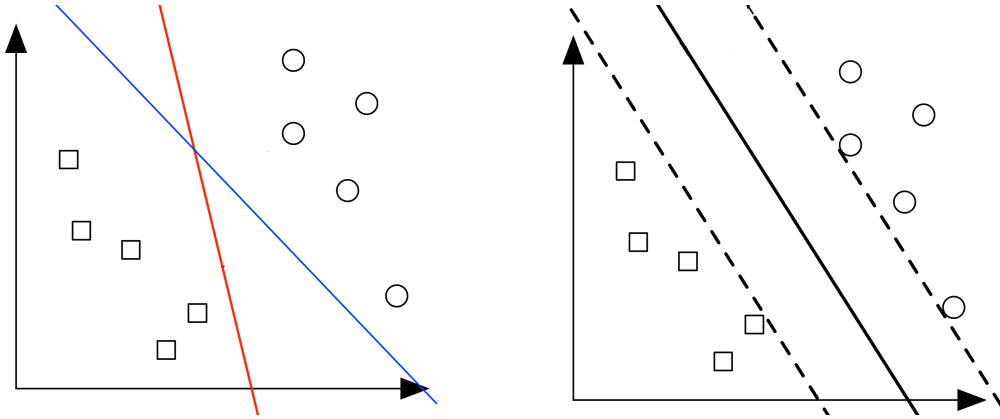


FIGURE 0.13 – Gauche : deux hyperplans différents séparant le même ensemble de données.
Droite : l'hyperplan à marge maximale.

Source : [Weinberger \(2018\)](#)

par l'hyperplan. Afin de ne pas obtenir trop de mauvaises classifications, l'utilisateur doit alors spécifier un coût de pénalisation pour chaque erreur commise, qui peut aussi être vu comme un paramètre de souplesse de la marge. Le choix de ce paramètre permet de contrôler le compromis entre les infractions à l'hyperplan et la largeur de la marge.

Une autre méthodologie utilisée dans les cas non séparables est l'astuce dite du noyau (*kernel trick* en anglais) qui consiste à projeter l'ensemble des données dans un espace de dimension supérieure. Un cas non séparable par un hyperplan peut alors le devenir après transformation des données dans une plus grande dimension. Un exemple graphique est présenté dans la Figure 0.14.

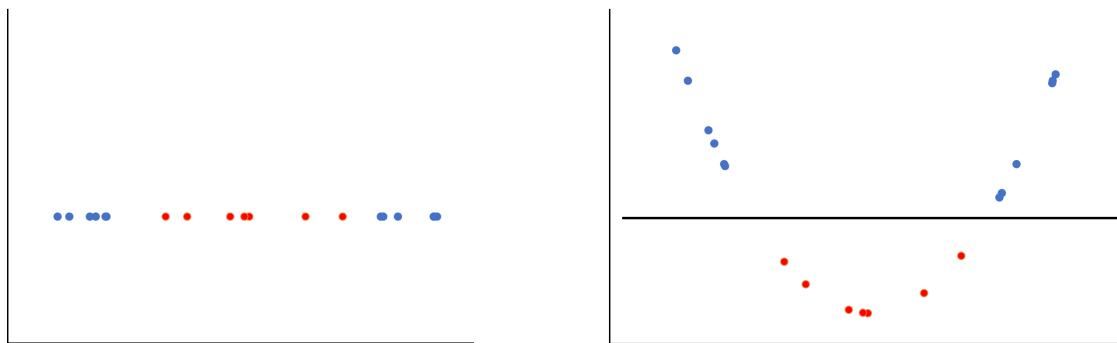


FIGURE 0.14 – Gauche : cas non séparable linéairement en dimension 1. Droite : séparation linéaire possible après projection de l'ensemble de données en dimension 2.

Source : [Noble \(2006\)](#)

Maintenant, donnons une description d'un point de vue statistique du fonctionnement du

SVM. Soit les données d'apprentissage $\{(y_i, \mathbf{x}_i)\}_1^N$, où $\mathbf{x}_i \in \mathbb{R}^p$ sont les variables explicatives et $y_i \in \{+1, -1\}$ la variable dépendante, l'algorithme du SVM cherche la solution du problème d'optimisation suivant

$$\min_{\omega, b, \xi} \frac{1}{2} \omega^\top \omega + C \sum_{i=1}^N \xi_i, \quad (0.34)$$

avec les contraintes

$$\begin{aligned} \forall i, y_i (\omega^\top \phi(\mathbf{x}_i) + b) &\geq 1 - \xi_i, \\ \text{et } \forall i, \xi_i &\geq 0. \end{aligned} \quad (0.35)$$

Dans ces notations, l'hyperplan est déterminé par le vecteur orthogonal ω et la constante b . Les erreurs de classification sont représentées par les $\xi_i > 1$, et sont associées au coût de pénalisation C . L'ensemble des données est projeté dans un espace de dimension supérieure grâce à la fonction ϕ dont le noyau sous-jacent est défini par $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$.

Beaucoup de noyaux ont été développés, cependant l'un des choix les plus communs est le noyau gaussien ou fonction de base radiale (RBF) :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \quad (0.36)$$

où $\gamma > 0$ est un paramètre du noyau. Dans la pratique, l'optimisation du paramètre de pénalisation d'erreur C et des paramètres du noyaux, seulement γ pour le cas du RBF, est effectuée par validation croisée.

Extreme Gradient Boosting

Les arbres de décision, et plus particulièrement les arbres de classification et de régression (CART pour *Classification and Regression Trees*) introduits par [Breiman et al. \(1984\)](#) sont des outils simples d'interprétation, rapides dans leur estimation et permettant de s'affranchir des structures linéaires. Le principe est de définir la valeur réponse de l'algorithme en suivant une suite de règles de décision sur les variables explicatives. Pour déterminer ces règles, l'espace des données est successivement séparé par des divisions binaires en sous-espaces disjoints. A chaque étape, ou nœud, de cette construction descendante, la variable explicative sur laquelle se fait la séparation ainsi que le point de division sont choisis pour optimiser un critère d'homogénéité, comme par exemple l'indice de Gini pour une classification ou la perte quadratique dans le cadre d'une régression. Bien qu'ils offrent une lecture facile, ces arbres sont aussi connus pour être relativement instables. Un faible changement dans les données initiales peut impliquer des règles de décisions fortement différentes ([Li et Belford, 2002](#)).

Pour pallier cette instabilité dans les prédictions du CART, les méthodes d'apprentissage ensemblistes proposent des réponses séduisantes. Au lieu d'entraîner un seul arbre, nous en utilisons alors une multitude : une "forêt". Cette intuition est assez ancienne, souvenons-nous du commandant athénien en 428 avant J.C. qui n'utilise pas un soldat mais toute une garnison pour estimer la hauteur du mur adverse. Cependant dans le cas des CART, estimer deux fois l'algorithme sur le même ensemble ne changera rien car leur construction est déterministe. Il faut donc modifier les données d'apprentissage.

Le principe du *bagging*, pour *bootstrap aggregating*, vient répondre à cette problématique. Développée par Breiman (1996), cette méthode consiste à entraîner chaque arbre sur un échantillon aléatoire généré par tirages avec remise à partir de l'ensemble des données initiales, comme dans le cas du bootstrap. L'agrégation de tous les arbres donne la prévision finale. Dans le cas d'une classification nous pouvons choisir le vote majoritaire (comme au siège de Platée), alors que pour une régression la moyenne sera généralement préférée. Ainsi, pour réduire la variance de la prédiction, de l'aléatoire a été introduit dans les observations. Une alternative, proposée par Breiman (2001), est d'introduire de l'aléatoire dans les variables explicatives utilisées : on parlera alors de "forêts aléatoires" (*random forests*). Chaque arbre de la forêt ne sera entraîné que sur un sous-ensemble aléatoire des covariables.

Les techniques de bagging et de forêts aléatoires permettent de réduire la variance de l'estimateur. Cependant, puisque les arbres du bagging reposent du bootstrap, le biais reste le même que pour un arbre simple. Schapire (1990) propose un algorithme d'apprentissage ensembliste répondant à cet objectif : le *boosting*. Le principe général est de combiner des modèles d'apprentissage "faibles" h (*weak learners*), par exemple des arbres avec une faible profondeur, pour en faire un modèle "fort" F (*strong learner*) avec une bonne précision. A la différence du bagging, où les arbres sont entraînés indépendamment sur les données initiales (à un bootstrap près), le boosting estime les modèles faibles itérativement sur les "résidus" du modèle précédent.

Vu que l'objectif de l'itération est d'améliorer la précision de prédiction, il est logique d'introduire une fonction de perte, notée L , qui pourra par exemple être logistique dans le cadre d'une classification ou quadratique pour une régression. Étant donné un ensemble d'apprentissage $D = \{(y_i, \mathbf{x}_i)\}_1^N$, nous cherchons un modèle fort F_M comme combinaison de modèles faibles h_m appartenant à une classe de modèles \mathcal{H} :

$$F_M = \sum_{m=0}^M h_m. \tag{0.37}$$

En tant que processus de construction itératif, supposons que l'étape m soit terminée et que nous ayons ainsi un modèle fort F_m . L'itération suivante se fait alors par

$$F_{m+1} = F_m + \arg \min_{h \in \mathcal{H}} \sum_{i=1}^N L(y_i, F_m(\mathbf{x}_i) + \nu h(\mathbf{x}_i)), \quad (0.38)$$

où ν est un paramètre de "shrinkage" permettant de limiter la vitesse d'apprentissage lors des étapes d'itération. Dans le reste de la thèse, nous considérons le cadre spécifique où les modèles faibles h sont des CART.

Dans le cas où la fonction de perte est différentiable et que la vitesse d'apprentissage est relativement lente, [Friedman \(2001\)](#) propose d'utiliser une descente de gradient pour estimer le modèle faible h_m . Cet algorithme dit de gradient boosting repose sur les approximations par séries de Taylor :

$$L(F_m + \nu h) \approx L(F_m) + \nu \langle \nabla L(F_m), h \rangle, \quad (0.39)$$

où par abus de notation $L(F) = \sum_{i=1}^N L(y_i, F(\mathbf{x}_i))$.

Récemment, [Chen et Guestrin \(2016\)](#) ont proposé une variante du gradient boosting en y incorporant d'autres méthodes d'apprentissage statistique afin d'améliorer le modèle fort final : l'*extreme gradient boosting* (XGBoost).

- A la manière du bagging, chaque arbre faible est entraîné sur un sous-ensemble d'observations tirées aléatoirement (sans remise). Cette amélioration, que l'on nomme gradient boosting stochastique, avait déjà été proposée par [Friedman \(2002\)](#).
- Comme pour les forêts aléatoires, chaque arbre faible est estimé sur un sous-ensemble des covariables.
- Une fonction de régularisation Ω , faisant intervenir des pénalisations de type lasso ou ridge, est ajoutée à la fonction de perte L afin de limiter le surapprentissage. L'optimisation se fait alors sur une fonction *objective* $O = L + \Omega$.

L'algorithme XGBoost est vite devenu très populaire dans le milieu de l'apprentissage statistique pour toutes les problématiques faisant intervenir des données structurées. En regroupant différentes techniques, il permet d'obtenir de très bonnes performances de prédiction comparativement aux autres méthodologies, notamment en offrant les outils pour traiter le compromis biais-variance avec précision. De plus, l'implémentation en langage informatique (R, Python, etc.) a été optimisée avec notamment l'utilisation du calcul en parallèle sur plusieurs cœurs, permettant ainsi un temps d'estimation relativement faible. Enfin, une grande liberté est laissée à l'utilisateur dans les choix des paramètres, ce que ne propose pas forcément d'autres bibliothèques de boosting.

Indice de végétation

Les données des agences spatiales gouvernementales sont une des sources d'information libre dont l'accès a été récemment simplifié. La NASA a par exemple mis à disposition une large catégorie d'observations atmosphériques (température, vent, précipitation, etc.) depuis le début des années 1980 par le biais du projet MERRA-2 (Gelaro *et al.*, 2017). De plus l'agence a dernièrement développé une plateforme permettant un accès groupé à l'ensemble de ses bases de données en accès libre : search.earthdata.nasa.gov. Parmi ces observations spatiales, nous nous intéressons aux indices de végétations, et plus particulièrement au *normalized difference vegetation index* (NDVI).

Pionnier dans la télédétection en agriculture, Colwell (1956) montre que des photographies aériennes dans le spectre de l'infrarouge permettent de détecter les maladies des cultures agricoles. Plus tard, Kumar et Silva (1973) analysent en détail le lien entre l'activité chlorophyllienne d'une plante et sa réflectivité, i.e. la proportion d'énergie électromagnétique (la lumière en l'occurrence) réfléchi à la surface d'un matériau. Ils montrent alors que le spectre de réflectivité de la végétation comporte une signature spécifique dans le proche infrarouge ($\lambda = 700 - 1300$ nm) et le rouge visible ($\lambda = 550 - 700$ nm). La Figure 0.15 montre graphiquement cette signature dans le cas général, et son changement selon la santé de la plante considérée. Nous remarquons une forte augmentation de la réflectivité entre le rouge et le proche infrarouge, l'amplitude de cet écart diminuant avec la santé de la plante. Par ailleurs, nous notons un faible pic de réflectivité dans le vert : c'est la raison pour laquelle nous voyons la végétation en général de cette couleur. L'augmentation de la réflectivité dans le rouge pour les plantes en mauvaise santé donne l'explication des couleurs automnales.

Introduit par Rouse *et al.* (1974), le NDVI repose sur cette caractéristique du spectre électromagnétique de la végétation. Il est défini par

$$NDVI = \frac{R_{NIR} - R_R}{R_{NIR} + R_R}, \quad (0.40)$$

où R_{NIR} et R_R sont la réflectivité respectivement dans le proche infrarouge et le rouge. Une forêt tropicale a alors des valeurs comprises entre 0,6 et 0,8, le sol donne des NDVI proches de 0, enfin les pierres ou la neige mènent à des valeurs négatives. Le NDVI mesurant l'activité chlorophyllienne, son niveau fluctue au cours de l'année en fonction de la phase de développement et de la santé des plantes lorsque celles-ci sont sensibles aux saisons, comme les cultures agricoles.

Le NDVI est aujourd'hui l'un des indices de végétation les plus utilisés dans la télédétection en agriculture. Il permet notamment de prédire les rendements agricoles des cultures. La régression linéaire est souvent employée pour les modèles de prédiction de rendement de maïs (Prasad *et al.*, 2006), de soja (Ma *et al.*, 2001) ou de blé (Mkhabela *et al.*, 2011). D'autres

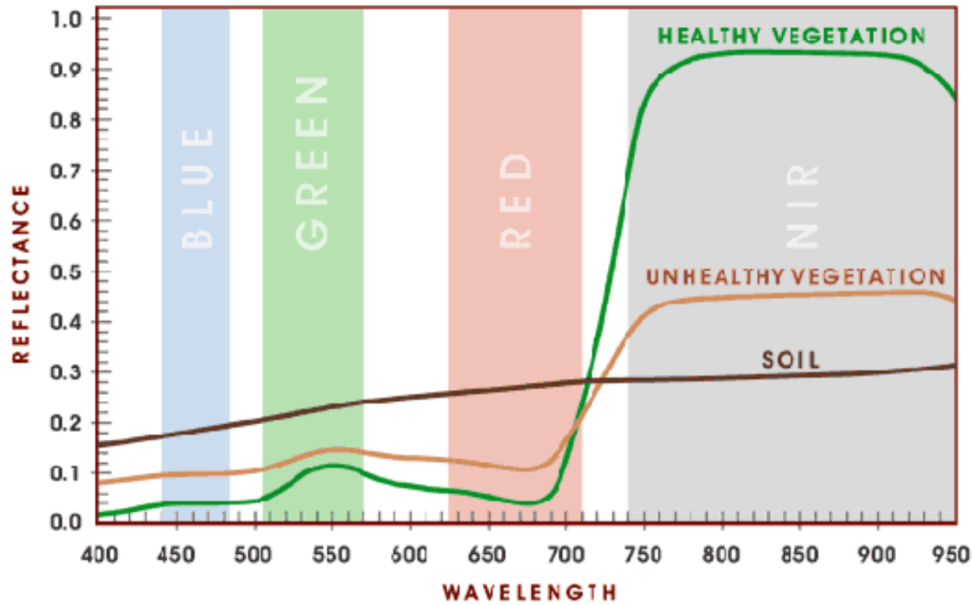


FIGURE 0.15 – Spectres de réflectivité de la végétation en bonne et mauvaise santé, et du sol.

Source : Giusti (2017)

algorithmes plus sophistiqués ont aussi été appliqués dans une moindre mesure, comme par exemple les réseaux de neurones (Li *et al.*, 2007). Le NDVI étant une série temporelle observée à plusieurs moments donnés durant la vie d'une culture, différentes variables d'agrégation peuvent être utilisées dans les modèles de projection de rendement. Ainsi, Mkhabela *et al.* (2011) exploitent la valeur moyenne du NDVI pendant la période de croissance, alors que Zhang *et al.* (2012) séparent la série temporelle en deux périodes distinctes : de la levée à la floraison, puis de la floraison à la maturation. Un grand nombre de modèles a été développé à cause des spécificités de chaque projet. Ces dernières peuvent être géographiques, des études ayant été menées dans beaucoup de pays comme, par exemple, le Zimbabwe (Svotwa *et al.*, 2014), la Hongrie (Ferencz *et al.*, 2004), la Chine (Ren *et al.*, 2008), ou les États-Unis (Becker-Reshef *et al.*, 2010b). Les particularités peuvent aussi être agronomiques, l'analyse pouvant être conduite sur différentes espèces de cultures : le maïs, le soja, le blé mais aussi l'orge, le riz, la tabac, les pommes de terre, la canne à sucre, etc. De plus, le NDVI peut parfois être associé à des données météorologiques, comme les précipitations ou des indices d'humidité, afin de d'améliorer la précision du modèle de prédiction (Prasad *et al.*, 2006). En général, ces études montrent une bonne prédiction des rendements agricoles par le NDVI, les modèles ayant régulièrement des R^2 jusqu'à 0,9.

Cette forte corrélation entre la santé des cultures et le NDVI a permis le développement d'outils de gestion de risque dans l'agriculture. L'observation de l'indice grâce au satellite a facilité le suivi de la progression des cultures partout dans le monde. Ainsi, des instruments de transfert de risques comme l'assurance indicielle sont désormais proposés aux agriculteurs. Les

indemnités reçues par ces derniers sont alors calculées en fonction de la série temporelle NDVI observée sur leur exploitation. Ces assurances peuvent porter sur les cultures dont nous venons de discuter (Turvey et McLaurin, 2012), mais aussi sur les prairies (Vroege *et al.*, 2019) ou même sur le bétail en mesurant la mortalité due à la sécheresse (Vrieling *et al.*, 2014). L'utilisation de cette structure indicielle est d'autant plus bénéfique pour les pays en voie de développement où elle permet aux agriculteurs la souscription d'assurance à un prix raisonnable (Smith et Watts, 2009; Miranda et Farrin, 2012; Jensen et Barrett, 2017). La récente démocratisation des drones, a amené à l'utilisation du NDVI en tant qu'outil de prévention des risques agricoles. Les drones associés à des spectroradiomètres adaptés permettent le suivi détaillé de chaque parcelle de culture. Ce suivi est plus spécifique que les images satellites, que ce soit au niveau spatial ou temporel, permettant alors une agriculture de précision (Puri *et al.*, 2017; Mogili et Deepak, 2018).

Avec l'amélioration de la résolution des données satellite, des sujets d'étude de télédétection, plus sophistiqués que la prédiction des rendements agricoles, ont émergés. Ainsi, certains chercheurs ont analysé la possibilité de la détection d'irrigation dans les cultures à l'aide des séries temporelles NDVI (Pervez et Brown, 2010; Peterson *et al.*, 2011). Cependant le sujet qui a le plus gagné en intérêt est la cartographie générale des cultures (e.g., Wardlow et Egbert, 2010; Zheng *et al.*, 2015; Skakun *et al.*, 2017). En effet, l'utilisation d'une telle cartographie permet d'améliorer significativement la précision des modèles de prédiction de rendements agricoles (Kastens *et al.*, 2005; Zhang *et al.*, 2019).

Plusieurs sources de données satellite NDVI sont disponibles en libre accès. Elles proviennent de différents programmes spatiaux, pour la plupart européens ou américains, et ont des résolutions spatiales et temporelles variées. Dans notre thèse, nous utilisons les informations transmises par les satellites *Terra* et *Aqua*, qui font partie du programme *Earth Observing System* (EOS) de la NASA. Ils embarquent tous les deux le *Moderate-Resolution Imaging Spectroradiometer* (MODIS) qui fournit, entre autres, des observations NDVI de l'ensemble de la terre depuis février 2000. L'accès aux données a été simplifié pour le public grâce au *Global Agriculture Monitoring Project* (Becker-Reshef *et al.*, 2010a), initié entre la NASA, l'Université du Maryland et le Service agriculture étrangère (FAS) de l'USDA. La résolution spatiale des données est de 250 mètres. Nous utilisons une résolution temporelle de 16 jours, i.e. tous les 16 jours nous observons la moyenne des valeurs NDVI sur la période passée. En effet, en tant qu'indice de réflectivité, le NDVI est sensible aux conditions météorologiques, comme par exemple les nuages, lorsqu'il est observé depuis un satellite (Whitcraft *et al.*, 2015).

Contributions

Pour conclure cette introduction générale, nous résumons désormais les contributions issues de cette thèse. Elles seront détaillées par la suite dans les Chapitres 1 à 4.

Chapitre 1 : Modèle VAR pénalisé pour la projection des taux d'amélioration de mortalité.

Dans ce chapitre, qui correspond à [Guibert *et al.* \(2019\)](#), nous proposons une modélisation VAR des logarithmes des taux de mortalité différenciés pour la projection de la longévité. Nous laissons ainsi une grande liberté dans la structure de dépendance spatio-temporelle à long terme des taux d'amélioration de mortalité. Afin de s'assurer de la parcimonie des matrices autorégressives, nous incluons une pénalisation de type elastic-net dans la procédure d'estimation, i.e. sans contrainte *a priori* sur la forme de la structure de dépendance. L'utilisation de cette régularisation permet au modèle de s'adapter aux spécificités de chaque population tout en évitant le surapprentissage. Nous soulignons quantitativement cette aptitude en effectuant une comparaison des performances de projection à court-moyen terme de notre approche sur neuf populations par rapport à cinq modèles de référence. Nos erreurs de prédiction sont en moyenne plus faibles, mais restent relativement proches de celles des modèles classiques comme le [Lee et Carter \(1992\)](#) ou le [Hyndman et Ullah \(2007\)](#). Au contraire, nous remarquons une augmentation significative de la stabilité des erreurs selon la dimension population. De plus, les projections à long terme sont cohérentes. Ces capacités de notre modèle en font une approche intéressante pour obtenir facilement de "bonnes" projections de mortalité, et ce quelque soit la population considérée.

La parcimonie des matrices de Granger, obtenue grâce à la régularisation lasso, permet de détecter graphiquement des effets démographiques. En particulier, notre modèle met en lumière les effets période et cohorte, déjà connus de la littérature. Cependant, nous remarquons qu'au contraire des approches déjà développées, l'effet cohorte est obtenu sans contrainte spécifique, i.e. nous ne forçons pas son estimation par des paramètres particuliers. En plus de ces effets période et cohorte, un autre type de motif récurrent est aussi observé dans les matrices autorégressive. Néanmoins, son interprétation ne correspond à aucun effet démographique étudié à notre connaissance, et nous n'en avons pas trouvé explicitement sa cause. Cette structure de dépendance estimée peut être causée par des phénomènes environnementaux ou sociétaux complexes, à moins qu'elle ne soit due à des anomalies dans les données de la HMD.

Enfin, nous montrons que notre approche peut être étendue à la modélisation de la mortalité dans le cadre de la multi-population sans que la question de la grande dimension ne vienne perturber la procédure d'estimation. Cette faculté est directement liée à la pénalisation utilisée. Bien que nous donnons les fondements d'une telle extension, nous n'analysons pas en détail cette application précise dans ce chapitre.

Chapitre 2 : Elastic-net et cohérence locale pour la modélisation des dynamiques de mortalité inter-population.

Le second chapitre de cette thèse s'inscrit dans la continuité du premier dans le sens où une problématique de grande dimension rencontrée dans le cadre de la modélisation des dynamiques de mortalité est traitée à l'aide d'un VAR estimé grâce à une pénalisation elastic-net. Néanmoins, nous nous concentrons ici sur les dynamiques inter-population : le VAR est appliqué sur les facteurs temporels issus de l'estimation de modèles Lee et Carter (1992) ou Li et Lee (2005). La difficulté réside alors dans la volonté de modéliser simultanément un nombre important de populations, nous prenons pour l'exemple de notre étude un groupe de 16 pays européens dont les populations sont divisées par sexes. Notre modèle propose en outre une nouvelle notion, que nous définissons de "locale", de la cohérence des projections de taux de mortalité dans le cadre de la multi-population. Au lieu qu'aucune population ne puisse diverger en terme de dynamique de mortalité (i.e. la définition de la cohérence), nous classifions les populations par groupe de cohérence : les dynamiques de mortalité au sein du même groupe ne peuvent diverger mais cette contrainte est relâchée pour les dynamiques inter-groupe.

En effet, la propriété de cohérence, qu'une large majorité de la littérature suppose pour les modèles multi-population, a récemment été remise en cause à la vue des données historiques à disposition (Li *et al.*, 2017). Pour étudier cette hypothèse dans le cadre de multiples populations, nous proposons une mesure de dispersion des taux de mortalité. Ainsi, nous montrons pour l'exemple des 32 populations européennes que le modèle cohérent Li-Lee implique des projections qualitativement insatisfaisantes de cette dispersion. Les dynamiques sont trop peu dispersées entre elles au vue de l'historique, créant alors une concentration artificielle en terme de risque de longévité. A l'opposé, des Lee-Carter estimés indépendamment provoquent une diversification exagérée. Notre modèle permet alors l'obtention d'un large spectre de possibilités intermédiaires entre ces deux extrêmes. Nous proposons deux exemples de classification : l'un fondé sur des jugements d'expert, l'autre défini par une approche plus statistique. Comme attendu, les projections de la dispersion sont plus satisfaisante.

Afin de montrer l'importance que ces choix de cohérence peuvent avoir pour la gestion des risques de longévité, nous nous plaçons dans le cadre d'un assureur vie exposé aux 32 populations pour un contrat de type retraite simplifié. Nous estimons alors le capital réglementaire (SCR) prescrit par la directive Solvabilité II grâce des scénarios stochastiques de mortalité générés par les différents modèles. Même si la provision est sensiblement la même, le SCR est significativement impacté selon l'hypothèse de cohérence retenue : globale, locale ou inexistante.

Finalement, nous étendons la localité de la propriété de cohérence du modèle à la dimension temporelle. En permettant aux populations de passer d'un groupe de cohérence à un autre, nous élargissons le spectre des scénarios stochastiques de mortalité possibles. Cette méthodologie innovante est particulièrement intéressante dans un cadre d'évaluation du risque de base des transferts du risque de longévité. Nous proposons par ailleurs l'analyse d'un exemple concret, le Longevity Divergence Index Value, à l'aide de cette approche.

Chapitre 3 : Apprentissage statistique et mesures économiques pour la gestion du risque de rachat.

Ce chapitre, correspondant à [Loisel *et al.* \(2019\)](#), analyse la segmentation du risque de rachat par rapport aux facteurs structurels d'un portefeuille de contrats d'assurance vie, environ 600 000, provenant d'une compagnie taïwanaise de taille moyenne. Pour ce faire nous y appliquons deux modèles de classification supervisée de référence : la régression logistique et le CART. Nous utilisons aussi des techniques plus sophistiquées d'apprentissage statistique : le SVM et le XGBoost. Afin de comparer la performance de classification des différents algorithmes, nous effectuons une validation croisée et analysons la précision totale des prédictions. Comme attendu, le SVM et le XGBoost offrent de meilleures prévisions.

En se référant à la littérature sur le risque d'attrition, nous proposons une nouvelle mesure des performances de prédiction des modèles. Cette dernière estime le profit économique d'efforts de rétention sur les assurés classifiés comme à risque par les algorithmes de segmentation. La comparaison des modèles en est alors fortement modifiée. En particulier, nous remarquons que le XGBoost, qui est significativement meilleur que le SVM en terme de précision statistique, amène désormais à des résultats peu différents par rapport à ce dernier. De plus, la régression logistique donne de meilleures classifications que le CART avec la vision économique, alors que nous observons le contraire lors de la comparaison basée sur la mesure purement statistique.

S'inspirant de [Lemmens et Gupta \(2017\)](#), nous proposons alors une méthodologie pour optimiser la mesure de profit économique. Nous transformons la problématique de la classification du risque de rachat, en une régression du profit attendu de l'effort de rétention. Pour ce faire nous modifions la fonction de perte utilisée dans l'algorithme du XGBoost. Si la prédiction du gain est positive, l'assuré est alors classifié comme risqué. La variable à optimiser ayant changé, nous constatons des modifications manifestes dans les performances de prédiction :

- la précision statistique diminue fortement, atteignant même des niveaux inférieurs à la régression logistique dans certains cas ;
- la mesure de profit augmente significativement par rapport aux algorithmes de classification classique.

Cette étude montre ainsi l'importance du choix de la fonction de perte selon l'objectif exact de l'analyse menée par l'actuaire.

Chapitre 4 : Potentiel des données satellite dans la gestion des risques financiers agricoles.

Dans ce chapitre, tiré de [Piette \(2019\)](#), nous explorons la possibilité de prédire une partie des informations contenues dans les rapports USDA à partir de données satellite en libre accès : les séries NDVI MODIS. Premièrement, grâce à des tests statistiques, nous montrons que le

marché des *futures* sur le maïs réagit à la publication des rapports WASDE et *Crop Production* pendant la période retenue (2000–2016), suggérant que des informations de valeur s’y trouvent pour les agents du marché. Ensuite, nous soulignons à l’aide d’une approche de régression, que ces dernières peuvent, en partie, être expliquées par les projections des rendements agricoles fournis par le NASS.

Dans un second temps, nous proposons un modèle de prédiction de ces mêmes rendements agricoles grâce aux séries temporelles NDVI. Au lieu d’envoyer des agents dans les champs et interroger les agriculteurs sur la progression de leur plantation comme peut le faire l’USDA, nous nous basons seulement sur des images satellites libre d’accès. Dans les deux cas, les observations sont fortement corrélées avec la santé des cultures. En appliquant la même méthodologie qu’avec les projections gouvernementales, nous trouvons alors que les informations obtenues grâce au NDVI sont rationnellement et significativement corrélées avec les réactions de marché. Ceci ouvre la possibilité du développement d’un outil de gestion de risques financiers fondé sur un modèle utilisant les images satellite et permettant de connaître en avance les informations contenues dans les rapports gouvernementaux.

Enfin, nous proposons des pistes d’amélioration à examiner avant une mise en pratique de ce potentiel. En effet, notre approche doit être plutôt vue en tant que démonstration de faisabilité (*proof of concept*), et non comme une méthodologie directement applicable. Plus généralement, cette étude montre que la gestion des risques peut bénéficier de l’intégration d’information issues de l’*open data* dans les modèles d’apprentissage statistique.

Bibliographie

- Philip ABBOTT, David BOUSSIOS et Jess LOWENBERG DEBOER : Valuing Public Information in Agricultural Commodity Markets : WASDE Corn Reports. *In NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management.*, St. Louis, MO, 2016.
- Michael K. ADJEMIAN : Quantifying the WASDE Announcement Effect. *American Journal of Agricultural Economics*, 94(1):238–256, janvier 2012. ISSN 0002-9092.
- Michael K. ADJEMIAN et Scott H. IRWIN : USDA Announcement Effects in Real-Time. *American Journal of Agricultural Economics*, 100(4):1151–1171, juillet 2018. ISSN 0002-9092.
- George A. AKERLOF : The Market for “Lemons” : Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, août 1970. ISSN 0033-5533.
- Marco ALEANDRI : Modeling Dynamic Policyholder Behavior through Machine Learning Techniques. *Working paper*, 2017. Available at <https://www.dss.uniroma1.it/it/node/6829>.
- Katrien ANTONIO, Sander DEVRIENDT, Wouter de BOER, Robert de VRIES, Anja DE WAEGENAERE, Hok-Kwan KAN, Egbert KROMME, Wilbert OUBURG, Tim SCHULTEIS, Erica SLAGTER, Marco van der WINDEN, Corné van IERSEL et Michel VELLEKOOP : Producing the Dutch and Belgian mortality projections : a stochastic multi-population standard. *European Actuarial Journal*, 7(2):297–336, décembre 2017. ISSN 2190-9741.
- Kenneth J. ARROW et Gerard DEBREU : Existence of an Equilibrium for a Competitive Economy. *Econometrica*, 22(3):265–290, 1954. ISSN 0012-9682.
- Eva ASCARZA : Retention Futility : Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, 55(1):80–98, février 2018. ISSN 0022-2437.
- Eva ASCARZA, Scott A. NESLIN, Oded NETZER, Zachery ANDERSON, Peter S. FADER, Sunil GUPTA, Bruce G. S. HARDIE, Aurelie LEMMENS, Barak LIBAI, David NEAL, Foster PROVOST et Rom SCHRIFT : In Pursuit of Enhanced Customer Retention Management : Review, Key Issues, and Future Directions. *Customer Needs and Solutions*, 5(1):65–81, mars 2018. ISSN 2196-2928.
- Louis BACHELIER : Théorie de la spéculation. *Annales scientifiques de l’Ecole normale supérieure*, 3(17):21–86, 1900.
- Anna Rita BACINELLO : Pricing Guaranteed Life Insurance Participating Policies with Annual Premiums and Surrender Option. *North American Actuarial Journal*, 7(3):1–17, juillet 2003. ISSN 1092-0277.

- Anna Rita BACINELLO : Endogenous model of surrender conditions in equity-linked life insurance. *Insurance : Mathematics and Economics*, 37(2):270–296, octobre 2005. ISSN 0167-6687.
- Kerry BACK : Insider Trading in Continuous Time. *The Review of Financial Studies*, 5 (3):387–409, juillet 1992. ISSN 0893-9454.
- Pauline BARRIEU, Harry BENSUSAN, Nicole EL KAROUI, Caroline HILLAIRET, Stéphane LOISEL, Claudia RAVANELLI et Yahia SALHI : Understanding, modelling and managing longevity risk : key issues and main challenges. *Scandinavian Actuarial Journal*, 2012(3):203–231, 2012. ISSN 0346-1238.
- Daniel BAUER, Jin GAO, Thorsten MOENIG, Eric R. ULM et Nan ZHU : Policyholder Exercise Behavior in Life Insurance : The State of Affairs. *North American Actuarial Journal*, 21 (4):485–501, octobre 2017. ISSN 1092-0277.
- Daniel BAUER, Rüdiger KIESEL, Alexander KLING et Jochen RUSS : Risk-neutral valuation of participating life insurance contracts. *Insurance : Mathematics and Economics*, 39(2):171–183, octobre 2006. ISSN 0167-6687.
- I. BECKER-RESHEF, C. JUSTICE, M. SULLIVAN, E. VERMOTE et C. TUCKER : The Global Agriculture Monitoring (GLAM) project. *Remote Sensing*, 2(6):1589–1609, 2010a.
- I. BECKER-RESHEF, E. VERMOTE, M. LINDEMAN et C. JUSTICE : A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment*, 114(6):1312–1323, 2010b.
- Rémi BELLINA : *Méthodes d'apprentissage appliquées à la tarification non-vie*. Mémoire d'actuaire, ISFA, 2014.
- Paul D. BERGER et Nada I. NASR : Customer lifetime value : Marketing models and applications. *Journal of Interactive Marketing*, 12(1):17–30, 1998. ISSN 1520-6653.
- Hans BÜHLMANN et Alois GISLER : *A Course in Credibility Theory and its Applications*. Springer Science & Business Media, mars 2006. ISBN 978-3-540-29273-9.
- Fischer BLACK et Myron SHOLES : The Valuation of Option Contracts and a Test of Market Efficiency. *The Journal of Finance*, 27(2):399–417, 1972. ISSN 1540-6261.
- Fischer BLACK et Myron SHOLES : The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654, mai 1973. ISSN 0022-3808.
- D. BLAKE, A. J. G. CAIRNS et K. DOWD : Living with Mortality : Longevity Bonds and Other Mortality-Linked Securities. *British Actuarial Journal*, 12(1):153–197, mars 2006. ISSN 2044-0456, 1357-3217.

- D. BLAKE, A. J. G. CAIRNS, K. DOWD et A. R. KESSLER : Still living with mortality : the longevity risk transfer market after one decade. *British Actuarial Journal*, 24, 2018. ISSN 1357-3217, 2044-0456.
- David BLAKE et William BURROWS : Survivor Bonds : Helping to Hedge Mortality Risk. *Journal of Risk and Insurance*, pages 339–348, 2001.
- Richard J. BOLTON et David J. HAND : Statistical Fraud Detection : A Review. *Statistical Science*, 17(3):235–249, 2002. ISSN 0883-4237.
- Bernhard E. BOSER, Isabelle M. GUYON et Vladimir N. VAPNIK : A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 978-0-89791-497-0. event-place : Pittsburgh, Pennsylvania, USA.
- Alexandre BOUMEZOUED : *Approches micro-macro des dynamiques de populations hétérogènes structurées par âge. Application aux processus auto-excitants et à la démographie*. PhD Thesis, Université Pierre et Marie Curie, avril 2016a.
- Alexandre BOUMEZOUED : Improving HMD mortality estimates with HFD fertility data. *Working paper*, février 2016b.
- Alexandre BOUMEZOUED, Nicole EL KAROUI et Stéphane LOISEL : Measuring mortality heterogeneity with multi-state models and interval-censored data. *Insurance : Mathematics and Economics*, 72:67–82, janvier 2017. ISSN 0167-6687.
- Leo BREIMAN : Bagging predictors. *Machine Learning*, 24(2):123–140, août 1996. ISSN 1573-0565.
- Leo BREIMAN : Random Forests. *Machine Learning*, 45(1):5–32, octobre 2001. ISSN 1573-0565.
- Leo BREIMAN, J. FRIEDMAN, R. OLSHEN et C. STONE : *Classification and Regression Trees*. Wadsworth, 1984. ISBN 978-1-351-46049-1.
- J. BUREZ et D. Van den POEL : Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, avril 2009. ISSN 0957-4174.
- Jonathan BUREZ et Dirk Van den POEL : CRM at a pay-TV company : Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2):277–288, février 2007. ISSN 0957-4174.
- Andrew J. G. CAIRNS, David BLAKE et Kevin DOWD : A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty : Theory and Calibration. *Journal of Risk and Insurance*, 73(4):687–718, décembre 2006. ISSN 1539-6975.

- Andrew J. G. CAIRNS, David BLAKE, Kevin DOWD, Guy D. COUGHLAN, David EPSTEIN, Alen ONG et Igor BALEVICH : A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35, janvier 2009. ISSN 1092-0277.
- Andrew J. G. CAIRNS, David BLAKE, Kevin DOWD et Amy R. KESSLER : Phantoms never die : living with unreliable population data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 179(4):975–1005, 2016. ISSN 1467-985X.
- Andrew J. G. CAIRNS, Malene KALLESTRUP-LAMB, Carsten ROSENSKJOLD, David BLAKE et Kevin DOWD : Modelling Socio-Economic Differences in Mortality Using a New Affluence Index. *ASTIN Bulletin : The Journal of the IAA*, pages 1–36, 2019. ISSN 0515-0361, 1783-1350.
- J. CAMPBELL, M. CHAN, K. LI, L. LOMBARDI, M. PURUSHOTHAM et A. RAO : Modeling of Policyholder Behavior for Life Insurance and Annuity Products : A Survey and Literature Review. *Society of Actuaries*, 2014.
- Rocco Roberto CERCHIARA, Matthew EDWARDS et Alessandra GAMBINI : Generalized Linear Models in Life Insurance : Decrements and Risk Factor under Solvency II. In *18th international AFIR colloquium*, Rome, 2008.
- Arthur CHARPENTIER : *Computational Actuarial Science with R*. Crc press édition, 2014.
- Arthur CHARPENTIER, Emmanuel FLACHAIRE et Antoine LY : Econometrics and Machine Learning. *Economie et Statistique*, 505(1):147–169, 2018.
- Rong-Chang CHEN, Ming-Li CHIU, Ya-Li HUANG et Lin-Ti CHEN : Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines. In Zheng Rong YANG, Hujun YIN et Richard M. EVERSON, éditeurs : *Intelligent Data Engineering and Automated Learning – IDEAL 2004*, Lecture Notes in Computer Science, pages 800–806. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-28651-6.
- Tianqi CHEN et Carlos GUESTRIN : XGBoost : A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. event-place : San Francisco, California, USA.
- Jin Wook CHOI : *An analysis of price responses to public information : a case study of the USDA corn crop forecasts*. Thèse de doctorat, Iowa State University, janvier 1982.
- Marcus C. CHRISTIANSEN, Evgeny SPODAREV et Verena UNSELD : Differences in European Mortality Rates : A Geometric Approach on the Age-Period Plane. *ASTIN Bulletin : The Journal of the IAA*, 45(3):477–502, septembre 2015. ISSN 0515-0361, 1783-1350.

- Peter CHRISTOFFERSEN et Kris JACOBS : The importance of the loss function in option valuation. *Journal of Financial Economics*, 72(2):291–318, mai 2004. ISSN 0304-405X.
- Phil L. COLLING et Scott H. IRWIN : The Reaction of Live Hog Futures Prices to USDA Hogs and Pigs Reports. *American Journal of Agricultural Economics*, 72(1):84–94, février 1990. ISSN 0002-9092.
- Phil L. COLLING, Scott H. IRWIN et Carl R. ZULAUF : Reaction of Wheat, Corn, and Soybean Futures Prices to USDA "Export Inspections" Reports. *Review of Agricultural Economics*, 18(1):127–136, 1996. ISSN 1058-7195.
- R. COLWELL : Determining the prevalence of certain cereal crop diseases by means of aerial photography. *Hilgardia*, 26(5):223–286, novembre 1956. ISSN 0073-2230.
- Andrea CONSIGLIO et Domenico De GIOVANNI : Pricing the Option to Surrender in Incomplete Markets. *Journal of Risk and Insurance*, 77(4):935–957, 2010. ISSN 1539-6975.
- Corinna CORTES et Vladimir VAPNIK : Support-vector networks. *Machine Learning*, 20(3):273–297, septembre 1995. ISSN 1573-0565.
- Guy COUGHLAN, David EPSTEIN, Amit SINHA et Paul HONIG : q-Forwards : Derivatives for transferring longevity and mortality risk. Rapport technique, London, 2007.
- Guy D. COUGHLAN, Marwa KHALAF-ALLAH, Yijing YE, Sumit KUMAR, Andrew J. G. CAIRNS, David BLAKE et Kevin DOWD : Longevity Hedging 101. *North American Actuarial Journal*, 15(2):150–176, avril 2011. ISSN 1092-0277.
- Kristof COUSSEMENT et Dirk Van den POEL : Churn prediction in subscription services : An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327, janvier 2008. ISSN 0957-4174.
- Samuel COX, H. et Yija LIN : Annuity Lapse Modeling : Tobit or not Tobit ? *Society of Actuaries*, 2006.
- Léopold CRESTEL et Philippe ESLING : Live Orchestral Piano, a system for real-time orchestral music generation. *arXiv :1609.01203 [cs]*, septembre 2016. arXiv : 1609.01203.
- I. D. CURRIE : Smoothing and forecasting mortality rates with P-splines, juin 2006.
- Iain D CURRIE, Maria DURBAN et Paul HC EILERS : Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298, 2004. ISSN 1471-082X.
- Ivan Luciano DANESI, Steven HABERMAN et Pietro MILLOSOVICH : Forecasting mortality in subpopulations using Lee–Carter type models : A comparison. *Insurance : Mathematics and Economics*, 62:151–161, mai 2015. ISSN 0167-6687.

- A. DAR et C. DODDS : Interest Rates, the Emergency Fund Hypothesis and Saving through Endowment Policies : Some Empirical Evidence for the U.K. *The Journal of Risk and Insurance*, 56(3):415–433, 1989. ISSN 0022-4367.
- Edouard DEBONNEUIL, Stéphane LOISEL et Frédéric PLANCHET : Do actuaries believe in longevity deceleration? *Insurance : Mathematics and Economics*, 78:325–338, janvier 2018. ISSN 0167-6687.
- Philippe DEPREZ, Pavel V. SHEVCHENKO et Mario V. WÜTHRICH : Machine learning techniques for mortality modeling. *Eur. Actuar. J.*, 7(2):337–352, décembre 2017. ISSN 2190-9741.
- Alexander DOKUMENTOV, Rob J. HYNDMAN et Leonie TICKLE : Bivariate smoothing of mortality surfaces with cohort and period ridges. *Stat*, 7(1):e199, 2018. ISSN 2049-1573.
- J. C. DOLLEY : Characteristics an procedure of common stock split-ups. *Harvard Business Review*, 11(3):316–326, 1933.
- Jeffrey H. DORFMAN et Berna KARALI : A Nonparametric Search for Information Effects from USDA Reports. *Journal of Agricultural and Resource Economics*, 40(1):124–143, 2015. ISSN 1068-5502.
- P. DOUKHAN, D. POMMERET, J. RYNKIEWICZ et Y. SALHI : A class of random field memory models for mortality forecasting. *Insurance : Mathematics and Economics*, 77:97–110, 2017. ISSN 0167-6687.
- Kevin DOWD, Andrew J. G. CAIRNS, David BLAKE, Guy D. COUGHLAN et Marwa KHALAF-ALLAH : A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, 15(2):334–356, 2011. ISSN 1092-0277.
- A. E. EIBEN, A. E. KOUDIJS et F. SLISSER : Genetic modelling of customer retention. In Wolfgang BANZHAF, Riccardo POLI, Marc SCHOENAUER et Terence C. FOGARTY, éditeurs : *Genetic Programming*, Lecture Notes in Computer Science, pages 178–186. Springer Berlin Heidelberg, 1998. ISBN 978-3-540-69758-9.
- EIOPA : Report on the fifth Quantitative Impact Study (QIS5) for Solvency II. Rapport technique EIOPA-TFQIS5-11/001, 2011.
- Martin ELING et Dieter KIESENBAUER : What Policy Features Determine Life Insurance Lapse? An Analysis of the German Market. *Journal of Risk and Insurance*, 81(2):241–269, 2014. ISSN 1539-6975.
- Martin ELING et Michael KOCHANSKI : Research on lapse in life insurance : what has been done and what needs to be done? *The Journal of Risk Finance*, 14(4):392–413, août 2013. ISSN 1526-5943.

- Vasil ENCHEV, Torsten KLEINOW et Andrew J. G. CAIRNS : Multi-population mortality models : fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342, janvier 2016. ISSN 0346-1238.
- Robert F. ENGLE : On the limitations of comparing mean square forecast errors : Comment. *Journal of Forecasting*, 12(8):642–644, 1993. ISSN 1099-131X.
- Robert F. ENGLE et C. W. J. GRANGER : Co-Integration and Error Correction : Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276, 1987. ISSN 0012-9682.
- Anne EYRAUD-LOISEL : *EDSR et EDSPR avec grossissement de filtration, problèmes d’asymétrie d’information et de couverture sur les marchés financiers*. Thèse de doctorat, Université Paul Sabatier, Toulouse 3, décembre 2005.
- C. FERENCZ, P. BOGNÁR, J. LICHTENBERGER, D. HAMAR, Gy TARCSAI†, G. TIMÁR, G. MOLNÁR, SZ PÁSZTOR, P. STEINBACH, B. SZÉKELY, O. E. FERENCZ et I. FERENCZ-ÁRKOS : Crop yield estimation by satellite remote sensing. *International Journal of Remote Sensing*, 25(20):4113–4149, octobre 2004. ISSN 0143-1161.
- T.R. FORTENBERY et D. A. SUMNER : The effects of USDA reports in futures and options markets. *Journal of Futures Markets*, 13:157–173, 1993.
- Jerome H. FRIEDMAN : Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.
- Jerome H. FRIEDMAN : Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, février 2002. ISSN 0167-9473.
- Philip GARCIA, Scott H. IRWIN, Raymond M. LEUTHOLD et Li YANG : The value of public information in commodity futures markets. *Journal of Economic Behavior & Organization*, 32(4):559–570, avril 1997. ISSN 0167-2681.
- Natalia S. GAVRILOVA et Leonid A. GAVRILOV : Are We Approaching a Biological Limit to Human Longevity? *The Journals of Gerontology : Series A*, juillet 2019.
- Ronald GELARO, Will McCARTY, Max J. SUÁREZ, Ricardo TODLING, Andrea MOLOD, Lawrence TAKACS, Cynthia A. RANDLES, Anton DARMENOV, Michael G. BOSILOVICH, Rolf REICHLÉ, Krzysztof WARGAN, Lawrence COY, Richard CULLATHER, Clara DRAPER, Santha AKELLA, Virginie BUCHARD, Austin CONATY, Arlindo M. da SILVA, Wei GU, Gi-Kong KIM, Randal KOSTER, Robert LUCCHESI, Dagmar MERKOVA, Jon Eric NIELSEN, Gary PARTYKA, Steven PAWSON, William PUTMAN, Michele RIENECKER, Siegfried D. SCHUBERT, Meta SIENKIEWICZ et Bin ZHAO : The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, mai 2017. ISSN 0894-8755.

- Guillaume GERBER, Yohann LE FAOU, Olivier LOPEZ et Michael TRUPIN : The impact of churn on client value in health insurance, evaluation using a random forest under random censoring. *Working paper*, juin 2018.
- Alessio GIUSTI : NDVI Index. *PhysicsOpenLab*, janvier 2017.
- Nicolas GLADY, Bart BAESENS et Christophe CROUX : Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197(1):402–411, août 2009. ISSN 0377-2217.
- Clara-Cecilie GÜNTHER, Ingunn Fride TVETE, Kjersti AAS, Geir Inge SANDNES et Ørnulf BORGAN : Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1):58–71, février 2014. ISSN 0346-1238.
- Benjamin GOMPERTZ : On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583, 1825.
- Christophe GOUEL : The Value of Public Information in Storable Commodity Markets : Application to the Soybean Market. In *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, Minneapolis, Minnesota, 2018.
- C. W. J. GRANGER : Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. ISSN 0012-9682.
- Anders GROSEN et Peter LØCHTE JØRGENSEN : Fair valuation of life insurance liabilities : The impact of interest rate guarantees, surrender options, and bonus policies. *Insurance : Mathematics and Economics*, 26(1):37–57, février 2000. ISSN 0167-6687.
- Orlen GRUNEWALD, Mark S. McNULTY et Arlo W. BIERE : Live Cattle Futures Response to Cattle on Feed Reports. *Am J Agric Econ*, 75(1):131–137, février 1993. ISSN 0002-9092.
- Quentin GUIBERT, Olivier LOPEZ et Pierrick PIETTE : Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance : Mathematics and Economics*, 88:255–272, septembre 2019. ISSN 0167-6687.
- G. GUO, S. Z. LI et K. CHAN : Face recognition by support vector machines. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 196–201, Grenoble, France, mars 2000.
- Sunil GUPTA, Dominique HANSENS, Bruce HARDIE, Wiliam KAHN, V. KUMAR, Nathaniel LIN, Nalini RAVISHANKER et S. SRIRAM : Modeling Customer Lifetime Value. *Journal of Service Research*, 9(2):139–155, novembre 2006. ISSN 1094-6705.

- Sunil GUPTA et Donald R. LEHMANN : Customers as assets. *Journal of Interactive Marketing*, 17(1):9–24, janvier 2003. ISSN 1094-9968.
- Sunil GUPTA et Donald R. LEHMANN : Customer Lifetime Value and Firm Valuation. *Journal of Relationship Marketing*, 5(2-3):87–110, octobre 2006. ISSN 1533-2667.
- Sunil GUPTA, Donald R. LEHMANN et Jennifer Ames STUART : Valuing Customers. *Journal of Marketing Research*, 41(1):7–18, février 2004. ISSN 0022-2437.
- John HADDEN, Ashutosh TIWARI, Rajkumar ROY et Dymitr RUTA : Churn Prediction : Does Technology Matter? *International Journal of Intelligent Technology*, 1(2):104–110, 2006. ISSN 1305-6417.
- Lukas Josef HAHN : A Bayesian Multi-Population Mortality Projection Model. Mémoire de D.E.A., Universität Ulm, Ulm, Germany, décembre 2014.
- Donatien HAINAUT : A Neural-Network Analyzer for Mortality Forecast. *ASTIN Bulletin : The Journal of the IAA*, 48(2):481–508, mai 2018. ISSN 0515-0361, 1783-1350.
- M. R. HARDY et H. H. PANJER : A Credibility Approach to Mortality Risk. *ASTIN Bulletin : The Journal of the IAA*, 28(2):269–283, novembre 1998. ISSN 0515-0361, 1783-1350.
- Scott E. HARRINGTON : The Financial Crisis, Systemic Risk, and the Future of Insurance Regulation. *Journal of Risk and Insurance*, 76(4):785–819, 2009. ISSN 1539-6975.
- Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, NY, 2nd edition édition, 2016. ISBN 978-0-387-84857-0.
- Caroline HILLAIRET : *Equilibres sur un marché financier avec asymétrie d'information et discontinuité des prix*. Thèse de doctorat, Université Paul Sabatier, Toulouse 3, novembre 2004.
- HMD : *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org (data downloaded on 2019-07-08)., 2019.
- Arthur E. HOERL et Robert W. KENNARD : Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, février 1970. ISSN 0040-1706.
- Ines HOLZMÜLLER : The United States RBC Standards, Solvency II and the Swiss Solvency Test : A Comparative Assessment. *Geneva Pap Risk Insur Issues Pract*, 34(1):56–77, janvier 2009. ISSN 1468-0440.

- Bingquan HUANG, Mohand Tahar KECHADI et Brian BUCKLEY : Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, janvier 2012. ISSN 0957-4174.
- Shin-Yuan HUNG, David C. YEN et Hsiu-Yu WANG : Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, octobre 2006. ISSN 0957-4174.
- M M HUYNEN, P MARTENS, D SCHRAM, M P WEIJENBERG et A E KUNST : The impact of heat waves and cold spells on mortality rates in the Dutch population. *Environ Health Perspect*, 109(5):463–470, mai 2001. ISSN 0091-6765.
- Hyunseok HWANG, Taesoo JUNG et Euiho SUH : An LTV model and customer segmentation based on customer value : a case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26(2):181–188, février 2004. ISSN 0957-4174.
- Rob J. HYNDMAN et Shahid ULLAH : Robust forecasting of mortality and fertility rates : A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956, 2007. ISSN 0167-9473.
- Scott H. IRWIN, Darrel L. GOOD et Jennifer K. GOMEZ : The Value of USDA Outlook Information : An Investigation using event Study Analysis. *In NCR Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, St. Louis MO., 2001.
- Olga ISENGILDINA, Scott H. IRWIN et Darrel L. GOOD : The Value of USDA Situation and Outlook Information in Hog and Cattle Markets. *Journal of Agricultural and Resource Economics*, 31(2):262–282, 2006. ISSN 1068-5502.
- Olga ISENGILDINA-MASSA, Scott H. IRWIN, Darrel L. GOOD et Jennifer K. GOMEZ : The Impact of Situation and Outlook Information in Corn and Soybean Futures Markets : Evidence from WASDE Reports. *Journal of Agricultural and Applied Economics*, 40(1):89–103, avril 2008. ISSN 1074-0708, 2056-7405.
- Etienne IZRAELEWICZ : L'effet moisson - l'impact des catastrophes vie sur la mortalité à long terme - Exemple de la canicule de l'été 2003. *Bulletin Français d'Actuariat*, 12(24):113–159, 2012.
- Salma JAMAL : Lapse Risk Modeling with Machine Learning Techniques : an Application to Structural Lapse Drivers. *Bulletin Francais d'Actuariat*, 17(33):27–91, 2017.
- Fanny JANSSEN : Advances in mortality forecasting : introduction. *Genus*, 74(1):21, décembre 2018. ISSN 2035-5556.

- Marijn JANSSEN, Yannis CHARALABIDIS et Anneke ZUIDERWIJK : Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29 (4):258–268, septembre 2012. ISSN 1058-0530.
- Søren Fiig JARNER et Esben Masotti KRYGER : Modelling Adult Mortality in Small Populations : The Saint Model. *ASTIN Bulletin : The Journal of the IAA*, 41(2):377–418, novembre 2011. ISSN 0515-0361, 1783-1350.
- Bjarke JENSEN, Peter Løchte JØRGENSEN et Anders GROSEN : A Finite Difference Approach to the Valuation of Path Dependent Life Insurance Liabilities. *Geneva Risk Insur Rev*, 26 (1):57–84, juin 2001. ISSN 1573-6954.
- Nathaniel JENSEN et Christopher BARRETT : Agricultural Index Insurance for Development. *Appl Econ Perspect Policy*, 39(2):199–219, juin 2017. ISSN 2040-5790.
- T. JEULIN : *Semi-martingales et grossissement d'une filtration*. Numéro 833 in Lecture Notes in Mathematics. Springer, 1980.
- Thierry JEULIN et Marc YOR : *Grossissements de filtrations : exemples et applications : Seminaire de Calcul Stochastique 1982/83 Universite Paris VI*. Numéro 11118 in Lecture Notes in Mathematics. Springer, 1985.
- Yusho KAGRAOKA : Modeling Insurance Surrenders by the Negative Binomial Model. *Working paper*, 2005.
- Aymric KAMEGA et Frédéric PLANCHET : Modélisation prospective en l'absence de données sur les tendances passées et mesure des risques associés. *Assurances et Gestion des Risques*, 80(2), 2011.
- Berna KARALI : Do USDA Announcements Affect Comovements Across Commodity Futures Returns? *Journal of Agricultural and Resource Economics*, 37(1):77–97, 2012. ISSN 1068-5502.
- Berna KARALI, Olga ISENGILDINA-MASSA, Scott H. IRWIN, Michael K. ADJEMIAN et Robert JOHANSSON : Are USDA reports still news to changing crop markets? *Food Policy*, mars 2019. ISSN 0306-9192.
- Jude H. KASTENS, Terry L. KASTENS, Dietrich L. A. KASTENS, Kevin P. PRICE, Edward A. MARTINKO et Re-Yang LEE : Image masking for crop yield forecasting using AVHRR NDVI time series imagery. *Remote Sensing of Environment*, 99(3):341–356, novembre 2005. ISSN 0034-4257.
- Dieter KIESENBAUER : Main Determinants of Lapse in the German Life Insurance Industry. *North American Actuarial Journal*, 16(1):52–73, janvier 2012. ISSN 1092-0277.

- Changki KIM : Modeling Surrender and Lapse Rates With Economic Variables. *North American Actuarial Journal*, 9(4):56–70, octobre 2005a. ISSN 1092-0277.
- Changki KIM : Surrender Rate Impacts on Asset Liability Management. *Asia-Pacific Journal of Risk and Insurance*, 1(1), 2005b.
- Torsten KLEINOW : A common age effect model for the mortality of multiple populations. *Insurance : Mathematics and Economics*, 63:147–152, 2015. ISSN 0167-6687.
- Alexander KLING, Frederik RUEZ et Jochen RUSS : The impact of policyholder behavior on pricing, hedging, and hedge efficiency of withdrawal benefit guarantees in variable annuities. *European Actuarial Journal*, 4(2):281–314, décembre 2014. ISSN 2190-9741.
- Dudyala Anil KUMAR et Vadlamani RAVI : Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1):4–28, 2008.
- R. KUMAR et L. SILVA : Light Ray Tracing Through a Leaf Cross Section. *Appl. Opt., AO*, 12(12):2950–2954, décembre 1973. ISSN 2155-3165.
- Weiyu KUO, Chenghsien TSAI et Wei-Kuang CHEN : An Empirical Study on the Lapse Rate : The Cointegration Approach. *Journal of Risk and Insurance*, 70(3):489–508, 2003. ISSN 1539-6975.
- Albert S. KYLE : Continuous Auctions and Insider Trading. *Econometrica*, 53(6):1315–1335, 1985. ISSN 0012-9682.
- Bart LARIVIÈRE et Dirk Van den POEL : Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2):472–484, août 2005. ISSN 0957-4174.
- Hervé LE BRAS : *Naissance de la mortalité. L'origine politique de la statistique et de la démographie*. Seuil, octobre 2013. ISBN 978-2-02-101512-6. Google-Books-ID : kU2tAQAAQBAJ.
- Ronald D. LEE et Lawrence R. CARTER : Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–671, septembre 1992. ISSN 0162-1459.
- Georg V. LEHECKA : The Value of USDA Crop Progress and Condition Information : Reactions of Corn and Soybean Futures Markets. *Journal of Agricultural and Resource Economics*, 39(1):88–105, 2014. ISSN 1068-5502.
- Aurelie LEMMENS et Christophe CROUX : Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, 43(2):276–286, 2006. ISSN 0022-2437.
- Aurelie LEMMENS et Sunil GUPTA : Managing Churn to Maximize Profits. SSRN Scholarly Paper ID 2964906, Social Science Research Network, Rochester, NY, mai 2017.

- Susanna LEVANTESI et Virginia PIZZORUSSO : Application of Machine Learning to Mortality Modeling and Forecasting. *Risks*, 7(1):26, mars 2019.
- Ainong LI, Shunlin LIANG, Angsheng WANG et Jun QIN : Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. *Photogrammetric Engineering & Remote Sensing*, 73(10):1149–1157, octobre 2007.
- Han LI, Colin O’HARE et Farshid VAHID : Two-Dimensional Kernel Smoothing of Mortality Surface : An Evaluation of Cohort Strength. *Journal of Forecasting*, 35(6):553–563, 2016. ISSN 1099-131X.
- Hong LI et Yang LU : Coherent Forecasting of Mortality Rates : A Sparse Vector-Autoregression Approach. *ASTIN Bulletin : The Journal of the IAA*, 47(2):563–600, mai 2017. ISSN 0515-0361, 1783-1350.
- Jing LI et Alexander SZIMAYER : The effect of policyholders’ rationality on unit-linked life insurance contracts with surrender guarantees. *Quantitative Finance*, 14(2):327–342, février 2014. ISSN 1469-7688.
- Johnny Siu-Hang LI, Wai-Sum CHAN et Rui ZHOU : Semicohherent Multipopulation Mortality Modeling : The Impact on Longevity Risk Securitization. *Journal of Risk and Insurance*, 84(3):1025–1065, 2017. ISSN 1539-6975.
- Johnny Siu-Hang LI et Mary R. HARDY : Measuring Basis Risk in Longevity Hedges. *North American Actuarial Journal*, 15(2):177–200, avril 2011. ISSN 1092-0277.
- Johnny Siu-Hang LI, Rui ZHOU et Mary HARDY : A step-by-step guide to building two-population stochastic mortality models. *Insurance : Mathematics and Economics*, 63:121–134, juillet 2015. ISSN 0167-6687.
- Nan LI et Ronald LEE : Coherent mortality forecasts for a group of populations : An extension of the lee-carter method. *Demography*, 42(3):575–594, août 2005. ISSN 1533-7790.
- Ruey-Hsia LI et Geneva G. BELFORD : Instability of Decision Tree Classification Algorithms. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’02, pages 570–575, New York, NY, USA, 2002. ACM. ISBN 978-1-58113-567-1. event-place : Edmonton, Alberta, Canada.
- Stéphane LOISEL, Pierrick PIETTE et Jason TSAI : Applying economic measures to lapse risk management with machine learning approaches. *arXiv :1906.05087 [stat]*, juin 2019. arXiv : 1906.05087.
- Olivier LOPEZ, Xavier MILHAUD et Pierre-E. THÉRON : Tree-based censored regression with applications in insurance. *Electron. J. Statist.*, 10(2):2685–2716, 2016. ISSN 1935-7524.

- B. L. MA, Lianne M. DWYER, Carlos COSTA, Elroy R. COBER et Malcolm J. MORRISON : Early Prediction of Soybean Yield from Canopy Reflectance Measurements. *Agronomy Journal*, 93(6):1227–1234, novembre 2001. ISSN 1435-0645.
- Anne MACKAY, Maciej AUGUSTYNIAK, Carole BERNARD et Mary R. HARDY : Risk Management of Policyholder Behavior in Equity-Linked Life Insurance. *Journal of Risk and Insurance*, 84(2):661–690, 2017. ISSN 1539-6975.
- Andrew M. MCKENZIE : Pre-Harvest Price Expectations for Corn : The Information Content of USDA Reports and New Crop Futures. *American Journal of Agricultural Economics*, 90(2):351–366, mai 2008. ISSN 0002-9092.
- Andrew M. MCKENZIE et Jessica L. DARBY : Information Content of USDA Rice Reports and Price Reactions of Rice Futures. *Agribusiness*, 33(4):552–568, 2017. ISSN 1520-6297.
- Kevin McNEW, P. et Juan Andres ESPINOSA : The Informational Content of USDA Crop Reports : Impacts on Uncertainty and Expectations in Grain Futures Markets. *The Journal of Futures Markets*, 14(4):475–492, 1994.
- Abigail McWILLIAMS et Donald SIEGEL : Event Studies In Management Research : Theoretical And Empirical Issues. *AMJ*, 40(3):626–657, juin 1997. ISSN 0001-4273.
- Robert C. MERTON : Theory of Rational Option Pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183, 1973. ISSN 0005-8556.
- Avery MICHAELSON et Jeff MULHOLLAND : Strategy for Increasing the Global Capacity for Longevity Risk Transfer : Developing Transactions That Attract Capital Markets Investors. *The Journal of Alternative Investments*, 17(1):18–27, juin 2014. ISSN 1520-3255, 2168-8435.
- Trent T. MILACEK et B. Wade BRORSEN : Trading Based on Knowing the WASDE Report in Advance. *Journal of Agricultural and Applied Economics*, 49(3):400–415, août 2017. ISSN 1074-0708, 2056-7405.
- Xavier MILHAUD : Exogenous and Endogenous Risk Factors Management to Predict Surrender Behaviours. *ASTIN Bulletin : The Journal of the IAA*, 43(3):373–398, septembre 2013. ISSN 0515-0361, 1783-1350.
- Xavier MILHAUD, Stéphane LOISEL et Véronique MAUME-DESCHAMPS : Surrender triggers in life insurance : what main features affect the surrender behavior in a classical economic context ? *Bulletin Français d'Actuariat*, 11(22):5–48, 2011.
- Nikolaos T. MILONAS : The effects of USDA crop announcements on commodity prices. *Journal of Futures Markets*, 7(5):571–589, 1987. ISSN 1096-9934.

- Mario J. MIRANDA et Katie FARRIN : Index Insurance for Developing Countries. *Applied Economic Perspectives and Policy*, 34(3):391–427, septembre 2012. ISSN 2040-5790.
- M. S. MKHABELA, P. BULLOCK, S. RAJ, S. WANG et Y. YANG : Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, 151(3):385–393, mars 2011. ISSN 0168-1923.
- Thorsten MOENIG et Daniel BAUER : Revisiting the Risk-Neutral Approach to Optimal Policyholder Behavior : A Study of Withdrawal Guarantees in Variable Annuities. *Review of Finance*, 20(2):759–794, mars 2016. ISSN 1572-3097.
- UM Rao MOGILI et B B V L DEEPAK : Review on Application of Drone Systems in Precision Agriculture. *Procedia Computer Science*, 133:502–509, janvier 2018. ISSN 1877-0509.
- Claire MOUMINOUX, Jean-Louis RULLIÈRE et Stéphane LOISEL : Obfuscation and Honesty Experimental Evidence on Insurance Demand with Multiple Distribution Channels. *Working paper*, juin 2018.
- M. C. MOZER, R. WOLNIEWICZ, D. B. GRIMES, E. JOHNSON et H. KAUSHANSKY : Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3):690–696, mai 2000. ISSN 1045-9227.
- Tony MULLEN et Nigel COLLIER : Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 412–418, Barcelona, Spain, juillet 2004. Association for Computational Linguistics.
- Shyam V NATH et Ravi S. BEHARA : Customer Churn Analysis in the Wireless Industry : A Data Mining Approach. *In Proceedings - annual meeting of the decision science institute*, volume 561, pages 505–510, 2003.
- J. A. NELDER et R. W. M. WEDDERBURN : Generalized Linear Models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3):370–384, 1972. ISSN 2397-2327.
- Scott A. NESLIN, Sunil GUPTA, Wagner KAMAKURA, Junxiang LU et Charlotte H. MASON : Defection Detection : Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 43(2):204–211, mai 2006. ISSN 0022-2437.
- Thien Hai NGUYEN, Kiyooki SHIRAI et Julien VELCIN : Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, décembre 2015. ISSN 0957-4174.
- Andrea NIGRI, Susanna LEVANTESI, Mario MARINO, Salvatore SCOGNAMIGLIO et Francesca PERLA : A Deep Learning Integrated Lee–Carter Model. *Risks*, 7(1):33, mars 2019.

- W. S. NOBLE : Support vector machine applications in computational biology. *In Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004. ISBN 978-0-262-19509-6.
- William S. NOBLE : What is a support vector machine ? *Nature Biotechnology*, 24(12):1565–1567, décembre 2006. ISSN 1546-1696.
- Helge A. NORDAHL : Valuation of life insurance surrender and exchange options. *Insurance : Mathematics and Economics*, 42(3):909–919, juin 2008. ISSN 0167-6687.
- J. François OUTREVILLE : Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance : Mathematics and Economics*, 9(4):249–255, décembre 1990. ISSN 0167-6687.
- Md Shahriar PERVEZ et Jesslyn F. BROWN : Mapping Irrigated Lands at 250-m Scale by Merging MODIS Data and National Agricultural Statistics. *Remote Sensing*, 2(10):2388–2412, octobre 2010.
- James PESANDO, E. : The Interest Sensitivity of the Flow of Funds Through Life Insurance Companies : An Econometric Analysis. *The Journal of finance*, 29(4):11105–1121, 1974.
- Dana PETERSON, Jerry WHISTLER, Stephen EGBERT et J Christopher BROWN : Mapping Irrigated Lands by Crop Type in Kansas. *Pecora 18-Forty Years of Earth Observation... Understanding a Changing World*, pages 14–17, 2011.
- Pierrick PIETTE : Can Satellite Data Forecast Valuable Information from USDA Reports ? Evidences on Corn Yield Estimates. *In Proceedings of NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, Minneapolis, Minnesota, avril 2019.
- Jean PINQUET, Montserrat GUILLÉN et Mercedes AYUSO : Commitment and Lapse Behavior in Long-Term Insurance : A Case Study. *Journal of Risk and Insurance*, 78(4):983–1002, 2011. ISSN 1539-6975.
- David Martin POWERS : Evaluation : from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *International Journal of Machine Learning Technology*, 2(1):37–63, 2011. ISSN 2229-3981.
- Anup K. PRASAD, Lim CHAI, Ramesh P. SINGH et Menas KAFATOS : Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33, janvier 2006. ISSN 0303-2434.
- Vikram PURI, Anand NAYYAR et Linesh RAJA : Agriculture drones : A modern breakthrough in precision agriculture. *Journal of Statistics and Management Systems*, 20(4):507–518, juillet 2017. ISSN 0972-0510.

- R CORE TEAM : *R : A Language and Environment for Statistical Computing*. Vienna, Austria, 2019.
- Jianqiang REN, Zhongxin CHEN, Qingbo ZHOU et Huajun TANG : Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation*, 10(4):403–413, décembre 2008. ISSN 0303-2434.
- A. E. RENSHAW et S. HABERMAN : Statistical analysis of life assurance lapses. *Journal of the Institute of Actuaries*, 113(3):459–497, décembre 1986. ISSN 2058-1009, 0020-2681.
- A. E. RENSHAW et S. HABERMAN : A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance : Mathematics and Economics*, 38(3):556–570, juin 2006. ISSN 0167-6687.
- Michael ROTHSCHILD et Joseph STIGLITZ : Equilibrium in Competitive Insurance Markets : An Essay on the Economics of Imperfect Information. *The Quarterly Journal of Economics*, 90(4):629–649, 1976. ISSN 0033-5533.
- J. W. ROUSE, R. H. HAAS, J. A. SCHELL et D. W. DEERING : Monitoring vegetation systems in the Great Plains with ERTS. *In Proceedings 3rd Earth Resources Technology Satellite (ERTS) symposium*, volume 1, pages 309–317, Washington, DC, USA, janvier 1974. NASA.
- Yahia SALHI et Stéphane LOISEL : Basis risk modelling : a cointegration-based approach. *Statistics*, 51(1):205–221, janvier 2017. ISSN 0233-1888.
- Yahia SALHI et Pierre-E. THÉRON : Age-Specific Adjustment of Graduated Mortality. *ASTIN Bulletin : The Journal of the IAA*, 48(2):543–569, mai 2018. ISSN 0515-0361, 1783-1350.
- Yahia SALHI, Pierre-E. THÉRON et Julien TOMAS : A credibility approach of the Makeham mortality law. *European Actuarial Journal*, 6(1):61–96, juillet 2016. ISSN 2190-9741.
- Robert E. SCHAPIRE : The strength of weak learnability. *Machine Learning*, 5(2):197–227, juin 1990. ISSN 1573-0565.
- R. D. SCHNEPF : Price Determination in Agricultural Commodity Markets : A Primer. CRS Report for Congress RL33204, 2005.
- B. SCHOLKOPF, KAH-KAY SUNG, C. J. C. BURGESS, F. GIROSI, P. NIYOGI, T. POGGIO et V. VAPNIK : Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, novembre 1997. ISSN 1053-587X.
- Sergii SKAKUN, Belen FRANCH, Eric VERMOTE, Jean-Claude ROGER, Inbal BECKER-RESHEF, Christopher JUSTICE et Nataliia KUSSUL : Early season large-area winter crop

- mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment*, 195:244–258, juin 2017. ISSN 00344257.
- V. SMITH et M. WATTS : Index based agricultural insurance in developing countries : Feasibility, scalability and sustainability. *Gates Foundation*, pages 1–40, 2009.
- Michael SPENCE et Richard ZECKHAUSER : Insurance, Information, and Individual Action. *The American Economic Review*, 61(2):380–387, 1971. ISSN 0002-8282.
- Daniel A. SUMNER et Rolf A. E. MUELLER : Are Harvest Forecasts News? USDA Announcements and Futures Market Reactions. *American Journal of Agricultural Economics*, 71(1):1–8, février 1989. ISSN 0002-9092.
- E. SVOTWA, A. J. MASUKA, B. MAASDORP, A. MURWIRA et M. SHAMUDZARIRA : Use of multi-spectral radiometer and MODIS satellite data correlations in developing models for tobacco yield forecasting. *International Journal of Agricultural Research*, 3(1):147–154, 2014.
- Ali TAMADDONI, Stanislav STAKHOVYCH et Michael EWING : Comparing Churn Prediction Techniques and Assessing Their Performance : A Contingent Perspective. *Journal of Service Research*, 19(2):123–141, mai 2016. ISSN 1094-6705.
- THUCYDIDE : Histoire de la guerre du Péloponnèse. 3(XX), 411 av. J.-C.
- Robert TIBSHIRANI : Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161.
- Laurent TOULEMON et Magali BARBIERI : The mortality impact of the August 2003 heat wave in France : Investigating the ‘harvesting’ effect and other long-term consequences. *Population Studies*, 62(1):39–53, mars 2008. ISSN 0032-4728.
- Chenghsien TSAI, Weiyu KUO et Wei-Kuang CHEN : Early surrender and the distribution of policy reserves. *Insurance : Mathematics and Economics*, 31(3):429–445, décembre 2002. ISSN 0167-6687.
- Chenghsien TSAI, Weiyu KUO et Derek Mi-Hsiu CHIANG : The Distributions of Policy Reserves Considering the Policy-Year Structures of Surrender Rates and Expense Ratios. *Journal of Risk and Insurance*, 76(4):909–931, 2009. ISSN 1539-6975.
- Chih-Fong TSAI et Yu-Hsin LU : Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547–12553, décembre 2009. ISSN 0957-4174.
- Calum G. TURVEY et Megan K. MCLAURIN : Applicability of the Normalized Difference Vegetation Index (NDVI) in Index-Based Crop Insurance Design. *Weather, Climate, and Society*, 4(4):271–284, août 2012. ISSN 1948-8327.

- USDA : *Production and distribution of the principal agricultural products of the world*. Washington, D.C. : U.S. Dept. of Agriculture, Division of Statistics, 1893.
- USDA : Crop Production Historical Track Records. Rapport technique, U.S. Dept. of Agriculture, National Agricultural Statistics Service, Washington, D.C., 2018.
- T. VAFEIADIS, K. I. DIAMANTARAS, G. SARIGIANNIDIS et K. Ch. CHATZISAVVAS : A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, juin 2015. ISSN 1569-190X.
- Julien VEDANI, Nicole EL KAROUÏ, Stéphane LOISEL et Jean-Luc PRIGENT : Market inconsistencies of market-consistent European life insurance economic valuations : pitfalls and practical solutions. *Eur. Actuar. J.*, 7(1):1–28, juillet 2017. ISSN 2190-9741.
- Wouter VERBEKE, Karel DEJAEGER, David MARTENS, Joon HUR et Bart BAESENS : New insights into churn prediction in the telecommunication sector : A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, avril 2012. ISSN 0377-2217.
- Andrés M. VILLEGAS et Steven HABERMAN : On the Modeling and Forecasting of Socio-economic Mortality Differentials : An Application to Deprivation and Mortality in England. *North American Actuarial Journal*, 18(1):168–193, janvier 2014. ISSN 1092-0277.
- Andrés M. VILLEGAS, Steven HABERMAN, Vladimir K. KAISHEV et Pietro MILLOSovich : A Comparative Study of Two-Population Models for the Assessment of Basis Risk in Longevity Hedges. *ASTIN Bulletin : The Journal of the IAA*, 47(3):631–679, septembre 2017a. ISSN 0515-0361, 1783-1350.
- Andrés M. VILLEGAS, Vladimir KAISHEV et Pietro MILLOSovich : StMoMo : An R Package for Stochastic Mortality Modelling. 2017b.
- Anton VRIELING, Michele MERONI, Apurba SHEE, Andrew G. MUDE, Joshua WOODARD, C. A. J. M. (Kees) de BIE et Felix REMBOLD : Historical extension of operational NDVI products for livestock insurance in Kenya. *International Journal of Applied Earth Observation and Geoinformation*, 28:238–251, mai 2014. ISSN 0303-2434.
- Willemijn VROEGE, Tobias DALHAUS et Robert FINGER : Index insurances for grasslands – A review for Europe and North-America. *Agricultural Systems*, 168:101–111, janvier 2019. ISSN 0308-521X.
- Brian D. WARDLOW et Stephen L. EGBERT : A comparison of MODIS 250-m EVI and NDVI data for crop mapping : a case study for southwest Kansas. *International Journal of Remote Sensing*, 31(3):805–830, février 2010. ISSN 0143-1161.

- Kilian Q. WEINBERGER : *Machine Learning for Intelligent Systems*. Cornell university édition, 2018.
- A. K. WHITCRAFT, E. VERMOTE, I. BECKER-RESHEF et C. JUSTICE : Cloud cover throughout the agricultural growing season : Impacts on passive optical earth observations. *Remote Sensing of Environment*, 156:438–447, 2015.
- R. C. WILLETS : The Cohort Effect : Insights and Explanations. *British Actuarial Journal*, 10(4):833–877, octobre 2004. ISSN 2044-0456, 1357-3217.
- Guo-en XIA et Wei-dong JIN : Model of Customer Churn Prediction on Support Vector Machine. *Systems Engineering - Theory & Practice*, 28(1):71–77, janvier 2008. ISSN 1874-8651.
- Bowen YANG, Jackie LI et Uditha BALASOORIYA : Cohort extensions of the Poisson common factor model for modelling both genders jointly. *Scandinavian Actuarial Journal*, 2016 (2):93–112, février 2016. ISSN 0346-1238.
- Jiahui YING, Yu CHEN et Jeffrey H. DORFMAN : Flexible Tests for USDA Report Announcement Effects in Futures Markets. *American Journal of Agricultural Economics*, 101(4):1228–1246, juillet 2019. ISSN 0002-9092.
- Marc YOR : Grossissement d’une filtration et semi-martingales : Theoremes generaux. *In Séminaire de Probabilités XII*, Lecture Notes in Mathematics, pages 61–69. Springer, Berlin, Heidelberg, 1978. ISBN 978-3-540-35856-5.
- Hongwei ZHANG, Huailiang CHEN et Guanhui ZHOU : The Model of Wheat Yield Forecast Based on Modis-Ndvi - A Case Study of Xinxiang. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1-7, 2012.
- Yinsuo ZHANG, Aston CHIPANSHI, Bahram DANESHFAR, Lauren KOITER, Catherine CHAMPAGNE, Andrew DAVIDSON, Gordon REICHERT et Frédéric BÉDARD : Effect of using crop specific masks on earth observation based crop yield forecasting across Canada. *Remote Sensing Applications : Society and Environment*, 13:121–137, janvier 2019. ISSN 2352-9385.
- Yu ZHAO, Bing LI, Xiu LI, Wenhuan LIU et Shouju REN : Customer Churn Prediction Using Improved One-Class Support Vector Machine. *In Xue LI, Shuliang WANG et Zhao Yang DONG*, éditeurs : *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 300–306. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31877-4.
- Baojuan ZHENG, Soe W. MYINT, Prasad S. THENKABAIL et Rimjhim M. AGGARWAL : A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *International Journal of Applied Earth Observation and Geoinformation*, 34:103–112, février 2015. ISSN 0303-2434.

Rui ZHOU, Yujiao WANG, Kai KAUFHOLD, Johnny Siu-Hang LI et Ken Seng TAN : Modeling Period Effects in Multi-Population Mortality Models : Applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167, janvier 2014. ISSN 1092-0277.

Hui ZOU et Trevor HASTIE : Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2):301–320, avril 2005. ISSN 1467-9868.

Chapitre 1

Modèle VAR pénalisé pour la projection des taux d'amélioration de mortalité

Ce chapitre reprend l'article "Forecasting mortality rate improvements with a high-dimensional VAR", co-écrit avec Quentin Guibert et Olivier Lopez.

Abstract

Forecasting mortality rates is a problem which involves the analysis of high-dimensional time series. Most of usual mortality models propose to decompose the mortality rates into several latent factors to reduce this complexity. These approaches, in particular those using cohort factors, have a good fit, but they are less reliable for forecasting purposes. One of the major challenges is to determine the spatial-temporal dependence structure between mortality rates given a relatively moderate sample size. This paper proposes a large vector autoregressive (VAR) model fitted on the differences in the log-mortality rates, ensuring the existence of long-run relationships between mortality rate improvements. Our contribution is threefold. First, sparsity, when fitting the model, is ensured by using high-dimensional variable selection techniques without imposing arbitrary constraints on the dependence structure. The main interest is that the structure of the model is directly driven by the data, in contrast to the main factor-based mortality forecasting models. Hence, this approach is more versatile and would provide good forecasting performance for any considered population. Additionally, our estimation allows a one-step procedure, as we do not need to estimate hyper-parameters. The variance-covariance matrix of residuals is then estimated through a parametric form. Secondly, our approach can be used to detect nonintuitive age dependence in the data, beyond the cohort and the period effects which are implicitly captured by our model. Third, our approach can be extended to model the several populations in long run perspectives, without raising issue in

the estimation process. Finally, in an out-of-sample forecasting study for mortality rates, we obtain rather good performances and more relevant forecasts compared to classical mortality models using the French, US and UK data. We also show that our results enlighten the so-called cohort and period effects for these populations.

Keywords: Mortality forecasting; High-dimensional time series; Vector Autoregression; Elastic-Net.

1.1 Introduction

Identifying patterns in the mortality dynamics of a population is a hard task due to the complex underlying phenomena that impact the death rates. This problem is of crucial interest for government policies, pension funds and insurance companies. A wide range of models has been developed since the introduction of the famous model proposed by [Lee and Carter \(1992\)](#). Most of these approaches rely on time-series modeling with past data and forecast the main factors influencing the force of mortality, see among other [Booth et al. \(2002\)](#); [Brouhns et al. \(2002\)](#); [Cairns et al. \(2006, 2009\)](#); [Renshaw and Haberman \(2008\)](#); [Plat \(2009\)](#); [Hunt and Blake \(2014\)](#). Some reviews are available in the literature, see e.g. [Booth and Tickle \(2008\)](#); [Cairns et al. \(2008\)](#); [Barrieu et al. \(2012\)](#). The purpose of the present paper is to provide a new flexible modeling for the evolution of the mortality. Our high-dimensional vector autoregressive (VAR) approach, combined with an elastic-net penalty estimation method, aims to capture complex demographic effects without imposing a too restricting shape for the dynamics.

In recent years, some advances have been made to improve the forecast of mortality rates compared to the traditional factor-based models inspired by [Lee and Carter \(1992\)](#). This innovation has been provoked in particular by practitioners need for managing longevity risk and responding to the Solvency II requirements in insurance. Indeed, traditional models, even when a cohort effect is considered, have a reasonable fit, but poorer forecasts, indicating that these models may overfit the data. In such a context, one of the major concerns is to avoid the divergence of mortality rates between adjacent ages and different countries. Such inconsistency in the forecasting is pointed out for example by [Börger et al. \(2014\)](#), who explain that these low performances are due to the fact that traditional models mainly focus on the central trajectory projection. Several directions have been explored to overcome this issue. [Li et al. \(2013\)](#) develop an approach letting the age coefficients rotate over time, based on an expert judgment. [Hunt and Villegas \(2015\)](#) add an additional constraint on the cohort effect extensions of the [Renshaw and Haberman \(2006\)](#) model to overcome the convergence and robustness issues induced by the two-stage fitting algorithm for parameters. Regarding mortality trends of multiple populations, a relatively wide literature is organized around the idea of a biological convergence at a long horizon, see [Dowd et al. \(2011\)](#); [Järner and Kryger \(2011\)](#); [Li and Lee \(2005\)](#); [Enchev et al. \(2016\)](#); [Cairns et al. \(2016b\)](#) among others. These

approaches estimate the mortality model by bringing together the data of several countries. Recently, [Bohk-Ewald and Rau \(2017\)](#) propose to approach the turning points of the mortality problem by combining trends of several countries.

Based on the observation that mortality rates are, in fact, noisy data, other alternative methods have emerged. If no exceptional event occurs, one can assume that the mortality surface is rather smooth over the age and time dimensions. Thus, functional data analysis and nonparametric smoothing techniques have been applied to mortality modeling, leading to a particular family of mortality models (see e.g. [Currie et al., 2004](#); [Hyndman and Ullah, 2007](#); [Li et al., 2016](#); [Dokumentov et al., 2018](#)). These models are known to have good fitting and forecasting performances, however they mostly consider future values as missing ones, making the stochastic generation of multiple prospective mortality scenarios non intuitive.

Other approaches focusing on the age-period dependency have recently been proposed with the constraint of being more data-driven. [Christiansen et al. \(2015\)](#) use spatial statistics to forecast age-period mortality rate improvements using a kriging method. Their approach is parsimonious and provides good performances for short-term projection. However, it seems that their long-term results are more questionable. [Doukhan et al. \(2017\)](#) also focus on the surface of mortality improvements and model it parsimoniously with an AR-ARCH specification for a random field memory model. A valuable feature of their approach is that both dependencies between cohorts and the conditional heteroscedasticity of mortality are taken into account. Although they have good forecasting results, it is difficult to justify the size of the neighborhoods used to specify the memory process. [Li and Lu \(2017\)](#) choose a VAR process to consider the spatial dependence of mortality rates between neighboring ages adapted to short-term and long-term perspectives. These authors account for sparsity and stationarity in their VAR model by constraining the shape of the Granger causality matrix ([Granger, 1969](#)) as a lower triangular. Their model is also able to consider multiple populations.

In this paper, we propose an alternative approach that newly forecasts the age-period dependency using a large VAR specification on the log-mortality improvements. Although a VAR model is suitable for mortality time-series and is able to capture both long-term relationships and short-term shocks (see e.g. [Salhi and Loisel, 2017](#)), it is difficult to estimate accurately such models using an ordinary least square (OLS) technique, as these series are highly correlated and histories of data are relatively short. To avoid overparameterization, existing forecasting approaches impose an *a priori* spatio-temporal dependency structure between mortality rates or mortality improvements, which implies that only some selected series can interact. In contrast, our main contribution is the introduction of an estimation framework allowing for a large and flexible VAR structure without excluding potentially relevant relationships. A great feature of such a VAR specification is that all classical mortality models could naturally be included in our specification, especially the so-called cohort and period effects as noted by [Li and Lu \(2017\)](#).

Following recent developments in economics and finance (Fan et al., 2011; Furman, 2014), we develop a penalized VAR method based on the elastic-net (Zou and Hastie, 2005), which allows to take into account the sparsity correctly. Indeed, such a VAR model has a sparse structure in high dimension, which requires an accurate estimation method for shrinking zero coefficients in the Granger causality matrix. Compared to a classical maximum likelihood estimation approach, the key idea behind the elastic-net is to incorporate a penalty, which constrains the parameters. This penalty is a combination of an \mathcal{L}_2 term (as in a ridge regression) to avoid ill-conditioning matrices, and an \mathcal{L}_1 term (as in a LASSO regression) to produce a sparse model. By sparse model, we mean that our data-driven automatic selection produces a model with a relatively small number of non-zero parameters. As noted by Furman (2014), this is an attractive alternative to Bayesian VAR procedures usually considered in an econometrical framework and developed for example by Hahn (2014) for multiple populations modeling. Indeed, such approaches require to introduce relevant priors and do not address the sparsity’s issue. The residuals are modeled as a Gaussian vector where the variance-covariance matrix is described using a parametric form for parsimony purposes.

Similarly to Doukhan et al. (2017), but contrary to Li and Lu (2017), our approach models the log of mortality improvements rather than the log of mortality rates. Several empirical elements have been advanced in the recent literature showing the interest of mortality improvements. Haberman and Renshaw (2012) show that a dual approach based on improvement rates can be followed for usual mortality models. They generally obtain quite comparable (but often better) forecasting results with this alternative route for the Lee-Carter model and its variants. As also noted by Bohk-Ewald and Rau (2017), mortality improvements seem to be easier to analyze, which facilitates the identification of divergences in mortality. As our approach is highly flexible, we expect that it can better capture complex patterns of mortality improvements. Another argument is that mortality improvements are generally stationary (see e.g. Chai et al., 2013), which is required for projection as our approach, contrary to Li and Lu (2017), does not impose constraints for guarantying stationarity.

We compare our high-dimensional VAR model to five different benchmark mortality forecasting models: the usual Lee-Carter model (Lee and Carter, 1992) and the M7 model developed by Cairns et al. (2009) which are standard factor-based models; a reference model in smoothing methodologies developed by Hyndman and Ullah (2007) and the more recent smoothing RESPECT model introduced by Dokumentov et al. (2018); and finally the STAR model, based like ours on a high-dimensional VAR method, developed by Li and Lu (2017). Using the root mean squared error measure, we show that our approach leads to general better fitting (in-sample) and forecasting (out-of-sample) of the mortality rate time series from the three countries we have focused our analysis on. Moreover, our data-driven model implies more stable errors over different countries while the benchmark models tend to have more variable predictive power depending on the considered population.

The remainder of this paper is organized as follows. In Section 1.2 we describe the VAR model we retained. The high-dimensional estimation of this model is then developed in Section 1.3. We present the data we used, different results that we obtained and a comparison to other standard mortality models in Section 1.4. Finally, Section 1.5 proposes an extension of the VAR model to multi-population modeling, and Section 1.6 considers some ways of improvement and concludes.

1.2 A Vector Autoregression approach for mortality rate improvements

In this section, we introduce an econometric model to describe the mortality improvement dynamics jointly. The mortality models we introduce in the literature in Section 1.1 are initially based on an analysis of the main factors explaining a common trend of mortality rates. For instance, many models have been developed in the past for capturing the cohort effect, observed in the residuals for improvement rates plots (Willets, 2004). Conversely, our approach only imposes an autoregressive structure, which encountered these classical models¹, as shown for example by Salhi and Loisel (2017) or Li and Lu (2017). In particular, the latter authors explain in details how the cohort and period effects can directly be captured in the VAR(1) representation, without defining specific factors explicitly.

Throughout this paper, we focus on the time series $y_{i,t} = \ln(m_{i,t})$, where $m_{i,t}$ is the crude annual death rate at age i and at date t . These rates can be easily computed thanks to annual risk exposures and count of deaths for a country of interest. Those series are usually not stationary, as a trend can be observed in mortality rates and life expectancy. Since we want to apply our vector autoregressive model on stationary series, we compute the first difference of the log-mortality rate $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$ or, in other words, the log-mortality improvement rates. By working on these quantities, we remove a linear trend in the $y_{i,t}$ series.

With this notation, we specify the mortality rate improvement process by a stationary vector autoregressive model of temporal lag p or a VAR(p). For a minimum age i_{\min} and a maximum age i_{\max} , we define the d -dimensional vector of log-mortality rate improvements, with $d = i_{\max} - i_{\min} + 1$, as $\Delta \mathbf{Y}_t = (\Delta y_{i_{\min},t}, \Delta y_{i_{\min}+1,t}, \dots, \Delta y_{i_{\max},t})^\top$. Next, we assume the following dependence structure dynamic,

$$\Delta \mathbf{Y}_t = \mathbf{C} + \sum_{k=1}^p \mathbf{A}_k \Delta \mathbf{Y}_{t-k} + \mathbf{E}_t, \quad (1.1)$$

where, for $k = 1, \dots, p$, \mathbf{A}_k are $d \times d$ -autoregressive matrices, \mathbf{C} is a d -dimensional vector of constants (an intercept), and \mathbf{E}_t is a d -dimensional Gaussian white noise with mean 0

1. More precisely, it is their alternatives using mortality improvements, as documented by Haberman and Renshaw (2012).

and Σ the related covariance matrix. We denote by $\epsilon_{i,t}$ its marginals. The matrices \mathbf{A}_k , $k \in \{1, \dots, p\}$, capture the relationship between current mortality improvements and the k^{th} lag of $\Delta \mathbf{Y}_t$. In other words, this corresponds to the Granger causality between different cohorts for the mortality improvement rates. As a result, for a VAR(1) model, the coefficients related to the first subdiagonal of the Granger causality matrix capture a cohort effect for individuals born in the same year. As also noted by [Li and Lu \(2017\)](#), the terms of the diagonal can be interpreted as a period effect, since they measure an effect for a fixed age between periods.

The VAR(p) model allows taking into account a more complex dependence structure than the usual mortality factor models. First, our model enables a larger flexibility in the long-term spatio-temporal dependence structure through the autoregressive matrices than the standard factor models. For a given square (i, t) in the Lexis diagram, we let the possibility for the improvement mortality rates $\Delta y_{i,t}$ to be dependent of all the ages among the d -dimensional space of ages, and through all the p temporal lags. In particular, we notice that this domain includes a cohort effect for these improvement rates. For each factor $\Delta y_{i-1,t-1}, \dots, \Delta y_{i-p,t-p}$, this effect is indeed captured by the loading coefficients positioned on the k^{th} -subdiagonal of the matrix \mathbf{A}_k for each $k \in \{1, \dots, p\}$. Hence, the VAR(p) structure permits for the shocks to propagate through different periods. Compared to the model proposed by [Li and Lu \(2017\)](#), the lag order p can take a value greater than 1, allowing to capture a more complex dependence structure.

Furthermore, it has the ability to enlighten some effects that are not captured in the standard mortality literature, e.g. between neighboring cohorts, as we do not impose any constraint on the matrices \mathbf{A}_k , $k \in \{1, \dots, p\}$. Compared to most of factor models, the second improvement of our model on the dependence flexibility is that it captures the long-term co-movement by the autoregressive matrices and the short-term dependence through the covariance matrix at the same time.

Nevertheless, the major issue of our VAR(p) model is that it is a natural high-dimensional problem. The number of parameters for the Granger causality matrices is pd^2 , without considering the covariance matrix and the constant vector. In mortality modeling studies, it is common to focus on the age range from 0 to 100, that is to say $d = 101$, while the historical data for estimation rarely exceed 70 years. Given this, a VAR(3) implies 30,704 parameters estimated on only 7,070 observations, which makes the ordinary least-squares estimation not feasible. To avoid over-fitting, additional constraints have to be added. In this direction, [Li and Lu \(2017\)](#) impose that some parameters have to be nil for guarantying that the model is sparse and stationary. Conversely, we choose a less arbitrary high-dimensional selection variables technique, developed in the next section, to ensure sparsity.

Similarly, the covariance matrix estimation is also a high-dimensional problem with $\frac{d(d+1)}{2}$ parameters. In order to estimate prediction intervals, we consider an additional specification

for the residuals. Although some high-dimensional techniques do exist for covariance estimation (see e.g. Schäfer and Strimmer, 2005; Opgen-Rhein and Strimmer, 2007; Bickel and Levina, 2008; Bien and Tibshirani, 2011), we rather choose a simple parametric form presented in the following part to reduce the number of parameters.

1.3 High-dimensional estimation of the VAR model

As highlighted in the previous section, the VAR(p) model estimation is a high-dimensional problem, especially with mortality data. The estimation can be decomposed into two parts: first, we estimate the pd^2 -dimensional autoregressive matrices, then the d^2 -dimensional covariance matrix. The dimension reduction in the autoregressive matrices is treated through an elastic penalization in Section 1.3.1. We tackle the problem of the covariance through the choice of a parametric form in Section 1.3.2.

1.3.1 Elastic-net

We now described the extension of the elastic-net regularization and variable selection method, proposed by Zou and Hastie (2005), for the high-dimensional estimation of our autoregressive matrices. This technique can be seen as the combination of the LASSO \mathcal{L}_1 -penalty, introduced by Tibshirani (1996), and the ridge \mathcal{L}_2 -penalty developed by Hoerl and Kennard (1970). Elastic-net has similar properties of variable selection as the LASSO. Moreover, it provides a grouping effect: highly correlated variables tend to be selected or dropped together. LASSO and elastic-net have already been extended to VAR model (Gefang, 2014; Basu and Michailidis, 2015), mostly with an economic application (see e.g. Song and Bickel, 2011; Furman, 2014).

Therefore, we estimate the VAR(p) model presented in Equation (1.1) with T observations of the process $\Delta\mathbf{Y}_t$ for $t = t_{\min}, \dots, t_{\max}$ by minimizing the criterion

$$L(\mathbf{C}, \mathbf{A}_1, \dots, \mathbf{A}_p) = \frac{1}{T-p} \sum_{t=t_{\min}+p}^{t_{\max}} \|\Delta\mathbf{Y}_t - \mathbf{C} - \sum_{k=1}^p \mathbf{A}_k \Delta\mathbf{Y}_{t-k}\|_2^2 - \alpha\lambda \sum_{k=1}^p \|\mathbf{A}_k\|_1 - \frac{(1-\alpha)\lambda}{2} \sum_{k=1}^p \|\mathbf{A}_k\|_2^2, \quad (1.2)$$

where we define for a d -dimensional vector $\mathbf{b} = (b_i)_{1 \leq i \leq d}$

$$\|\mathbf{b}\|_2^2 = \sum_{i=1}^d |b_i|^2,$$

and for a $d \times d$ -dimensional matrix $\mathbf{B} = (b_{i,j})_{1 \leq i \leq d, 1 \leq j \leq d}$

$$\|\mathbf{B}\|_1 = \sum_{i=1}^d \sum_{j=1}^d |b_{i,j}| \text{ and } \|\mathbf{B}\|_2^2 = \sum_{i=1}^d \sum_{j=1}^d |b_{i,j}|^2.$$

The parameter $\alpha \in [0, 1]$ is a hyper-parameter which determines the mix between ridge and LASSO penalties. We use a 10-folds cross-validation method to choose the penalty coefficient λ . It determines the strength of the penalties, for example in the LASSO case, the higher λ gets, the fewer number of variables are selected. The algorithm we used is described in [Friedman et al. \(2010\)](#). In theory, the LASSO \mathcal{L}_1 -penalty part forces most of the coefficients to 0. Nevertheless, in a more practical approach, the algorithm employed does not lead to exact zeroes. Thus, the R-package `glmnet` ([Friedman et al., 2010](#)) applies a threshold on the coefficients. Furthermore, following [Chatterjee and Lahiri \(2011\)](#), the `sparsevar` R-package ([Vazzoler et al., 2016](#)) enables to apply a more tailor-made threshold for time-series estimation which equals to $\frac{1}{\sqrt{pd \ln T}}$, that we retain for our model.

The hyper-parameter α is determined through a grid search. For every value α_h of a pre-defined grid $\{\alpha_1, \dots, \alpha_H\}$, we estimate the parameters of the VAR model $\{\hat{\mathbf{C}}, \hat{\mathbf{A}}_1, \dots, \hat{\mathbf{A}}_p, \hat{\lambda}\}_{\alpha_h}$, as explained just before, and deduce the residuals $\hat{E}_{\alpha_h, t}$ for $t \in \{t_{\min} + p, \dots, t_{\max}\}$.

In the applications, we estimate the tuning parameters by minimizing the prediction error, that we obtain by computing the root-mean-square error

$$\text{RMSE}(\alpha_h) = \sqrt{\frac{1}{d(T-p)} \sum_{t=t_{\min}+p}^{t_{\max}} \|\hat{E}_{\alpha_h, t}\|_2^2}. \quad (1.3)$$

In our application, we considerate the grid $\{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ in order to impose a larger weight to the LASSO penalty for sparsity purposes.

The choice of the lag order p for our VAR elastic-net (VAR-ENET) model differs significantly from the usual lag order selection in the standard VAR models. The parameter p does not fully determine the number of parameters, since the LASSO penalty force the less significant coefficients to zero. By increasing the lag order, some non-null coefficients can be forced to zero in favor of other coefficients in autoregressive matrices of higher lag order. Moreover, if there is no significant coefficient above a certain lag order, all autoregressive matrices above the limit order are largely forced to zero. Thus, we chose a relatively large p to capture eventual high order lag effects, without being worried of over-fitting.

1.3.2 Variance-covariance estimation

The autoregressive matrices are not the only high-dimensional problem of the VAR(p) model, the variance-covariance matrix estimation has $\frac{d(d+1)}{2}$ parameters. This number can quickly get higher than the number of observations while dealing with mortality modeling, and then can cause overfitting, as noted e.g. by [Li and Lu \(2017\)](#).

To overcome this issue, we propose an approach to estimate the covariance matrix with a parametric covariance function, in a manner similar to [Spodarev et al. \(2013\)](#). Firstly, for

each age i , we estimate the standard empirical variance $\hat{\sigma}_i^2$ of the residual, and for each couple of ages $(i, j) \in \{i_{\min}, \dots, i_{\max}\}^2$, we estimate the empirical correlation

$$\hat{r}_{i,j} = \frac{\hat{\sigma}_{i,j}}{\hat{\sigma}_i \hat{\sigma}_j},$$

where $\hat{\sigma}_{i,j}$ is the empirical covariance. Then, guided by the form of the empirical correlation matrices and by the approach of [Christiansen et al. \(2015\)](#), we use a parametric form close to the stable family of covariance functions

$$r_{i,j} = \beta e^{-(a_i + a_j) \times |i-j|} \times \mathbb{1}_{\{i \neq j\}} + \mathbb{1}_{\{i=j\}}, \quad (1.4)$$

with $\beta \geq 0$ and $a_i \geq 0$ for each age i . We fit the model based on the empirical correlation as the OLS solution. Thus, after determining $\hat{\beta}$ and $(\hat{a}_{i_{\min}}, \dots, \hat{a}_{i_{\max}})$, we compute our parametric correlation $\tilde{r}_{i,j}$ given by Equation (1.4). Finally, for each couple of ages (i, j) we estimate the covariance by

$$\tilde{\sigma}_{i,j} = \tilde{r}_{i,j} \times \hat{\sigma}_i \hat{\sigma}_j.$$

1.4 Empirical analysis

In this section, we apply our high-dimensional VAR-ENET model to real data and show its strengths in estimating and forecasting populations. Different populations are considered and we analyze both our in-sample and out-sample results compared to those obtained with retained benchmark models. In the following, the computations are carried out with the R software ([R Core Team, 2019](#)). Our scripts are available upon request.

1.4.1 Data

The dataset that we analyze comes from the [HMD \(2019\)](#). We choose to illustrate our approach with historical mortality data from the England and Wales (UK), the United States (US) and France (FR), as these populations have been largely studied, but have specific features. At first, the overall population is considered, and then both males and females are segregated. We select the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ which was available for these 3 countries when the data was extracted. We begin our analysis by a visual inspection of our data. Figure 1.1 describes the shape of the period log-mortality improvements for populations on a Lexis diagram where the trajectory of one cohort follows a 45 degree line. Different cohort effects can be observed for these countries with pink (resp. green) shades for positive (resp. negative) improvements, indicating a lower (resp. higher) survival.

On the center, UK improvement rates do exhibit some significant diagonal patterns corresponding to the so-called cohort effects. A diagonal stands out, more precisely for individuals aged 45 in 1965. Several vertical patterns corresponding to period effects are also observed, especially for the older ages. Diagonal and vertical structures associated with cohort and

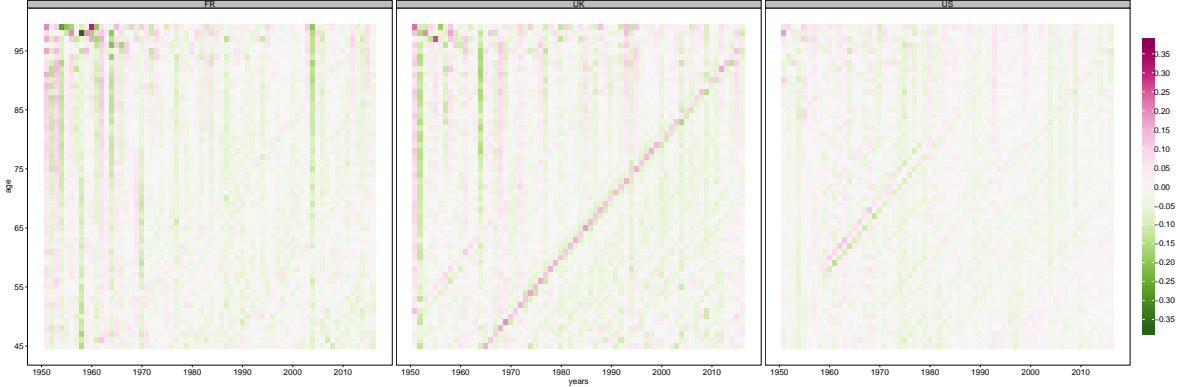


Figure 1.1 – The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for overall populations.

period effects also stand out in the American data, even if the patterns are less marked than in the English data. Contrary to English and American mortality, the French mortality data doesn't clearly display any diagonal structures, but only period effects. We note that the cohort effects (also observable on residual plots), which used to appear in the French data, were strongly reduced with the correction developed by Boumezoued (2016), thanks to fertility rates. Female and male data are displayed in Appendix 1.7.1.

We have chosen to apply the VAR-ENET to the first difference of log-mortality rates series because they are known in the literature as stationary time series. In order to verify this point, we perform a Phillips-Perron test (Perron, 1988) and an augmented Dickey-Fuller test (Said and Dickey, 1984) on every age mortality series for each of the nine populations of interest. All of these time series satisfy the Phillips-Perron test at a confidence level of 1%, and 93% of them are considered as stationary by the augmented Dickey-Fuller test at a level of 5%. These results strengthen our choice to focus on the first difference of time-series.

1.4.2 Benchmark models

In this section, we present the benchmark mortality models that we compare to the VAR-ENET. First, we retain two models from the standard factor-based family:

- the usual Lee-Carter (LC) model (Lee and Carter, 1992), estimated with the approach of Brouhns et al. (2002) and given by

$$y_{i,t} = \alpha_i + \beta_i \kappa_t, \tag{1.5}$$

with the hyper-parameters α_i and β_i , and the mortality trend κ_t ,

- the M7 model developed by Cairns et al. (2009), which considers a quadratic and a

cohort effect, i.e.

$$y_{i,t} = \kappa_t^{(1)} + (i - \bar{i}) \kappa_t^{(2)} + \kappa_t^{(3)} \left((i - \bar{i})^2 - \hat{\sigma}_i^2 \right) + \gamma_{t-i}, \quad (1.6)$$

where $\kappa_t^{(j)}$, $j = 1, 2, 3$, are period effects, γ_{t-i} is a cohort effect, \bar{i} is the average age in the data, and $\hat{\sigma}_i^2$ is the average value of $(i - \bar{i})^2$.

These last models are estimated using the R-package **StMoMo** (Villegas et al., 2017) following their usual two-stage fitting procedure: first, we estimate the factor coefficients of each model, and then we forecast it using univariate ARIMA processes, automatically selected by the R-package using an AIC criterion. To be comparable with our one-stage fitting approach, the results for these models are those obtained after fitting the time-series parameters.

We also consider two recent models based on smoothing methodologies :

- the classical model proposed by Hyndman and Ullah (2007) (HU) which estimates a non-parametric smoothing function $f_t(i)$ for every period t that smooths mortality rates over the age dimension, and is then decomposed

$$f_t(i) = \mu(i) + \sum_{k=1}^K \beta_{t,k} \phi_k(i), \quad (1.7)$$

where $\mu(i)$ is a measure of location of $f_t(i)$, $(\phi_k(i))_{i=1, \dots, K}$ is a set of orthonormal basis functions of dimension $K \geq 1$. This model is applied thanks to the R-package **demography** (Hyndman, 2019) based on weighted penalized regression splines for smoothing.

- The recent RESPECT model, developed by Dokumentov et al. (2018) and implemented in the R-package **smoothAPC** (Dokumentov and Hyndman, 2018), which uses \mathcal{L}_1 -regularized bivariate smoothing over the age and period dimensions, and further allows identification of period and cohort effects on the smoothing residuals.

Finally, we retain a model closer to the methodology of the VAR-ENET, which is based on a spatial-temporal autoregressive framework: the STAR method, introduced by Li and Lu (2017). It models the dynamic of the log mortality rates through a large first-order VAR, of which autoregressive matrix's parameters are forced to a sparse estimation by the following constraints:

$$y_{i_{min}, t+1} = y_{i_{min}, t} + m_{i_{min}}, \quad (1.8)$$

$$y_{i_{min}+1, t+1} = (1 - \alpha_{i_{min}+1}) y_{i_{min}+1, t} + \alpha_{i_{min}+1} y_{i_{min}, t} + m_{i_{min}+1}, \quad (1.9)$$

and

$$y_{i+1, t+1} = (1 - \alpha_{i+1} - \beta_{i+1}) y_{i+1, t} + \alpha_{i+1} y_{i, t} + \beta_{i+1} y_{i-1, t} + m_{i+1}, \quad (1.10)$$

for $i \in \{i_{min} + 2, \dots, i_{max}\}$, m_{i+1} a parameter, and α_{i+1} and β_{i+1} two positive parameters that are smaller than 1. The model is estimated using the benchmark ordinary least square proposed by the authors.

1.4.3 In-sample analysis

In this section, we present the results of our empirical estimations with the VAR-ENET model for each population. We especially focus on the study of the estimated Granger causality matrices that describe the long-term underlying mortality dynamic of the model. The goodness of fit is analyzed by comparing the in-sample results with the benchmark models presented in Section 1.4.2.

Parameters estimation

Let us present our estimated results on the period 1950 – 2016. The parameters are estimated as described in Section 1.3. For the lag order p , we choose the value 7, which represents between 10% and 15% of the observation, depending if we analyze the in-sample or, as in the latter sections, out-sample. Table 1.1 reports the list of the estimated hyper-parameters for each population of interest. We note that, for some populations, we retain the value 1 for α , i.e. we estimate the model with the LASSO constraint only.

Table 1.1 – The estimated VAR-ENET hyper-parameters.

| Country | Population | α | λ |
|---------|------------|----------|-----------|
| FR | Female | 0.8 | 0.0012 |
| FR | Male | 0.6 | 0.0003 |
| FR | Total | 0.8 | 0.0006 |
| US | Female | 0.6 | 0.0010 |
| US | Male | 1.0 | 0.0005 |
| US | Total | 1.0 | 0.0005 |
| UK | Female | 0.6 | 0.0004 |
| UK | Male | 1.0 | 0.0003 |
| UK | Total | 0.8 | 0.0005 |

Note: This table displays the estimated hyper-parameters α and λ in Equation (1.2) for the VAR(7) models. We consider males, females and the overall populations for FR, UK and US.

The first Granger causality matrix \mathbf{A}_1 for each population is displayed in Figure 1.2. These estimated matrices are sparse, i.e. most of the coefficients are estimated to 0 while minimizing the criterion given in Equation (1.2). We identify two main structures by observing the non-zero coefficients. We interpret these patterns in terms of demographic effects, basing our explanations on the underlying mortality dynamic of the model induced by the matrices and described in Equation (1.1).

First, an expected cohort effect, induced by individuals belonging to the same generation, is

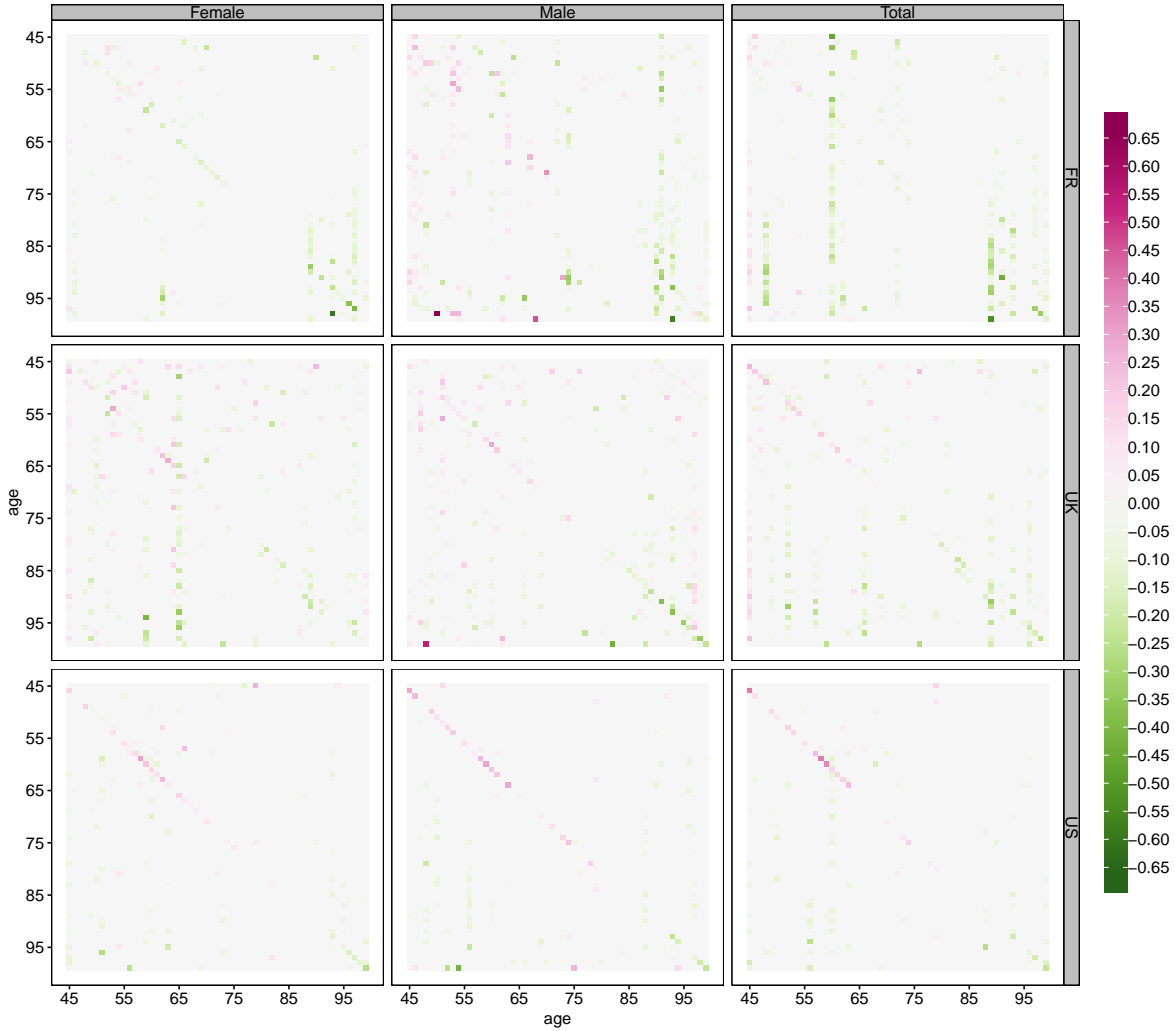


Figure 1.2 – The Granger causality matrix \mathbf{A}_1 for England and Wales (UK), the United States (US) and France (FR) on the age range 45-99 for females, males and the overall population.

highlighted by allocated coefficients on the k^{th} subdiagonal for \mathbf{A}_k for $k \in \{1, \dots, p\}$. Indeed, those coefficients in the VAR model describe the Granger causality of $\Delta y_{i-k,t-k}$ on $\Delta y_{i,t}$. This effect appears positively mainly for the younger ages of our different samples. It is more diffuse between the ages of 65 and 85 years old. In Figure 1.2, the difference between countries appears clearly and the so-called cohort effect is relatively strong for the US population, compared to the FR and the UK. The cohort effect is also clearly visible for $k \in \{2, \dots, 7\}$ (not shown here).

Second, negative period effects are observed on the main diagonals in just about any population, especially between the ages of 85 and 95. Females in France are more impacted by this effect also for younger ages, whereas almost no cohort effect appears for this group. An opposite situation emerges for the US, where the period effect remains limited to the very

older ages.

Third, we notice some age-specific effects corresponding to vertical structure of non-zero coefficients. This third type of patterns reveal non-trivial interactions between different cohorts which are non-necessary within a close neighboring. More concretely, a vertical pattern on the i^{th} column of \mathbf{A}_1 reveals a persistent effect from the term $\Delta y_{i,t-1}$ on $\{\Delta y_{j_1,t}, \dots, \Delta y_{j_l,t}\}$, with $\{j_1, \dots, j_l\} \subset \{i_{\min}, \dots, i_{\max}\}$. It means that the mortality improvement of a single specific age i seems to impact the mortality improvement on a group of ages $\{j_1, \dots, j_l\}$ one year after. Similar structures can be observed on the matrices \mathbf{A}_k for $k \in \{2, \dots, 7\}$ (not shown here).

This latter effect, underlined by our data-driven approach, has not been well documented in the literature yet to our knowledge. These patterns are quite difficult to interpret and to explain within the demographic framework with the available dataset. Disaggregated data would be very useful to explore these effects further. Indeed, these patterns could result from biological, environmental or societal causes, unless it is due to some anomalies in the HMD (Cairns et al., 2016a; Boumezoued, 2016). At this point, we are unable to conclude on the very causes of such age-effects.

On the contrary, the observed cohort and period effects have already been well studied in the literature. However, our model highlights this result in a more data-driven way. Indeed, the existing models either detect these effects in the residuals (e.g. Lee and Carter, 1992; Dokumentov et al., 2018), or force the estimation of specific parameters (e.g. Cairns et al., 2009; Li and Lu, 2017). In our case, we notice these effects by analyzing the parameters estimated without imposing any specific constraints. To exhibit how our data-driven approach can adapt to these effects, we estimate two VAR-ENET(1) models on the French male population over the period 1950-2012, but with the data downloaded from the HMD at two different dates: 2nd October 2017 and 28th January 2019. Indeed, between these two dates, the HMD data had been updated, following the work of Cairns et al. (2016a); Boumezoued (2016) using fertility rates. With this correction, the residual plots display a cohort effect which is substantially lessened. The two Granger causality matrices estimated are displayed in Figure 1.3. In the old version, we clearly remark a subdiagonal in the estimated parameters. In the new estimation, where many false cohort effects had been removed, the pattern on the first subdiagonal is virtually nil, whereas the negative period effect on the main diagonal is only slightly reduced.

In-sample model comparison

We now study how our approach fits well and captures better the mortality pattern on these subsets compared to the benchmark models we retained in Section 1.4.2. Table 1.2 contains the values of the RMSE, as defined in Equation (1.3) for each model and each population. Table 1.3 displays summary statistics of RMSE values over all the populations considered for each model. Benchmark models and the VAR-ENET have quite comparable results. Although

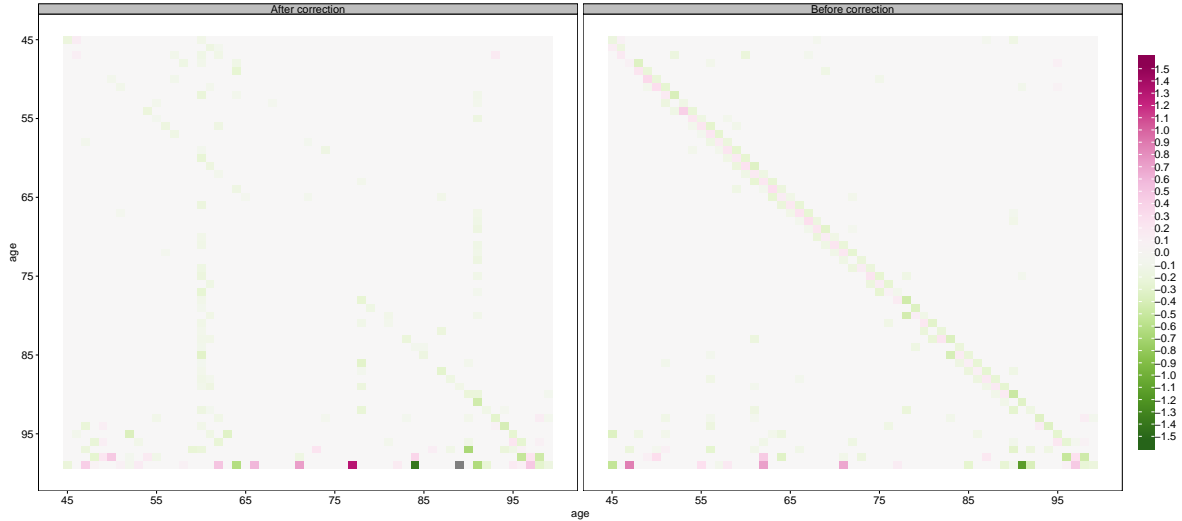


Figure 1.3 – The Granger causality matrix \mathbf{A}_1 for the French males on the age range 45-99 estimated with a VAR-ENET(1) over the period 1950-2012, on the HMD data downloaded on the 2nd October 2017 (Before correction) and the 28th January 2019 (After correction).

the VAR-ENET has not the lowest value for each population, it globally leads to one of the best in-sample results with the HU and RESPECT model (around 2% in average). More particularly, the RESPECT model outperforms the VAR-ENET on the US populations.

The RMSE value is also computed for each age and each year, and the results are displayed respectively on Figures 1.4 and 1.5 for all of the overall populations. First, we note that, on top of having one the lowest RMSE with the smoothing models, the VAR-ENET leads to a more stable error over the age. This is clearly noticeable on the higher ages, especially for the French and English data. For example we observe that the M7's fitting error drastically increases for ages above 95. More generally, on these two populations, all the benchmark model tend to have an increasing age-marginal RMSE starting from 90 years old, while our model's fitting error stays relatively stable. On the American data, the results are more nuanced: whereas we still notice an increase of RMSE values at higher ages for the stochastic benchmark models (LC, M7 and STAR), the smoothing ones and the VAR-ENET lead to stable errors.

The RMSE patterns across periods are more erratic. While the errors of smoothing models (HU and RESPECT) are quite stable over the periods, many peaks are observed for all the stochastic models. More particularly, we note that these peaks tend to occur at the same period for all the concerned models, especially on the French and American data. We remark that, among the stochastic models, the VAR-ENET is the one producing the peaks of the lowest amplitude, and is the closest of the smoothing models in terms of goodness-of-fit. Finally, we notice that, on the American data, the errors of the VAR-ENET and the STAR model are highly correlated, mainly starting from 1985. This shows the methodological closeness of

Table 1.2 – The RMSE of the VAR-ENET and benchmark models.

| Country | Model | RMSE Female | RMSE Male | RMSE Overall |
|---------|---------|-------------|-----------|--------------|
| FR | VAR | 0.032 | 0.013 | 0.021 |
| FR | HU | 0.022 | 0.024 | 0.017 |
| FR | LC | 0.054 | 0.050 | 0.041 |
| FR | M7 | 0.076 | 0.071 | 0.065 |
| FR | RESPECT | 0.021 | 0.025 | 0.022 |
| FR | STAR | 0.042 | 0.045 | 0.038 |
| UK | VAR | 0.014 | 0.015 | 0.018 |
| UK | HU | 0.024 | 0.029 | 0.021 |
| UK | LC | 0.052 | 0.059 | 0.050 |
| UK | M7 | 0.058 | 0.049 | 0.044 |
| UK | RESPECT | 0.022 | 0.027 | 0.016 |
| UK | STAR | 0.040 | 0.044 | 0.035 |
| US | VAR | 0.020 | 0.017 | 0.017 |
| US | HU | 0.017 | 0.016 | 0.014 |
| US | LC | 0.045 | 0.049 | 0.042 |
| US | M7 | 0.048 | 0.050 | 0.046 |
| US | RESPECT | 0.010 | 0.009 | 0.006 |
| US | STAR | 0.025 | 0.024 | 0.023 |

Note: This table reports the RMSE values obtained after fitting the VAR-ENET and the considered benchmark models. We compare this indicator for males, females and the overall populations for FR, UK and US.

Table 1.3 – Summary statistics for the RMSE of the VAR-ENET and the benchmark models.

| Model | Mean | Standard Deviation | Minimum | Maximum |
|---------|-------|--------------------|---------|---------|
| VAR | 0.019 | 0.006 | 0.013 | 0.032 |
| HU | 0.020 | 0.005 | 0.014 | 0.029 |
| LC | 0.049 | 0.006 | 0.041 | 0.059 |
| M7 | 0.056 | 0.012 | 0.044 | 0.076 |
| RESPECT | 0.018 | 0.008 | 0.006 | 0.027 |
| STAR | 0.035 | 0.009 | 0.023 | 0.045 |

Note: This table reports statistics of the RMSE values obtained after fitting the VAR-ENET and the considered benchmark models over the males, females and the overall populations for FR, UK and US.

these two methods. The results for females and males are given in Appendix 1.7.2 with similar findings.

Some of the period peaks may be explained by specific events that have an unexpected impact on the mortality rates, such as an influenza epidemic or a heat wave (Huynen et al., 2001). Indeed, this type of exogenous stresses is difficult to predict with only mortality rate series, which explains why the peaks are observable for the three models. For example, we try to explain the relatively high RMSE in France in 2004. In 2003 a heat wave led to one of the hottest summer ever recorded in France and, as a direct consequence, to higher mortality rates during this year, especially for the elderly. Then, the mortality was much lower in 2004 due to the so-called harvesting effect (Toulemon and Barbieri, 2008; Izraelewicz, 2012). On the contrary, in the calculation of the RMSE, the mortality rates of the year 2004 are forecasted from the observation of the 2003 rates in accordance with the temporal dynamics we have imposed; in this way, the 2004 mortality was expected to be relatively high. This must explain why we observe a RMSE peak in 2004 for the French population.

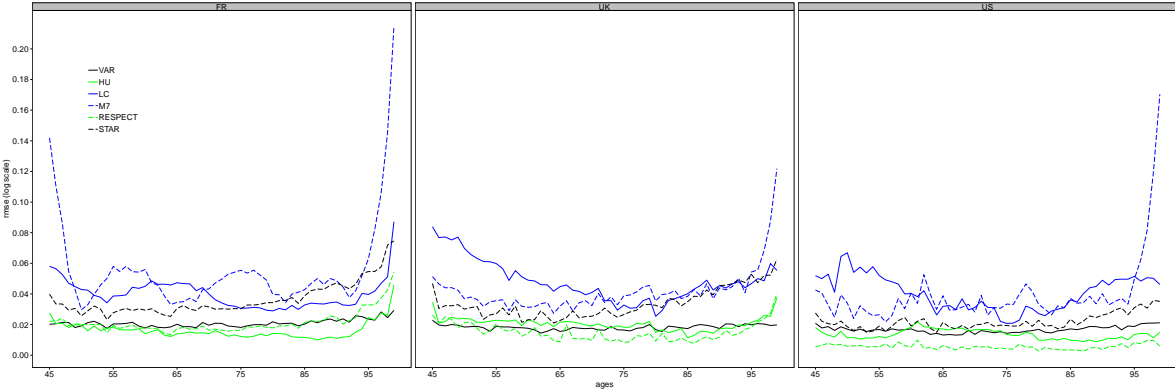


Figure 1.4 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the overall populations.

1.4.4 Out-of-sample performance

For risk management in insurance or more generally for demographers or public policy purposes, mortality rates require being predicted based on the past information. A quite usual test for accuracy is to analyze how the model is able to reproduce the mortality rates correctly. Note that this objective is more demanding than measuring the prediction power on the residual life expectancy. A reasonable model should be able to predict a kind of convergence for mortality at a similar level.

We focus on the prediction power of the VAR-ENET model compared to the benchmark models through an analysis of the out-sample forecasting performance on the same age-period space. To this end, we first estimate each model based on the observations from 1950 to 2000, then we

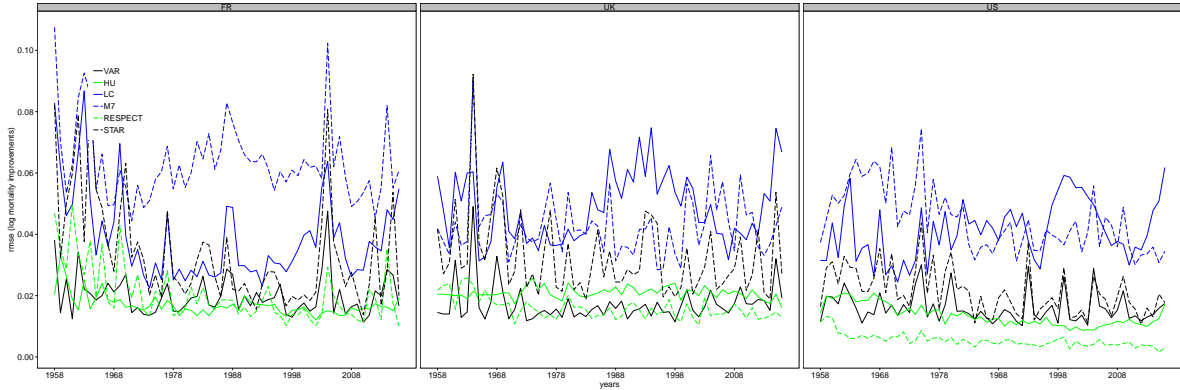


Figure 1.5 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the overall populations.

forecast the mortality rates for the period 2001-2016. Note of course that the mortality rates can be easily calculated using the VAR model, based on the predicted improvement rates and the initial values known for the last year of the training sample. We choose a similar measure to the one taken for the in-sample analysis, the root mean squared forecast error (RMSFE) that we define for a projection horizon h as

$$\text{RMSFE} = \sqrt{\frac{1}{dh} \sum_{i=t_{\min}}^{i_{\max}} \sum_{t=t_0+1}^{t_0+h} (y_{i,t} - \hat{y}_{i,t})^2}, \quad (1.11)$$

where t_0 is the year 2000 and h equals to 16 years in our study.

We compare the predictive power of the different models on the period 2001 – 2016 for the 3 populations (overall, female, male) of the 3 countries of interest. The results are displayed in Table 1.4 and 1.5 (we also display the results for different estimation years but with the same forecasting period in Appendix 1.7.3). We note in Table 1.5 that the average RMSFE is smaller with the VAR-ENET than with the benchmark models, indicating that the former one has higher predictive power in general. However, it is locally outperformed by other models for some populations. Thus, we note for example that the RESPECT model slightly outperforms the VAR-ENET on the French data, and the STAR’s RMSFE value on the US male population (5%) is significantly lower than the one obtained with the VAR-ENET (11.6%) and the other models.

Furthermore, we observe that, in the VAR-ENET applications, all the forecasting errors are of the same order of magnitude, no matter the selected population. This point is highlighted by the standard deviation of RMSFE values over the 9 populations displayed in Table 1.5 for each model. Although the RESPECT and STAR models locally outperform the VAR-ENET, they are more sensible to population considered, leading to a significantly higher standard deviation (respectively 9.4% and 11.3% against only 1.4% for our model). The M7, and to a lesser extent

the LC and HU models, also tend to have more variable forecasting errors, depending on the considered population. These results highlight the stability of the prediction error of the VAR-ENET over different populations compared to the other benchmark models due to the more data-driven approach of the first one, allowing it to better capture the features of each populations' mortality dynamic. By analyzing the results displayed in Appendix 1.7.3, we also notice a better stability of the VAR-ENET over different estimation periods.

Table 1.4 – The RMSFE of the VAR and the benchmark models estimated on the period 1950 – 2000.

| Country | Model | RMSFE Female | RMSFE Male | RMSFE Overall |
|---------|---------|--------------|------------|---------------|
| FR | VAR | 0.088 | 0.110 | 0.078 |
| FR | HU | 0.082 | 0.112 | 0.067 |
| FR | LC | 0.111 | 0.113 | 0.067 |
| FR | M7 | 0.676 | 0.193 | 0.257 |
| FR | RESPECT | 0.083 | 0.091 | 0.071 |
| FR | STAR | 0.098 | 0.127 | 0.417 |
| UK | VAR | 0.095 | 0.087 | 0.080 |
| UK | HU | 0.109 | 0.138 | 0.122 |
| UK | LC | 0.142 | 0.141 | 0.138 |
| UK | M7 | 0.228 | 0.099 | 0.145 |
| UK | RESPECT | 0.281 | 0.230 | 0.296 |
| UK | STAR | 0.083 | 0.115 | 0.156 |
| US | VAR | 0.078 | 0.116 | 0.078 |
| US | HU | 0.061 | 0.141 | 0.085 |
| US | LC | 0.085 | 0.122 | 0.087 |
| US | M7 | 0.237 | 0.144 | 0.135 |
| US | RESPECT | 0.110 | 0.081 | 0.075 |
| US | STAR | 0.071 | 0.050 | 0.049 |

Note: This table reports the out-of-sample performance *via* the RMSFE values for the HU, the LC, the M7, the RESPECT, the STAR and the VAR-ENET models estimated on the period 1950 – 2000. We compare this indicator for males, females and the overall populations for FR, UK and US.

Table 1.5 – Summary statistics for the RMSFE of the VAR-ENET and the benchmark models.

| Model | Mean | Standard Deviation | Minimum | Maximum |
|---------|-------|--------------------|---------|---------|
| VAR | 0.090 | 0.014 | 0.078 | 0.116 |
| HU | 0.102 | 0.030 | 0.061 | 0.141 |
| LC | 0.112 | 0.027 | 0.067 | 0.142 |
| M7 | 0.235 | 0.174 | 0.099 | 0.676 |
| RESPECT | 0.146 | 0.094 | 0.071 | 0.296 |
| STAR | 0.129 | 0.113 | 0.049 | 0.417 |

Note: This table reports statistics of the out-of-sample performance *via* the RMSFE values for the VAR-ENET and the considered benchmark models estimated on the period 1950 – 2000 over the males, females and the overall populations for FR, UK and US.

We plot the RMSFE in Figure 1.6 for the overall population of our three countries of interest. The results for female and male populations are postponed in Appendix 1.7.3. First, we note

that the forecasting errors from most models tend to converge with the projection horizon, suggesting that obtaining a significant enhancement of the forecasting accuracy on the long term seems very challenging. However, we observe that for specific population, some models fail to capture the mortality dynamics, therefore the RMSFE strongly diverges, see e.g. M7 and STAR on the French data or HU on the English data. Nevertheless, our model doesn't suffer from this drawback. Furthermore, it always belongs among the best models for any population and any period considered in this study.

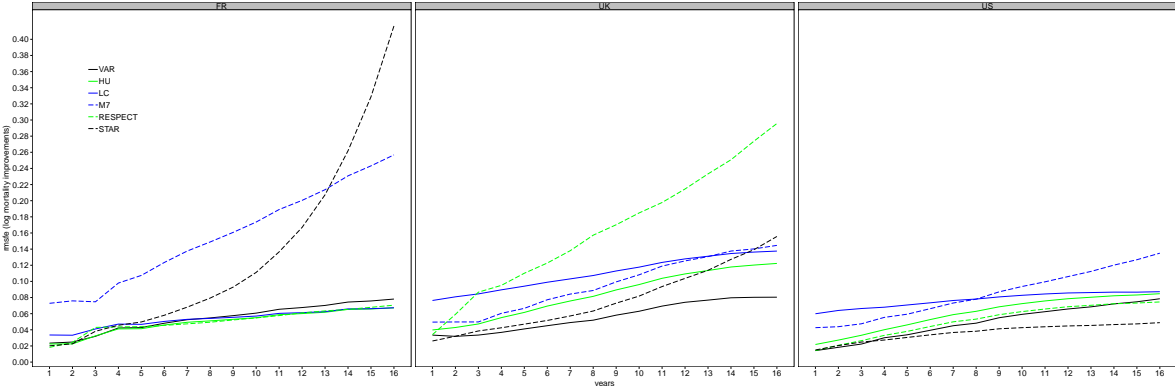


Figure 1.6 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the overall populations.

We now focus on the forecasting error by age groups. We choose to separate the age dimension into 5 classes and compute the RMSFE at a projection horizon of 10 years. By doing so, we compare the predictive power over the different ages. Indeed, depending on the purpose of the mortality forecasting application, one could be more interested in producing accurate predictions for some specific ages. We show the results of the three models in Figure 1.7.

Yet again, we observe that the VAR-ENET is the most stable model over the age classes, when analyzing the forecasting errors at a 15 years projection horizon. While the M7 consistently fails to capture the mortality dynamic at higher ages, the other models have more local issues. For example, we note poorer predictions for the STAR on the French age class 85-94. On the English mortality, the two models LC and HU, and the RESPECT methodology, have respectively higher RMSFE on the age classes 65-74 and 85-94. This point highlights the capacity of our model to uniformly forecast the mortality rates over the age dimension for any of the considered populations.

The results presented in this section suggest that our model slightly outperforms the benchmark methodologies in average for the considered data, even though it doesn't lead to the best accuracy for every population. In all the tested situation, it is at least a credible competitor compared to the best model. More importantly, the VAR-ENET seems to be more stable according to the selected population. This last point is the most noticeable difference between

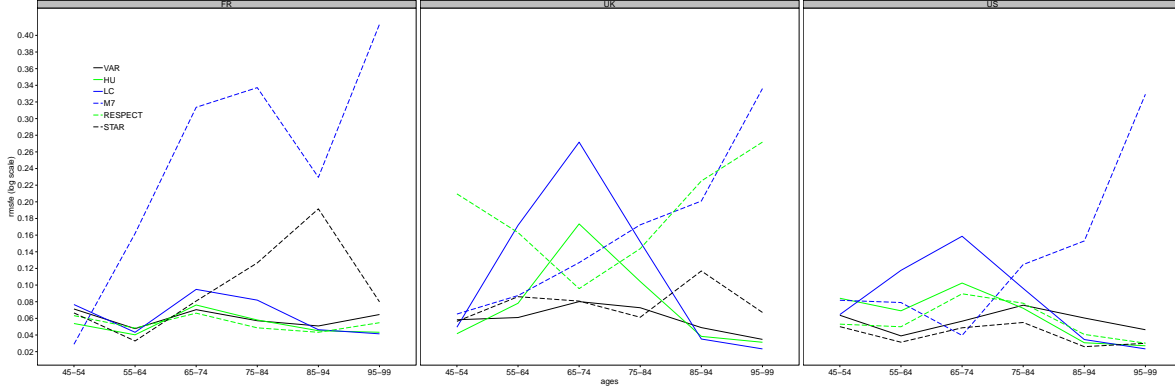


Figure 1.7 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the overall populations.

our in and out-of-sample results. Whereas, in our in-sample study, the models’ fit seems to be relatively equally stable in respect of the considered population, the out-of-sample analysis emphasizes a noticeable heterogeneity in the outcomes depending on the selected dataset. In that regard, the VAR-ENET tends to provide significantly more consistent forecasts.

1.4.5 Forecasting application

Figure 1.8 displays the median forecasts of the log of death rates for ages in $\{45, 65, 85, 95\}$ using the VAR-ENET model from 2017 to 2066. We note that the trends seem rather realistic. We remark that the male and female mortality rates tend to converge rapidly for the UK population. This forecasting result has already been observed in the literature with other models with a similar estimation period (see e.g. [Bohk-Ewald and Rau \(2017\)](#)). We note on the forecasted series that there are some limited shocks during the first projection years, followed by a linear trend. This effect is characteristic of the VAR model and shows how it can propagate innovation shocks among a cohort for example.

Figure 1.9 compares the median forecasts of the log of death rates of two popular models LC, the HU model with the VAR-ENET model from 2013 to 2062. First, we note that for many of the forecasted series, the three models produce very similar projections, especially on the female populations at higher ages. On the contrary, for the British male mortality dynamics, the forecasts are noticeably different. While the LC and HU models predict a stabilization of the mortality rates, and even a slight increase at age 45 for the smoothing methodology, our VAR-ENET forecasts a decrease consistent with the average longevity improvement over the last decades. The two benchmark models seems to be more impacted by the slowdown of this enhancement observed during the very recent years. To a lesser extent, we also notice a comparable results on the French male population.

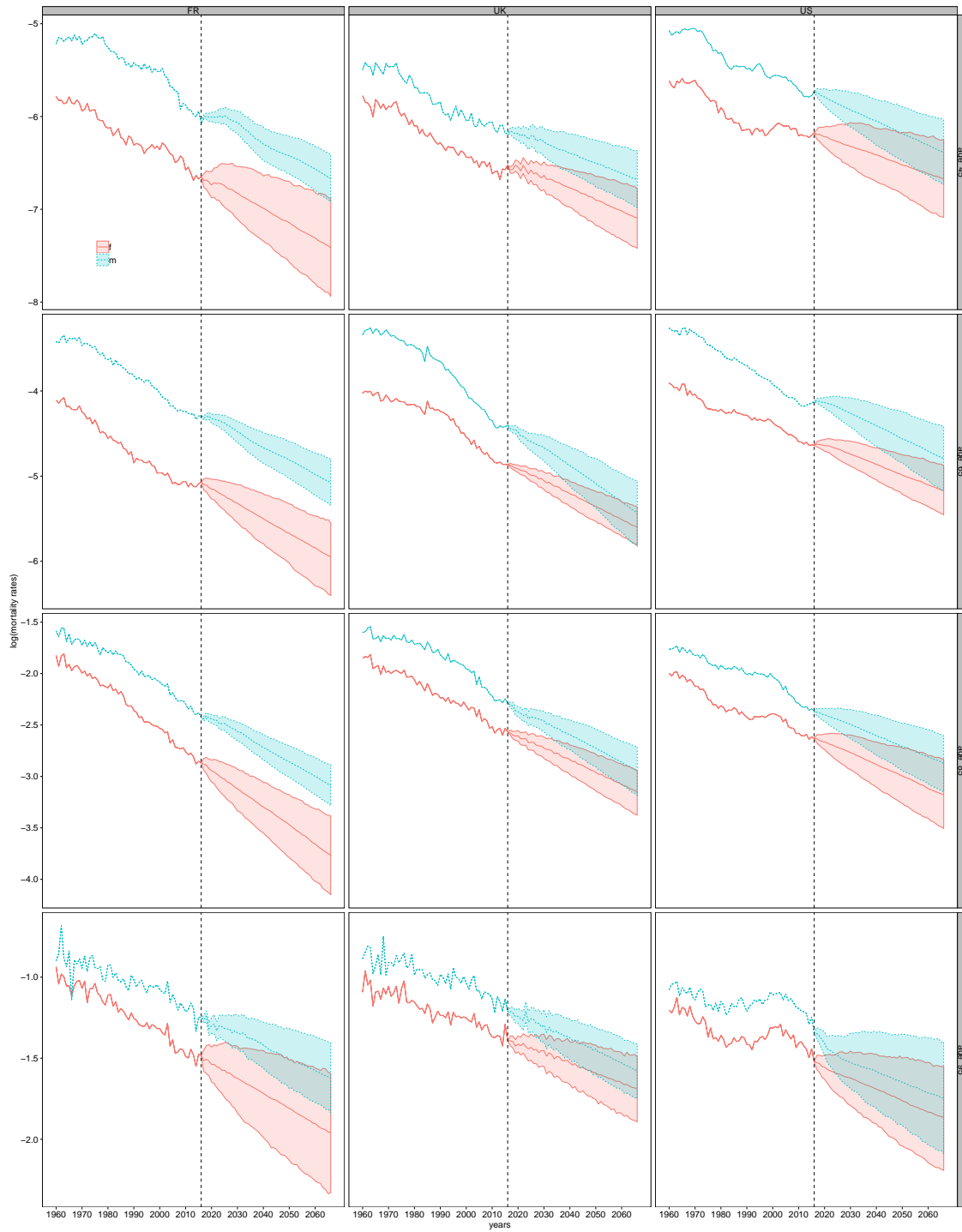


Figure 1.8 – The observed and the projected log of death rates for British (UK), American (US) and French(FR) females and males with the 97.5% prediction intervals, obtained from the VAR-ENET model.

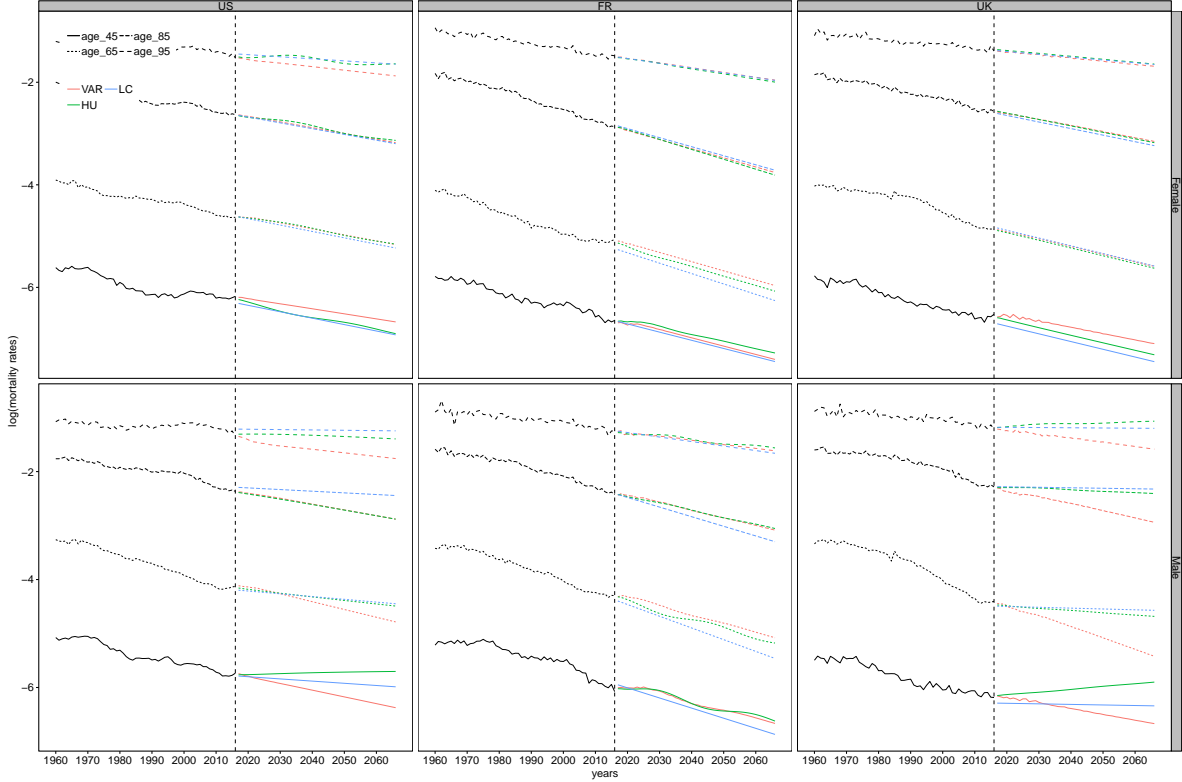


Figure 1.9 – The observed and the projected log of death rates for British (UK), American (US) and French (FR) females and males. This figure compares trends obtained with the HU, the LC and the VAR-ENET models.

1.5 A multi-population extension

Some of the standard mortality forecasting models can be extended to multi-population. However, many of these extensions suffer from limits. One of the recurrent limits in multi-population mortality modeling is the restriction of the extension to only 2 populations. For example, we can note the GRAVITY model of [Dowd et al. \(2011\)](#) or the Bayesian model of [Cairns et al. \(2011\)](#). Another restriction imposed by some existing multi-population models is the necessity to determine a dominant population and sub-populations, see e.g. the SAINT model of [Jarner and Kryger \(2011\)](#), or a common trend for the different populations like in [Li and Lee \(2005\)](#). In this section, we explore the possibility of extending our model for multi-population mortality forecasting and we give the needed details.

We denote M the number of selected populations and $y_{m,i,t}$ the log of mortality rates for the m^{th} population. We suppose that the pair (i_{\min}, i_{\max}) is the same for all the populations to avoid exaggerated notations, although we could have chosen M different pairs of age limits. Thus, we define M different d -dimensional vectors $\Delta \mathbf{Y}_{m,t}$ that we concatenate into a single Md -dimensional vector $\Delta \mathbf{Y}_t = (\Delta y_{1,i_{\min},t}, \dots, \Delta y_{1,i_{\max},t}, \Delta y_{2,i_{\min},t}, \dots, \Delta y_{M,i_{\max},t})^\top$. We then

apply the same model as in Equation (1.1) except that the dimension equals now to Md .

The $(Md) \times (Md)$ -dimensional autoregressive matrices and the Md -dimensional vector of constants are estimated through the same elastic-net methodology as in the single population problem. However, the covariance matrix estimation needs to be extended since its structure may change significantly compared to the single population case.

In the multi-population context, the covariance matrix Σ is $(Md) \times (Md)$ -dimensional. Firstly, we propose to consider this matrix as a block matrix broken into M^2 different $d \times d$ -dimensional submatrices noted $\Sigma_{m,n}$ for each population couple $(m, n) \in \{1, \dots, M\}^2$. Then, for each population m , we estimate the diagonal submatrix $\Sigma_{m,m}$ through the same methodology as in the single population, obtaining in this way $\tilde{\Sigma}_{m,m}$. Since $\tilde{\Sigma}_{m,m}$ is a positive-definite matrix, we define its Cholesky decomposition

$$\tilde{\Sigma}_{m,m} = \tilde{R}_m^\top \tilde{R}_m,$$

where \tilde{R}_m is an upper triangular matrix. Finally, for a couple of populations (m, n) we estimate the covariance submatrix as

$$\tilde{\Sigma}_{m,n} = \rho_{m,n} \times \tilde{R}_m^\top \tilde{R}_n,$$

where $\rho_{m,n} \in [-1, 1]$ is the empirical Pearson's correlation coefficient between the observations of residuals $(\epsilon_{m,i,t})_{(i,t)}$ and $(\epsilon_{n,i,t})_{(i,t)}$ for $(i, t) \in \{i_{\min}, \dots, i_{\max}\} \times \{t_{\min} + p, \dots, t_{\max}\}$.

Thus, our extended covariance model has only $M(2d + 1) + \frac{M(M-1)}{2}$ parameters. The number of ages, d , being generally much larger than the number of populations in mortality modeling, we note that $M(2d + 1) \gg \frac{M(M-1)}{2}$, meaning that we do not add many parameters for covariance estimation while modeling M populations together compared to fitting M single models.

1.6 Conclusion

In this paper, we have proposed a vector-autoregression elastic-net (VAR-ENET) model on the differentiated log-mortality, leading to three key results. First, this new high-dimensional time series analysis outperforms in fitting the mortality rate series of each of the nine populations we considered, compared with the three stochastic benchmark models (LC, M7 and STAR). Moreover, in average, it leads to in-sample errors of same order as the two smoothing benchmark models (HU and RESPECT). Even though our model doesn't produce the most accurate forecasts on each population, it leads to relatively close results compared to the best model each time. In addition, the average RMSFE over the 9 population is lower than the one obtained with any other benchmark models. Furthermore, thanks to its data-driven approach, the VAR-ENET leads to more stable errors than the benchmark models over populations, showing its power of adaptability to the specific mortality dynamics of different populations.

Compared to the usual strategy which requires to compare a variety of possible models and then select the best for a particular age-period population, our approach gives directly and with little effort a serious candidate for a consistent modeling of the mortality, regardless of the population features. The second key result is that, although we let a large freedom in the spatio-temporal dependence structure without imposing a priori constraints, the VAR-ENET model enlightens three main effects: the so-called cohort and period effects and a specific age effect. While the first two models have already been well studied in many papers on mortality modeling, we develop in this paper a new ways for detect the effects for any population. The last effect is less known or possibly even unknown in the literature. Future researches are needed, probably on a finest dataset to understands such a phenomena. Finally, the proposed extension of the VAR-ENET to multi-population mortality modeling seems a priori straightforward, without raising unavoidable issues on the number of populations or on the hierarchy between them, considering the estimation process.

Some points should however be improved and need further researches. The first one concerns the interpretation of the results given by our VAR-ENET model. Although it seems to have a better forecasting and adaptability power than the standard factor-based models, the last ones do benefit from a greater interpretability. Indeed, even if most of the coefficients in the autoregressive matrices are estimated to zero in the VAR-ENET and that the non-null coefficients seems to form specific patterns, the comprehension of the underlying dynamics remains complex. On the contrary, it is much easier to understand the mortality dynamics in terms of period, age and cohort effects, which are directly visible through the use of the classical factor-based models.

Second, we are also aware that some of the hyper-parameter selection techniques we applied can be improved. Firstly, we imposed the lag order p equals to 7 for all the population. In a sensibility analysis, we have noted that according to the considered population, the highest predictive power of the VAR-ENET(p) model is not reached at the same lag order p . These results suggest that an optimization on the hyper-parameter p could be developed. The second hyper-parameter to be improved is the mixing weight parameter α between the LASSO and ridge penalties. In the general context of elastic-net regression, it is usually selected with a grid search. However, [Friedman et al. \(2010\)](#) propose to optimize it through a cross-validation by following the same methodology as the λ selection.

Third, the log-mortality rates series $y_{i,t}$ are known to not be stationary, but also to be cointegrated (see e.g. [Chen et al., 2015](#); [Salhi and Loisel, 2017](#); [Li and Lu, 2017](#)). In our paper, we choose to study the first difference in the log mortality-rates and, by doing so, we lose some information about the long-term co-movement. Another way that we can deal with the non-stationarity and co-integration is to rather select the Vector Error Correcting Model (VECM). Nevertheless, although high-dimensional VAR model has been relatively well studied and recently documented especially in financial econometrics, VECM sparse estimation

with elastic-net or other techniques seems to be a new field (see e.g. [Wilms and Croux, 2016](#)), and could be developed further for the mortality projection. A major improvement of our model would be to implement the elastic-net procedure to VECM estimation and apply it to the log-mortality series.

Finally, even though we introduce an extension to multi-population mortality forecasting of our model, we don't show any empirical studies on that subject in this paper, which rather focuses on the single population case. Many points need to be analyzed in greater detail to correctly assess the behavior of the VAR-ENET model applied to multi-population. It notably includes the examination of a broader list of countries, the specific case of sub-regional populations and the comparison of forecasts to recent multi-population models. Further more specific studies should be conducted to examine the multi-population model.

1.7 Appendix

1.7.1 Improvement rates data for females and males

Figures 1.10 and 1.11 describe the log-mortality improvements for females and males.

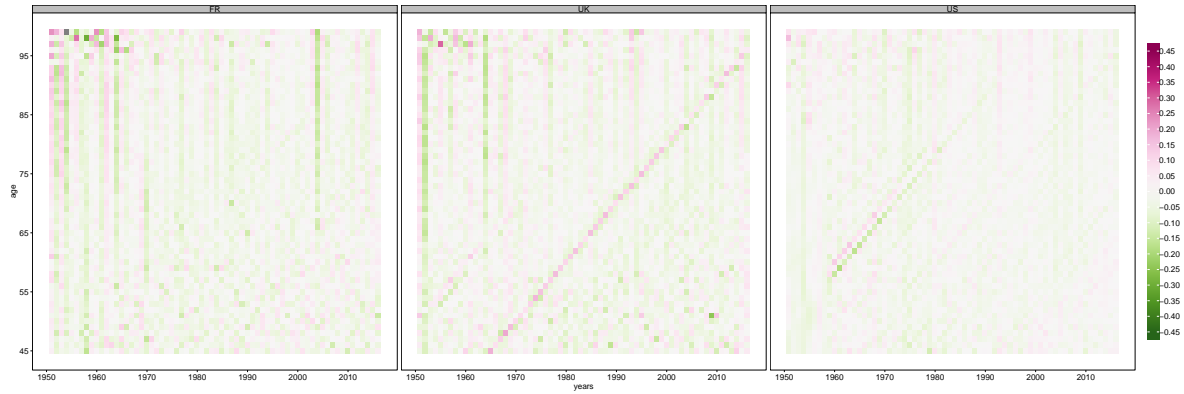


Figure 1.10 – The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for females.

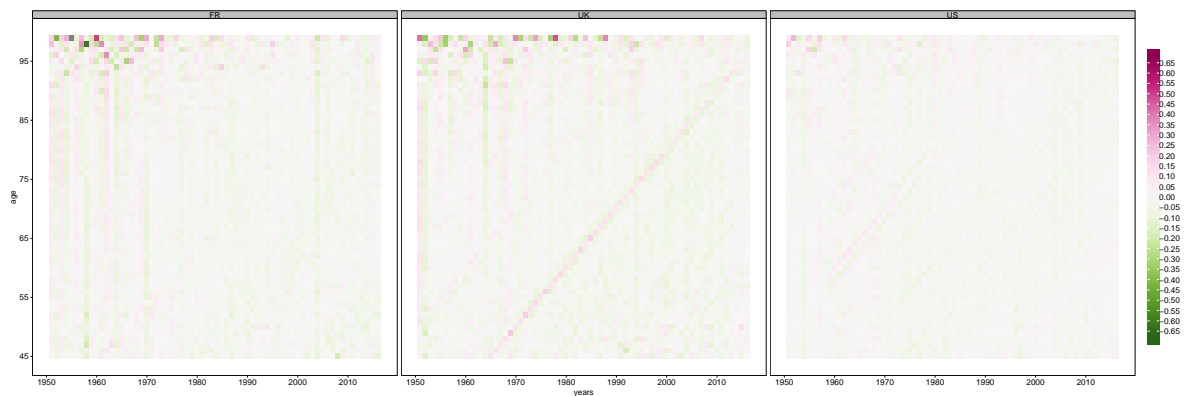


Figure 1.11 – The period log-mortality improvements for England and Wales (UK), the United States (US) and France (FR) on the age-period observation $\{45, \dots, 99\} \times \{1950, \dots, 2016\}$ for males.

1.7.2 In-sample analysis for females and males

Figures 1.12 and 1.13 present the in-sample performance in terms of RMSE for females. Figures 1.14 and 1.15 present the in-sample performance in terms of RMSE for males.

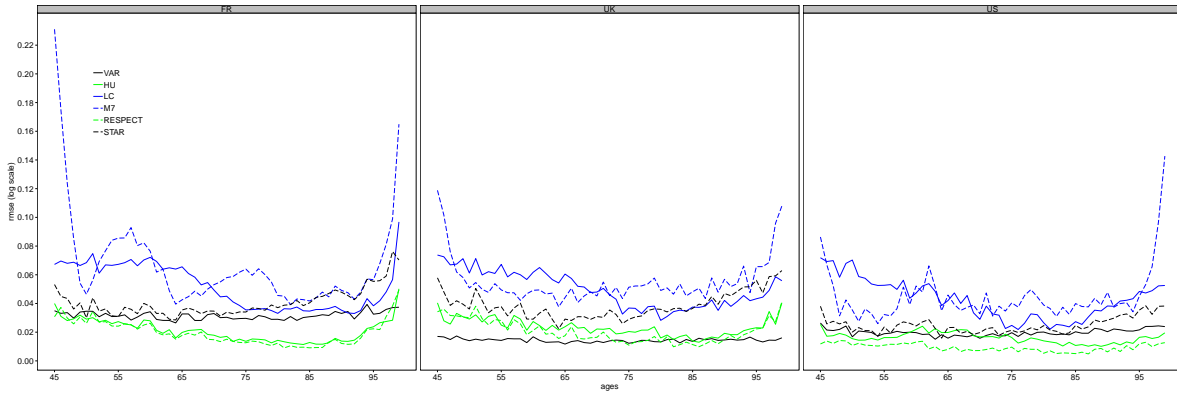


Figure 1.12 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the female populations.

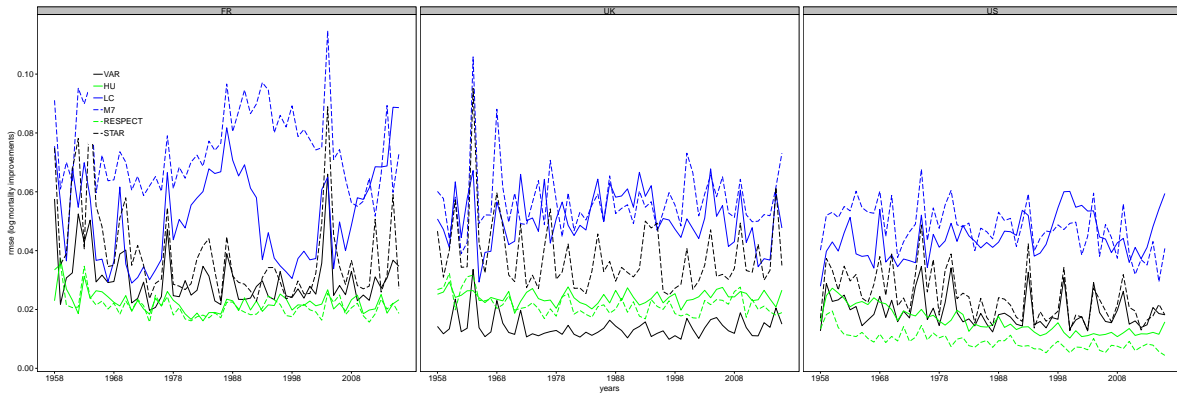


Figure 1.13 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the female populations.

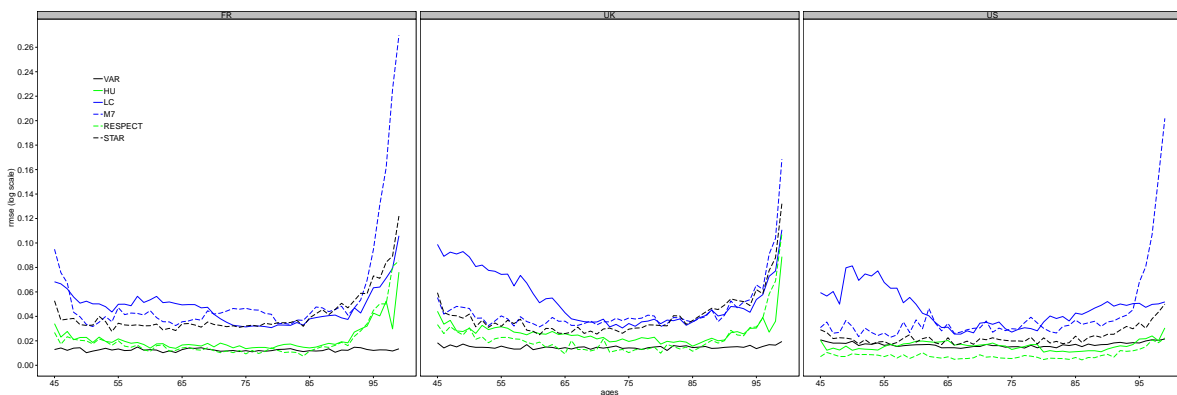


Figure 1.14 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the male populations.

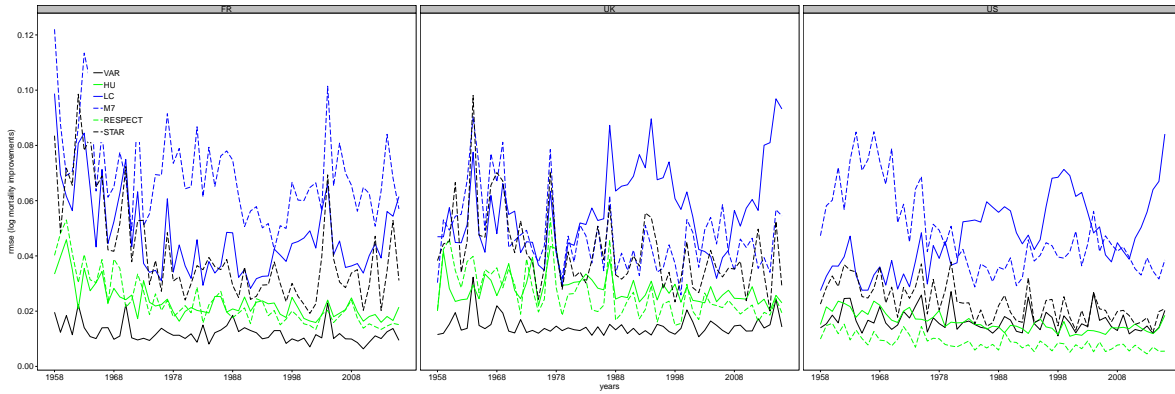


Figure 1.15 – The RMSE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the male populations.

1.7.3 Out-of-sample analysis for females and males

Tables 1.6 and 1.8 present the out-of-sample performance in terms of RMSFE for different model estimation period, respectively $\{1970, \dots, 2000\}$ and $\{1980, \dots, 2000\}$. Tables 1.7 and 1.9 outlines the global statistics of out-sample performance in terms of RMSFE.

Figures 1.16 and 1.17 present the out-of-sample performance in terms of RMSFE for females.

Figures 1.18 and 1.19 present the out-of-sample performance in terms of RMSFE for males.

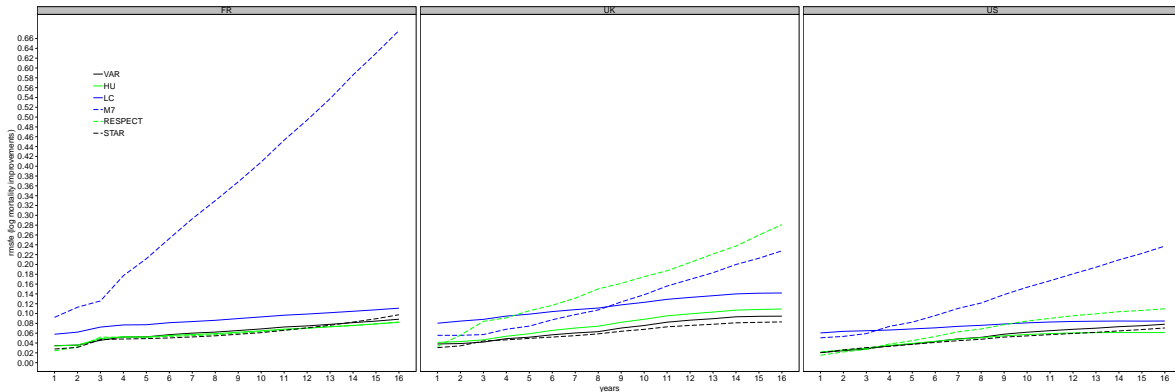


Figure 1.16 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the female populations.

Table 1.6 – The RMSFE of the VAR and the benchmark models estimated on the period 1970 – 2000.

| Country | Model | RMSFE Female | RMSFE Male | RMSFE Overall |
|---------|---------|--------------|------------|---------------|
| FR | VAR | 0.088 | 0.092 | 0.068 |
| FR | HU | 0.081 | 0.085 | 0.058 |
| FR | LC | 0.085 | 0.086 | 0.055 |
| FR | M7 | 0.266 | 0.137 | 0.139 |
| FR | RESPECT | 0.077 | 0.085 | 0.074 |
| FR | STAR | 0.101 | 0.110 | 0.172 |
| UK | VAR | 0.074 | 0.091 | 0.067 |
| UK | HU | 0.116 | 0.134 | 0.126 |
| UK | LC | 0.133 | 0.146 | 0.129 |
| UK | M7 | 0.148 | 0.111 | 0.097 |
| UK | RESPECT | 0.321 | 0.232 | 0.260 |
| UK | STAR | 0.073 | 0.083 | 0.069 |
| US | VAR | 0.105 | 0.116 | 0.092 |
| US | HU | 0.105 | 0.118 | 0.087 |
| US | LC | 0.133 | 0.123 | 0.090 |
| US | M7 | 0.192 | 0.133 | 0.133 |
| US | RESPECT | 0.148 | 0.081 | 0.077 |
| US | STAR | 0.120 | 0.070 | 0.079 |

Note: This table reports the out-of-sample performance *via* the RMSFE values for the HU, the LC, the M7, the RESPECT, the STAR and the VAR-ENET (4) models estimated on the period 1970–2000. We compare this indicator for males, females and the overall populations for FR, UK and US.

Table 1.7 – Summary statistics for the RMSFE of the VAR-ENET and the benchmark models estimated on the period 1970 – 2000.

| Model | Mean | Standard Deviation | Minimum | Maximum |
|---------|-------|--------------------|---------|---------|
| VAR | 0.088 | 0.017 | 0.067 | 0.116 |
| HU | 0.101 | 0.025 | 0.058 | 0.134 |
| LC | 0.109 | 0.031 | 0.055 | 0.146 |
| M7 | 0.151 | 0.051 | 0.097 | 0.266 |
| RESPECT | 0.151 | 0.096 | 0.074 | 0.321 |
| STAR | 0.097 | 0.033 | 0.069 | 0.172 |

Note: This table reports statistics of the out-of-sample performance *via* the RMSFE values for the VAR-ENET and the considered benchmark models estimated on the period 1970 – 2000 over the males, females and the overall populations for FR, UK and US.

Table 1.8 – The RMSFE of the VAR and the benchmark models estimated on the period 1980 – 2000.

| Country | Model | RMSFE Female | RMSFE Male | RMSFE Overall |
|---------|---------|--------------|------------|---------------|
| FR | VAR | 0.092 | 0.092 | 0.077 |
| FR | HU | 0.071 | 0.093 | 0.068 |
| FR | LC | 0.068 | 0.088 | 0.065 |
| FR | M7 | 0.219 | 0.113 | 0.125 |
| FR | RESPECT | 0.091 | 0.086 | 0.072 |
| FR | STAR | 0.295 | 0.081 | 0.764 |
| UK | VAR | 0.088 | 0.095 | 0.079 |
| UK | HU | 0.106 | 0.124 | 0.110 |
| UK | LC | 0.114 | 0.125 | 0.113 |
| UK | M7 | 0.111 | 0.099 | 0.097 |
| UK | RESPECT | 0.286 | 0.269 | 0.197 |
| UK | STAR | 0.236 | 0.065 | 0.288 |
| US | VAR | 0.122 | 0.122 | 0.109 |
| US | HU | 0.106 | 0.121 | 0.096 |
| US | LC | 0.107 | 0.121 | 0.097 |
| US | M7 | 0.161 | 0.129 | 0.141 |
| US | RESPECT | 0.101 | 0.085 | 0.080 |
| US | STAR | 0.942 | 0.088 | 0.322 |

Note: This table reports the out-of-sample performance *via* the RMSFE values for the HU, the LC, the M7, the RESPECT, the STAR and the VAR-ENET (4) models estimated on the period 1980–2000. We compare this indicator for males, females and the overall populations for FR, UK and US.

Table 1.9 – Summary statistics for the RMSFE of the VAR-ENET and the benchmark models estimated on the period 1980 – 2000.

| Model | Mean | Standard Deviation | Minimum | Maximum |
|---------|-------|--------------------|---------|---------|
| VAR | 0.097 | 0.017 | 0.077 | 0.122 |
| HU | 0.099 | 0.020 | 0.068 | 0.124 |
| LC | 0.100 | 0.022 | 0.065 | 0.125 |
| M7 | 0.133 | 0.038 | 0.097 | 0.219 |
| RESPECT | 0.141 | 0.086 | 0.072 | 0.286 |
| STAR | 0.342 | 0.309 | 0.065 | 0.942 |

Note: This table reports statistics of the out-of-sample performance *via* the RMSFE values for the VAR-ENET and the considered benchmark models estimated on the period 1980 – 2000 over the males, females and the overall populations for FR, UK and US.

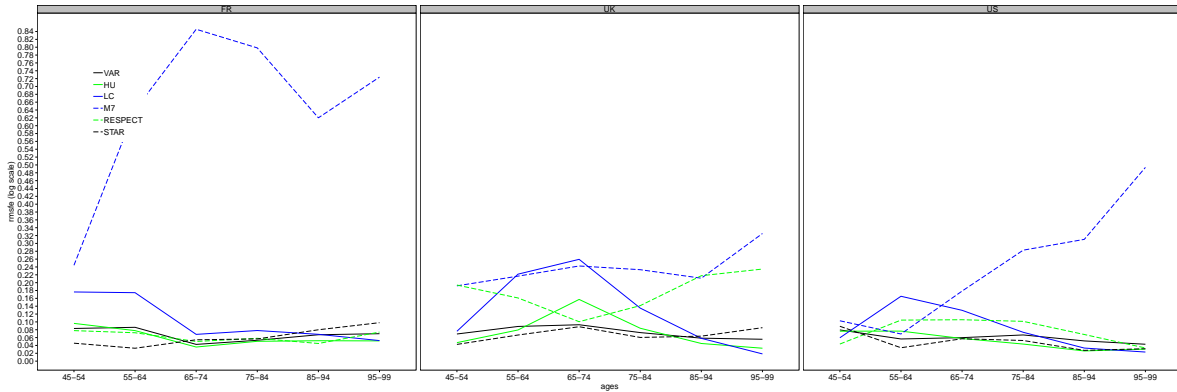


Figure 1.17 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the female populations.

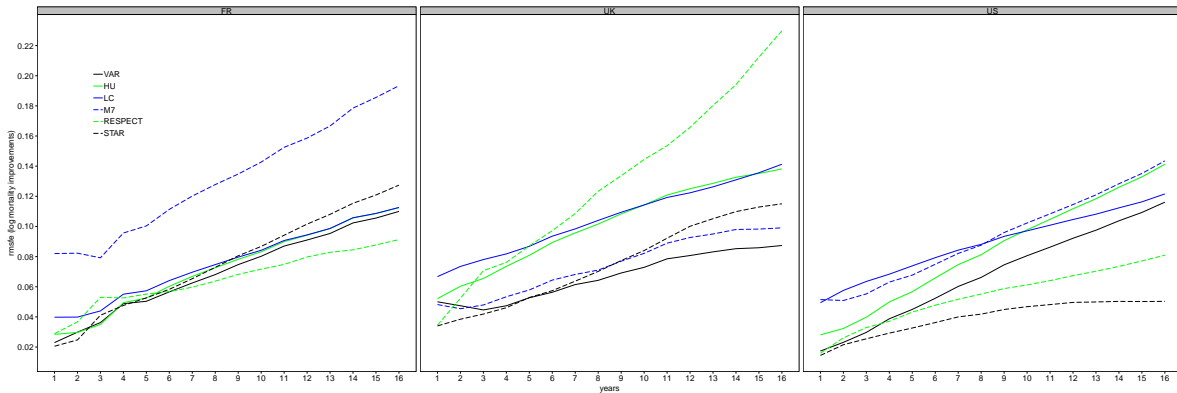


Figure 1.18 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by period for the male populations.

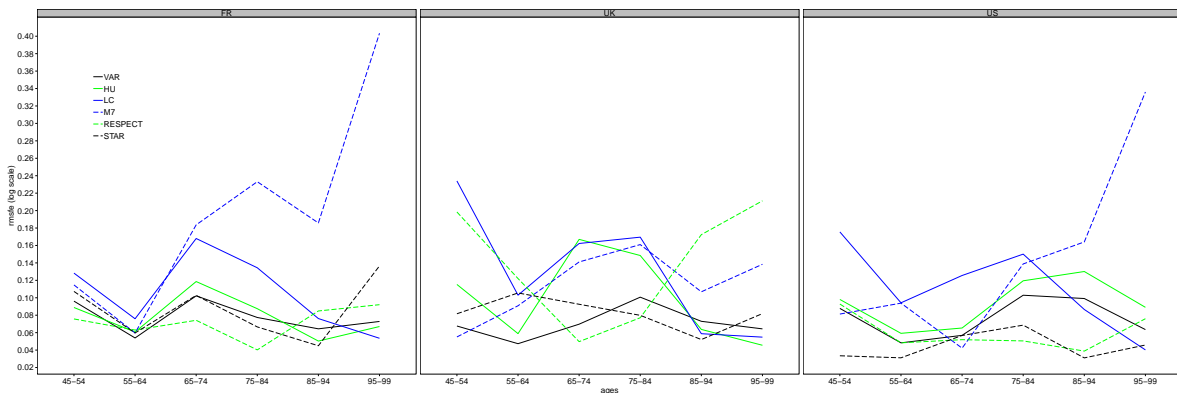


Figure 1.19 – The RMSFE for England and Wales (UK), the United States (US) and France (FR) grouped by age for the male populations.

Bibliography

- Pauline Barrieu, Harry Bensusan, Nicole El Karoui, Caroline Hillairet, Stéphane Loisel, Claudia Ravanelli, and Yahia Salhi. Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian Actuarial Journal*, 2012(3):203–231, 2012. ISSN 0346-1238. doi: 10.1080/03461238.2010.511034.
- Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, August 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1315.
- Peter J. Bickel and Elizaveta Levina. Covariance Regularization by Thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008. ISSN 0090-5364.
- Jacob Bien and Robert J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, December 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr054.
- Christina Bohk-Ewald and Roland Rau. Probabilistic mortality forecasting with varying age-specific survival improvements. *Genus*, 73(1):1, December 2017. ISSN 2035-5556. doi: 10.1186/s41118-016-0017-8.
- H. Booth and L. Tickle. Mortality Modelling and Forecasting: a Review of Methods. *Annals of Actuarial Science*, 3(1-2):3–43, September 2008. ISSN 1748-5002, 1748-4995. doi: 10.1017/S1748499500000440.
- Heather Booth, John Maindonald, and Len Smith. Applying Lee-Carter under conditions of variable mortality decline. *Population studies*, 56(3):325–336, 2002. doi: 10.1080/00324720215935.
- Alexandre Boumezoued. Improving HMD mortality estimates with HFD fertility data. *Working paper*, February 2016.
- Matthias Börger, Daniel Fleischer, and Nikita Kuksin. Modeling the Mortality Trend under Modern Solvency Regimes. *ASTIN Bulletin*, 44(01):1–38, January 2014. ISSN 1783-1350. doi: 10.1017/asb.2013.24.
- Natacha Brouhns, Michel Denuit, and Jeroen K. Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393, 2002. doi: 10.1016/S0167-6687(02)00185-3.
- Andrew J. G. Cairns, David Blake, and Kevin Dowd. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, 73(4):687–718, December 2006. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2006.00195.x.

- Andrew J. G. Cairns, David Blake, and Kevin Dowd. Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, 2008(2-3):79–113, 2008. ISSN 0346-1238. doi: 10.1080/03461230802173608.
- Andrew J. G. Cairns, David Blake, Kevin Dowd, Guy D. Coughlan, David Epstein, Alen Ong, and Igor Balevich. A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States. *North American Actuarial Journal*, 13(1): 1–35, January 2009. ISSN 1092-0277. doi: 10.1080/10920277.2009.10597538.
- Andrew J. G. Cairns, David Blake, Kevin Dowd, and Amy R. Kessler. Phantoms never die: living with unreliable population data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(4):975–1005, 2016a. ISSN 1467-985X. doi: 10.1111/rssa.12159.
- Andrew J. G. Cairns, Malene Kallestrup-Lamb, Carsten PT Rosenskjold, David Blake, Kevin Dowd, and others. Modelling Socio-Economic Differences in the Mortality of Danish Males Using a New Affluence Index. Technical report, Department of Economics and Business Economics, Aarhus University, 2016b.
- Andrew J.G. Cairns, David Blake, Kevin Dowd, Guy D. Coughlan, and Marwa Khalaf-Allah. Bayesian Stochastic Mortality Modelling for Two Populations. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 41(01):29–59, 2011. ISSN 1783-1350. doi: 10.2143/AST.41.1.2084385.
- Celeste M. H. Chai, Tak Kuen Siu, and Xian Zhou. A double-exponential GARCH model for stochastic mortality. *European Actuarial Journal*, 3(2):385–406, December 2013. ISSN 2190-9733, 2190-9741. doi: 10.1007/s13385-013-0077-5.
- A. Chatterjee and S. N. Lahiri. Bootstrapping Lasso Estimators. *Journal of the American Statistical Association*, 106(494):608–625, June 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10159.
- Hua Chen, Richard MacMinn, and Tao Sun. Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, 63:135–146, July 2015. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2015.03.022.
- Marcus C. Christiansen, Evgeny Spodarev, and Verena Unseld. Differences in European Mortality Rates: A Geometric Approach on the Age-Period Plane. *ASTIN Bulletin: The Journal of the IAA*, 45(3):477–502, September 2015. ISSN 0515-0361, 1783-1350. doi: 10.1017/asb.2015.13.
- Iain D Currie, Maria Durban, and Paul HC Eilers. Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298, 2004. ISSN 1471-082X. doi: 10.1191/1471082X04st0800a.

- A. Dokumentov and R.J. Hyndman. smoothAPC: Smoothing of Two-Dimensional Demographic Data, Optionally Taking into Account Period and Cohort Effects. 2018.
- Alexander Dokumentov, Rob J. Hyndman, and Leonie Tickle. Bivariate smoothing of mortality surfaces with cohort and period ridges. *Stat*, 7(1):e199, 2018. ISSN 2049-1573. doi: 10.1002/sta4.199.
- P. Doukhan, D. Pommeret, J. Rynkiewicz, and Y. Salhi. A class of random field memory models for mortality forecasting. *Insurance: Mathematics and Economics*, 77:97–110, 2017. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2017.08.010.
- Kevin Dowd, Andrew J. G. Cairns, David Blake, Guy D. Coughlan, and Marwa Khalaf-Allah. A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, 15(2):334–356, 2011. ISSN 1092-0277. doi: 10.1080/10920277.2011.10597624.
- Vasil Enchev, Torsten Kleinow, and Andrew J. G. Cairns. Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342, January 2016. ISSN 0346-1238. doi: 10.1080/03461238.2015.1133450.
- Jianqing Fan, Jinchi Lv, and Lei Qi. Sparse High Dimensional Models in Economics. *Annual review of economics*, 3:291–317, September 2011. ISSN 1941-1383. doi: 10.1146/annurev-economics-061109-080451.
- Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22, 2010. ISSN 1548-7660.
- Yoel Furman. VAR Estimation with the Adaptive Elastic Net. SSRN Scholarly Paper ID 2456510, Social Science Research Network, Rochester, NY, June 2014.
- Deborah Gefang. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *International Journal of Forecasting*, 30(1):1–11, January 2014. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2013.04.004.
- C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. ISSN 0012-9682. doi: 10.2307/1912791.
- Steven Haberman and Arthur Renshaw. Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics*, 50(3):309–333, 2012. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2011.11.005.
- Lukas Josef Hahn. A Bayesian Multi-Population Mortality Projection Model. Master’s thesis, Universität Ulm, Ulm, Germany, December 2014.
- HMD. *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 2019-01-28)., 2019.

- Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, February 1970. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634.
- Andrew Hunt and David Blake. A General Procedure for Constructing Mortality Models. *North American Actuarial Journal*, 18(1):116–138, January 2014. ISSN 1092-0277. doi: 10.1080/10920277.2013.852963.
- Andrew Hunt and Andrés M. Villegas. Robustness and convergence in the Lee–Carter model with cohort effects. *Insurance: Mathematics and Economics*, 64:186–202, 2015. doi: 10.1016/j.insmatheco.2015.05.004.
- M M Huynen, P Martens, D Schram, M P Weijenberg, and A E Kunst. The impact of heat waves and cold spells on mortality rates in the Dutch population. *Environ Health Perspect*, 109(5):463–470, May 2001. ISSN 0091-6765.
- Rob J. Hyndman. demography: Forecasting Mortality, Fertility, Migration and Population Data. 2019.
- Rob J. Hyndman and Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956, 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2006.07.028.
- Etienne Izraelewicz. L’effet moisson - l’impact des catastrophes vie sur la mortalité à long terme - Exemple de la canicule de l’été 2003. *Bulletin Français d’Actuariat*, 12(24):113–159, 2012.
- Søren Fiig Jarner and Esben Masotti Kryger. Modelling Adult Mortality in Small Populations: The Saint Model. *ASTIN Bulletin: The Journal of the IAA*, 41(2):377–418, November 2011. ISSN 0515-0361, 1783-1350. doi: 10.2143/AST.41.2.2136982.
- Ronald D. Lee and Lawrence R. Carter. Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–671, September 1992. ISSN 0162-1459. doi: 10.1080/01621459.1992.10475265.
- Han Li, Colin O’hare, and Farshid Vahid. Two-Dimensional Kernel Smoothing of Mortality Surface: An Evaluation of Cohort Strength. *Journal of Forecasting*, 35(6):553–563, 2016. ISSN 1099-131X. doi: 10.1002/for.2399.
- Hong Li and Yang Lu. Coherent Forecasting of Mortality Rates: A Sparse Vector-Autoregression Approach. *ASTIN Bulletin: The Journal of the IAA*, 47(2):563–600, May 2017. ISSN 0515-0361, 1783-1350. doi: 10.1017/asb.2016.37.

- Nan Li and Ronald Lee. Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, 42(3):575–594, August 2005. ISSN 1533-7790. doi: 10.1353/dem.2005.0021.
- Nan Li, Ronald Lee, and Patrick Gerland. Extending the Lee-Carter method to model the rotation of age patterns of mortality-decline for long-term projection. *Demography*, 50(6): 2037–2051, December 2013. ISSN 0070-3370. doi: 10.1007/s13524-013-0232-2.
- Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37, 2007. ISSN 1752-0509. doi: 10.1186/1752-0509-1-37.
- Pierre Perron. Trends and random walks in macroeconomic time series. *Journal of Economic Dynamics and Control*, 12(2):297–332, June 1988. ISSN 0165-1889. doi: 10.1016/0165-1889(88)90043-7.
- Richard Plat. On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404, 2009. doi: 10.1016/j.insmatheco.2009.08.006.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019.
- A. E. Renshaw and S. Haberman. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570, June 2006. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2005.12.001.
- A. E. Renshaw and S. Haberman. On simulation-based approaches to risk measurement in mortality with specific reference to Poisson Lee–Carter modelling. *Insurance: Mathematics and Economics*, 42(2):797–816, 2008. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2007.08.009.
- Said E. Said and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, December 1984. ISSN 0006-3444. doi: 10.1093/biomet/71.3.599.
- Yahia Salhi and Stéphane Loisel. Basis risk modelling: a cointegration-based approach. *Statistics*, 51(1):205–221, January 2017. ISSN 0233-1888. doi: 10.1080/02331888.2016.1259806.
- Juliane Schäfer and Korbinian Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1175.
- Song Song and Peter J. Bickel. Large Vector Auto Regressions. *arXiv:1106.3915 [q-fin, stat]*, June 2011. arXiv: 1106.3915.

- Evgeny Spodarev, Elena Shmileva, and Stefan Roth. Extrapolation of stationary random fields. *arXiv preprint arXiv:1306.6205*, 2013.
- Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x.
- Laurent Toulemon and Magali Barbieri. The mortality impact of the August 2003 heat wave in France: Investigating the ‘harvesting’ effect and other long-term consequences. *Population Studies*, 62(1):39–53, March 2008. ISSN 0032-4728. doi: 10.1080/00324720701804249.
- Simone Vazzoler, Lorenzo Frattarolo, and Monica Billio. sparsevar: A Package for Sparse VAR/VECM Estimation. Technical report, 2016.
- Andrés M. Villegas, Vladimir Kaishev, and Pietro Millossovich. StMoMo: An R Package for Stochastic Mortality Modelling. 2017.
- R. C. Willets. The Cohort Effect: Insights and Explanations. *British Actuarial Journal*, 10(4):833–877, October 2004. ISSN 2044-0456, 1357-3217. doi: 10.1017/S1357321700002762.
- Ines Wilms and Christophe Croux. Forecasting using sparse cointegration. *International Journal of Forecasting*, 32(4):1256–1267, October 2016. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2016.04.005.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x.

Chapitre 2

Elastic-net et cohérence locale pour la modélisation des dynamiques de mortalité inter-population

Ce chapitre reprend l'article "Bridging the Li-Carter's gap : a locally coherent mortality forecasting approach", co-écrit avec Quentin Guibert, Stéphane Loisel et Olivier Lopez.

Abstract

Countries with common features in terms of social, economic and health systems generally have mortality trends which evolve in a similar manner. Drawing on this, many multi-population models are built on a coherence assumption which inhibits the divergence of mortality rates between two populations, or more, on the long run. However, this assumption may prove to be too strong in a general context, especially when it is imposed to a large collection of countries. We also note that the coherence hypothesis significantly reduces the spectrum of achievable mortality dispersion forecasts for a collection of populations when comparing to the historical observations. This may distort the longevity risk assessment of an insurer. In this paper, we propose a new model to forecast multiple populations assuming that the long-run coherent principle is verified by subgroups of countries that we call the "locally coherence" property. Thus, our specification is built on a trade-off between the Lee-Carter's diversification and Li-Lee's concentration features and allows to fit the model to large number of populations simultaneously. A penalized vector autoregressive (VAR) model, based on the elastic-net regularization, is considered for modeling the dynamics of common trends between subgroups. Furthermore, we apply our methodology on 32 European populations mortality data and discuss the behavior of our model in terms of simulated mortality dispersion. Within the Solvency II directive, we quantify the impact on the longevity risk solvency capital requirement of an insurer for a simplified pensions product. Finally, we extend our model by allowing

populations to switch from one coherence group to another. We then analyze its incidence on longevity hedges basis risk assessment.

Keywords: Coherent mortality forecasting; Multipopulation; Basis risk; Vector Autoregression; Elastic-Net.

2.1 Introduction

Around the world, the life expectancy of many countries has significantly risen over the past decades. This general augmentation implicitly suggests that common drivers are shared between populations such as lifestyle factors, the level of socioeconomic inequalities or the quality of the health system. However, this improvement is not equally shared among all groups of populations, precisely because the driving features are heterogeneous at an individual level. The links and differences between groups are important to understand for demographic and actuarial studies, since they may cause issues in managing social security system, pension funds or in defining longevity risk hedging solution. This explains the growing interest in mortality forecasting of multiple populations, especially to understand the effect of sub-groups on an overall population (Danesi et al., 2015; Bergeron-Boucher et al., 2018; Cairns et al., 2019) or to enhance the projections robustness for populations with not well-known trend (see e.g Li et al., 2018).

Since the introduction of the augmented common factor approach by Li and Lee (2005), many multi-population models have been developed. Most of them are extensions of the Lee and Carter (1992) model or the CDB models (Cairns et al., 2006), and rely on a so-called coherence principle (see Villegas et al., 2017, for a review). This concept, introduced by Li and Lee (2005), consists in imposing that the mortality rates of two populations (or more) do not diverge in the long run. In such a specification, common age and period effects are estimated for a group of mortality data, the population specific trend being therefore modeled via mean-reverting process. This principle is also introduced in the so-called GRAVITY model (Dowd et al., 2011) or the Common Age Effect (CAE) model (Kleinow, 2015; Enchev et al., 2016). Although this assumption may be relevant for close populations, Li et al. (2017) have pointed out that it is often only suitable for specific populations and over limited time windows. Hence, this approach is more likely to become unsatisfactory when it is applied to a large collection of populations. Furthermore, by inhibiting the divergence between populations, an irrelevant use of the mortality convergence principle may thereby distort the projections spectrum of mortality dispersion among the populations. In that context, extending the current mortality forecasting models for a large number of populations is challenging as it involves to cluster together similar countries and separately forecasting those with diverging mortality dynamics, as noted by Richman and Wuthrich (2018).

Motivated by the need of applying stochastic mortality models to a large collection of populations, our aim in this paper is to find a trade-off between the full coherence principle between populations and the noncoherence situation where single stochastic mortality models are considered independently. These two situations can be considered as border cases of the inter-group dependence possibilities, and are both undesirable, in a general context, for life insurers which are assisted by stochastic mortality multi-population models in their longevity risk solvency capital assessment. Indeed, a fully coherent model can overestimate the capital requirement through the concentration between the populations it imposes, whereas the use of independent single-population models may induce an underestimation of this capital. Similar problem appears also for the pricing and the hedging of longevity risk transfer solutions where an accurate risk assessment is required, especially for the corresponding basis risk.

To overcome this issue, [Li et al. \(2017\)](#) introduce the concept of semicoherence which contains the mortality differentials into a tolerance corridor. In other words, the two population mortality dynamics can diverge over certain periods of time, but not permanently. Their semicoherent mortality projection is based on a threshold vector autoregressive process, which can be seen as a vector autoregressive (VAR) model characterized by a parameter set and autoregressive order which change according of the time series regime switch. [Zhou et al. \(2019\)](#) recently improve this approach by introducing a threshold vector error correction model and taking into account long-term equilibrium between time series. However, these models are developed for only two populations and their VAR structure contained many parameters, which is a serious limitation in the large-scale mortality forecasting framework.

In this paper, we propose an alternative approach, named "local coherence", which allows us to model simultaneously a large collection of populations by assuming the mortality coherence principle is satisfied for several subgroups of populations. Although our approach is mainly based on existing models, such as the Lee-Carter or the Li-Lee models, it intends to gain in flexibility by offering a trade-off between diverging mortality forecasts, obtained with a single stochastic mortality model, and fully coherence forecasts. In that way, our goal is similar to the [Li et al. \(2017\)](#) semicoherence one, i.e. offering a compromise between the fully coherence and the independence situations, even if the approach differs. Notably, in our case the size of the populations collection is not an issue and the proposed model includes the two boarder cases.

Within a locally coherent group, the dynamics of the mortality is based on a [Li and Lee \(2005\)](#) model. To capture both long-term relationships and short-term interactions between groups, a VAR process models the improvements in the common mortality trends of each group. A similar specification for a small group of populations is proposed by several authors in the literature (see [Zhou et al., 2014](#); [Kleinow, 2015](#); [Enchev et al., 2016](#), among others). However, a VAR model may suffer of over-fitting issue when the size of the group increases, since historical series are relatively short and strongly correlated. For our large-scale general

context, we then use a VAR-ENET approach. Based on an elastic-net regularization technique, this VAR estimation is relevant with high dimensional time series and offers good in-sample and out-of-sample performances in a mortality forecasting framework (Guibert et al., 2019).

For illustrative purpose, we estimate our locally coherent model on a collection of 16 European countries, by distinguishing the gender. We propose two approaches to cluster these populations in coherent groups, regarding to the historical mortality rates data or expert judgments. Our results highlight a better control of the mortality rates dispersion forecasting among the populations. In addition, we show the importance of having an accurate level of variability when considering the solvency capital requirement of an European life insurer exposed to longevity risks from multiple populations.

Furthermore, we extend our model to a dynamic point of view of the coherence property by allowing the populations to switch from one group of coherence to another. Thus, the coherence is local not only in the spatial dimension (i.e. according to the populations) but also in the temporal dimension. Such switch can be caused by changes in socio-economic features driving the mortality dynamics. We underpin the importance this extension by analyzing the impact of a specific switch on a Longevity Divergence Index Value, similar to the one that the Kortis bond is based on.

The remainder of this paper is organized as follows. In Section 2.2, we recall the usual Lee-Carter and Li-Lee models and show the mortality dispersion forecasting issue. Section 2.3 introduces our locally coherent model and shows how it fills the gap between a fully coherent and an independent approach. We also present the VAR-ENET estimation method retained in the paper. Section 2.4 presents two clustering approaches for identifying coherent groups: one based on expert judgments and a data-driven approach. The results obtained on the groups composed of 32 European populations is discussed in Section 2.5, and demonstrate the interest for the risk-based capital requirement of a life insurer in presence of multipopulation exposures. Section 2.6 extends our model to a dynamic local coherence, and illustrates its impact on a basis risk assessment example. Finally, Section 2.7 concludes the paper and gives several improvements for future researches.

2.2 The mortality dispersion issue

2.2.1 Notation

As noted in the Section 2.1, a large range of models have been developed for mortality modeling and forecasting within the multipopulation framework. In this paper, we do not limit our study to the specific two-population context: we consider a more general scope of a collection \mathcal{I} of populations, where its cardinal I is potentially high. Furthermore, we note \mathcal{X} and \mathcal{T} the collections of integer ages and calendar years respectively.

For $(i, x, t) \in \mathcal{I} \times \mathcal{X} \times \mathcal{T}$, we note $D_{x,t}^{(i)}$ the number of people in the i -th population who die in year t and aged x at their last birthday. The so-called "exposure to risk", noted $E_{x,t}^{(i)}$, represents the amount of *person years* lived by people of population i aged $[x, x + 1)$ in year $[t, t + 1)$. Thus, the central death rate $m_{x,t}^{(i)}$ is given by

$$m_{x,t}^{(i)} = \frac{D_{x,t}^{(i)}}{E_{x,t}^{(i)}}. \quad (2.1)$$

In the two populations framework, the coherence can be analyzed by focusing on the difference of log-mortality rates at a specific age x , which we note $\delta_{x,t} = \ln m_{x,t}^{(1)} - \ln m_{x,t}^{(2)}$. The coherence hypothesis specifies that for all $x \in \mathcal{X}$, the series $(\delta_{x,t})_t$ do not diverge. We extend this measure to the general multipopulation case where I is potentially important. Thus, for an age and time (x, t) , and a collection \mathcal{I} of populations, we define $\delta_{x,t}^{\mathcal{I}}$ the dispersion of mortality rates:

$$\delta_{x,t}^{\mathcal{I}} = \sqrt{\frac{1}{I-1} \sum_{i \in \mathcal{I}} \left(\ln m_{x,t}^{(i)} - \overline{\ln m_{x,t}} \right)^2}, \quad (2.2)$$

where $\overline{\ln m_{x,t}} = \frac{1}{I} \sum_{i \in \mathcal{I}} \ln m_{x,t}^{(i)}$ is the mean of log mortality rates of the different populations from the collection \mathcal{I} . In a way, this measure allows us to evaluate the heterogeneity of the mortality levels among a set of population. The specific case of a coherent multipopulation model corresponds to a situation where for all $x \in \mathcal{X}$ the dispersion $(\delta_{x,t}^{\mathcal{I}})_t$ is controlled and can not diverge.

2.2.2 The Lee-Carter model

There is a broad variety of mortality models that have been introduced in the actuarial science and demography literature. In this paper, we choose to focus on the Lee-Carter model, which is one of the most used by practitioners. We first quickly recall the model using the previous notation and adopting a multipopulation point of view. Thus, [Lee and Carter \(1992\)](#) propose to model the central death rates such as

$$\ln m_{x,t}^{(i)} = \alpha_x^{(i)} + \beta_x^{(i)} \kappa_t^{(i)} + \epsilon_{x,t}^{(i)}, \quad (2.3)$$

where $\alpha_x^{(i)}$ and $\beta_x^{(i)}$ are age-specific effect parameters, $\kappa_t^{(i)}$ represent the temporal mortality dynamics, and $\epsilon_{x,t}^{(i)}$ are residual terms. For the sake of the uniqueness of the solution we impose the usual constraints in the estimation process:

$$\sum_{t \in \mathcal{T}} \kappa_t^{(i)} = 0 \text{ and } \sum_{x \in \mathcal{X}} \beta_x^{(i)} = 1. \quad (2.4)$$

Thus, $\alpha_x^{(i)}$ are set to be the averages of log mortality central rates over the period \mathcal{T} considered, and the series of parameters $\beta_x^{(i)}$ and $\kappa_t^{(i)}$ are estimated thanks to the application of the singular value decomposition (SVD) method.

Furthermore, we model the dynamic of the time series $\kappa_t^{(i)}$ as a random walk with a drift $c^{(i)}$:

$$\kappa_t^{(i)} = c^{(i)} + \kappa_{t-1}^{(i)} + e_t^{(i)}, \quad (2.5)$$

where the innovations $\left(e_t^{(i)}\right)_{i \in \mathcal{I}}^\top$ is an I -dimensional Gaussian white noises with mean 0 and a diagonal variance-covariance matrix Σ_1 .

The Lee-Carter (LC) being a single population mortality model, we note that no relationships between the different populations are taken into account. As a consequence, the derived mortality forecasts are independent and can diverge, creating a large diversification effect when forecasting the longevity risk within a stochastic multi-population framework.

2.2.3 The coherent Li-Lee model

On the contrary, [Li and Lee \(2005\)](#) extend the Lee-Carter model to the multi-population scope by imposing a common trend $B_x^\mathcal{I} K_t^\mathcal{I}$ to the collection \mathcal{I} of considered populations:

$$\ln m_{x,t}^{(i)} = \alpha_x^{(i)} + B_x^\mathcal{I} K_t^\mathcal{I} + \beta_x^{(i)} \kappa_t^{(i)} + \epsilon_{x,t}^{(i)}. \quad (2.6)$$

The coherence is then enforced by modeling the dynamic of time series $\kappa_t^{(i)}$ to be a mean-reverting process. We retain here a first order autoregressive (AR) model:

$$\kappa_t^{(i)} = a^{(i)} \kappa_{t-1}^{(i)} + r_t^{(i)}, \quad (2.7)$$

where the innovations $\left(r_t^{(i)}\right)_{i \in \mathcal{I}}^\top$ are an I -dimensional Gaussian white noises with mean 0 and a diagonal variance-covariance matrix Σ_2 .

We denote $m_{x,t}^\mathcal{I}$ the mortality rates of the global population, i.e. the gathering of all populations from the collection \mathcal{I} :

$$m_{x,t}^\mathcal{I} = \frac{D_{x,t}^\mathcal{I}}{E_{x,t}^\mathcal{I}}, \quad (2.8)$$

where $D_{x,t}^\mathcal{I} = \sum_{i \in \mathcal{I}} D_{x,t}^{(i)}$ and $E_{x,t}^\mathcal{I} = \sum_{i \in \mathcal{I}} E_{x,t}^{(i)}$. Thus, the fitting of a Lee-Carter model on these series, under the usual constraints $\sum_{t \in \mathcal{T}} K_t^\mathcal{I} = 0$ and $\sum_{x \in \mathcal{X}} \beta_x^\mathcal{I} = 1$, gives the common trend

factor estimates. Like in the single population context, the age–population specific series of parameters $\alpha_x^{(i)}$ are set to be the averages of log-mortality central rates over the training period \mathcal{T} . Finally, the population–specific temporal dynamic factors $\beta_x^{(i)} \kappa_t^{(i)}$ are obtained thanks to the SVD method, with the constraints $\sum_{t \in \mathcal{T}} \kappa_t^{(i)} = 0$ and $\sum_{x \in \mathcal{X}} \beta_x^{(i)} = 1$, for each $i \in \mathcal{I}$.

Similarly to the Lee–Carter model, time series $K_t^{\mathcal{I}}$ is modeled as a random walk with drift:

$$K_t^{\mathcal{I}} = c^{\mathcal{I}} + K_{t-1}^{\mathcal{I}} + e_t^{\mathcal{I}}, \quad (2.9)$$

where $e_t^{\mathcal{I}}$ is a Gaussian white noise with mean 0 and related variance $\sigma_{\mathcal{I}}^2$.

Thus defined, the Li–Lee (LL) extension enforces the specific mortality rates of the whole collection \mathcal{I} of populations to follow the same trend and converge in the long-run. Even if the population–specific temporal factors $\kappa_t^{(i)}$ allow a population to slightly deviate from this trend for a short period of time, the mean–reverting property brings it back to the common dynamic. From a longevity risk management point of view, it then implies a limited diversification among the population in a stochastic evaluation context.

2.2.4 Mortality dispersion on the HMD

Let us now focus on a group of 16 western European countries, namely: Austria (AUT), Belgium (BEL), Switzerland (CHE), West Germany (DEUW), Denmark (DNK), Spain (ESP), Finland (FIN), France (FRA), Ireland (IRL), Italy (ITA), Luxembourg (LUX), Netherlands (NLD), Norway (NOR), Portugal (PTR), Sweden (SWE) and Great Britain (GBR). Furthermore, we consider separately the male and female populations for each country. Thus, our collection \mathcal{I} of interest is composed of $I = 32$ populations, each one of them being defined by a couple (country, sex). The data is extracted from the Human Mortality Database (HMD, 2019). For the age–period training set, we retain $\mathcal{X} \times \mathcal{T} = \{45, \dots, 90\} \times \{1960, \dots, 2014\}$, and we fit single LC models as well as a LL model, as described in Sections 2.2.2 and 2.2.3. The Figure 2.1 represents the dispersion $\delta_{85,t}^{\mathcal{I}}$ estimated on the historical data, i.e. \mathcal{T} , and on the projections over 50 years obtained with the stochastic mortality models. For the latter, we plot the median together with the 95% prediction intervals derived from 500 simulations.

First, we remark that the dispersion measure at age 85 has nearly doubled between the 1960’s and the 2000’s, and seems more stable in recent years. As expected, the predicted dispersion is significantly different between the results from the LC and the LL models. In the first case, both the median and the width of the prediction intervals of the dispersion increase linearly with the projection horizon. In other words, the mortality rates at age 85 of the considered populations become more and more dissimilar in average, and this heterogeneity significantly varies from one stochastic scenario to another. On the contrary, the LL’s forecasts exhibit a strongly stable dispersion. Not only the median is constant over time, but also the confidence intervals are very narrow comparing to the historical data and their bandwidth does not

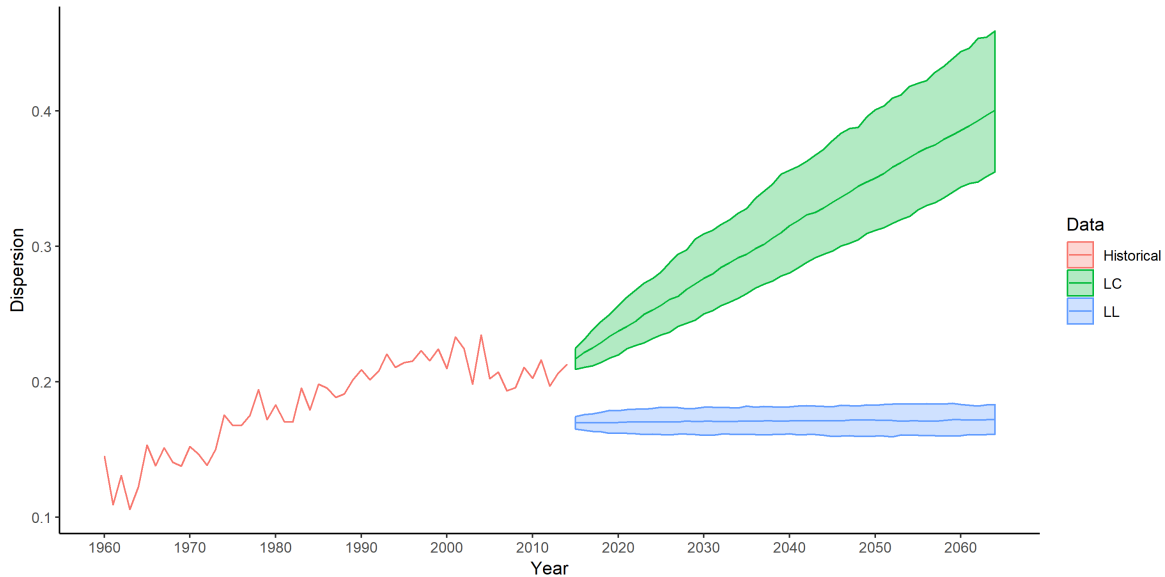


Figure 2.1 – Dispersion at age 85 in the western European populations: historical data and median projections by Lee-Carter and Li-Lee models with the corresponding 95% prediction intervals.

augment with the projection horizon. In term of generated mortality scenarios, it implies that a considerable majority of them lead to very similar pattern of mortality rates, even in the long term.

Both the LC and LL models present drawbacks when analyzing the projected dispersion in a longevity risk management framework. By not imposing any relationships between the population mortality dynamics, the LC model artificially creates some diversification in term of longevity risk. Contrariwise, the LL model imposes a strong coherence hypothesis on all the populations, thus leaving no significant scope for even a limited deviation of the mortality between the populations: it assumes a high concentration risk in the longevity evaluation. Intuitively, a longevity risk manager may want to consider some intermediate scenarios between the exaggerated LC’s diversification and LL’s over-concentration.

2.3 A locally coherent approach

To bridge the gap between the Lee-Carter’s diversification and the Li-Lee’s concentration, we propose a new model based on the idea that the populations are coherent by homogeneous groups, and not all together at the same time. In the populations space, it can be seen as a locality of the coherence property.

2.3.1 The model

Keeping the notation introduced before, we now denote \mathcal{J} a partition of the populations collection \mathcal{I} in J distinct groups. Let $\phi : \mathcal{I} \rightarrow \mathcal{J}$ be the corresponding classification function that labels a population to a specific group. We then propose the following model for each population $i \in \mathcal{I}$:

$$\ln m_{x,t}^{(i)} = \alpha_x^{(i)} + B_x^{\phi(i)} K_t^{\phi(i)} + \mathbb{1}_{\{\#\phi(i)>1\}} \beta_x^{(i)} \kappa_t^{(i)} + \epsilon_{x,t}^{(i)}, \quad (2.10)$$

where $\# : \mathcal{J} \rightarrow \mathbb{N}$ is the cardinal function of a group, and $\phi(i)$ denotes the coherence group of the population i .

Similarly to the LL model, for each group $j \in \mathcal{J}$, the local common trend factors B_x^j and K_t^j are estimated by fitting a Lee–Carter thanks to the SVD method on the group log mortality rates data,

$$m_{x,t}^j = \frac{D_{x,t}^j}{E_{x,t}^j}, \quad (2.11)$$

where $D_{x,t}^j = \sum_{i \in j} D_{x,t}^{(i)}$ and $E_{x,t}^j = \sum_{i \in j} E_{x,t}^{(i)}$, with the usual uniqueness constraints. For each population i , the series of age factors $\alpha_x^{(i)}$ are set to be the averages of log mortality rates over the estimation period \mathcal{T} . Finally, the population–specific mortality dynamic parameters $\beta_x^{(i)}$ and $\kappa_t^{(i)}$ are again estimated through SVD, if the considered population i is assumed coherent with at least one other population.

To impose the local coherence, we keep the AR(1) dynamic of Equation (2.7) for the population–specific period factors, but, in the present more general definition, we do not impose the variance–covariance matrix Σ_2 of the innovations $\left(r_t^{(i)}\right)_{i \in \mathcal{I}}^\top$ to be diagonal anymore. Thus, we catch short–term relationships between the populations.

Likewise, we also generalize the temporal dynamics of the period group factors to capture possible relationships between the different clusters of populations. Following the underlying ideas of Kleinow (2015) and Zhou et al. (2014), we retain a vector autoregressive (VAR) model. For $j \in \mathcal{J}$, we note $\Delta K_t^j = K_t^j - K_{t-1}^j$ the common mortality improvement of the cluster. Given a temporal lag p , the dynamic of the VAR(p) is then given by

$$\Delta \mathbf{K}_t = \mathbf{C} + \sum_{k=1}^p \mathbf{A}_k \Delta \mathbf{K}_{t-k} + \mathbf{E}_t, \quad (2.12)$$

where $\Delta \mathbf{K}_t = \left(\Delta K_t^j\right)_{j \in \mathcal{J}}^\top$ is the vector of mortality improvements at a group level, for $k =$

$1, \dots, p$, \mathbf{A}_k are $J \times J$ -autoregressive matrices, \mathbf{C} is a J -dimensional vector of drifts, and \mathbf{E}_t is a J -dimensional Gaussian white noise with mean 0 and Σ_1 the related covariance matrix. The matrices \mathbf{A}_k , $k \in \{1, \dots, p\}$, capture the long-run relationships of mortality improvements between groups of coherent populations, while the variance-covariance matrix Σ_1 estimates the short-term dependence structure of innovation shocks.

2.3.2 Border cases

Before continuing to the practical application of the model and its assessment, we focus on two special border cases. Indeed, by choosing some particular clusters and setting specific parameters, we note that our model includes the original Lee-Carter and Li-Lee models.

To recover the LC model we need to impose the following specifications.

- All the coherence clusters contain only one population, i.e. $\forall i \in \mathcal{I}, \phi(i) = \{i\}$. In other words we do not impose any coherence between the populations of interest.
- The temporal lag p of the VAR dynamic is set to 0. This point can also be approximated by applying a high λ in the elastic-net estimation process. The vector \mathbf{C} of Equation (2.12) is then equivalent to the concatenation of the population-specific drifts $c^{(i)}$ from Equation (2.5).
- A diagonal structure is enforced to the variance-covariance matrix Σ_1 , i.e. no short-term dependence structure is estimated between the populations.

Similarly, we can also retrieve the LL framework from our model.

- By grouping all the populations into a single cluster, i.e. $\forall i \in \mathcal{I}, \phi(i) = \mathcal{I}$, the coherence property is imposed to the whole demographic set.
- The temporal lag p of the VAR dynamic is set to 0. In particular, we remark that it is a 1-dimensional VAR due to the partition specification.
- A diagonal structure is enforced to the variance-covariance matrix Σ_2 . Further more, Σ_1 being a 1×1 -matrix, we recover the $\sigma_{\mathcal{I}}^2$ of the LL's common temporal trend dynamic.

Hence, our model framework can be seen as a bridge between the Lee-Carter and the Li-Lee ones. Thereby, for the rest of this paper, we denote it by LC-LL.

2.3.3 VAR Elastic-Net

In a general context, one can consider a large number J of groups by assuming that the coherence hypothesis is too strong for the populations of interest, thereby increasing dramatically the number of parameters to be estimated in the VAR(p). From a statistic point of view, it may lead to high-dimensional problems that we want to avoid. Thus, we decide to apply the extension of the elastic-net regularization and variable selection method, proposed by [Zou and Hastie \(2005\)](#), for the high-dimensional estimation of our autoregressive matrices. This technique can be seen as the combination of the LASSO \mathcal{L}_1 -penalty, introduced by [Tibshirani \(1996\)](#), and the ridge \mathcal{L}_2 -penalty developed by [Hoerl and Kennard \(1970\)](#). Elastic-net has

similar properties of variable selection as the LASSO. Moreover, it provides a grouping effect: highly correlated variables tend to be selected or dropped together. LASSO and elastic-net have already been extended to VAR model (Gefang, 2014; Basu and Michailidis, 2015), mostly with an economic application (see e.g. Song and Bickel, 2011; Furman, 2014) and more recently for mortality forecasting purpose (Guibert et al., 2019).

Therefore, we estimate the VAR(p) model presented in Equation (2.12) with T observations of the process $\Delta \mathbf{K}_t$ for $t \in \mathcal{T} = \{t_{\min}, \dots, t_{\max}\}$ by minimizing the criterion

$$\begin{aligned}
 L(\mathbf{C}, \mathbf{A}_1, \dots, \mathbf{A}_p) &= \frac{1}{T-p} \sum_{t=t_{\min}+p}^{t_{\max}} \left\| \Delta \mathbf{K}_t - \mathbf{C} - \sum_{k=1}^p \mathbf{A}_k \Delta \mathbf{K}_{t-k} \right\|_2^2 \\
 &\quad - \alpha \lambda \sum_{k=1}^p \|\mathbf{A}_k\|_1 - \frac{(1-\alpha)\lambda}{2} \sum_{k=1}^p \|\mathbf{A}_k\|_2^2,
 \end{aligned} \tag{2.13}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively denote the \mathcal{L}_1 and \mathcal{L}_2 norms, $\lambda > 0$ determines the strength of the penalization, and $\alpha \in [0, 1]$ represents the mix between ridge ($\alpha = 0$) and LASSO ($\alpha = 1$) penalties. In particular, when α is set such as some LASSO penalty is enforced, i.e. $\alpha > 0$, the higher λ gets, the fewer number of variables is selected.

For sparsity purpose we impose a larger weight to the LASSO penalty by setting α to 0.9. The hyper-parameter λ is then estimated thanks to a 10-folds cross-validation method. To apply this estimation process, we use the `sparsevar` R-package (Vazzoler et al., 2016), which is based on the `glmnet` one (Friedman et al., 2010). Furthermore, we choose to retain a temporal lag $p = 4$.

2.4 Population clustering

In this section, we examine some possibilities of retained clustering method to determine the coherence groups among the population collection. The aim here is rather to give an outlook of some alternatives rather than describing an exhaustive list of options and how to optimize the choice, if an optimal solution can be defined / found. At first glance, the methods can roughly be split between those based on expert opinions and those more data-driven. In the following, we give one example of each for the collection of western European countries studied in the Section 2.2.4.

2.4.1 Coherence by country

In our case, the set of countries we are focusing on tend to strengthen their gender policy. Thus, in the long run, one can assume that within a single country both male and female populations have similar lifestyle with access to the same level of education, wealthiness and

healthcare, leaving mostly only biological differences. Thereby, we define the partition \mathcal{J}_{MF} such as it clusters the populations by countries. The number of groups J equals then to 16. The corresponding model is noted LC–LL(MF).

An interesting point of this approach, is that the VAR model estimates a dependence structure between the countries, allowing a rather easy reading of the autoregressive matrices that represent the underlying dynamics. Indeed, the common period factors $\left(K_t^j\right)_{j \in \mathcal{J}_{MF}}$ are equivalent to the ones obtained by fitting single population LC models on the overall country. We display the Granger causality matrices in Figures 2.2a to 2.2d.

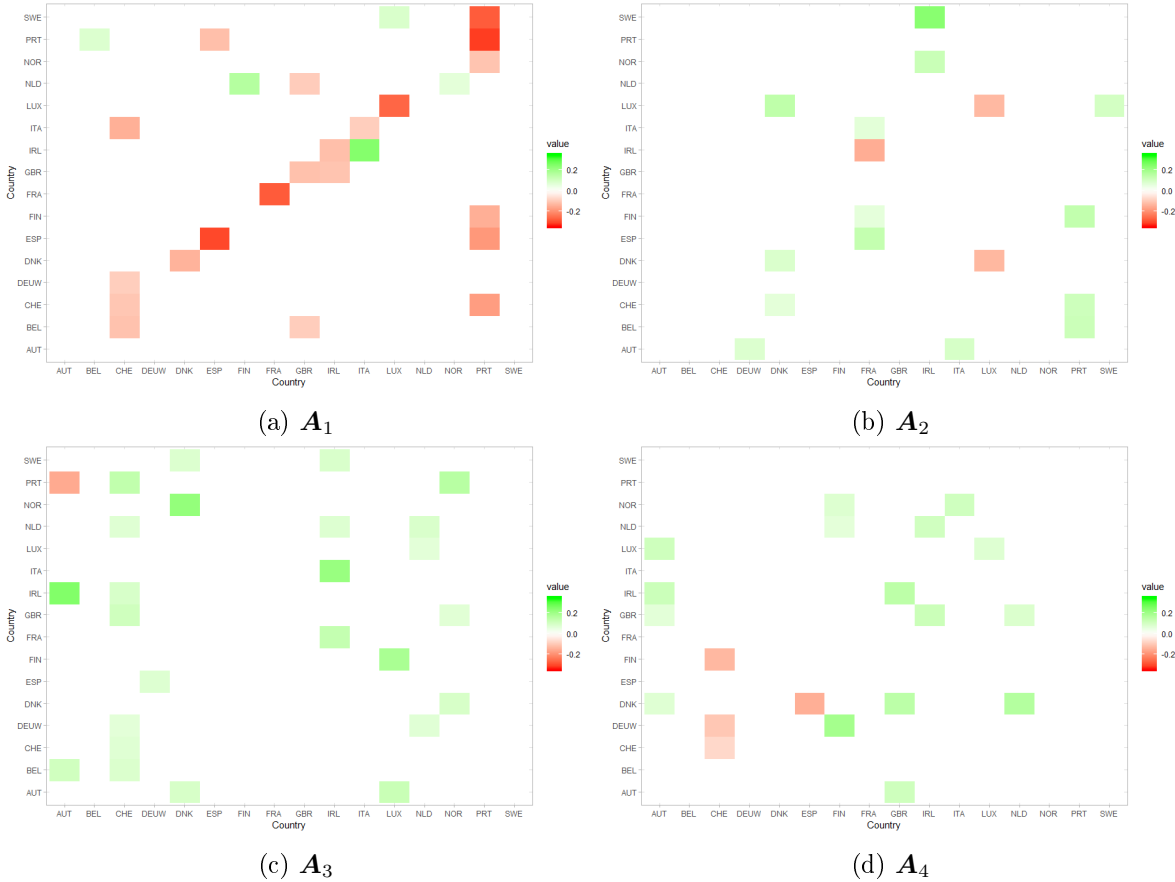


Figure 2.2 – Estimated autoregressive matrices of the VAR-ENET from the LC–LL(MF) model.

The matrix \mathbf{A}_1 (Figure 2.2a) exhibits mostly negative coefficients, shaped into a diagonal structure. We hereby capture the country–specific period effect. Indeed, when modelling the single population LC’s $\kappa_t^{(i)}$ dynamics by an AR(1), we generally obtain negative autoregressive estimates. Displayed in a matrix way for a multipopulation point of view, it leads to a negative diagonal matrix. This result, outlined by the use of elastic–net estimation process, recall the diagonal structure found by Guibert et al. (2019) in a single population study.

We also remark some vertical patterns within the Granger causality matrices. In term of

mortality dynamics, this should be interpreted as a persistent effect of a single country to a group countries. For example, in the autoregressive matrix \mathbf{A}_3 (Figure 2.2c), we note the existence of a positive vertical structure on the Swiss population. Hence, the Swiss mortality improvement seems to Granger cause the mortality improvement on 6 other countries, namely BEL, DEUW, GBR, IRL, NLD and PRT. A chock in the mortality rates in the Switzerland would then reverberate 3 years later in the 6 countries. In this way, the vertical structure can be seen as some kind of leader effect.

2.4.2 A data mining approach

The main idea of the coherence in the Li–Lee model is to impose a common trend $(K_t^{\mathcal{I}})_{t \in \mathcal{T}}$ in the mortality dynamics of each population i . This is done by fitting a Lee–Carter on the overall population of the retained collection \mathcal{I} (see Section 2.2.3). Hence, for a single population i , we consider its time series $(\kappa_t^{(i)})_{t \in \mathcal{T}}$, derived from the LC fitting, as a signature of the mortality dynamics. To determine the coherence groups, we then apply an unsupervised clustering method on these signatures. We choose the well-known hierarchical cluster analysis (HCA) method. In particular, we retain the Euclidean metric and Ward’s criterion as dissimilarity measure and linkage criterion respectively. We display the corresponding dendrogram in the Figure 2.3.

We remark that, unlike the previous clustering, the populations are not gathered by country. It even seems to be the contrary: the sex tends to be one of the major splitting criteria. Yet, we note one exception to this comment: Denmark. The two Danish populations are clustered together, apart from other countries, suggesting that the Danish mortality dynamic exhibits some specific features when comparing to other western European countries.

For the following numerical applications, we choose to retain 8 different clusters, which are emphasized in the Figure 2.3 through the colors and recalled in the Table 2.1. We denote the LC–LL mortality model based on this clustering by LC–LL(HCA8), and the corresponding partition by $\mathcal{J}_{\text{HCA8}}$.

Table 2.1 – 8 groups of populations determined by HCA on the LC’s κ_t .

| Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 | Group 8 |
|---------|---------|---------|---------|---------|---------|---------|---------|
| DEUW M | NOR M | GBR M | NOR F | DNK M | PRT F | DEUW F | FRA F |
| BEL M | NLD M | ITA M | NLD F | DNK F | FIN M | BEL F | CHE F |
| FRA M | SWE M | AUT M | SWE F | | AUT F | | FIN F |
| GBR F | | LUX M | | | IRL F | | ITA F |
| PRT M | | IRL M | | | CHE M | | ESP F |
| ESP M | | | | | LUX F | | |

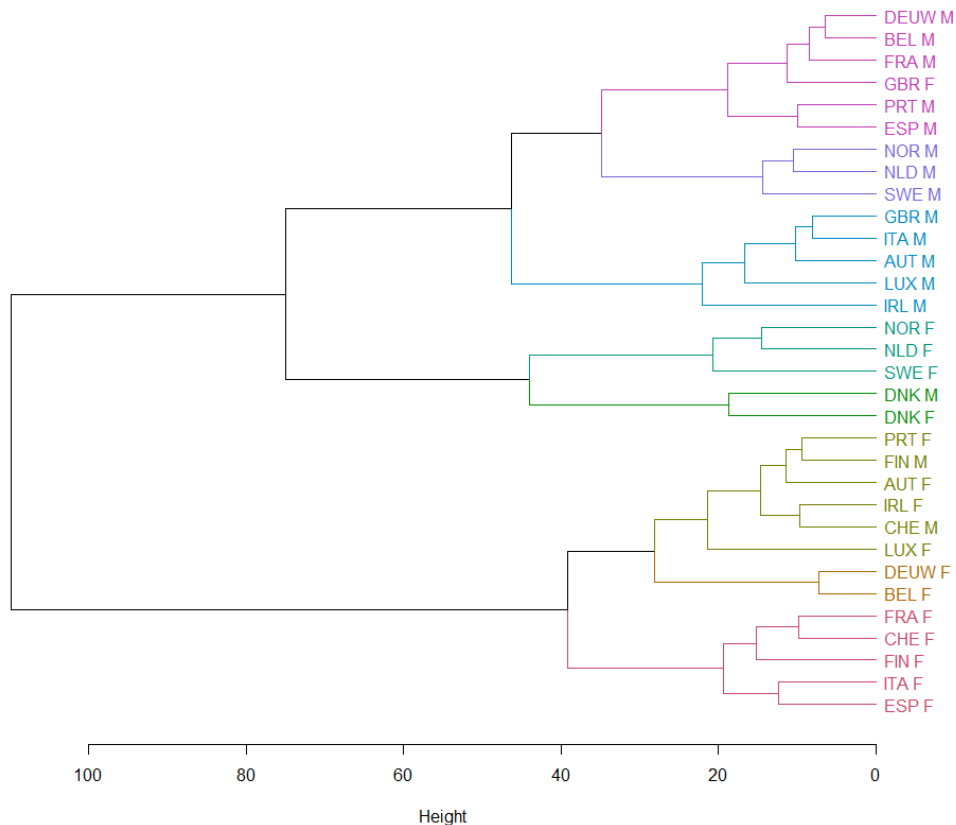


Figure 2.3 – Dendrogram associated with the HCA applied on the LC’s κ_t , colored according to 8 final clusters.

2.5 Numerical applications

In this section, we compare the forecasting results from the different models that we introduced: LC, LL, LC–LL(MF) and LC–LL(HCA8). Rather than focusing on the main mortality trends that can be derived from these models, in the scope of this paper we are more interesting to the stochastic scenarios spectrum they are producing.

2.5.1 Dispersion forecasting

Following the Section 2.2.4, we display on Figures 2.4 and 2.5 the historical dispersion $\delta_{85,t}^{\mathcal{I}}$ for the 32 populations of \mathcal{I} at age 85, derived from the historical data, the LC and LL model forecasts but, this time, together with the projection from the LC–LL(MF) and LC–LL(HCA8) respectively. The same graphs are displayed in Appendix 2.8 for age 70, 75, 80 and 90. As expected, the dispersion predictions produced by the LC–LL models are contained between the LC and LL ones.

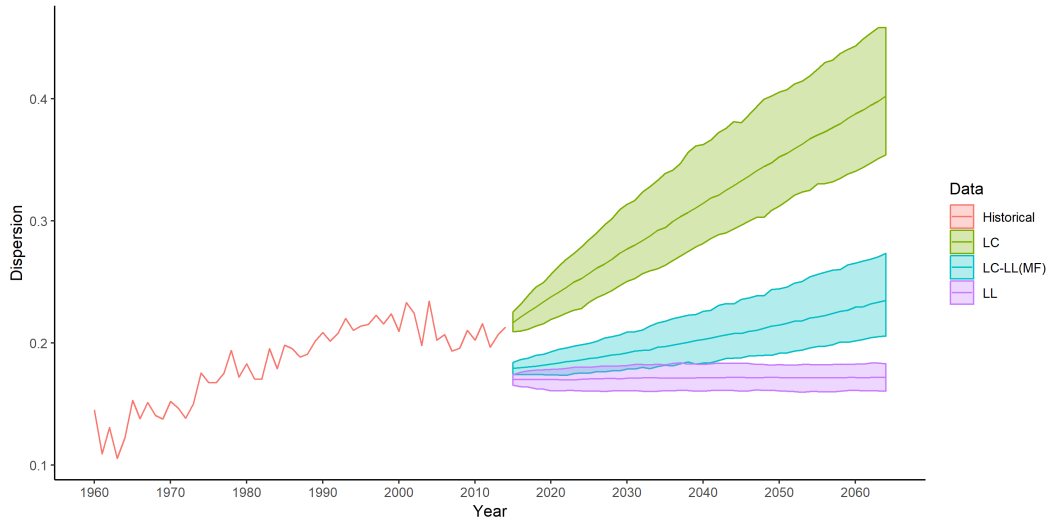


Figure 2.4 – Dispersion at age 85 in the western European populations: historical data and median projections by LC, LL and LC–LL(MF) models with the corresponding confidence intervals at 5% and 95%.

In the LC–LL(MF) case, we note that the dispersion level is comparable to the LL’s projection during the first years, however both the mean dispersion and the width of the confidence intervals are increasing with the projection horizon like in the LC model. This increase is yet significantly slower than the one obtained with the single population framework. While the latter causes a doubling of the dispersion within 50 years, the LC–LL(MF) leads to a dispersion spectrum relatively close to the historical observations of this measure since the 1990’s.

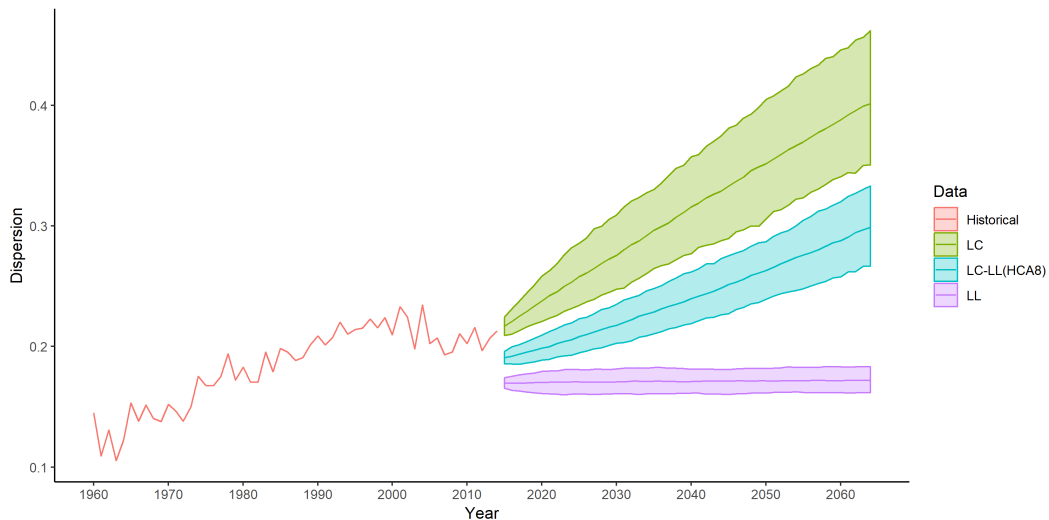


Figure 2.5 – Dispersion at age 85 in the western European populations: historical data and median projections by LC, LL and LC–LL(HCA8) models with the corresponding confidence intervals at 5% and 95%.

When focusing on the LC–LL(HCA8), we observe that the dispersion dynamic is similar to the LC: both of the models leads to a significant augmentation of the dispersion compared to the historical data. However, in the multipopulation model, the increases of the main trend and confidence intervals are more limited. In term of longevity risk, it leads to a significant lower diversification level among the populations.

2.5.2 Solvency Capital Requirement evaluation

We now quantify the impact that the different models could have for a global (re)insurance company in her solvency capital requirement evaluation of the longevity risk. To do so, we analyze a simplified situation where the liabilities are pensions to be paid between ages 60 to 90 for each of the 32 populations. Moreover, we focus only on the cohort aged 59 in 2014 and assume that at the date of evaluation, i.e. end of 2014, the number of pensioners is the same in every population. We retain an annual pension of 100 paid at the end of the year, and a discount rate of 1%.

Hence, to evaluate the corresponding provision the actuary needs to forecast the mortality rates $m_{59+t, 2014+t}^{(i)}$ for $(i, t) \in \mathcal{I} \times \{1, \dots, 32\}$. Following the Solvency II framework, the associated longevity risk solvency capital requirement (SCR) is obtained by estimating the Value at Risk (VaR) at a 99.5% level. This is done by simulating 2,000 longevity scenarios through the stochastic mortality models. We display in the Table 2.2 the results of such approach for the LC, LL, LC–LL(MF) and LC–LL(HCA8) models.

Table 2.2 – Provisions and Solvency capital requirement estimates.

| | LC | LL | LC–LL(MF) | LC–LL(HCA8) |
|-------------------|--------|--------|-----------|-------------|
| Provisions (mean) | 67,513 | 67,577 | 67,675 | 67,485 |
| VaR 99.5 % | 68,059 | 69,591 | 68,999 | 68,954 |
| SCR (VaR - mean) | 547 | 2,014 | 1,323 | 1,469 |

First, we remark that the choice of the model, among the collection we explore in this paper, does not impact much the provision. The difference between the highest provision estimate and the lowest one is of 190, which represents less than 0.3% of the latter. On the contrary, the range of SCR is significantly wider. As expected, the LC leads to the lowest value because the mortality scenarios are independent from one population to another, thus creating diversification. In comparison, the SCR estimated by the LL is nearly 3.7 times bigger. Indeed, in this case the mortality rate dynamics of all populations follow the same trend, thereby implying a concentration risk on this trend. Finally, between these two extreme modellings, our locally coherent approach leads to intermediate levels of SCR. For the coherent groups determined in Section 2.4, we retrieve approximately the average between the LC and the LL.

2.6 Toward a dynamic local coherence

In our model introduced in Section 2.3.1, we have supposed that the classification function $\phi : \mathcal{I} \rightarrow \mathcal{J}$ is constant through time. In other words, a population indefinitely belongs to the same group of coherence. However, in a more general longevity risk assessment framework, it could be interesting to study scenarios in which populations are allowed to switch from one coherence group to another. One of the motivations for such sensitivities is driven by the evaluation of basis risk in some longevity risk hedges. In this Section, the main interest is to describe the stochastic forecasts of the mortality rates, we do not tackle the statistical or expert-based fitting methods for the extended model.

2.6.1 Trend switching model

Hereafter, the groups of local coherence $j \in \mathcal{J}$ are rather to be considered as dominant mortality trends which lead the mortality dynamics of their related specific populations. Thus, as long as a group of populations are lead by the same dominant trend, they are coherent between them. On the contrary to the model of Section 2.3.1, \mathcal{J} is not necessarily a partition of the retained collection of populations \mathcal{I} anymore. However it can always be viewed as a partition of a larger collection Ω , which is not fully observed. Thus, depending of the time period, some of the dominant trends $j \in \mathcal{J}$ may not be linked to any retained population $i \in \mathcal{I}$. During such periods, they can be viewed as dominant trends exogenous from the retained collection of populations. In the same way, we can suppose that $\forall j \in \mathcal{J}, \#j > 1$, even if it implies assuming the existence of populations in Ω not present in the retained collection of populations \mathcal{I} .

Thus, for $(i, x, t) \in \mathcal{I} \times \mathcal{X} \times \mathcal{T}$, we propose the following dynamics for the central mortality rates:

$$\ln m_{x,t}^{(i)} = \alpha_x^{(i)} + B_x^{\phi_t(i)} K_t^{\phi_t(i)} + \beta_x^{(i)} \kappa_t^{(i)} + \text{ad}_{x,t}^{(i)}, \quad (2.14)$$

where $\phi_t : \mathcal{I} \rightarrow \mathcal{J}$ is a set of classification functions, $\alpha_x^{(i)}$ are specific population mortality levels, (B_x^j, K_t^j) define the dominant mortality trends, $(\beta_x^{(i)}, \kappa_t^{(i)})$ represent specific population mortality dynamics, $\text{ad}_{x,t}^{(i)}$ are adjustment mortality levels. Considering the time series, we keep the dynamics of the previous model, i.e. $\Delta \mathbf{K}_t$ follows a VAR(p) model (see Equation 2.12) and $\kappa_t^{(i)}$ are driven by AR(1) models.

The fact that the classification function ϕ_t is now dynamic, it may happen that the switch from one dominant trend to another creates a significant leap in the mortality dynamics. Indeed, for two different dominant trends (j_1, j_2) and a couple age-period (x, t) , we do not impose the term $B_x^{j_2} K_t^{j_2} - B_x^{j_1} K_t^{j_1}$ to be limited since it describes the difference between two non-coherent

mortality dynamics. Thus, to avoid such jumps in the mortality level each time a population changes of dominant trend, we need to add adjustment mortality levels:

$$\text{ad}_{x,t}^{(i)} = \sum_{s=t_0+1}^t B_x^{\phi_{s-1}^{(i)}} K_s^{\phi_{s-1}^{(i)}} - B_x^{\phi_s^{(i)}} K_s^{\phi_s^{(i)}}, \quad (2.15)$$

where t_0 is a time such as $\text{ad}_{x,t_0}^{(i)} = 0$ for all ages $x \in \mathcal{X}$. In a practical point of view, $\phi_{t_0}^{(i)}$ defines the initial dominant trend of the population i . Furthermore, we note that, in the non-switching model case, we have $\forall s \in \mathcal{T}$, $\phi_{s-1}^{(i)} = \phi_s^{(i)} = \phi^{(i)}$, so $\forall t \in \mathcal{T}$, $\text{ad}_{t,x}^{(i)} = 0$. Hence, we retrieve the dynamics of the LC-LL with constant groups of coherence (cf. Equation 2.10).

2.6.2 Kortis bond

As previously stated, this extension is motivated by the assessment of basis risk in longevity risk hedges. We choose to focus on one practical example: the Longevity Divergence Index Value (LDIV), which the Swiss Re Kortis bond is based on (Hunt and Blake, 2015).

Issued in December 2010, the Kortis bond was designed to hedge the basis risk resulting from Swiss Re's "partial natural hedge" in longevity risk. Indeed, the reinsurer was globally exposed to

- mortality risk on the US male population aged between 55 and 65;
- longevity risk on the UK male population aged between 75 and 85.

Even though these two exposures create a natural longevity hedge, a significant basis risk remains between them two, especially if the mortality trends of the corresponding populations diverge. In our model, it would be implied by the fact that the two populations are not linked to the same dominant trend.

The payout of the Kortis bond is based on the level of the LDIV in 2016. To compute this value, we first need to observe the annualized mortality improvement of each population i :

$$\text{Improvement}_n^{(i)}(x, t) = 1 - \left[\frac{m_{x,t}^{(i)}}{m_{x,t-n}^{(i)}} \right]^{\frac{1}{n}}, \quad (2.16)$$

where n is the averaging period of the bond, equals to 8 years in the Kortis bond's case. Then, an averaged improvement index is computed over the age classes exposed to the risk

$$\text{Index}(t, i) = \frac{1}{1 + x_2^{(i)} - x_1^{(i)}} \sum_{x=x_1^{(i)}}^{x_2^{(i)}} \text{Improvement}_n^{(i)}(x, t), \quad (2.17)$$

where for the Kortis bond the age classes are $(x_1^{(\text{UK})}, x_2^{(\text{UK})}) = (75, 85)$ and $(x_1^{(\text{US})}, x_2^{(\text{US})}) = (55, 65)$. Finally, the LDIV at time t is obtained by

$$\text{LDIV}(t) = \text{Index}(t, i_1) - \text{Index}(t, i_2), \quad (2.18)$$

where in the Swiss Re case $(i_1, i_2) = (\text{UK}, \text{US})$ and $t = 2016$.

2.6.3 European LDIV

To illustrate the impact on the basis risk of a switch of dominant trend in our model, we consider an European LDIV. More specifically, we retain:

- the Swiss female population, aged between 75 and 85, for the longevity risk exposure i_1 ;
- the French female population, aged between 55 and 65, for the mortality risk exposure i_2 ;
- a risk period n of 8 years, which ends at year $t = 2024$.

Moreover we suppose that at $t_0 = 2014$, the mortality dynamics of the collection of populations \mathcal{I} follow the model of Section 2.4.2, i.e. the LC-LL(HCA8). Thus, the classification function ϕ_{t_0} can be described thanks to Table 2.1. Finally, we suppose that at a time $T > t_0$, the French female population switches from its original dominant trend to the dominant trend that the Belgian and German female populations are initially linked to. No further switch are considered. Hence, for $t > t_0$, we have:

$$\phi_t(i) = \begin{cases} \phi_{t_0}(i) & \text{if } i \neq \text{FRA F}, \\ \phi_{t_0}(\text{FRA F}) & \text{if } i = \text{FRA F and } t < T, \\ \phi_{t_0}(\text{BEL F}) & \text{if } i = \text{FRA F and } t \geq T. \end{cases} \quad (2.19)$$

Figure 2.6 displays the sensibility of our European LDIV(2024) to the switching time T . To obtain the results, we simulate 10,000 stochastic mortality scenarios thanks to our model for each switching time between 2016 (i.e. the start of the risk period) and 2024 (i.e. the end of the risk period). We then retain the median scenario in term of LDIV(2024). As expected, the sooner the switch occurs, the more significant the Longevity Divergence Index Value is. In particular, we remark that the LDIV is 1.8 times higher when the switch happens in 2016 than in 2024. Indeed, the longer the French and Swiss female populations are linked to different dominant trends, i.e. the longer they are not coherent, the more they diverge.

Thus, our extended model allows the longevity risk manager to evaluate the impact of a wide spectrum of stochastic longevity scenarios on her portfolio and the potential hedges, together with the corresponding basis risks. In a context where basis risk is one the main challenges

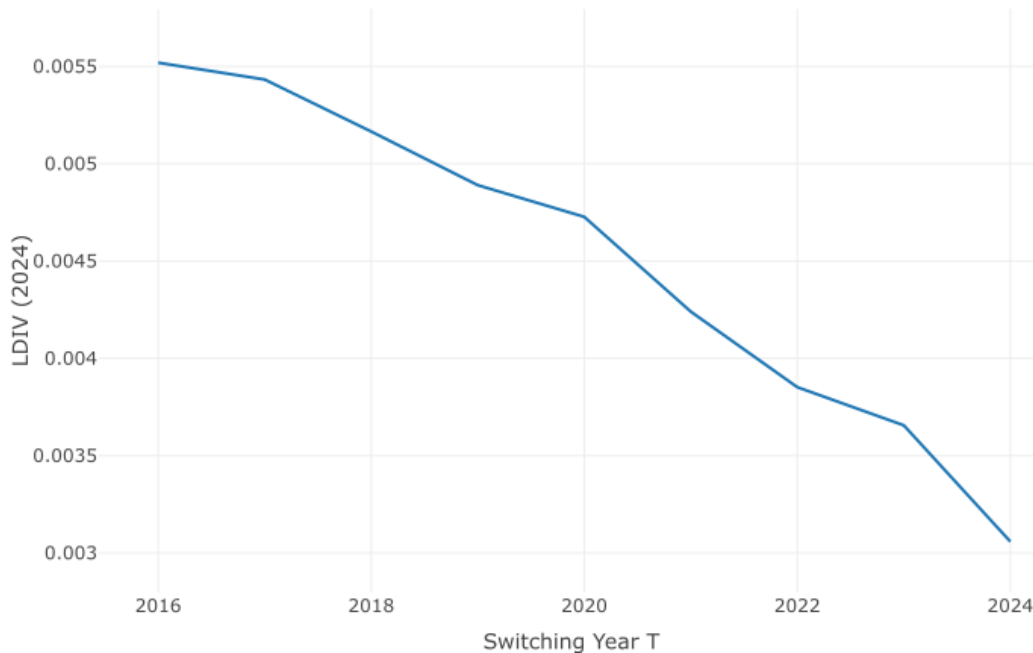


Figure 2.6 – Median European LDIV(2024) according to the switching time of the French female population.

that the longevity risk transfer market is facing (Blake et al., 2018), our model provides an innovative point of view for its assessment.

2.7 Conclusions

In this paper we have proposed a new notion for multipopulation mortality forecasting that we have named local coherence. Indeed, we remark that full independence and "full" coherence approaches both suffer from drawbacks when considering the projections. To highlight this point we have compared the dispersion over 32 European populations of mortality rate forecasts derived from Lee and Carter (1992) and Li and Lee (2005) models. In the first case, the diversification in the stochastic scenarios is overestimated, whereas, in the second method, the concentration is exaggerated.

To overcome this issue, our locally coherent mortality forecasting model allows to cluster populations by coherence groups. Within a specific group, the projections are coherent. However, the long-term relationships between inter-group populations are only modelled through a VAR process without any coherence constraints. Thus, it offers a trade-off between the two boarder coherence cases, which are included in our model. The practitioner has then the possibility to simulate a larger spectrum of longevity scenarios to evaluate her risks compared to the usual models cited before.

To assess the behavior of our model, we have compared it with the Lee-Carter and Li-Lee's ones over the set of European countries for some numerical applications. We notably show that the retained coherence hypothesis can have a major impact on the longevity risk solvency capital requirement for a global life insurance company within the Solvency II framework. Thereby, the large freedom of our model allows to evaluate more precisely this crucial risk measure.

Finally, we have extended the locality coherence property of our model to the temporal dimension. By allowing populations to switch from one group of coherence to another, we expand the spectrum of possible stochastic mortality scenarios. This innovative methodology is particularly interesting in a longevity hedges basis risk assessment framework. Thus, we propose a new tool that should be useful for the development of the longevity risk transfer market.

2.8 Appendix

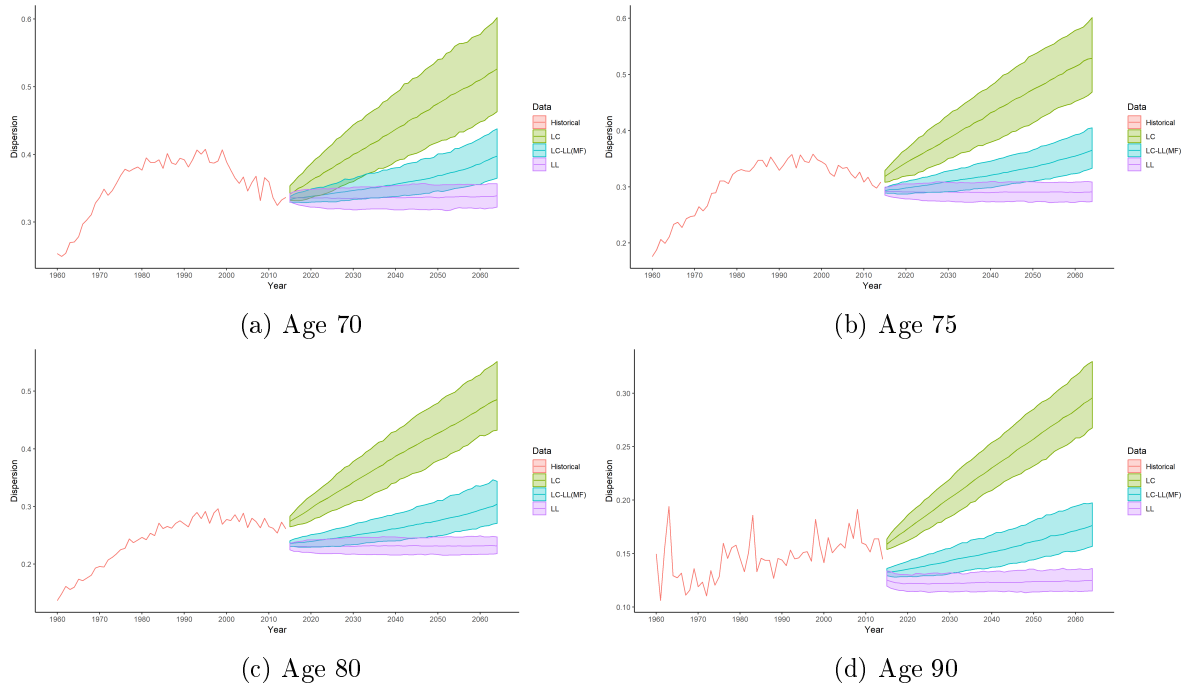


Figure 2.7 – Dispersion in the western European populations: historical data and median projections by LC, LL and LC-LL(MF) models with the corresponding confidence intervals at 5% and 95%.

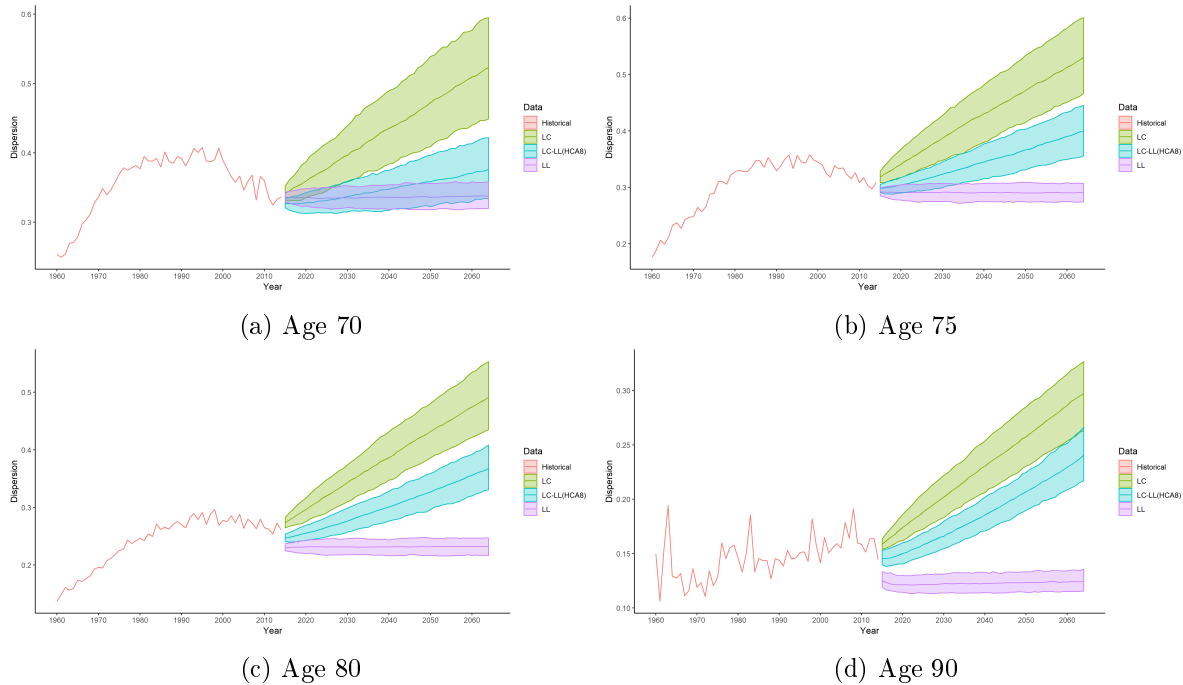


Figure 2.8 – Dispersion in the western European populations: historical data and median projections by LC, LL and LC–LL(HCA8) models with the corresponding confidence intervals at 5% and 95%.

Bibliography

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, August 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1315.

Marie-Pier Bergeron-Boucher, Violetta Simonacci, Jim Oeppen, and Michele Gallo. Coherent Modeling and Forecasting of Mortality Patterns for Subpopulations Using Multiway Analysis of Compositions: An Application to Canadian Provinces and Territories. *North American Actuarial Journal*, 22(1):92–118, January 2018. ISSN 1092-0277. doi: 10.1080/10920277.2017.1377620.

D. Blake, A. J. G. Cairns, K. Dowd, and A. R. Kessler. Still living with mortality: the longevity risk transfer market after one decade. *British Actuarial Journal*, 24, 2018. ISSN 1357-3217, 2044-0456. doi: 10.1017/S1357321718000314.

Andrew J. G. Cairns, David Blake, and Kevin Dowd. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, 73(4):687–718, December 2006. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2006.00195.x.

Andrew J. G. Cairns, Malene Kallestrup-Lamb, Carsten Rosenskjold, David Blake, and Kevin

- Dowd. Modelling Socio-Economic Differences in Mortality Using a New Affluence Index. *ASTIN Bulletin: The Journal of the IAA*, pages 1–36, 2019. ISSN 0515-0361, 1783-1350. doi: 10.1017/asb.2019.14.
- Ivan Luciano Danesi, Steven Haberman, and Pietro Millossovich. Forecasting mortality in subpopulations using Lee–Carter type models: A comparison. *Insurance: Mathematics and Economics*, 62:151–161, May 2015. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2015.03.010.
- Kevin Dowd, Andrew J. G. Cairns, David Blake, Guy D. Coughlan, and Marwa Khalaf-Allah. A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, 15(2):334–356, 2011. ISSN 1092-0277. doi: 10.1080/10920277.2011.10597624.
- Vasil Enchev, Torsten Kleinow, and Andrew J. G. Cairns. Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342, January 2016. ISSN 0346-1238. doi: 10.1080/03461238.2015.1133450.
- Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22, 2010. ISSN 1548-7660.
- Yoel Furman. VAR Estimation with the Adaptive Elastic Net. SSRN Scholarly Paper ID 2456510, Social Science Research Network, Rochester, NY, June 2014.
- Deborah Gefang. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *International Journal of Forecasting*, 30(1):1–11, January 2014. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2013.04.004.
- Quentin Guibert, Olivier Lopez, and Pierrick Piette. Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics*, 88:255–272, September 2019. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2019.07.004.
- HMD. *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 2019-01-28)., 2019.
- Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, February 1970. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634.
- Andrew Hunt and David Blake. Modelling longevity bonds: Analysing the Swiss Re Kortis bond. *Insurance: Mathematics and Economics*, 63:12–29, July 2015. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2015.03.017.
- Torsten Kleinow. A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63:147–152, 2015. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2015.03.023.

- Ronald D. Lee and Lawrence R. Carter. Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–671, September 1992. ISSN 0162-1459. doi: 10.1080/01621459.1992.10475265.
- Hong Li, Yang Lu, and Pintao Lyu. Coherent Mortality Forecasting for Less Developed Countries. *SSRN Scholarly Paper*, July 2018.
- Johnny Siu-Hang Li, Wai-Sum Chan, and Rui Zhou. Semicoherent Multipopulation Mortality Modeling: The Impact on Longevity Risk Securitization. *Journal of Risk and Insurance*, 84(3):1025–1065, 2017. ISSN 1539-6975. doi: 10.1111/jori.12135.
- Nan Li and Ronald Lee. Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, 42(3):575–594, August 2005. ISSN 1533-7790. doi: 10.1353/dem.2005.0021.
- Ronald Richman and Mario V. Wuthrich. A Neural Network Extension of the Lee-Carter Model to Multiple Populations. SSRN Scholarly Paper ID 3270877, Social Science Research Network, Rochester, NY, October 2018.
- Song Song and Peter J. Bickel. Large Vector Auto Regressions. *arXiv:1106.3915 [q-fin, stat]*, June 2011. arXiv: 1106.3915.
- Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x.
- Simone Vazzoler, Lorenzo Frattarolo, and Monica Billio. sparsevar: A Package for Sparse VAR/VECM Estimation. Technical report, 2016.
- Andrés M. Villegas, Steven Haberman, Vladimir K. Kaishev, and Pietro Millosovich. A Comparative Study of Two-Population Models for the Assessment of Basis Risk in Longevity Hedges. *ASTIN Bulletin: The Journal of the IAA*, 47(3):631–679, September 2017. ISSN 0515-0361, 1783-1350. doi: 10.1017/asb.2017.18.
- Rui Zhou, Yujiao Wang, Kai Kaufhold, Johnny Siu-Hang Li, and Ken Seng Tan. Modeling Period Effects in Multi-Population Mortality Models: Applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167, January 2014. ISSN 1092-0277. doi: 10.1080/10920277.2013.872553.
- Rui Zhou, Guangyu Xing, and Min Ji. Changes of Relation in Multi-Population Mortality Dependence: An Application of Threshold VECM. *Risks*, 7(1):14, March 2019. doi: 10.3390/risks7010014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x.

Chapitre 3

Apprentissage statistique et mesures économiques pour la gestion du risque de rachat

Ce chapitre reprend l'article "Applying economic measures to lapse risk management with machine learning approaches", co-écrit avec Stéphane Loisel et Cheng-Hsien Jason Tsai.

Abstract

Modeling policyholders lapse behaviors is important to a life insurer since lapses affect pricing, reserving, profitability, liquidity, risk management, as well as the solvency of the insurer. Lapse risk is indeed the most significant life underwriting risk according to European Insurance and Occupational Pensions Authority's Quantitative Impact Study QIS5. In this paper, we introduce two advanced machine learning algorithms for lapse modeling. Then we evaluate the performance of different algorithms by means of classical statistical accuracy and profitability measure. Moreover, we adopt an innovative point of view on the lapse prediction problem that comes from churn management. We transform the classification problem into a regression question and then perform optimization, which is new for lapse risk management. We apply different algorithms to a large real-world insurance dataset. Our results show that XGBoost and SVM outperform CART and logistic regression, especially in terms of the economic validation metric. The optimization after transformation brings out significant and consistent increases in economic gains.

Keywords: lapse; machine learning; SVM; XGBoost; life insurance; loss function.

3.1 Introduction

Lapse risk is the most significant risk associated with life insurance when compared with longevity risk, expenses risk, and catastrophe risk. Policyholders of life insurance may choose to surrender their policies at any time for cash values, or opt to stop paying premiums and leave policies to become invalid eventually. Lapses have significant impacts on the profitability, or even on the solvency, of a life insurer as many studies demonstrate. They may reduce expected profits (Hwang and Tsai, 2018), cause underwriting expenses unrecovered (Tsai et al., 2009; Pinquet et al., 2011), impair the effectiveness of an insurer’s asset-liability management (Kim, 2005c; Eling and Kochanski, 2013) and bring in liquidity threats as experienced by US life insurers in the late 1980s.

When lapses vary with interest rates as suggested by Dar and Dodds (1989); Kuo et al. (2003); Kim (2005a,b); Cox and Lin (2006), they become even more detrimental to life insurers (Tsai et al., 2009). Many papers argue that the option to surrender a policy for the cash value might account for a large proportion of the policy value, e.g., Albizzati and Geman (1994); Grosen and Løchte Jørgensen (2000); Bacinello (2003); Bauer et al. (2006); Gatzert and Schmeiser (2008); Consiglio and Giovanni (2010). The above reasoning and finding may be the reasons why the fifth Quantitative Impact Study (QIS5), conducted by the European Insurance and Occupational Pensions Authority (EIOPA) in 2011 regarding the implementation of Solvency II, reports that lapse risk accounts for about 50% of the life underwriting risks.

The significance of lapse risk draws attentions of scholars to study what causes policyholders to lapse their policies. We may classify the literature into being macro- or micro-oriented. Macro-oriented papers (e.g., Dar and Dodds, 1989; Kuo et al., 2003; Kim, 2005a,b; Cox and Lin, 2006) focus on how lapse rates (the proportion of lapsed policies to the total number of sampled policies within a period of time) are affected by environmental variables such as interest rates, unemployment rates, gross domestic product, and returns in capital markets, as well as by company characteristics like size and organizational form.

Micro-oriented papers secure data from insurers on individual policies to investigate the determinants of the lapse propensities/tendencies. The identified determinants include the characteristics of policyholders and the features of life insurance products/policies (see Renshaw and Haberman, 1986; Kagraoka, 2005; Cerchiara et al., 2008; Milhaud et al., 2011; Pinquet et al., 2011; Eling and Kiesenbauer, 2014, among others) . Eling and Kochanski (2013); Campbell et al. (2014) provide extensive reviews of the literature on lapses¹.

This paper extends the micro-oriented line of literature in two ways. Firstly, we introduce

1. There are some papers on the subject of modeling early terminations that do not fit our macro-micro classification on empirical, explanatory studies. They impose specific assumptions on the transition probabilities to early terminations (Buchardt et al., 2015), the early terminations’ intensity (Barsotti et al., 2016), or the early termination rates (Loisel and Milhaud, 2011; Milhaud, 2013).

machine learning algorithms including Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) to lapse behavior modeling. These two advanced algorithms have their merits over other approaches used in the literature such as generalized linear models (i.e., binomial and Poisson models and logistic regression), Classification and Regression Tree (CART) analysis, and the proportional hazards model. Secondly, we adopt economic measures in addition to statistical accuracy in evaluating the performance of different algorithms. Such an adoption better demonstrates how different algorithms may benefit the insurer.

Thirdly, we transform the optimization objective from classification accuracy to economic gains to demonstrate the benefit of integrating modeling with profit maximization. Such an integration can increase life insurers' profitability, improve insurers' customer management through taking preventive measures to reduce lapses, and retain more of the so-called Contractual Service Margin (CSM) in International Financial Reporting Standard (IFRS) 17. It also links us to the literature on churn management and its impact on the customer lifetime value (e.g., [Neslin et al., 2006](#); [Lemmens and Croux, 2006](#); [Lemmens and Gupta, 2017](#)).

The results from applying different algorithms to a large dataset consisting of more than six hundred thousand life insurance policies show that XGBoost and SVM outperform CART and logistic regression with respect to statistic accuracy. The results further show that XGBoost is the most robust across training samples.

The advantages of XGBoost and SVM are more apparent with respect to retention gains. The retention gain takes into account the costs of providing incentives to policyholders to reduce their propensities towards lapses, the benefits of retaining policies, and the costs of false alarms. XGBoost and SVM generate much higher retention gains than logistic regression and CART do.

Last but not least, we confirm that economic gains can be further enhanced when the optimization is done on a function linked to the gains rather than on statistic accuracies. The resulted retention gains are 126% of those from applying XGBoost to pursue classification accuracies, and the increase in retention gains remains to be significant under an alternative policyholder retention scheme. An insurer, therefore, should apply robust machine learning algorithms like XGBoost to its economic objective to achieve optimal lapse management.

The organization of the paper is as follows. Section 3.2 contains explanations about XGBoost and SVM, followed by brief descriptions on CART and logistic regression. In Section 3.3 we delineate two performance metrics to be used. One is the commonly seen accuracy, i.e., a statistical validation metric, while the other one is an economic metric considering the expected profits and costs of lapse management. We describe the data obtained from a medium-sized life insurer in Section 3.4. Section 3.5 displays the comparison results across the four algorithms in terms of the statistical and economic metrics. We explain how to integrate algorithms with the profit maximization goal at the beginning of Section 3.6, and then compare the results

from optimizing profit objectives with those from optimization statistic accuracy. Section 3.7 summarizes and concludes the paper.

3.2 Binary classification algorithms

The problem that we want to tackle is detecting whether a policyholder will lapse her/his policy or not, i.e., $y_i \in \{0, 1\}$. Popular predictive models include logistic regression and CART models. More advanced machine learning models that we introduce in this paper are SVM and XGBoost.

3.2.1 XGBoost

XGBoost is an extension of the gradient boosting introduced by Friedman (2001). The gradient boosting tree is an ensemble method, i.e., multiple weak learners h are combined to become a strong learner F in order to achieve a better predictive performance. The following descriptions are summarized from Friedman (2002)).

Given a training sample $\{y_i, \mathbf{x}_i\}_1^N$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$, one would like to find a strong learner $F^*(\mathbf{x})$ which minimizes a loss function $\Psi(y, F(\mathbf{x}))$:

$$F^*(\mathbf{x}) = \arg \min_{F(\mathbf{x})} E_{y, \mathbf{x}}[\Psi(y, F(\mathbf{x}))]. \quad (3.1)$$

The strong learner is an additive expansion of weak learners $h(\mathbf{x}, \{R_{lm}\}_1^L, \bar{y}_{lm})$ that will be a L -terminal node regression tree in our case:

$$\begin{aligned} F_M(\mathbf{x}) &= \sum_{m=0}^M \beta_m h(\mathbf{x}, \{R_{lm}\}_1^L, \bar{y}_{lm}) \\ &= \sum_{m=0}^M \sum_{l=1}^L \beta_m \bar{y}_{lm} \mathbb{1}(\mathbf{x} \in R_{lm}), \end{aligned} \quad (3.2)$$

where $\{R_{lm}\}_1^L$ and \bar{y}_{lm} are the L -disjoint regions and the corresponding split points determined by the m th regression tree, respectively, and β_m are the expansion coefficients. This strong learner is estimated through a stage-wise method that begins with an initial guess $F_0(\mathbf{x})$. Then the pseudo-residuals for $m = 1, 2, \dots, M$ are computed:

$$\tilde{y}_{im} = - \left[\frac{\delta \Psi(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \quad (3.3)$$

The regions $\{R_{lm}\}_1^L$ are obtained by estimating the m th L -terminal node regression tree on the sample $\{\tilde{y}_i, \mathbf{x}_i\}_1^N$. The product $\beta_m \bar{y}_{lm} = \gamma_{lm}$ is set to optimize the loss function Ψ :

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \quad (3.4)$$

At the final stage, the strong learner is updated,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} \mathbb{1}(\mathbf{x} \in R_{lm}), \quad (3.5)$$

where $\nu \in (0, 1]$ is a shrinkage parameter that controls how much information is used from the new tree.

The gradient boosting tree method may be summarized as the following algorithm extracted from [Friedman \(2002\)](#).

| Algorithm 1: Gradient_TreeBoost | |
|--|--|
| 1 | $F_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^N \Psi(y_i, \gamma)$ |
| 2 | For $m = 1$ to M do: |
| 3 | $\tilde{y}_{im} = - \left[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N$ |
| 4 | $\{R_{lm}\}_1^L = L - \text{terminal node tree}(\{\tilde{y}_{im}, \mathbf{x}_i\}_1^N)$ |
| 5 | $\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$ |
| 6 | $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} \mathbb{1}(\mathbf{x} \in R_{lm})$ |
| 7 | endFor |

Figure 3.1 – Pseudocode of the Gradient Tree Boosting algorithm.

Source : [Friedman \(2002\)](#)

Inspired by previous general works on statistical learning, many extensions to the gradient boosting tree method have been developed. The stochastic gradient boosting technique ([Friedman, 2002](#)) is based on the same principle as the bagging technique ([Breiman, 1996](#)). It introduces randomness in the observation: given a random permutation π of the integers $\{1, \dots, N\}$ and $\tilde{N} < N$, the new weak learner tree is estimated on the random subsample $\{\tilde{y}_{\pi(i)m}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}}$. Another way to inject randomness that has been popularized by [Breiman \(2001\)](#) is randomly selecting a subspace of the explanatory variables. More specifically, given a random permutation π^* of integers $\{1, \dots, n\}$ and $\tilde{n} < n$, the new weak learner tree is estimated on $\{\tilde{y}_{im}, P^*(\mathbf{x})_i\}_1^N$ in which $P^*(\mathbf{x}) = \{x_{\pi^*(1)}, \dots, x_{\pi^*(\tilde{n})}\}$.

To avoid overfitting, some extensions follow the general idea of the ridge regression ([Hoerl and Kennard, 1970](#)) and lasso regression ([Tibshirani, 1996](#)) and adopt the penalized optimization

point of view. Instead of optimizing a loss function $\Psi(y, F(\mathbf{x}))$, the problem is modified as the optimization on an “objective” function \mathcal{O} that is the sum of a loss function Ψ and a regularization term Ω :

$$\mathcal{O}(y, F(\mathbf{x})) = \Psi(y, F(\mathbf{x})) + \Omega(F). \quad (3.6)$$

Among all the boosting packages that have been developed, the XGBoost system (Chen and Guestrin, 2016) has become the most popular due to its flexibility and computing performances. It has also become the most popular machine learning algorithm in data science challenges such as Kaggle for structured data. We list the main parameters that need to be tuned, using the package’s terminology and the notation of (Friedman, 2002), as follows.

1. *nrounds* is the number of trees to grow: M ;
2. *eta* is the shrinkage parameter: $M\nu$;
3. *gamma* is the regularization parameter which is used in Ω ;
4. *max_depth* is the number of nodes of a tree: L ;
5. *min_child_weight* is the minimal number of observations in a node and $\min_{l,m} \sum_{i=1}^N \mathbb{1}(\mathbf{x}_i \in R_{lm})$ should be higher than this value;
6. *subsample* is the relative size of the random subsample used in the case of a stochastic gradient boosting: \tilde{N}/N ;
7. *colsample_bytree* is the relative size of the random subspace of explanatory variables selected at each new tree: \tilde{n}/n .

Since we are interested in a binary classification in this paper, we use the logistic loss function:

$$\Psi(y, \hat{y}) = \sum_{i=1}^N \left[y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i}) \right], \quad (3.7)$$

and the error function as the metric for cross-validation:

$$error(y, \hat{y}) = \frac{\sum_{i=1}^N \mathbb{1}(y_i \neq round(\hat{y}_i))}{N}, \quad (3.8)$$

where $round(\hat{y}_i) = \begin{cases} 1 & \text{if } \hat{y}_i \geq 0.5, \\ 0 & \text{if } \hat{y}_i < 0.5. \end{cases}$

The tuning method that we adopt consists of two nested cross-validations. We first perform a grid search on the parameters except *nrounds* with a 2-folds cross-validation (the grid of values is reported in Appendix 3.8.1). Then we determine the best *nrounds* through a 5-folds cross-validation up to 200 for every possible set of parameters in the grid.

3.2.2 SVM

The theory of SVM was introduced in the 1990's by [Boser et al. \(1992\)](#); [Cortes and Vapnik \(1995\)](#). It has become a popular algorithm for classification problems and for churn prediction in particular (e.g., [Zhao et al., 2005](#); [Xia and Jin, 2008](#)) Its predictive power is rather good compared to other classification algorithms (e.g., [Vafeiadis et al., 2015](#); [Wainer, 2016](#)).

The SVM algorithm can be described by geometrical terms. The main idea is to find a hyperplane that separates the observation space into two homogeneous subspaces that is as far apart from each other as possible. This solution is defined as the maximum-margin hyperplane. To deal with misclassifications, a soft margin (i.e., a penalty determined by the user) is imposed upon the SVM. Another way to deal with classification errors is to project the data to a higher-dimensional space through a kernel function. A more complete geometrical description of SVM can be found in [Noble \(2006\)](#).

In the following, we adopt a formula-based description of the SVM by using the notation of [Hsu et al. \(2003\)](#). Given a training sample $\{y_i, \mathbf{x}_i\}_1^N$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{+1, -1\}$, the SVM algorithm is the solution of the following optimization problem:

$$\min_{\omega, b, \xi} \frac{1}{2} \omega^\top \omega + C \sum_{i=1}^N \xi_i, \quad (3.9)$$

with the constraints

$$y_i (\omega^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0. \quad (3.10)$$

The separating hyperplane is determined by the orthogonal vector ω and constant b . The soft margin penalty cost is denoted as C . The data may be projected to a higher dimension space by the function ϕ , and the underlying kernel function is defined by $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$.

In our case we choose to consider the radial basis function kernel (also called RBF kernel) that is the most commonly used in practice and determined by

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \quad (3.11)$$

with $\gamma > 0$ being the kernel parameter.

Then we use the `e1071` R package ([Meyer et al., 2015](#)) to implement the SVM algorithm. To tune the SVM parameters (C, γ) , we perform a grid search on a 2-folds cross-validation and adopt the misclassification error function as the validation metric. The grid of values is reported in [Appendix 3.8.2](#).

3.2.3 CART

CART was first introduced by [Breiman et al. \(1984\)](#). The underlying idea is straight forward: defining a class by following a list of decision rules on the explanatory variables. To determine these rules, the data space is iteratively separated by binary split into two disjointed subspaces. At each step or node of this top-down construction, the explanatory variable and the dividing point are chosen to minimize the Gini impurity of the node.

More specifically, given a node l of N_l observations of response $y_i \in \{0, 1\}$ with $i \in l$, the proportion of observations in the node is defined by $p_l = \frac{1}{N_l} \sum_{i \in l} y_i$. Then use an algorithm to partition the parent node into two nodes l_L and l_R by maximizing

$$I_G(l) - [I_G(l_L) + I_G(l_R)], \quad (3.12)$$

where I_G is the Gini impurity of the node and computed by

$$I_G(l) = N_l p_l (1 - p_l). \quad (3.13)$$

This construction is applied up to obtaining a node for every observation point. The tree obtained is thus designated as the saturated model. Although fitting the response on the training sample perfectly, it generally leads to low predictive performance when applied to new samples. Hence the tree needs to be pruned, i.e., the number of final nodes needs to be reduced to increase its predictive power.

Many criteria can be used to prune the tree, e.g., the minimum number of observations in a final node. We choose L , the number of terminal nodes, that minimizes the misclassification error:

$$error(y, \hat{y}) = \frac{\sum_{i=1}^N \mathbb{1}(y_i \neq \hat{y}_i)}{N}. \quad (3.14)$$

L is estimated by a 10-folds cross-validation methodology. We use the *rpart* R package ([Therneau et al., 2018](#)) to implement CART.

3.2.4 Logistic Regression

The logistic regression is a special case of the generalized linear models ([Nelder and Wedderburn, 1972](#)) obtained with the Bernoulli distribution. The goal is to model the probability of a binary event such as the lapse probability p_i of the policyholder i . Given a training sample $\{y_i, \mathbf{x}_i\}_1^N$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$, the regression model is specified as:

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}. \quad (3.15)$$

The parameters $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^n$ can be estimated by the maximum-likelihood method:

$$\mathcal{L} = \prod_{i=1}^N \left(\frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)^{1-y_i}. \quad (3.16)$$

When applying the estimated logistic regression model to a classification problem, it doesn't directly lead to labeled responses but to estimated probabilities. To determine the forecasted class, we chose the common threshold of 0.5, i.e.,

$$\hat{y}_i^* = \begin{cases} 1 & \text{if } \hat{y}_i \geq 0.5, \\ 0 & \text{if } \hat{y}_i < 0.5. \end{cases} \quad (3.17)$$

3.3 Validation metrics

For each policy, the observed lapse y_i and the forecasted lapse \hat{y}_i are binary variables: $(y_i, \hat{y}_i) \in \{0, 1\}^2$. The four different outputs of a binary classification model are named true positive (1, 1), true negative (0, 0), false positive (0, 1) and false negative (1, 0) while the number of each case is usually laid out in the so-called confusion matrix. Denote $N(j, k)$ as the coefficients of the confusion matrix in which $j \in \{0, 1\}$ stands for the observed lapse indicator and $k \in \{0, 1\}$ the predicted lapse indicator. Given a set of response variables $\{y_i, \hat{y}_i\}_1^N$, we estimate $N(j, k)$ as:

$$N(j, k) = \sum_{i=1}^N \mathbf{1}(y_i, \hat{y}_i = k). \quad (3.18)$$

3.3.1 Statistical metric

Based on the confusion matrix, different metrics can be developed. We first focus on the accuracy metric, the ratio of correctly classified predictions over the total number of predictions:

$$\begin{aligned} accuracy(y, \hat{y}) &= \frac{N(1, 1) + N(0, 0)}{N} \\ &= 1 - error(y, \hat{y}). \end{aligned} \quad (3.19)$$

3.3.2 Economic metric

Although we adopt mathematical algorithms to predict lapses, the risk is an economic issue after all. We thus would like to analyze and compare the classification algorithms by an

economic metric. More specifically, we will estimate the impacts of different classification results on the expected profits from policies, also called customer lifetime values. In order to do so, we plan to adopt an economic model inspired by Neslin et al. (2006); Gupta et al. (2006).

Suppose that policy i stays Θ_i years in the portfolio ($\Theta_i \in \mathbb{N}$). The profitability ratio at time t can be represented by $p_{i,t}$ and the face amount by $F_{i,t}$. The lifetime value for policy i is computed as:

$$CLV_i = \sum_{t=0}^{\Theta_i} \frac{p_{i,t} F_{i,t}}{(1 + d_t)^t}, \quad (3.20)$$

where d_t is the discount rate.

Assuming a deterministic time horizon $T \in \mathbb{N}$, we define the $(T + 1)$ -dimensional real vectors \mathbf{p}_i , \mathbf{F}_i , \mathbf{r}_i and \mathbf{d} for profitability ratios, face amounts, retention probabilities, and interest rates respectively. Given the four vectors, the customer lifetime value is

$$CLV_i(\mathbf{p}_i, \mathbf{F}_i, \mathbf{r}_i, \mathbf{d}) = \sum_{t=0}^T \frac{p_{i,t} F_{i,t} r_{i,t}}{(1 + d_t)^t}, \quad (3.21)$$

The lapse management strategy is modelled by the offer of an incentive $\boldsymbol{\delta}_i \in \mathbb{R}^{T+1}$ to policyholder i who is contacted with a cost c . The incentive is accepted with the probability γ_i , and the acceptance will change the vector of the probabilities of staying in the portfolio from \mathbf{r}_i to $\mathbf{r}_i^* \in \mathbb{R}^{T+1}$. We further make the following simplifying assumptions:

1. \mathbf{p}_i are the same for all policies and denoted as \mathbf{p} hereafter;
2. $\boldsymbol{\delta}_i$ are the same for all contacted policies and denoted as $\boldsymbol{\delta}$ hereafter;
3. $p_{i,t}$, $F_{i,t}$ and d_t remain constant across time;
4. \mathbf{r}_i equals to \mathbf{r}_{lapse} or $\mathbf{r}_{stay} = (1, 1, \dots, 1)$ and \mathbf{r}_{lapse} is estimated on the dataset and will be given in Section 3.5.2;
5. if $\mathbf{r}_i = \mathbf{r}_{stay}$, the incentive is accepted with probability $\gamma_i = 1$ and $\mathbf{r}_i^* = \mathbf{r}_{stay}$;
6. if $\mathbf{r}_i = \mathbf{r}_{lapse}$, the incentive is accepted with probability $\gamma_i = \gamma$ and $\mathbf{r}_i^* = \mathbf{r}_{stay}$.² Policyholders who reject the offers (probability = $1 - \gamma$) will lapse their policies, i.e. $\mathbf{r}_i^* = \mathbf{r}_{lapse}$.

2. These simplifications assume that the profitability ratio, the incentive, and the probability to accept the incentive is the same across policyholders, respectively. Upon the availability of data, we may compute an expected profitability ratio for each policy. The incentive offered to each policyholder can then be set as a function of the policy's profitability. The probability of accepting the offer can also be a function of the incentive, but such a function is difficult to estimate in practice. Face amount may be variable for some products, which increases the difficulty in estimating the expected profitability ratio. The retention probabilities may change with time, and this calls for a dynamic model of lapse propensities

The application of a segmentation algorithm to the tested samples produces two confusion matrices: one with respect to number of policies while the other in term of face amount. For the latter matrix, we denote $F(j, k)$ as the coefficients of the matrix with regard to face amount, where j stands for the indicator of the policyholder's lapse in real life, k the indicator by the algorithm's prediction, and $(j, k) \in \{0, 1\}^2$. More specifically,

$$F(j, k) = \sum_{i=1}^N F_i \cdot \mathbf{1}(y_i, \hat{y}_i = k), \quad (3.22)$$

while N is defined in Equation (3.18).

We define the reference portfolio value (RPV) as the customer lifetime value of all policies if no customer relationship management about lapses are carried out to be:

$$RPV = CLV(\mathbf{p}, F(0, 0) + F(0, 1), \mathbf{r}_{stay}, \mathbf{d}) + CLV(\mathbf{p}, F(1, 0) + F(1, 1), \mathbf{r}_{lapse}, \mathbf{d}). \quad (3.23)$$

Given a segmentation algorithm, we compute the lapse managed portfolio value (LMPV) by

$$\begin{aligned} LMPV(\boldsymbol{\delta}, \gamma, c) = & CLV(\mathbf{p}, F(0, 0), \mathbf{r}_{stay}, \mathbf{d}) + CLV(\mathbf{p}, F(1, 0) + (1 - \gamma)F(1, 1), \mathbf{r}_{lapse}, \mathbf{d}) \\ & + CLV(\mathbf{p} - \boldsymbol{\delta}, F(0, 1) + \gamma F(1, 1), \mathbf{r}_{stay}, \mathbf{d}) - c(N(0, 1) + N(1, 1)). \end{aligned} \quad (3.24)$$

Then we define the economic metric of the algorithm as the retention gain:

$$RG(\boldsymbol{\delta}, \gamma, c) = LMPV(\boldsymbol{\delta}, \gamma, c) - RPV, \quad (3.25)$$

that can be simplified as

$$\begin{aligned} RG(\boldsymbol{\delta}, \gamma, c) = & \gamma [CLV(\mathbf{p} - \boldsymbol{\delta}, F(1, 1), \mathbf{r}_{stay}, \mathbf{d}) - CLV(\mathbf{p}, F(1, 1), \mathbf{r}_{lapse}, \mathbf{d})] \\ & + CLV(\boldsymbol{\delta}, F(0, 1), \mathbf{r}_{stay}, \mathbf{d}) - c(N(0, 1) + N(1, 1)). \end{aligned} \quad (3.26)$$

3.4 Data

Our data come from a medium-size life insurance company in Taiwan that had total assets over 15 billion US dollars at the end of 2013. The data contain 629,331 life insurance policies sold during the period from 1998 to 2013. The data-providing insurer tracked changes in the

statuses of policies including death and lapse. The last tracking date is 31/08/2013. 243,152 policies out of all samples were lapsed, and 5,486 insureds died during the sampling period.

We specify several variables based on the literature and the data provided by the insurer as input to the algorithms of Section 3.2. Firstly we are able to identify from the data the age, gender, and occupation of an insured at the time when the policy was issued. Female is designated as 1 while male 0 for the dummy variable Gender. Then we designate the dummy variable Occupation as 1 for the occupations that the insurers in Taiwan would undertake extra screening/underwriting. The data also record whether the insured is required to have a physical examination when purchasing life insurance and how many non-life policies (health and long-term care) a person are listed as the insured (since a person may purchase multiple policies).

The data also contain the inception date and face amount of each policy. There are three types of policies. The most popular type is traditional policies like term life, whole life, and endowment. Investment-linked and interest-adjustable types of products appeared in 2000s. We also able to identify whether a policy is a single-premium one or not. There are three cases with regard to participation. It was not until 2004 that insurers were allowed to sell non-participating policies. The policies sold by the end of 2003 are thus designated as Mandatory Participating. Starting from 2004, policies may be classified into participating and non-participating. Most policies sold in Taiwan are dominated in New Taiwan Dollar (NTD) ; there are some policies dominated in other currencies.

We further set up two nominal variables. Firstly, we categorize distribution channels as Tied Agents (denoted by TA), Direct Marketing (DM), and Banks (BK)³. Secondly, premium paying methods are classified into three ways: collected by the personnel of the insurer (denoted as Insurer), automatic transfers from banks or payments by credit cards (B&C)⁴, and going to the post office or convenient stores in person (P&C).

Table 3.1 and 3.2 present the descriptive statistics of the above explanatory variables. The average age of the sampled insureds is 28 and the standard deviation of the insureds' age is 17. The minimum, medium, and maximum age is 0, 27, and 80, respectively. The samples consist of relatively equivalent portions of male and female insureds. About 20% of the insureds work in riskier occupations that call for extra underwriting. Most insureds (over 96%) were not required to go through physical examination in purchasing life insurance. Many insureds are associated with multiple non-life policies so that the average number of non-life policies a person are listed as the insured is 1.2. There is a person who is listed as the insured for 33 non-life policies.

3. Few policies are also sold by independent agents, brokers that we gather in the same category.

4. Paying premiums by automatic transfers from bank accounts or by recurring payments of credit cards is indifferent to policyholders. We thus regard these two automatic/recurring payment methods as one.

The mean and medium of policy inception dates are in the second quarter of 2005, and the standard deviation around this quarter is almost 5 years. The face amount of the sampled policies has an average of 17,165 US dollars⁵ with big variations: the largest policy reaches 2 million dollars, the smallest one is only 333 dollars⁶, and the standard deviation is about twenty-eight thousand dollars. Around 3% of the samples are single-premium policies. 46.6% of samples are mandatory-participating policies while 37.2% are non-participating ones. Almost all policies are traditional types of products ; interest-adjustable and investment-linked types of products are merely 3% of our samples. 88% of policies are dominated in NTD.

Table 3.1 also shows that selling life insurance through tied agents is the major way (94%) of this insurer while the sampled policies sold through direct marketing are smaller than 3%. It further shows that the most popular way of paying premiums is through automatic/recurring transfers from bank accounts or credit cards (71%). Since post offices and convenient stores providing money transferring services are conveniently around, about 10% of our samples have premiums paid in places like these.

3.5 Result with respect to statistical and economic metrics

Our focus is on the predictive performance of different algorithms. We thus conduct out-of-sample tests using the following procedure. First, we randomly split the dataset D into 10 subsamples $\{D_1, \dots, D_{10}\}$ of equal size and then train an algorithm on D_k , $k \in \{1, \dots, 10\}$. The estimated model is subsequently applied to the other subsamples to obtain forecasts \hat{y} of lapses. In the last step, we compare these predictions with the observed lapses y by the validation metric $\rho(y, \hat{y})$ to measure the predictive performance of the algorithm. This procedure enables us to make sure that every observation is used, at some point of an algorithm, as both training and testing samples. It is similar to the k -fold cross-validation technique in which the training subsample is composed of $D - D_k$ and the testing subsample is set to D_k . We use the k -fold cross-validation to tune parameters in training some of the algorithms.

3.5.1 Results with respect to the statistical metric

The mean accuracy computed using the above cross-validation procedure is displayed in the Table 3.3 and Figure 3.2 for each binary classification algorithm. As expected, the more sophisticated the model is, the more accurate the predictions will be. XGBoost ranks number one, followed by SVM, CART, and logistic regression (LR). XGBoost surpasses logistic regression by 2.24% on average, which represents a significant improvement of 12,684 correctly classified policies. Moreover, the smallest standard deviation of accuracy of the XGBoost,

5. The exchange rate used in the paper is 30 NTD/1 USD.

6. This policy is a whole life insurance with a one-year old insured and the death benefit of ten thousand NTD (a little over three hundred USD). There are other small policies with death benefits smaller than three thousand USD. These policies constitute less than one percent of our samples.

Table 3.1 – Descriptive statistics of categorical explanatory variables.

| Variables | Category | Percentage |
|-----------------------|---------------------------|-------------------|
| Gender | Female | 48 |
| | Male | 52 |
| Occupation | Tier one | 80.5 |
| | Requiring extra screening | 19.5 |
| Physical Examination | Exempted | 96.4 |
| | Required | 3.6 |
| Distribution Channel | TA | 93.9 |
| | BK | 3.4 |
| | DM | 2.4 |
| | Others ⁷ | 0.3 |
| Premium Payment | Single premium | 3.1 |
| | Non single premium | 96.9 |
| Premium Paying Method | Insurer | 18.8 |
| | B&C | 70.8 |
| | P&C | 10.4 |
| Participation | Non-participating | 37.2 |
| | Participating | 16.2 |
| | Mandatory participating | 46.6 |
| Product Type | Interest-adjustable | 1.7 |
| | Investment-linked | 1.2 |
| | Traditional | 97.1 |
| Currency Domination | NTD | 88.1 |
| | Others | 11.9 |

Table 3.2 – Descriptive statistics of continuous explanatory variables.

| | Mean | Medium | St. Dev. | Minimum | Maximum |
|------------------------|-------------|---------------|-----------------|----------------|----------------|
| Age | 28.3 | 27 | 16.8 | 0 | 80 |
| # of non-life policies | 1.2 | 0 | 2 | 0 | 33 |
| Inception date | 06/06/2005 | 21/04/2005 | 4.8 (years) | 01/01/1998 | 31/07/2013 |
| Face Amounts (in USD) | 17,165 | 10,000 | 28,050 | 333 | 2,000,000 |

0.03%, indicates that XGBoost is less prone to sample selection. This is visible in the box plot of Figure 3.2.

Table 3.3 – Cross-validated statistical accuracies.

| | LR | CART | SVM | XGB |
|--------------------|--------|--------|--------|-------|
| Mean Accuracy | 76.64% | 77.15% | 77.82% | 78.8% |
| Standard Deviation | 0.07% | 0.10% | 0.08% | 0.03% |

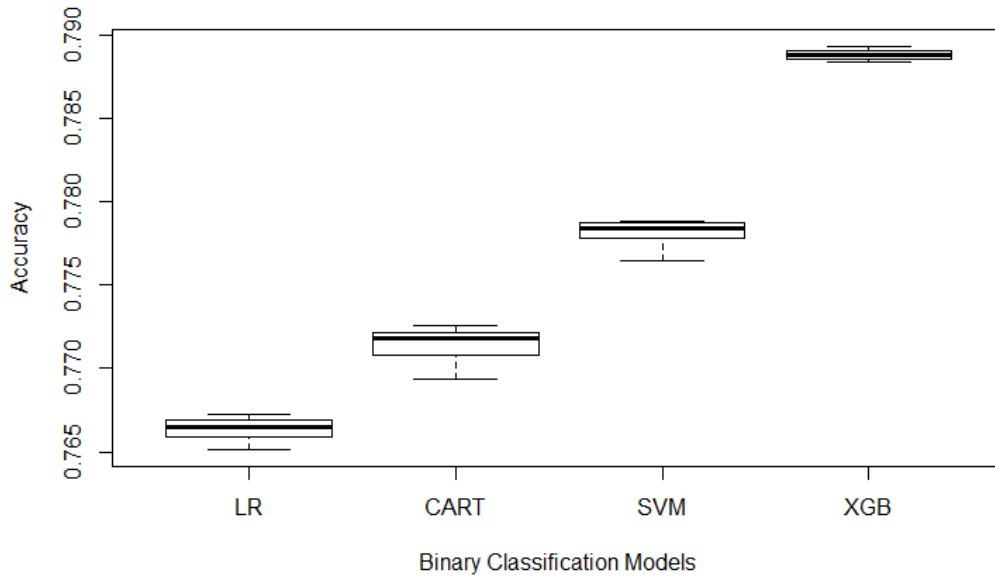


Figure 3.2 – Box plot of statistical accuracies.

Looking at the entire confusion matrices in Tables 3.4 to 3.7, we find that CART predicts the most lapses ($191,869 = 51,241 + 140,628$) from which it identifies the most lapses correctly (140,628) but also signals the most false alarms (51,241). SVM predicts the most stays ($398,597 = 310,258 + 88,339$) in which it identifies the most stays correctly (310,258) while produces many false security cases (88,339). XGBoost is rather robust on the other hand. It is ranked the second in terms of all aspects: correctly identifying lapses (137,660), correctly identifying stays (309,111), not producing false alarms (38,450), and not producing false securities (81,177).

3.5.2 Results with respect to the economic metric

To evaluate the algorithms by the economic metric, we first need to specify the parameters of the cash flows model. Since no data is available for us to estimate these parameters, we have to make assumptions. We had conducted sensitivity analyses and confirmed that the comparison results remain the same in general.

Table 3.4 – Average confusion matrix of XGB.

| | | Predicted | |
|---------------|-------|------------------|---------|
| | | Stay | Lapse |
| Actual | Stay | 309,111 | 38,450 |
| | Lapse | 81,177 | 137,660 |

Table 3.5 – Average confusion matrix of SVM.

| | | Predicted | |
|---------------|-------|------------------|---------|
| | | Stay | Lapse |
| Actual | Stay | 310,258 | 37,303 |
| | Lapse | 88,339 | 130,498 |

Table 3.6 – Average confusion matrix of CART.

| | | Predicted | |
|---------------|-------|------------------|---------|
| | | Stay | Lapse |
| Actual | Stay | 296,320 | 51,241 |
| | Lapse | 78,209 | 140,628 |

Table 3.7 – Average confusion matrix of LR.

| | | Predicted | |
|---------------|-------|------------------|---------|
| | | Stay | Lapse |
| Actual | Stay | 304,025 | 43,537 |
| | Lapse | 88,775 | 130,062 |

The time horizon T is set to 12 years according to the length of the sampling period. We estimate the retention probability vector \mathbf{r}_{lapse} from the dataset and obtain the vector displayed in Table 3.8.

Table 3.8 – Estimated retention probability \mathbf{r}_{lapse} .

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Retention probability | 0.96 | 0.87 | 0.67 | 0.37 | 0.27 | 0.21 | 0.15 | 0.12 | 0.10 | 0.08 | 0.06 | 0.05 | 0.04 |

Other parameters are set as follows:

- the profitability ratio $p = 0.5\%$;
- the discount rate $d = 2\%$;
- the cost to contact a policyholder $c = 10$ USD.

We propose two different incentive strategies: an aggressive one and a moderate one. The incentive vectors are described in Table 3.9

Table 3.9 – Incentive strategies.

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---------------------|---|---|-----|-----|---|---|-----|-----|----|----|----|----|----|
| Incentive 1 (in bp) | 0 | 0 | 3 | 3 | 6 | 6 | 9 | 9 | 12 | 12 | 15 | 15 | 18 |
| Incentive 2 (in bp) | 0 | 0 | 1.5 | 1.5 | 3 | 3 | 4.5 | 4.5 | 6 | 6 | 6 | 6 | 6 |

We further assume that the probabilities of accepting the incentives for a would-lapse policyholder are $\gamma_1 = 20\%$ and $\gamma_2 = 10\%$ respectively.

The results from comparing different classification algorithms by the economic metric with the aggressive incentive strategy are displayed in Table 3.10 and Figure 3.3. The winner looks to be XGBoost: it has the highest retention gain with the smallest standard deviation across subsampling. Figure 3.3 further illustrates that XGBoost and SVM lead to similar retention gain compared to logistic regression and CART.

Notice that the differences across the algorithms are wider in terms of the economic metric than the statistical metric. The accuracies of the models are between 76.64% and 78.88%, which means an improvement ratio of 2.9%. The retention gains, on the other hand, range from 2.7 and 5.2 million USD, indicating an enhancement of 96%. Therefore, choosing a good algorithm is more important in terms of economic reality (dollar amount) than by statistical accuracy. It appears that CART produces the lowest retention gain: \$2,680,012. This is mostly because CART has the highest false alarm rate (cf. Table 3c) which means offering the incentive to many policyholders who have no intention to lapse their policies. Furthermore, CART leads to the highest contacting cost since it predicts the highest lapses. The profits are thus reduced.

Table 3.10 – Cross-validated retention gains with the aggressive strategy.

| | LR | CART | SVM | XGB |
|---------------------|-----------|-----------|-----------|-----------|
| Mean Retention Gain | 4,046,602 | 2,680,012 | 5,028,737 | 5,243,913 |
| Standard Deviation | 133,993 | 209,220 | 139,102 | 115,415 |

Then we look at algorithms' performances when the incentive strategy is moderate and leads to lower acceptance probabilities. The results are displayed in the Table 3.11 and the Figure 3.4. We first notice XGB and SVM remains to be ranked No. 1 and No. 2, respectively. Next we observe that the improvement ratio of the best algorithm over the worst is smaller but remains to be significant (56%). Thirdly, retention gains are significantly lower with the moderate incentive strategy. For instant, XGB achieves a gain of 5.2 million dollars with the aggressive incentive strategy but the gain reduces to 3.3 million dollars when incentives offered to policyholders are moderate. Under our assumptions, the company should rather set the aggressive incentive strategy up to optimize her gains. However, in practice, one would

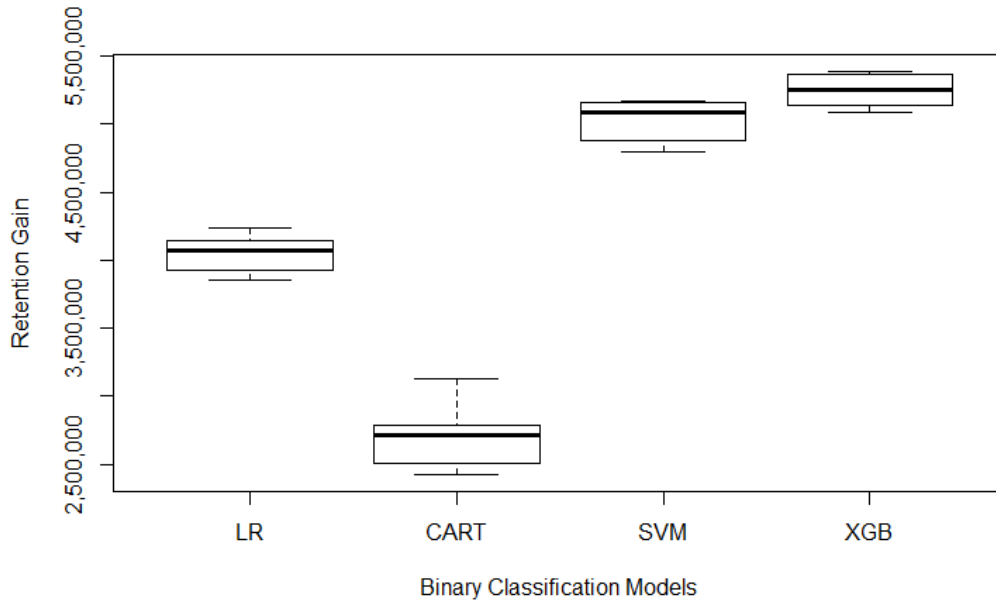


Figure 3.3 – Box plot of retention gains with the aggressive strategy.

need a more complete sensitivity study on the incentive to be offered and the corresponding acceptance probability to fully optimize the lapse management.

Table 3.11 – Cross-validated retention gains with the moderate strategy.

| | LR | CART | SVM | XGB |
|---------------------|-----------|-----------|-----------|-----------|
| Mean Retention Gain | 2,618,396 | 2,085,599 | 3,113,900 | 3,261,029 |
| Standard Deviation | 63,693 | 85,184 | 54,169 | 45,928 |

In summary, XGB and SVM consistently perform better than CART and LR no matter which performance index, statistical accuracy or retention gains with alternative incentive strategies, is used. The drawbacks of XGB and SVM relative to CART and LR that we may think of are not related to performance. For instance, XGB and SVM are less transparent, more complex, demanding more computing power, and more difficult to be comprehended by inexperienced persons than CART and LR.

3.6 Optimization on profitability instead of classification

It is obvious that insurers would not seek to optimize the classification accuracy but focus on economic gains resulted from the classification algorithms when forming a lapse management strategy. When our aim is to maximize the profitability of the lapse management strategy, binary classifications might be unsuitable since they are not designed to meet such a need. [Ascarza et al. \(2018\)](#) emphasize the difference between the at-risk population (e.g., customers

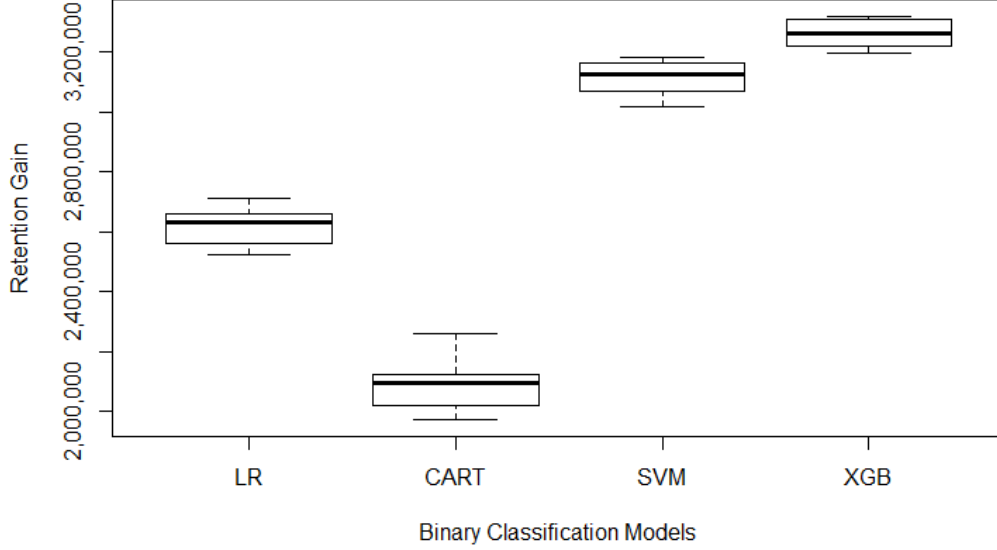


Figure 3.4 – Box plot of retention gains with the moderate strategy.

with high churn probabilities) and the targeted population (e.g., customers that the company should focus her retention campaign on in order to optimize her profits) from an economic point of view. Along this line of churn literature, [Lemmens and Gupta \(2017\)](#) modify the usual loss function into a profit-based function to optimize economic gains. They obtain a significantly increase in the expected profit of a retention campaign. Learning from the churn literature, we transform the above classification problem into a regression question in this section.

3.6.1 Methodology

Let the new response variable $z_i^{R_j}$ represents the retention gain or loss resulting from proposing the incentive $j \in \{1, 2\}$ (cf. Section 3.5.2) to policyholder i . More specifically, we define $z_i^{R_j}$ as

$$z_i^{R_j} = \begin{cases} -CLV(\boldsymbol{\delta}_j, F_i, \mathbf{r}_{stay}, \mathbf{d}) - c & \text{if } y_i = 0, \\ \gamma_j \cdot [CLV(\mathbf{p} - \boldsymbol{\delta}_j, F_i, \mathbf{r}_{stay}, \mathbf{d}) - CLV(\mathbf{p}, F_i, \mathbf{r}_{lapse}, \mathbf{d})] - c & \text{if } y_i = 1. \end{cases} \quad (3.27)$$

Then we may apply the XGBoost algorithm to $\{z_i^{R_j}, \mathbf{x}_i\}_1^N$ and use the mean squared error as the loss function

$$\Psi(z^{R_j}, \hat{z}^{R_j}) = \frac{1}{N} \sum_{i=1}^N [z_i^{R_j} - \hat{z}_i^{R_j}]^2, \quad (3.28)$$

and as the metric for cross-validation.

In the last step, lapse \hat{y}_i is forecasted if the estimated gain is positive:

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{z}^{R_j} \geq 0, \\ 0 & \text{if } \hat{z}^{R_j} < 0, \end{cases} \quad (3.29)$$

By this way we can apply the same metrics described in previous sections. Here \hat{y}_i is better to be understood as the estimation of the profitability about offering an incentive to the policyholder i rather than the forecast on the policyholder's lapse.

The two new classifications are denoted as XGB_R1 and XGB_R2, respectively, for applying XGBoost to z^{R_1} and z^{R_2} . The tuning method that we apply to estimating the parameters is described in Appendix 3.8.3.

3.6.2 Results

Table 3.12 and Figure 3.5 display the prediction accuracies. Table 3.12 shows that XGB_R1 and XGB_R2 produce relatively low mean accuracy of respectively 76.7% and 75.7%. While XGB_R2 is clearly the worst model in term of accuracy, XGB_R1 generates similar results to the logistic regression which is the worst binary classification model regarding the accuracy measure. These seemingly unsatisfied results are understandable since both XGB_R1 and XGB_R2 are not designed to predict whether a policy would be lapsed or not. What they aim for are economic gains.

Table 3.12 – Cross-validated statistical accuracies obtained with the economic loss functions.

| | LR | CART | SVM | XGB | XGB_R1 | XGB_R2 |
|--------------------|--------|--------|--------|-------|--------|--------|
| Mean Accuracy | 76.64% | 77.15% | 77.82% | 78.8% | 76.67% | 75.71% |
| Standard Deviation | 0.07% | 0.10% | 0.08% | 0.03% | 0.07% | 0.06% |

The numbers in Table 3.13 and 3.14 tell us more about why XGB_R1 and XGB_R2 performs badly in statistical accuracy. They result in the smallest correct identifications on lapses (resp. 104,889 and 99,432) and produce the most false-sense-of-security (resp. 113,948 and 119,405). However, we will see very soon that XGB_R1 and XGB_R2 stand out when we switch focus to retention gain.

Table 3.15 and Figure 3.6 show that XGB_R1 generates a significantly larger average retention gain with the aggressive incentive strategy (\$6,586,357) than other algorithms as well as a

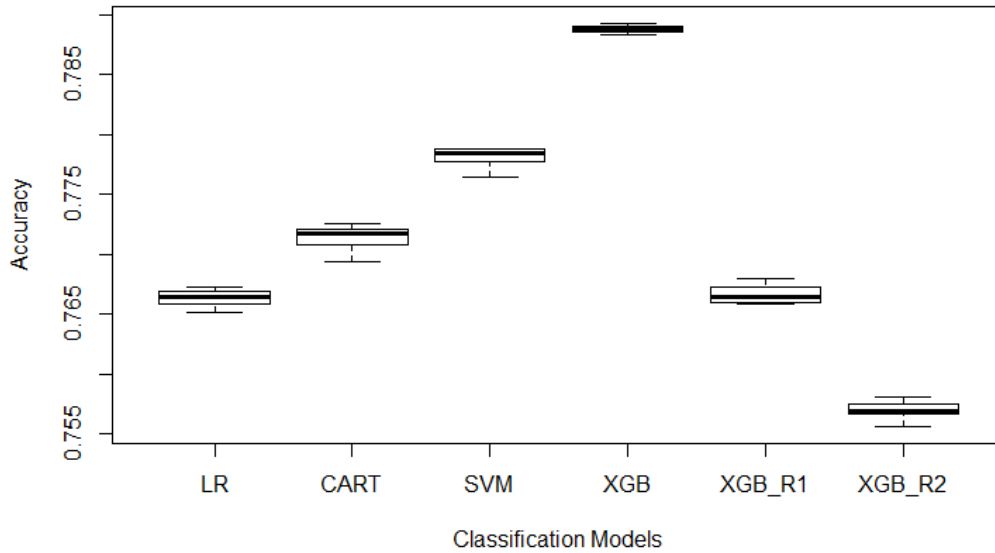


Figure 3.5 – Box plot of statistical accuracies obtained with the economic loss functions.

Table 3.13 – Average confusion matrix of XGB_R1.

| | | Predicted | |
|--------|-------|-----------|---------|
| | | Stay | Lapse |
| Actual | Stay | 329,357 | 18,204 |
| | Lapse | 113,948 | 104,889 |

Table 3.14 – Average confusion matrix of XGB_R1.

| | | Predicted | |
|--------|-------|-----------|--------|
| | | Stay | Lapse |
| Actual | Stay | 329,413 | 18,149 |
| | Lapse | 119,405 | 99,432 |

significantly lower standard deviation (\$53,460). The increase in retention gain is 26% (1.3 million USD) higher than that generated by XGB (the second-best algorithm) and 146% (3.9 million USD) better than that produced by CART. Looking back to Table 3.13, we see that XGB_R1 leads to reduce the number of false alarms (18,204) in optimizing the retention gain, even if this also reduces the correct detection (104,889). The good results of XGB_R1 in achieving retention gain demonstrate the benefit of integrating the algorithm with the goal to be achieved. The objective function for XGB_R1 to minimize, Equation (3.28), is

about predicting retention gains. XGB_R1 therefore would naturally perform the best when compared with other algorithms optimizing other objectives (such as classification accuracies).

Table 3.15 – Cross-validated retention gains with the aggressive strategy.

| | LR | CART | SVM | XGB | XGB_R1 |
|---------------------|-----------|-----------|-----------|-----------|-----------|
| Mean Retention Gain | 4,046,602 | 2,680,012 | 5,028,737 | 5,243,913 | 6,586,357 |
| Standard Deviation | 133,993 | 209,220 | 139,102 | 115,415 | 53,460 |

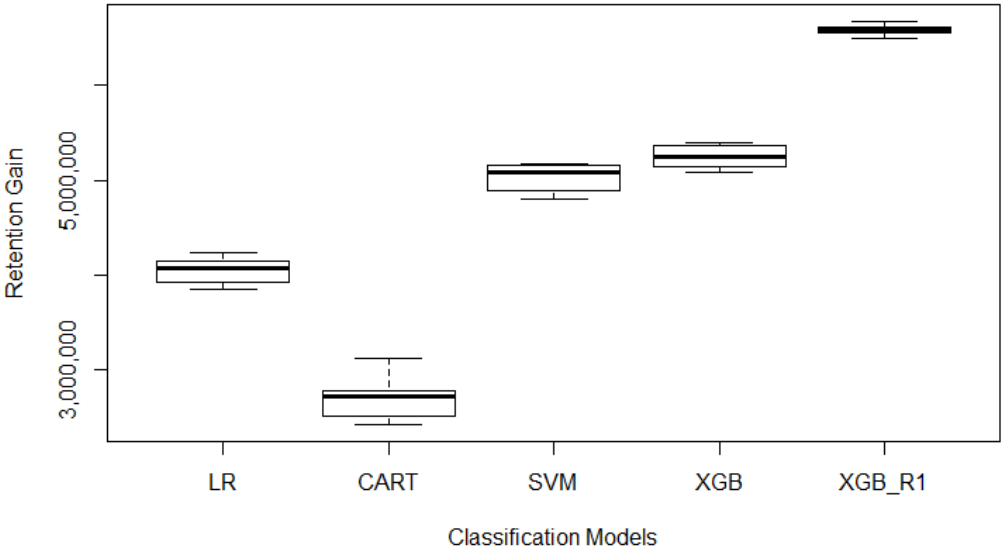


Figure 3.6 – Box plot of retention gains with the aggressive strategy.

We expect that the benefit of integrating the algorithm with the goal is robust across incentive strategies. This is confirmed by the results in Table 3.16 and Figure 3.7. XGB_R2 generates retention gain of 3.9 million dollars that is nearly 600 thousand dollars more than that achieved by the second place XGB. The increase in retention gains is 18%. The increases with respect to the commonly seen LR and CART reach 47% and 85%.

Table 3.16 – Cross-validated retention gains with the moderate strategy.

| | LR | CART | SVM | XGB | XGB_R2 |
|---------------------|-----------|-----------|-----------|-----------|-----------|
| Mean Retention Gain | 2,618,396 | 2,085,599 | 3,113,900 | 3,261,029 | 3,852,782 |
| Standard Deviation | 63,693 | 85,184 | 54,169 | 45,928 | 39,163 |

The results in this section demonstrate the benefit of having a specific objective. If senior managers of an insurer are able to specify an objective to be optimized (e.g., maximizing retention gain), the staff should apply an advanced algorithm like XGB directly to such an

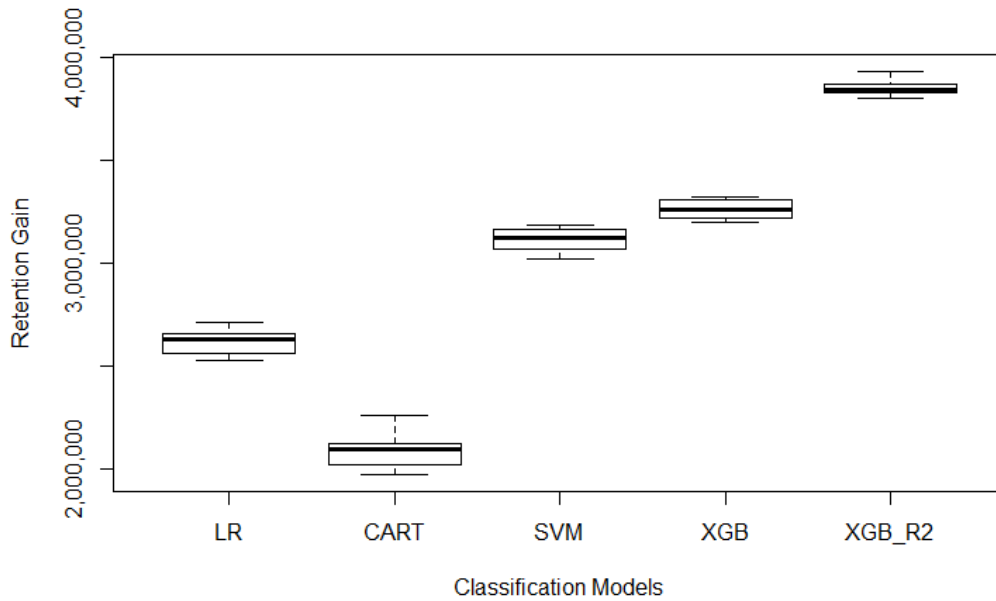


Figure 3.7 – Box plot of retention gains with the moderate strategy.

objective to achieve the optimum. The enhanced gain relative to the case having no specific objective other than classification accuracy can be substantial.

3.7 Conclusion

Lapse risk is the most significant risk associated with life insurance. Lapses may cause losses, reduce expected profits, lead to stringent liquidity, result in mis-pricing, impair the risk management, or even pose solvency threats. Employing a good algorithm to model policyholder lapse behavior is therefore valuable.

In this study, we adopt innovative viewpoints on lapse management in addition to introducing machine learning algorithms to lapse prediction. Applying XGBoost and SVM to predicting whether a policyholder will lapse her/his policy is new to the literature. Secondly, we adopt not only a statistical metric in evaluating algorithms' prediction performance but also an economic metric based on customer lifetime value and retention gains.

The goal of classification accuracy has no direct link to the insurer's costs and profits. It thus might lead to a biased strategy (Powers, 2011). Following the churn literature, we define a specific validation metric based on the economic gains. This constitutes our third contribution: we are the first to set up a profit-based loss function so that we may directly optimize the economic gains. More specifically, we change the usual statistical idea of classification to a gain regression in which profits are to be maximized.

The two machine learning algorithms, XGBoost and SVM, perform a little bit better than classic CART and logistic regression in terms of statistical accuracy on a large dataset consisting of more than six hundred thousand life insurance policies with information on policy terms and policyholders' characteristics. XGBoost has another advantage over other algorithms: it is less dependent upon the choice of training samples.

The advantages of XGBoost and SVM are more apparent with respect to retention gains. The retention gains incorporate the costs of providing incentives to policyholders to reduce lapse propensities and the benefits of retaining policies. XGBoost and SVM generate much higher retention gains than logistic regression and CART do. For instance, XGBoost produces 1.2 to 2.6 million dollars more economic gains than CART.

In the last section, we demonstrate that the economic gains can be further enhanced when the optimization is done on a function linked to economic gains rather than on statistic accuracies. The results show that the retention gains with an aggressive incentive strategy resulted from XGB_R1 is 126% of those from applying XGBoost to pursue classification accuracies, in particular by reducing the false alarm rates. An insurer should therefore apply advanced machine learning algorithms like XGB to its economic objective so that lapse management can be really optimized.

3.8 Appendix

3.8.1 XGBoost Tuning - Binary Classification

The values of the parameters tested in the grid search for the tuning of XGBoost are as follows:

- *eta*: 0.005, 0.1, 0.15;
- *gamma*: 0, 5, 10;
- *max_depth*: 10, 15, 20, 25, 30;
- *min_child_weight*: 15, 20, 25;
- *subsample*: 1;
- *colsample_bytree*: 0.4, 0.5, 0.6.

The values of the grid search are chosen by a previous sensitivity study in which we apply the same methodology on a subsample of the whole database but with a coarser grid. Then we focus on a finer grid to obtain better results within a reasonable time period. In addition, the fact that we only test subsample with the value of 1 means that we do not adopt the stochastic gradient boosting of [Friedman \(2002\)](#).

3.8.2 SVM Tuning

The values of the parameters tested in the grid search for the tuning of SVM are as follows:

- *Cost*: 0.5, 1, 2, 5, 10;
- *gamma*: 0.25, 0.5, 0.75, 1, 1.25.

Similar to the previous section, the values of the grid search are chosen by a previous sensitivity study in which we apply the same methodology on a subsample of the whole database but with a coarser grid. Then we focus on a finer grid to obtain better results. This is necessary so that the computing can be done within a reasonable time period.

3.8.3 XGBoost Tuning - Profitability

We adopt the values of most parameters generated by a previous sensitivity study as:

- *eta*: 0.005;
- *gamma*: 1;
- *max_depth*: 15;
- *min_child_weight*: 15;
- *subsample*: 0.7;
- *colsample_bytree*: 0.8.

Then, we determine the best *nrounds* through a 5-folds cross-validation with this parameter tested up to 1,000.

Bibliography

- Marie-Odile Albizzati and Hélyette Geman. Interest Rate Risk Management and Valuation of the Surrender Option in Life Insurance Policies. *The Journal of Risk and Insurance*, 61(4): 616–637, 1994. ISSN 0022-4367. doi: 10.2307/253641.
- Eva Ascarza, Scott A. Neslin, Oded Netzer, Zachery Anderson, Peter S. Fader, Sunil Gupta, Bruce G. S. Hardie, Aurelie Lemmens, Barak Libai, David Neal, Foster Provost, and Rom Schrift. In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Customer Needs and Solutions*, 5(1):65–81, March 2018. ISSN 2196-2928. doi: 10.1007/s40547-017-0080-0.
- Anna Rita Bacinello. Pricing Guaranteed Life Insurance Participating Policies with Annual Premiums and Surrender Option. *North American Actuarial Journal*, 7(3):1–17, July 2003. ISSN 1092-0277. doi: 10.1080/10920277.2003.10596097.
- Flavia Barsotti, Xavier Milhaud, and Yahia Salhi. Lapse risk in life insurance: Correlation and contagion effects among policyholders’ behaviors. *Insurance: Mathematics and Economics*, 71:317–331, November 2016. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2016.09.008.
- Daniel Bauer, Rüdiger Kiesel, Alexander Kling, and Jochen Ruß. Risk-neutral valuation of participating life insurance contracts. *Insurance: Mathematics and Economics*, 39(2):171–183, October 2006. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2006.02.003.
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 978-0-89791-497-0. doi: 10.1145/130385.130401. event-place: Pittsburgh, Pennsylvania, USA.
- Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 1573-0565. doi: 10.1007/BF00058655.
- Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.
- Leo Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 978-1-351-46049-1. doi: 10.1201/9781315139470.
- Kristian Buchardt, Thomas Møller, and Kristian Bjerre Schmidt. Cash flows and policyholder behaviour in the semi-Markov life insurance setup. *Scandinavian Actuarial Journal*, 8:1–29, November 2015. ISSN 0346-1238. doi: 10.1080/03461238.2013.879919.
- J. Campbell, M. Chan, K. Li, L. Lombardi, M. Purushotham, and A. Rao. Modeling of Policyholder Behavior for Life Insurance and Annuity Products: A Survey and Literature Review. *Society of Actuaries*, 2014.

- Rocco Roberto Cerchiara, Matthew Edwards, and Alessandra Gambini. Generalized Linear Models in Life Insurance: Decrements and Risk Factor under Solvency II. In *18th international AFIR colloquium*, Rome, 2008.
- Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. event-place: San Francisco, California, USA.
- Andrea Consiglio and Domenico De Giovanni. Pricing the Option to Surrender in Incomplete Markets. *Journal of Risk and Insurance*, 77(4):935–957, 2010. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2010.01358.x.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.
- Samuel Cox, H. and Yija Lin. Annuity Lapse Modeling: Tobit or not Tobit ? *Society of Actuaries*, 2006.
- A. Dar and C. Dodds. Interest Rates, the Emergency Fund Hypothesis and Saving through Endowment Policies: Some Empirical Evidence for the U.K. *The Journal of Risk and Insurance*, 56(3):415–433, 1989. ISSN 0022-4367. doi: 10.2307/253166.
- Martin Eling and Dieter Kiesenbauer. What Policy Features Determine Life Insurance Lapse? An Analysis of the German Market. *Journal of Risk and Insurance*, 81(2):241–269, 2014. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2012.01504.x.
- Martin Eling and Michael Kochanski. Research on lapse in life insurance: what has been done and what needs to be done? *The Journal of Risk Finance*, 14(4):392–413, August 2013. ISSN 1526-5943. doi: 10.1108/JRF-12-2012-0088.
- Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.
- Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, February 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2.
- Nadine Gatzert and Hato Schmeiser. Assessing the Risk Potential of Premium Payment Options in Participating Life Insurance Contracts. *Journal of Risk and Insurance*, 75(3): 691–712, 2008. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2008.00280.x.
- Anders Grosen and Peter Løchte Jørgensen. Fair valuation of life insurance liabilities: The impact of interest rate guarantees, surrender options, and bonus policies. *Insurance: Mathematics and Economics*, 26(1):37–57, February 2000. ISSN 0167-6687. doi: 10.1016/S0167-6687(99)00041-4.

- Sunil Gupta, Dominique Hanssens, Bruce Hardie, William Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. Modeling Customer Lifetime Value. *Journal of Service Research*, 9(2):139–155, November 2006. ISSN 1094-6705. doi: 10.1177/1094670506293810.
- Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, February 1970. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634.
- Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classification. *Working paper*, 2003. ISSN 1385-2256, 1573-1618. doi: 10.1007/s11119-014-9370-9.
- Yawen Hwang and Chenghsien Tsai. Differentiating Surrender Propensity from Lapse Propensity across Life Insurance Products. *Taiwan Risk Theory Seminar*, 2018.
- Yusho Kagraoka. Modeling Insurance Surrenders by the Negative Binomial Model. *Working paper*, 2005.
- Changki Kim. Modeling Surrender and Lapse Rates With Economic Variables. *North American Actuarial Journal*, 9(4):56–70, October 2005a. ISSN 1092-0277. doi: 10.1080/10920277.2005.10596225.
- Changki Kim. Report to the Policyholder Behavior in the Tail Subgroups Project. *Society of Actuaries*, 2005b.
- Changki Kim. Surrender Rate Impacts on Asset Liability Management. *Asia-Pacific Journal of Risk and Insurance*, 1(1), 2005c. doi: 10.2202/2153-3792.1004.
- Weiyu Kuo, Chenghsien Tsai, and Wei-Kuang Chen. An Empirical Study on the Lapse Rate: The Cointegration Approach. *Journal of Risk and Insurance*, 70(3):489–508, 2003. ISSN 1539-6975. doi: 10.1111/1539-6975.t01-1-00061.
- Aurelie Lemmens and Christophe Croux. Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, 43(2):276–286, 2006. ISSN 0022-2437.
- Aurelie Lemmens and Sunil Gupta. Managing Churn to Maximize Profits. SSRN Scholarly Paper ID 2964906, Social Science Research Network, Rochester, NY, May 2017.
- Stéphane Loisel and Xavier Milhaud. From deterministic to stochastic surrender risk models: Impact of correlation crises on economic capital. *European Journal of Operational Research*, 214(2):348–357, October 2011. ISSN 0377-2217. doi: 10.1016/j.ejor.2011.04.038.
- D. Meyer, E. Dimitriadou, K. Hornik, F. Leisch, A. Weingessel, C. Chang, and C. Lin. Misc Functions of Department of Statistics, Probability, Theory Group (Formerly : E1071). *R package version 1.7-0.*, 2015.

- Xavier Milhaud. Exogenous and Endogenous Risk Factors Management to Predict Surrender Behaviours. *ASTIN Bulletin: The Journal of the IAA*, 43(3):373–398, September 2013. ISSN 0515-0361, 1783-1350. doi: 10.1017/asb.2013.2.
- Xavier Milhaud, Stéphane Loisel, and Véronique Maume-Deschamps. Surrender triggers in life insurance: what main features affect the surrender behavior in a classical economic context? *Bulletin Français d'Actuariat*, 11(22):5–48, 2011.
- J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972. ISSN 2397-2327. doi: 10.2307/2344614.
- Scott A. Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 43(2):204–211, May 2006. ISSN 0022-2437. doi: 10.1509/jmkr.43.2.204.
- William S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, December 2006. ISSN 1546-1696. doi: 10.1038/nbt1206-1565.
- Jean Pinquet, Montserrat Guillén, and Mercedes Ayuso. Commitment and Lapse Behavior in Long-Term Insurance: A Case Study. *Journal of Risk and Insurance*, 78(4):983–1002, 2011. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2011.01420.x.
- David Martin Powers. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *International Journal of Machine Learning Technology*, 2(1):37–63, 2011. ISSN 2229-3981.
- A. E. Renshaw and S. Haberman. Statistical analysis of life assurance lapses. *Journal of the Institute of Actuaries*, 113(3):459–497, December 1986. ISSN 2058-1009, 0020-2681. doi: 10.1017/S0020268100042566.
- T. Therneau, B. Aktinson, and B. Ripley. Recursive partitioning and regression trees. *R package version 4.1-13.*, 2018.
- Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x.
- Chenghsien Tsai, Weiyu Kuo, and Derek Mi-Hsiu Chiang. The Distributions of Policy Reserves Considering the Policy-Year Structures of Surrender Rates and Expense Ratios. *Journal of Risk and Insurance*, 76(4):909–931, 2009. ISSN 1539-6975. doi: 10.1111/j.1539-6975.2009.01324.x.

- T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, June 2015. ISSN 1569-190X. doi: 10.1016/j.simpat.2015.03.003.
- Jacques Wainer. Comparison of 14 different families of classification algorithms on 115 binary datasets. *Working paper*, June 2016. arXiv: 1606.00930.
- Guo-en Xia and Wei-dong Jin. Model of Customer Churn Prediction on Support Vector Machine. *Systems Engineering - Theory & Practice*, 28(1):71–77, January 2008. ISSN 1874-8651. doi: 10.1016/S1874-8651(09)60003-X.
- Yu Zhao, Bing Li, Xiu Li, Wenhuan Liu, and Shouju Ren. Customer Churn Prediction Using Improved One-Class Support Vector Machine. In Xue Li, Shuliang Wang, and Zhao Yang Dong, editors, *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 300–306. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31877-4.

Chapitre 4

Potentiel des données satellite dans la gestion des risques financiers agricoles

Ce chapitre reprend l'article "Can satellite data forecast valuable information from USDA reports? Evidences on corn yield estimates".

Abstract

On the one hand, recent advances in satellite imagery and remote sensing allow one to easily follow in near-real time the crop conditions all around the world. On the other hand, it has been shown that governmental agricultural reports contain useful news for the commodities market, whose participants react to this valuable information. In this paper, we investigate whether one can forecast some of the newsworthy information contained in the USDA reports through satellite data. We focus on the corn futures market over the period 2000-2016. We first check the well-documented presence of market reactions to the release of the monthly WASDE reports through statistical tests. Then we investigate the informational value of early yield estimates published in these governmental reports. Finally, we propose an econometric model based on MODIS NDVI time series to forecast this valuable information. Results show that market rationally reacts to the NASS early yield forecasts. Moreover, the modeled NDVI-based information is significantly correlated with the market reactions. To conclude, we propose some ways of improvement to be considered for a practical implementation.

Keywords: Commodities market; Data Enrichment; Market information; MODIS; NDVI; USDA reports.

4.1 Introduction

The value of public information in agricultural commodity markets has been a topic of great attention for many years in the literature (e.g., [Sumner and Mueller, 1989](#); [Garcia et al., 1997](#); [Isengildina-Massa et al., 2008](#); [Dorfman and Karali, 2015](#); [Gouel, 2018](#)). More recently, with the improvement of data access and the emergence of the *Big Data* era, the interest in this subject has enhanced, especially on the possible declining value of USDA reports ([Karali et al., 2019](#); [Tack et al., 2017](#); [Ying et al., 2017](#)). Indeed, the recent advances in satellite data and remote sensing allow one to easily follow in near-real time the crop progress all over the world thanks to weather data or vegetation index (e.g., [Prasad et al., 2006](#); [Mkhabela et al., 2011](#)). Nevertheless, until now, studies show that USDA reports still have a significant impact on the commodities market. The purpose of this paper is to provide evidences that some valuable information contained in the governmental reports can be forecasted by using satellite data. Our regression analysis shows significant correlation between our satellite data-based forecasted information and the corn futures market reactions to the report releases. We also suggest some needed enhancements of our methodology for a practical application.

The monthly World Agricultural Supply and Demand Estimates (WASDE) report, provided by the United States Department of Agriculture (USDA), together with, for some specific months, the National Agricultural Statistics Service (NASS) Crop Production report, are the most valuable public sources of information for the U.S. commodities market. An extensive literature already examines the impact of these governmental report releases on the commodities market.

[Sumner and Mueller \(1989\)](#) were the first to use an event-study framework to explore the impact of USDA reports on corn and soybean future prices over the period 1961-1982. They conclude that the information included in the governmental forecasts are considered as new and reliable by the market participants, since the changes in prices on days just following the publication are significantly higher than on other days. [Isengildina-Massa et al. \(2008\)](#) find consistent results by testing differences in variance of financial returns over the period 1985-2006 for the same commodities. Other similar studies have been conducted, and a large majority of them conclude on the fact that the commodities market is significantly impacted by the WASDE report releases ([Garcia et al., 1997](#); [Irwin et al., 2001](#); [McKenzie, 2008](#)).

On the contrary to the previous investigations, which rely on statistical tests such as F or Chi-squared test, other studies base their analysis on a regression model. Thus, [Fortenbery and Sumner \(1993\)](#) regress the changes in prices on, amongst other, a zero-one dummy variable for report dates. [Lehecka \(2014\)](#) quantifies the futures price reaction by using the crop condition information included in the weekly Crop Progress reports over the period 1986-2012. The author regresses the close-to-open return on the change in the percentages of the crop in excellent and good condition. The results highlight a significant and negative influence of the

crop condition improvement on the prices change, which is a rational outcome in respect with the supply and demand theory.

More recently, other concerns have been raised with the significant increase of available information from multiple sources and the essor of *Big Data*. Intuitively, the augmentation of existing data sources should lower the impact of the USDA reports on the commodities market. However, [Karali et al. \(2019\)](#) and [Ying et al. \(2017\)](#) examine this hypothesis and conclude that the effects are not decreasing. Furthermore, it even seems that the informational value of some reports tends to enhance. Alternatively, [Milacek and Brorsen \(2017\)](#), based on the assumption that a private firm would have developed a WASDE reports forecasting model for trading purpose, determine the informational value of this latter. [Abbott et al. \(2016\)](#) also quantify the value of corn information contained in the WASDE reports to \$301 million, and, in particular, assess the corn yield information to \$188 million.

One of the data sources that private firms might use to forecast the USDA reports is satellite imagery. In this paper we focus on a specific vegetation index: the Normalized Difference Vegetation Index (NDVI), which is a reflectance index. [Colwell \(1956\)](#) was the first to explore the detection of crops healthiness by aerial infrared photographs. [Kumar and Silva \(1973\)](#) study into details the link between one crop reflectance and its chlorophyll activity. Indeed, they remark a specific signature in the near-infrared. Introduced by [Rouse et al. \(1974\)](#), the NDVI is nowadays one of the most used and studied vegetation index and is defined as follows

$$NDVI = \frac{R_{NIR} - R_R}{R_{NIR} + R_R}, \quad (4.1)$$

where R_{NIR} and R_R are the reflectance in the near-infrared and in red, respectively. Thus, a dense tropical forest NDVI value is positive from 0.6 to 0.8, while the bare soil leads to lower value around 0.1, and rock or snow have negative NDVI values.

The NDVI is widely used in agricultural remote sensing, and more specifically in crop yield forecasting. Many forecasting models have been developed based on simple linear regression, in particular for corn ([Prasad et al., 2006](#)), soybeans ([Ma et al., 2001](#)) and wheat ([Mkhabela et al., 2011](#)). Other more sophisticated algorithms have also been applied, such as [Li et al. \(2007\)](#) who use neural networks. Different variables, derived from the NDVI time series, can be used for yield forecasting. Thus, [Mkhabela et al. \(2011\)](#) estimate crop yield thanks to the mean NDVI value over the growth period, while [Zhang et al. \(2012\)](#) split the series into two distinct periods: from re-greening to heading and from heading to maturity. Numerous models exist due the specificity of each research project. Indeed, studies have been conducted in many countries such as, for examples, Zimbabwe ([Svotwa et al., 2014](#)), Hungary ([Ferencz et al., 2004](#)), China ([Ren et al., 2008](#)) or the U.S. ([Prasad et al., 2006](#); [Becker-Reshef et al., 2010b](#)). Different data sources can be used, like the Advanced Very High Resolution Ra-

diometer (AVHRR) (Rasmussen, 1997), the Moderate Resolution Imaging Spectroradiometer (MODIS) (Doraiswamy et al., 2005), or Sentinel-2 (Skakun et al., 2017b). Applications cover a large number of crop types (corn, soybeans, wheat, barley, rice, tobacco, potato, sugarcane, etc.). Furthermore, in some studies, authors associate NDVI with weather data, like rainfall estimates and humidity index (Prasad et al., 2006), to improve the forecasting model accuracy. In general, these studies conclude to a high and significant correlation between NDVI and crop yield, with a R^2 regularly up to 0.9.

More recently, with the improvement of satellite data resolution, more sophisticated remote sensing studies have emerged. Thus, Pervez and Brown (2010) manage to determine if a land is irrigated or not with an accuracy of 92% in California, while Peterson et al. (2011) detect the irrigation for different crops in Kansas with an accuracy of 88%. Beyond irrigation detection, crop mapping studies have gained even more interests (e.g., Wardlow and Egbert, 2008, 2010; Skakun et al., 2017a; Gao et al., 2017; Zhong et al., 2019). The development of such maps is useful since it significantly improve the crop yield forecasting models (Maselli and Rembold, 2001; Kastens et al., 2005).

In this paper, we investigate whether one can forecast some of the valuable information from the USDA reports through satellite open-access data. We focus on the corn yield early estimates information from the NASS reports of August, September and October over the period 2000-2016, by observing the reaction of the corn futures with a maturity in December. The chosen satellite data is the NDVI values derived from MODIS aboard the NASA's satellites Terra and Aqua.

First, we test, on the data we base our study on, that the commodity market rationally reacts to the early yield estimates from the NASS reports. To do so, we follow the general econometric methodology of Lehecka (2014) to model the valuable information contained in the USDA announcements. Then, during the growing season, we use the MODIS NDVI data to forecast the corn yield through linear regression models, trained on the final crop yields. Finally, we focus on three specific NDVI-based estimates, which are the ones obtained around two weeks before the publication of the WASDE report, therefore forecasting the valuable information contained in the next governmental report. Following again the methodology of Lehecka (2014), we test the correlation between our NDVI-based forecasts of valuable information and the commodity market reactions.

Our findings show significant and rational correlations between our modeled information contained in the reports and the financial returns of corn futures. This result is in accordance with the existing literature (Irwin et al., 2001; Lehecka, 2014). Moreover, the correlation between the financial returns and our NDVI-based forecasted information is also significant and in line with the supply and demand theory. Since the NDVI data needed is available, at least, one week before USDA announcements, it highlights the possibility of knowing in advance some

of the valuable information contained in the reports.

The remainder of this paper is organized as follows. In Section 4.2 we describe the data we base our study on. The methodology we apply is defined in Section 4.3. We present the results that we obtained in Section 4.4. Furthermore, Section 4.5 considers future possible improvements of the current work for a practical implementation. Finally, Section 4.6 concludes.

4.2 Data

4.2.1 NDVI Data

In this study, we use the MODIS NDVI data, available in open facilitated access thanks to the Global Agriculture Monitoring Project (Becker-Reshef et al., 2010a) initiated between the NASA, the University of Maryland and the USDA Foreign Agriculture Service (FAS). The geographic resolution is of 250 meters and the temporal one is 16 days (MOD44 16-days product). Thus, every 16 days, we obtain a mean image of the period. This methodology is due to the fact that NDVI is a reflectance index, and is therefore sensitive to weather conditions such as clouds when estimated from satellite (Whitcraft et al., 2015). The release of these images is done at a constant pace in the year (cf. Table 4.1). We also apply the standard water and crop mask (MOD 12) to focus on the crop conditions. MODIS data are available since February 2000.

4.2.2 Futures Data

On the contrary to Irwin et al. (2001) and Isengildina-Massa et al. (2008), we don't explore the market reaction through the futures whose maturity are the closest to the session of interest. We rather base our study on the corn futures with the December maturity of the very year. Since our results may be useful for the agricultural insurance companies' risk management, we focus on the price at risk for the revenue protection products, which is the new-crop harvest price, i.e., December for corn in the U.S.

4.2.3 Crop yield estimates

Finally, to model the valuable information contained in the NASS reports, we use the early corn yield estimates of 10 states from the Corn Belt, namely Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Nebraska, Ohio, South Dakota and Wisconsin. Indeed, these states represent between 82 and 85% of the corn production in the United States. More particularly, we explore the market reaction to the release of August, September and October yield estimates over the period 2000-2016.

In order to train our NDVI-based yield forecasting model, we also use the final corn yield estimates, communicated by the NASS in the January following the harvest.

Table 4.1 – NDVI image periods and corresponding calendar dates.

| Period | Starting Date | Ending Date |
|--------|---------------|-------------|
| 1 | 01 Jan. | 16 Jan. |
| 2 | 17 Jan. | 01 Feb. |
| 3 | 01 Feb. | 17 Feb. |
| 4 | 18 Feb. | 05 Mar. |
| 5 | 06 Mar. | 21 Mar. |
| 6 | 22 Mar. | 06 Apr. |
| 7 | 07 Apr. | 22 Apr. |
| 8 | 23 Apr. | 08 May |
| 9 | 09 May | 24 May |
| 10 | 25 May | 09 Jun. |
| 11 | 10 Jun. | 25 Jun. |
| 12 | 26 Jun. | 11 Jul. |
| 13 | 12 Jul. | 27 Jul. |
| 14 | 28 Jul. | 12 Aug. |
| 15 | 13 Aug. | 28 Aug. |
| 16 | 29 Aug. | 13 Sep. |
| 17 | 14 Sep. | 29 Sep. |
| 18 | 30 Sep. | 15 Oct. |
| 19 | 16 Oct. | 31 Oct. |
| 20 | 01 Nov. | 16 Nov. |
| 21 | 17 Nov. | 02 Dec. |
| 22 | 03 Dec. | 18 Dec. |
| 23 | 19 Dec. | 03 Jan. |

4.3 Methodology

4.3.1 Event study analysis

Following already existing event study methodologies (Irwin et al., 2001; Isengildina-Massa et al., 2008), we first check that the commodity market reacts to the release of the reports over the period of interest. We focus our research on the future returns to test if the variability is higher in the trading session just after the WASDE release than normal, i.e., sessions in a temporal window of 5 days before and after. Since we carry out our study over the period 2000-2016, we need to pay attention to the futures return definition. Indeed, before the year 2013, USDA reports were released at 8:30 a.m., while, since January 1, 2013, the statistical reports have now been published at 12:00 p.m.¹ Thus, if a market reaction exists, its timing might have changed over the years.

Therefore, we define the futures return of interest as

1. <https://www.usda.gov/media/press-releases/2012/09/19/usda-announces-change-release-time-key-statistical-reports>

$$r_{t,i,N} = \ln \left(\frac{p_{t,i,N}^O}{p_{t-1,i,N}^C} \right) \times 100, \quad (4.2)$$

if the release happened strictly before 2013, and

$$r_{t,i,N} = \ln \left(\frac{p_{t,i,N}^C}{p_{t,i,N}^O} \right) \times 100, \quad (4.3)$$

if the report publication occurred after January 1, 2013; and where $p_{t,i,N}^O$ and $p_{t,i,N}^C$ are respectively the opening and the closing price of corn futures with a maturity in December of year N for the session t of the event (i, N) , i.e., the WASDE release of month i of year N . Descriptive statistics of the defined financial returns are displayed in Table 4.2.

Then, we perform a F -test of equality of variances to compare the variability of the returns in the session just following the release of the report ($t = 0$) with the sessions of the event window ($t \in \{-5, \dots, -1, 1, \dots, 5\}$). Under market efficiency assumption, if the market participants consider that the reports contain new and reliable information, the futures price variability should be higher on the announcement date.

Table 4.2 – Descriptive statistics of corn future returns over the period 2000-2016.

| | r | $ r $ |
|----------|---------|--------|
| Mean | 0.009 | 0.772 |
| Median | -0.044 | 0.456 |
| 1st Qu. | -0.468 | 0.189 |
| 3rd Qu. | 0.442 | 0.442 |
| Min | -17.477 | 0.000 |
| Max | 24.429 | 24.429 |
| Variance | 1.939 | 1.343 |

4.3.2 Regression analysis - USDA reports

The tests described in Section 4.3.1 are only designed to detect that new and reliable information is contained in the reports. However, this methodology doesn't allow us to know to which information the market exactly reacts to. Indeed, WASDE and NASS reports contain a large amount of statistics. In particular, [Abbott et al. \(2016\)](#) show that corn yield forecasts represent a significant value of \$188 million for the market. Inspired by [Lehecka \(2014\)](#) methodology, we test if the crop yield estimates are considered as valuable information.

Indeed, the early yield estimates are crucial data for the commodity market in the growing season. With the harvested area, it gives an early forecast of the next harvest production and, therefore, an outlook of the supply. All other factors being equal, a change in the crop

yield forecasts should impact the futures price. More precisely, if the USDA scales up her expectation in terms of yield, the futures price should decrease, since the supply augments.

In the NASS report of month i , published in the year N , new early corn yield estimates, noted $Y_{k,i,N}$, are displayed for each state of interest k . We also note $Y_{k,N}$ the final yield estimates of state k for the year N . We focus on the reports released in August, September and October. In the current paper, we suppose that no other information is integrated by the market between two WASDE reports publication (this latter hypothesis will be discussed later in Section 4.5). Thus, for September and October, we model the new information provided by the report $i \in \{9, 10\}$ of year N as

$$X_{k,i,N} = \ln \left(\frac{Y_{k,i,N}}{Y_{k,i-1,N}} \right) \times 100. \quad (4.4)$$

However, this modeling cannot be applied to the August release since it is the first early NASS yield forecast of the year. We therefore define the information by

$$X_{k,8,N} = \ln \left(\frac{Y_{k,8,N}}{\overline{Y_{k,N}}} \right) \times 100, \quad (4.5)$$

where $\overline{Y_{k,N}}$ represents the Olympic mean of the final yields from the last 5 years before N , computed as follows

$$\overline{Y_{k,N}} = \frac{1}{3} \left(\sum_{n=1}^5 Y_{k,N-n} - \max_{n \in \{1, \dots, 5\}} \{Y_{k,N-n}\} - \min_{n \in \{1, \dots, 5\}} \{Y_{k,N-n}\} \right) \quad (4.6)$$

The 5-years Olympic mean is a major index, and is notably used in the Agricultural Risk Coverage Program (ARC) (Kim et al., 2015).

Finally, and similarly to Lehecka (2014), we perform a regression analysis to determine the possible market reaction to the news by

$$r_{0,i,N} = \beta_0 + \beta_1 X_{k,i,N}. \quad (4.7)$$

With this equation, one can assess the significance of the linear correlation between the modeled USDA information contained in the reports and the future returns. However, the market reaction might not necessary be linear with the governmental news. Thus, we also estimate the Kendall rank correlation coefficient between $r_{0,i,N}$ and $X_{k,i,N}$ to release the linear relation hypothesis.

4.3.3 Regression analysis - NDVI forecasts

In this section, we explore the possibility of forecasting the NASS reports information through the MODIS NDVI data. These images are highly correlated to the vegetation conditions, and therefore to corn conditions. The World Agricultural Outlook Board (WAOB) and the Foreign Agricultural Service (FAS) teams actually already use satellite imagery and weather analysis to monitor crop conditions in order to prepare the WASDE². Hence, it is logical to use similar data when aiming to predict its information.

First, we develop corn yield forecasting models based on MODIS NDVI time series for each of the 10 states considered in the study. We recall the period notation used in the MODIS NDVI time series in Table 4.1. We note $V_{k,P,N}$ the mean value of NDVI over the state k during the period P of the year N . Then, for each period between 9 and 17, we define two variables derived from the MODIS NDVI time series: the growing phase total NDVI value, defined as follows,

$$G_{k,P,N} = \sum_{p=9}^P V_{k,p,N}, \quad (4.8)$$

and the maximum NDVI peak value reached during the season,

$$M_{k,P,N} = \max_{p \leq P} V_{k,p,N}. \quad (4.9)$$

Then, we model the NDVI-based yield forecasts by the following linear regressions for each state

$$Y_{k,N} = \beta_{k,P,0} + \beta_{k,P,1} \cdot N + \beta_{k,P,2} \cdot M_{k,P,N} + \beta_{k,P,3} \cdot (G_{k,P,N} - M_{k,P,N}) + \epsilon_{k,P,N}, \quad (4.10)$$

where $Y_{k,N}$ is the final yield estimates of year N for the State k and $\epsilon_{k,P,N}$ is an error term. We estimate the coefficients through the ordinary least square methods, leading to $\hat{\beta}_{k,P,l}$ for each state k , $l \in \{0, 1, 2, 3\}$ and $P \in \{13, 15, 17\}$. Thus, we obtain NDVI-based early yield estimates

$$\hat{Y}_{k,P,N}^{NDVI} = \hat{\beta}_{k,P,0} + \hat{\beta}_{k,P,1} \cdot N + \hat{\beta}_{k,P,2} \cdot M_{k,P,N} + \hat{\beta}_{k,P,3} \cdot (G_{k,P,N} - M_{k,P,N}). \quad (4.11)$$

In this methodology, we note two major points. Firstly, the NASS early yield estimates are never used to train the NDVI-based yield forecasting models: we only rely on the governmental final yield. We chose to do so to avoid overfitting issue that training on the NASS early yield

2. <https://www.usda.gov/oce/commodity/wasde/prepared.htm>

estimates may have raised. Secondly, the three NDVI periods we train our models on, are ending on the July 27, August 28 and September 29. The data needed is therefore available before the WASDE releases of August, September and October respectively.

Hence, following the same underlying idea developed in Section 4.3.2, we model the new information provided in the WASDE reports to the market participants by

$$\hat{X}_{k,8,N}^{NDVI} = \left(\ln \left(\hat{Y}_{k,13,N}^{NDVI} \right) - \ln \left(\overline{Y_{k,N}} \right) \right) \times 100, \quad (4.12)$$

$$\hat{X}_{k,9,N}^{NDVI} = \left(\ln \left(\hat{Y}_{k,15,N}^{NDVI} \right) - \ln \left(\hat{Y}_{k,13,N}^{NDVI} \right) \right) \times 100, \quad (4.13)$$

$$\hat{X}_{k,10,N}^{NDVI} = \left(\ln \left(\hat{Y}_{k,17,N}^{NDVI} \right) - \ln \left(\hat{Y}_{k,15,N}^{NDVI} \right) \right) \times 100, \quad (4.14)$$

respectively for the August, September and October releases.

Finally, we perform a regression analysis to determine if this modelled information is correlated enough with the information contained in the governmental reports. In other words, we test whether the NDVI-based information forecasts could have been statistically considered as information regarding the corn future price changes at the WASDE release dates. Similarly to Section 4.3.2, we test this hypothesis thanks to the following regression,

$$r_{0,i,N} = \beta_0 + \beta_1 \hat{X}_{k,i,N}^{NDVI}. \quad (4.15)$$

4.4 Results

4.4.1 Market reaction

In this section, we present the results obtained about the impact of the report releases on the commodities market. First, we display in Figures 4.1, 4.2 and 4.3 the mean absolute value of the financial returns according to session t relative to the report publication. For the months of August (Figure 4.1) and October (Figure 4.3), we note a higher value of the mean absolute return of the futures for the session just following the WASDE reports. However, for the September reports (Figure 4.2), results are less convincing: even though the reports session has one of the highest value, the difference is not clear when comparing to other sessions. Intuitively, we would conclude that the August and October reports contain new and reliable information for the commodity market, which therefore reacts, while the September ones have less impacts.

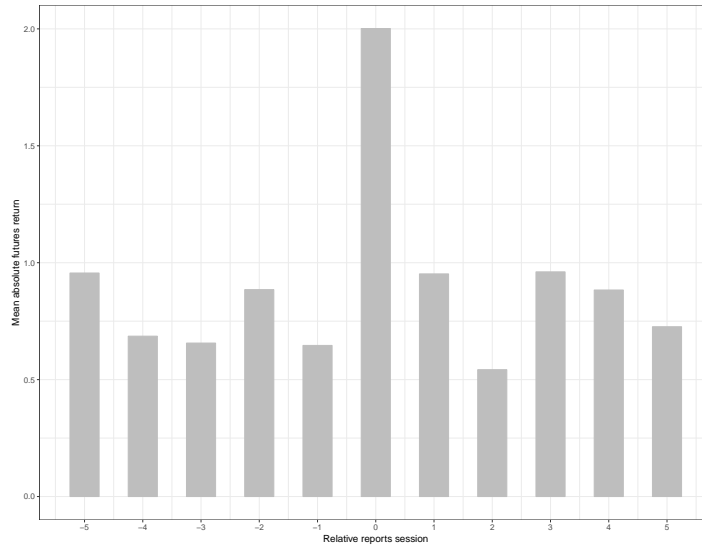


Figure 4.1 – Mean absolute future returns relative to the August reports session during the period 2000-2016

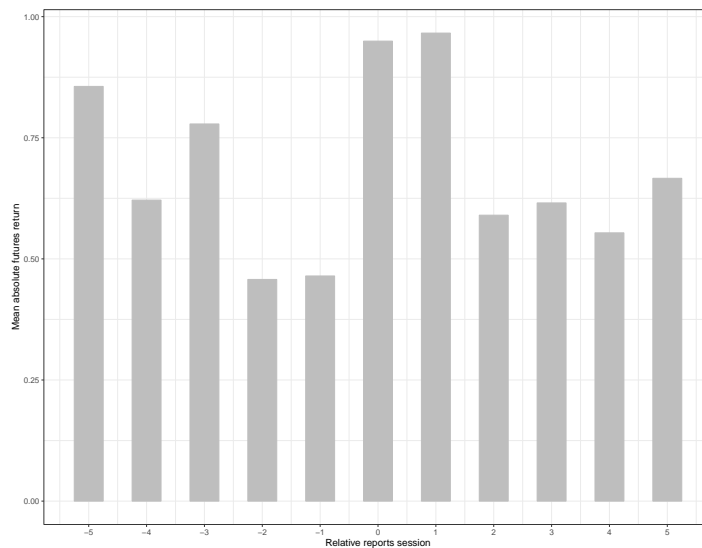


Figure 4.2 – Mean absolute future returns relative to the September reports session during the period 2000-2016

To validate this graphical intuition, we perform a F -test, whose results are presented in Table 4.3. Return variance for all reports session considered is 4.2 times more important than pre and post reports return variance, and this difference is significant at a 0.001% level. However, we note that all the reports don't seem to have the same impact on the commodity market. Thus, the ratio of variances is around 5 for the releases of August and October, and only of 2.4 for September reports. This result is highlighted by the p -values, which indicate levels of confidence lower than 0.001% for both August and October, while the September one is around 1%.

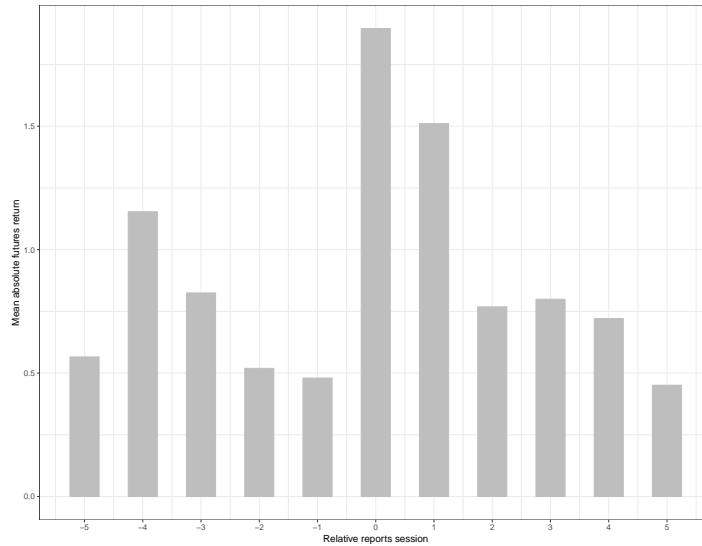


Figure 4.3 – Mean absolute future returns relative to the October reports session during the period 2000-2016

Table 4.3 – Future return volatility test results for WASDE reports in Corn Markets, August-October, 2000-2016.

| Reports | Reports Variance | Pre/Post Reports Variance | F -statistic | p -value |
|-------------|------------------|---------------------------|----------------|---------------|
| August | 7.554 | 1.467 | 5.152 | $< 1.10^{-5}$ |
| September | 2.073 | 0.879 | 2.362 | 0.0068 |
| October | 7.822 | 1.737 | 4.503 | $< 1.10^{-5}$ |
| All Reports | 5.727 | 1.360 | 4.211 | $< 1.10^{-5}$ |

We now test more specifically whether the information contained in the NASS early yield reestimation have an impact on the corn futures market. Results of the Equation (4.7) are displayed in Table 4.4, together with the Kendall rank correlation coefficient. First, we note that, for all the statistics, the estimates are negative with levels of confidence at least of 2%. This suggests that not only the future prices react to NASS early yield changes, but also that the reaction is rational. The significant negative sign of the estimated coefficient is consistent with the economic theory: a reduction (res. augmentation) of the crop yield expectations leads to an increase (res. decrease) of the commodity prices. Nevertheless, we remark that the significance levels differ depending of the publication month. As anticipated in the variances test, the reaction is less marked for September releases. Indeed, the p -values of Pearson's r and Kendall's τ are respectively only of 1.8% and 1.1%, while for other months the tests lead to a minimum of 0.03% level of confidence.

The results obtained in this market reaction study are consistent with the existing literature, like for example [Isengildina-Massa et al. \(2008\)](#) who also detect a lower, while still statistically significant, impact of the September reports on the market. Thus, we can conclude on the fact

Table 4.4 – Regression results of future prices impacts to early corn yield estimate changes, and their Kendall τ , August to October, 2000-2016

| Reports | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | | Kendall τ | |
|-------------|-----------------|-----------------|-----------------|-----------------|----------------|-----------------|
| | Estimates | <i>p</i> -value | Estimates | <i>p</i> -value | Estimates | <i>p</i> -value |
| August | 0.835 | 5.10^{-5} | -7.784 | 2.10^{-5} | -0.21 | 5.10^{-5} |
| September | -0.021 | 0.843 | -8.001 | 0.018 | -0.14 | 0.011 |
| October | -0.273 | 0.196 | -24.3 | 3.10^{-4} | -0.20 | 3.10^{-4} |
| All reports | 0.162 | 0.122 | -7.869 | $< 1.10^{-5}$ | -0.15 | $< 1.10^{-5}$ |

that, during the period 2000-2016, the market participants considered that the WASDE reports of August, September and October, contain new and reliable information, and therefore react to it. Furthermore, we show evidences that, in particular, the NASS yield estimation changes provide valuable information to the commodity market, which seems to rationally react to it. This last result is line with the findings of [Abbott et al. \(2016\)](#) who estimate the corn yield informational value of WASDE reports to \$188 million.

4.4.2 Forecasting valuable information

In this section, we display the results from the NDVI-based analysis. First, we present in Table 4.5 estimates of the early yield forecasting models for each state based on the MODIS NDVI time series described in Equation (4.10). We note that the end of September models (period 17) achieve high adjusted- R^2 values, between 0.82 and 0.91, for almost every state. On the contrary, Kansas corn forecast model fail to fulfill such level of accuracy, even though it leads to a rather correct R^2 of 0.61. Similarly, the adjusted- R^2 increases throughout the growing season for all the states except Kansas and, to a lesser extent, Wisconsin. Thus, for example in Minnesota, at end of July the models lead to an accuracy of 0.54, which augments up to 0.77 around end of August and finally achieves an adjusted- R^2 of 0.85 at end of September. The special case of Kansas must be due to the fact that, contrary to other states where the two main crop types are corn and soybeans, Kansas is also a large winter wheat producer (cf. Table 4.6 for example). Hence, MODIS NDVI time series from wheat interfere with the corn ones when estimating the NDVI at a state level.

The NDVI-based early yield forecasting model results are in line with the WASDE's aim at providing accurate information about future yields, and perfecting it along the growing season. Thus, it is a logical candidate to model the information given by the WASDE reports. We therefore plot the modeled NDVI-based information (estimated with the Equations (4.12) to (4.14)) to the NASS-based one (estimated with the Equation (4.4) and (4.5)) in Figure 4.4. We graphically observe a positive correlation, which is statistically validated. The Pearson's r and Kendall's τ correlation coefficients are equal respectively to 0.68 and 0.35, both with a significance level of 0.1%.

Table 4.5 – Regression results of final corn yield estimate to NDVI time series, 2000-2016

| States | Period | Dependant variable estimates | | | | Adjusted- R^2 |
|--------------|--------|------------------------------|--------|---------|-------------|-----------------|
| | | Intercept | Year | Maximum | ($G - M$) | |
| Illinois | 13 | -3809* | 1.8* | 681*** | -81* | 0.60 |
| | 15 | -2863** | 1.2* | 575*** | 25 | 0.84 |
| | 17 | -2042. | 0.78 | 415** | 59* | 0.86 |
| Indiana | 13 | -1784 | 0.84 | 458*** | -49* | 0.61 |
| | 15 | -3219** | 1.4** | 623*** | 33* | 0.85 |
| | 17 | -2656** | 1.1* | 472*** | 56* | 0.87 |
| Iowa | 13 | -3085* | 1.5* | 407* | -46 | 0.46 |
| | 15 | -2903** | 1.2* | 693*** | 1.9 | 0.77 |
| | 17 | -2352** | 0.97* | 487** | 32. | 0.82 |
| Kansas | 13 | -3777 | 0.15 | 292* | 20 | 0.60 |
| | 15 | 303** | -0.19 | 57 | 58* | 0.64 |
| | 17 | 964 | -0.52 | 90 | 39* | 0.61 |
| Michigan | 13 | -4284*** | 2.2*** | 224** | -51 | 0.81 |
| | 15 | -3478*** | 1.6** | 407*** | 40* | 0.87 |
| | 17 | -3033** | 1.3** | 397** | 45* | 0.88 |
| Minnesota | 13 | -3186* | 1.5* | 265 | 17 | 0.54 |
| | 15 | -1799. | 0.6 | 613** | 54** | 0.77 |
| | 17 | -1576. | 0.52 | 477** | 63*** | 0.85 |
| Nebraska | 13 | -3483** | 1.7** | 175* | 28 | 0.71 |
| | 15 | -3103** | 1.5** | 108 | 39. | 0.77 |
| | 17 | -3072*** | 1.5*** | 0.7 | 47** | 0.89 |
| Ohio | 13 | -3055 | 1.5 | 201. | 64 | 0.53 |
| | 15 | -1157 | 0.34 | 569*** | 48** | 0.91 |
| | 17 | -1620. | 0.57 | 293. | 79** | 0.91 |
| South Dakota | 13 | -4122* | 2.0* | 224* | 5.8 | 0.64 |
| | 15 | -3701** | 1.8* | 233* | 21 | 0.73 |
| | 17 | -3517** | 1.7** | 98 | 49** | 0.84 |
| Wisconsin | 13 | -3124*** | 1.5** | 207** | 36 | 0.80 |
| | 15 | -3001** | 1.4** | 310. | 47* | 0.79 |
| | 17 | -3167*** | 1.4*** | 215 | 50** | 0.83 |

Note: The levels of significance are noted as: * * * for 0.1%, ** for 1%, * for 5% and . for 10%.

Table 4.6 – Area planted for corn, soybeans and winter wheat in Iowa and Kansas in 2016 (thousands of acres).

| State | Crop | Planted area |
|--------|----------|--------------|
| Iowa | Corn | 13,900 |
| Iowa | Soybeans | 9,500 |
| Iowa | Wheat | 25 |
| Kansas | Corn | 5,100 |
| Kansas | Soybeans | 4,050 |
| Kansas | Wheat | 8,500 |

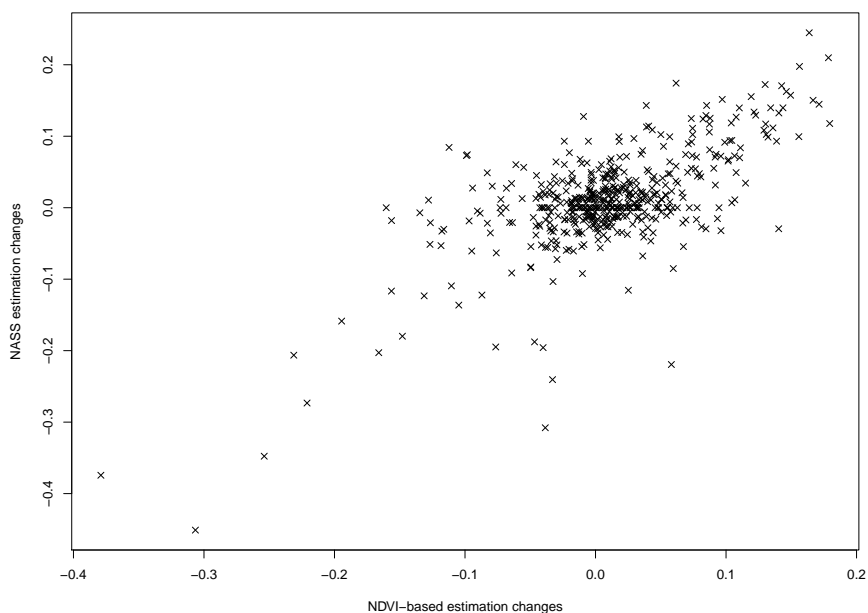


Figure 4.4 – NASS yield estimation changes to NDVI-based yield estimation changes, August to October, 2000-2016

Although the NDVI-based information is significantly correlated to the early NASS yield estimation changes, the main goal of our study is to determine whether this correlation is strong enough for our NDVI-based yield estimation changes to be significantly correlated with the commodities market reactions. To assess this point we present the results of the Equation (4.15), together with the Kendall's τ , in Table 4.7. We remark that, when considering the three reports, the Pearson's correlation is significant at a 0.2% level. Furthermore, the $\hat{\beta}_1$ estimate is negative (-5), and therefore in line with the expectations. However, when focusing on specific month, we note that for the September one, no significant estimation emerge, the estimate even appear to be positive. This poor performance on the September reports is not completely a surprise. Indeed, these releases lead to the less important, but still statistically significant, market reactions quantified through the variance or regression analysis (cf. Table 4.3 and 4.4 and Figure 4.2). Kendall's τ study also indicates that the correlation is negative and significant for the August and October reports, while, for the September release, no significance arises. However, and on the contrary to the linear model, the rank correlation is not statistically significant when considering all reports together.

4.5 Perspectives

The results we present in the current paper are promising, however we believe that the methodology needs to be enhanced before a concrete and effective application on the commodities

Table 4.7 – Regression results of future prices impacts to early NDVI-based corn yield estimate changes, and their Kendall τ , August to October, 2000-2016

| Reports | $\hat{\beta}_0$ | | $\hat{\beta}_1$ | | Kendall τ | |
|-------------|-----------------|-----------------|-----------------|-----------------|----------------|-----------------|
| | Estimates | <i>p</i> -value | Estimates | <i>p</i> -value | Estimates | <i>p</i> -value |
| August | 0.841 | 9.10^{-5} | -6.828 | 0.002 | -0.12 | 0.023 |
| September | -0.003 | 0.815 | 1.352 | 0.550 | 0.049 | 0.354 |
| October | -0.400 | 0.054 | -27.25 | 1.10^{-4} | -0.17 | 0.002 |
| All reports | 0.127 | 0.233 | -5.001 | 0.002 | -0.012 | 0.689 |

market. Our statistical study aims at highlighting evidences that there is a possibility to forecast some of the valuable information contained in the USDA reports through the use of satellite data. These first findings should incite the private sector to invest in the development of similar methodologies for their risk management. We therefore propose some improvements that should be further investigate for a practical implementation.

First, we have only used NDVI to obtain our early yield estimates. To improve the accuracy of our model, a relative easy upgrade would be to also consider weather data (Mkhabela et al., 2011), or other vegetation index such as Enhanced Vegetation Index (Johnson et al., 2016). The use of a larger period to train the models could also be useful. In the current paper, we focus on the MODIS data, i.e. from 2000, while older data sources exist such as the AVHRR. However, the instruments and resolution being different, data first needs to be reprocessed (Pedelty et al., 2007). This issue will surely raise again in the future with the continuous enhancement of the satellite resolution (Skakun et al., 2018). More sophisticated data upgrades can be applied. Indeed, even though we apply a crop mask (MOD 12) to the MODIS NDVI data, we still conduct our study at a state level. Yet, many different crop types can be grown in a single state. It, therefore, leads to interferences between crops in the resulting NDVI time series. For example, Figure 4.5 displays the MODIS NDVI time series over the year 2016 in Kansas. We observe two peaks: one around the period 9 (mid May) and one around the period 15 (mid August). This results from the fact that Kansas is an important winter wheat producer (first peak), but also of corn and soybeans (second peak), as shown in Table 4.6. Recent studies highlight the possibility to elaborate a crop mapping thanks to remote sensing methods (Skakun et al., 2017a,b). Applying such a map should result to a better yield forecasting accuracy (Maselli and Rembold, 2001; Kastens et al., 2005; Zhang et al., 2019). Furthermore, if the map can be estimated early enough, it could also be used to forecast another valuable information from the USDA reports: the planted area. Indeed, the value to the market participants of such information is estimated to \$145 million (Abbott et al., 2016).

The multiple crops problematic raises another limit of our paper: we focus on the corn future market only. The WASDE reports also contain valuable information about other agricultural commodities such as soybeans, wheat, barley, rice, etc. Moreover, as the "W" suggests, the

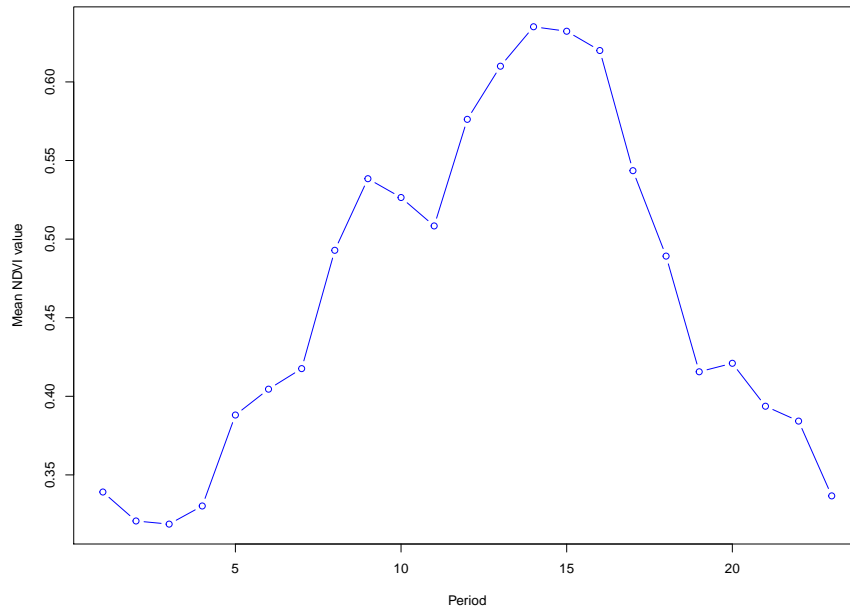


Figure 4.5 – MODIS NDVI time series of Kansas over the year 2016.

WASDE reports provide overviews of the supply and demand for these crops all around the world. For example, a careful attention is given to European (European Union, Ukraine and Russia) wheat market or Brazilian soybeans market. The agricultural commodities we focus on being storable, the different financial markets are linked one to another. Hence, to improve the forecasting performance of the market reactions to the WASDE releases, one would need to develop a more global model, in term of both crop type and localisation. Since the MODIS data cover the entire world, no major data problems should be expected.

Thirdly, in this research we focus on the public information contained in the USDA. More specifically, we evaluate the market reaction to the reestimation of yields as if no other information sources, public or private, were considered as newsworthy by the market participants. However, it is known that other governmental reports, published between two WASDE ones, contain valuable information on the crop conditions, notably the Crop Progress reports (Lehecka, 2014). In addition, many market participants have access to private forecasts on the future agricultural supply and demand (Karali et al., 2019). Therefore, it would be beneficial to take into account other sources of information in the modelling of the commodities market reaction to the WASDE release.

Last but not least, we highlight significant and rational correlations in our results. For a practical implementation in the commodities market, the general idea of the current paper has to be converted into a trading strategy. Moreover, the performance should be assessed through a backtesting methodology before being put into practice. This method is an out-sample one,

contrary to the study we conduct here. Thus, overfitting mentioned in Section 4.3.3 is no longer an issue. Hence, to increase the accuracy of the governmental valuable information forecasts, one should train the NDVI-based early yield model (Equation (4.10)) not with the NASS final yield estimates $Y_{k,N}$, but rather with the corresponding NASS early yield estimates $Y_{k,i,N}$.

4.6 Conclusion

In this paper, we retrieve, over the period 2000-2016, the well-known result that corn futures market reacts to the WASDE report releases. Then, we evaluate the informational value contained in the early yield estimations from the NASS, published together with the WASDE. We find out that the changes in these estimations are significantly and rationally correlated to the financial returns of the corn futures. Then, we propose an econometric model to obtain early yield forecasts based on MODIS NDVI time series, available around two weeks before the actual publication of the WASDE reports. Finally, we show that changes between two of the NDVI-based early yield estimates are also significantly and rationally correlated to the corn future returns. Therefore, it seems possible to forecast some valuable information from the USDA announcements thanks to MODIS NDVI data.

Even though we do not pursue our work on this subject, our study can inspire commodity traders or agricultural insurance companies to optimize their financial and risk management strategies. Indeed, the value of the information is important for the market participants. Moreover, since the databases are provided in open access, the data acquisition cost can be considered as null. Hence, the incentives seem high enough for the private sector to invest in developing remote sensing tools for the commodities price risk management. Indeed, the approach presented in the current paper is only to be seen as a proof of concept rather than a directly useable methodology. In this perspective, we have proposed some improvements that one would need to focus on before putting the method into practice.

Nevertheless, giving that the data is provided by the NASA, i.e. a reliable governmental institution, the derived information can be considered as public. Hence, in the long term, if all market participants have developed such methodology, it may decrease the impact of the USDA reports.

Bibliography

- Philip Abbott, David Boussios, and Jess Lowenberg DeBoer. Valuing Public Information in Agricultural Commodity Markets: WASDE Corn Reports. In *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management.*, St. Louis, MO, 2016.
- I. Becker-Reshef, C. Justice, M. Sullivan, E. Vermote, and C. Tucker. The Global Agriculture Monitoring (GLAM) project. *Remote Sensing*, 2(6):1589–1609, 2010a.
- I. Becker-Reshef, E. Vermote, M. Lindeman, and C. Justice. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment*, 114(6):1312–1323, 2010b.
- R. Colwell. Determining the prevalence of certain cereal crop diseases by means of aerial photography. *Hilgardia*, 26(5):223–286, November 1956. ISSN 0073-2230. doi: 10.3733/hilg.v26n05p223.
- Paul C. Doraiswamy, Thomas R. Sinclair, Steven Hollinger, Bakhyt Akhmedov, Alan Stern, and John Prueger. Application of MODIS derived parameters for regional crop yield assessment. *Remote Sensing of Environment*, 97(2):192–202, July 2005. ISSN 0034-4257. doi: 10.1016/j.rse.2005.03.015.
- Jeffrey H. Dorfman and Berna Karali. A Nonparametric Search for Information Effects from USDA Reports. *Journal of Agricultural and Resource Economics*, 40(1):124–143, 2015. ISSN 1068-5502.
- C. Ferencz, P. Bognár, J. Lichtenberger, D. Hamar, Gy Tarcsai†, G. Timár, G. Molnár, SZ Pásztor, P. Steinbach, B. Székely, O. E. Ferencz, and I. Ferencz-Árkos. Crop yield estimation by satellite remote sensing. *International Journal of Remote Sensing*, 25(20): 4113–4149, October 2004. ISSN 0143-1161. doi: 10.1080/01431160410001698870.
- T.R. Fortenbery and D. A. Sumner. The effects of USDA reports in futures and options markets. *Journal of Futures Markets*, 13:157–173, 1993.
- Feng Gao, Martha C. Anderson, Xiaoyang Zhang, Zhengwei Yang, Joseph G. Alfieri, William P. Kustas, Rick Mueller, David M. Johnson, and John H. Prueger. Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery. *Remote Sensing of Environment*, 188:9–25, January 2017. ISSN 0034-4257. doi: 10.1016/j.rse.2016.11.004.
- Philip Garcia, Scott H. Irwin, Raymond M. Leuthold, and Li Yang. The value of public information in commodity futures markets. *Journal of Economic Behavior & Organization*, 32(4):559–570, April 1997. ISSN 0167-2681. doi: 10.1016/S0167-2681(97)00013-9.

- Christophe Gouel. The Value of Public Information in Storable Commodity Markets: Application to the Soybean Market. In *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, Minneapolis, Minnesota, 2018.
- Scott H. Irwin, Darrel L. Good, and Jennifer K. Gomez. The Value of USDA Outlook Information: An Investigation using event Study Analysis. In *NCR Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, St. Louis MO., 2001.
- Olga Isengildina-Massa, Scott H. Irwin, Darrel L. Good, and Jennifer K. Gomez. The Impact of Situation and Outlook Information in Corn and Soybean Futures Markets: Evidence from WASDE Reports. *Journal of Agricultural and Applied Economics*, 40(1):89–103, April 2008. ISSN 1074-0708, 2056-7405. doi: 10.1017/S1074070800027991.
- Michael D. Johnson, William W. Hsieh, Alex J. Cannon, Andrew Davidson, and Frédéric Bédard. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, 218-219:74–84, March 2016. ISSN 0168-1923. doi: 10.1016/j.agrformet.2015.11.003.
- Berna Karali, Olga Isengildina-Massa, Scott H. Irwin, Michael K. Adjemian, and Robert Johansson. Are USDA reports still news to changing crop markets? *Food Policy*, March 2019. ISSN 0306-9192. doi: 10.1016/j.foodpol.2019.02.005.
- Jude H. Kastens, Terry L. Kastens, Dietrich L. A. Kastens, Kevin P. Price, Edward A. Martinko, and Re-Yang Lee. Image masking for crop yield forecasting using AVHRR NDVI time series imagery. *Remote Sensing of Environment*, 99(3):341–356, November 2005. ISSN 0034-4257. doi: 10.1016/j.rse.2005.09.010.
- Sanghyo Kim, Carl Zulauf, Matthew Roberts, and Kevin Cook. Performance of 5-Year Olympic Moving Average in Forecasting U.S. Crop Year Revenue for Program Crops. In *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, St. Louis MO., 2015.
- R. Kumar and L. Silva. Light Ray Tracing Through a Leaf Cross Section. *Appl. Opt., AO*, 12(12):2950–2954, December 1973. ISSN 2155-3165. doi: 10.1364/AO.12.002950.
- Georg V. Lehecka. The Value of USDA Crop Progress and Condition Information: Reactions of Corn and Soybean Futures Markets. *Journal of Agricultural and Resource Economics*, 39(1):88–105, 2014. ISSN 1068-5502.
- Ainong Li, Shunlin Liang, Angsheng Wang, and Jun Qin. Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. *Photogrammetric Engineering & Remote Sensing*, 73(10):1149–1157, October 2007. doi: 10.14358/PERS.73.10.1149.

- B. L. Ma, Lianne M. Dwyer, Carlos Costa, Elroy R. Cober, and Malcolm J. Morrison. Early Prediction of Soybean Yield from Canopy Reflectance Measurements. *Agronomy Journal*, 93(6):1227–1234, November 2001. ISSN 1435-0645. doi: 10.2134/agronj2001.1227.
- Fabio Maselli and Felix Rembold. Analysis of GAC NDVI Data for Cropland Identification and Yield Forecasting in Mediterranean African Countries. *Photogrammetric Engineering & Remote Sensing*, 67(5):593–602, 2001.
- Andrew M. McKenzie. Pre-Harvest Price Expectations for Corn: The Information Content of USDA Reports and New Crop Futures. *American Journal of Agricultural Economics*, 90(2):351–366, May 2008. ISSN 0002-9092. doi: 10.1111/j.1467-8276.2007.01117.x.
- Trent T. Milacek and B. Wade Brorsen. Trading Based on Knowing the WASDE Report in Advance. *Journal of Agricultural and Applied Economics*, 49(3):400–415, August 2017. ISSN 1074-0708, 2056-7405. doi: 10.1017/aae.2017.8.
- M. S. Mkhabela, P. Bullock, S. Raj, S. Wang, and Y. Yang. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, 151(3):385–393, March 2011. ISSN 0168-1923. doi: 10.1016/j.agrformet.2010.11.012.
- Jeffrey Pedelty, Sadashiva Devadiga, Edward Masuoka, Molly Brown, Jorge Pinzon, Compton Tucker, Eric Vermote, Stephen Prince, Jyotheshwar Nagol, Christopher Justice, David Roy, Junchang Ju, Crystal Schaaf, Jicheng Liu, Jeffrey Privette, and Ana Pinheiro. Generating a long-term land data record from the AVHRR and MODIS Instruments. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 1021–1025, Barcelona, Spain, 2007. IEEE. ISBN 978-1-4244-1211-2. doi: 10.1109/IGARSS.2007.4422974.
- Md Shahriar Pervez and Jesslyn F. Brown. Mapping Irrigated Lands at 250-m Scale by Merging MODIS Data and National Agricultural Statistics. *Remote Sensing*, 2(10):2388–2412, October 2010. doi: 10.3390/rs2102388.
- Dana Peterson, Jerry Whistler, Stephen Egbert, and J Christopher Brown. Mapping Irrigated Lands by Crop Type in Kansas. *Pecora 18-Forty Years of Earth Observation... Understanding a Changing World*, pages 14–17, 2011.
- Anup K. Prasad, Lim Chai, Ramesh P. Singh, and Menas Kafatos. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33, January 2006. ISSN 0303-2434. doi: 10.1016/j.jag.2005.06.002.
- M. S. Rasmussen. Operational yield forecast using AVHRR NDVI data: Reduction of environmental and inter-annual variability. *International Journal of Remote Sensing*, 18(5):1059–1077, March 1997. ISSN 0143-1161. doi: 10.1080/014311697218575.

- Jianqiang Ren, Zhongxin Chen, Qingbo Zhou, and Huajun Tang. Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation*, 10(4):403–413, December 2008. ISSN 0303-2434. doi: 10.1016/j.jag.2007.11.003.
- J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains with ERTS. In *Proceedings 3rd Earth Resources Technology Satellite (ERTS) symposium*, volume 1, pages 309–317, Washington, DC, USA, January 1974. NASA.
- Sergii Skakun, Belen Franch, Eric Vermote, Jean-Claude Roger, Inbal Becker-Reshef, Christopher Justice, and Nataliia Kussul. Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment*, 195:244–258, June 2017a. ISSN 00344257. doi: 10.1016/j.rse.2017.04.026.
- Sergii Skakun, Eric Vermote, Jean-Claude Roger, and Franch Belen. Combined use of Landsat-8 and Sentinel-2a images for winter crop mapping and winter wheat yield assessment at regional scale. *AIMS Geoscience*, 3(2):163–186, 2017b. doi: 10.3934/geosci.2017.2.163.
- Sergii Skakun, Christopher O. Justice, Eric Vermote, and Jean-Claude Roger. Transitioning from MODIS to VIIRS: an analysis of inter-consistency of NDVI data sets for agricultural monitoring. *International Journal of Remote Sensing*, 39(4):971–992, February 2018. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2017.1395970.
- Daniel A. Sumner and Rolf A. E. Mueller. Are Harvest Forecasts News? USDA Announcements and Futures Market Reactions. *American Journal of Agricultural Economics*, 71(1): 1–8, February 1989. ISSN 0002-9092. doi: 10.2307/1241769.
- E. Svatwa, A. J. Masuka, B. Maasdorp, A. Murwira, and M. Shamudzarira. Use of multi-spectral radiometer and MODIS satellite data correlations in developing models for tobacco yield forecasting. *International Journal of Agricultural Research*, 3(1):147–154, 2014.
- Jesse Tack, Keith Coble, Robert Johansson, Ardian Harri, and Barry Barnett. The Potential Implications of 'Big Ag Data' for USDA Forecasts. SSRN Scholarly Paper ID 2909215, Social Science Research Network, Rochester, NY, January 2017.
- Brian D. Wardlow and Stephen L. Egbert. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sensing of Environment*, 112(3):1096–1116, March 2008. ISSN 0034-4257. doi: 10.1016/j.rse.2007.07.019.
- Brian D. Wardlow and Stephen L. Egbert. A comparison of MODIS 250-m EVI and NDVI data for crop mapping: a case study for southwest Kansas. *International Journal of Remote Sensing*, 31(3):805–830, February 2010. ISSN 0143-1161. doi: 10.1080/01431160902897858.

- A. K. Whitcraft, E. Vermote, I. Becker-Reshef, and C. Justice. Cloud cover throughout the agricultural growing season: Impacts on passive optical earth observations. *Remote Sensing of Environment*, 156:438–447, 2015. doi: <https://doi.org/10.1016/j.rse.2014.10.009>.
- Jiahui Ying, Yu Chen, and Jeffrey H. Dorfman. Is the Value of USDA Annoucement Effects Declining over Time ? In *NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, St. Louis MO., 2017.
- Hongwei Zhang, Huailiang Chen, and Guanhui Zhou. The Model of Wheat Yield Forecast Based on Modis-Ndvi - A Case Study of Xinxiang. *ISPREs Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1-7, 2012.
- Yinsuo Zhang, Aston Chipanshi, Bahram Daneshfar, Lauren Koiter, Catherine Champagne, Andrew Davidson, Gordon Reichert, and Frédéric Bédard. Effect of using crop specific masks on earth observation based crop yield forecasting across Canada. *Remote Sensing Applications: Society and Environment*, 13:121–137, January 2019. ISSN 2352-9385. doi: [10.1016/j.rsase.2018.10.002](https://doi.org/10.1016/j.rsase.2018.10.002).
- Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 221:430–443, February 2019. ISSN 0034-4257. doi: [10.1016/j.rse.2018.11.032](https://doi.org/10.1016/j.rse.2018.11.032).

Conclusion

L'objet de notre conclusion n'est pas de détailler une nouvelle fois les contributions issues de cette thèse. En effet ces dernières ont déjà été abordées dans le résumé global, l'introduction générale et dans les conclusions de chaque chapitre. Nous proposons dans cette conclusion générale quelques perspectives de recherche qui s'inscrivent dans la continuité des travaux présentés.

Dans le Chapitre 1, nous appliquons une régularisation elastic-net lors de l'estimation du modèle VAR dans une optique de projection de la mortalité. La problématique de grande dimension est avant tout d'éviter le sur-apprentissage. Néanmoins, l'obtention de matrices autorégressives parcimonieuses, nous permet aussi de détecter des effets démographiques connus, tels que les effets période ou cohorte, par les structures des coefficients ainsi estimés. Les matrices deviennent alors une signature de la dynamique des taux de mortalité et rendent possible les comparaisons entre populations. La parcimonie impliquée par la pénalisation lasso peut alors être vu comme un outil de classification non-supervisée plutôt que de régression et de projection.

En particulier, la parcimonie des matrices met empiriquement en avant des structures verticales de coefficients ne correspondant pas à des effets démographiques qui nous sont connus. Ces motifs sont observés, plus ou moins nettement, dans plusieurs populations de la Human Mortality Database. L'application pratique du modèle sur cette importante base d'information nous porte à supposer qu'il puisse s'agir d'anomalies dans les données que notre modèle VAR-ENET permettrait de mettre en évidence. Une étude plus approfondie, se fondant notamment sur des données simulées, conforterait cette hypothèse.

La problématique de classification de populations en fonction de la dynamique des taux de mortalité est aussi un sujet de préoccupation du Chapitre 2. En effet, dans ce dernier nous regroupons les populations de manière à créer des groupe de cohérence. Bien que nous ne nous attardons pas beaucoup sur cette classification dans cette thèse, notre contribution reposant essentiellement sur l'introduction de la notion de cohérence locale et ses implications, ce problème n'en reste pas moins attrayant. Les futures recherches se devront de développer des méthodes de classification de séries temporelles adaptées à la détermination des groupes de cohérence.

Enfin, dans le Chapitre 3, nous avons effectué un nombre important d'hypothèses sur les paramètres du modèle économique permettant d'évaluer le gain de la campagne de rétention en terme de Customer lifetime value. Néanmoins, dans la pratique, l'entreprise ne connaît pas exactement ces derniers, ce qui provoque de l'incertitude sur le profit attendu. Or, nous proposons une méthode qui optimise ce profit, et non la précision statistique, à l'aide d'une fonction de perte spécifique se fondant justement sur ces paramètres. L'interrogation réside alors dans la robustesse de cette approche en fonction du degré d'incertitude. En d'autre terme, il sera bon d'étudier à quel point devons nous personnaliser la fonction de perte selon la précision de l'estimation des paramètres du modèle économique.